Towards a Web of Structured Knowledge: Methods, Applications and Perspectives

Konstantin Todorov

UNIVERSITY OF MONTPELLIER

HABILITATION À DIRIGER DES RECHERCHES

Towards a Web of Structured Knowledge: Methods, Applications and Perspectives

---

Konstantin Todorov

Laboratoire d'Informatique, Robotique et Microélectornique de Montpellier
Group FADO
Montpellier, France

December 2, 2019

JURY

Maria-Esther VIDAL	Professor, Univ. Simon Bolivar (Venezuela) / TIB - Hannover (Germany)
Fabian SUCHANEK	Professor, Télécom Paris University
Jérôme EUZENAT	Research Director, INRIA Grenoble Rhone-Alpes
Catherine FARON-ZUCKER	Assoc. Professor, HDR, University of Nice Sophia Antipolis
Nathalie PERNELLE	Assoc. Professor, HDR, University Paris Sud, University Paris Saclay
Pascal PONCELET	Professor, University of Montpellier

Acknowledgements

I would like to thank the members of the jury and reviewers of this thesis, Maria-Esther Vidal, Jérôme Euzenat, Fabian Suchanek, Catherine Faron-Zucker, Nathalie Pernelle and Pascal Poncelet for taking from their time to read my work and report on it and for their participation in the jury. I am honored to be given the possibility to present this work in front of them.

I thank very warmly all my colleagues that I had the privilege and pleasure to work and publish with over the years. The full list being very long, I refer here to my main co-authors: in the first place my PhD thesis supervisors Peter Geibel and Kai-Uwe Kühnberger, followed by my postdoc project leader Céline Hudelot and my colleagues Stefan Dietze, Zohra Bellahsene, Pavlos Fafalios, Danai Symeonidou, Pasquale Lisena, Nicolas James, Hoa Duy Ngo, Elena Demidova, Andon Tchechmedjiev, Francois Scharffe, Raphaël Troncy, Katarina Boland. I would particularly like to thank the students who worked under my supervision during their PhD or masters theses and made possible the publication of a number of joint works: Abdelnacer Tigrine, Mohamed Ben Ellefi, Manel Achichi, Houssammedine Farah, Théophile Mandon, Mikael Vygo, Malo Gasquet, Darlene Brechtel, Imène Chentli.

I am equally thankful to all my colleagues at LIRMM and other research institutes in Montpellier such as CIRAD, INRA, IRD and IGH. In particular, I thank Clément Jonquet, Pierre Larmande, Mathieu Roche, Anne Laurent, Mathieu Lafourcade, Dino Ienco, Sofia Kossida, Madalina Croitoru, Marie-Laure Mugnier, Sandra Bringay, Michel Chein and many others with whom I had the occasion to co-supervise master students, submit joint grant applications or have a nice chat during a coffee break.

I also thank all my colleagues at the Computer Science department at the University of Montpellier and particularly those with whom I had the pleasure to work on various pedagogical projects: Federico Ulliana, Annie Chateau, Pierre Pompidor, Isabelle Mougenot, Marie-Louise Zinsou, Marianne Huchard, Sylvain Daudé, William Puech, Christophe Dony, Anne-Muriel Chifolleau, Michel Leclère. My gratitude also goes to all project partners from various academic and non-academic institutions from France and Europe with whom I had the occasion to work. And finally, I thank all bachelor and master students that have attended my courses over the past years from whom I was able to learn a lot.

Contents

Acknowledgements iii

1 An Overview and a Guide to this Thesis 1

2 From Unstructured Data to Knowledge Graphs: A Map of My Research Activities 3
2.1 The Lifecycle of Linked Open Data . . . 3
2.2 Ontology Matching . . . 6
2.2.1 A Brief State-of-the-Art . . . 9
2.2.2 Instance-based Ontology Matching . . . 11
2.2.3 Ontology Matching with Background Knowledge as a Mediator 12
2.3 Data Linking and Knowledge Extraction for Data Linking Applications 15
2.3.1 A Brief State-of-the-Art on Data Linking . . . 16
Knowledge Graph Heterogeneities . . . 16
Automating the Data Linking Process . . . 17
2.3.2 Linking and Disambiguating Entities across Heterogeneous RDF Graphs . . . 19
2.3.3 RDF Dataset Profiling and Profile-based Dataset Recommendation for Data Linking . . . 21
2.3.4 Key Discovery and Ranking . . . 24
2.4 Data Modeling and Knowledge Graph Building . . . 26
2.4.1 Conceptual Models and Knowledge Graphs for Music Meta-Data . . . 27
2.4.2 Conceptual Models and Knowledge Graphs for Fact-checked Claims . . . 28

3 A Web of Linked Data and Beyond: Three Novel Research Axes 31
3.1 Towards a Bottom-up Data Linking Paradigm . . . 31
3.2 Data Linking in the Lack of Shared Context . . . 36
3.3 Combining Knowledge Graphs with Machine Learning for the Analysis of Online Discourse Data . . . 38

4 Five Selected Articles: Contributions, Impact and Collaborations 45
4.1 Data and Knowledge Integration . . . 45
4.2 Knowledge Extraction and Graph Profiling . . . 47

5 Conclusion 49

Bibliography 51

A Extended Curriculum Vitae 63

B Five Selected Publications 77

C Extended Abstract in French **205**

Chapter 1

An Overview and a Guide to this Thesis

The invention of the Web has created unprecedented opportunities to access the immense wealth of information shared on the Internet. This wealth, however, cannot be explored without means to filter the precious bits out of the noise, without search engines that are able to correctly interpret user information needs and return meaningful results to queries that may contain ambiguous entities. It has been for a long time now a goal of the computer science (CS) community to endow computers with *knowledge* and thus enable them to serve better in the information search process. The advent of the Semantic Web technologies together with advances in related fields such as statistical natural language processing (NLP) and machine learning (ML) are getting us closer today to that goal than we have ever been before [147]. The provision of rich knowledge graphs (KG)—those machine readable artefacts that structure the knowledge and underlying data in various domains of life—has shown to be the key brick towards the construction of a more (artificially) intelligent Web.

The building of such knowledge graphs is guided by a number of principles, known in the community under the acronym FAIR: produce and publish *findable*, *accessible*, *interoperable* and *reusable* data [149]. As we will see in the following chapter, the harvesting of knowledge and "encapsulating" it in FAIR knowledge graphs is a complex and challenging process, comprising several components, each of which constitutes a research area in its own right (Section 2.1, Chapter 2). These include information extraction on various levels (entities, relations, types, properties, etc.) from textual or semi-structured data, providing conceptual data models that describe the types of things there exist and how are they related, linking entities across knowledge graphs in order to enable federated search and data integration, ensuring maintenance and quality of the provided structured knowledge, providing tools for access to the data and, finally, using all that knowledge as a background for a number of (statistical or logical) inference tasks that will eventually lead to the discovery of more and new knowledge.

Since I started researching in the fields of Semantic Web and machine learning as a PhD student in 2005 up until now, I had the opportunity to work on a number of challenging issues along the knowledge harvesting and structuring pipeline, described above (I refer to it as the Web data lifecycle in the following chapter). My early contributions address a set of problems in the field of ontology matching (Section 2.2, Chapter 2). Guided by the need to provide methods for integration of ontologies that annotate textual content, I have introduced approaches for instance-based ontology matching using machine learning variable selection techniques. In order to reply to challenges related to the consideration of the inherent vagueness of ontological concepts and concept mappings, I have worked on the development of

approaches that rely on fuzzy set theory and background knowledge in the matching process with applications in the field of multimedia information retrieval and cross-lingual data integration, towards opening up the way to a multilingual Web of data.

Later on I took interest in the related problem of data linking (Section 2.3, Chapter 2), where in the context of a number of projects, I was able to provide both novel methodological solutions and tools. Taking a more holistic perspective to the data linking task, I have investigated the knowledge graph profiling problem with applications to the task of dataset recommendation – providing a ranked list of datasets that are likely to contain entities of interest for interlinking with a given input dataset. In an attempt to reduce the human user effort in the process of tuning a data linking tool, I have proposed approaches for selecting key properties that are valid on two Resource Description Framework (RDF) knowledge graphs simultaneously and ranking them with respect to their likelihood to produce a large number of links when used as parameters of established data linking tools.

Finally, I had the opportunity to steer the construction of several knowledge graphs in the fields of music and computational social science (fact-checking). Those projects included taking and tackling many of the above described challenges, as well as addressing a number of data modelling and related information extraction problems (Section 2.4, Chapter 2), as part of the KGs building process.

I have been asked to select five of my publications that cover important aspects of the research described above, listed and summarised in Chapter 4. These summaries, beyond the articles content, focus and detail on their current or expected impact, as well as the collaboration networks that have enabled their publication. Indeed, I was very lucky to be able to carry out my research in a close collaboration with a large number of colleagues and students from both computer science and other research fields (e.g. sociology or biology), both from within and outside academia. The acknowledgements section of this manuscript contains their names. I would like to underline here again that the achieved results are the fruits of many collaborative efforts, which only made their realisation possible.

To paraphrase Matthiew Walker,[1] every answer is a door to a new question. Building on the acquired experience, I am excited to pursue my endeavours in a number of novel research directions that I will describe in more detail in Chapter 3. In particular, I will address the challenging issue of going beyond a Web of documents and beyond a Web of structured knowledge towards a Web containing, in addition to documents and facts, structured knowledge about *claims* and other online discourse data, such as viewpoints and stances, controversial topics and related contextual information.

[1] A neuroscientist, author of the book “Why we sleep" – a great read.

Chapter 2

From Unstructured Data to Knowledge Graphs: A Map of My Research Activities

2.1 The Lifecycle of Linked Open Data

In the past years, we have witnessed a growing effort in building and publishing structured data on the Web in the form of graph-like datasets, also called knowledge graphs, often using RDF and Semantic Web formalisms and technologies for their construction and sharing. The most prominent example is the Linked Open Data (LOD) project, which currently gathers hundreds of datasets from different fields and domains of application, made openly accessible on the Web. Large knowledge graphs, such as the Google Knowledge Graph or the Knowledge Vault [29] have been devised in order to structure the information available on the Web, infer new knowledge and improve information retrieval and access. In the same time, initiatives such as schema.org[1] have been successfully put forward by W3C[2] working groups and private corporations (such as Google) in order to provide means for annotating Web content and making information easier to find for search engines [155]. We are witnessing a shift in the way of publishing and consuming data on the web, moving from unstructured information made available in a decentralised manner that is searched over by the help of keywords-like queries, towards a "novel" Web, which, without replacing the Web of today, extends it with an extra structured layer, resembling more and more an immense database, where question answering algorithms allow to retrieve precise answers to user queries, very much like in a relational database fashion (cf. Fig. 2.1).

Beyond ordinary Web users, experts from particular fields of science, culture or society at large benefit from these efforts. Computer science researchers work in close cooperation with various domain experts in order to foster the development of structured and openly accessible data in their respective fields, aiming to enhance and facilitate the experts access to data, allow them to uncover new knowledge and free them from the technological burden in that process. Examples are various initiatives in the fields of biomedicine [141], agronomy[3], or music [3], including projects like D2KAB,[4] DOREMUS[5] or ClaimsKG[6], to which I have or am currently contributing. The set of principles that guide the development of structured knowledge, its

[1] https://schema.org/
[2] The World Wide Web Consortium, https://www.w3.org/
[3] http://agrold.southgreen.fr/agrold/
[4] http://d2kab.mystrikingly.com/
[5] https://www.doremus.org/
[6] https://data.gesis.org/claimskg/site/#about

sharing and reuse have recently been known as *FAIR*, standing for Findable, Accessible, Interoperable and Reusable data [149].[7]

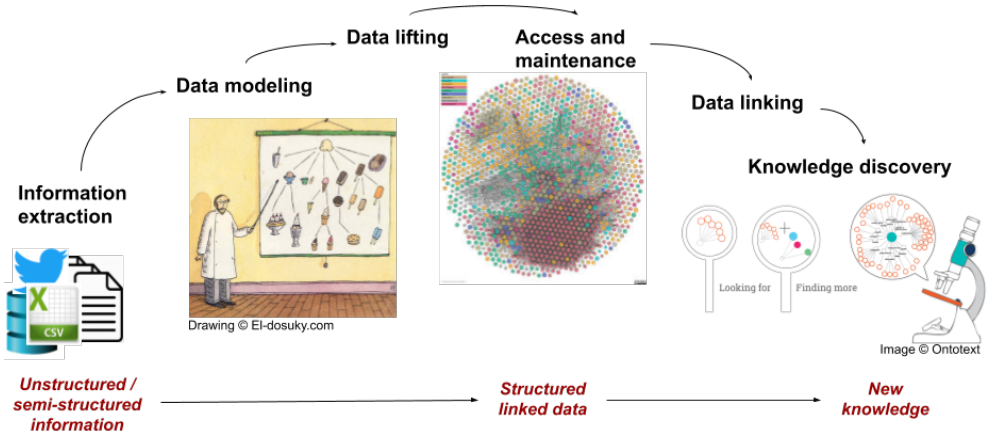


FIGURE 2.1: A perspective on the linked open data lifecycle.

Doing FAIR Web data opens numerous scientific challenges for artificial intelligence (AI) and CS research. My research activities since the beginning of my PhD thesis have been driven by the need to respond to some of these challenges. Therefore, I will proceed to explain in more detail the lifecycle of (FAIR) data in the process of their opening and publishing on the Web in the particular case of using as a background technological layer the Semantic Web technologies. Based on visions proposed in [48, 26, 142], or the LOD2 Linked Open Data Lifecycle introduced in [10], I summarise the main stages of publishing a new FAIR dataset as linked data as follows (also illustrated in Fig. 2.1):

- **Information extraction.** This component comprises extracting and transforming information (entities and relations) from the raw data source to RDF data. The original data source can be in unstructured, structured or semi-structured form, i.e., a CSV or an XML file, a relational database, a corpus of text documents (Web pages, social network documents like tweets). A number of techniques from natural language processing are applied in order to extract the salient entities and the relations that hold among them.

- **Data modelling.** The model, also called an ontology or a vocabulary, is the conceptual backbone that is used to describe and represent the data in a field of interest: what types of things there are and how they are related to one another, what are their properties. It is crucial that the final ontology reuses relevant existing vocabulary terms whenever applicable. Coming up with an appropriate data model for the domain of interest can be a particularly laborious and time consuming task since it implies the interactions between different communities, where CS and in particular Semantic Web experts work closely together with experts from a specific domain (e.g. cultural heritage or sociology, depending on the application).

- **Data lifting.** This step brings the outcomes of the first two steps together. This comprises assigning namespaces, or else—giving names—to all resources and properties in our RDF relational data according to the developed model,

[7] https://www.go-fair.org/fair-principles/

disambiguating and linking the named entities and making these resources accessible via unique identifiers (URIs). In other words, the entities and relations extracted in the first step will be now annotated by the concepts and properties names from our vocabulary developed in the second step.

- **Access and maintenance.** This component consists in hosting the linked dataset and its metadata publicly, making it accessible via SPARQL endpoints and, optionally, non-CS user-friendly tools for exploration and search. Guaranteeing the perseverance of the published resources is an important issue that has to be attended to.

- **Data linking.** This step is meant to fully unlock the potential of the linked data project by allowing the federated access to data elements distributed across various sources. It involves the (automatic) linking of the newly published knowledge graph to other datasets already published as linked data on the Web by establishing typed relations between resources.

- **Knowledge discovery.** Finally, the linked knowledge graph can be used as background knowledge for the discovery of new relations, links and new knowledge by the help of machine learning (link prediction, clustering or pattern mining) algorithms or symbolic inference.

These different stages are not in a strict sequence nor do they exist in isolation, but are in a process of mutual enrichment. Each of these steps raises a set of specific challenges. Extracting entities and relations from unstructured (web pages, social networks) or semi-structured (Wikipedia) sources in order to populate and build knowledge bases and thus provide structured knowledge on the Web has been of central interest in the NLP, the Web data mining and the Semantic Web communities over the past decades, focusing on a variety of tasks such as named entity recognition, entity linking, relation extraction or word sense disambiguation. The extensive research in this field has lead to a very broad range of works, surveyed, for example, in [41, 71, 97, 147].

The data *modeling* process is challenging in that it requires a significant effort from both metadata designers *and* domain experts in order to, in the first place, come up with an appropriate conceptual model of the field of interest, and then address issues raised by the need to identify suitable terms from already existing vocabularies in order to reuse them following the Linked data modeling best practices [15]. For example, if one is modeling data about political figures' claims, one may want to reuse properties from the schema.org *ClaimReview* class.

Further challenges include semantic links discovery between different RDF graphs, both in terms of entities (known as *data linking*) and schema (what we call *ontology matching*), which is manually unfeasible, considering the scale of the problem. Usually, among the different kinds of semantic links that can be established, one is particularly interested in that of identity stating that two different URIs (identifying ontological concepts or knowledge graph entities) refer to the same real object or group of objects. For example, DBpedia uses the URI `http://dbpedia.org/resource/Montpellier` to identify the city of *Montpellier*, while Geonames uses the URI `http://www.geonames.org/2992166` for the same entity and the two are described differently in both resources, by the help of different attributes, labels and sometimes even languages. Data linking or ontology matching techniques and tools are being developed in order to deal with this problem automatically, as surveyed in [76, 111]. However, these tasks still require human involvement, notably in the

laborious task of instance matching tools configuration, or that of the identification of candidate datasets to link to, where the search for target ones should be done almost by an exhaustive search of all datasets in the different catalogues. The latter challenges stand in front of the user even before the data linking process begins.

Due to the quest to automate as much as possible the linked data production process, many graphs available out there are a result of the application of automatic tools for knowledge extraction and data lifting or linking. While this has facilitated the publication of a large number of linked datasets on the Web, automatic approaches have raised many questions regarding the quality, the currentness and the completeness of the contained information. Hence, the major challenge at the outcome of this process concerns the assessment of the data quality [12]. In that respect, several issues arise after publishing linked data. On the one hand, data publishers bear the responsibility to ensure a continued maintenance of the published dataset in terms of quality, i.e., access, dynamicity (different versions, mirrors), and other. On the other hand, from the linked data consumers side, there is a need for ensuring continued feedback for data maintainers. In those terms, approaches to create automatically profiles of knowledge graphs and linked datasets appear as an important endeavour towards acquiring a better understanding of the nature and status of a given dataset with potential to enhance discoverability and reuse.

Finally, once the knowledge graphs are built, of attested quality, published and linked to other data resources—in other words they are truly FAIR—, their true potential in service of domain experts or scientists can be unfolded. For that reason, the Web and knowledge community has to ensure that these non-CS users have the appropriate tools for data access, exploration, analysis and retrieval. This implies the development of user-friendly environments that do not require the knowledge of formal query languages in order to access the data. And ultimately, thanks to reasoning or machine learning techniques (such as pattern mining, clustering or relation prediction) one is given the possibility to discover new knowledge by using the currently available linked data resources as background knowledge in that process [140, 139].

In the following sections, I will focus in more detail on several of the above mentioned challenges that constitute some of my research contributions over the past decade. As shown in Fig. 3.2, I had the opportunity to contribute to most of the stages of the linked data lifecycle, including knowledge extraction and data harmonisation, data integration (schema and instance matching) and knowledge graph building. This research has been carried out in collaboration with a number of academic and non-academic partners from the field of computer science and other fields, such as (computational) sociology or biology with applications in the fields of cultural heritage, human genetics, plant biology and social sciences, within the framework of a number of national and international research projects. I refer the reader to the “Publication strategy" section in my extended CV (see Appendix A) for more detailed description of my projects and cross-community interactions.

I will now open the "map" of my research contributions, presenting the different topics, which I had the occasion and pleasure to work on, by roughly following the order in which I took interest in them.

2.2 Ontology Matching

Conceptual models of data, also known as ontologies, schemata or vocabularies, describe what exists and is of interest for a particular application in a given domain

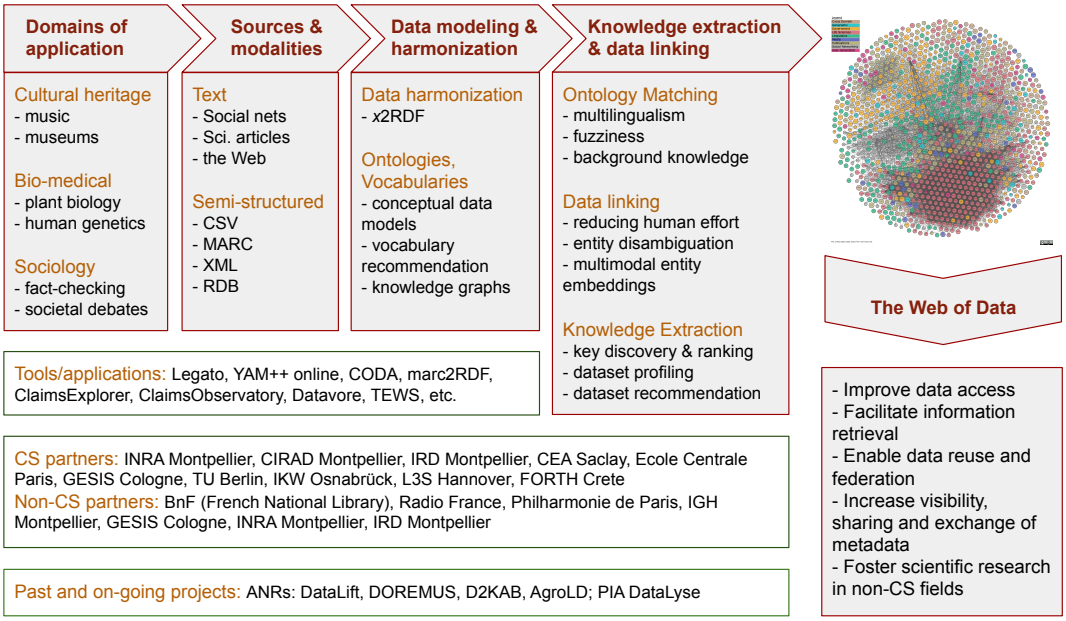


FIGURE 2.2: Doing linked Web data: a synthesis of my research activities.

in a shared and formal manner, enabling a common understanding of the concepts of importance in that domain as well as their properties and relations. The process of development of data models of this kind since the early years of the Semantic Web has been and largely remains decentralised and biased by specific individual or inter-community understandings of these domains. This process is also highly dependent on the perspective, with which one regards a particular domain—for example talking about biology from a scientific or librarian perspective would yield two different definitions of the concepts of interest and their relations. These phenomena have lead to the creation of vocabularies that describe identical, highly similar or intersecting domains of interest by using different conceptualisations in terms of terminology, structure and semantics. We refer to any difference in the description of a given set of concepts across two or more ontologies as *ontology heterogeneity*. Examples of such heterogeneities are the use of different terms to label the same concept or structuring concepts differently across two schemata in terms of how they are related to one another or what sets of properties describe them. The Semantic Web field of ontology matching or alignment, or also schema matching or reconciliation, has taken the challenge of providing approaches and tools to automatically deal with the heterogeneities that two or more vocabularies manifest by establishing typed relations between cross-ontology concepts [111, 109],[8] with the underlying understanding that heterogeneous ontologies can and should (co-)exist peacefully. These approaches promise to open the way to data integration and sharing across wider communities and to enable federated access to a wealth of distributed sources.

Ontology matching was the central research and application context of my PhD thesis. I proceed to provide a definition of the problem before introducing my own contributions. The formal definition of an ontology that I will use extends the one given in [116].

[8]Equivalence relations are of prevailing interest in the majority of research into the question, although for completeness note that one may be interested in establishing relations of other types (e.g. disjointness or subsumption).

Definition 1 (Ontology) *Let*

- *C be a set of concepts,*
- *$is\_a \subseteq C \times C$,*
- *R be a set of relations on C,*
- *I be a set of instances,*
- *and $g : C \rightarrow 2^I$ be a function that assigns subsets of instances from I to each concept in C.*

We require that is_a and g are compatible, i.e., that

$$is\_a(A, A') \leftrightarrow g(A) \subseteq g(A')$$

holds for all $A, A' \in C$. In particular, this entails that is_a has to be a partial order. A terminological knowledge component is defined as the triple $< C, is\_a, R >$ and a factual knowledge component as the couple $< I, g >$. With these definitions, the pair

$$O = (< C, is\_a, R >, < I, g >)$$

forms an ontology.

In this way, a concept is *intensionally* modeled by its relations to other concepts, and *extensionally* by a set of *instances* assigned to it via the function g. The partial order provides the hierarchical backbone of the ontology and is an essential part of its semantic structure. In R, we can have additional semantic concept-level relations like meronymy, antonymy, similarity, and also meta-level domain relations. A concept instance can be any entity, document, structure or, generally, piece of data annotated by this concept [116].

A number of different ways to formally define the ontology matching process have been proposed [109, 156]. In my work, I have been mainly[9] interested in establishing relations of equivalence between cross-ontology concepts. I will call *an ontology matching procedure* the automatic process that allows to discover such relations. In that context, an ontology matching procedure requires a concept similarity measure (that can be an aggregation of several measures) and an alignment algorithm which applies this measure for two ontologies by taking into account optionally additional meta data and their structures. The concept similarity measure assesses the extent to which two ontology entities are related via a particular relation of interest (for example, equivalence or subsumption). Depending on the application field and the kind of ontologies that we are working with, the concept similarity measure can be based on different characteristics, such as linguistic, structural, instance-based or other [33]. Optionally, background knowledge in the form of reference vocabularies, dictionaries, multilingual knowledge bases or web-corpora can be called upon in that process, just as the human user can be involved by requiring their feedback.

[9] We will see that in one particular application I have looked into hierarchical relations, as well.

2.2.1 A Brief State-of-the-Art

Ontology heterogeneity[10] occurs when two or more ontologies are created independently from one another over similar domains. Heterogeneity may be observed on *linguistic or terminological* level (concepts which represent the same real world entities but have different names), on *conceptual* level (mismatches in level of detail, coverage or scope) [33] or on *extensional* level (differences in the population). Whenever heterogeneity of any of these kinds is observed over a set of ontologies, these ontologies will be referred to as *heterogeneous*.

Different matching techniques have been introduced in the past years in order to resolve different types of heterogeneity relying on methods coming from fields as various as machine learning, NLP, graph-theory, relational algebra and logics [40].

Terminological methods comprise two major groups of approaches: those that use strings in order to match names of entities, and those that rely on linguistic information contained in dictionaries and thesauri combined with techniques from natural language processing in order to compare the similarity of terms and their relations and overcome problems evolving from synonymy and polysemy. Terminological methods for ontology alignement have been studying in more detail in [78].

The *structure* of an ontology can be taken into account in order to define a similarity measure. This can be done on two levels with respect to either a single ontology element, known as internal structure, or the way in which a set of elements are related, known as relational structure. Methods based on the former structure type look into similarities of the sets of properties of two elements, the datatypes used to describe them or their properties, the cardinalities that sets of values of two properties are allowed to reach, etc. These approaches are suited to schema matching problems where one disposes readily with an internal structure of the database entities. Relational structure approaches are concerned with comparing blocks of elements together with the relations that hold between them. Structure and terminology are often used in a combined manner as this has been done in [70, 72, 90].

Instance-based, or *extensional ontology matching* is based on the idea that ontology concepts can be represented as sets of related instances and the similarity measured on these sets of instances reflects the semantic similarity between the concepts that these instances populate. In particular, these instance sets together reflect structural and terminological differences and similarities [49]. Among the basic assumptions of such approaches is that two ontologies use the same instances and when this is not so, mechanisms for extracting instances (from text corpora or other external sources) should be made available, as in [116]. Other techniques rely on estimating the concept similarity by measuring class-means distances [60] or estimating joint probabilities by the help of machine learning techniques [27]. A matching approach, which does not rely on intersections of instance sets, nor on the estimation of joint probabilities is proposed in [127]. An application of the approach in the multimedia domain is given in [128].

Dealing with uncertainty and imprecision. Handling vague knowledge, uncertainty and imprecision in ontology construction and matching are important real life issues, which have been addressed in the literature recently. The theory of fuzzy sets

[10] This section presents a brief state-of-the-art of the original topic of my research—that of Ontology Matching—to which I have not been contributing for a number of years now and, therefore, have only loosely been following the latest advances. It is, therefore, not complete nor up-to-date. In return, it corresponds to a particular point in time to which my contributions are relevant. Of course, I include pointers to recent surveys on the field. I hope my readers will agree that this is a fair approach to structure this thesis.

and logics provides a suitable framework for handling imprecise information in ontologies. A general definition of a fuzzy ontology is given as one *which uses fuzzy logic to provide a natural representation of imprecise and vague knowledge, and eases reasoning over it* [16]. Papers by Sanchez, Calegari and colleagues [19, 20, 101] form an important body of work in this field. These authors have been motivated by the observation that crisp reasoning through two-valued logics is not suited to deal with uncertain or imprecise information available in a real world context. They define every ontology concept as a fuzzy set on the domain of instances, denoted by E. Relations are defined as fuzzy mappings of the kind $R : E^n \longmapsto [0,1]$. Particularly, subsumption is handled by the fuzzy taxonomic relation $\tau : E^2 \longmapsto [0,1]$. In that, $\tau(i,j)$ expresses the fact that the entity j is a specification of i to a certain degree between 0 and 1.

Although there exists a solid body of work on fuzzy ontologies on one hand and on ontology matching on the other hand, only few authors have addressed *fuzzy ontology matching*. Work in this field can be classified into two families: (1) approaches extending crisp ontology matching to deal with fuzzy ontologies and (2) approaches addressing imprecision of the matching of (crisp or fuzzy) concepts.

Based on the work on approximate concept mapping by Stuckenschmidt [115] and Akahani *et al.* [6], Xu *et al.* [152] suggested a framework for the mapping of fuzzy concepts between fuzzy ontologies. Their approach is based on finding the best approximations in an ontology for all the concepts in another ontology. The approximations (least upper approximation and greatest lower approximation) are defined by using fuzzy concept subsumption and an iterative algorithm is proposed to find a simplified least upper bound. With a similar objective, Bahri *et al.* [11] proposed a framework to define similarity relations (*More General, LessGeneral, Equivalent, Disjoint, Overlap*) among fuzzy ontology components based on their intentional definitions (i.e. a set of description logics formulas that represent the meaning of a component).

The second family of fuzzy matching approaches is characterized by the representation of imprecision of the matching itself, even with crisp ontologies. For instance, Ferrara *et al.* [36] propose a fuzzy approach which handles mapping uncertainty and provides criteria for its validation. The principle of this approach is to interpret and translate each crisp matching result as a set of fuzzy assertions and perform fuzzy reasoning over this set. An ontology mapping approach based on fuzzy conceptual graphs and rules is proposed by Buche and colleagues in [17].

To define new intra-ontology concept similarity measures, Cross *et al.* [25] model a concept as a fuzzy set of its ancestor concepts and itself. As a membership degree function, the authors use the information content (IC) of concept with respect to its ontology. IC can be measured by using external text corpus (more occurrences of the concept suggest less informativeness) or by using the ontology structure – the number of ancestors and the depth of the concept in its ontology. Cross *et al.* suggest a number of intra-ontology concept similarity measures based on these fuzzy set representations.

In what follows, I will proceed to summarize my contributions to the ontology matching field, with a particular accent on instance-based matching, multimedia ontologies, as well as fuzzy ontology matching by the use of background knowledge and its applications to the problem of multilingualism.

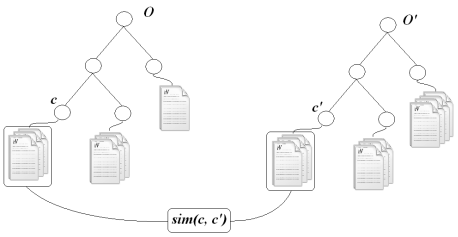
2.2.2 Instance-based Ontology Matching

Ontologies annotating text documents. The core contribution of my PhD thesis consists in the development of approaches for instance-based ontology matching. That comprises relying on ontological instances in order to establish correspondances between the classes (or concepts) of two or more ontologies. My contribution focuses on the particular case where ontology concepts are used to annotate text documents (e.g. in the case of Web-directories such as Yahoo or Amazon). In that, a concept is modeled as a set of instances (text documents) and the equivalence relation between two concepts is established based on a measure of similarity on their corresponding sets of documents (cf. Fig. 2.3a). The matching process allows to align heterogeneous annotations of similar Web content.

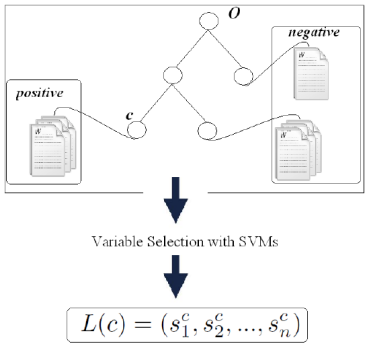
In that context, together with my PhD thesis supervisors, I have developed several approaches, described in my early publications, e.g., [129, 131, 133]. Here, I focus in a little more detail on the approach published in [134]. In this work, we have relied on two main components in order to assess the relatedness of two cross-ontology concepts populated with text documents: (1) modeling text documents as vectors of features and (2) applying machine learning techniques in order to define a similarity measure on these vectors as a proxy for the similarity of two concepts. The component (1) relies on a rather straightforward approach where many feature modeling techniques can be applied, such as for example, well-known text indexing methods creating vector space models for text documents by using weighting schemes based on term frequencies (e.g. TF*IDF). My main contribution, therefore, is centered on (2) and it answers the question of how to define and apply similarity measures on concepts by using their documents' vector space model representations.

The similarity measures I propose are based on learning a classifier (relying on the Support Vector Machines (SVMs) family of models) for each concept that allows to discriminate the respective concept from the remaining concepts in its ontology. In particular, I develop and compare experimentally two new measures: (a) one based on comparing the sets of support vectors from the learned SVMs and (2) one which considers the list of discriminating variables for each concept. This is illustrated in Fig. 2.3b, where L denotes the learned list of discriminant features for a given concept c. The lists of features that describe two cross ontology concepts (e.g., c and c' in Fig. 2.3a) can be considered as random variable vectors of ordered observations and thus can be compared by taking them as inputs for non-parametric correlation measures. In my work, I have tested Spearman, Pearson and Kendall's tau [24, 133], as explained in the paper. The thus defined measures look for re-occurring elements in two lists of top k-scored variables. The proposed measures are compared to two standard approaches (Jaccard similarity and class-means distance). A complementary contribution of the approach, pertaining purely to the machine learning field, is the development of a method to determine discriminative data features by using a novel variable selection approach for the SVMs based on their VC-dimension parameter.

Multimedia information retrieval. In a natural continuation of my PhD work, during the first period of my postdoctoral research, I proceeded to apply and extend my results obtained on text-document-populated ontologies on multimedia data, in particular dealing with the task of aligning annotations of images in order to enhance multimedia information retrieval. Together with my colleagues from the Ecole Centrale Paris, I have proposed methods to align multimedia concepts coming



(A) Instance-based ontology matching: the use-case of concepts annotating Web documents.



(B) Variable selection on a binary classification task for concept description.

FIGURE 2.3: Modeling concepts annotating text documents as sets of discriminative features.

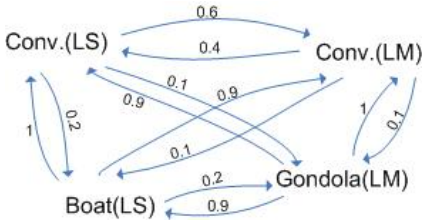
from established multimedia annotation ontologies (such as LabelMe or LSCOM) to common sense knowledge, e.g., contained in large online encyclopaedia or lexical databases like WordNet. In that, I have adapted and enhanced the instance-based matching techniques presented above in order to deal with images. The proposed approaches, disseminated within the multimedia information retrieval community [136, 54, 53, 132], showed promising results in their first applications for the linguistic annotation of images and image retrieval more generally.

2.2.3 Ontology Matching with Background Knowledge as a Mediator

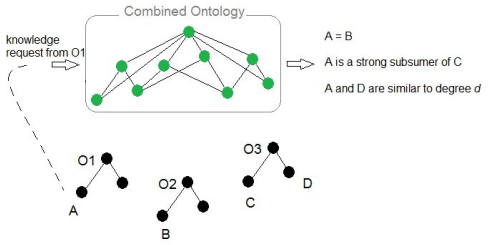
Using background knowledge in the ontology matching process was one of the challenges identified in the survey paper of 2012 by Euzenat and Shvaiko [110]. In a response to that challenge, I had the opportunity to provide a number of propositions that enhance ontology matching and open possibilities for novel applications. A particularity of the proposed solutions is that they attempt to model the inherent vagueness and imprecision of any possible alignment, produced by a machine or by a human, borrowing a theoretical apparatus from the fuzzy sets theory and they rely on background knowledge for that purpose. I will proceed to describe these contributions briefly.

Fuzzy ontology matching. By the second half of my three years postdoctoral fellowship, I started taking interest in the consideration of the inherent fuzziness and imprecision of the alignment of ontologies. Together with my colleagues, we asked ourselves the question, from a more general standpoint, to what *degree* two data models share a conceptual overlap and along what axes this overlap can be broken down. Can we account formally for the partial overlap of knowledge and explain it via distributional representations?

These questions made me turn towards the theory of fuzzy sets and the use of a reference knowledge base as a source of background knowledge in the process of ontology matching. Together with my team at the Ecole Centrale Paris, I worked on the development of an ontology alignment framework with two core features: the use of background knowledge and the ability to account for the vagueness of the resulting concept alignments, allowing to a certain extent to enhance the explainability of the results. The background knowledge is considered to be contained in a reference vocabulary of some kind, which is used for *fuzzifying* the ontologies to be



(A) A fragment of the common fuzzy ontology of the two multimedia ontologies LSCOM (LS) and LabelMe (LM).



(B) Example of a fuzzy matching across three input ontologies leading to the construction of a combined knowledge body and the kinds of relations that can be derived.

FIGURE 2.4: Fuzzy ontology matching by using background knowledge: examples of applications.

matched. Note that the choice and selection of a background knowledge source according to particular domains or applications is a subject of follow up research that I carried out later on at LIRMM [126] (not described in this thesis). For the purposes of the studies that I summarise here, we have relied on general-purpose reference vocabularies such as Wikipedia and BabelNet that can be used in many cases, and even allow cross language matchings (I detail on that particular aspect below).

In the first step of the approach, each *domain* concept (meaning that it belongs to one of the two input ontologies) is represented as a fuzzy set of *reference* concepts (meaning that they belong to the background knowledge source of reference). This fuzzification is obtained by scoring each of the domain concept with respect to its similarity to each of the reference concepts, allowing in that way to represent this concept as a fuzzy distribution over the set of reference concepts. In other words, each input concept is projected to the space defined by the set of reference concepts. In that, the concepts of both input ontologies will live in the same reference-concept space, where their proximity can be measured.

The fuzzy set formalism implies modelling these vector representations as fuzzy membership functions allowing in this way to infer fuzzy relations between cross-ontology concepts instead of crisp ones and, additionally, explain these relations by the relative contribution of each reference concept to the obtained alignment. Technically speaking, the fuzzified domain concepts are matched to one another by using a fuzzy set version of a dot product or an euclidean distance (for details, please see the related papers [137, 135]), resulting in fuzzy descriptions of the matches of the original concepts.

As an application of that framework, based on the fuzzy representations of the mappings, we propose an algorithm that produces a merged fuzzy ontology that captures what is common to the source ontologies. The example in Fig. 2.4a, showing four concepts from two different ontologies (*Conveyance* and *Boat* from the LSCOM (LS) ontology and *Conveyance* and *Gondola* from the LabelMe (LM) vocabulary), illustrates the relation of subsumption computed as a fuzzy degree on the cartesian product of these four concepts. We can see that the equivalent concepts of Conveyance are related via the subsumption relation with almost identical degrees (close to 0.5) while hierarchically related concepts, such as Gondola and Conveyance have a non-symmetrical distribution of these degrees that depends on the direction of the relationship, implying that Gondola is a Conveyance with a score of 0.9 while a Conveyance is a Gondola with a 0.1 score. Note that the subsumption relation is thus measured and inferred for cross-ontology concepts allowing to go beyond links of

identity in the matching process. In the general case, any given relation that is derived from the matching process will be assigned naturally a degree, as this is shown in the example. In addition, every established correspondence between two given input concepts can be explained by the degree of "contribution" of the reference concepts to that particular correspondence—we are thus able to identify a number of reference concepts that explain the obtained match.

The proposed approach allows to align multiplie input ontologies and, more interestingly, thanks to the fuzzy set relationships that are computed over the set of concepts from these ontologies, derive a common (merged) fuzzy ontology. In that, relations of different kinds can be inferred for any pair of concepts coming from any pair of input ontologies, as illustrated in Fig. 2.4b.

The developed approaches and the obtained results gave raise to a series of journal and conference publications that gather communities in the intersection of the fields of semantics and uncertainty / fuzziness (e.g. [137, 135]).

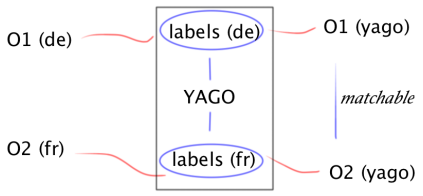
Multilingual Ontology Matching. The vision of a multilingual Web of Data was presented back in 2012 in a paper by Garcia et al. [44], where the provision of links across multi- and cross-lingual vocabularies was identified as one of the challenges towards the realisation of this vision. Two ontologies are considered *cross-lingual* if they are described in two different natural languages, while each of them is monolingual. They are said to be *multilingual* if each of them is described by the help of labels in different natural languages [114]. The presence of multilingual labels renders, indeed, the alignment process more challenging, since it adds a supplementary heterogeneity to be handled, in addition to standard differences in terminology, scope or structure.

Taking up that challenge, I have applied and extended my work on fuzzy ontology matching by the help of background knowledge to the problem of multilingual and cross-lingual ontology alignment. This work was carried out by the end of my postdoctoral fellowship and continued at LIRMM in the first years of my appointment as an Assoc. Professor. This topic was one of the main themes of the dissertation of my PhD student Abdelnacer Tigrine.[11]

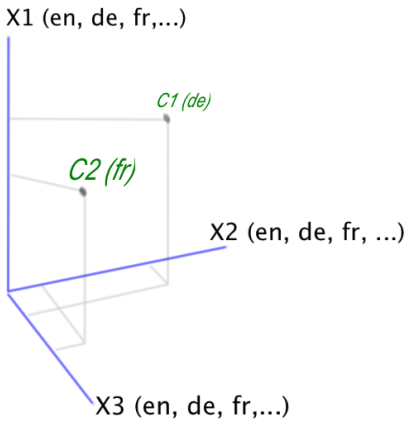
The majority of the existing approaches at that time relied on machine translation to deal with this problem.[12] Inherent problems of machine translation are imprecision and ambiguity. Together with my colleagues, I proposed a novel approach to the cross-lingual ontology matching task, relying on large multilingual semantic networks, such as YAGO and BabelNet, as a source of background knowledge to assist the matching process. The approach leans upon the idea described above, consisting in representing the concepts of the input ontologies as vectors in a space defined by the concepts of the reference vocabulary (cf. Figs. 2.5a and 2.5b). The particularity here is that the input ontologies come with labels in different languages. Therefore, using a multilingual reference knowledge resource as background knowledge comes as a natural choice. The approach is implemented under the form of a prototype named LYAM++ (Yet Another Matcher-Light)—an end-to-end cross-lingual ontology matching system that does not rely on machine translation. The reported results in [125] show that LYAM++ outperforms considerably the best techniques in

[11] Abdelnacer, unfortunately, had to abandon his thesis into his fourth year due to personal and health issues, but his work resulted in a number of publications in Web data and semantics conferences such as ODBase and EKAW.

[12] Note that nowadays more performant methods based on multilingual embeddings and neural language models have been successfully applied to this task [100].



(A) Aligning cross-lingual ontologies.



(B) Projection of two cross lingual concepts (C1 labeled in German and C2 – in French) to a common multilingual vector space.

FIGURE 2.5: An illustration of a cross-lingual concept alignment by using the YAGO multilingual knowledge base as a mediator.

the state-of-the-art according to the obtained results on the MultiFarm datasets[13] of the Ontology Alignment Evaluation Initiative 2014.[14]

2.3 Data Linking and Knowledge Extraction for Data Linking Applications

A strongly related to the ontology matching field is that of data linking, defined as the process of establishing typed relations between entities (or instances or resources) across knowledge graphs (or what we also will call datasets or RDF graphs) in an automatic manner. Just as in the case of the ontology matching problem, the motivating observation is that two entities in two different datasets can be described quite differently (e.g., by using different labels, sets of properties or identifiers) raising again the heterogeneity issue. The identity relation between resources, allowing to declare that they refer to the same real-world entity, is of particular interest since it allows to create bridges across isolated silos of data enabling federated queries and enhancing data sharing.

Research in that field aims at addressing two major challenges that can be tackled jointly in certain cases: (1) reducing data heterogeneities by the help of the development of generic data linking approaches and (2) reducing the human effort in that process by rendering the task as automatic as possible. I will discuss state-of-the-art works in the field with respect these two challenges, which provide an appropriate framework to situate my contributions. My work in that field follows two main directions: *(a)* directly addressing the data linking problem in terms of both challenges (1) and (2) defined above by the development of a novel data linking system, and *(b)* proposing approaches indirectly related to the linking task for: knowledge graph profiling, dataset recommendation and key ranking. These approaches allow to discover new knowledge about the datasets at hand before the data linking process starts. This knowledge aims to facilitate the choice of linking candidates or the configuration of generic data linking tools.

[13] https://www.irit.fr/recherches/MELODI/multifarm/
[14] http://oaei.ontologymatching.org/2014/

2.3.1 A Brief State-of-the-Art on Data Linking

Knowledge Graph Heterogeneities

In the context of Web data linking, we will refer to data heterogeneity as any difference in the expression of a given piece of information about entities across two knowledge graphs, observed in terms of schema (classes, properties), values, or general data structure. Understanding data heterogeneity in its multiple forms allows to identify and analyse the origins of the data linking problem and hence propose better solutions. Together with my PhD student Manel Achichi, we have propesed a novel heterogeneity types classification [4], extending the one given in [38]. It allows to present a number of data linking approaches along its axes.

Value Dimension. Datatype properties are an ample source of heterogeneities, both when it comes to string or numerical attributes. In that, we can observe *Terminological heterogeneity*, referred to as differences between the lexical labels used to denote the same information across graphs. This comprises well-known issues related to synonymy, polysemy or variations in spelling. The problems of synonymy and polysemy have been largely addressed in the literature by term disambiguation techniques assisted by lexico-semantic resources [75]. A long tradition of research in the field of string similarity measures, at the core of the instance comparison modules of state-of-the-art systems [79, 143, 35], has allowed to handle the case of orthographical variations among labels [78, 22]. Several works propose solutions allowing to find the full form of an acronym or an abbreviation [153, 65]. *Lingual heterogeneity*, in turn, concerns the multilingualism problem discussed in [44]. Several studies propose solutions based on machine translation [104, 62] or, alternatively, relying on a lexico-semantic resources such as BabelNet [75] as a mediator to bridge the language gap, as proposed in [63]. More recent method call upon concept embeddings models [74]. Finally, the *types of properties* induce a particular heterogeneity type. A piece of information can be given by a string literal via a datatype property, or by a URI, that is used to identify the same element in a controlled (SKOS) vocabulary.[15] Since the two objects are not directly comparable, a linking tool needs to access the string label associated to the URI in the respective vocabulary before proceeding to the comparison. In the same line of thought, comparing object property values (different URI's identifying the same object) can potentially lead to a similar type of difficulty.

Ontological Dimension. This dimension concerns schema-related differences across RDF graphs. *Vocabulary heterogeneity* has been amply discussed in Section 2.2 of this work. Ontology matching techniques surveyed there, have been adopted internally by certain data linking systems [87], but can also be applied independently. *Structural heterogeneity*, or the description of an entity with different levels of granularity across two graphs is, to the best of our knowledge, only partially resolved by using inverted indexes and NLP techniques by a several data linking approaches [143, 62, 63, 99, 105]. A vector space model is used to represent each resource description and select linking candidates on the basis of vectors proximity. *Property depth heterogeneity* refers to the same piece of information being found at different distances to the resource in two different graphs. This problem can also be solved by indexing the scope of literals describing each resource. For each entity, the distance at which the literals are collected can be fixed (e.g., [59] choose a distance of 1). *Descriptive*

[15] Examples of such vocabularies: https://github.com/DOREMUS-ANR/knowledge-base/blob/master/vocabularies/

heterogeneity refers to a resource described with more information (a larger set of properties and types) in one dataset than in another. The lack of information narrows down the intersection of the resources respective descriptions and hence the common ground where to look for commonalities or differences. Recent link prediction [57] and knowledge graph completion methods [67] can be applied to recover the missing shared of description. Finally, *Key heterogeneity* regards mismatches of key-properties across datasets. Key identification algorithms [119, 117, 112] aim to discover discriminative properties on two datasets independently and thus identify potential candidates for link specifications of property-based state-of-the-art tools [79, 143]. However, in certain cases, the values of such properties are not comparable and comparing them may lead to the generation of false negatives. We take two examples: (1) properties that are valued by unstructured textual information (e.g., a free-text description) and (2) properties used to provide dataset-specific individual identifiers, e.g., the ID's of bibliographical entries of two libraries. In both cases the values of these key properties are not comparable across datasets.

Logical Dimension. In a number of cases, the equivalence between two pieces of information across two datasets is implicit but can be inferred. For example, this is the case of two resources belonging to different classes for which an explicit or an implicit hierarchical relationship is defined. Moreover, two instances referring to the same object can belong to two different subclasses of a given class. On the property level, the equivalence between two values can deduced in a reasoning task: $<< i2 >, - : composed, "Moonlight\ sonata"@en >, << i4 >, - : composedBy, < i1 >>, << i4 >, - : title, "Sonate\ au\ clair\ de\ lune" >$.

Data Quality Dimension. Quality related issues can be observed on any of the levels discussed above, therefore, we consider these aspects as a separate (transversal) category. The topic of data quality has been of interest for many years to the semantic Web community [14, 32]. We will provide several examples of heterogeneities related to data quality that can potentially hinder the instance matching task. *Transgression to best practices*. Data representations can differ depending on the degree to which the Semantic Web best practices are respected in the data publishing process (e.g. missing language tags, the introduction of inappropriate symbols that are supposed to replace missing information (while a good practice would be to ignore what we do not know), the use of a string literal instead of a URI to identify an object, and so forth). *Value type heterogeneity*. This concerns differences in encoding data, as for example, representing an *age*-value as a string or as a number, or a date as a string. The benchmark data generators SPIMBENCH [103] and LANCE [102] focus on these issues by applying value transformations. *Dataset currentness*. The temporal evolution (or the lack thereof) of data and its dynamicity [32] can lead to conceptual discrepancies across datasets. For example similar or identical types in terms of semantics can be applicable to a given group of instances only during well-defined periods of time.

Automating the Data Linking Process

The data linking process commonly follows a pipeline consisting of three main steps [37]: (1) preprocessing, where data is prepared for linking and a number of system parameters are set, (2) matching, where instances are compared by the help of an aggregation of similarity measures and (3) post-processing, where erroneous links are removed and / or new links are inferred. For extensive surveys of data

linking approaches, we refer the reader to [1, 37, 77]. Here, we focus in more detail on the phase that takes place before the actual instance comparison. We argue that the preparation of data and the configuration of the linking tool constitute a major part of the effort with regard to the linking task. Moreover, this effort is often required from the user, leading to a pressing need of automation of this process. Therefore, we pay particular attention to approaches that propose (semi-)automatic solutions to the preprocessing and configuration tasks.

Several of the most commonly used linking tools [55, 56, 79] require prior knowledge provided by either the user or another tool in order to proceed to the linking task. This knowledge is expressed in the form of *linking rules*, describing under which conditions two instances should be compared and linked. There are two main configuration groups of elements to feed to the linking tool: (1) types (classes) of instances to align as well as a set of properties across the two datasets whose values to compare, and (2) a set of similarity measures, together with thresholds and possibly an aggregation function. We discuss these two groups in the following subsections.

Selecting Classes and Properties. The choice of types of instances that defines the pool of linking candidates is often left to the user (considering a dataset as a set of resources belonging all to the same class), although certain systems attempt to identify the equivalent classes automatically by applying ontology matching techniques [87, 56].

The properties to compare are selected manually or by the help of key discovery tools—this choice is crucial for it predetermines the outcome of the linking task. Intuitively, instances having common values for highly discriminant sets of properties (keys) are likely to be representing the same real-world objects. While many approaches to automatic key discovery from RDF data exist [119, 117, 8, 113, 9, 118, 121], their use for data linking is not always straightforward. Most of these tools produce large numbers of keys valid on a single dataset with no assessment given of their likelihood to discover links. For example, a property containing a record's identifier in a bibliographical database will be identified as a key in two datasets containing the entries of musical works of two libraries independently on one another, but it will be of no use for the linking task, since the two libraries use different identifiers for the same work. An exception is [117], which considers keys valid on two datasets. In addition, key discovery systems do not consider the heterogeneity of the properties used to describe instances across datasets, which compromises the usefulness of the keys for the linking task. In an attempt to overcome this issue, the authors of [8] present measures of the quality of link keys, valid on two datasets, in order to facilitate their selection. Two recent studies [2, 34] propose approaches that attempt to close the gap between key discovery and data linking tools, allowing to produce a list of keys, valid on two datasets simultaneously and ranked with respect to their usefulness for the particular data linking task at hand.

Learning Link Specifications. Link specification is defined in [83] as (i) the setting of the elements to compare from two knowledge bases, (ii) the setting of a complex similarity metric via the combination of several atomic similarity measures, and (iii) the setting of thresholds for these similarity measures. The (semi-)automatic link specification approaches of which we know have focused predominantly on (ii) and (iii)—configuration parameters of type (2) that can either be set by the user or learned from data in a semi–supervised or unsupervised manner. Two main categories of *semi-supervised* learning methods emerge: *active* [80, 83, 82] and *batch* [50,

52] approaches. *Batch* approaches require a large amount of candidate links as input to learn the classifiers while *active* approaches proceed iteratively and for each iteration the user is asked to label a set of generated links until the maximal number of iterations is reached or the fitness value is greater than a given threshold. *Unsupervised learning* methods attempt to surpass the necessity of human labeled examples [58]. A method based on a refinement operator that only needs positif examples that are more often available than negative ones is proposed in [106], while [85] propose an approach implemented in KnoFuss [88], based on a genetic programming algorithm learning iteratively the optimal similarity parameters. However, it is required from the user to set the fitness function and to specify the fitness measures, thresholds and the maximum number of iterations. Certain releases of the well-known data linking tool LIMES [79] include both EAGLE [80] and WOMBAT [106] as link specification algorithms,[16] while SILK [55] includes ActiveGenLink [50].

2.3.2 Linking and Disambiguating Entities across Heterogeneous RDF Graphs

In order to reply to several of the challenges raised above, particularly regarding reducing a larger number of heterogeneities and automating the matching process, I have introduced in collaboration with my PhD student Manel Achichi and her co-supervisor Zohra Bellahsene, the data linking system *Legato*.[17] The tool has been primarily developed in the context of the ANR DOREMUS project and motivated by the use cases of the project. It is however a generic and standalone instance matching system and the underlying methodological approach constitues one of the major contributions of Manel's PhD thesis.

The global framework of the approach is illustrated in Figure 2.6. The tool takes as an input two RDF graphs (two sets of instances of the same type). The datasets are automatically preprocessed and prepared for comparison, and then a set of links is generated, as a result of an instance matching, instance disambiguation and link selection (or link merging) procedures. Note that in its default release, the system takes one parameter as input: a pair of types (classes) of instances and optionally a global similarity threshold value. However, for the data-aware user, a customisation of *Legato* is possible with regard to two additional parameters, giving rise to two versions of the tool - an automatic and a manual one. I will here proceed to detail on the automatic default release of *Legato*.

Property Filtering. As we have seen above, *key heterogeneities* hinder the resources comparison, mainly because properties concerned with this heterogeneity type are erroneously likely to be considered as linking rules parameters. If a linking tool uses these keys to compare instances, it will fail to find a correspondence. A way of going around this problem is to remove properties with such values, that we will call *problematic properties*, before proceeding to data comparison. We propose to identify automatically these properties by discovering all mono-property keys that are valid over *both* datasets to be linked (in that we consider the union of the two input datasets as a single dataset), i.e., each object for such a property has at most one subject in *both* graphs.

Main Matching Module. A core feature of our approach is the representation of instances as text documents and their projection to a vector space. Particularly, each resource is represented by a set of literals considered as relevant to its description, based on a choice of a *CBD* subgraph. CBD, for concise bounded description, is

[16] E.g., *limes-core-1.2.1*.

[17] The source code of the tool is openly accessible at `https://github.com/DOREMUS-ANR/legato`

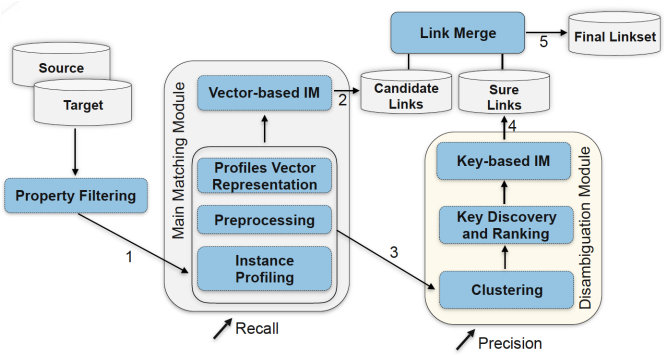


FIGURE 2.6: Processing pipeline of *Legato*. (Figure from the PhD thesis of Manel Achichi.)

a subgraph centered around a given resource that includes all nodes that are directly linked to the resource or separated by it via a blank node or a chain of blank nodes.[18] We extend the *CBD* definition by considering the descriptions of neighbouring nodes of a resource r in its graph. A *CBD* sub-graph is a directed one, where the orientation is implied by the order of the subject and the object in a triple. This allows us to introduce the notion of successors of a node r (the nodes that are in a triple of which r is a subject) and the predecessors of r (the nodes that are in a triple of which r is an object). In that, we consider four different kinds of CBDs: the classically defined one, the subgraph including also the CBDs of the successors of the resource, the subgraph including also the CBDs of the predecessors of the resource and, finally, the subgraph containing both the CBDs of the successors and the predecessors of the resource. We refer to an instance representation obtained on that basis as an *instance profile*, defined as the set of literals collected on the CBD-based subgraph that has been chosen to describe the resource. This results in the creation of a pseudo document that relates and describes each of the graph instances and enables their comparison based on comparing the textual document that profile them. Avoiding property-based comparison addresses the remaining ontology-level heterogeneities introduced above.

Once all resources in both datasets are profiled, the resulting documents are processed in order to prepare data for the matching task. The set of *instance profiles* in both datasets are indexed in a standard manner. We project the instance profiles to a vector space and weight them by using their TF*IDF scores per instance. The distances between the vectors of the resources, expressed by the cosine similarity measure, is used as a proxy for the similarity of resources. We empirically fix te similarity measure threshold to 0.2 (deliberately low) in order to capture a large number of links and ensure high recall. As an outcome of this process, a first linkset (a short for “set of links") is produced, called *candidate links* (Fig. **??**).

The main matching module ensures high recall. To improve the matching quality and precision, we perform a post-processing step, described next, allowing to filter out erroneous links that may have been generated at this step and add new quality links.

Instance Disambiguation Module. Taking as an input the vector space representations of the indexed instance profiles, the algorithm proceeds to cluster *within each dataset* highly similar instances by relying on a *Hierarchical Clustering Algorithm* [98]. A cluster matching procedure across the two datasets, using a distance metric on the

[18] https://www.w3.org/Submission/CBD/

cluster centroids, allows to isolate pairs of corresponding clusters, where the first one belongs to the source dataset and the second one—to the target dataset. Each pair of corresponding clusters is then analyzed separately and their respective instances are compared, this time on a property basis. The effectiveness of the comparison process depends on the quality of the selected properties. We apply key ranking algorithms (described in detail below) in order to discover keys that are valid on *two* datasets, ranked with respect to the performance achieved by a linking system using these keys in its configuration. This allows to select the set of properties over two graphs that guarantee the best linking result. We apply such an algorithm independently on each pair of corresponding clusters. In that, we identify the discriminant properties among these clusters that would have remained "diluted" in the global graphs. This allows to disambiguate the highly similar instances in each pair of clusters and maximises the rate of correct alignments. As a result of this process, we end up with a second set of links, that we call *sure links*.

Link Merge. Finally, a merge operation is performed on the two linksets generated previously. The set of *sure links* will be taken as a catalyser on the links in the *candidate links* set and directly fed to the final linkset, because of the high precision in the process of generation of the links that it contains. For each link between two resources r_s and r_t l=(r_s, r_t) in the set of *candidate links*, the module searches over the set of *sure links* for a link between a source resource r_s and a target resource $r'_t \neq r_t$. If found, the link l=(r_s, r_t) is deleted from the *candidate links* set. The remaining links in the *candidate links* set, merged with those from the *sure links* set, are then fed to the *final linkset*.

2.3.3 RDF Dataset Profiling and Profile-based Dataset Recommendation for Data Linking

As this story goes, we now have tools to establish links between datasets, both on schema level and on instance level. The specificity of these tools is that they take as an input either two ontologies or two RDF graphs. In certain cases, we will have our datasets to align readily available—for example, the catalogs of two libraries within a common data integration project involving the two institutions that provide the data. However, it might be useful to be able to create links to datasets that are beyond our particular use-case, datasets already published on the Web, but which contain identical or related resources as those described in our data, datasets of whose existence we might not even know.

How do we discover such datasets, that are potentially good linking candidates with respect to a given RDF graph? In order to reply to that question, I have worked, together with my PhD student Mohamed Ben Ellefi and colleagues from France and Germany, on the development of methods for dataset profiling and recommendation. *Profiling* here is understood as the task of defining a set of characteristic features that best describe a dataset and also allow to separate it maximally from the other datasets. *Recommendation* is seen in the context of discovering linking candidates among a set of already published datasets that are likely to contain identical or related resources to those of a given input dataset and ranking these datasets with respect to this likelihood. I will detail on these notions and the ensuing approaches that we have developed in the sequel. The overall profiling and recommendation process is represented conceptually in Fig. 2.7. For a given input dataset D_S and a set of target RDF graphs,[19] we define and construct profiles of D_S and all datasets in

[19]In the figure, this set is denoted by LOD, for the Linked Open Data cloud, but in principle this can be any let of already published RDF graphs.

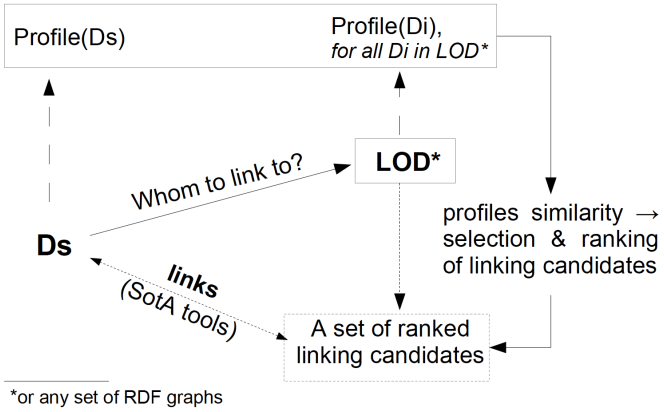


FIGURE 2.7: Profiling and recommendation of datasets within the data linking frame.

the target set. The similarity measured on their profiles is used in order to select and rank a subset of target datasets that will be then recommended for linking to D_S.

Profiles can be defined in many different ways. For example, an RDF dataset can be described by the set of its most frequently used vocabulary concepts and properties, its structure, indicators of the data quality (completeness, errors), etc. One of the contributions of Mohamed Ben Ellefi's thesis was the publication of a thorough survey of RDF graph features, the methods for their extraction, applications and existing vocabularies for their representation [12]. This survey is discussed in more detail in Chapter 4, where I present five of my major publications. We note here that a feature-based representation of a dataset of this kind, that we call a profile, allows to compute distances or measure similarities between datasets (or for that matter profiles). Among various applications of this idea, discussed in [12], this allows to devise methods for the recommendation of datasets, which meet specific criteria. This task is of particular interest when addressing issues such as entity retrieval, semantic search and data linking. As discussed earlier, the latter application is of particular importance for the discovery of good quality linking candidates that would allow to go beyond established datasets such as DBpedia (the usual "suspect") when creating links between a newly published dataset and graphs already existing on the Web of data. Within the PhD thesis project of Mohamed, we have focused on that issue, resulting in two independent dataset recommendation approaches both based on the notion of dataset profiling: (1) a topic-based approach using a collaborative filtering algorithm for recommendation [30] and (2) an intensional approach relying on schema overlap as a recommendation criterium [31]. I will briefly present these two approaches in the sequel.

(1) A collaborative filtering approach. The main principle of collaborative filtering suggests that one can rely on the link among "like-minded" people in the recommendation process, e.g., if two persons have similar tastes they can recommend items to each other and each one will be happy with the recommendations made by his or her peer. In that, similarity with other users is based on distance measurements based on user profiles. We translate this setting to the data linking problem: if a dataset D_S is strongly similar to a dataset D_T, we consider that D_S and D_T have the same connectivity behaviour, meaning that they are likely to be linked to the same or similar resources and thus are likely to share related resources.

A previous approach by Fetahu et al. [39] has suggested to build topical profiles for datasets. The authors sample resource instances randomly and then extract entities and topics from these samples by using state of the art tools (e.g. DBpedia Spotlight[20] and its extensions). The topics are then weighted with respect to each dataset and the ratings, or weights, are normalized. This allows to construct a bipartite graph containing two sets of nodes: topics and datasets, linked via the pair-wise strength of relation between the elements of the two.

Building on this approach and integrating it in a novel dataset recommendation algorithm, we have proposed to represent each topic as a distribution over a set of datasets. By using the dataset topic profile of a newly arriving dataset as obtained via the approach in [39], we then can assign to each topic the top k datasets of that topic's distribution. In that, we propose a list of ranked datasets of relevance together with information of the topic via which they have been recommended. For example, we will know that "economy" and "taxes" are the most prominent topics that allow to link a new dataset about finance to its linking candidate datasets that could be also of a broader scope.

Our approach [30] has been evaluated by using all accessible at the time LOD datasets. We have used as ground truth for the evaluation the topology of links of the LOD at the time of publication. Due to the incompleteness of the links between LOD datasets, a natural problem that we had to face during the evaluation is the overestimation of false positives (i.e. predicting links that do not exist currently in the LOD ground truth data, but that turn out to be valid links).

(2) An intensional approach. In a follow up paper [31], we have introduced an alternative and simpler dataset recommendation approach, also based on profiles, but of different nature. Here, we are interested in intensional dataset characteristics in the form of a set of keywords together with their definitions in order to build our profiles. We provide the following definitions.

Definition 2 (Dataset Label Profile) *The* **label profile** *of a dataset D, denoted by $\mathcal{P}_l(D)$, is defined as the set of n schema concept labels corresponding to D: $\mathcal{P}_l(D) = \{L_i\}_{i=1}^{n}$.*

The representativity of the labels in $\mathcal{P}_l(D)$ with respect to D is improved by filtering out certain types, like too popular concepts (such as foaf:Person) or, alternatively, types with too few instances in a dataset. Each of the concept labels in $\mathcal{P}_l(D)$ can be mapped to a text document consisting of the label itself and a textual description of this label (e.g., the definition of the concept in its ontology, or an external description of the terms from the label). We define a document profile of a dataset in the following way.

Definition 3 (Dataset Document Profile) *The* **document profile** *of a dataset D, $\mathcal{P}_d(D)$, is defined as a text document constructed by the concatenation of the labels in $\mathcal{P}_l(D)$ and the textual descriptions of the labels in $\mathcal{P}_l(D)$.*

The document profile is an extended label profile with more terms, allowing to project the profiles onto a richer vector space by indexing the documents and using a term weighting scheme of some kind (e.g., TF*IDF). Applying an euclidean-distance-based measure of similarity on the vectors will serve as a proxy for profile—and thereof dataset—similarity.

By the help of these two definitions, a profile can be constructed for any given dataset in a simple and inexpensive way, independent on its connectivity properties

[20] https://github.com/dbpedia-spotlight/dbpedia-spotlight

on the LOD. We rely on the simple intuition that datasets with similar intension have extensional overlap. Therefore, it suffices to identify at least one pair of semantically similar types in the schema of two datasets in order to select these datasets as potential linking candidates. We are interested in the semantic similarity of concept labels in the dataset label profiles. We have focused on the well known semantic measures Wu Palmer [151] and Lin's [66], as well as the UMBC [45] measure that combines semantic distance in WordNet with frequency of occurrence and co-occurrence of terms in a large external corpus (the Web).

With a concept label similarity measure at hand, we introduce the notion of dataset comparability, based on the existence of shared intension.

Definition 4 (Comparable Datasets) *Two datasets D' and D'' are comparable if there exists L_i and L_j such that $L_i \in \mathcal{P}_l(D')$, $L_j \in \mathcal{P}_l(D'')$ and $sim_{UMBC}(L_i, L_j) \geq \theta$, where $\theta \in [0, 1]$.*

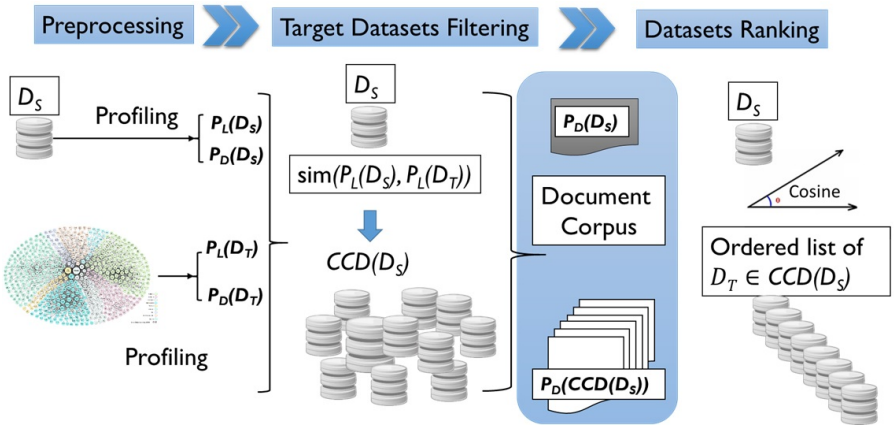


FIGURE 2.8: Pipeline of the intensional recommendation approach (figure taken from the PhD thesis of my student Mohamed Ben Ellefi).

Putting these elements together in an integral recommendation framework, we introduce the pipeline illustrated in Fig. 2.8. The preprocessing step consists of constructing the label and document profiles of our input dataset, D_S, and all target datasets D_T - in the example we see the LOD cloud diagram from 2016 to represent the pool of target datasets. The following step, named Target Datasets Filtering, selects a subset of the target datasets, named cluster of comparable datasets (CCD) that are such that their semantic profile similarity to D_S is sufficiently high, or in our definition this will be the set of comparable datasets with respect to D_S. Finally, the set of datasets in the CCD of D_S are ranked by using a standard text document similarity measurement appraoch based on a TF-IDF weighting scheme and the cosine similarity.

Again as in the previous approach, we have used the totality of the responsive LOD dataset in order to evaluate the recommendation method. I refer the reader to [31] for a more detailed discussion.

2.3.4 Key Discovery and Ranking

We have reached a bit farther: in addition to data linking approaches, we have now in hand tools for selecting the most suitable datasets that we should link our own one to. As mentioned in the previous sections, key properties (defined, for example, in [119, 117]) of an RDF graph are useful to conduct the link discovery task,

particularly if we rely on popular state-of-the-art systems like SILK [55] or LIMES [79], which adopt a property-based philosophy for data linking. As outlined in Section 2.3.1, however, data linking approaches remain to a large extent disconnected from key discovery approaches. While the latter aim at outlining key properties that promise to be useful for parametrizeing DL tools of reference, they come short in doing so due two main reasons: (1) the majority of these algorithms, with very few exceptions [8], operate on a single dataset, thus not accounting for the potential commonalities of the two datasets to be linked, (2) they often return a very large number of keys, reaching on certain data several hundreds, with no intuition provided about their usefulness for the data linking task at hand.

In the framework of a collaboration with INRA Montpellier and in particular with Danai Symeonidou, I have carried out work aiming to close the gap between automatic key discovery and data linking approaches. This work has lead to the introduction of two approaches for key ranking, named RANKey [2] and Key Ranker [34], the former being developed in collaboration with my PhD student Manel Achichi and the latter - with my master student Houssameddine Farah. The two methods suggest to unlock the potential of key-based techniques by providing the user a list of *ranked keys* valid on a set of datasets, well-suited to a particular instance matching task. As compared to automatic link specification algorithms, e.g. [83, 86], our approaches can be seen as complementary: we focus on the identification of a limited set of properties that can be used to effectively link datasets, while leaving the choice of the similarity measures, their combination and tuning to the user, or to the auto-configuring link specification methods just cited and discussed in more detail in Section 2.3.1.

I will proceed to briefly describe the KeyRanker approach,[21] since it followed the RANKey tool and was demonstrated to outperform it [34]. The overall pipeline of the approach is given in Fig. 2.9. We consider all resources from a given type C of a source dataset DS_S and a target dataset DS_S (we suppose that there is a known correspondance between types across the two datasets). In order to harmonize data by reducing certain trivial forms of terminological and structural heterogeneities between the explored datasets, a data cleaning step takes place, including downcasing, eliminating special characters, alphabetical reordering of the tokens of a literal are applied depending on the nature of the data. Data are then rewritten (or transformed) in order to prepare it as an input for generic key discovery algorithms – in our approach we use SAKey [119]. Given a set of keys generated by SAKey on the two RDF datasets, we merge and rank these keys so as to favor those most likely to produce identity links between the datasets. The ranks correspond to the *individual* likelihood of each key to produce links when used in a link specification file of a system such as SILK [55] or LIMES [79]. In that process, both datasets are processed simultaneously to ensure the applicability of the keys on the two datasets (and hence their usefulness for the linking task). Our method represents instances by using their Concise Bounded Descriptions,[22] exploring information found on different depths of the graphs. This allows to efficiently handle object properties having URI values, which bypasses a major shortcoming of key discovery tools. We take into account various data heterogeneities at the key selection step by exploring partial similarity of string literals. In that, we preselect key properties based on the presence of string similarity between their values across datasets. We define the notion of a *ranking*

[21] The source code and datasets are openly accessible at https://github.com/HFarah/KEYRANKER-2017

[22] See Section 2.3.2 for a definition.

support as the number of source instances that have at least one corresponding resource in the target dataset. KeyRanker is a generic tool, independent both on the used key discovery or data linking systems and on the data.

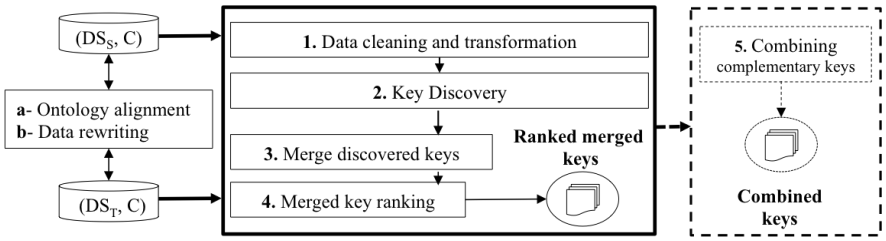


FIGURE 2.9: Pipeline of the KeyRanker approach.

Additionally and separately, we propose a method allowing for the combination of a selected subset of keys that, applied together, increases the number of discovered correct links. To this end, we introduce and apply the notion of key *complementarity*. Intuitively, merging the sets of identity links produced by several top ranked keys yields a larger set of identity links, containing more links than any of the sets produced by each of these keys individually. The complementarity notion is defined based on the ability of keys to discover non-overlapping sets of instances.

We evaluate our approaches on diversified data coming from different Instance Matching tracks of the Ontology Alignment Evaluation Initiative (OAEI) 2010 and 2016,[23] as well as on real-world data about classical music.

2.4 Data Modeling and Knowledge Graph Building

We have seen that many challenges stand on the way of linked open data provision: from schema alignment and instance matching, to dataset profiling and recommendation, key discovery and ranking. The ultimate goal of this process is to structure data and knowledge and make them available to large communities of data consumers. Knowledge graphs—a term coined by Google in 2012—designate those data artefacts that describe in a structured manner the entities of interest in a given domain together with their relations and can be seen as among the end “products" in this long pipeline that we have named “Web data lifecycle" in the beginning.

Knowledge graphs, particularly when they are openly accessible and built abiding by the FAIR principles, are powerful tools that enable data reuse and federation thus improving information retrieval and facilitating research and knowledge discovery in various fields of life, including science, cultural heritage or education. As also discussed in the introductory section of this chapter, their construction implies the orchestration of several knowledge engineering tasks, from data collection and harmonisation to knowledge extraction, data modelling and linking, to the provision of tools for data access and exploration tailored to non-computer science experts. This process can largely depend on the domain of interest and, therefore, often requires the development of specific approaches that are best fit to the data and domain at hand. In addition, the building of such knowledge graphs poses a number of scientific challenges related in particular to data modelling and knowledge extraction. I will focus on these two broad aspects in the following describing two recent efforts towards the construction of large knowledge graphs in two different fields. I will first introduce the DOREMUS knowledge graph of linked musical

[23] http://oaei.ontologymatching.org

works [3], which gathers metadata about classical music coming from three major French cultural institutions. This work has been predominantly performed in collaboration with my PhD student Manel Achichi and EURECOM (Raphael Troncy and his student Pasquale Lisena). Further on, I will present recent work on the construction of ClaimsKG—a knowledge graph of fact-checked claims, which facilitates structured queries about their truth values, authors, dates, journalistic reviews and other kinds of metadata and provides ground truth data in support of research into the analysis of societal debates on the Web [122]. The latter is a result of ongoing collaboration with colleagues from the computational social science institute GESIS in Germany.

2.4.1 Conceptual Models and Knowledge Graphs for Music Meta-Data

The ANR DOREMUS project gathers computer and social scientists together with librarian experts from three major French cultural institutions—the French National Library (BnF), Radio France and the Philharmonie de Paris—in order to develop shared methods to describe semantically their catalogs of music works and events and create tools for exploration and recommendation based on analysis of user needs and types of usages. This process comprises the construction of knowledge graphs representing the data contained in these catalogs following a novel agreed upon ontology that extends existing models, the linking of these graphs and their open publication on the Web. A number of specialized tools that allow for the reproduction of this process are developed, as well as Web applications for easy access and navigation through the data.[24] The main outcome of this project is the DOREMUS knowledge graph, consisting of three linked datasets describing classical music works and their associated events (e.g., performances in concerts). This resource fills an important gap between library content description and music metadata. I will briefly present the DOREMUS pipeline for lifting and linking the data, the tools developed for these purposes, as well as a search application allowing to explore the data. For details I refer to [3].[25]

As a result of collaborative work between library experts from the three cultural institution and Semantic Web experts (lead by R. Troncy's team), we have developed the DOREMUS model—an ontology for the description of music catalogs. It is an extension for the music domain of the FRBRoo model for describing librarian information, which has in turn been born as a dialog of the librarian FRBR model and the CIDOC-CRM ontology[26] for representing museum information, putting together the distinction between Work, Expression and Manifestation with the centrality of creation events for describing the cultural object lifecycle coming from the latter [28]. DOREMUS imports the Work-Expression-Event triplet pattern of FRBRoo: the abstract intention of the author (Work) exists only through an Event (i.e. the composition) that realizes it in a distinct series of choices called Expression(s). On top of the FRBRoo original classes and properties, specific ones have been added in order to describe aspects of a work that are related to music, such as the musical key, the genre, the tempo, the medium of performance (MoP), etc. [23]. The model is ready to be used for describing the interconnection of different arts: it is the case of the soundtrack of a movie, or a song that uses the text of a poem. In addition to the DOREMUS ontology, I have participated in the development and alignment of a number of controlled vocabularies expressed in the SKOS formalism, all of which

[24] Open access to the data and the tools is granted at https://github.com/DOREMUS-ANR.
[25] http://data.doremus.org/
[26] http://www.cidoc-crm.org/

are relevant to the field of music. Some of these vocabularies were already available and in use by the community: in this case our contribution consists in their collection, conversion to SKOS (if needed) and alignment. For the alignment purposes, we have relied upon the vocabulary matching and mapping validation tool YAM++ online, developed in collaboration with my colleagues at LIRMM.[27] As a result, we collected, implemented and published 17 controlled vocabularies belonging to 7 different categories (musical keys, types of derivation, modes, thematic catalogs, functions, musical genres and MoP) [69].

The data lifting process, as described in the beginning of this chapter, in general and in the DOREMUS case implies the extraction of entities and relations from the raw data and their attribution of names and identifiers from our ontology. This laborious process has been handled in close collaboration with the librarian experts. As a result, we have come up with several knowledge graphs describing musical works and their associated events - one per institution. The linking of these graphs, and in particular of the music works that they describe, has been performed by the help of the *Legato* system, introduced in Section 2.3.2 - one of the major outcomes of the PhD thesis of my student Manel Achichi.

Currently, the DOREMUS dataset includes more than 16 million triples, which describe over 3 million distinct entities. The classes and properties used come mostly from the DOREMUS ontology, FRBRoo and CIDOC-CRM, counting in total 57 distinct classes and 120 distinct properties.

2.4.2 Conceptual Models and Knowledge Graphs for Fact-checked Claims

In the framework of a recent and ongoing collaboration with colleagues from FORTH (Greece), GESIS and L3S (Germany) and ITM-Mines Alès (France), I have steered the creation of a knowledge graph of fact-checked claims, named *ClaimsKG* [122] that I will proceed to describe in the sequel.[28]

Various research areas at the intersection of computer and social sciences require a ground truth of contextualised claims labelled with their truth values in order to facilitate supervision, validation or reproducibility of approaches dealing, for example, with fact-checking or analysis of societal debates in a broader sense. On the other hand, non-computer science experts, like social scientists or journalists, need to be given tools for enhanced information retrieval about claims, allowing to answer complex information needs. Some of these needs might currently require looking up multiple desperate resources, such as various fact-checking portals and online encyclopaedias, in order, for example, to retrieve all false claims by Donald Trump mentioning journalists. We have observed that no reasonably large, up-to-date and queryable corpus of structured information about claims and related metadata was publicly available. In an attempt to fill this gap, together with my colleagues, I have introduced ClaimsKG, a knowledge graph of fact-checked claims,[29] which facilitates structured queries about their truth values, authors, dates, journalistic reviews and other kinds of metadata.

ClaimsKG is generated through a semi-automated pipeline, which harvests data from popular fact-checking websites (we currently crawl 8 websites, for a full list we refer to our website given in the link above) on a regular basis. The entities contained in the claims and their journalistic reviews are then annotated with related entities from DBpedia. The data is then lifted to RDF by using an RDF/S model that we have

[27] http://yamplusplus.lirmm.fr

[28] https://data.gesis.org/claimskg/site

[29] Within this work, we define a claim as a statement reviewed by a fact-checking organisation.

developed for the purposes of our task that relies on established vocabularies such as schema.org and NIF, represented in Fig. 2.10. In order to harmonise data originating from diverse fact-checking sites, a normalised rating scheme is introduced that maps the various assessments specific to the individual fact-checking portals to four basic categories: “true", “false", “mixture" and “other". We also introduce a simple claims coreference resolution strategy that allows to establish links of identity across a number of identical claims published at different times at different websites.

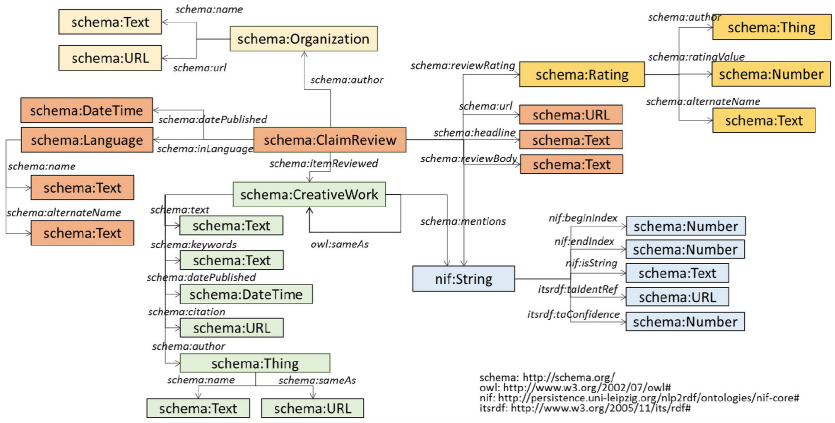


FIGURE 2.10: A model for fact-checked claims based largely on the ClaimReview class of schema.org.

The current knowledge graph, extensible to new information, consists of more than 32K claims published since 1996, amounting to more than 6,606,000 triples. To facilitate data access and retrieval, we have developed and provided open source Web applications for exploration of our resource and its statistics [42].[30]

[30] https://data.gesis.org/claimskg/explorer/

Chapter 3

A Web of Linked Data and Beyond: Three Novel Research Axes

In the current chapter, I will present three research axes, along which I plan to direct my efforts in near future in terms of collaboration networks, funding opportunities and doctoral and master students supervision. The first and third axes are very broad, composed each of several more specific research challenges, that I outline briefly in the sequel, while the second axis is based on a specific problem arising from experience with real-world data.

The first axis is a direct follow up of my long standing contributions in the field of data linking and knowledge integration that I have started already during my PhD thesis, but proposes a novel way of approaching this problem arguing that it has the potential of overcoming limitations of current approaches. Note that a paper describing this axis of research has been recently accepted to ISWC[1] 2019's "Outrageous ideas" track [130].[2] This track of the conference provides a forum for visionary ideas, long term challenges and opportunities for the Semantic Web field.[3]

The second axis is also situated in the field of data linking, but, contrarily to axis 1, is inspired by a very specific problem from that field - the issue of linking RDF graphs where the needed shared description context of entities is missing, which hinders the application of state of the art tools.

The third axis consists in the development of more recent research topics that I have started working on only since 2 years. It comprises the combination of semantic approaches for conceptual representation of knowledge and data with machine learning techniques in order to enable a better understanding and reuse of online discourse data, such as claims on controversial topics, their sources and associated viewpoints, going beyond the current *fact*-centered paradigm of Web search, data provision and consumption. While machine learning has been the topic of my master studies and in part also underlies my PhD thesis contributions, this axis can be seen both as a novel thematic field of research and as a return to my original area. The application field of the planned research along this axis is rooted in the broad context of the analysis of societal debates on the Web and is framed by an ongoing recent collaboration with computational social scientists from Cologne (Germany).

3.1 Towards a Bottom-up Data Linking Paradigm

The efforts of the Semantic Web community are largely directed towards the realisation of a shift of the paradigm regarding the way we publish and consume data

[1] The International Semantic Web Conference

[2] The paper won the CCC Blue Sky first prize award at ISWC'19 amounting to 1K$.

[3] https://iswc2019.semanticweb.org/call-for-outrageous-papers/

on the Web moving from a network of unstructured Web documents towards a Web of knowledge, where knowledge graphs and linked data play crucial role to enable entity centric search and knowledge discovery. And while the knowledge consolidation and KG construction methods are largely bottom up and data-centric (I refer here, for example, to the Wikidata project[4] or the ever larger adoption of shema.org markup), can the same be said about the methods applied to do linked data? I will argue that the answer to that question is "no", while enabling a data-centric discovery of links across RDF datasets can be potentially beneficial.

Where are we now. As discussed in the previous chapter, the problem of data linking is defined as that of automatically establishing typed links between entities of two or more RDF datasets or graphs. A variety of data linking systems have been proposed over the past 15 years within the Web community with interactions with government, cultural or research institutions as major linked data consumers and providers. As a result, vast amounts of linked data already exist on the Web (I refer, for example, to the LOD project[5]). A number of benchmarks are developed and shared publicly in order to provide frameworks for the evaluation of data linking systems, driven by the well-known OAEI campaign, or the more industry-oriented EU HOBBIT project.[6]

Recalling our discussion from the previous chapter, state-of-the-art research into data linking [77] goes in two main[7] directions: (1) proposing novel generic data linking systems and (2) developing methods for automatic link specification by (semi-) supervised machine learning techniques, in order to assist the configuration and tuning of established tools. Several of the most common systems, such as SILK [55] and LIMES [79], adopt a property-based link-discovery strategy: a set of predicates has to be selected before the system proceeds to compare their values by the help of (an aggregation of) similarity measures that also need to be selected and tuned. This configuration task can be demanding in terms of user involvement in real-world scenarios. Hence, a number of methods have been proposed to assist the users in the configuration process. Properties to compare can be selected by the help of key discovery tools. While many approaches exist (e.g., [120]), their use for data linking is not straightforward because they often produce a large number of keys that are valid on a single dataset with no assessment of their likelihood to discover links. On the other hand, automatic link specification learning approaches develop (semi-)supervised techniques to select and combine similarity measures and fix their thresholds. Systems like EAGLE [81] and WOMBAT [106] are included in LIMES, just as ActiveGenLink [51] is part of SILK.

Most existing linking approaches have in common the fact that they attempt to solve the problem from a generic stance by remaining vastly agnostic to the nature of the underlying data [77]. Many systems achieve good results on dedicated benchmarks [5], but fail to take into account the particularities of the various domains and/or data generation practices that raise very specific heterogeneity issues calling for a significant user input. In particular, the user is required to have an in-depth understanding of both their data *and* the internals of the linking system of choice in order to achieve satisfactory results, as shown in [3]. The quest to fully automate the linking task, on which recent research has departed, remains rather challenging,

[4] https://www.wikidata.org/wiki/Wikidata:Main_Page
[5] https://lod-cloud.net/
[6] http://oaei.ontologymatching.org, https://project-hobbit.eu/
[7] I insist here on the word "main", since the granularity of research methods in reality is way higher.

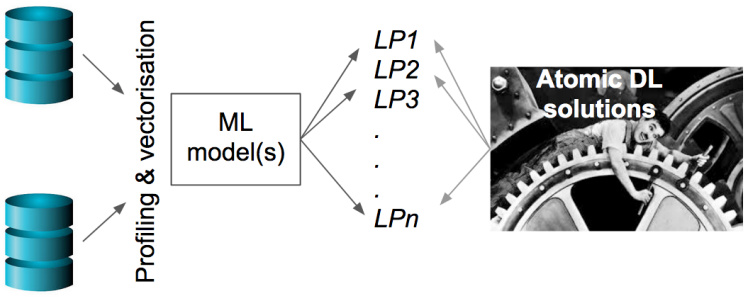


FIGURE 3.1: A pipeline for the proposed framework. LP = Linking Problem (type), DL = Data Linking.

as investigated in [4]: a heavy machinery of learning link specification rules is being developed to only partially assist the users in the selection of parameters in the pre-processing step of the linking task.

Rethinking the data linking process. I argue that the current generic approach to develop data linking solutions has reached its limits and suggest that a paradigm shift in the way we look onto this task needs to take place. I propose to enable the development of data-centric approaches for bottom-up linking, rather than investing efforts in devising incremental generic solutions: time has come to step back and look at what we can learn from the large amount of existing cross-dataset links *and* linking systems.

I formulate the hypothesis that there exist a finite number of identifiable and generalisable *types of linking problems*, defined as heterogeneity types that two to-be-linked datasets can manifest, e.g. differences in terminology, natural languages or structure, as presented in my earlier work (together with my PhD students M. Achichi and M. Ben Ellefi) [4] and also discussed in the previous chapter. Additionally, I hypothesise that these linking problems can be detected automatically by the help of machine learning (ML) models trained on sufficiently large amounts of quality linked data. On the other hand, state-of-the-art linking tools are based on modules (I will call them *atomic* or *modular* solutions), that allow to handle separately many of these linking problem types (e.g., measure the string or semantic similarity of entities). On these bases, I redefine the data linking task as that of the *automatic identification via ML techniques of the linking problem type(s) that two datasets manifest and the application of an automatically generated combination of atomic linking solutions that are best fit for the datasets at hand*. I propose to lean upon the wealth of existing linked data sets, particularly those coming from real-world scenarios, in order to enable training and validation of ML models, while a number of RDF graph profiling and graph embedding methods will be applied in order to extract the necessary features for these models. This is illustrated shcematically in Fig. 3.1.

Based on the hypotheses formulated above, I propose to direct future research and engineering effort into the development of a data-centric bottom-up linking framework that channelizes and consolidates existing disparate efforts. This will allow to build on the wealth of linked RDF graphs via their in-depth analysis and consolidation and take advantage of years of research and practice in the field (cf. Fig. 3.2). I identify a set of research axes and associated challenges on the way to realize this framework.

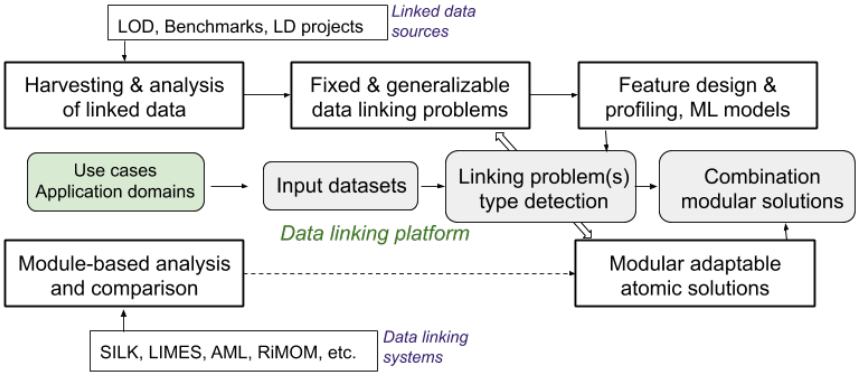


FIGURE 3.2: Conceptual overview of the proposed framework.

(1) Linked data harvesting and analysis. This axis consists in the consolidation of a large amount of already existing quality linked Web data, benchmarks, evaluation campaigns and linked data projects from a large variety of domains. An in-depth statistical analysis of these data will allow to identify a number of limited and generalisable linking problem types, discover correlations between application domains and data structure or quality, or between heterogeneity types and link density. This will inform the feature design in (2) and will generate training data for the automatic classification of pairs of datasets according to their linking problem type(s). *Challenge:* We need to ensure high quality of the links from which we will learn. Hence, I propose to rely on existing real-world benchmarks as a starting point, where datasets are often grouped according to specific heterogeneity criteria (terminology, logics, structure, etc.) that can be mapped to the identified linking problem types. Alternatively, one can apply existing linking methods of high precision in a preprocessing step. Relying on a large variety of data sets and domains is important to guarantee representativity. A particular difficulty is ensuring completeness of the training datasets in terms of available links in order to avoid the generation of false negatives. This problem has been discussed in more detail in one of my student's ESWC papers [31] in the context of dataset recommendation.

(2) Joint datasets feature design. This axis involves data linking-oriented graph profiling and feature extraction via state-of-the-art RDF graph profiling techniques [12] and joint graph embeddings methods (learning vector representations jointly on a pair of graphs). It will generate the set of features that describe jointly the linking candidate datasets and are indicative of the heterogeneities that they manifest, necessary to train the ML algorithms. Thus, we aim to answer the question of what is it that discriminates between two pairs of datasets manifesting two different types of linking problems. *Challenge:* A large plethora of RDF dataset profiling methods and tools exist (reviewed in our recent work [12]), allowing to extract and represent the graphs in terms of a number of "profile features", such as their domains, connectivity, representative instances, quality, provenance, statistics, dynamicity, etc. Under the hypothesis that these features in combination account for describing the datasets from aspects that match the linking problems identified in (1), a significant challenge consists in identifying the set of features that are necessary and sufficient in order to design efficient linking problem classification ML models. In addition, from a practical viewpoint, the application of the methods that allow for the extraction of these features is not straightforward, as outlined in [12]. Finally, we will be interested in extracting *joint* profiles for a *pair* of datasets, which is not explicitly addressed in the literature. In that respect, profile features can be coupled with graph embeddings

learned *jointly* on a pair of RDF graphs, which is a novel problem in the community [145].

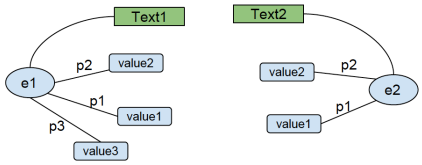
(3) Learning and applying ML models. This axis will rely on the training data harvested in (1) and the features extracted in (2) in order to define and apply classification models for linking problem type detection. *Challenge:* We identify here the standard challenge of selection and tuning of ML model(s) from a set of supervised algorithms. In addition, the multitude of possible classes will lead us beyond the standard binary classification task.

(4) Construct a library of automatic, adaptable, modular solutions for each of the linking problems identified in (1). We rely on the premise that a linking problem type is fine-grained enough so that a particular modular solution can be applied to it (for example, relying on lexical synset intersection for synonymy-type heterogeneity). These modular solutions will be identified by a comparative analysis of the modules and respective performances of a large spectrum of existing data linking systems, as this has been in part performed in [77]. *Challenge:* A significant effort will be involved in the association of state-of-the-art atomic solutions to the data linking problem types identified in (1). In the lack of training data, unsupervised ML models have to be divised, enhanced by a *human-in-the-loop* approach.

In a summary, instead of trying to fit a generic solution to any linking problem and dataset type, I suggest to enable a better understanding of the underlying data before applying a targeted solution best suited to the particular datasets at hand. I rely on the premise that the in-depth analysis of large amounts of linked data will allow to isolate a limited number of identifiable data linking problems that ML models based on datasets profiles will help detect automatically. The moment is appropriate to take this approach for reasons of, one the one hand, the large and growing availability of linked data in an ever greater number of domains and, on the other hand, the existence of a large plethora of data linking tools, result of decades of research and practice. Channelizing these decentralised endeavours will foster and facilitate the application of linked data technologies within and across an even larger variety of domains and will ultimately free the domain expert of the technological burden.

Application Fields, Use-Cases and Realisation of the Research Project For the realisation of this research axis, I will target the field of cultural heritage, because of the large adoption of linked data in that field and the potential of building rich use-cases within an already existing collaboration network. Indeed, I will build on my experience gained via the DOREMUS project in order to continue my contributions to reinforcing linked data production and publication in the cultural heritage field by expanding my current collaborations (including Radio France and the BnF) beyond the national boundaries. I will seek to federate a broader network of both Semantic Web and machine learning experts and data providers from the cultural heritage field. In particular, I am in the process of establishing a collaboration with Antoine Isaac, manager of the Europeana project[8] that gathers a number of European cultural actors (libraries and museums). I plan to submit a joint Horizon Europe proposal in 2021 (an ANR MRSEI funding will be requested based on the proposal described here), in which cultural institutions partners of Europeana will provide use-cases for the realisation of the envisioned research.

[8] https://www.europeana.eu/portal/fr



(A) Two non-complementary entities.

(B) Two complementary entities.

FIGURE 3.3: Illustration of the case of non-complementarity (on the left) and complementarity (on the right) .

3.2 Data Linking in the Lack of Shared Context

As discussed in the state-of-the-art sub-section of Section 2.3, a number of commonly used data linking systems rely on the comparison of values of selected properties across instances of two different knowledge graphs. The premise of this approach is that two instances from the same classes from two different graphs are described by the help of more or less the same (or semantically equivalent) properties, or at least that the intersection of the sets of properties that describe them is non-empty and within that intersection one is able to find characteristic properties (for example keys). With that hypothesis at hand, the focus falls on coming up with ways to compare the values of those semantically equivalent properties in order to use their similarity as a proxy for the similarity of the resources that they describe.

Dataset complementarity as a challenge for the data linking task. I had the occasion to come across real-world scenarios, where this hypothesis simply does not hold, both in the case of the DOREMUS project (from the field of music) and in the case of the AgroLD use-case in the domain of plant biology. For reasons that remain specific to the way data is generated, collected and structured locally, different providers choose to describe their data by the help of different or very little intersecting sets of properties. In addition, under the open world assumption missing values are largely present in the case of large knowledge graphs. For example, in the case of AgroLD, among several of the involved datasets, one can find numerous examples where the same plant gene is described by the help of non-intersecting sets of properties across the two datasets. I will call *complementary datasets* two datasets that are such that identical entities across them are described by complementary sets of properties (see Fig. 3.3a and 3.3b for an illustration).

I suggest to deal with this problem by developing an approach in two steps (illustrated in Fig. 3.4), where a number of research questions arise and will be addressed within each of the two steps.

- Step 1: Graph enrichment. The premise is that each entity is potentially described in external sources, such as scientific articles, other knowledge graphs or social networks. The missing descriptive intersection then can be recreated by looking up those external sources or by enriching the initial graphs by adding this additional information that will allow for the linking. The following research questions will be addressed:

  Q1: Where to look for the missing information?

  Q2: How to recreate the missing description intersection?

  In order to tackle Q1, I propose to continue my previous work on selection of background knowledge [126] and dataset recommendation (described in

Chapter 2), by situating and adapting the approaches to the case of background knowledge recommendation for particular scientific domains. As a starting point, I will rely on the expertise of domain users: for example, in the case of dealing with agronomy data, these will be biologists, who will identify a set of relevant scientific articles or datasets where the missing background knowledge can be found.

My main effort will be directed towards tackling Q2. The difficulty is to relate the entities from the graphs to be linked to the external resources identified in Q1. In the framework of the masters internship of my student Serge Sonfack (defended in November 2019), I have started exploring an approach which consists in (1) the identification of the entities of interest in a set of scientific articles via named entity recognition or entity linking tools tailored to the field of biology, (2) learning of word embeddings for these entities in the text corpus that those scientific articles constitute and (3) applying the learned vector representations in order compute entity alignments, towards replying Q3 (described below). For (1), we have relied on all terms that are found in the description of the entities within each of the two graphs. In that way, we transfer the linking problem from the graphs to the text corpus, contrarily to approaches based on knowledge base completion or augmentation [148, 43, 67]. Our first results showed to be promising, where two concrete concerns arise: (a) how to weigh the terms that describe the graph entities, used to anchor them to the textual corpus, in a way that these weights are transferred to the resulting vector representations and (b) how to optimise the process in terms of time efficiency. My focus in future research on this axis will therefore fall on these two issues.

- Step 2: Enhanced semantic entity matching. Under the premise that the missing knowledge has been completed by looking up external sources, the question arises of how to represent the semantically enriched entities in a way that allows for their comparison. This gives rise to the following question:

  Q3: How to exploit representations derived from text and from relational data (knowledge graphs) in order to compute semantic similarity between entities?

  The computed entity embeddings can be used within a standard vector distance measure (e.g. the dot product) in order to come up with an estimation of the entities' similarity, as suggested above. However, given the multimodal approach, where we will rely on data structured as graphs but also on textual documents, a challenging question would be how to combine representations learned on different types of data in a common distance or similarity measure [73].

Application Fields, Use-Cases and Realisation of the Research Project. In order to pursue work in that direction, I will target the field of agronomy. I will continue and enforce my ongoing collaboration with the agro-community in Montpellier, which becomes a major provider of semantisized, rich linked data, via projects like the ANR D2KAB (lead by Clément Jonquet), for example. I will carry on my current collaboration with Pierre Larmande and the AgroLD project. These projects will provide important use-cases for the realization and application of the framework proposed above, facilitated by the understanding of the inherent problems of this field over the past couple of years. Together with Clément and Pierre, I will be looking for funding opportunities on regional level.

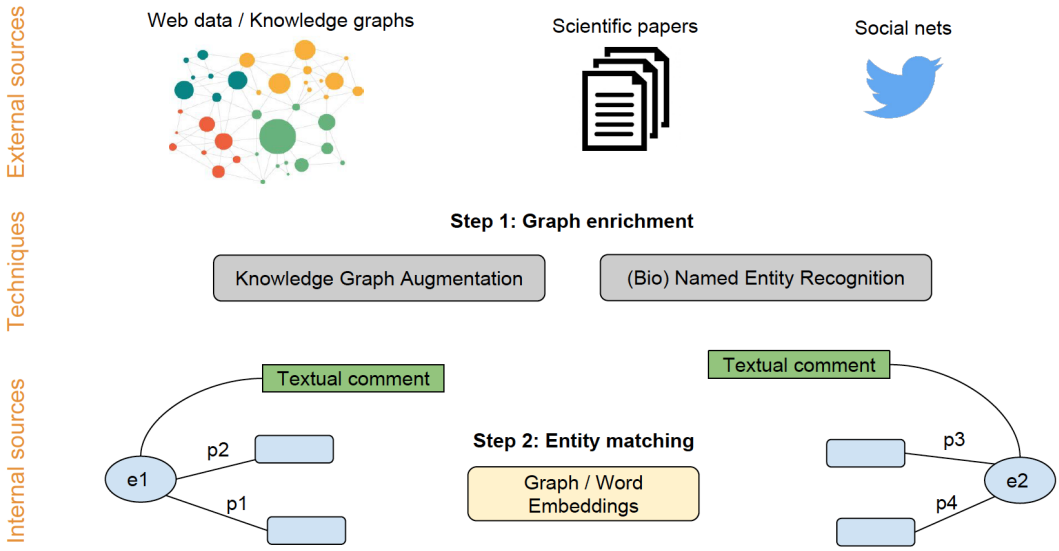


FIGURE 3.4: Linking complementary datasets via knowldege graph enrichment methods

3.3 Combining Knowledge Graphs with Machine Learning for the Analysis of Online Discourse Data

The Web has evolved into a ubiquitous platform where many people are given the opportunity to be publishers, express opinions and interact with one another. While this potentially facilitates democratic and bottom-up knowledge construction, conditions are created to manipulate the public discourse by spreading misinformation or biased narratives, bearing a great potential to influence society [144]. The (long-term) consequences cannot yet be assessed in their entirety, but harmful effects for the democratic public discourse are widely assumed [7]. As a result, we are witnessing the production of an abundance of Online Discourse Data (ODD), such as claims and viewpoints on controversial topics, their sources and contexts (e.g., related events and entities). This creates unprecedented possibilities to gain insights into societal debates about controversial topics, where ODD are a valuable source for studies into misinformation spread, bias reinforcement, echo chambers or political agenda setting.

Consequently, there is an increasing need for methods and datasets that can facilitate the analysis of ODD and societal debates for scientists from both within and outside the computer science (CS) and (computational) social science communities. Current studies often suffer from limited generalisability and representativity due to restricted and disparate sets of analyzed sources, topics, media types, or timeframes [89]. At the same time, there are no well-structured and publicly accessible datasets describing information about previous or ongoing societal debates, which could be exploited in current and future research. Knowledge graphs like DBpedia widely used for tasks like entity-based annotation of online documents or Web-search, capture **factual** information without keeping track of the diversity, connection or temporal evolution of viewpoints, aspects (subtopics), claims and sources associated to particular entities and topics.

Being able to integrate versatile data coming both from the Web (what we called

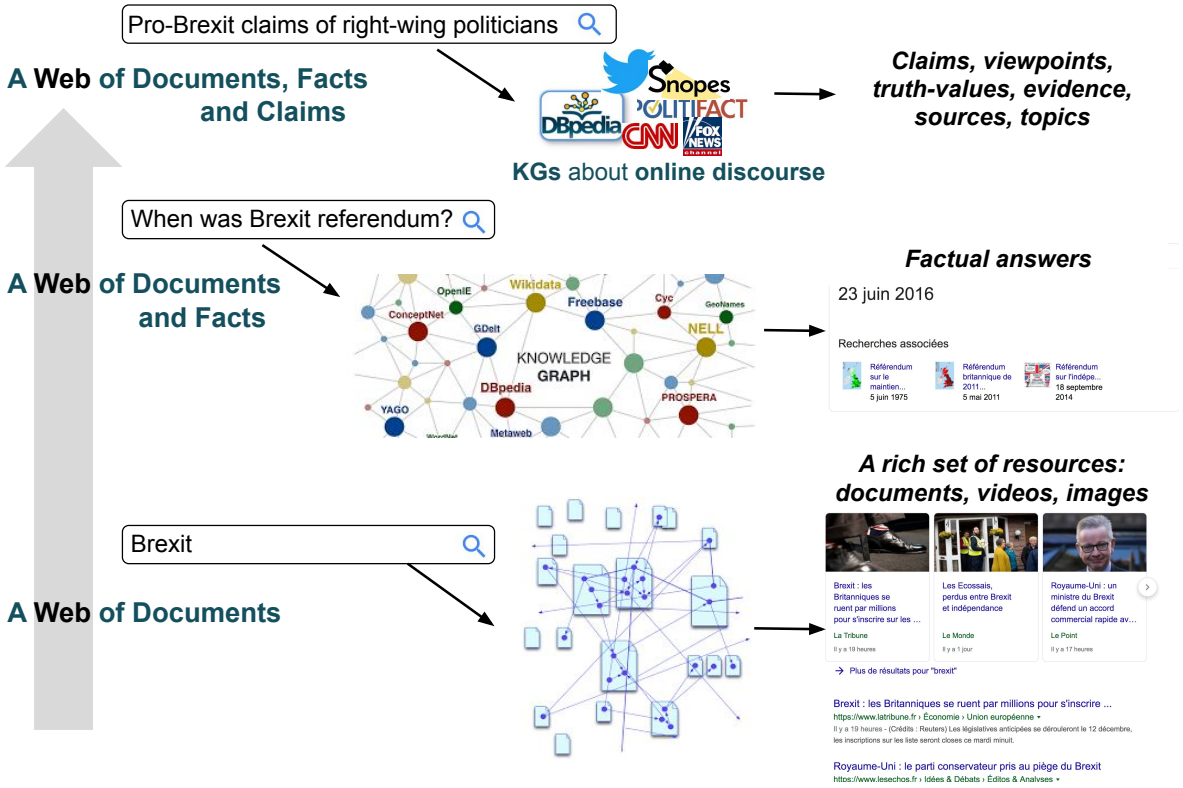


FIGURE 3.5: From a Web of documents to a Web of claims and facts: enhancing search and reuse of online discourse data.

ODD), but also from scientific archives or libraries and census institutes into semantically structured and openly available knowledge resources will empower data-driven analysis of controversial topics towards improving the democratic practices of our societies. Given the high societal importance of these issues, providing *automated tools for extraction and structured access* to the relevant information poses a pressing task for the CS and artificial intelligence (AI) communities.

The current section develops my research project within this axis in two main closely linked directions:

1. Providing methods and open source tools for the construction of large knowledge graphs about online discourse data: controversial topics, related claims, sources, viewpoints and other meta-data by enhancing and going beyond the on-going work on ClaimsKG described in the previous chapter and in [122].

2. Developing approaches at the intersection of semantics and machine learning for the computational modelling and analysis of claims, their veracity assessment and their relatedness measurement that will use the resources developed in (1) in order to discover new knolwedge.

My hypothesis is that these approaches will ultimately contribute to the enrichment of the Web in a way, where claims and claim-related information will be given machine understandable form enabling search engines to interpret claims in relation to controversies, just as the current Web enables us to understand and interpret facts. This will allow to go beyond the Web of documents and beyond the KGs empowered Web of facts onto a novel layer of the Web including *claims*, where online discourse data are made accessible, reusable and machine-readable in service of science and society at large (see Fig. 3.5 for an illustration).

Where are we now. State-of-the-art research into the analysis of societal debates comprises as major challenges the assessment of the truthfulness of claims on controversial topics taken from online debates, e.g. in social networks. We define a claim as a verifiable statement supported by one or more people/ organizations. Claim truthfulness verification methods [21] fall into the following (non-disjoint) groups. (1) Reference approaches (the largest group) rely on background knowledge to contextualize and look up the information [46]. Often claims are given structured representations, e.g. in the form of triples [108, 155] or queries [150], allowing for their matching against certified facts. (2) Inference approaches apply reasoning techniques in order to infer the truth value from structured resources [29, 13]. (3) ML approaches apply supervised learning mainly on text claims [94, 146]. Topic models and NLP filters are used to construct a feature space, relying on highly contextualized features on document level [47] or linguistic and stylistic cues [96]. Additional features such as provenance, time and sources are considered in [94].

Large knowledge graphs represent powerful resources in support of entity-centric tasks in CS and computational social science. Facts are of main interest in the Web community, forming the core elements of knowledge bases of reference, such as DBpedia or YAGO. The construction of such resources implies the extraction, verification and representation of facts. Projects like DBpedia, LinkedGeoData[9] and OntoWiki[10] have established themselves as sources of reference, where DBpedia [61], complemented by a variety of tools and techniques, facilitates large-scale publication of KGs. Knowledge base augmentation methods aim to discover and add new facts relevant to the entities of a KG, which also implies their validation [84, 93]. The approaches for fusing entity-centric Web markup data introduced in [154, 155] address the fact verification problem for heterogeneous schema.org markup. The Knowledge vault [29] and the Voldemort KG [138] also rely on markup annotations in order to validate and match statements to established KGs. Recent work, to which I had the opportunity to contribute, deals with the construction of a KG of fact-checked claims, enabling queries about their truth values, dates and entities (cf. the discussion about ClaimsKG[11] in the previous chapter). Knowledge graphs in combination with ML techniques have been largely used in order to infer new facts or validate newly discovered ones. The authors of [107] define the fact verification problem as a supervised link prediction task in a KG with connectivity patterns features. I outline [29] as one of the few transversal approaches, applied to both structured and unstructured data and relying on a diverse set of features.

A limitation of current claim verification approaches is their lack of transparency and explainability, often rooted in the lack of common theoretical understanding of the problem, as well as lack of large and shared benchmark data for the evaluation of such approaches. Regarding methods from the Semantic Web community for extracting, verifying and storing facts cited above, they do not capture explicit knowledge about the provenance of these facts or their relatedness to controversial topics, nor the diversity of the associated viewpoints and claims and the mechanics of their validation. There is, therefore a need to go beyond the mere publication of verified facts by enriching them with context and provenance information and thus open a new pathway for fact-verification methods and related research into the analysis of ODD and societal debates in general.

[9] http://linkedgeodata.org/About
[10] http://ontowiki.net/
[11] https://data.gesis.org/claimskg/site

Enabling a better understanding of online debates via the combination of knowledge graph-based and machine learning approaches. I propose to make progress towards the analysis of societal debates on controversial topic on the Web by combining a (1) Semantic Web approach for knowledge extraction and engineering resulting in the construction of knowledge graphs of linked claims, topics and sources with (2) machine learning knowledge discovery techniques that will use that KG as a background knowledge source in order to uncover latent relations and patterns (e.g., relation between sources and veracity labels of claims) and infer new knowledge and thus support research into the fact-verification task, among other related problems.

I outline several related directions of research and their challenges.

(1) Constructing knowledge graphs about ODD on controversial topics. The first objective of this research axis is thus to provide Web-accessible structured data (KGs) that connect information on the Web about controversial topics and associated viewpoints, claims, entities, features and other metadata, enabling users and applications to access, query and reason about these data. I will labor for the provision of the means to (1) dynamically build such KGs by building on W3C standards and established vocabularies (schema.org/ClaimReview, PROV-DM, Dublin Core, SIOC), via a reproducible open source knowledge extraction, representation and linking pipeline and (2) apply these KGs in a variety of use-case scenarios in order to enable the discovery of new knowledge. The KGs will allow efficient interpretation and assessment of topic-related information, like the trustworthiness of claims that support a particular viewpoint over a given period of time, in relation to a given event or the evolution of viewpoints and claims related to a given topic over time. I hypothesize that these resources will enable a variety of data-driven studies on societal topics of debate in support of AI-researchers, social and computational social scientists, fostering both knowledge discovery and new research in these fields. In addition to Web data (social and media networks and the Web at large), I will be looking into the possibility to integrate knowledge from diverse and heterogeneous sources including *scientific publications*, facilitated by the unique assets made available by library, census and research institutes with which I am currently collaborating (e.g. TIB and GESIS in Germany and the Cirad in France). The building of such resources implies the development of novel tools for knowledge extraction, claim veracity assessment and linking, by using in a combined manner structured semantic data and machine learning approaches.

(2) Modeling claims. The notion of a “claim" is central in a number of related studies into misinformation propagation, computational fact-checking, bias reinforcement or in the emergent field of argumentation mining [68, 18]. Existing definitions vary considerably not only across different fields but also even within a single community. At the same time, different communities use the same terminology to refer to different concepts. There are few studies that have focused on the computational and conceptual modeling of the notion of a claim. I will attempt to contribute to these areas by proposing, jointly with my computational social science colleagues from GESIS (Germany) and Semantic Web experts from FORTH (Greece), a novel conceptual model for claims and related entities, such as viewpoints, topics and annotations, that aims to take into consideration their inherent complexity, distinguishing between their meaning, linguistic representation and context.

(3) Claim veracity verification and explainability. While I plan to rely on and adapt existing claim extraction methods [46, 64], I intend to contribute to the claim verification field by (1) consolidating the theoretical foundations of the problem (a challenge formulated in [21]) and (2) investigating the explainability and transferability

of a diverse set of features and models. The latter promises to account for the lack of transparency and the context/source type specificity of existing models and address the problem of assessing emerging claims. Building on [29, 94, 13] and on my previous work on data linking and fusion, I propose to collect signals of claim truthfulness from Web and scientific documents in order to exploit associations to topics and, in particular, to viewpoints, which will help contextualize a claim and its veracity verdict. I will rely on the ongoing collaboration with Pavlos Fafalios (FORTH) and Mohand Boughanem (IRIT) for the development and application of approaches for the extraction and modeling of viewpoints specific topics (cf. their related publications [124, 95]).

(4) Claim matching. Claims and other extracted entities may differ in their expressions across sources and disciplines (e.g. scientific publication vs. social media, social sciences vs. natural sciences), although being semantically identical. Therefore, I will work on the development of tools and approaches for the linking and fusion of entities represented in the controversial topics KG (outcome of *(1)*) as part of its construction, enhancing the applicability of the resource. I have had the occasion to gather experience in the field of data linking, through a number of projects (the ANRs DataLift[12], DOREMUS[13]) [79, 4], in addition to my current collaborators sound expertise in that field (TIB, GESIS) and joint work with LGI2P (Ecoles des Mines d'Ales) on lexico-semantic knowledge integration [123]. However, automatically linking claims-related data sourced from the Web is an under-researched issue. On the one hand, I plan to connect claims—a problem that I am currently investigating with FORTH, GESIS and LGI2P—in order to establish links of identity and relatedness. The challenge here consists in studying the different dimensions across which two claims can be related — topical (depending on a certain granularity), contextual, textual, etc., depending on the claim model defined in *(2)*. I will rely on and evaluate a fusion approach of standard overlap measures and of sentence embeddings from the representation learning field [92]. I will further on develop methods for measuring the relatedness of topics building on our previous work on entity linking [4, 91] and extending it for the case of topics where often multiple entities are involved. Additionally, I will propose methods to establish links between source mentions and entities, as part of the PhD thesis of Katarina Boland (starting in 2020). Finally, I will work on the establishment of links to entity-centric KGs, such as DBpedia, YAGO, or EventKG[14] enriching them with currently missing contextual knowledge and anchoring our resources to the LOD cloud.

Application Fields, Use-Cases and Realisation of the Research Project For the realisation of this research project, I will rely on the ongoing and long term collaboration with various European partners, namely The Institute of Social Sciences GESIS (Cologne, Germany) and The National Technical Library TIB (Hannover, Germany), the Computer Science Institute FORTH (Greece), LGI2P (France), Cirad (France) and IRIT (France). In that, I will be guided by pilot scenarios realised in close collaboration with social and computational social scientists in two fields: communication science and agronomy. To give a few examples, we will be investigating (1) social media phenomena: public and scientific spheres and counter-public spheres (echo-chambers), (2) their practical implications and remedies: search/media bias, in collaboration with GESIS and (3) viewpoint shifts in concrete debates, e.g., land policy

[12] https://datalift.org
[13] www.doremus.org/
[14] http://eventkg.l3s.uni-hannover.de/, developed by my partners from L3S

in the South of France, in collaboration with the Cirad.

To support that research, I have submitted a grant application (with LIRMM as a coordinator) to a joint ANR-DFG PRCI French-German call including the afore mentioned partners for a total amount of 860K euros, including 4 PhD students and 2 postdocs. The project has not been selected for funding, but the encouraging reviews and evaluation led us to the decision to submit a H2020 FET Open proposal (deadline May 2020) with LIRMM as coordinator. Independently on that, the PhD thesis of Katarina Boland, co-supervised by Stefan Dietze (GESIS) and myself will start in 2020 on topics closely related to this research axis.

Chapter 4

Five Selected Articles: Contributions, Impact and Collaborations

In the current chapter, I will briefly describe five of my publications that I consider as being the most important ones. These descriptions contain, beyond a summary of the content of each paper, the justification of leading the respective work, its collaborative context, as well as its (current or expected) impact. I will present the papers in a thematic and not chronological order.

4.1 Data and Knowledge Integration

I outline three publications in the field knowledge extraction and mining in the context of data and knowledge integration of web resources.

Ngo, D., Bellahsene, Z., & Todorov, K. (2013). Opening the black box of ontology matching.

- **Published at:** Extended Semantic Web Conference (ESWC 2013), rank A.

The paper can be seen as a particular and non-standard kind of an empirical survey of the field of ontology matching. Traditional surveys would look onto OM systems as integral units and compare them methodologically and empirically. In that process, the internals of the studied systems would remain closed within the black box that each of these systems represents and the conclusions would be drawn based on the overall comparative performance of the systems (in terms of measures of Precision, Recall and F1 score). As the paper's title suggest, here we aim at opening these black boxes by analysing separately a large set of components that are commonly considered as parts of an OM system's pipeline. We look into different kinds of ontology heterogeneities (e.g., terminological, structural or semantic) and the respective modular solutions that are applied in order to solve separately each of these heterogeneities. In addition, a mapping selection module is introduced to filter out the most likely mapping candidates. We conduct an in-depth empirical study of the interaction between these components separately and working together inside an ontology matching system. Our results enable to provide insights on the impact of each of these interrelated components on the performance on the ones that depend on it further on over the matching pipeline. To carry out our evaluation, we have used a large number of evaluation datasets made available publicly in the annual Ontology Alignment Evaluation Initiative campagne. These are benchmark datasets of reference in the field that ensure the reproducibility of our results.

The work on this paper allowed me to lean upon years of experience in the field of ontology matching that I had the occasion to consolidate during my PhD thesis and postdoctoral research. In the same time, this work was my first collaboration with members of the LIRMM laboratory, namely with Zohra Bellahsene and her back then PhD student Hoa Duy Ngo, that started immediately upon my arrival and allowed for my fast integration in the working environment of LIRMM. The paper was published at ESWC 2013, which took place in Montpellier that year, and I had the occasion to present it.

Todorov, K., Hudelot, C., Popescu, A., & Geibel, P. (2014). Fuzzy ontology alignment using background knowledge.

- **Published at:** International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, impact factor 1.286.

The paper presents a novel paradigm of aligning ontologies by using a reference resource (in our case Wikipedia) as a background knowledge. The novelty of the approach is the fact that input ontologies' concepts are modelled as fuzzy sets of reference concepts. The fuzzified input concepts are matched to one another, resulting in fuzzy descriptions of the matches. Based on these concept matches, we propose an algorithm that produces a merged fuzzy ontology that captures what is common to the source ontologies. The paper describes experiments in the domain of multimedia by using ontologies containing tagged images, as well as an evaluation of the approach in an information retrieval setting. We show that this approach can also handle multilingual ontologies (i.e., such that are described in different natural languages), as long as the reference background knowledge source contains multilingual labels.

Although this paper has appeared at the time when I was already working as an associate professor at LIRMM, it presents collaborative work that I have carried out in the second half of my postdoctoral stay at the Ecole Centrale Paris. The particularity of this international collaboration that I had the occasion to steer is that it gathers both my back then postdoctoral supervisor from the Ecole Centrale Céline Hudelot and my former PhD thesis co-supervisor from the Technical University in Berlin, Peter Geibel, in addition to Adrian Popescu, researcher at CEA - Paris, providing for a large spectrum of cross-disciplinary profiles ranging from machine learning and data mining (P. Geibel and me), to knowledge engineering and extraction (myself), to information retrieval (A. Popescu, C. Hudelot) and fuzzy sets and logics (C. Hudelot and myself). This was also among my first publications in a major journal on fuzzy sets and logics, while that was a new topic that I have started developing during my postdoctoral period. The paper deals with the use of background knowledge for the ontology alignment task, which is among the novel research axes that I have brought to the LIRMM laboratory upon my arrival and particularly to my working group (what was then to become the FADO team). Particularly, this topic was central to the PhD thesis of Nacer Tigrine that I co-supervised with Z. Bellahsene (resulting in two main publications [125, 126], although not defended) and also the PhD thesis of Amina Annane, supervised by Z. Bellahsene and C. Jonquet (I have not participated in the supervision of Amina's thesis).

Achichi, M., Bellahsene, Z., Ellefi, M. B. & Todorov, K. (2019). Linking and disambiguating entities across heterogeneous RDF graphs.

- **Published at:** Journal of Web Semantics, impact factor 2.429.

This article presents the main contributions of the PhD thesis of Manel Achichi that I co-supervised with Zohra Bellahsene, defended in 2018, funded on the ANR DOREMUS project that I have co-lead as a WP leader and scientific manager for LIRMM. The work has multiple contributions. In the first place (1), it provides a survey that analyses in-depth the various data heterogeneity problems that are at the origine of the data linking task, proposing a novel classification of these types of heterogeneities. The second part (2) of the paper overviews the state of the art in terms of doing linked data with less human effort. This includes works on automatic link specification and key discovery for the data linking task. Building on these two parts, the paper proposes a novel solution, implemented in the open-source tool *Legato*, that attempts to overcome a major part of the heterogeneities studies in (1), but also to reduce the human effort in parametrizing a data linking tool, based on the analysis of the existing solutions carried out in (2). The solution thus consists in representing instances semantically by appropriately chosen subgraphs (referred to as Concise Bounded Descriptions) and then modelling these subgraphs as bags of words containing the literals gather on the different nodes of these subgraphs. The approach has the advantage of being able to handle very similar yet distinct instances by efficiently disambiguating them. The work includes a thorough empirical evaluation of the novel approach component per component by using datasets of reference in the field, both generic and domain specific, that are all made publicly available together with the system's code in order to guarantee reproductivity of the results and help foster future research in the field.

This paper represents my first major contribution to the field of data linking, which, although strongly related to the ontology matching field, represents a different and a broader and more multi-level problem. A particularity of this work in terms of collaboration is that it provided the possibility for my two PhD students (Manel and Mohamed) to work jointly on a journal submission.

4.2 Knowledge Extraction and Graph Profiling

Ben Ellefi, M., Bellahsene, Z., Breslin, J. G., Demidova, E., Dietze, S., Szymanski, J., & Todorov, K. (2018). RDF dataset profiling: a survey of features, methods, vocabularies and applications.

- **Published at:** Semantic Web Journal, impact factor 3.524.

This survey paper represents the main contribution in terms of state-of-the-art study of my PhD student Mohamed Ben Ellefi, that I co-supervised with Zohra Bellahsene, defended in 2016, funded on the PIA DATALYSE project. The paper builds on the premise that reuse and take-up of the rich data sources growingly available on the web is often limited and focused on a few well-known and established RDF datasets, because of the lack of reliable and up-to-date information about the characteristics of these datasets. The survey then goes into the analysis and classification of (RDF) dataset features related to quality, provenance, interlinking, licenses, statistics and dynamics, in relation to particular tasks and applications such as entity linking, distributed query, search or question answering. An inventory of the existing methods and tools for the extraction of these features is then presented, together with a set of vocabularies (ontologies) that allow for their formal description. The work overviews more than 80 papers, 40 vocabularies and 30 tools by specifying their type in a particular taxonomy, their accessibility and applicability to a particular task, thus relating together the different components of the survey.

This work involves an international team of seven co-authors from four different institutes and countries (France, Germany, the UK and Poland) that I had the pleasure to steer. The most active kernel was formed by my PhD student and his supervisors on the one hand, and the German team lead by S. Dietze on the other. The paper is a natural follow-up on our common work with S. Dietze on the topic of RDF dataset recommendation for the data linking task – a research effort that laid ground for my long-lasting and on-going collaboration with Prof. Dietze (note that a jointly supervised PhD project will begin in 2019), but also allowed my student Mohamed to take a central role in the preparation of a major journal paper together with a large international team. This work is an important contribution to the community with a potential to foster ongoing and future research in the field, as already testify its 35 citations within less than two years (according to Google Scholar, October 2019).

Ellefi, M. B., Bellahsene, Z., Dietze, S., & Todorov, K. (2016). Dataset recommendation for data linking: An intensional approach.

- **Published at:** Extended Semantic Web Conference. Rank A.

This paper presents one of the contributions of the PhD thesis of my student Mohamed Ben Ellefi. We witness a growing number of publicly available RDF datasets on the web (see, for example, the LOD project[1]). In order to enable data sharing and federation, there is an identified necessity to establish links between any newly arriving dataset and the ones already published. The problem that this paper tackles is that of proposing an automatic algorithm that allows to identify and recommend candidate datasets for interlinking with a given novel dataset. In the presented solution, we propose a definition of a dataset profile based on the salient schema concepts that describe each dataset. We apply these profiles in order to model the datasets and measure their relatedness. We identify schema overlap by the help of a semantico-frequential concept similarity measure and a ranking criterium based on the tf*idf cosine similarity. We conduct extensive experiments over all available datasets in the LOD cloud (as of 2016), where in turn each dataset is taken as a "new" one and all the others - as linking candidates. In that, the current link topology of the LOD is used as a ground truth, where a particular discussion is dedicated to the incompleteness if these links, which may tend to lead to an overestimation of false positives in our results.

This work is the immediate result of the collaboration with the L3S institute in Hannover that had started the year before, following the research visit of S. Dietze to LIRMM at my invitation. His input and expertise in the field was particularly important and motivating for my student Mohamed. Our mutual results helped frame our discussions on the topic of dataset profiling in a more general setting, which later resulted in the publication of the survey paper cited above.

[1] https://lod-cloud.net/

Chapter 5

Conclusion

The Web of today provides unprecedented means to discover information and resources related to a multitude of topics. For example, as shown in Fig. 3.5 from Chapter 3, when looking up "Brexit" in a search engine, one comes across a rich set of resources, including articles, documents, pictures and videos related to the long lasting event of the UK leaving the European Union. Thanks to knowledge graphs, we are getting closer to a Web of structured knowledge, where question answering enables the provision of precise responses to user queries of various complexity (e.g., "when was the Brexit referendum?"). My research contributions so far, described in the chapters above, have been oriented towards of the development of methodological approaches and tools for the realisation of this Web of structured knowledge. They span a palette of research themes from ontology matching and data linking, to knowledge extraction, data modelling and knowledge graph building in a number of application domains, such as cultural heritage and social sciences.

Knowledge graphs, even if widely used for entity-centric Web-search, capture factual information mostly without keeping track of the diversity, connection or temporal evolution of viewpoints, stances and claims associated to particular entities and (controversial) topics or the contexts, in which those emerge and evolve. It is impossible to query the Web of today and retrieve a set of false claims by right-wing leaning politicians on the topic of Brexit that express an anti-immigration viewpoint. As opposed to facts stored in KGs, claims are inherently more complex, where their interpretation usually is strongly dependent on their context. A claim carries a variety of intentional or unintended meanings: subtle changes in the wording or context can have significant effects on its validity. The used terminology and the underlying conceptual understandings are still strongly diverging across communities from computational social science, to argumentation mining, automatic fact-checking, discourse analysis, or viewpoint and stance detection. There is, thus, a clear need for a shared understanding and structured knowledge about claims and related meta-data to enable machine-interpretation, discoverability and reuse, leveraging semantic web technologies, in support of scientific or journalistic studies into the analysis of online discourse, misinformation spread or biases reinforcement.

In the future, I will continuing work towards the creation of methods for doing linked open data and linked knowledge graphs by resolving various challenges in that field, as outlined in Chapter 3. For example, I have proposed to develop a framework for data linking that considers the linking problem differently, taking a bottom up and data-centric approach. A number of concrete issues, observed while working with real world data in applied scenarios, raise specific challenges that I hope to have the occasion to address, such as the problem of establishing links between knowledge graphs in the lack of shared context among their entities. Going beyond a Web of facts, it is also in the direction of claim-centric knowledge representation that I will focus my research efforts in the upcoming years. This implies

taking a research direction, where the need to account for the evolution and diversity of controversial topics, their related viewpoints and claims will drive the development of novel knowledge extraction methods and data models. They will come to enhance knowledge graphs of today with these currently missing aspects, allowing to discover, query in a meaningful way and reuse online discourse data, in service of scientific or journalistic studies and the society at large.

Bibliography

[1] Manel Achichi, Zohra Bellahsene, and Konstantin Todorov. "A survey on web data linking". In: *Revue des Sciences et Technologies de l'Information - ISI* (2016).

[2] Manel Achichi, Mohamed Ben Ellefi, Danai Symeonidou, and Konstantin Todorov. "Automatic key selection for data linking". In: *European Knowledge Acquisition Workshop*. Springer. 2016, pp. 3–18.

[3] Manel Achichi, Pasquale Lisena, Konstantin Todorov, Raphaël Troncy, and Jean Delahousse. "DOREMUS: A graph of linked musical works". In: *ISWC*. Springer. 2018, pp. 3–19.

[4] Manel Achichi, Zohra Bellahsene, Mohamed Ben Ellefi, and Konstantin Todorov. "Linking and disambiguating entities across heterogeneous RDF graphs". In: *J. of Web Semantics* (2019).

[5] Manel Achichi, Michelle Cheatham, Zlatan Dragisic, Jérôme Euzenat, et al. "Results of the ontology alignment evaluation initiative 2017?" In: *OM at ISWC*. CEUR-WS. 2017.

[6] J.-I. Akahani, K. Hiramatsu, and T. Satoh. "Approximate query reformulation based on hierarchical ontology mapping". In: *In Proc. of Intl Workshop on SWFAT*. 2003, pp. 43–46.

[7] Hunt Allcott and Matthew Gentzkow. "Social media and fake news in the 2016 election". In: *Journal of Economic Perspectives* 31.2 (2017), pp. 211–36.

[8] Manuel Atencia, Jérôme David, and Jérôme Euzenat. "Data interlinking through robust linkkey extraction." In: *ECAI*. 2014, pp. 15–20.

[9] Manuel Atencia, Jerome David, and Francois Scharffe. "Keys and Pseudo-Keys Detection for Web Datasets Cleansing and Interlinking". In: *EKAW*. 2012, pp. 144–153.

[10] Sören Auer and Jens Lehmann. "Creating knowledge out of interlinked data". In: *Semantic Web* 1.1, 2 (2010), pp. 97–104.

[11] A. Bahri, R. Bouaziz, and F. Gargouri. "Dealing with similarity relations in fuzzy ontologies". In: *Fuzzy Systems Conference, 2007. FUZZ-IEEE 2007. IEEE International*. IEEE. 2007, pp. 1–6. ISBN: 1424412099.

[12] Mohamed Ben Ellefi, Zohra Bellahsene, John G Breslin, Elena Demidova, Stefan Dietze, Julian Szymański, and Konstantin Todorov. "RDF dataset profiling–a survey of features, methods, vocabularies and applications". In: *Semantic Web* (2018).

[13] Valentina Beretta, Sébastien Harispe, Sylvie Ranwez, and Isabelle Mougenot. "Combining Truth Discovery and RDF Knowledge Bases to Their Mutual Advantage". In: *International Semantic Web Conference*. Springer. 2018, pp. 652–668.

[14] Christian Bizer and Richard Cyganiak. "Quality-driven information filtering using the WIQA policy framework". In: *Web Semantics: Science, Services and Agents on the World Wide Web* 7.1 (2009), pp. 1–10.

[15] Christian Bizer, Tom Heath, and Tim Berners-Lee. "Linked data: The story so far". In: *Semantic services, interoperability and web applications: emerging concepts*. IGI Global, 2011, pp. 205–227.

[16] F. Bobillo. "Managing vagueness in ontologies". PhD thesis. PhD thesis, University of Granada, Spain, 2008.

[17] P. Buche, J. Dibie-Barthélemy, and L. Ibanescu. "Ontology Mapping Using Fuzzy Conceptual Graphs and Rules". In: *ICCS Supplement*. 2008, pp. 17–24.

[18] Elena Cabrio and Serena Villata. "Five Years of Argument Mining: A Data-driven Analysis". In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. IJCAI'18. Stockholm, Sweden: AAAI Press, 2018, pp. 5427–5433. ISBN: 978-0-9992411-2-7. URL: http://dl.acm.org/citation.cfm?id=3304652.3304780.

[19] S. Calegari and D. Ciucci. "Fuzzy Ontology, Fuzzy Description Logics and Fuzzy-OWL". In: *Applications of Fuzzy Sets Theory*. Ed. by F. Masulli, S. Mitra, and G. Pasi. Vol. 4578. LNCS. Springer Berlin / Heidelberg, 2007, pp. 118–126.

[20] S. Calegari and E. Sanchez. "A Fuzzy Ontology-Approach to improve Semantic Information Retrieval". In: *Knowledge Creation Diffusion Utilization 7* (2007). Ed. by F. Bobillo, P. C. Da Costa, C. D'Amato, N. Fanizzi, F. Fung, T. Lukasiewicz, T. Martin, M. Nickles, Y. Peng, M. Pool, and et al. Editors, p. 6.

[21] Sylvie Cazalens, Philippe Lamarre, Julien Leblay, Ioana Manolescu, and Xavier Tannier. "A content management perspective on fact-checking". In: *The Web Conference 2018-alternate paper tracks" Journalism, Misinformation and Fact Checking"*. 2018, pp. 565–574.

[22] Michelle Cheatham and Pascal Hitzler. "String similarity metrics for ontology alignment". In: *ISWC 2013*. Springer. 2013, pp. 294–309.

[23] Pierre Choffé and Françoise Leresche. "DOREMUS: Connecting Sources, Enriching Catalogues and User Experience". In: *IFLA World Library and Information Congress*. 2016.

[24] William Jay Conover and William Jay Conover. "Practical nonparametric statistics". In: (1980).

[25] V. Cross and X. Yu. "A Fuzzy set framework for ontological similarity measures". In: *WCCI 2010, FUZZ-IEEE 2010*. IEEE Compter Society Press, 2010, pp. 1 –8.

[26] Michael Hausenblas; Richard Cygankiak. "Linked Data Life cycles". In: *formerly at http://linked-data-life-cycles.info/* ().

[27] A. Doan, J. Madhavan, P. Domingos, and A. Halevy. "Learning to map between ontologies on the semantic web". In: *WWW'02*. ACM Press, 2002, pp. 662–673. ISBN: 1581134495.

[28] Martin Doerr, Chryssoula Bekiari, and Patrick LeBoeuf. "FRBRoo: a conceptual model for performing arts". In: *CIDOC Annual Conference*. 2008, pp. 6–18.

[29] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. "Knowledge vault: A web-scale approach to probabilistic knowledge fusion". In: *ACM SIGKDD*. ACM. 2014, pp. 601–610.

[30] Mohamed Ben Ellefi, Zohra Bellahsene, Stefan Dietze, and Konstantin Todorov. "Beyond established knowledge graphs-recommending web datasets for data linking". In: *International Conference on Web Engineering*. Springer. 2016, pp. 262–279.

[31] Mohamed Ben Ellefi, Zohra Bellahsene, Stefan Dietze, and Konstantin Todorov. "Dataset recommendation for data linking: An intensional approach". In: *European Semantic Web Conference*. Springer. 2016, pp. 36–51.

[32] Mohamed Ben Ellefi, Zohra Bellahsene, J Breslin, Elena Demidova, Stefan Dietze, Julian Szymanski, and Konstantin Todorov. "Rdf dataset profiling-a survey of features, methods, vocabularies and applications". In: *Semantic Web* (2017).

[33] J. Euzenat and P. Shvaiko. *Ontology Matching*. 1st ed. Springer-Verlag, 2007. ISBN: 3540496114.

[34] Houssameddine Farah, Danai Symeonidou, and Konstantin Todorov. "KeyRanker: Automatic RDF key ranking for data linking". In: *Proceedings of the Knowledge Capture Conference*. ACM. 2017, p. 7.

[35] Daniel Faria, Catia Pesquita, Booma S Balasubramani, Catarina Martins, Joao Cardoso, Hugo Curado, Francisco M Couto, and Isabel F Cruz. "OAEI 2016 results of AML". In: *Ontology Matching ISWC, CEUR*. Vol. 1766. 2016.

[36] A. Ferrara, D. Lorusso, G. Stamou, G. Stoilos, V. Tzouvaras, and T. Venetis. "Resolution of conflicts among ontology mappings: a fuzzy approach". In: OM'08 at ISWC, 2008.

[37] Alfio Ferrara, Andriy Nikolov, and Francois Scharffe. "Data linking for the semantic web". In: *Semantic Web: Ontology and Knowledge Base Enabled Tools, Services, and Applications* 169 (2013).

[38] Alflo Ferrara, Davide Lorusso, Stefano Montanelli, and Gaia Varese. "Towards a benchmark for instance matching". In: *Ontology Matching-Volume 431*. CEUR-WS. org. 2008, pp. 37–48.

[39] Besnik Fetahu, Stefan Dietze, Bernardo Pereira Nunes, Marco Antonio Casanova, Davide Taibi, and Wolfgang Nejdl. "A scalable approach for efficiently generating structured dataset topic profiles". In: *European Semantic Web Conference*. Springer. 2014, pp. 519–534.

[40] A. Gal and P. Shvaiko. "Advances in Web Semantics I". In: ed. by T. S. Dillon, E. Chang, R. Meersman, and K. Sycara. Berlin, Heidelberg: Springer-Verlag, 2009. Chap. Advances in Ontology Matching, pp. 176–198.

[41] Aldo Gangemi. "A comparison of knowledge extraction tools for the semantic web". In: *Extended semantic web conference*. Springer. 2013, pp. 351–366.

[42] M. Gasquet, D. Brechtel, M. Zloch, K. Boland, P. Fafalios, A. Tchechmedjiev, S. Dietze, and K. Todorov. "Exploring Fact-checked Claims and their Descriptive Statistics". In: *ISWC Poster and Demo*. Springer. 2019.

[43] Daniel Gerber, Sebastian Hellmann, Lorenz Bühmann, Tommaso Soru, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. "Real-time RDF extraction from unstructured data streams". In: *International semantic web conference*. Springer. 2013, pp. 135–150.

[44] Jorge Gracia, Elena Montiel-Ponsoda, Philipp Cimiano, Asunción Gómez-Pérez, Paul Buitelaar, and John McCrae. "Challenges for the multilingual web of data". In: *Web Semantics: Science, Services and Agents on the World Wide Web* 11 (2012), pp. 63–71.

[45] Lushan Han, Abhay L. Kashyap, Tim Finin, James Mayfield, and Johnathan Weese. "UMBC_EBIQUITY-CORE: Semantic Textual Similarity Systems". In: *Proc. of the *SEM*. Association for Computational Linguistics. 2013.

[46] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. "Claimbuster: The first-ever end-to-end fact-checking system". In: *Proceedings of the VLDB Endowment* 10.12 (2017), pp. 1945–1948.

[47] Naeemul Hassan, Bill Adair, James T Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. "The quest to automate fact-checking". In: *world* (2015).

[48] Bernadette Hyland, BV Terrazas, and S Capadisli. "Cookbook for Open Government Linked Data". In: *W3C, W3C Task Force-Government Linked Data Group* (2011).

[49] A. Isaac, L. van der Meij, S. Schlobach, and S. Wang. "An Empirical Study of Instance-Based Ontology Matching". In: *The Semantic Web* (2008), pp. 253–266.

[50] Robert Isele and Christian Bizer. "Active learning of expressive linkage rules using genetic programming". In: *J. Web Sem.* 23 (2013), pp. 2–15.

[51] Robert Isele and Christian Bizer. "Active learning of expressive linkage rules using genetic programming". In: *Web Semantics* 23 (2013), pp. 2–15.

[52] Robert Isele and Christian Bizer. "Learning linkage rules using genetic programming". In: *Ontology Matching-Volume 814*. CEUR-WS. org. 2011, pp. 13–24.

[53] Nicolas James, Konstantin Todorov, and Céline Hudelot. "Combining visual and textual modalities for multimedia ontology matching". In: *International Conference on Semantic and Digital Media Technologies*. Springer. 2010, pp. 95–110.

[54] Nicolas James, Konstantin Todorov, and Céline Hudelot. "Ontology matching for the semantic annotation of images". In: *International Conference on Fuzzy Systems*. IEEE. 2010, pp. 1–8.

[55] Anja Jentzsch, Robert Isele, and Christian Bizer. "Silk-generating rdf links while publishing or consuming linked data". In: *ISWC*. 2010.

[56] Ernesto Jimenez-Ruiz and Bernardo Cuenca Grau. "Logmap: Logic-based and scalable ontology matching". In: *ISWC*. Springer. 2011, pp. 273–288.

[57] Seyed Mehran Kazemi and David Poole. "Simple embedding for link prediction in knowledge graphs". In: *Advances in Neural Information Processing Systems*. 2018, pp. 4284–4295.

[58] Mayank Kejriwal and Daniel P. Miranker. "An unsupervised instance matcher for schema-free RDF data". In: *J. Web Sem.* 35 (2015), pp. 102–123.

[59] Mayank Kejriwal and Daniel P. Miranker. "Semi-supervised Instance Matching Using Boosted Classifiers". In: *ESWC*. 2015, pp. 388–402.

[60] M. S. Lacher and G. Groh. "Facilitating the Exchange of Explicit Knowledge Through Ontology Mappings". In: *Proceedings of the 14th FLAIRS Conf.* AAAI Press, 2001, pp. 305–309.

[61] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. "DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia". In: *Semantic Web* 6.2 (2015), pp. 167–195.

[62] Tatiana Lesnikova, Jerome David, and Jerome Euzenat. "Interlinking English and Chinese RDF Data Sets Using Machine Translation". In: *ESWC workshop Know@ LOD*. Vol. 2013. 2014.

[63] Tatiana Lesnikova, Jerome David, and Jerome Euzenat. "Interlinking English and Chinese RDF Data Using BabelNet". In: *ACM DocEng*. 2015, pp. 39–42.

[64] Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. "Context dependent claim detection". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 2014, pp. 1489–1500.

[65] Chao Li, Lei Ji, and Jun Yan. "Acronym Disambiguation Using Word Embedding". In: *AAAI*. 2015, pp. 4178–4179.

[66] Dekang Lin. "An Information-Theoretic Definition of Similarity". In: *Proc. of ICML*. 1998, pp. 296–304.

[67] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. "Learning entity and relation embeddings for knowledge graph completion". In: *Twenty-ninth AAAI conference on artificial intelligence*. 2015.

[68] Marco Lippi and Paolo Torroni. "Argumentation mining: State of the art and emerging trends". In: *ACM Transactions on Internet Technology (TOIT)* 16.2 (2016), p. 10.

[69] Pasquale Lisena, Konstantin Todorov, Cecile Cecconi, Francoise Leresche, Isabelle Canno, Frederic Puyrenier, Martine Voisin, and Raphael Troncy. "Controlled Vocabularies for Music Metadata". In: *ISMIR*. 2018.

[70] J. Madhavan, P. A. Bernstein, and E. Rahm. "Generic Schema Matching with Cupid". In: *The VLDB Journal*. 2001, pp. 49–58.

[71] Jose L Martinez-Rodriguez, Aidan Hogan, and Ivan Lopez-Arevalo. "Information extraction meets the semantic web: a survey". In: *Semantic Web* Preprint (2018), pp. 1–81.

[72] D. L. McGuinness, R. Fikes, J. Rice, and S. Wilder. "An Environment for Merging and Testing Large Ontologies". In: *Proc. 17th Intl. Conf. on Principles of Knowledge Representation and Reasoning (KR'2000)*. Colorado,USA, 2000, pp. 483–493.

[73] Jose G Moreno, Romaric Besançon, Romain Beaumont, Eva D?hondt, Anne-Laure Ligozat, Sophie Rosset, Xavier Tannier, and Brigitte Grau. "Combining word and entity embeddings for entity linking". In: *European Semantic Web Conference*. Springer. 2017, pp. 337–352.

[74] Fedelucio Narducci, Matteo Palmonari, and Giovanni Semeraro. "Cross-lingual link discovery with TR-ESA". In: *Inf. Sci.* 394 (2017), pp. 68–87.

[75] Roberto Navigli and Simone Paolo Ponzetto. "BabelNet: Building a very large multilingual semantic network". In: *48th annual meeting of the association for computational linguistics*. ACL. 2010, pp. 216–225.

[76] Markus Nentwig, Michael Hartung, Axel-Cyrille Ngonga Ngomo, and Erhard Rahm. "A survey of current Link Discovery frameworks". In: *Semantic Web* (2015), pp. 1–18.

[77] Markus Nentwig, Michael Hartung, Axel-Cyrille Ngonga Ngomo, and Erhard Rahm. "A survey of current link discovery frameworks". In: *Semantic Web* 8.3 (2017), pp. 419–436.

[78] DuyHoa Ngo, Zohra Bellahsene, and Konstantin Todorov. "Extended tversky similarity for resolving terminological heterogeneities across ontologies". In: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer. 2013, pp. 711–718.

[79] Axel-Cyrille Ngonga Ngomo and Sören Auer. "LIMES - a time-efficient approach for large-scale link discovery on the web of data". In: *IJCAI*. 2011.

[80] Axel-Cyrille Ngonga Ngomo and Klaus Lyko. "Eagle: Efficient active learning of link specifications using genetic programming". In: *ESWC*. Springer. 2012, pp. 149–163.

[81] Axel-Cyrille Ngonga Ngomo and Klaus Lyko. "EAGLE: Efficient Active Learning of Link Specifications Using Genetic Programming". In: *ESWC*. 2012, pp. 149–163.

[82] Axel-Cyrille Ngonga Ngomo, Klaus Lyko, and Victor Christen. "COALA - Correlation-Aware Active Learning of Link Specifications". In: *ESWC*. 2013, pp. 442–456.

[83] Axel-Cyrille Ngonga Ngomo, Jens Lehmann, Sören Auer, and Konrad Höffner. "Raven-active learning of link specifications". In: *Ontology Matching*. CEUR-WS. org. 2011, pp. 25–36.

[84] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. "A review of relational machine learning for knowledge graphs". In: *Proceedings of the IEEE* 104.1 (2015), pp. 11–33.

[85] Andriy Nikolov, Mathieu d'Aquin, and Enrico Motta. "Unsupervised Learning of Link Discovery Configuration". In: *ESWC*. 2012, pp. 119–133.

[86] Andriy Nikolov, Mathieu dquin, and Enrico Motta. "Unsupervised learning of link discovery configuration". In: *ESWC*. Springer. 2012, pp. 119–133.

[87] Andriy Nikolov, Victoria Uren, and Enrico Motta. "KnoFuss: A comprehensive architecture for knowledge fusion". In: *K-Cap*. ACM. 2007, pp. 185–186.

[88] Andriy Nikolov, Victoria S. Uren, Enrico Motta, and Anne N. De Roeck. "Integration of Semantically Annotated Data by the KnoFuss Architecture". In: *EKAW*. 2008, pp. 265–274.

[89] Gerret von Nordheim, Karin Boczek, and Lars Koppers. "Sourcing the Sources". In: *Digital Journalism*. 2018, pp. 807–828.

[90] N. Noy and M. Musen. "Anchor-PROMPT: Using Non-Local Context for Semantic Matching". In: *Workshop on Ontologies and Information Sharing at IJCAI*. 2001, pp. 63–70.

[91] Bernardo Pereira Nunes, Stefan Dietze, Marco Antonio Casanova, Ricardo Kawase, Besnik Fetahu, and Wolfgang Nejdl. "Combining a co-occurrence-based and a semantic measure for entity linking". In: *Extended Semantic Web Conference*. Springer. 2013, pp. 548–562.

[92] Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. "Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features". In: *NAACL*. 2018.

[93] Heiko Paulheim. "Knowledge graph refinement: A survey of approaches and evaluation methods". In: *Semantic web* 8.3 (2017), pp. 489–508.

[94] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. "Where the truth lies: Explaining the credibility of emerging claims on the web and social media". In: *WWW*. 2017, pp. 1003–1012.

[95] Mainul Quraishi, Pavlos Fafalios, and Eelco Herder. "Viewpoint Discovery and Understanding in Social Networks". In: *ACM Web Science*. 2018, pp. 47–56.

[96] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. "Truth of varying shades: Analyzing language in fake news and political fact-checking". In: *EMNLP*. 2017, pp. 2931–2937.

[97] Petar Ristoski and Heiko Paulheim. "Semantic Web in data mining and knowledge discovery: A comprehensive survey". In: *Web semantics: science, services and agents on the World Wide Web* 36 (2016), pp. 1–22.

[98] Lior Rokach and Oded Maimon. "Clustering Methods". In: *The Data Mining and Knowledge Discovery Handbook*. 2005, pp. 321–352.

[99] Shu Rong, Xing Niu, Evan Wei Xiang, Haofen Wang, Qiang Yang, and Yong Yu. "A machine learning approach for instance matching based on similarity metrics". In: *ISWC*. Springer, 2012, pp. 460–475.

[100] Sebastian Ruder, Ivan Vulić, and Anders Søgaard. "A survey of cross-lingual word embedding models". In: *arXiv preprint arXiv:1706.04902* (2017).

[101] E. Sanchez and T. Yamanoi. "Fuzzy Ontologies for the Semantic Web". In: *Flexible Query Answering Systems* (2006), pp. 691–699.

[102] Tzanina Saveta, Evangelia Daskalaki, Giorgos Flouris, Irini Fundulaki, Melanie Herschel, and Axel-Cyrille Ngonga Ngomo. "LANCE: Piercing to the Heart of Instance Matching Tools". In: *ISWC*. 2015, pp. 375–391.

[103] Tzanina Saveta, Evangelia Daskalaki, Giorgos Flouris, Irini Fundulaki, Melanie Herschel, and Axel-Cyrille Ngonga Ngomo. "Pushing the limits of instance matching systems: A semantics-aware benchmark for linked data". In: *WWW*. ACM. 2015, pp. 105–106.

[104] Francois Scharffe, Yanbin Liu, and Chuguang Zhou. "Rdf-ai: an architecture for rdf datasets matching, fusion and interlink". In: *IJCAI 2009 workshop IR-KR*. 2009.

[105] Chao Shao, Linmei Hu, Juan-Zi Li, Zhichun Wang, Tong Lee Chung, and Jun-Bo Xia. "RiMOM-IM: A Novel Iterative Framework for Instance Matching". In: *J. Comput. Sci. Technol.* 31.1 (2016), pp. 185–197.

[106] Mohamed Ahmed Sherif, Axel-Cyrille Ngonga Ngomo, and Jens Lehmann. "W ombat–a generalization approach for automatic link discovery". In: *ESWC*. Springer. 2017, pp. 103–119.

[107] Baoxu Shi and Tim Weninger. "Discriminative predicate path mining for fact checking in knowledge graphs". In: *Knowledge-based systems* 104 (2016), pp. 123–133.

[108] Prashant Shiralkar, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. "Finding streams in knowledge graphs to support fact checking". In: *ICDM*. IEEE. 2017, pp. 859–864.

[109] Pavel Shvaiko and Jérôme Euzenat. "A survey of schema-based matching approaches". In: *Journal on data semantics IV*. Springer, 2005, pp. 146–171.

[110] Pavel Shvaiko and Jérôme Euzenat. "Ontology matching: state of the art and future challenges". In: *IEEE Transactions on knowledge and data engineering* 25.1 (2011), pp. 158–176.

[111] Pavel Shvaiko and Jérôme Euzenat. "Ontology matching: state of the art and future challenges". In: *IEEE Transactions on knowledge and data engineering* 25.1 (2013), pp. 158–176.

[112] Tommaso Soru, Edgard Marx, and Axel-Cyrille Ngonga Ngomo. "ROCKER: A Refinement Operator for Key Discovery". In: *WWW*. 2015, pp. 1025–1033.

[113] Tommaso Soru, Edgard Marx, and Axel-Cyrille Ngonga Ngomo. "ROCKER: a refinement operator for key discovery". In: *WWW*. ACM. 2015, pp. 1025–1033.

[114] Dennis Spohr, Laura Hollink, and Philipp Cimiano. "A machine learning approach to multilingual and cross-lingual ontology matching". In: *International Semantic Web Conference*. Springer. 2011, pp. 665–680.

[115] H. Stuckenschmidt. "Approximate Information Filtering on the Semantic Web". In: *KI 2002*. Ed. by M. Jarke, G. Lakemeyer, and J. Koehler. Vol. 2479. LNCS. Springer Berlin / Heidelberg, 2002, pp. 195–228.

[116] G. Stumme and A. Maedche. "FCA-Merge: Bottom-up merging of ontologies". In: *IJCAI*. 2001, pp. 225–230.

[117] Danai Symeonidou, Nathalie Pernelle, and Fatiha Saïs. "KD2R: A Key Discovery Method for Semantic Reference Reconciliation". In: *On the Move to Meaningful Internet Systems: OTM 2011 Workshops*. 2011, pp. 392–401.

[118] Danai Symeonidou, Isabelle Sanchez, M. Croitoru, P. Neveu, N. Pernelle, F. Sais, A. Roland-Vialaret, P. Buche, A.-R. Muljarto, and R. Schneider. "Key Discovery for Numerical Data: Application to Oenological Practices". In: *ICCS*. 2016, pp. 222–236.

[119] Danai Symeonidou, Vincent Armant, Nathalie Pernelle, and Fatiha Sais. "SAKey: Scalable Almost Key Discovery in RDF Data". In: *ISWC*. 2014, pp. 33–49.

[120] Danai Symeonidou, Vincent Armant, Nathalie Pernelle, and Fatiha Saïs. "Sakey: Scalable almost key discovery in RDF data". In: *ISWC*. Springer. 2014, pp. 33–49.

[121] Danai Symeonidou, Luis Galarraga, Nathalie Pernelle, Fatiha Saïs, and Fabian Suchanek. "VICKEY: Mining Conditional Keys on Knowledge Bases". In: *ISWC*. 2017.

[122] Andon Tchechmedjiev, Pavlos Fafalios, Katarina Boland, Stefan Dietze, Benjamin Zapilko, and Konstantin Todorov. "ClaimsKG - A Knowledge Graph of Fact-checked Claims". In: *ISWC*. 2019.

[123] Andon Tchechmedjiev, Théophile Mandon, Mathieu Lafourcade, Anne Laurent, and Konstantin Todorov. "Ontolex JeuxDeMots and Its Alignment to the Linguistic Linked Open Data Cloud". In: *International Semantic Web Conference*. Springer. 2017, pp. 678–693.

[124] Thibaut Thonet, Guillaume Cabanac, Mohand Boughanem, and Karen Pinel-Sauvagnat. "Vodum: A topic model unifying viewpoint, topic and opinion discovery". In: *European Conference on Information Retrieval*. Springer. 2016, pp. 533–545.

[125] Abdel Nasser Tigrine, Zohra Bellahsene, and Konstantin Todorov. "Lightweight cross-lingual ontology matching with LYAM++". In: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer. 2015, pp. 527–544.

[126] Abdel Nasser Tigrine, Zohra Bellahsene, and Konstantin Todorov. "Selecting optimal background knowledge sources for the ontology matching task". In: *European Knowledge Acquisition Workshop*. Springer. 2016, pp. 651–665.

[127] K. Todorov, P. Geibel, and K.-U. Khnberger. "Mining Concept Similarities for Heterogeneous Ontologies". In: *Advances in Data Mining. Applications and Theoretical Aspects*. Ed. by P. Perner. Vol. 6171. LNCS. Springer Berlin / Heidelberg, 2010, pp. 86–100.

[128] K. Todorov, N. James, and C. Hudelot. "Multimedia Ontology Matching by Using Visual and Textual Modalities". In: *Multimedia Tools and Applications* (2011), pp. 1–25. ISSN: 1380-7501.

[129] Konstantin Todorov. "Combining structural and instance-based ontology similarities for mapping web directories". In: *2008 Third International Conference on Internet and Web Applications and Services*. IEEE. 2008, pp. 596–601.

[130] Konstantin Todorov. "Datasets First! A Bottom-up Data Linking Paradigm". In: *Procs of the ISWC 2019 Satellite Tracks (Outrageous Ideas)*. 2019, pp. 338–342.

[131] Konstantin Todorov and Peter Geibel. "Ontology mapping via structural and instance-based similarity measures". In: *The 7th International Semantic Web Conference*. 2008, p. 224.

[132] Konstantin Todorov, Peter Geibel, and Céline Hudelot. "A framework for a fuzzy matching between multiple domain ontologies". In: *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer. 2011, pp. 538–547.

[133] Konstantin Todorov, Peter Geibel, and Kai-Uwe Kuehnberger. "Extensional ontology matching with variable selection for support vector machines". In: *2010 International Conference on Complex, Intelligent and Software Intensive Systems*. IEEE. 2010, pp. 962–967.

[134] Konstantin Todorov, Peter Geibel, and Kai-Uwe Kühnberger. "Mining concept similarities for heterogeneous ontologies". In: *Industrial Conference on Data Mining*. Springer. 2010, pp. 86–100.

[135] Konstantin Todorov, Celiné Hudelot, and Peter Geibel. "Fuzzy and cross-lingual ontology matching mediated by background knowledge". In: *Uncertainty Reasoning for the Semantic Web III*. Springer, 2012, pp. 142–162.

[136] Konstantin Todorov, Nicolas James, and Céline Hudelot. "Multimedia ontology matching by using visual and textual modalities". In: *Multimedia tools and applications* 62.2 (2013), pp. 401–425.

[137] Konstantin Todorov, Celine Hudelot, Adrian Popescu, and Peter Geibel. "Fuzzy ontology alignment using background knowledge". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 22.01 (2014), pp. 75–112.

[138] Alberto Tonon, Victor Felder, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. "Voldemortkg: Mapping schema. org and web entities to linked open data". In: *ISWC*. Springer. 2016, pp. 220–228.

[139] Ignacio Traverso-Ribón and Maria-Esther Vidal. "GARUM: a semantic similarity measure based on machine learning and entity characteristics". In: *International Conference on Database and Expert Systems Applications*. Springer. 2018, pp. 169–183.

[140] Sahar Vahdati, Guillermo Palma, Rahul Jyoti Nath, Christoph Lange, Sören Auer, and Maria-Esther Vidal. "Unveiling scholarly communities over knowledge graphs". In: *International Conference on Theory and Practice of Digital Libraries*. Springer. 2018, pp. 103–115.

[141] Maria-Esther Vidal, Kemele M Endris, Samaneh Jozashoori, Farah Karim, and Guillermo Palma. "Semantic data integration of big biomedical data for supporting personalised medicine". In: *Current Trends in Semantic Web Technologies: Theory and Practice*. Springer, 2019, pp. 25–56.

[142] Boris Villazón-Terrazas, Luis M Vilches-Blázquez, Oscar Corcho, and Asunción Gómez-Pérez. "Methodological guidelines for publishing government linked data". In: *Linking government data*. Springer, 2011, pp. 27–49.

[143] Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. "Silk-A Link Discovery Framework for the Web of Data." In: *LDOW* 538 (2009).

[144] Soroush Vosoughi, Deb Roy, and Sinan Aral. "The spread of true and false news online". In: *Science* 359.6380 (2018), pp. 1146–1151.

[145] Shangsi Wang, Jesus Arroyo, Joshua T Vogelstein, and Carey E Priebe. "Joint embedding of graphs". In: *arXiv preprint arXiv:1703.03862* (2017).

[146] William Yang Wang. "Liar, liar pants on fire: A new benchmark dataset for fake news detection". In: *AMACL* (2017), pp. 422–426.

[147] Gerhard Weikum, Johannes Hoffart, and Fabian Suchanek. "Knowledge harvesting: achievements and challenges". In: *Computing and Software Science*. Springer, 2019, pp. 217–235.

[148] Gerhard Weikum and Martin Theobald. "From information to knowledge: harvesting entities and relationships from web sources". In: *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM. 2010, pp. 65–76.

[149] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific data* 3 (2016).

[150] You Wu, Pankaj K Agarwal, Chengkai Li, Jun Yang, and Cong Yu. "Toward computational fact-checking". In: *Proceedings of the VLDB Endowment* 7.7 (2014), pp. 589–600.

[151] Zhibiao Wu and Martha Palmer. "Verbs Semantics and Lexical Selection". In: *Proc. of the 32Nd ACL*. 1994, pp. 133–138.

[152] B. Xu, D. Kang, J. Lu, Y. Li, and J. Jiang. "Mapping Fuzzy Concepts Between Fuzzy Ontologies". In: *Knowledge-Based Intelligent Information and Engineering Systems*. Ed. by R. Khosla, R. J. Howlett, and L. C. Jain. Vol. 3683. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2005, pp. 199–205.

[153] Yasunori Yamamoto, Atsuko Yamaguchi, Hidemasa Bono, and Toshihisa Takagi. "Allie: a database and a search service of abbreviations and long forms". In: *Database* 2011 (2011).

[154] Ran Yu, Ujwal Gadiraju, Besnik Fetahu, and Stefan Dietze. "Fusem: Query-centric data fusion on structured web markup". In: *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*. IEEE. 2017, pp. 179–182.

[155] Ran Yua, Ujwal Gadirajua, Besnik Fetahua, Oliver Lehmbergb, Dominique Ritzeb, and Stefan Dietzea. "KnowMore-Knowledge Base Augmentation with Structured Web Markup". In: *Semantic Web J., IOS Press* (2017).

[156] Antoine Zimmermann, Markus Krotzsch, Jérôme Euzenat, and Pascal Hitzler. "Formalizing ontology alignment and its operations with category theory". In: *Proc. 4th International Conference on Formal Ontology in Information Systems*. 2006.

Appendix A

Extended Curriculum Vitae

Curriculum Vitae

Konstantin Todorov

Associate professor in Computer Science
PhD in Cognitive Science

Personal
Born on June 6, 1981 in Bulgaria

20 rue Saint Guilhem
34000 Montpellier -France
konstantin.get@gmail.com
+33 6 99438352

Professional
Université Montpellier
LIRMM UMR 5506, CC477
161 rue Ada
34095 Montpellier Cedex 5 - France
konstantin.todorov@umontpellier.fr
konstantin.todorov@lirmm.fr

Homepage: *https://www.lirmm.fr/users/utilisateurs-lirmm/konstantin-todorov*
Languages: Bulgarian (9/10), English (8/10), French (8/10), German (7/10), Spanish (6/10)

Areas of scientific interest:
Web Data Science, Knowledge Extraction, Linked Data, Knowledge Graphs, Fuzzy and Multilingual Ontology Matching, Semantics and Multimedia

Methodologies:
Machine Learning, Natural Language Processing, Fuzzy Sets and Theory

Application fields:
Cultural heritage, Plant biology and agronomy, Computational sociology, Fact-checking

Professional appointments

Since 2012	***Associate professor*** (Maître de conférences) at the university of Montpellier, LIRMM (Laboratoire d'informatique, robotique et microélectronique de Montpellier), member and co-founder of the FADO team (Fuzzinness, Alignments, Data and Ontologies)
2017 - 2018	***A 6 months paid sabbatical leave*** (*CRCT*) dedicated to the development of a novel research direction (analysis of societal debates on the web) and the building of a new scientific network. Resulted in the submission of a joint ANR-DFG (Germany) PRCI proposal.
2009 - 2012	***Postdoctoral fellow*** at the MAS Laboratory, Ecole Centrale Paris, France. Employed on the ANR Collaviz project in the field of multimedia semantics.

Education & degrees

2012	***Qualification for the function of assoc. professor*** / *Qualification aux fonctions de MC, CNU 27*

2009 ***PhD in Cognitive Science***
Institute of Cognitive Science (IKW), University of Osnabrück, Germany
Title: "Ontology Matching by Combining Instance-based Concept Similarity Measures with Structure"
Supervised by Prof. Dr Kai-Uwe Kühnberger (IKW) and PD Dr. Peter Geibel (Technical University of Berlin)

Co-funded on a DAAD (the German Academic Exchange Service) grant and a Research Assistant grant from Lower Saxony, Germany

2004 ***M.Sc. / DEA in Applied Mathematics*** (Probabilities and Statistics), option Statistical Inference & Machine learning
University of Provence, Marseille, France

2003 ***Bachelor degree studies in Mathematical Modelling***
Ecole Centrale Marseille, France

Funded on an Erasmus scholarship on a bilateral agreement between the Technical University of Sofia, Bulgaria and ESM2 Marseille (Ecole Centrale Marseille)

2000 - 2003 ***Bachelor degree studies in Applied Mathematics***
Technical University of Sofia, Bulgaria

Best student award (2002) at the Faculty of Applied Mathematics and Informatics, Technical University of Sofia, Bulgaria

2000 ***High school graduation***
High School of Natural Science, Veliko Tarnovo, Bulgaria

Teaching activities

University of Montpellier (since 2012):

Courses

Master: Data mining, Machine learning, Semantic Web, Statistical Data Analysis
Bachelor: Discrete Mathematics

Student project supervision

1st year master projects in research and development (TER) (approx. 3 groups of 4 students per year since 2013)

Ecole Centrale Paris (2009 – 2012):

Courses

Algorithmics and Programming (Python), Information retrieval

Student project supervision

1st year students projects in Information Retrieval (two groups)

University of Osnabrück

Seminars

Machine learning, Concepts and words

Administrative responsibilities at the University of Montpellier

Since 2014	Head of the master's program IPS (Computer science for other sciences) at the University of Montpellier. This responsibility implies selecting the students among approx. 400 applications annually, establishing jointly with other master programs the study plan, participating at bi-weekly faculty staff meetings, organisation of bi-semestrial meetings with the students, handling various individual issues, preparation of the accreditation of the program (every 4 years). The master's program accepts approximately 25 students in first year and 15 students in second year, implying the management of 40 students annually on average.
Since 2014	Organizer of 1st year master projects for research and development (TER) for IPS (approx. 25 students per year). This responsibility implies the collect of project proposals from colleagues, the organisation of an introductory meeting for the students in order for them to choose among the proposed projects, supervision and management of potential problems in terms of communication between students and tutors, organisation of defenses and heading the defense jury, organisation of a poster session presenting the different realised projects at the end of the term.
Since 2012	Pedagogical manager of the following master courses: Data mining, Data Science I, Data Science II, Semantic Web, Networks I, Networks II. These responsibilities imply, beyond teaching and defining the study programme, preparing the content and schedule of each course and managing the pedagogical team (including both faculty members and external speakers).

Supervision

Doctoral supervision

Mohamed Ben Ellefi Defended December, 2016	Title: "Profile-based Dataset Recommendation for Data Linking" Co-supervised with Zohra Bellahsene (LIRMM)
Manel Achichi Defended February, 2018	Title: "Linking and Publishing Heterogeneous Open Data in the Musical Field" Co-supervised with Zohra Bellasene
Abdelnasser Tigrine Abandoned after the 3rd year without defense in 2017	Title: "Using Background Knowledge for Facilitating the Ontology Matching Task" Co-supervised with Zohra Bellasene
Katarina Boland planned to start in 2019	Title: "Identifying Claim Sources and Their Types for the Analysis of Societal Debates" Co-supervised with Stefan Dietze (University of Düsseldorf)

Master (research) theses supervision (6 months)

Manel Achichi Defended June, 2014	Title: "Extraction de données liées à partir de tweets pour leur publication sur le web de données" Co-supervised with Z. Bellahsene and Dino Ienco (CIRAD Montpellier)
Imène Chentli Defended August, 2015	Title: "Facilitation de l'accès aux données biologiques structurées" Co-supervised with Pierre Larmande (IRD Montpellier)
Sara Remini Defended August, 2016	Title: "Acquisition automatique de connaissances à partir de textes scientifiques" Co-supervised with Pierre Larmande (IRD Montpellier)
Tayeb Ghofri Defended June, 2016	Title: "Interlinking Heterogeneous Musical Data on the Web of Data"
Houssemeddine Farah Defended June, 2017	Title: "Key Ranking for Data Linking" Co-supervised with Danai Symeonidou (INRA Montpellier)
Théophile Mandon Defended June, 2017	Title: "Transformation de la ressource langagière JeuxDeMots en RDF et interconnexion au web de données" Co-supervised with Mathieu Laforcade and Anne Laurent (LIRMM)
Ahmed Sayadi Defended June, 2018	Title: "Linking complementary RDF datasets" Co-supervised with Pierre Larmande
Jérémy Bressant Defended June, 2018	Title: "Fake news detection: identifying salient features" Co-supervised with Mathieu Roche (CIRAD Montpellier) and Stefan Dietze (University of Düsseldorf)
Serge Sonfack defense in December 2019	Title: "Multimodal embeddings for complementary data linking" Co-supervised with Pierre Larmande
Olivier Kadima on-going	Title: "Predicting climate change impact on wine production data" Co-supervised with Danai Symeonidou
Gaoussou Sanou to start in February 2020	Title: "Une base de connaissance en immunogénétique pour la découverte de la nouvelle connaissance scientifique" Co-supervised with Patrice Duroux, Véronique Giudicelli and Sofia Kossida (IGH Montpellier)
Rémi Cérès to start in February 2020	Title: "Hybridation des méthodes d'apprentissage et Web sémantique pour l'optimisation et la planification de cultures maraîchères en agro-écolo" Co-supervised with Clément Jonquet (LIRMM), Florence Amardeihl (société Elzeard)

Master's first year internships supervision (2-3 months, no defense)

Mykael Vigo 2015	Title: "Twitter Event Detection and Modeling." Co-supervised with Dino Ienco
Marius Vekleber 2018	Title: "Linking music-related metadata"
Quentin Monod 2018	Title: "Building a pivot knowledge graph of linked data"
Leonardo Moros 2019	Title: "Extracting salient features from fact-checked claims" Co-supervised with Mathieu Roche
Bastien Carbonier 2019	Title: "Structuring and semantizing human genetics data" Co-supervised with Sofia Kossida (IGH, Montpellier)
Reda Souadi 2019	Title: "Fact-checking web-site crawling and structuring of related meta-data."
Malo Gasquet 2019	Title: "A web-application for exploring fact-checked data" Co-supervised with Andon Tchechmedjiev (Ecole des Mine, Alès)

Research activities

Third-Party Funding

Since 2014, I have brought over 240K euros for the University of Montpellier and the LIRMM laboratory via national and international projects that I have either co-lead or participated to. I provide a list of the projects, in which I have been involved over the past years with a short description of their objectives and my role in terms of research and resource management.

ANR DOREMUS (ANR-14-CE24-0020) 2014-2018
Coordinated by Laurent Bouvier-Ajam (OUROUK, Paris)

Role: PI, WP leader (Data Lifting (semanticization) and Linking), co-supervisor of the PhD student Manel Achichi. In that context, I worked on the development of tools for data linking and vocabulary alignment that are both generic and tailored to the needs of the project.
Resources for LIRMM/UM: A PhD student salary over 3 years (105K euros), 3 master internships (9K euros) and traveling budget (approx. 25K euros)
Description: Three major French cultural institutions—the French National Library (BnF), Radio France and the Philharmonie de Paris—have joint efforts with computer scientists (LIRMM and EURECOM) in order to develop shared methods to describe semantically their catalogs of music works and events. This process comprises the construction of knowledge graphs representing the data contained in these catalogs following a novel agreed upon ontology that extends CIDOC-CRM and FRBRoo, the linking of these graphs via novel data integration solutions and their open publication on the web.

LDCT PHC Amadeus (N° 33791NF) between LIRMM and IST Innsbruck (Austria), 2015-2016

Role: Coordinator.
Resources for LIRMM/UM: traveling budget for two research visits in IST Innsbruck
Description: The project aims to develop shared tools for the integration of cultural and tourism in common applications, leaning upon IST's expertise in the field of tourism and LIRMM's experience in cultural heritage gained in the DOREMUS project.

PIA DATALYSE (FSN-AAP Big Data n3), 2013-2016

Role: Participant. My mission was to provide for the open data needs of the project, particularly the semantization of the data of the City of Grenoble, as well as the development of a dataset recommendation engine. I have co-supervised the PhD student Mohamed Ben Ellefi funded on the project.
Resources for LIRMM/UM: A PhD student salary over 3 years (105K euros) and traveling budget
Description: In the context of collaboration between academia and industry, the project aims at creating Big Data applications for the development of intelligent warehouses of heterogeneous and massive customer/consumer and public data.

ANR PratikPharma (ANR-15-CE23-0028) 2016 - 2019
Coordinated by Adrien Coulet (Loria, Nancy)

Role: Participant. I have been involved in WPs and tasks related to the multilingual alignment of semantic resources by the use of background knowledge. My PhD student Nacer Tigrine has worked in collaboration with the project consortium
Description: PractiKPharma develops computer science approaches to extract, compare, validate state of the art knowledge in the biomedical domain of Pharmacogneomics (PGx) - the study of how genetics impacts drug response phentoypes. Units of knowledge in PGx typically have the form of ternary relationships gene variant–drug–adverse event, and can be formalized in various manners, in particular with the help of (biomedical) ontologies semantic web technologies.

ANR D2KAB started in 2019
Coordinated by Clément Jonquet (LIRMM, Montpellier)

Role: Participant, involved in WPs and tasks related to data linking and ontology alignment.
Description: The project creates a framework to turn agronomy and biodiversity data into knowledge –semantically described, interoperable, actionable, open– and investigate scientific methods and tools to exploit this knowledge for applications in science & agriculture.

Non-funded projects
In addition, I have coordinated or participated in the submission of several unsuccessful proposals to the following calls: ANR PRCI French-German collaboration with DFG (2019) (coordinator), I-site MUSE (2018), DigitAG (2019), H2020-MSCA-ITN (2019). The ANR-DFG PRCI and the H2020-ITN proposals are currently in the process of their resubmission.

Scientific forums organisation

- OAEI Instance Matching track co-organiser in 2016 and 2017.
- Linked Data tutorial at J-DEV 2017, Marseille (co-organiser)
- DOREMUS tutorial at ESWC 2016 (co-organiser). The lifecycle of music bibliographical data: transformation to RDF, model creation, data linking and publication on the web of data

- Special Session “Uncertainty and imprecision on the web of data” at IPMU 2014, International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems
- OMIR workshop -International Workshop on Ontologies for Multimedia Interpretation and Retrieval- at SAMT 2010, International Conference on Semantic and Digital Media Technologies

International Programme Committees and International Journals Reviewing (a selection)

- ISWC 2017- 2019 - International Semantic Web Conference 2017-2019 (Research, Resource)
- ESWC 2018 - 2020 - Extended Semantic Web Conference 2018-2020 (Research, Knowledge Graphs)
- WWW 2017-2020 - The Web Conference 2017-2020 (Semantics and Knowledge)
- SEMANTiCS 2016-2019 - International Conference on Semantic Systems (Research, Posters&Demo)
- ACM Web Science 2017 (Poster)
- COOPiS 2019 - 27th International Conference on Cooperative Information Systems (Research)
- IPMU 2016 - International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (Research)
- Frequent reviewing for the following international journals: SWJ (The Semantic Web Journal), JWS (Journal of Web Semantics), JoDS (Journal of Data Semantics), FSS (Fuzzy Sets and Systems), IJUFKS (International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems), Discrete Applied Mathematics.

National programme committees and grant applications reviewing
- IC (Ingénierie de connaissance), 2019
- Reviewer for the ANR MRSEI call in October 2018

Keynote and invited talks
- Invited talk at the MIDI seminars at ENSEA, Laval University, Cergy, 2012
- Keynote at SoWeDo at IC2016, Montpellier, France, 2016
- Invited plenary talk at J-DEV, CNRS, Marseille, France, 2017
- Invited talk at LSIS, Marseille, France, 2017
- Keynote at SMDI at EKAW, Nancy, France, 2019
- Invited talk at GESIS, Cologne, Germany, 2019
- Invited talk at the CONTROVERSE workshop, Montpellier, 2019

Research visits (approx. duration 1 week)
- Inria Grenoble (France), visiting Jérôme Euzenat and the Moex team, 2009
- L3S Hannover (Germany), visit funded by an EU COST STSM grant, 2017
- GESIS Cologne (Germany), visit funded by Prof. Stefan Dietze and his department, 2019

Grants & awards
- PEDR grant for excellent research and supervision - 20K€ (2018-2022)
- CRCT (payed research sabbatical leave) (2018-2019)
- DAAD PhD grant (2007-2008)
- EU COST STSM traveling grant (2017)
- ISWC2019 Blue Sky Ideas first prize for “Outrageous ideas” paper - 1000$ (2019)

Administrative duties
- Member of the committee for local promotions (section 27) at LIRMM since 2018

Juries

- Selection committee member for the position of associate professor (Maître de conférences) at Ecole Centrale Paris - Supelec, 2018
- Frequent member of juries of master theses defenses since 2014
- Frequent participation at PhD theses mid-term evaluation committees (*comités de suivi individuel*), approximately 4 committees per year since 2015 both at LIRMM and at external institutes (e.g., EspaceDev, Ecole des Mines - Alès).

Tools and applications

I have steered the development of a number of open-source applications and tools in support of data integration or information retrieval for both computer (data) scientists and domain experts with no computer science background. These tools have been developed in close collaboration with my PhD and master students, leading to the publication of research or demo papers in top-tier conferences.

- *Claims Explorer:* an engine to conduct ad-hoc/faceted search over a knowledge graph (ClaismKG) of fact-checked claims https://data.gesis.org/claimskg/explorer/
- *Claims Statistical Observatory:* a tool allowing to extract and visualize detailed statistics of the ClaimsKG data https://data.gesis.org/claimskg/observatory/
- *Legato*: a data linking system https://github.com/DOREMUS-ANR/legato
- *CODA:* web-interface for expert validation of sameAs links https://coda.lirmm.fr/SameAsValidator
- *YAM++ online:* a tool for ontology and thesaurus matching and mapping validation http://yamplusplus.lirmm.fr
- *Keyranker*: a system for determining and ranking jointly key properties for data linking https://github.com/HFarah/KEYRANKER-2017
- *Datavore:* a system for vocabulary recommendation for data modelling http://www.lirmm.fr/benellefi/Datavore_VideoDemo
- *marc2rdf:* an RDF converter for bibliographic MARC-data https://github.com/DOREMUS-ANR/marc2rdf

Publications

Publication strategy

I will proceed to describe briefly my publication strategy over the years, which reflects the strategy of my research.

Topic-wise, I have contributed to the AI and Web-data science fields of data integration, knowledge graph construction and data modeling, knowledge extraction and information retrieval both in terms of the development of generic approaches and methodologies, as well as in terms of applications driven by diverse use-cases, such as integrating music-metadata or plant-biology scientific datasets or the analysis of societal debates on the web. I have given priority to highly ranked international conferences and journals in the web community, e.g., the Semantic Web Journal, the Journal of Web Semantics, the International Semantic Web Conference or the Extended Semantic Web Conference (ISWC and ESWC). In a few cases, I have also published in French-speaking conferences or journals, such as IC, EGC or ISI, where this has been an occasion for both me and my students to connect more closely to the French data and knowledge engineering community. I have published several survey papers in top-tier conferences or journals (e.g. Web Semantics or

ESWC) driven by the will to share gathered over the years expertise, knowledge and understanding of the challenges of a particular subfield and thus contribute to fostering new research in the community. In order to share preliminary results or ideas, I have targeted well-referenced workshops held jointly with top-tier conferences in the field (e.g. Ontology Matching OM@ISWC or PROFILES@ESWC). Just recently, a paper, which describes one of my current research projects, appeared and won an award at the Outrageous Ideas track of ISWC 2019, which provides a forum for the publication of novel visionary ideas and long-term challenges.

Content-wise, my research papers, as this is common in the field, contain a methodological contribution, which is then evaluated empirically by the help of appropriately designed experiments and publicly available datasets and benchmarks. Indeed, I have committed to making sure to the extent possible that both the datasets used for the experiments, as well as the code that implements the evaluation pipeline are made publicly available, in order to ensure reusability and reproducibility, abiding to the principles of open science.

In addition to research papers, I have also committed myself to publishing resource and demo papers. Resource papers are gaining more and more popularity recently - they give the opportunity to describe published and openly accessible datasets (in my case these are knowledge graphs in different domains, such as music or fact-checking) that can have an impact in support of reproducible research and open science in the community. Demo papers aim to describe tools and applications and are a convenient way of valorizing applied solutions, often developed by master students under my supervision, which gives them the opportunity to have a first experience in scientific publication. Indeed, most of my publications since 2013 are co-authored with my PhD students, but also with my master students, as it has been my goal to provide the possibility to master students under my supervision to have already at that stage their first experience in writing, submitting and presenting a scientific paper (these include both full research papers and demo/workshop papers). In many of the cases, the papers co-authored by my students have also been presented by them at the corresponding forums.

I have been looking to expand my scientific network both in terms of size, subject and geography. Testifying for the latter is the large number of collaborators that I have had the pleasure to publish with over the past years (over 20) from different institutes from France (LIRMM, CIRAD, INRA, Ecole Centrale Paris, CEA-Saclay, etc.), Germany (IKW Osnabrück, L3S Hannover, HHU Düsseldorf, TU Berlin), Ireland (NUI Galway) and other countries. In that, pluri- and interdisciplinarity are characteristic for my publication and research strategy, as my collaborators also include experts from renowned cultural institutions (BnF, Philharmonie de Paris, Radio France), social scientists (GESIS Cologne, GERiiCo Lille) or bioinformaticians (IRD Montpellier). Therefore, when appropriate, I have been looking to disseminate my work beyond the boundaries of the web-data community, as various publications in music-related or digital libraries venues (e.g. ISMIR or BFP (Germany)) testify.

I have 470 citations with an H-index of 12, according to Google Scholar (as of November 2019) for a total of 49 publications, listed below and summarized in Table 1. Most of my publications are also indexed by DBLP (https://dblp.uni-trier.de/pers/hd/t/Todorov:Konstantin) and Google Scholar (https://scholar.google.fr/citations?user=eWFy2AwAAAAJ&hl=fr). The co-authors order for the majority of my recent papers (since 2013) published together with my PhD or master students (a total of 26 such papers) follows the usual convention of having the student as first co-author, while I often appear as last co-author in my role of supervisor. The conferences ranks are collected from the CORE web portal (http://103.1.187.206/core) in July 2019, while the impact factors of the journals are gathered from their respective websites. I will use the following tags for the bibliographical entries, while in each category the references are grouped in blocks per year:

[BC] Book Chapter
[IJ] International Journal
[NJ] National Journal (France or Germany)
[IC] International Conference
[NC] National Conference (France)
[IW] International Workshop, Poster or Demo with peer reviews

Book chapters

[BC-3] Dietze, S., Demidova, E., Todorov, K. (2019). RDF Dataset Profiling. Encyclopedia of Big Data Technologies, Springer, 1378-1385

[BC-2] Achichi, M., Ben Ellefi, M., Bellahsene, Z., & Todorov, K. (2018). Doing Web Data: from Dataset Recommendation to Data Linking. *NoSQL Data Models: Trends and Challenges*, *1*, 57-91.

[BC-1] Todorov, K., Hudelot, C., & Geibel, P. (2012). Fuzzy and cross-lingual ontology matching mediated by background knowledge. In *Uncertainty Reasoning for the Semantic Web III - Revised Selected Papers* (pp. 142-162). Springer, Cham.

International journals

[IJ-4] Achichi, M., Bellahsene, Z., Ellefi, M. B., & Todorov, K. (2019). Linking and disambiguating entities across heterogeneous RDF graphs. *Journal of Web Semantics*, *55*, 108-121. **Impact factor 2.429**

[IJ-3] Ben Ellefi, M., Bellahsene, Z., Breslin, J. G., Demidova, E., Dietze, S., Szymański, J., & Todorov, K. (2018). RDF dataset profiling–a survey of features, methods, vocabularies and applications. *Semantic Web Journal*, 1-29. **Impact factor 3.524**

[IJ-2] Todorov, K., Hudelot, C., Popescu, A., & Geibel, P. (2014). Fuzzy ontology alignment using background knowledge. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *22*(01), 75-112. **Impact factor 1.286**

[IJ-1] Todorov, K., James, N., & Hudelot, C. (2013). Multimedia ontology matching by using visual and textual modalities. *Multimedia tools and applications*, *62*(2), 401-425. **Impact factor 2.101**

National journals (German or French)

[NJ-2] Lisena, P., Achichi, M., Choffé, P., Cecconi, C., Todorov, K., Jacquemin, B., & Troncy, R. (2018). Improving (Re-) Usability of Musical Datasets: An Overview of the DOREMUS Project. *BFP: Bibliothek Forschung und Praxis*, *42*(2)

[NJ-1] Achichi, M., Bellahsene, Z., Todorov, K. (2016). A survey on web data linking. *Ingénierie des Systèmes d'Information* 21(5-6): 11-29. **Impact factor 0.45**

International peer-reviewed conferences

[IC-21] Tchechmedjiev, A., Fafalios, P., Boland, K., Gasquet, M., Zloch, M., Zapilko, B., Dietze, S., Todorov, K. (2019). ClaimsKG: A Knowledge Graph of Fact-Checked Claims. In n *International Semantic Web Conference* (*ISWC)*. Springer. **Rank A**
[IC-20] Todorov, K. (2019). Datasets First! A Bottom-up Data Linking Paradigm. In *International Semantic Web Conference* (*ISWC) - Outrageous Ideas*. **Rank A**

[IC-19] Lisena, P., Todorov, K., Cecconi, C., Leresche, F., Canno, I., Puyrenier, F. & Troncy, R. (2018). Controlled vocabularies for music metadata. In *19th International Society for Music Information Retrieval Conference (ISMIR), Paris, France*.
[IC-18] Achichi, M., Lisena, P., Todorov, K., Troncy, R., & Delahousse, J. (2018). DOREMUS: A graph of linked musical works. In *International Semantic Web Conference* (*ISWC*) (pp. 3-19). Springer. **Rank A**

[IC-17] Farah, H., Symeonidou, D., & Todorov, K. (2017). Keyranker: Automatic rdf key ranking for data linking. In *Proceedings of the Knowledge Capture Conference (K-CAP)* (p. 7). ACM. **Rank A**
[IC-16] Tchechmedjiev, A., Mandon, T., Lafourcade, M., Laurent, A., & Todorov, K. (2017). Ontolex JeuxDeMots and Its Alignment to the Linguistic Linked Open Data Cloud. In *International Semantic Web Conference* (*ISWC*) (pp. 678-693). Springer. **Rank A**

[IC-15] Achichi, M., Ellefi, M. B., Symeonidou, D., & Todorov, K. (2016). Automatic key selection for data linking. In *EKAW* (pp. 3-18). Springer, Cham. **Rank B**
[IC-14] Tigrine, A. N., Bellahsene, Z., & Todorov, K. (2016). Selecting optimal background knowledge sources for the ontology matching task. In *EKAW* (pp. 651-665). Springer, Cham. **Rank A**
[IC-13] Ellefi, M. B., Bellahsene, Z., Dietze, S., & Todorov, K. (2016). Dataset recommendation for data linking: An intensional approach. In *Extended Semantic Web Conference* (*ESWC*) (pp. 36-51). Springer, Cham. **Rank A**
[IC-12] Ellefi, M. B., Bellahsene, Z., Dietze, S., & Todorov, K. (2016). Beyond established knowledge graphs-recommending web datasets for data linking. In *International Conference on Web Engineering (ICWE)* (pp. 262-279). Springer, Cham. **Rank B**

[IC-11] Tigrine, A. N., Bellahsene, Z., & Todorov, K. (2015). Light-weight cross-lingual ontology matching with LYAM++. In *ODBase "On the Move to Meaningful Internet Systems"* (pp. 527-544). Springer, Cham.

[IC-10] Ngo, D., Bellahsene, Z., & Todorov, K. (2013). Opening the black box of ontology matching. In *Extended Semantic Web Conference* (pp. 16-30). Springer, Berlin, Heidelberg. **Rank A**
[IC-9] Ngo, D., Bellahsene, Z., & Todorov, K. (2013). Extended tversky similarity for resolving terminological heterogeneities across ontologies. In *ODBase "On the Move to Meaningful Internet Systems"* (pp. 711-718). Springer, Berlin, Heidelberg.

[IC-8] Todorov, K., Geibel, P., & Hudelot, C. (2011). A framework for a fuzzy matching between multiple domain ontologies. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (pp. 538-547). Springer, Berlin, Heidelberg. **Rank B**

[IC-7] Todorov, K., Geibel, P., & Kuehnberger, K. U. (2010). Extensional ontology matching with variable selection for support vector machines. In *2010 International Conference on Complex, Intelligent and Software Intensive Systems* (pp. 962-967). IEEE. **Rank C**
[IC-6] James, N., Todorov, K., & Hudelot, C. (2010). Ontology matching for the semantic annotation of images. In *International Conference on Fuzzy Systems* (FUZZ-IEEE) (pp. 1-8). IEEE. **Rank A**
[IC-5] Todorov, K., Geibel, P., & Kühnberger, K. U. (2010). Mining concept similarities for heterogeneous ontologies. In *Industrial Conference on Data Mining* (pp. 86-100). Springer, Berlin, Heidelberg.
[IC-4] James, N., Todorov, K., & Hudelot, C. (2010). Combining visual and textual modalities for multimedia ontology matching. In *International Conference on Semantic and Digital Media Technologies* (SAMT) (pp. 95-110). Springer, Berlin, Heidelberg. **Rank C**

[IC-3] Todorov, K. (2009). Detecting ontology mappings via descriptive statistical methods. In *2009 Fourth International Conference on Internet and Web Applications and Services* (pp. 177-182). IEEE. **Rank C**
[IC-2] Todorov, K., & Geibel, P. (2009). Variable Selection as an Instance-Based Ontology Mapping Strategy. In *SWWS* (pp. 3-9). **Rank C**

[IC-1] Todorov, K. (2008). Combining structural and instance-based ontology similarities for mapping web directories. In *2008 Third International Conference on Internet and Web Applications and Services* (pp. 596-601). IEEE. **Rank C**

National peer-reviewed conferences (France)

[NC-4] Achichi, M., Lisena, P., Todorov, K., Troncy, R., & Delahousse, J. (2019). DOREMUS: A graph of linked musical works. In *IC: Ingénierie des Connaissances*.

[NC-3] Chentli, I., Larmande, P., & Todorov, K. (2016). Construction d'un gold standard pour les données agronomiques. In *IC: Ingénierie des Connaissances*.
[NC-2] Destandau, M., Troncy, R., Todorov, K. *et al. (*2016). Using Linked Data for Structuring and Interlinking Music Catalogs. In *Data In Libraries: The Big Picture, IFLA*.

[NC-1] Achichi, M.,Bellahsene, Z., Ienco, D., Todorov, K. (2015). Towards Linked Data Extraction from Tweets. In *EGC Extraction et Gestion de Connaissances 2015:* 383-388 **Rank C**

International peer-reviewed poster, demo and workshop papers

[IW-15] Boland, K., Fafalios, P., Tchechmedjiev, A., Todorov, K. and Dietze, S. (2019) Modeling and contextualizing claims. In Contextual Knowledge Graphs @ ISWC2019, to appear.
[IW-14] Gasquet, M., Brechtel, D., Zloch, M., Boland, K., Fafalios, P., Tchechmedjiev, A., Dietze, S., Todorov, K. (2019) Exploring Fact-checked Claims and their Descriptive Statistics. In *International Semantic Web Conference (Posters & Demos)*

[IW-13] Bellahsene, Z., Emonet, V., Ngo, D., & Todorov, K. (2017). YAM++ Online: a web platform for ontology and thesaurus matching and mapping validation. In *European Semantic Web Conference (Posters & Demos)* (pp. 137-142). Springer.
[IW-12] Lisena, P., Troncy, R., Todorov, K., & Achichi, M. (2017). Modeling the complexity of music metadata in semantic graphs for exploration and discovery. In *4th International Workshop on Digital Libraries for Musicology* (pp. 17-24). ACM.
[IW-11] Achichi, M., Todorov, K. et al. (2017). Results of the ontology alignment evaluation initiative 2017. In *OM: Ontology Matching* (pp. 61-113).
[IW-10] Achichi, M., Bellahsene, Z., & Todorov, K. (2017). Legato results for OAEI 2017. In *OM: Ontology Matching @ISWC* (pp. 146-152).

[IW-9] Achichi, M., Todorov, K. et al. (2016). Results of the ontology alignment evaluation initiative 2016. In *OM: Ontology matching* (pp. 73-129).
[IW-8] Lisena, P., Achichi, M., Fernández, E., Todorov, K., & Troncy, R. (2016). Exploring Linked Classical Music Catalogs with OVERTURE. In *International Semantic Web Conference (Posters & Demos)*.

[IW-7] Achichi, M., Bailly, R., Cecconi, C., Destandau, M., Todorov, K., & Troncy, R. (2015). Doremus: Doing reusable musical data. In *International Semantic Web Conference (Posters & Demos)*.
[IW-6] Ellefi, M. B., Bellahsene, Z., & Todorov, K. (2015). Datavore: a vocabulary recommender tool assisting Linked Data modeling. In *International Semantic Web Conference (Posters & Demos)*.
[IW-5] Vigo, M., Bellahsene, Z., Ienco, D., & Todorov, K. (2015). Twitter event detection and modeling with TEWS. In *14th International Semantic Web Conference (Posters & Demos)*.
[IW-4] Tigrine, A. N., Bellahsene, Z., & Todorov, K. (2015). Lyam++ results for OAEI 2015. In *International Semantic Web Conference (OM workshop)*.

[IW-3] Ellefi, M. B., Bellahsene, Z., Scharffe, F., & Todorov, K. (2014). Towards semantic dataset profiling. In *Extended Semantic Web Conference (PROFILES workshop).*

[IW-2] Todorov, K., Geibel, P., & Hudelot, C. (2011, October). Building A Fuzzy Knowledge Body for Integrating Domain Ontologies. In *International Semantic Web Conference (URSW worskhop)* (pp. 3-14).

[IW-1] Todorov, K., & Geibel, P. (2008). Ontology mapping via structural and instance-based similarity measures. In *International Semantic Web Conference (OM workshop)* (p. 224).

Table 1: Summary of the publications (P/D=Poster/Demo, WS=Workshop)

Year	**BC**	**IJ**	**NJ**	**IC (rank-wise)**				**NC**	**P/D**	**WS**	**Total**
				A	**B**	**C**	**NA**				
2019	1	1	1	2					1	1	**7**
2018	1	1		1			1	1			**5**
2017				2					1	3	**6**
2016			1	2	2			2	1	1	**9**
2015							1	1	3	1	**6**
2014		1								1	**2**
2013		1		1			1				**3**
2012	1										**1**
2011					1					1	**2**
2010				1		2	1				**4**
2009						2					**2**
2008						1				1	**2**
Total	**3**	**4**	**2**	**9**	**3**	**5**	**4**	**4**	**6**	**9**	**49**

Appendix B

Five Selected Publications

Opening the Black Box of Ontology Matching

DuyHoa Ngo, Zohra Bellahsene, Konstantin Todorov

Université Montpellier 2, INRIA, LIRMM
161 rue Ada, 34095, Montpellier, France
{firstname.lastname@lirmm.fr}

Abstract. Due to the high heterogeneity of ontologies, the combination of many methods is necessary in order to discover correct semantic correspondences between their elements. An ontology matching tool can be seen as a collection of several matching components, each implementing a specific method dealing with a specific heterogeneity type (terminological, structural or semantic). In addition, a mapping selection module is introduced to filter out the most likely mapping candidates. This paper proposes an empirical study of the interaction between these components working together inside an ontology matching system. By the help of datasets from the Ontology Alignment Evaluation Initiative, we have carried out several experimental studies. In the first place, we have been interested in the impact of the mapping selection module on the performance of terminological and structural matchers revealing the advantage of using global methods vs. local ones. Further, we have carried an extensive study on the flaw of the performance of a structural matcher in the presence of noisy input coming from a terminological method. Finally, we have analyzed the behavior of a structural and a semantic component with respect to inputs taken from different terminological matchers.

1 Introduction

The field of ontology matching has matured considerably as a result of more than a decade of research and practice. Many ontology matching approaches and systems have been developed dealing with the semantic heterogeneity problem by taking into account various aspects of this problem [20]. Methodologically speaking, these approaches rely on techniques from fields as diverse as machine learning, graph matching, information retrieval, relational algebra, logics, – each of these fields providing a framework to deal with a certain heterogeneity type. In this respect, a standalone ontology matching system is a successful combination of several matching components. We consider that time has come to pay attention to the way that these components connect to each other within a matching system and how this interaction impacts the overall quality of the produced alignments.

Many challenges stand in front of the ontology matching community – a full picture is given in [20]. By this study, we contribute to the solution of matcher selection and combination problems, which are fundamental for the development of a good quality matching system. A matching system can be seen as a combination of four main components: a terminological, a structure-based and a semantics-based matcher accompanied by a mapping selection module. Although they exploit different features of the entities

of an ontology to discover mappings, they are not independent. Commonly, a structure-based matcher takes as an input the mappings resulting from a terminological matcher [1, 7, 24]; a semantics-based matcher may take as an input the mappings resulting from either a terminological [5, 9], or a structure-based matcher, or a combination of the mappings resulting from both [4, 6]. Therefore, challenges and difficulties can arise not only inside of each component but also due to the interaction between them. We take into consideration several of these difficulties.

A ***mapping selection*** module is usually introduced in order to filter out the best mapping candidates, at each of the different matching levels. The interaction of this module with the matchers is, therefore, among the basic issues to be addressed.

A ***terminological matcher*** discovers mappings by comparing annotations (i.e., labels, comments) of entities. To these ends, it may use many different similarity measures. The difficulty is, on the one hand how to select the most appropriate similarity measures and, on the other hand, how to effectively combine them.

A ***structure-based matcher*** discovers mappings between entities by analyzing the similarity of the structural patterns, which these entities are part of. However, according to [3], almost all methods of this type are not stable and do not improve the matching quality when the structures of the ontologies are different. Moreover, structural matchers are error-prone, since they strongly depend on initial mappings provided by a terminological matcher and on the specific settings of the mapping selection component.

A ***semantic matcher*** is mainly used to refine candidate mappings [5, 6, 9]. It exploits the semantic constraints between entities in the ontologies in order to discover conflicts between potential mappings and remove them from the list of candidate mappings. To do that, in some tools [5, 9], the semantic module requires a confidence value for each mapping candidate. Then, it applies a global optimization method in order to find the minimal inconsistent set of mappings. Therefore, similarly to structural methods, semantic methods are error-prone because they also depend on the confidence values of the mappings obtained at previous steps.

This empirical study aims to investigate the complex interconnections between the different components in an ontology matching system. Our intention has been to make explicit the relations between these components by showing how one impacts the other and thus guide practitioners and researchers in the choice of the matchers with regard to the global quality of the matching system.

The rest of the paper is organized in the following manner. In the next section, we present a generic ontology matching system architecture together with an evaluation scenario for the ontology matching task. We continue by presenting two basic studies. With respect to different settings of the mapping selection module, we evaluate the performance of different terminological methods (Sections 3) and different structural methods (Section 4). We emphasize the superiority of global methods as compared to local ones. Further, we go into a detailed study of the interaction between terminological, structural and semantic methods. We first study the performance of different structural matchers at the presence of noisy input (Section 5) and then the behavior of structural and semantic matchers with respect to mappings produced by different terminological methods that they take as an input (Section 6). Sections 3 to 6 present one independent study each and are structured in an uniform manner: the matching methods

are presented first, followed by a presentation of the evaluation data and strategy and, finally, the results are given and supported by an in-depth discussion.

2 A Generic Framework for Ontology Matching and Evaluation

The main components of a self-dependent ontology matching system are depicted in the lower part of Fig. 1. As discussed in the introduction, the three core matching components are based on terminology, structure or semantics. The role of these matchers is to discover correct mappings or remove incorrect ones according to specific features extracted from the entities in the input ontologies. Additionally, a mapping selection module is introduced to act as a filter which selects the best candidate mappings. We define a matching strategy as the way the matcher and selection components work together in order to produce an alignment.

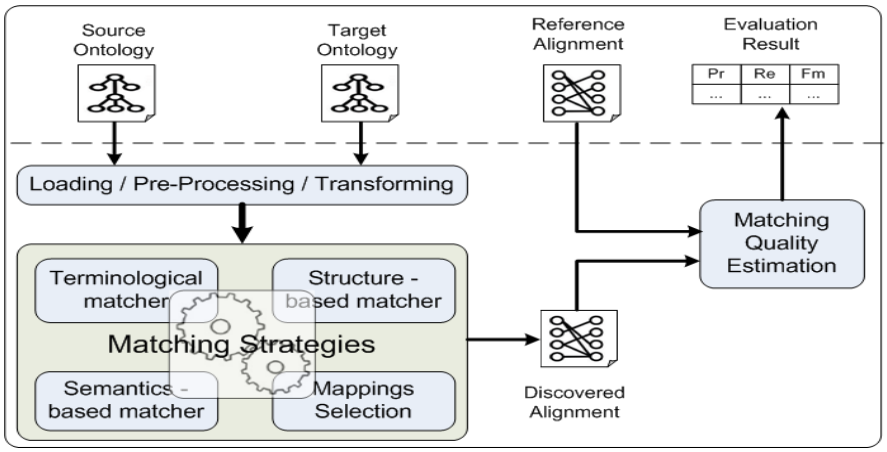


Fig. 1: Ontology Matching: System Architecture and Evaluation Scenario

To perform an evaluation of the quality of the different matching strategies, the ontology matching system requires matching scenarios as an input (upper part of Fig. 1). A matching scenario consists of a source and a target ontology and a reference alignment provided by a domain expert. Given a matching scenario, input ontologies are loaded, pre-processed and transformed into internal data structures (Loading/Pre-Processing/Transforming component). The Matching Quality Estimation module evaluates the quality of a given matching strategy by comparing discovered alignments with the reference alignment. It outputs three evaluation values corresponding to Precision (Pr), Recall (Re) and F-measure (Fm). In this study, we compute the harmonic means of precision, recall and F-measures on a set of n tests. These evaluation measures are used in the OAEI campaign. They are given in a standard manner as follows.

$$H(p) = \frac{\sum_{i=1}^{n} |C_i|}{\sum_{i=1}^{n} |A_i|}; \quad H(r) = \frac{\sum_{i=1}^{n} |C_i|}{\sum_{i=1}^{n} |R_i|}; \quad H(fm) = \frac{2 * H(p) * H(r)}{H(p) + H(r)}.$$

For the *i*th test, $|A_i|$ denotes the total number of mappings discovered by a matching system, $|C_i|$ is the number of correct mappings, and $|R_i|$ denotes the number of reference mappings provided by a domain expert. In the sequel, all results will be given by considering F-measures only.

By following this generic architecture, we have developed the YAM++ system[1]. Various matching methods inside of the three matcher components and several filtering methods used in the mapping selection module have been implemented. The system is described in detail in [17]. Because of the broad scope and diversity of the techniques employed by YAM++, as well as its excellent results in the OAEI campaigns[2], we have used this system in order to evaluate the different matching strategies based on the interaction of the matchers. More detail about the setup of the matching and filtering methods for each matching strategy will be given in each experiment in the succeeding sections.

3 Terminological Matchers and Mapping Selection

In this evaluation, we focus on terminological matchers and we study their interaction with the mapping selection module. According to a classification found in [3], the terminological matching approaches are divided into ***local*** and ***global*** methods. Local methods focus on the similarity between individual entities, whereas global methods combine local ones in order to produce an alignment, taking into account the semantic context that these entities belong to. Within this context, we are particularly interested in the comparison between local and global methods.

3.1 Methods

We have considered several state-of-the-art local methods as well as advanced global methods, some of which have been proposed specifically within the YAM++ system.

Local Terminology-Based Methods We have implemented more than fifty local methods used for terminology-based matching. We divide them into three main groups based on the algorithm for computing similarity between strings that they rely on. For more details, we refer to [18]. To economize space, the following representative methods will be used in this experiment:

Edit distance-based methods. The similarity of two strings is computed based on the number of edit operations. We have considered Levenstein and ISUB [21].

Token-based methods . These methods split strings into sets of tokens and then compare tokens by string-based methods. We have considered QGrams and TokLev (using Levestein to compare tokens).

Hybrid methods. Methods in this group split strings into sets of tokens and then compare tokens by a combination method of a string-based and linguistic-based methods. We have taken as examples HybLinISUB and HybJCLev. HybLinISUB uses a

[1] YAM++ - (not) Yet Another Matcher for Ontology Matching task.

[2] In OAEI 2012 YAM++ was first on the Conference, Multifarm, Benchmark and Bio-Medical track, and second on the Anatomy track.

combination of ISUB and Lin [11] to compare tokens; HybJCLev relies on the Levenstein and the Jiang-Corath [8] methods.

Global Terminology-Based Methods In our experiments, we have implemented the following global methods:

Weighted Average with Local Confidence (LC). Each local method is assigned a local confidence value. These values are used as weights in a weighted sum average function to compute the final similarity score between entities. More details can be find in [1].

Harmony-based Adaptive Similarity Aggregation (HADAPT). Here, each local method is assigned a weight which is computed by the harmony estimation algorithm [12]. Then, a weighted sum aggregation method is used to produce a final similarity score between entities.

Machine Learning-Based Approach (ML). This method combines all local methods and constructs a classification function on the basis of given training data. In a machine learning setting, the training dataset consists of pairs of entities (seen as training examples) for which the confidence value of their similarity is known. Based on these training examples, a classifier learns a function which is able to predict the confidence value of an unseen pair of entities. In that way, the ontology matching task is transformed into a classification task. After testing the performance of over 15 machine learning techniques, we have seen that J48 decision tree is the most appropriate one for the ontology matching task [17].

Information Retrieval-Based Approach (IR). This method judges the similarity between two entities by the amount of overlap of the information content of their labels [16]. It splits all labels of entities into tokens and calculates the information content of each token in the whole ontology. Then, IR extends Tversky's similarity measure [23] with weight of tokens to compute a similarity score between labels of entities. The method compares similarity of two labels by using not only the sequence of characters themselves, but also their information content in an ontology. We will illustrate this idea by examples in this experiment.

3.2 Matching and Evaluation Strategy

To perform this experiment, we have chosen the Conference dataset from the OAEI including 21 test cases[3]. The reason for this choice is that this dataset consists of moderate-sized real-world ontologies describing the same domain. These ontologies are highly heterogeneous since they were developed by different people, hence, the same concept is often labeled differently. We assume that high matching quality of a system on these tests guarantees similar results of the system when applied to other real matching scenarios.

The matching evaluation strategy works as follows. For each matching method (including local and global ones) in the terminological matching module, we compute a similarity score for all pairs of entities of the input ontologies. The mapping selection module then selects candidate mappings according to the filter threshold value. At

[3] http://oaei.ontologymatching.org/2012/conference/index.html

each level of this threshold, a H-mean F-measure is computed over all test cases in the dataset.

3.3 Results and Discussion

Fig. 2 shows the results of this comparison. As this can be seen, almost all local methods (except for QGrams and ISUB) improve the F-measure when the threshold value of the filter increases. When the filter threshold gets high, it seems that only highly similar or identical labels will be passed. Therefore, as Fig. 2 shows, local methods closely converge to results of the Identical method (≈ 0.55) when the filter threshold reaches 0.95 and 0.97. The same trend is observed for HADAPT and LC because they are linear combinations of local methods.

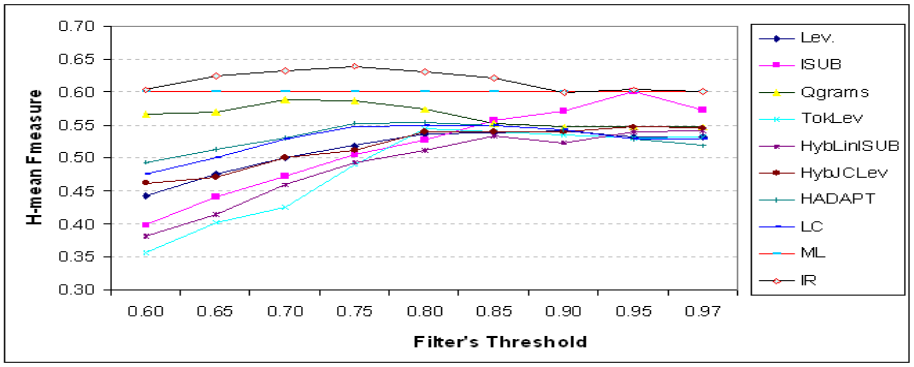


Fig. 2: Mapping Selection for the Terminological Matcher Module

The experiment shows that the two global methods, ML and IR outperform the other techniques within the terminological matcher module. Therefore, in what follows, we will discuss in more detail these two methods.

Performance of ML A machine learning method requires training data on which to learn a classification function and test data on which to apply this function. To create training data independent on the Conference dataset on which the evaluation will be performed, we have used data from the OAEI Benchmark 2009[4] and I3CON[5] datasets. We have constructed 10 different training datasets by using these two sources and we have trained the decision tree ML method on each of these 10 training sets. At each time, the learned classification algorithm has been applied to the Conference dataset, providing 10 different results. The result given in Fig. 2 is obtained by taking the average over these 10 results.

We note that the ML method does not depend on the filter threshold since no candidate mapping selection takes place. As it can be seen from the figure, ML returns a

[4] http://oaei.ontologymatching.org/2009/benchmarks
[5] http://www.atl.external.lmco.com/projects/ontology/i3con.html

better matching quality than LC, HADAPT and all local methods. For example, the ML method discovers (cmt.owl#Co-author $\equiv$ conference.owl#Contribution_co-author) in the `cmt.owl` and `conference.owl` ontologies, whereas local methods return a low similarity score between these labels (Levenstein(Co-author,Contribution_co-author) = 0.4; QGrams(Co-author,Contribution_co-author) = 0.6). This is explained by the fact that ML does not use arithmetic combination functions like LC and HADAPT, instead, it extracts the combination rules on local methods from training data. ML is able to find many patterns in the training data similar to the current example (e.g., (networkA.rdf#Office $\equiv$ networkB#OfficeSoftware), (russia1#payment $\equiv$ russia2#means_of_payment), etc.). However, the ML method strongly depends on the training data. With different training data, different machine learning models will be generated and, therefore different matching results will be produced. For instance, with some training data, ML can discover (cmt.owl#Co-author $\equiv$ conference.owl#Contribution_co-author), but not with other. Moreover, for a given training data this mapping is discovered by ML, but (cmt.owl#Document $\equiv$ conference.owl#Conference_document) is not, even though the latter seems similar to the former. To address this problem, we have designed the IR method, which is discussed in the sequel.

Performance of IR The IR method proposed in YAM++ [16] outperforms all other methods in the experiment. We analyze this fact by giving an example with two entities: `cmt.owl`#Co-author and `conference.owl`#Contribution_co-author. After splitting and normalizing the labels, we have the following two sets of tokens: {`coauthor`} and {`coauthor`, `contribution`}. Token `coauthor` appears in each input ontology only once, whereas token `contribution` appears 10 times among 60 concepts in the `conference.owl` ontology. Therefore, the information content of the token `contribution` is lower than that of token `coauthor`. In particular, the normalized *tf-idf* weights of each token inside the input ontologies are equal: $\{w_{coauthor} = 1.0\}$, $\{w_{coauthor} = 1.0, w_{contribution} = 0.34\}$. The two sets of tokens share only the token `coauthor`, hence the similarity computed by Tversky's method is $\frac{1.0+1.0}{1.0+1.0+0.34} = 0.855$. Similarly, we have the similarity between (`Document`,`Conference_document`) equaling `0.91`. In this pair, the token `conference` appears 15 times in the `conference.owl` ontology. Therefore, this token brings little information for this ontology and, consequently, this pair of entities represents a likely match.

It is difficult to give a clear indication which of these two best performing methods to use – ML or IR. Clearly, in the absence of training data, the choice will go for IR. Even in the presence of training data, the IR method appears to be more suitable because it reaches an overall higher performance than ML for low values of the mapping selection threshold. However, an advantage of the ML method is that it does not depend on the setting of a filter threshold. Both methods can be combined within the architecture of an ontology matching tool.

4 Structural Matchers and Mapping Selection

In this evaluation, we are interested in the behavior of structural similarity methods with respect to the mapping selection module.

4.1 Methods

The following standard structural matching methods have been considered within this study: ***ANCESTORS*** (two entities are similar if all or most of their ancestor entities are already similar), ***DESCENDANTS*** (two entities are similar if all or most of their descendant entities are already similar), ***LEAVES*** (two entities are similar if all or most of their leaf entities are already similar [2]), ***ADJACENTS*** (two entities are similar if all or most of their adjacent entities (parents, children, siblings, domains, ranges) are already similar), ***ASCOPATH*** (two entities are similar if all or most of entities in the paths from the root to the entities in question are already similar [10]). ***DSIPATH (Descendant's Similarity Inheritance)*** (two entities are similar if the total contribution of entities in the paths from the root to them is higher than a specific threshold [22]), and ***SSC (Sibling's Similarity Contribution)*** (two entities are similar if the total contribution of their sibling entities is higher than a specific threshold [22]).

Additionally, we have considered the ***SP (Similarity Propagation)*** method. This method is proposed in the system YAM++ as an extension of the well-known similarity flooding algorithm [15]. The basic idea of the method is as follows. Assume that the entities A_1 and A_2 in one ontology are related by a directed relation P and the entities B_1 and B_2 in another ontology are related by the same directed relation. Then, if we discover that (A_1, B_1) is a match, the SP method would imply that (A_2, B_2) is a match, too. The similarity values between the two pairs are propagated to each other at each iteration of algorithm. The approach is described in detail in [19].

4.2 Matching and Evaluation Strategy

To perform this experiment, we have used the Benchmark 2011 dataset from the OAEI campaign including 103 test cases. These datasets are acquired by taking an original ontology and altering the names of some of its entities by using random strings (no variation by naming convention or synonym words). The entities whose labels have not been altered are kept as in the original ontology. Therefore, a matching scenario which takes as an input the original ontology and the altered one is appropriate for evaluating the performance of structural methods. As an input to these structural methods, we use the alignment produced by an identical metric (discovering only identical strings as correct mappings), since the non-altered string names are identical in both ontologies. An additional characteristics of this dataset is that in some tests, not only the names of entities are altered but also the ontology structure by flattening, extension and other structure modifying operations.

The matching and evaluation strategy used in the experiment is given as follows. Only three modules will be used: a terminological matcher, a structure-based matcher and a mapping selection module.

The ***terminology-based*** matcher provides input mappings to the structural matcher. It uses the *identical metric* to compute a similarity score between two entities, which is equal to 1 if the entity names are the same, 0 otherwise.

Each ***structure-based*** matcher corresponding to each of the selected structural measures above produces a similarity matrix for all pairs of entities from the two input ontologies.

In the ***mapping selection*** module, we vary the threshold $(0.01 - 0.9)$ to filter out the mappings discovered by this matcher. The mappings obtained by the structural matcher are combined with mappings obtained by the terminological matcher to produce the set of candidate mappings. Then, a greedy selection method [14] is used to extract the final alignment.

4.3 Results and Discussion

As this can be seen from Fig. 3, when the threshold varies from 0.6 to 0.9, the structural methods converge to the INIT-MAPPINGS line where H-mean Fmeasure = 0.68 (4463 correct mappings, 27 incorrect mappings, 4342 unfound). This means that the structural methods did not discover additional correct mappings or they discovered correct mappings, which already exist in the init mappings. This is natural, because most of the structural methods compute similarity between two entities based on the overlap of their structural patterns (i.e., adjacent, ancestor, etc.) by using, for instance, the Jaccard measure. Therefore, the higher the filter threshold, the lower the possibility of discovering new mappings.

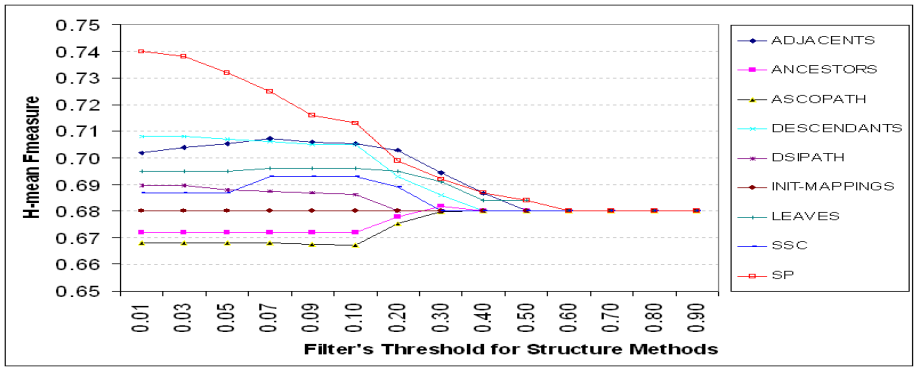


Fig. 3: Mapping Selection for Structural Methods

We note that the corresponding matching qualities of the structural methods differ significantly when the filter threshold is set to small values. We will consider methods that perform poorly and such that perform well.

For threshold values between 0.01 and 0.09, ASCOPATH and ANCESTORS discover many incorrect mappings. For example, when the threshold is equal to 0.01, ACSOPATH discovers 90 $(= 4733 - 4643)$ additional correct mappings but 453 $(= 480 - 27)$ incorrect mappings in comparison to the init mappings. This can be explained as follows. After observing the ontologies in the Benchmark 2011 dataset, we see that the maximum depth and also maximum number of ancestors of an entity in the ontology hierarchy is 5. Assume that two entities have only one common entity among their ancestors. Then their similarity score is equal to $1/10 = 0.1$ at least. If two entities do not have any common entities, then their similarity is equal to 0. Therefore, with a threshold in the range of 0.01 to 0.09, any pair of entities having at least one common

ancestor will be considered as a match. Since sibling entities have the same ancestors and paths to these ancestors, they will have the same structural patterns. Therefore, many pairs of entities will have the same similarity scores. Moreover, one entity may have many descendant entities so many pairs of entities can be coupled, consequently, many incorrect mappings will be produced.

In contrast, other methods such as DESCENDANTS, LEAVES, DSIPATH and SSC provide better results with small thresholds than the methods discussed above. They discover more additional correct mappings and, consequently, improve the overall quality of the matching. For example, with a threshold equal to 0.01, DESCENDANTS discovers 494 $(= 5137 - 4643)$ additional correct mappings and 175 $(= 202 - 27)$ incorrect mappings in comparison to the init mappings. Similarly to the ASCOPATH and ANCESTORS methods, with low threshold filter, many pairs of entities are passed. However, these methods clearly distinguish the structural patterns of entities. For instance, in DESCENDANTS and LEAVES, different entities have different sets of leaves / descendants; in DSIPATH and SSC, they use different contribution percentage of entities according to how much one entity is important to another [22]. Therefore, by running greedy selection, which always selects the pair of entities having high similarity score with 1:1 cardinality, most of the selected mappings are correct.

Performance of SP The similarity propagation (SP) method that we propose differs from the other structural methods discussed above in several aspects. Note that the similarity scores produced by SP are not absolute but relative values due to the normalization process at the end of each running iteration. SP propagates similarity values from one pair of entities to another, hence, if two entities have a similarity score higher than 0, then they are considered as similar to a certain degree. Thus, with a low threshold filter, SP discovers more correct mappings than with a high threshold value. Moreover, the similarity score of a pair of entities depends not only on their current status but also on the status of other related (neighboring) pairs. The more neighbors with high similarity a pair of entities has, the likelier that they are matched. For example, when the threshold is set to 0.01, SP discovers 1298 $(= 5941 - 4643)$ additional correct mappings and 247 $(= 274 - 27)$ incorrect ones in comparison with the init mappings. Therefore, SP distinguishes well correct and incorrect mappings by ranking the similarity scores which is the main reason why this method outperforms the other local structural methods discussed above when the filter threshold is low.

5 Impact of Noisy Input on Structural Matchers

In this experiment, we evaluate the behavior of different structural matchers when we add noise into the mappings that these methods take as an input from a terminological matcher. Here, we call "noise" a pair of dissimilar entities labeled as similar by the terminological matcher. Indeed, in real matching scenarios, a terminological method rarely produces 100% precision, consequently, it rarely provides input mappings without noise to the structural methods. We will study the impact of this noise on the mappings discovered by several structural methods with the aim to outline the most stable among these methods with respect to the presence of noise.

5.1 Methods and Evaluation Strategy

For these experiments, we have used the Benchmark 2011 dataset from the OAEI campaign. The justification of this choice given in Section 4 holds here, as well, since we are dealing with structural methods again.

At ***terminological level***, we use the identical metric. To produce noise, we add a number of random incorrect mappings, which correspond to $N\%$ of the size of the original init mappings, with $N \in \{0, 10, .., 100\}$.

At ***structural level***, the matcher takes input mappings from the terminological matcher. According to the experiments in Section 4, we select the best threshold filter for each structural method. For example, $\theta_{SP} = 0.01, \theta_{DESCENDANTS} = 0.01, \theta_{ADJACENTS} = 0.07$, etc.

At each iteration, we count the total number of correct mappings and the total number of incorrect mappings that a structure method produces over all 103 test cases contained in the Benchmark 2011 dataset.

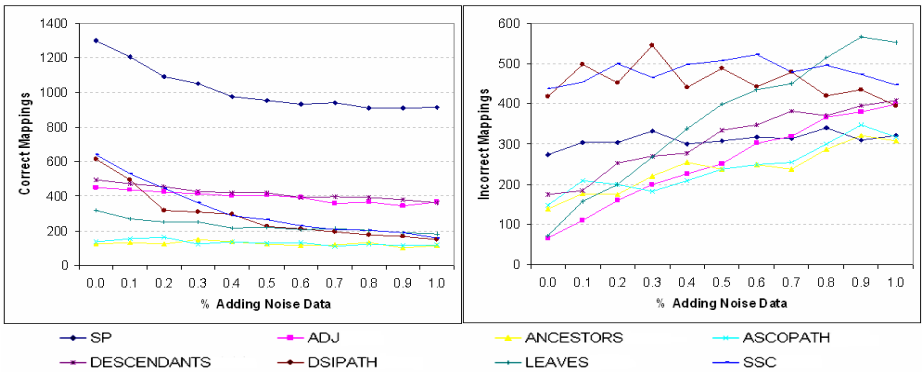


Fig. 4: Impact of input noise on structural matchers.

5.2 Results and Discussion

Fig. 4 shows the total number of correct and incorrect mappings produced by the structural methods at each time when noisy data are added to the input. When noisy data are added, the number of correct mappings discovered by all the methods decreases. Regarding the number of incorrect mappings, it increases for all methods, except for DSIPATH and SSC. Note that DSIPATH and SSC differ from the other local structural methods in terms of the interaction between the entities in an ontology. For example, the similarity of two entities computed by DSIPATH strongly depends on the similarity provided by input mappings and decreasingly depends on the similarity of their parents, grandparents, etc. Consider two entities of two input ontologies. If noise appears at the same level in their paths to the root, their similarity will be impacted by this noise, otherwise, it will not. Therefore, the impact of noise in discovering further mappings depends on the position of the entities in the hierarchies of the input ontologies. Because noise is generated randomly, its impact is hard to predict for these methods. Other structural

methods use set operations (i.e., intersection, union) with no hierarchical consideration for the elements. When noise appears in the set of ancestors or descendants of two entities, the noise will directly propagate errors to them. Therefore, as seen in Fig. 4, the number of incorrect mappings increases in almost all structural methods of this type.

This experiment shows the dominance of similarity propagation (SP) over other structural methods in terms of stability. When noisy data reaches 100%, SP still discovers 913 additional correct mappings in comparison to the init mappings. Note that the maximum number of correct mappings discovered by the other methods is only 612 mappings with no noise added. Moreover, from 0% to 100% of the noisy data, SP produces only 57 $(321 - 274)$ additional incorrect mappings. In contrast, for example, the LEAVES method produces 481 $(553 - 72)$. This is explained by the fact that SP takes into account all kinds of semantic relations of entities such as concept-concept, concept-property and property-property, which reduces the impact of noise.

6 Interaction of Terminological Matchers with Structural and Semantic Matchers

In this evaluation, we are going to study the impact of the quality of the input mappings provided by several terminological methods on the matching quality of structural and semantic matchers. More precisely, we are interested in discovering which are the terminological methods that provide best performance of the structural and semantic matchers for a given mapping selection threshold.

6.1 Methods and Evaluation Strategy

To carry out this experiment, we have used the Conference dataset from the OAEI campaign, which is a real world dataset from the domain of scientific publishing. Our evaluation strategy is described as follows.

At ***terminological level***, we have used three different methods to produce initial mappings. The choice of these matchers has been motivated by the study described in Section 3 and the results shown in Fig. 2. We have chosen QGrams representing token-based methods and ISUB representing edit-based methods because they show different behaviors when the terminology-based filter threshold changes as compared to the other methods. In addition, we have included IR, representing global methods, which is the best performing among these methods.

At ***structural level***, we have considered SP which takes input from the terminological matchers and performs similarity propagation. This choice is justified by the fact that this method has shown to perform best in the experiments in Section 4.

At ***semantic level***, we use the global diagnosis optimization method proposed in [13] which refines input terminological mappings in order to remove inconsistent ones.

We have studied the performance of each of the terminological methods when used alone and when used as an input for the structural and the semantic methods, respectively. At each iteration, the matching quality is evaluated by comparing the discovered alignment to a reference alignment.

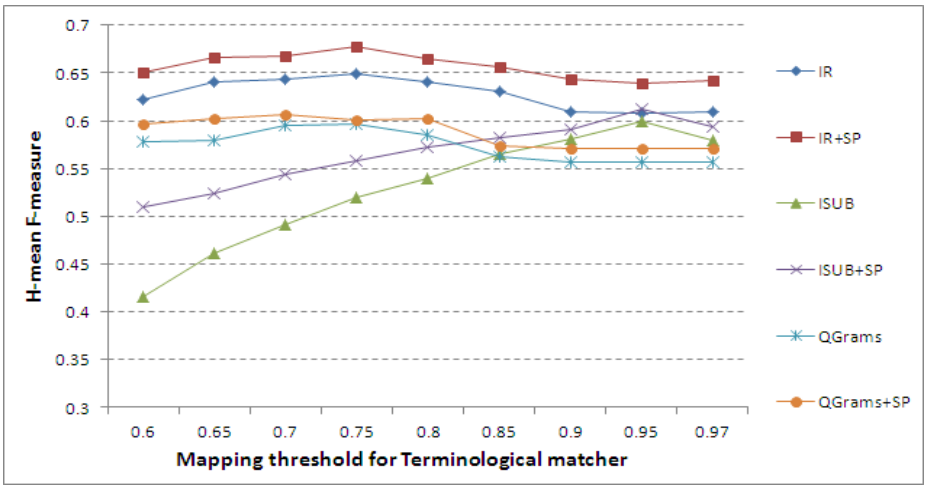


Fig. 5: Interaction of terminological methods with a structural matcher (SP) w.r.t. different values of the mapping selection filters.

6.2 Results and Discussion

Fig. 5 shows the performance of the terminological methods used alone and in combination with a structural matcher (SP) again as a function of the mapping selection threshold. Fig. 6 shows the behavior of the same terminological methods, this time taken as an input by a semantic matcher. The first observation is that the structural and the semantic methods combined with terminological matchers have similar behaviors, therefore the following analysis will encompass both.

Globally, the combined methods outperform the single terminological methods. Similarly to the previous experiments, the overall performance increases by increasing the threshold value. Quite straightforwardly, the quality of the combined methods increases simultaneously with the quality of the single terminological methods.

Further, we notice that the methods based on QGrams tend to be more stable over the variations of the filter threshold and provide high quality results already at low filter values. This is explained by the fact that the QGrams measure is based on Jaccard similarity computation and as soon as the threshold value reaches 0.6, the matcher already accounts for two third of the overlapping tokens. The methods based on ISUB have a different behavior – they have an almost linear growth of the performance as a function of the filter threshold, reaching higher values of the F-measure than the ones of the QGrams methods for thresholds above 0.9, both for structural and semantic approaches.

We explain that by the fact that at a certain level of the threshold value, the number of incorrect mappings is always higher than the number of correct mappings, especially due to the 1:1 cardinality. Therefore, when the threshold value increases, the number of removed incorrect mappings will get higher than the number of removed correct mappings. Thus, the overall quality increases. However, after surpassing the threshold of 0.95 the quality decreases again. This is due to the fact that when the threshold is

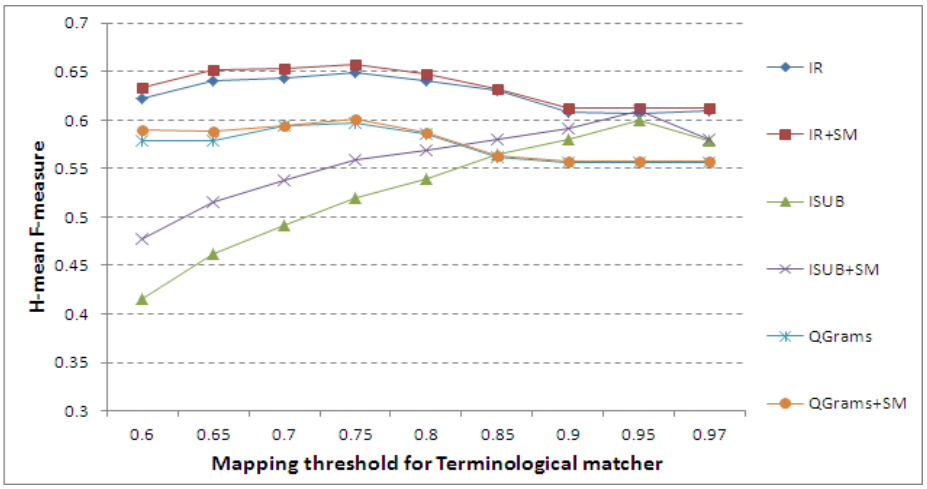


Fig. 6: Interaction of terminological methods with a semantic matcher (SM) w.r.t. different values of the mapping selection filters.

that high, only identical or nearly identical strings are passed (i.e. the overall number of passed entities decreases).

Finally, we note that the mapping selection component is a very important intermediate level between the terminology matchers and the structural or semantic ones in order to select output of each matching component. Indeed, the quality of the produced alignments is much worse if no mapping selection is performed. In the experiments, the role of mapping selection is shown by varying the value of the filter threshold.

As a general conclusion, we outline the fact that both the structural and the semantic matchers boost the performance of both local and global terminological methods, but perform best by taking input from the global IR method.

7 Conclusion

In this empirical study, we have presented experiments and evaluations analyzing the interaction between the components of an ontology matching system, seen as a chain in which the resulting mapping of a given module is the input to another. To these ends, we have used evaluation data from the OAEI campaign.

In the first place, we have been interested in the impact of the mapping selection module on the performance of terminological and structural methods revealing the advantage of using global methods vs. local ones. Further, we have carried an extensive study on the flaw of the performance of a structural method in the presence of noisy input coming from a terminological method. Finally, we have analyzed the behavior of a structural and a semantic matcher with respect to different inputs taken from different terminological methods at different values of the mapping selection filter.

The results of this study confirm explicitly the straightforward hypothesis that the quality of the overall system depends on the quality of its components. More impor-

tantly, this paper presents an in-depth analysis of the reasons of the observed behaviors, opening as much as possible the black box inside which the mechanism behind most of the ontology matching tools usually remains.

References

[1] I. F. Cruz, C. Stroe, M. Caci, F. Caimi, M. Palmonari, F. P. Antonelli, and U. C. Keles. Using agreementmaker to align ontologies for oaei 2010. In *OM*, 2010.
[2] R. Dieng and S. Hug. Comparison of personal ontologies represented through conceptual graphs. In *ECAI*, pages 341–345, 1998.
[3] J. Euzenat and P. Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg, 2007.
[4] F. Giunchiglia and P. Shvaiko et al. S-match: an algorithm and an implementation of semantic matching. In *ESWS*, pages 61–75, 2004.
[5] J. Huber, T. Sztyler, J. Nößner, and C. Meilicke. Codi: Combinatorial optimization for data integration: results for oaei 2011. In *OM*, 2011.
[6] Y. R. Jean-Mary and M. R. Kabuka. Asmov: Results for oaei 2008. In *OM*, 2008.
[7] N. Jian, W. Hu, G. Cheng, and Y. Qu. Falconao: Aligning ontologies with falcon. In *Integrating Ontologies*, 2005.
[8] J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *CoRR*, 1997.
[9] E. Jiménez-Ruiz, A. Morant, and B. Cuenca Grau. LogMap results for OAEI 2011. In *OM*, 2011.
[10] B. Thanh Le, R. Dieng-Kuntz, and F. Gandon. On ontology matching problems. In *ICEIS (4)*, pages 236–243, 2004.
[11] D. Lin. An information-theoretic definition of similarity. In *ICML*, pages 296–304, 1998.
[12] M. Mao, Y. Peng, and M. Spring. A harmony based adaptive ontology mapping approach. In *SWWS*, pages 336–342, 2008.
[13] C. Meilicke. Alignment incoherence in ontology matching. In *Thesis*, 2011.
[14] C. Meilicke and H. Stuckenschmidt. Analyzing mapping extraction approaches. In *OM*, 2007.
[15] S. Melnik, H. Garcia-Molina, and E. Rahm. Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In *ICDE*, pages 117–128, 2002.
[16] D.H. Ngo. *Enhancing Ontology Matching by Using Machine Learning, Graph Matching and Information Retrieval Techniques*. PhD thesis, Univrsity of Montpellier 2, 2012 (In print).
[17] D.H. Ngo, Z. Bellahsene, and R. Coletta. A flexible system for ontology matching. In *Caise 2011 LBIP*, pages 79–94, 2011.
[18] D.H. Ngo, Z. Bellasene, and R. Coletta. A generic approach for combining linguistic and context profile metrics in ontology matching. In *ODBASE*, 2011.
[19] D.H. Ngo, Z. Bellasene, and R. Coletta. Yam++ results for oaei 2011. In *OM*, 2011.
[20] P. Shvaiko and J. Euzenat. Ontology matching: State of the art and future challenges. *IEEE TKDE*, 99, 2011.
[21] Giorgos Stoilos, Giorgos B. Stamou, and Stefanos D. Kollias. A string metric for ontology alignment. In *ISWC Conference*, pages 624–637, 2005.
[22] W. Sunna and I. F. Cruz. Structure-based methods to enhance geospatial ontology alignment. In *GeoS*, pages 82–97. Springer, 2007.
[23] A. Tversky. Features of similarity. *Psychological Review*, 84:327–352, 1977.
[24] P. Wang and B. Xu. Lily: Ontology alignment results for oaei 2009. In *OM*, 2009.

Linking and Disambiguating Entities Across Heterogeneous RDF Graphs

Manel Achichi[a], Zohra Bellahsene[a], Mohamed Ben Ellefi[b], Konstantin Todorov[a,*]

[a]*LIRMM, University of Montpellier, CNRS*
161 rue Ada, 34000 Montpellier, France
[b]*LIS, Aix-Marseille University,*
163 Avenue de Luminy, 13288 Marseille, France

Abstract

Establishing identity links across RDF datasets is a central and challenging task on the way to realising the Data Web project. It is well-known that data supplied by different sources can be highly heterogeneous—two entities referring to the same real world object are often described, structured and valued differently, or in a complementary fashion. In this paper, we explore the origins and the multiplicity of *data heterogeneity* problems, proposing a novel classification that allows to isolate challenges and to position our and future work. Many state-of-the-art data linking approaches rely on sets of discriminative properties, provided by the user or by specialised tools, which, in the lack of knowledge of the nature of the data, do not allow to account automatically for a large number of structural heterogeneities. In addition, similarity measures and thresholds need to be selected and tuned manually or learned by specialized algorithms. We propose a solution covering an important number of heterogeneities, attempting to reduce the user configuration effort, based on: (i) *Property filtering*, or automatic data cleaning of "problematic" attributes; (ii) *Instance profiling* allowing to represent each resource by a sub-graph considered relevant for the comparison task; and (iii) *Instance vector representation* allowing to compare resources. To reduce the false positives rate, we apply a (iv) *Post-processing* step based on hierarchical clustering and key ranking techniques aiming to disambiguate highly similar, though not identical instances. This pipeline is implemented in *Legato*—a data linking tool, showing to outperform or to perform as well as state-of-the-art tools on highly heterogeneous and diverse benchmark datasets, yet keeping the user configuration effort low.

Keywords: RDF Data Linking, Knowledge Graphs, Linked Open Data, Data Heterogeneities

1. Introduction

Linked data and its underlying technologies have been gaining popularity over the past years, due to the means they offer for data reuse and federation, increased visibility and data sharing on the web and facilitated exchange of metadata. The web of data, and particularly the Linked Open Data (LOD) project,[1] has been growing in size over the past years, with hundreds of datasets published following the semantic web principles. To fully unlock the potential of the open data project, related resources across datasets need to be linked together—a process that cannot be handled manually at the web scale. Data linking is the semantic web research field that has taken the challenge of proposing methods and providing tools for the automatic detection of relations between cross-dataset resources. A plethora of data linking approaches and systems has been proposed in the past years, surveyed in [1, 2, 3]. The majority of these approaches attempt to solve the problem of discovering identity relations (often of arity 1:1), declared as `owl:sameAs` statements, between similarly-typed resources of two RDF datasets.

Preparing the datasets prior to linking and configuring the linking tools are challenging problems that often require in-depth knowledge of the data. Several state-of-the-art tools [4, 5]

*Correspodning author
Email address: `todorov@lirmm.fr` (Konstantin Todorov)

[1] `http://linkeddata.org`

require link specification files where one has to indicate property names, select and tune similarity measures. This process is handled either manually, or by specialised tools [6, 7]. Making a linking tool self-dependent in that respect is among the challenging issues that the community faces.

On the instance matching level, a data linking system has to be able to deal with a large variety of data heterogeneities, taking into account mismatches in the descriptions on value, ontological and logical level, as well as differences in the quality of the input datasets. While heterogeneities on literals are rather well-handled by similarity measures and data unification techniques, ontological discrepancies (regarding structure and properties) appear to be more challenging.

Finally, as we show in our experiments, most current approaches fail to handle correctly datasets containing blocks of highly similar in their descriptions, but yet distinct resources. Datasets with such characteristics are prone to the generation of false positives in the linking process (for example, two datasets containing all piano sonatas of Beethoven, where two works differ very little in their descriptions).

In this paper, we attempt to address the challenges given above. We aim at reducing the difficulty of manual configuration when it comes to data-related parameters, such as properties to compare. With respect to similarity measures or threshold settings, we rely on an empirical approach which shows satisfactory results on our tests, although we make no assumptions on its generalisation properties. We propose a system, called *Legato*, which, contrarily to property-based instance matchers, applies indexing techniques that allow to project each instance in a vector space defined by an appropriately chosen set of literals that describe that instance, derived from the Concise Bounded Descriptions (CBD) of the resources.[2] In addition to avoiding the selection of properties, this representation addresses in its mechanism a number of data heterogeneities without requiring user input. An automatic property filtering module allows to decrease noise prior to instance matching. We pay particular attention to discriminating highly similar, yet distinct resources, implementing an unsupervised learning post-processing strategy combined with a key selection and ranking algorithm [8] that reduces the number of false positives and increases precision.

Legato has been conceived in the framework of the DOREMUS project,[3] which develops methods to describe, publish, connect and contextualize music catalogs from major cultural institutions[4] on the web of data [9]. The data collected and handled in this project has served as a main motivation for the development of this system, which aimed to respond to the difficulties of linking these highly heterogeneous datasets, while remaining as generic as possible. We evaluate *Legato* on benchmarks from the Ontology Alignment Evaluation Initiative (OAEI)[5] instance matching tracks form 2015, 2016 and 2017. Note that two of the benchmarks of these campaigns released in 2016 and 2017 are real-world music-metadata datasets issued from the DOREMUS project. *Legato* has participated to the 2017 edition of OAEI. The experimental results show that our system performs as good as the state-of-the-art link discovery tools and it outperforms them particularly on heterogeneous real-world data and in the presence of difficult to disambiguate cross-graph instances. In addition, we show that *Legato* achieves better results in terms of F-measure than established matching tools that implement an automatic link specification learning strategy. Among the drawbacks of the system, we underline certain scaling issues encountered on several datasets. *Legato* is an open source, freely available system.[6] For the sake of reproducibility of our experiments, all datasets and configuration files (where applicable) used in the evaluation are made available (links and references are provided in the evaluation section).

To wrap up, the contributions of this paper are as follows:

- A classification of the different *data heterogeneity* types based on a large number of examples from real-world and synthetic datasets.

[2] https://www.w3.org/Submission/CBD/

[3] http://www.doremus.org/

[4] The French National Library, Philharmonie de Paris and Radio France

[5] http://oaei.ontologymatching.org

[6] https://github.com/DOREMUS-ANR/legato

- A new CBD-based instance profiling framework allowing to represent and compare resources at the matching phase.
- A novel preprocessing strategy aiming at the automatic identification and removal of "problematic" properties across two datasets.
- A new post-processing mechanism to select and repair erroneous links generated in the matching step.
- A multi-facet reproducible empirical evaluation on a large variety of openly available benchmarks.
- An open source implementation of our system with a simple user interface.

The rest of the paper is structured as follows. In Section 2, we present our account on RDF web data heterogeneity types, while in Section 3, we focus on challenges related to the problem of reducing the user effort in the data linking tool configuration process. These two sections are intended to be read as an account of different challenging issues that allow to structure the related work proposed to deal with these challenges and to identify open issues and problems. On these bases, Section 4 introduces the approach and workflow of *Legato*, which is discussed and positioned with respect to the related work in Section 5. We report on our experiments in Section 6 before we conclude and draw further directions of work and discuss lessons learned in Section 7.

2. RDF Datasets Heterogeneity Types

Understanding data heterogeneity in its multiple forms allows to identify and analyse the origins of the data linking problem and hence propose better solutions. In the context of web data linking, we will refer to data heterogeneity as any difference in the expression of a given piece of information across two graphs, observed in terms of schema (classes, properties), values, or general data structure. Ferrara et al. [10] consider three major levels of data heterogeneity (or requirements, as given in the paper): *value*, *structural* and *logical* levels. We base ourselves on this classification and extend it in an attempt to provide a comprehensive inventory of data heterogeneity types. Our classification emerges as a result of observations and tests on two types of data: (1) highly heterogeneous real-world data about classical music,[7] (2) a number of synthetic benchmark datasets released between 2015 and 2017 by the Instance Matching track of OAEI (IM@OAEI). Two different methodologies have been followed in the two cases. In case (1), we have worked tightly with librarian experts and archivists from the BnF, Radio France and the Philharmonie de Paris (partners on the DOREMUS project). They have identified collaboratively and listed a set of possible heterogeneities, given their expert knowledge of the data.[8] (As a matter of fact, this work is the basis of the creation of the DOREMUS benchmarks at OAEI 2016 and 2017, as discussed below.) Regarding (2), we have considered the respective benchmark generation strategies (e.g., altering string values or value types) as a basis and completed with these the list of heterogeneities identified in (1). Finally, the resulting list has been expanded by additional cases that we have observed through our work with the data. In order to form the taxonomical skeleton of our classification, we have used and extended the three axes identified by [10].

The resulting classification reflects the authors consensus and does not claim universality. For illustration purposes, we will use a fictional example, given in Figure 1, showing the descriptions of a real-world entity (the composer Ludwig van Beethoven) in two different graphs. For readability reasons, the example outlines only several of the heterogeneity types given below.

2.1. Value Dimension

Datatype properties are an ample source of heterogeneities, both when it comes to string or numerical attributes.

Terminological heterogeneity. We refer to differences between the lexical labels used to denote the same information across graphs. This comprises well-known issues related to

[7] https://github.com/DOREMUS-ANR/knowledge-base/tree/master/data

[8] Throughout the working process, these heterogeneities have been outlined together with concrete examples in a table that can be found here (in French): https://docs.google.com/spreadsheets/d/19dLjabt_fggTVNuM7XW9CUZkuB1JgoH9xKWQgLZv_Y4/edit#gid=1271677916

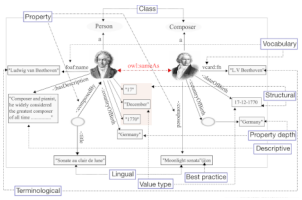


Figure 1: Open web data heterogeneity types: several examples (in blue-lined boxes).

synonymy, polysemy or variations in spelling (typos, acronyms, abbreviations, etc.). The problems of synonymy and polysemy have been largely addressed in the literature by term disambiguation techniques assisted by lexico-semantic resources [11]. A long tradition of research in the field of string similarity measures, at the core of the instance comparison modules of state-of-the-art systems [4, 12, 13], has allowed to handle the case of orthographical variations among labels [14, 15]. Several works propose solutions allowing to find the full form of an acronym or an abbreviation [16, 17].

Lingual heterogeneity. Multilingualism has been outlined as a major challenge to the open data community, already several years ago [18]. Looking at the datasets about music openly published as RDF graphs by the BNF (French National Library) [19] and Freebase [20] in French and English, respectively, one notices that very few string literals are directly comparable across these graphs. Several studies propose solutions based on machine translation [21, 22] or, alternatively, relying on a lexico-semantic resource such as BabelNet [11] as a mediator to bridge the language gap, as proposed in [23] (in a similar spirit as an earlier ontology matching method [24]). The latter approach anchors the resources as vectors of BabelNet identifiers where each of them represents a sense of a term allowing to compute vector distances as a proxy for instance similarity. Combining machine translation with concept embeddings, [25] translates each resource description to English and then a Wikipedia-based representation (a set of concepts) is generated for the resources in order to compare them.

Datatype properties vs. object properties. A piece of information (e.g., the genre of a music work) can be given by a string literal via a datatype property, or by a URI, that is used to identify the same element in a controlled vocabulary (e.g., a SKOS vocabulary of music genres[9]). Since the two objects are not directly comparable, a linking tool needs to access the string label associated to the URI in the respective vocabulary before proceeding to the comparison. In the same line of thought, comparing object property values (different URI's identifying the same object) can potentially lead to a similar type of difficulty.

2.2. Ontological Dimension

We discuss schema-related differences across RDF graphs.

Vocabulary heterogeneity. A recurrent discrepancy across data providers is the use of different ontologies. This is a challenging problem in the context of the open web of data, where we have an abundance of models and vocabularies/ontologies with different degrees of explicit rigour of their semantics, leading to different interpretations and usages.[10] Ontology matching techniques have been adopted by certain data linking systems [26], but can also be applied independently.

Structural heterogeneity. The description of an entity can be done at different levels of granularity, as is the case of the birth date of Beethoven in Fig. 1, given in a single information field or distributed over multiple properties—a challenge for property-based instance matchers. To the best of our knowledge, this heterogeneity type is only partially resolved by using inverted indexes and Natural Language Processing (NLP) techniques by several data linking approaches [12, 22, 23, 27, 28]

[9] Examples of such vocabularies: `https://github.com/DOREMUS-ANR/knowledge-base/blob/master/vocabularies/`

[10] The Linked Open Vocabularies catalog `http://lov.okfn.org/dataset/lov/` (LOV) allows to browse over 600 vocabularies from the linked open data cloud, but, although very useful, it is to date not exhaustive (certain established graphs such as Yago, Freebase or MusicBrainz are not indexed).

that propose to index each resource by the literals collected at a given distance $n \geqslant 1$ in the RDF graph.[11] Then, a vector space model is used to represent each resource description and select linking candidates on the basis of vectors proximity. By doing so, the resources are compared with respect to their literals without taking into account the properties describing them. The quality of the alignment depends strongly on the distance parameter.

Property depth heterogeneity. The same piece of information can be found at different distances to the resource in two different graphs. In our example, we can observe this problem with regard to the name of the country of birth of Beethoven. This problem can also be solved by indexing the scope of literals describing each resource. For each entity, the distance at which the literals are collected can be fixed (e.g., [29] choose a distance of 1). Again, the trade-off while setting this parameter is between too small a distance not allowing for a complete description and a too large distance, increasing the likelihood to collect irrelevant information.

Descriptive heterogeneity. A resource can be described with more information (a larger set of properties and types) in one dataset than in another, as we can see in our running example. The lack of information narrows down the intersection of the resources respective descriptions and hence the common ground where to look for commonalities or differences.

Key heterogeneity. Key identification algorithms [6, 30, 31] aim to discover discriminative properties on two datasets independently and thus identify potential candidates for link specifications of property-based state-of-the-art tools [4, 12]. However, in certain cases, comparing the values of such properties will lead to deciding negatively on the equality of two identical instances and to the generation of false negatives. We take two examples: (1) a property that is valued by unstructured textual information (e.g., a free-text description as the one given by the *-:hasDescription* property in our example) and (2) a property used to provide dataset-specific individual identifiers, e.g., the ID's of bibliographical entries of two libraries. In both cases the values of these key properties are not comparable across datasets.

[11] A distance in an RDF graph is defined as the minimal number of edges (properties) connecting two resources or a resource and a literal.

2.3. Logical Dimension

In a number of cases, the equivalence between two pieces of information across two datasets is implicit but can be inferred. We outline two main heterogeneity problems in that group.

Class heterogeneity. This is the case of two resources belonging to different classes for which an explicit or an implicit hierarchical relationship is defined ("*Person*" and "*Composer*" in Fig. 1). Moreover, two instances referring to the same object can belong to two different subclasses of a given class.

Property heterogeneity. At this level, the equivalence between two values is deduced after performing a reasoning task (cf. Fig. 1): $<< i2 >, - : composed, \text{“}Moonlight\ sonata''@en >$, $<< i4 >, - : composedBy, < i1 >>, << i4 >, - : title, \text{“}Sonate\ au\ clair\ de\ lune'' >$. The comparison process has to go beyond the value and property levels by comparing explicitly and implicitly specified values of the two entities.

2.4. Data Quality Dimension

Quality related issues can be observed on any of the three levels discussed above and can appear as a source of heterogeneity, therefore, we consider these aspects as a separate (transversal) category. The topic of data quality has been of interest for many years to the semantic web community [32, 33]. We will provide several examples of heterogeneities related to data quality that can potentially hinder the instance matching task.

Transgression to best practices. Data representations can differ depending on the degree to which the semantic web best practices are respected in the data publishing process. The list of transgressions is long: a missing language tag, the introduction of inappropriate symbols such as '♯' that are supposed to replace missing information (while a good practice would be to ignore what we do not know), the use of a string literal instead of a URI to identify an object, and so forth.

Value type heterogeneity. This heterogeneity type concerns differences in encoding data, as for example, representing an

age-value as a string or as a number, or not representing the date in a standard date format, but as a string. Multiple data unification techniques can be applied to solve this problem. The benchmark data generators SPIMBENCH [34] and LANCE [35] focus on these issues by applying value transformations.

Dataset currentness. The temporal evolution (or the lack thereof) of data and its dynamicity [33] can lead to conceptual issues across datasets. For example similar or identical classes in terms of semantics can be applicable to a given group of instances only during defined periods of time (e.g., "Orchestra-Conductor").

In Section 5, we discuss the positioning of our approach with regard to the heterogeneities and the methods for their resolution presented here.

3. Doing Linked Open Data with Less User Effort

The data linking process commonly follows a pipeline consisting of three main steps [1]: (1) preprocessing, where data is prepared for linking and a number of system parameters are set, (2) matching, where instances are compared by the help of an aggregation of similarity measures and (3) post-processing, where erroneous links are removed and / or new links are inferred. For extensive surveys of data linking approaches, we refer the reader to [1, 2, 3]. Here, we focus in more detail on the phase that takes place before the actual instance comparison. We argue that the preparation of data and the configuration of the linking tool constitute a major part of the effort with regard to the linking task. Moreover, this effort is often required from the user, leading to a pressing need of automation of this process. Therefore, we pay particular attention to approaches that propose (semi-)automatic solutions to the preprocessing and configuration tasks.

Several of the most commonly used linking tools [4, 5, 36] require prior knowledge provided by either the user or another tool in order to proceed to the linking task. This knowledge is expressed in the form of *linking rules*, describing under which conditions two instances should be compared and linked. There are two main configuration groups of elements to feed to the linking tool: (1) types (classes) of instances to align as well as a set of properties across the two datasets whose values to compare, and (2) a set of similarity measures, together with thresholds and possibly an aggregation function. We discuss these two groups in the following subsections.

3.1. Selecting Classes and Properties

The choice of types of instances that defines the pool of linking candidates is often left to the user (considering a dataset as a set of resources belonging all to the same class), although certain systems attempt to identify the equivalent classes automatically by applying ontology matching techniques [26, 36].

The properties to compare are selected manually or by the help of key discovery tools—this choice is crucial for it predetermines the outcome of the linking task. Intuitively, instances having common values for highly discriminant sets of properties (keys) are likely to be representing the same real-world objects. While many approaches to automatic key discovery from RDF data exist [6, 30, 37, 38, 39, 40, 41], their use for data linking is not always straightforward. Most of these tools produce large numbers of keys valid on a single dataset with no assessment given of their likelihood to discover links. For example, a property containing a record's identifier in a bibliographical database will be identified as a key in two datasets containing the entries of musical works of two libraries independently on one another, but it will be of no use for the linking task, since the two libraries use different identifiers for the same work. An exception is [30], which considers keys valid on two datasets. In addition, key discovery systems do not consider the heterogeneity of the properties used to describe instances across datasets, which compromises the usefulness of the keys for the linking task. In an attempt to overcome this issue, the authors of [37] present measures of the quality of link keys, valid on two datasets, in order to facilitate their selection. Two recent studies [8, 42] propose approaches that attempt to close the gap between key discovery and data linking tools, allowing to produce a list of keys, valid on two datasets simultaneously and ranked with respect to their usefulness for the particular data

linking task at hand.

3.2. Learning Link Specifications

Link specification is defined in [43] as (i) the setting of the elements to compare from two knowledge bases, (ii) the setting of a complex similarity metric via the combination of several atomic similarity measures, and (iii) the setting of thresholds for these similarity measures. The (semi-)automatic link specification approaches of which we know have focused predominantly on (ii) and (iii)—configuration parameters of type (2) that can either be set by the user or learned from data in a semi–supervised or unsupervised manner. Two main categories of *semi-supervised* learning methods emerge: *active* [44, 43, 45] and *batch* [7, 46] approaches. *Batch* approaches require a large amount of candidate links as input to learn the classifiers while *active* approaches proceed iteratively and for each iteration the user is asked to label a set of generated links until the maximal number of iterations is reached or the fitness value is greater than a given threshold. *Unsupervised learning* methods attempt to surpass the necessity of human labeled examples [47]. A method based on a refinement operator that only needs positif examples that are more often available than negative ones is proposed in [48], while [49] propose an approach implemented in KnoFuss [50], based on a genetic programming algorithm learning iteratively the optimal similarity parameters. However, it is required from the user to set the fitness function and to specify the fitness measures, thresholds and the maximum number of iterations. Certain releases of the well-known data linking tool LIMES [4] include both EAGLE [44] and WOMBAT [48] as link specification algorithms,[12] while SILK [5] includes ActiveGenLink [7].

We discuss how our approach positions with respect to end-user configuration effort reduction in Section 5, right after presenting it.

[12] E.g., *limes-core-1.2.1*.

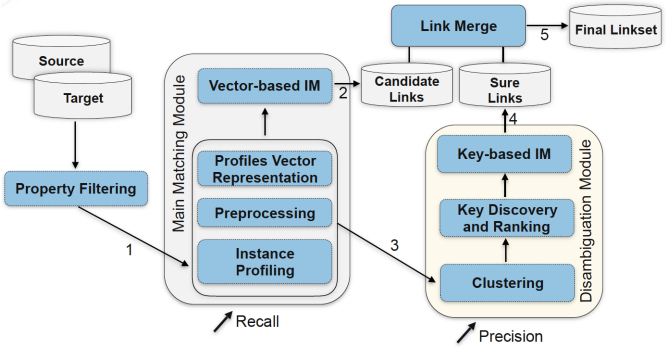


Figure 2: Processing pipeline of *Legato*.

4. Data Linking with *Legato*

We proceed to present the *Legato* framework, illustrated in Figure 2. The system takes as an input two RDF graphs (particularly, instances of the same type). The datasets are automatically preprocessed and prepared for comparison, and then a set of links is generated, as a result of an instance matching, instance disambiguation and link selection (or link merging) procedures. Note that in its default release, the system takes one parameter as input: a pair of types (classes) of instances and optionally a global similarity threshold value. However, for the data-aware user, a customisation of *Legato* is possible with regard to two additional parameters, giving rise to two version of the tool - an automatic and a manual one (see below for details). The approach and the results presented in this paper concern the automatic default release of *Legato*.

Before we proceed, we introduce definitions and notation. In the context of this work, an RDF graph or an RDF dataset[13] is defined in accordance to the dataset definition in the Vocabulary of Interlinked Datasets (VoID):[14] "A dataset is a set of RDF triples that are published, maintained or aggregated by a single provider". An RDF triple is a set of elements $< s, p, o >$, where s is a subject (a URI or a blank node), o is an object (a URI, a literal or a blank node) and p is a predicate (property, relation) (identified by a URI). We use the terms "instance" or

[13] The two terms are used interchangeably. To improve readability, we use the term "dataset" as a shortcut to "RDF dataset" and the term "graph" as a shortcut to "RDF graph".

[14] `http://vocab.deri.ie/void`

"resource" to identify an entity of interest described in a graph (e.g., a musical work or a composer).

A key is a central notion to the functioning of our system. We comply to the open world assumption-compatible definition of a key used in [6].

Definition 4.1 (RDF Dataset Key)**.** Let G be an RDF graph, let $subj(G)$ be the set of resources in G and let $pred(G)$ be the set of properties in G. We define a *key* denoted by K as the set $K = \{P : P \subseteq pred(G)$ and $\nexists s_1, s_2 \in subj(G)$ such as $p(s_1) = p(s_2)$ $\forall p \in P\}$, where $p(s_1)$ and $p(s_2)$ are the values of the property p for the resources s_1 and s_2, respectively. A set of properties P is considered as a *minimal key* if it is a key and there is no subset of P, which is a key.

A *CBD*, for Concise Bounded Description, allows to represent a given resource r by a subgraph such that all triples of this subgraph have as a subject r or a blank node connected to r or are reifications of statements of that subgraph. We cite the definition given by w3c.[15]

Definition 4.2 (Concise Bounded Description)**.** The Concise Bounded Description CBD of a resource r in an RDF graph G is a subgraph of G denoted by $CBD(r)$ identified as follows:

- Include in $CBD(r)$ all statements in G where the subject of the statement is r;
- Recursively, for all statements identified in $CBD(r)$ thus far having a blank node object, include in $CBD(r)$ all statements in G where the subject of the statement is the blank node in question and which are not already included in $CBD(r)$.
- Recursively, for all statements included in $CBD(r)$ thus far, for all reifications of each statement in the source graph, include the CBD(rdf:Statement) of each reification.

Definition 4.3 (Data Linking)**.** Given two graphs G and G' containing two equivalent classes C and C', respectively, the data linking problem consists in discovering all relations of identity across the instances of these classes. The outcome of this task is a set of links declared by `owl:sameAs` statements on a subset of the cartesian product of the elements of C and C'. We refer to G and G' as a source and target dataset, respectively, while the resources of C and C' are referred to as source and target resources, respectively.

Note that the given definition restricts the linking task to identity relations only, which are in the scope of this study. Relations of arbitrary types can be of interest in the general case. In that, Definition 4.3 serves the purposes of this paper which deals with the special data linking problem of deduplication, and therefore provides a particular case of a larger problem.

With these definitions at hand, we proceed to describe the framework of *Legato*.

Property Filtering. As we have seen in Section 2, *key heterogeneities* hinder the resources comparison, mainly because properties concerned with this heterogeneity type are erroneously likely to be considered as linking rules parameters, as discussed in Section 3.1 (certain OAEI datasets from 2016 and 2017 are rich with such examples[16]). If a linking tool uses these keys to compare instances, it will fail to find a correspondence. A way of going around this problem is to remove properties with such values, that we will call *problematic properties*, before proceeding to data comparison. We propose to identify automatically these properties by discovering all mono-property keys that are valid over *both* datasets to be linked (in that we consider the union of the two input datasets as a single dataset), i.e., each object for such a property has at most one subject in *both* graphs.

Note that the property filtering module can be seen as a preprocessing step. We analyse its impact on the global linking quality in our experiments (Section 6).

Main Matching Module. The main matching module consists of the following components.

CBD-based Instance Profiling. A core feature of our ap-

[15] `https://www.w3.org/Submission/CBD/`

[16] `http://islab.di.unimi.it/content/im_oaei/2016`

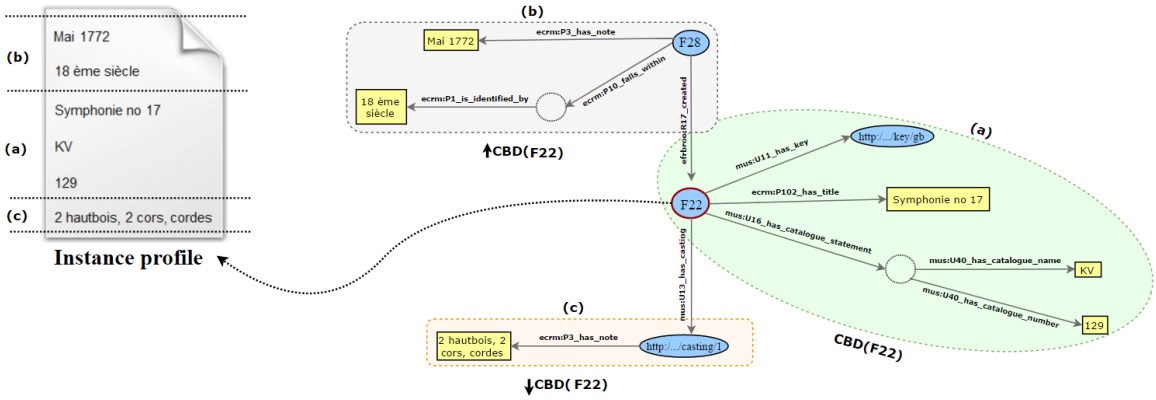


Figure 3: Constructing an instance profile by using the CBD of a resource and the CBDs of its successors and predecessors in the graph.

proach is the representation of instances as text documents and their projection to a vector space. Particularly, each resource is represented by a set of literals considered as relevant to its description, based on a choice of *CBD* subgraphs (Fig. 3). We extend the *CBD* definition by considering the descriptions of neighbouring nodes of a resource r in its graph. A *CBD* subgraph is a directed one, where the orientation is implied by the order of the subject and the object in a triple. This allows us to introduce the notion of successors of a node r (the nodes that are in a triple of which r is a subject) and the predecessors of r (the nodes that are in a triple of which r is an object). We provide the following definitions.

- $\uparrow CBD(r)$ defines the scope of resource description including $CBD(r)$ and the CBDs of its direct *predecessors*.
- $\downarrow CBD(r)$ defines the scope of resource description including $CBD(r)$ and the CBDs of its direct *successors*.
- $\updownarrow CBD(r)$ defines the scope of resource description including $CBD(r)$, $\uparrow CBD(r)$ and $\downarrow CBD(r)$.
- $CBD^*(r)$ defines the scope of resource description including one of the *CBDs* cited above, i.e., $CBD(r)$, $\uparrow CBD(r)$, $\downarrow CBD(r)$ or $\updownarrow CBD(r)$.

We refer to an instance representation obtained on that basis as an *instance profile*, defined as follows.

Definition 4.4 (Instance Profile)**.** Let G be an RDF graph, let r be a resource in G and let $L(G)$ be the set of literals found in G. We define the *instance profile* of r as the set

$$f(r) = \{l_r : l_r \in L(G) \wedge l_r \in CBD^*(r)\}.$$

Fig. 3 provides an illustration for an instance of the class F22 from the DOREMUS ontology[17] (a class of music works, example taken from the DOREMUS OAEI 2017 data). Selecting the most relevant profile depends on the way the resources are modeled in the graph. Without any user intervention, the default setting of *Legato* represents the instances only by literals found in their *CBDs*, which shows to produce good results in terms of F-measure on all benchmarks in our evaluation (Section 6). Note however, that this parameter is modifiable in the open source release of our system. Avoiding property-based comparison addresses the remaining ontology-level heterogeneities introduced in Section 2.

Instance Profiles to Vectors. Once all resources in both datasets are profiled, the resulting documents are processed in order to prepare data for the matching task. This includes tokenization and stop-words removal by applying NLP filters. The set of *instance profiles* in both datasets are indexed in a standard manner by using all remaining terms. We project the instance profiles to a vector space of a dimension limited to the number of these terms and weight them by using their TF-IDF (Term Frequency-Inverse Document Frequency) scores per instance.

Vector-based Instance Matching. The correlation between the vectors of the resources, expressed by the cosine similarity

[17] www.data.doremus.org

measure, is used as a proxy for the similarity of resources. The use of a similarity measure is tangled to the choice of a similarity threshold. In this step, it is empirically fixed to 0.2 (deliberately low) in order to capture a large number of links and ensure high recall. In order to ensure 1 : 1 type of matching, for each instance from the source dataset, the one from the target dataset that has the highest similarity score greater than the threshold is selected. In case of ties, the instances are handled by the disambiguation module described below. As an outcome of this process, a first linkset (a short for "set of links") is produced, called *candidate links* (Fig. 2).

The main matching module ensures high recall. To improve the matching quality and precision, we perform a post-processing step, described next, allowing to filter out erroneous links that may have been generated at this step and add new quality links.

Instance Disambiguation Module. Taking as an input the vector space representations of the indexed instance profiles, the algorithm proceeds to cluster *within each data set* highly similar (in terms of their vector space similarity) instances by relying on the generic agglomerative bottom-up *hierarchical clustering algorithm* [51]. This results in the formation of a number of clusters of instances within each dataset. A cluster matching procedure across the two datasets, using a distance metric on the cluster centroids, allows to isolate pairs of corresponding clusters, where the first one belongs to the source dataset and the second one—to the target dataset. Each pair of corresponding clusters is then analyzed separately and their respective instances are compared, this time on a property basis. The effectiveness of the comparison process depends on the quality of the selected properties. The RANKey algorithm [8] is developed to discover keys that are valid on *two* datasets, ranked with respect to the performance achieved by a linking system using these keys in its configuration. This allows to select the set of properties over two graphs that guarantee the best linking result. We apply that algorithm independently on each pair of corresponding clusters, considering them as a source and target datasets to be linked. In that, we identify the discriminant properties among these clusters that would have remained "diluted" in the global graphs. This allows to disambiguate the highly similar instances in each pair of clusters and maximises the rate of correct alignments. This component of *Legato* is illustrated in Fig. 4. As a result of this process, we end up with a second set of links, that we call *sure links* (the choice of this name is motivated by the fact that the instance clustering and property-based link discovery ensures high precision and high quality of these links).

Note that this component of *Legato* shares certain similarities with the well-known blocking techniques used in ontology matching, although it differs in its mechanism, motivation and application. Blocking aims at isolating disjoint sets of potential matching candidates based on certain property values so as to ensure that comparison is performed only among comparable entities and thus reduce the search space and computational effort. As an example, works by the same composer would form a block. Our instance disambiguation module has a different motivation: we aim at creating clusters of instances that, rather than being similar with respect to a small set of property values, are different with respect to only very few property values. An example would be all piano sonatas by Beethoven that would only differ by their music keys (e.g., G minor vs. A major, although same composer, genre, title, instrument, etc.). The motivation comes from observations on real-life data containing many blocks of highly similar entities. The comparison of instances belonging to such clusters allows to determine the discriminating property (the music key in our example), which would have been difficult to determine by taking the entire dataset or a block (of all works by Beethoven).

Link Merge. Finally, a merge operation is performed on the two linksets generated previously. The set of *sure links* will be taken as a catalyser on the links in the *candidate links* set and directly fed to the final linkset, because of the high precision in the process of generation of the links that it contains. For each link between two resources r_s and r_t l=(r_s, r_t) in the set of *candidate links*, the module searches over the set of *sure links* for a link between a source resource r_s and a target resource $r'_t \neq r_t$.

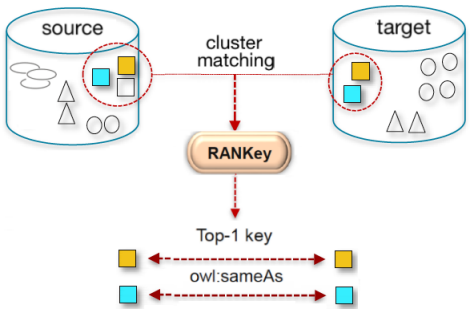


Figure 4: Instance disambiguation by clustering, cluster matching and key ranking techniques.

If found, the link $l=(r_s, r_t)$ is deleted from the *candidate links* set. The remaining links in the *candidate links* set, merged with those from the *sure links* set, are then fed to the *final linkset*.

The disambiguation module followed by the link merge procedure can be seen as post-processing steps. In our experiments, we analyse their impact on the global linking quality.

5. Discussion and Positioning

After having introduced our approach, in this section we discuss how it stands with respect to state of the art methods and issues presented in Sections 2 and 3.

5.1. Positioning with Respect to Dataset Heterogeneities

Most of the heterogeneities on *value level*, as well as certain quality-related heterogeneities (such as the value-type heterogeneities) are relatively well-understood and assisted by various similarity measures, external resources and data unification methods [15]. A large number of instance matching tools of reference rely on string similarity measures and their combinations, coupled with thresholds that condition the matching decision [4, 12, 13]. *Legato* is only partially an exception to this tradition—we do consider instance similarities, but similarity between strings is not explicitly computed. Instead, we represent instances as bags of words, which allows their information retrieval kind of indexing and a projection onto a vector space, in a similar spirit as [22, 28]. We rely more strongly on the data presentation, or instance profiling (taking place at the preprocessing step) and matching selection (post-processing) than on the actual instance similarity computation.

While *logical heterogeneities* are in the realm of reasoning—a field of research in its own right—*Legato* focuses on issues from the *ontological level* that appear to be more challenging for the majority of the linking tools. The representation of instances that we adopt allows to avoid a number of structure- and property-related heterogeneities found on ontological level. Particularly, the use of CBD vs. a fixed distance n to the resource, as proposed for example in [29], tackles heterogeneities on ontological level (excluding key heterogeneity discussed below), allowing to ensure that literal values of relevance will be included in the description of an instance. Note that [49] defines and compares attribute sets across graphs that are equivalent to CBDs. In contrast, we also look into the CBDs of the neighbouring nodes of the resources of interest, in order to ensure that information found at a greater depth is taken into consideration.

The presence of unstructured information, such as properties containing textual descriptions, although common in real-world examples is, to our knowledge, not directly handled by linking systems. In the same line of thought, other types of single-property keys, such as instance identifiers specific to a single data provider, present a hindrance for property-based linking systems. *Legato* adresses these issues by the help of a property filtering method, allowing to automatically identify and remove from the process "problematic" attributes. This issue pertains to the key heterogeneity category outlined above.

Finally, the paradigm adopted by state-of-the-art systems traditionally stands on the plausible premise that heterogeneities are a major hindrance to the instance comparison task. However, the problem of disambiguating between very similar descriptions of yet different resources has received less attention. An example can be two different music works by the same composer, written in the same genre and key for the same instrument. Comparing their property values is likely to lead to the generation of false positives. The identification of the discriminative properties for such groups of instances can be a difficult task for automatic key-discovery methods, because these groups of instances are found in larger datasets where these

properties are not keys. Therefore, in addition to data heterogeneity, *Legato* pays attention to the problem of data similarity, in an attempt to disambiguate cross-datasets entities by introducing an unsupervised learning post-processing module. It aims to identify and isolate clusters of highly similar instances across the two datasets in order to enable efficient key identification on these clusters.

5.2. Level of Automaticity

We position our approach with respect to user configuration effort reduction methods, reviewed in Section 3.

With regard to property selection, in contrast to state-of-the-art tools [4, 5], our system takes a holistic stance by representing an instance as the set of literals collected from a CBD-defined subgraph. In that, no property selection is required from the user at the main matching module. At the instance disambiguation and link selection phase, to the best of our knowledge, *Legato* is the first tool of its kind to apply a key discovery algorithm combined with a key ranking tool in its internal mechanics, which allows to automatically select the discriminative set of properties that guarantee the best linking performance. The types of instances to compare has to be manually specified.

As we have seen in the discussion in Section 3, fully automatic link specification remains a challenge, even for specialised approaches. In all cited methods, the user input is required under one form or another (we detail on that in Section 6.6). With regard to the choice of similarity measures and thresholds, we take a different and simpler approach as compared to machine learning-based techniques. In the first place, given the vector space framework that we adopt, the choice of the cosine similarity combined with a TF-IDF weighting scheme that allows to take into account the overall textual content of the instance descriptions, appears to be natural. *Legato* depends on a single threshold for the cosine similarity. It has been fixed as an outcome of an extensive empirical analysis, although the user is given the possibility to easily modify this parameter. In a similar spirit as [52], during the main matching phase, the similarity threshold is deliberately kept very low, so that the system can discover a large number of candidate links, ensuring high recall. Improving precision is handled at the time by the preprocessing module (particularly the filter on problematic properties), and by the instance disambiguation and link merging module (see details in the preceding section).

While we do not claim that the tool provides a fully automatic solution, we have attempted to offer punctual solutions to several issues that are not integrally handled automatically by a couple of the most popular state-of-the-art tools, such as (1) property filtering, (2) property selection and (3) similarity measures combination and tuning. A direct comparison in terms of user configuration to other tools is difficult, because of the varying underlying principles of these tools. However, we attempt to compare *Legato* to several other popular and freely available systems - SILK, LIMES, AML, as well as three automatic configuration versions of LIMES (relying on batch, active or Unsupervised learning) (cf. Table 1). The comparison criteria in the table come from the union of the sets of parameters in the configuration files of SILK and LIMES, as well as the set of "potential" parameters of *Legato* and AML. We use the word "potential" to indicate that a number of parameters are currently hard-coded both in AML and Legato (giving rise to automatic and manual versions of these tools). For fairness of comparison, we have chosen to include both versions of the two systems to the table. Regarding AML, note that there is no officially released user-tuneable version of the system: in the current release all parameter values are hard-coded except for the type restriction and a threshold. We therefore consider as a manual version of the tool a one containing all parameters that could be included in a configuration file of the tool or could be set by a user with programming skills.[18] The table allows us to conclude that *Legato* and AML are least demanding in terms of user intervention, while the experiments reported in the following section give an advantage to *Legato* in terms of performance.

[18] Source: personal exchanges with the authors of AML.

Configuration element	*Legato* (manual)	*Legato* (automatic)	Silk, Limes (standard)	Limes + Batch L.	Limes + Active L.	Limes + Unsupervised L.	AML (manual)	AML (automatic)
Types	y (=yes)	y	y	y	y	y	y	y
Properties	y	n (=no)	y	y	y	y	n	n
Similarity measures	n	n	y	n	n	n	y	n
Local sim. thresholds	n	n	y	n	n	n	y	n
Global sim. threshold	y	y	y	y	y	y	y	y
Instance profile type	y	n	n	n	n	n	n	n
Machine learning alg.	n	n	n	y	y	y	n	n
Training data	n	n	n	y	n	n	n	n
Label link candidates	n	n	n	n	y	n	n	n
Matching strategy	n	n	n	n	n	n	y	n
Total ratio	4/10	2/10	5/10	5/10	5/10	4/10	5/10	2/10

Table 1: End-user configuration effort comparison.

6. Experimental Evaluation

The experiments reported in this section aim (1) to assess the effectiveness of the internal modules of *Legato*, (2) compare the system to other linking tools and (3) compare it to approaches for automatic link specification. *Legato* was implemented in Java 8 and the experiments were conducted on a machine running under Windows 10 over an Intel Core i5-5300U, with 2.30 GHz CPU and 16 GBytes RAM. Note that the automatic (default) version of *Legato* is evaluated here. The system is available as an open source release at `https://github.com/DOREMUS-ANR/legato`.

6.1. Experimental Setting

We begin by describing the evaluation framework that we have established.

Datasets. For the various experiments carried out and reported in this paper, we have relied on data coming from the Instance Matching evaluation campaigns of OAEI (IM@OAEI) from 2015 to 2017.

• **DOREMUS datasets.** One of the main results of the DOREMUS project is the representation of the catalogs of three French cultural institutions as knowledge graphs following a specifically designed for this purpose model [9] and their publication on the web. This has resulted in the creation of (currently) three knowledge graphs—one per partner institution.[19] The DOREMUS benchmarks have been built together with librarian experts form the BnF and the Philharmonie de Paris. The basis for the construction of the benchmark is a set of pairs of identical works (given in a synthetic table) that have been manually selected by the experts such that one work in each pair belongs to the catalog of the BnF and the other - to that of the Philharmonie. In that process, the experts have identified the heterogeneities that each pair of works manifest. As it can be seen in the table,[20] this resulted in a set of heterogeneity types, specific to the DOREMUS data (note that these types have been generalized in our heterogeneity types categorization provided in Section 2): (1) numbers vs. letters in the tiles, arabic vs. roman numbers in the titles, (2) differences in spelling, (3) missing catalog numbers, (4) different catalogues, (5) multilingual titles, (6) specific characters, (7) differences in the lengths of the property chains that lead to the value of interest (graph depths), (8) different property names for the same entity types, (9) missing descriptions (missing property values), (10) missing titles, (11) use of synonyms. Whenever a given pair of works manifests one of these heterogeneity types, this has been indicated

[19] `https://github.com/DOREMUS-ANR/knowledge-base/tree/master/data`

[20] `https://docs.google.com/spreadsheets/d/19dLjabt_fggTVNuM7XW9CUZkuB1JgoH9xKWQgLZv_Y4/edit#gid=1271677916`

in the table. This resulted in two RDF datasets benchmarks released by OAEI in 2016 and 2017 (the datasets are not identical, the underlying model as well as their sizes have evolved from one year to the other):

- DOREMUS 2016 consists of three datasets of different sizes and scopes: **9-HT**, **4-HT** (for heterogeneities) and **FP-trap** (for false positives trap). The data is available and described at `http://islab.di.unimi.it/content/im_oaei/2016/#doremus`.
- DOREMUS 2017 consists of **HT** (for heterogeneities) and **FPT** (for false positives trap). The data is available and described at `http://islab.di.unimi.it/content/im_oaei/2017/#doremus`.

The particularity of these benchmarks is that they contain datasets that were particularly designed to challenge the capacity of linking tools to correctly disambiguate highly similar in their descriptions instances (FP-trap (2016) and FPT (2017)). All data follow the same model and therefore share significant number of vocabulary terms. Nonetheless, these datasets are highly heterogeneous in terms of all other ontology-dimension heterogeneities and many value-dimension heterogeneities (Section 2) .

• **Synthetic datasets.** We additionally evaluated *Legato* on synthetic benchmark datasets from three consecutive years of the OAEI campaign.

- SPIMBENCH 2015: This includes three benchmark datasets generated through the Semantic Publishing Instance Matching Benchmark (SPIMBENCH) [34] by transforming the source instances based on their values and semantics (the **Val-Sem** dataset), on their values and structures (the **Val-Struct** dataset) and on their values, structures and semantics (the **Val-Struct-Sem** dataset). The data is available and described at `http://oaei.ontologymatching.org/2015/im/index.html`.
- SPIMBENCH 2016: This comprises **SPIMBENCH small**, that we denote **SB-s** and **SPIMBENCH large**, that we denote **SB-l**, two datasets of different sizes, produced by following the same strategy as described above. The data is available and described at `http://islab.di.unimi.it/content/im_oaei/2016/#synthetic`.
- SPIMBENCH 2017: This includes **SPIMBENCH sandbox**, that we denote **SB-s** (the year of edition helps disambiguate the two SB-s notations) and **SPIMBENCH mainbox**, that we denote **SB-m**, datasets of different sizes, produced by following the SPIMBENCH transformation patterns.

Scenarii. We consider five evaluation scenarii. First, we evaluate *Legato* with respect to three of its core components, by assessing (1) the efficiency of the automatic property filtering module by measuring the impact of automatically identified problematic properties on the quality of the generated links (Section 6.2), (2) the impact of the choice of instance profile (Section 6.3) and (3) the use of keys to efficiently disambiguate instances and assess and select links by improving recall (Section 6.4). Then, (4) we assess the overall performance of *Legato* by comparing it to state-of-the-art systems and participant-systems to the IM@OAEI campaigns on a large variety of datasets (Section 6.5). Finally, (5) we confront *Legato* with EAGLE and WOMBAT (in its two versions)—two automatic link specification methods implemented with LIMES (Section 6.6).

We use three well-known performance measures: Precision (P), Recall (R) and F-Measure (F-m). P and R evaluate the *correctness* and the *completeness* of the generated links, respectively, while F-m is their harmonic mean.

Tuning. We have conducted a series of experiments by varying the cosine similarity threshold value (cf. Section 4) observing its impact on F-m, P and R. We observed that the best results of *Legato* on all data were achieved with a threshold of 0.2. We report as an example the results obtained on the DOREMUS 2017 HT dataset in the form of couples (F-m, threshold): (0.92, 0.1), (0.93, 0.2), (0.81, 0.3), (0.59, 0.4), (0.29, 0.5), (0.14, 0.6), (0.1, 0.7), (0.02, 0.8), (0.0, 0.9). Analogical behaviour has been

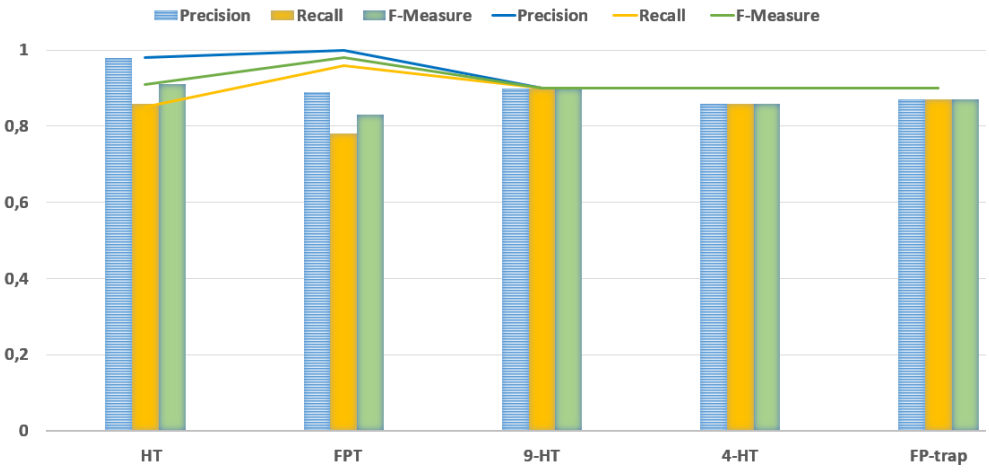


Figure 5: Property filtering evaluation on DOREMUS datasets: without (histograms) and with (curves) automatic removal of problematic properties.

observed with other datasets. This threshold is low, which guarantees a fair trade-off between the matching module (ensuring high recall) and the disambiguation and merge module (improving precision).

6.2. *Effectiveness of Property Filtering*

In this experiment (scenario 1), instances are compared first by considering all the properties and then after removing the automatically identified problematic ones. An improved performance is expected after property filtering. We experiment on the DOREMUS datasets from 2016 and 2017, as reported in Fig. 5, articulating the assessment in two phases: without (histograms) and with (curves) the automatic removal of problematic properties. The experiments show that applying the automatic property filtering module allows to improve the linking quality for all datasets except for HT and 9-HT, where this module has no impact. Further experiments and analysis on these two datasets reveal that removing all the other properties separately does not improve the results either, which indicates that no problematic property has been "missed" by the module.

6.3. *Effectiveness of Instance Profiling*

In these experiments (scenario 2), we analyse the behaviour of *Legato* with respect to different choices of instance profiles, expressed as four *CBD*-based instance representations (Section 4): *CBD*, $\uparrow$ *CBD*, $\downarrow$ *CBD* or $\updownarrow$ *CBD*. We have selected two real-world datasets from OAEI 2017, as well as four synthetic ones from OAEI 2015 and 2016. As expected, the effectiveness of instance profiling depends on how the data is modelled (Figure 6). Particularly, these tests show that the choice of a $\updownarrow$ *CBD* profile is relevant for the real-world datasets HT and FPT data, as the highest F-measure scores are achieved with that representation (91% and 98% for HT and FPT, respectively), while for the synthetic datasets the relevant information is located in their direct *CBD*s. For those datasets, we can also deduce that taking into account the description of predecessors does not impact the matching decision. The results do not allow to conclude on the choice of an instance profile in the general case. An understanding of how data is modelled (where to look for important information) is needed in order to guarantee a choice of a profile that maximises the outcome. Based on our results, we set $\updownarrow$ *CBD* and *CBD* as profile parameters for the real-world and the synthetic datasets, respectively.

6.4. *Effectiveness of Post-processing*

We evaluate the efficiency of the post-processing step (scenario 3) of *Legato* consisting of an instance disambiguation and a link merge module. In that, we execute *Legato* with and without performing this step. By taking as a reference the set of *candidate links*, generated at the main matching step, we measure the proportion of links that fall on its intersection with the *sure links* set, dubbed **#safe_links**, as well as the propor-

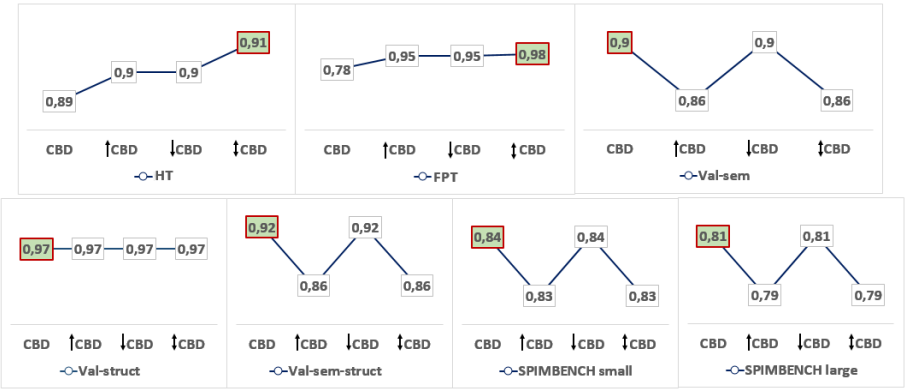


Figure 6: CBD-based instance profiling evaluation on reference datasets (F-measures).

dataset / profile		#safe_links	#deleted_links	#added_links
HT	*CBD*	$\simeq$ 10%	$\simeq$ 2%	0%
	$\downarrow$ *CBD*	$\simeq$ 14%	$\simeq$ 2%	0%
	$\uparrow$ *CBD*	10.5%	$\simeq$ 2%	0%
	$\updownarrow$ *CBD*	$\simeq$ 5%	$\simeq$ 1%	0%
FPT	*CBD*	$\simeq$ **55**%	$\simeq$ **18**%	$\simeq$ 3%
	$\downarrow$ *CBD*	$\simeq$ 21%	$\simeq$ 3%	0%
	$\uparrow$ *CBD*	$\simeq$ 15%	0%	0%
	$\updownarrow$ *CBD*	$\simeq$ 15%	0%	0%

Table 2: Post-processing evaluation on OAEI 2017 datasets.

tion of links deleted or added by the merge module, dubbed **#deleted_links** and **#added_links**, respectively, in order to form the *final linkset*. Table 2 shows the results on the DOREMUS 2017 data. We observe, as hypothesised, that performing the post-processing step is significant in the presence of highly similar instances (dataset FPT). We also notice that the precision boost is pronounced in the case of a simple *CBD* profile, which (or variants of which) is most commonly used in the existing instance representation approaches (cf. Section 3). This shows the potential of the post-processing module of *Legato* to make up for possible flaws in the instance representation.

6.5. General Evaluation

In this experiment (scenario (4)), we assess the performance of *Legato* in its complete automatic version, performing both *property filtering* and *post-processing*. We compare *Legato* to the participant tools to IM@OAEI on all benchmarks from the years 2015, 2016 and 2017 and additionally with SILK [5] for the DOREMUS 2017 data. In 2015, two systems participated on the three tasks: STRIM [53] and LogMap[21] [36]. In 2016, the systems LogMapIm [54], AML[22] [55, 52] and RiMOM [28] participated on the two proposed tasks. In 2017, in addition to *Legato*, which participated to OAEI for the first time, AML, I-Match [56] and LogMap participated to the SPIMBENCH task, while AML, I-Match, NjuLink[23] [57] and LogMap participated to the DOREMUS task. In addition to the participant systems, we have included to the comparison SILK[24] in its 2.6.1 version on the 2017 data. Note that the comparison to SILK was made by using the best keys in its link specification as identified by the RANKey algorithm [8], namely the

[21] https://github.com/ernestojimenezruiz/logmap-matcher/
[22] https://github.com/AgreementMakerLight/AML-Project
[23] https://github.com/nju-websoft/njuLink
[24] https://github.com/silk-framework/silk

Benchmark (year)	System	P	R	F-m	size
HT (2017)	*Legato*	0.930	0.920	0.930	476
	AML	0.851	0.479	0.613	
	I-Match	0.680	0.071	0.129	
	LogMap	0.406	0.882	0.556	
	NjuLink	**0.966**	**0.945**	**0.955**	
	SILK	0.34	0.12	0.18	
FPT (2017)	*Legato*	**1.000**	**0.980**	**0.990**	150
	AML	0.914	0.427	0.582	
	I-Match	**1.000**	0.053	0.101	
	LogMap	0.119	0.880	0.210	
	NjuLink	0.959	0.933	0.946	
	SILK	0.45	0.2	0.27	
SB-s (2017)	*Legato*	**0.980**	0.730	0.840	$\simeq$ 1800
	LogMap	0.938	0.763	0.841	
	AML	0.849	**1.000**	0.918	
	I-Match	0.854	0.997	**0.920**	
SB-m (2017)	*Legato*	**0.970**	0.700	0.810	$\simeq$ 1800
	LogMap	0.893	0.709	0.790	
	AML	0.855	**1.000**	**0.922**	
	I-Match	0.856	0.997	0.921	
9-HT (2016)	*Legato*	0.9	**0.9**	0.9	60
	AML	**0.96**	0.87	**0.91**	
	RIMOM	0.81	0.81	0.81	
4-HT (2016)	*Legato*	0.9	**0.9**	**0.9**	400
	AML	**0.93**	0.77	0.84	
	RIMOM	0.74	0.74	0.74	
FP-trap (2016)	*Legato*	0.9	**0.9**	**0.9**	80
	AML	**0.92**	0.85	0.88	
	RIMOM	0.7	0.7	0.7	
SB-s (2016)	*Legato*	**0.98**	0.74	0.84	$\simeq$ 380
	LogMapIm	0.95	0.76	0.85	
	AML	0.9	0.74	0.82	
	RiMOM	**0.98**	**1.0**	**0.99**	
SB-l (2016)	*Legato*	0.96	0.71	0.81	$\simeq$ 1800
	LogMapIm	0.98	0.69	0.81	
	AML	0.9	0.74	0.81	
	RiMOM	**0.99**	**1.0**	**0.99**	

Table 3: Results on different benchmark datasets for *Legato*, compared to other linking tools.

U16_has_catalogue_statement property from the DOREMUS ontology.[25] For the sake of reproducibility, we make available the SILK configuration files used for the HT and the FPT data at the following link: https://github.com/manoach/SILK-Evaluation.

Note that, among the cited systems RIMOM and I-Match do not have openly available source code or executable versions. The results reported for these tools are taken from the OAEI web site for the year 2015 or via the SEALS platform to which we had access as co-organisers of the OAEI tracks of 2016 and 2017. The runtimes of the systems participating in the different OAEI campaign editions are not reported, for which reason comparison with respect to that criterion to *Legato* is not feasible (runtimes of our system are reported in the following experiment).

The results of the evaluation are presented in Table 3. For the sake of readability, we provide only the results of the evaluation on data from OAEI 2016 and 2017. The complete results can be found at https://github.com/DOREMUS-ANR/legato/blob/master/Legato-Results.png.

We can observe that *Legato* has comparable (and in certain cases better) performance as the state-of-the-art systems on all benchmarks. These results show in particular that *Legato* is well-suited for dealing with real-world data containing difficult to disambiguate instances (FPT and FP-trap), which confirms the effectiveness of the post-processing module. Overall, *Legato* performs well when data heterogeneity is related to *descriptive* differences and all remaining heterogeneity types of the *ontological dimension* (Section 2) thanks to its *property filtering* and *indexing* techniques. As expected, our system is less effective on the synthetic data transformed with SPIMBENCH. This is explained by the fact that the heterogeneities of *value dimension* are not considered by *Legato* (for example, we did not implement string unification methods in the indexing process). Nevertheless, we consider the results of *Legato* satisfactory on these data, given that it provides comparable results as the systems, to which it was confronted, but requires significantly less user-tuning effort than many of these tools.

Table 4: Automatic LS approaches vs. *Legato* on SPIMBENCH data.

Method	System	F-m	P	R	Execution (ms)
SB-s (2016)					
UNSUPERVISED	EAGLE	0.62	0.63	0.61	74531
	WOMBAT simple	0.64	0.61	0.67	14905
	WOMBAT complete	0.64	0.61	0.67	64275
ACTIVE	EAGLE	0.36	0.28	0.50	100373
	WOMBAT simple	0.64	0.61	0.67	15958
	WOMBAT complete	0.64	0.61	0.67	61612
BATCH	EAGLE	0.64	0.77	0.55	93857
	WOMBAT simple	0.64	0.61	0.67	14079
	WOMBAT complete	0.64	0.61	0.67	61730
Legato		**0.84**	**0.98**	**0.74**	**7000**
SB-l (2016)					
UNSUPERVISED	EAGLE	0.43	0.32	0.63	1597173
	WOMBAT simple	0.43	0.32	0.63	413450
	WOMBAT complete	0.43	0.32	0.63	364643
ACTIVE	EAGLE	0.43	0.32	0.63	1668132
	WOMBAT simple	0.43	0.32	0.63	282928
	WOMBAT complete	0.43	0.32	0.63	254423
BATCH	EAGLE	0.45	0.38	0.55	1032956
	WOMBAT simple	0.43	0.32	0.63	249196
	WOMBAT complete	0.43	0.32	0.63	226922
Legato		**0.81**	**0.96**	**0.71**	**31000**

6.6. Comparison to Automatic Link Specification Methods

These experiments (scenario (5)) confront *Legato* with two approaches to automatic links specification (LS), EAGLE [44] and WOMBAT [48], that are included in the linking system LIMES. We have used the *limes-core-1.2.1* release of the system,[26] where two versions of WOMBAT are implemented: simple and complete. The simple version of WOMBAT learns links specifications and combines them to improve F-measure, while the complete version implements a refinement operator, which guarantees the best specification.

Regarding the level of automation of the linking process, in the case of the three methods a manual configuration is required. At the beginning, the user is asked to select the preferred link specification method (EAGLE or WOMBAT) and afterwards a number of parameters need to be manually fed to the system. In the first place, the user has to select the type of learning approach to apply (supervised, active or batch, see Section 3). Then, the names of the properties to compare across the source and the target datasets (types) have to be specified.

[25] http://data.doremus.org/ontology#U16_has_catalogue_statement

[26] https://github.com/dice-group/LIMES/releases

Table 5: Automatic LS approaches vs. *Legato* on DOREMUS data.

Method	System	F-m	P	R	Execution (ms)
HT (2017)					
UNSUPERVISED	EAGLE	0.45	0.97	0.29	1882
	WOMBAT simple	0.45	1.0	0.29	390
	WOMBAT complete	0.45	1.0	0.29	472
ACTIVE	EAGLE	0.44	1.0	0.28	1227
	WOMBAT simple	0.45	1.0	0.29	395
	WOMBAT complete	0.45	1.0	0.29	565
BATCH	EAGLE	0.44	1.0	0.28	7733
	WOMBAT simple	0.45	1.0	0.29	**387**
	WOMBAT complete	0.45	1.0	0.29	459
Legato		**0.93**	0.93	**0.92**	69000
FPT (2017)					
UNSUPERVISED	EAGLE	0.57	1.0	0.4	911
	WOMBAT simple	0.57	1.0	0.4	**232**
	WOMBAT complete	0.57	1.0	0.4	294
ACTIVE	EAGLE	0.57	1.0	0.4	1399
	WOMBAT simple	0.57	1.0	0.4	238
	WOMBAT complete	0.57	1.0	0.4	294
BATCH	EAGLE	0.57	1.0	0.4	813
	WOMBAT simple	0.57	1.0	0.4	235
	WOMBAT complete	0.57	1.0	0.4	294
Legato		**0.99**	1.0	**0.98**	16000

For each pair of properties to compare, the user does not have to select the similarity measure and its associated threshold, but a global similarity threshold needs to be manually set. For the batch algorithm, a small sample of `owl:sameAs` links over the data is also required. In our experiments, the choice of properties to compare has been done again by using the RANKey key selection system (just as described in the previous subsection). This resulted in the choice of the following keys on each of the datasets:

SB-s: {bbc[27]:bbc/primaryContentOf, bbc:creativework/description}

SB-l: {bbc:creativework/title, bbc:creativework/about}

HT & FPT: {mus[28]:#U16_has_catalogue_statement}

The results are given in Tables 4 and 5. For the sake of reproducibility of the experiments, all configuration files used for the different LS tools for LIMES, together with the data and the results are made available.[29] In that experiment, we also report on the respective execution times of the evaluated systems. We see that *Legato*, in its fully automatic version where only the classes to compare need to be specified, achieves better results than the three tested methods in terms of performance. In terms of execution times, our system outperforms the other methods on the synthetic data but scales much worse on the DOREMUS real-world datasets.

We explain this difference by the characteristics of the datasets. Particularly, the sizes of the respective *CBD*-based profiles and hence documents generated in each of the two cases are different. The number of literals collected by the *CBD*-profile of the DOREMUS resources (using both successors and predecessors) is larger than those of the SPIMBENCH data (using only the *CBD* of the resource), leading to larger in size documents and hence vectors of larger dimension, which has a direct impact on the computational efficiency.

7. Conclusion and Future Work

In this work, we propose, implement and evaluate *Legato*, an open source framework for discovery of identity links across RDF graphs in the context of the web of open data. We provide an extensive inventory of dataset heterogeneities, which lie at the origin of the data linking problem. We show that *Legato* addresses efficiently many of these heterogeneities, particularly those at ontological level, by implementing efficient property filtering (data cleaning) and link selection modules, acting, respectively, at the pre- and post-processing steps of the linking pipeline. A core feature of *Legato* is its capacity to avoid the generation of false positives by disambiguating effectively highly similar instances across datasets by the help of a clustering method and a key selection and ranking algorithm. In addition, the system has the advantage of proposing a fully automatic version (only the types of instances to compare need to be indicated) - in that, nor property, neither similarity selection are performed manually.

Fully automating the data linking process is a highly challenging task. As we have seen in our related work discussions, several approaches have been proposed to handle automatic system configurations at different levels (e.g., selecting properties, or tuning similarity measures and thresholds). These approaches, even when included in the internals of a tool, do not

[27] bbc=http://www.bbc.co.uk/ontologies/

[28] mus=http://data.doremus.org/ontology

[29] `https://github.com/manoach/LIMES-Evaluation`

guarantee a process with zero user implication—to our knowledge, there does not exist to date an entirely automatic data linking system, that is also agnostic to the underlying data. In reality, there is a multitude of factors that make it difficult (or even futile) to think of a generic fully automatic linking system, and these factors strongly relate to the specificities of data with respect to domains, structure, coverage, their degree of heterogeneity, and varieties of presentation. For that reason, we have left the possibility to the alerted user to set two parameters in a customised version of *Legato*: the threshold for the similarity measure and the shape of the CBD-based instance profile.

We evaluate *Legato* on real-world music-related data and on synthetic datasets coming from OAEI 2015, 2016 and 2017. The results showed that our system is in competition with state-of-the-art tools, outperforming them on datasets containing highly heterogeneous or difficult to disambiguate instances. Additionally, we evaluate the performance of *Legato* with respect to a version of LIMES that uses automatic link specification modules and obtain better results.

Scalability is an important factor given the sizes of datasets that one has to deal with. *Legato* can be improved in that respect. One way that we deal with this issue in real-wolrd matching tasks is to partition the data with respect to the values of a property that we know that will generate groups of potential linking candidates across the two datasets and treat these groups separately (in the case of the DOREMUS data, this property is the name of the composer of a music work).[30] It is a subject of future work to automate this process and thus reduce the execution time of the system.

As observed before, our approach is sensitive to the underlying data model, particularly with regard to the CBD-based profiling. Our experiments do not allow to conclude on which CBD presentation is most appropriate in the general case. More experiments need to be performed on larger sets of benchmarks (for example, those coming form the HOBBIT project[31]) in order to confirm this observation. The choice of centroid vectors to represent clusters of instances in the disambiguation method requires further evaluation as well, under the hypothesis that such representation would provide skewed results in the presence of outliers.

In future work, we also aim to address the problem of information complementarity across datasets—how to handle entities that are described by complementary sets of properties in two different graphs and therefore have little information in common in order to compare them. We will explore the application of knowledge graph augmentation techniques in order to reconstitute the missing descriptions intersection.

Acknowledgements

This work has been partially supported by the French National Research Agency (ANR) within the DOREMUS Project, under grant number ANR-14-CE24-0020.

References

[1] A. Ferrara, A. Nikolov, F. Scharffe, Data linking for the semantic web, Semantic Web: Ontology and Knowledge Base Enabled Tools, Services, and Applications 169.

[2] M. Achichi, Z. Bellahsene, K. Todorov, A survey on web data linking, Revue des Sciences et Technologies de l'Information - ISI.

[3] M. Nentwig, M. Hartung, A.-C. Ngonga Ngomo, E. Rahm, A survey of current link discovery frameworks, Semantic Web 8 (3) (2017) 419–436.

[4] A. N. Ngomo, S. Auer, LIMES - A time-efficient approach for large-scale link discovery on the web of data, in: IJCAI, 2011, pp. 2312–2317.

[5] A. Jentzsch, R. Isele, C. Bizer, Silk-generating rdf links while publishing or consuming linked data, in: ISWC, 2010.

[6] D. Symeonidou, V. Armant, N. Pernelle, F. Sais, Sakey: Scalable almost key discovery in RDF data, in: ISWC, 2014, pp. 33–49.

[7] R. Isele, C. Bizer, Active learning of expressive linkage rules using genetic programming, J. Web Sem. 23 (2013) 2–15.

[8] M. Achichi, M. Ben Ellefi, D. Symeonidou, K. Todorov, Automatic key selection for data linking, in: EKAW, Springer, 2016, pp. 3–18.

[9] M. Achichi, P. Lisena, K. Todorov, R. Troncy, J. Delahousse, Doremus: A graph of linked musical works, in: International Semantic Web Conference, Springer, 2018, pp. 3–19.

[30] Note that this manipulation has not been applied to the experiments reported in this paper and therefore does not impact the execution times reported here.

[31] `https://project-hobbit.eu/`

[10] A. Ferrara, D. Lorusso, S. Montanelli, G. Varese, Towards a benchmark for instance matching, in: Ontology Matching-Volume 431, CEUR-WS. org, 2008, pp. 37–48.

[11] R. Navigli, S. P. Ponzetto, Babelnet: Building a very large multilingual semantic network, in: 48th annual meeting of the association for computational linguistics, ACL, 2010, pp. 216–225.

[12] J. Volz, C. Bizer, M. Gaedke, G. Kobilarov, Silk-a link discovery framework for the web of data., LDOW 538.

[13] D. Faria, C. Pesquita, B. S. Balasubramani, C. Martins, J. Cardoso, H. Curado, F. M. Couto, I. F. Cruz, OAEI 2016 results of AML, in: Ontology Matching ISWC, CEUR, Vol. 1766, 2016.

[14] D. Ngo, Z. Bellahsene, K. Todorov, Extended tversky similarity for resolving terminological heterogeneities across ontologies, in: OTM On the Move to Meaningful Internet Systems", Springer, 2013, pp. 711–718.

[15] M. Cheatham, P. Hitzler, String similarity metrics for ontology alignment, in: ISWC 2013, Springer, 2013, pp. 294–309.

[16] Y. Yamamoto, A. Yamaguchi, H. Bono, T. Takagi, Allie: a database and a search service of abbreviations and long forms, Database 2011.

[17] C. Li, L. Ji, J. Yan, Acronym disambiguation using word embedding, in: AAAI, 2015, pp. 4178–4179.

[18] J. Gracia, E. Montiel-Ponsoda, P. Cimiano, A. Gomez-Perez, P. Buitelaar, J. McCrae, Challenges for the multilingual web of data, Web Semantics: Science, Services and Agents on the World Wide Web 11 (2012) 63–71.

[19] A. Simon, R. Wenz, V. Michel, A. D. Mascio, Publishing bibliographic records on the web of data: Opportunities for the bnf (french national library), in: ESWC 2013, 2013, pp. 563–577.

[20] K. D. Bollacker, R. P. Cook, P. Tufts, Freebase: A shared database of structured general human knowledge, in: AAAI, 2007, pp. 1962–1963.

[21] F. Scharffe, Y. Liu, C. Zhou, Rdf-ai: an architecture for rdf datasets matching, fusion and interlink, in: IJCAI 2009 workshop IR-KR, 2009.

[22] T. Lesnikova, J. David, J. Euzenat, Interlinking english and chinese rdf data sets using machine translation, in: ESWC workshop Know@ LOD, Vol. 2013, 2014.

[23] T. Lesnikova, J. David, J. Euzenat, Interlinking english and chinese RDF data using babelnet, in: ACM DocEng, 2015, pp. 39–42.

[24] A. N. Tigrine, Z. Bellahsene, K. Todorov, Light-weight cross-lingual ontology matching with lyam++, in: OTM: On the Move to Meaningful Internet Systems, Springer, 2015, pp. 527–544.

[25] F. Narducci, M. Palmonari, G. Semeraro, Cross-lingual link discovery with TR-ESA, Inf. Sci. 394 (2017) 68–87.

[26] A. Nikolov, V. Uren, E. Motta, Knofuss: A comprehensive architecture for knowledge fusion, in: K-Cap, ACM, 2007, pp. 185–186.

[27] S. Rong, X. Niu, E. W. Xiang, H. Wang, Q. Yang, Y. Yu, A machine learning approach for instance matching based on similarity metrics, in: ISWC, Springer, 2012, pp. 460–475.

[28] C. Shao, L. Hu, J. Li, Z. Wang, T. L. Chung, J. Xia, Rimom-im: A novel iterative framework for instance matching, J. Comput. Sci. Technol. 31 (1) (2016) 185–197.

[29] M. Kejriwal, D. P. Miranker, Semi-supervised instance matching using boosted classifiers, in: ESWC, 2015, pp. 388–402.

[30] D. Symeonidou, N. Pernelle, F. Saïs, KD2R: A key discovery method for semantic reference reconciliation, in: On the Move to Meaningful Internet Systems: OTM 2011 Workshops, 2011, pp. 392–401.

[31] T. Soru, E. Marx, A. N. Ngomo, ROCKER: A refinement operator for key discovery, in: WWW, 2015, pp. 1025–1033.

[32] C. Bizer, R. Cyganiak, Quality-driven information filtering using the wiqa policy framework, Web Semantics: Science, Services and Agents on the World Wide Web 7 (1) (2009) 1–10.

[33] M. B. Ellefi, Z. Bellahsene, J. Breslin, E. Demidova, S. Dietze, J. Szymanski, K. Todorov, Rdf dataset profiling-a survey of features, methods, vocabularies and applications, Semantic Web.

[34] T. Saveta, E. Daskalaki, G. Flouris, I. Fundulaki, M. Herschel, A.-C. Ngonga Ngomo, Pushing the limits of instance matching systems: A semantics-aware benchmark for linked data, in: WWW, ACM, 2015, pp. 105–106.

[35] T. Saveta, E. Daskalaki, G. Flouris, I. Fundulaki, M. Herschel, A. N. Ngomo, LANCE: piercing to the heart of instance matching tools, in: ISWC, 2015, pp. 375–391.

[36] E. Jimenez-Ruiz, B. C. Grau, Logmap: Logic-based and scalable ontology matching, in: ISWC, Springer, 2011, pp. 273–288.

[37] M. Atencia, J. David, J. Euzenat, Data interlinking through robust linkkey extraction., in: ECAI, 2014, pp. 15–20.

[38] T. Soru, E. Marx, A.-C. Ngonga Ngomo, Rocker: a refinement operator for key discovery, in: WWW, ACM, 2015, pp. 1025–1033.

[39] M. Atencia, J. David, F. Scharffe, Keys and pseudo-keys detection for web datasets cleansing and interlinking, in: EKAW, 2012, pp. 144–153.

[40] D. Symeonidou, I. Sanchez, M. Croitoru, P. Neveu, N. Pernelle, F. Sais, A. Roland-Vialaret, P. Buche, A. Muljarto, R. Schneider, Key discovery for numerical data: Application to oenological practices, in: ICCS, 2016, pp. 222–236.

[41] D. Symeonidou, L. Galarraga, N. Pernelle, F. Saïs, F. Suchanek, VICKEY: Mining Conditional Keys on Knowledge Bases, in: ISWC, 2017.

[42] H. Farah, D. Symeonidou, K. Todorov, Keyranker: Automatic rdf key ranking for data linking, in: K-Cap, ACM, 2017, p. 7.

[43] A.-C. N. Ngomo, J. Lehmann, S. Auer, K. Höffner, Raven-active learning of link specifications, in: Ontology Matching, CEUR-WS. org, 2011, pp. 25–36.

[44] A.-C. N. Ngomo, K. Lyko, Eagle: Efficient active learning of link specifications using genetic programming, in: ESWC, Springer, 2012, pp. 149–163.

[45] A. N. Ngomo, K. Lyko, V. Christen, COALA - correlation-aware active learning of link specifications, in: ESWC, 2013, pp. 442–456.

[46] R. Isele, C. Bizer, Learning linkage rules using genetic programming, in: Ontology Matching-Volume 814, CEUR-WS. org, 2011, pp. 13–24.

[47] M. Kejriwal, D. P. Miranker, An unsupervised instance matcher for

schema-free RDF data, J. Web Sem. 35 (2015) 102–123.

[48] M. A. Sherif, A.-C. N. Ngomo, J. Lehmann, Wombat–a generalization approach for automatic link discovery, in: ESWC 2017, Springer, 2017, pp. 103–119.

[49] A. Nikolov, M. d'Aquin, E. Motta, Unsupervised learning of link discovery configuration, in: ESWC, 2012, pp. 119–133.

[50] A. Nikolov, V. S. Uren, E. Motta, A. N. D. Roeck, Integration of semantically annotated data by the knofuss architecture, in: EKAW, 2008, pp. 265–274.

[51] L. Rokach, O. Maimon, Clustering methods, in: The Data Mining and Knowledge Discovery Handbook, 2005, pp. 321–352.

[52] D. Faria, B. S. Balasubramani, V. R. Shivaprabhu, I. Mott, C. Pesquita, F. M. Couto, I. F. Cruz, Results of AML in OAEI 2017, in: OM@ISWC, 2017, pp. 122–128.

[53] A. Khiat, M. Benaissa, M. A. Belfedhal, Strim results for oaei 2015 instance matching evaluation., in: OM, 2015, pp. 208–215.

[54] E. Jiménez-Ruiz, B. C. Grau, V. Cross, Logmap family participation in the OAEI 2017, in: OM@ISWC, 2017, pp. 153–157.

[55] D. Faria, C. Pesquita, E. Santos, M. Palmonari, I. F. Cruz, F. M. Couto, The agreementmakerlight ontology matching system, in: OTM: On the Move to Meaningful Internet Systems, Springer, 2013, pp. 527–541.

[56] A. Khiat, M. Mackeprang, I-match and ontoidea results for OAEI 2017, in: OM@ISWC, 2017, pp. 135–137.

[57] X. Lyu, Q. Zhang, W. Hu, Z. Sun, Y. Qu, njulink: results for instance matching at OAEI 2017, in: OM@ISWC, 2017, pp. 158–165.

International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems

FUZZY ONTOLOGY ALIGNMENT USING BACKGROUND KNOWLEDGE

KONSTANTIN TODOROV
LIRMM, University of Montpellier 2,
121 rue Ada, Montpellier
konstantin.todorov@lirmm.fr

CELINE HUDELOT
Laboratory MAS, École Centrale Paris,
Grande Voie des Vignes, 92290 Châtenay-Malabry, France
celine.hudelot@ecp.fr

ADRIAN POPESCU
CEA, LIST, Vision & Content Engineering Laboratory,
Gif-sur-Yvettes, France,
adrian.popescu@cea.fr

PETER GEIBEL
Institute of Medical Informatics, Charité – Universitätsmedizin Berlin,
Hindenburgdamm 30, 12200 Berlin, Germany
peter.geibel@charite.de

Received (received date)
Revised (revised date)

We propose an ontology alignment framework with two core features: the use of background knowledge and the ability to handle vagueness in the matching process and the resulting concept alignments. The procedure is based on the use of a generic reference vocabulary, which is used for fuzzifying the ontologies to be matched. The choice of this vocabulary is problem-dependent in general, although Wikipedia represents a general-purpose source of knowledge that can be used in many cases, and even allows cross language matchings. In the first step of our approach, each domain concept is represented as a fuzzy set of reference concepts. In the next step, the fuzzified domain concepts are matched to one another, resulting in fuzzy descriptions of the matches of the original concepts. Based on these concept matches, we propose an algorithm that produces a merged fuzzy ontology that captures what is common to the source ontologies. The paper describes experiments in the domain of multimedia by using ontologies containing tagged images, as well as an evaluation of the approach in an information retrieval setting. The undertaken fuzzy approach has been compared to a classical crisp alignment by the help of a ground truth that was created based on human judgment.

1. Introduction

With the growing demand and acceptance of ontology-based applications, we have witnessed the creation of multiple ontologies describing similar or even identical fractions of real world knowledge. These ontologies, complementary or redundant in part, have an impaired collaborative functionality because of the decentralized nature of their creation resulting in mismatches in terms of scopes and application purposes, syntax and terminology. However, often the sharing, integration and interoperability of these resources is desirable in real life application scenarios, such as semantic web applications, semantic information retrieval, biomedical information systems, and geo-spatial applications.

In recent years, a considerable research effort has been directed towards the development of procedures for *ontology matching*, *alignment* and *merging*[14], which aim at attaining a certain degree of reconciliation between various knowledge resources by automatically detecting correspondences between the elements of these resources. Although the field of ontology matching has matured considerably after more than a decade of research and practice, many challenges still stand in front of the scientific community[43]. Among these, this paper contributes particularly to solutions related to the use of background knowledge, as well as the consideration of vagueness and imprecision within the ontologies and the resulting alignments.

A current research issue in aligning real-world ontologies is handling both imprecise and vague information in the matching process. An example we will come back to in the experimental section of this paper is the match between two related concepts, say Atmospheric_Phenomenon and Natural_Hazard. As we will see, a crisp method fails in finding any correspondence between these related but still different concepts, whereas our fuzzy approach relates these entities assigning to their match a degree of 0.1 representing the overlap of the two concepts in a numerical manner. This result is more satisfactory with respect to user expectations and common sense than the result obtained by the crisp match since it is better able to capture a more vague idea of relatedness between the two concepts. In addition to vagues resulting from the matching process, many concept definitions are inherently vague, which needs to be dealt with in the alignment process[19].

The different ontology alignment approaches are often complemented by the use of background knowledge, represented by a *reference ontology*. This is well adapted to various real-world matching problems (e.g., dealing with weakly structured models)[2,3,36,38]. Following this paradigm, we suggest a procedure for matching the concepts of two (or more) domain ontologies, referred to as source ontologies, by integrating fuzziness in their concepts representation and, consequently, in the match itself. In its first phase, the proposed procedure consists in *anchoring* the source concepts onto a general purpose reference ontology[2,36]. In our experiments, we have used Wikipedia as such a reference knowledge resource. We argue that it is a suitable choice in many contexts and application scenarios because of its availability, size and generality. The anchoring process is used in order to redefine the

source concepts as fuzzy sets of the reference concepts, which enables the application of a number of similarity measures defined on fuzzy sets. Using a fuzzy set representation accounts for the inherent vagueness in the definition of concepts. In consequence, vagueness and imprecision in concept matching are addressed as well since the similarity function[a] for two source concepts takes as input two fuzzy membership functions. Finally, the resulting match of two concepts is also described as a fuzzy set of the reference concepts, which allows incremental ontology alignment.

In order to demonstrate the usefulness of our proposal, we provide experimental results in the domains of multimedia and text. Aligning text-populated ontologies is related to the important problem of web-directory mapping that is relevant for enhancing information retrieval. Regarding multimedia resources, several important applications of ontology alignment can be found, such as narrowing the well known semantic gap[44] and improving image retrieval by query reformulation and expansion[52]. Since the chosen application fields comprise ontologies which come equipped with instance sets (e.g., annotated images or text), it is natural to apply an instance-based concept similarity measure within the fuzzification phase. Note however, that our approach can also be applied to ontologies without instances by using other notions of concept similarity (e.g., based on concept labels and the ontology structure).

The article is structured as follows. We discuss related approaches from the fields of classical and fuzzy ontology alignment in Section 2. Basic definitions from the field of fuzzy sets and logics that are important for our approach are given in Section 3. We define crisp ontologies and propose a crisp ontology alignment framework in Section 4. In Section 5, we describe our fuzzy alignment method. Section 6 provides an evaluation of the proposed techniques on data from two different domains - text and multimedia (videos and photographs). We analyze empirically the impact of the size and kind of reference ontology on the precision of the alignments and we describe a use-case study in the information retrieval domain. The conclusions are found in Section 7.

2. Related Work

The current section refers to related interdisciplinary work from the fields of classical and fuzzy ontology alignment.

2.1. *Aligning Heterogeneous Ontologies*

An ontology describes that what exists in a given domain of interest. It consists of a set of *concepts* and *relations* defined on these concepts, which provide explicit and formal knowledge about this domain[22,48]. The most important of these relations is the subclass/superclass relation, also known as *is_a* relation in the field of Artificial Intelligence.

[a]There are many possible choices discussed in Section 5.2.2

Ontology heterogeneity occurs when two or more ontologies are created independently from one another over similar domains. It may be observed on *linguistic or terminological* level (concepts which represent the same real world entity but have different names), on *conceptual* level (mismatches in level of detail, coverage, scope and structure) or on *extensional* level (differences in the population). Whenever heterogeneity of any of these kinds is observed over a set of ontologies, these ontologies will be referred to as *heterogeneous*.

Ontology matching is understood as the process of establishing relations between the elements of two heterogeneous ontologies which results in an alignment between these ontologies. Different alignment techniques have been introduced in the past years in order to resolve different types of heterogeneity relying on methods coming from fields as various as machine learning, linguistics, graph theory, relational algebra and logics[19].

Terminological methods comprise two major groups of approaches: those that use strings in order to match names of entities, and those that rely on linguistic information contained in dictionaries and thesauri combined with techniques from natural language processing in order to compare the similarity of terms and their relations and overcome problems evolving from synonymy and polysemy.

In addition to terminological information, the *structure* of an ontology can be taken into account in order to define a similarity measure. This can be done on two levels with respect to either a single ontology element, known as internal structure, or to the way in which a set of elements are related, known as relational structure[14]. Methods based on the former structure type are based on similarities of the sets of properties of two elements, the datatypes used to describe them or their properties, the cardinalities that sets of values of two properties are allowed to reach, etc. These approaches are suited to schema matching problems where one disposes readily with an internal structure of the database entities. Relational structure approaches are concerned with comparing groups of elements together with the relations that hold between them. Structure and terminology are often used in a combined manner as this has been done by Noy *et al.*[36], Madhavan *et al.*[29] or McGuiness[30]. In the (crisp) ontology matching framework that we describe in Section 4, the matching algorithm relies on results that have been long known in graph theory. In the field of ontology matching, this algorithm relates closely to similarity propagation approaches, such as the well-known similarity flooding algorithm[31] or its improved version introduced in the YAM++ system[34].

Instance-based, or *extensional ontology alignment* is based on the idea that ontology concepts can be represented as sets of instances. The similarity measured on these sets of instances reflects the semantic similarity between the concepts that the sets of instances populate. In particular, these instance sets together reflect structural and terminological differences and similarities[25]. Certain approaches assume that two ontologies use the same sets of instances, and – when this is not the case – that mechanisms for extracting common instances (from corpora or other external sources) are available (FCA-MERGE[48]). Other techniques rely on estimating the

concept similarity by measuring distances of class centers (CAIMAN[26]) or estimating joint probabilities by the help of machine learning techniques (GLUE[12]). In previous work[51], we proposed an approach which does not rely on intersections of instance sets, nor on the estimation of joint probabilities, but on the selection of descriptive variables for each concept. An application of the approach in the multimedia domain is given in an article by Todorov *et al.*[52].

Since the fuzzy approach described in this paper is based on using similarities, it can incorporate terminological, structural and instance-based information. In what follows, however, we focus on an instance-based method combined with structure, assuring that the alignment is evidence-driven. Depending on the nature of the reference ontology, no common set of instances is needed. For instance, if the reference ontology characterizes each concept with both texts *and* images, it is in principle possible to match image and text ontologies with our approach. This is so because the concepts from the "image" ontology and the concepts from the "texts" ontology will be fuzzified independently from one another by the help of the reference vocabulary.

2.2. *Dealing with Imprecision and Uncertainty*

As discussed in the introduction, handling vague and imprecise knowledge, as well as uncertainty in ontology construction and alignment are important real life issues, which have been addressed in the literature recently. This section discusses several of these endeavors. Mind that there is an important distinction to be made between imprecise and uncertain knowledge – often confused, but well explained in a paper by Dubois and Prade[13].

2.2.1. *Fuzzy Ontologies*

The theory of fuzzy sets and logics provides a suitable framework for handling vagueness and imprecision in ontologies. A general definition of a fuzzy ontology is given as one *which uses fuzzy logics to provide a natural representation of imprecise and vague knowledge, and eases reasoning over it*[5]. Papers by Sanchez, Calegari and colleagues[7,8,42] form an important body of work in this field. These authors have been motivated by the observation that crisp reasoning through two-valued logics is not suited to deal with imprecise and vague information available in a real world context. They define every ontology concept as a fuzzy set on the domain of instances. Relations on concepts are defined as fuzzy mappings. Particularly, subsumption is handled by a fuzzy taxonomic relation that expresses the fact that a concept is a specification of another concept up to a certain degree between 0 and 1. In the definition of a fuzzy ontology, which is given later on, we will follow a similar approach.

2.2.2. *Fuzzy Ontology Alignment*

Although there exists a solid body of work on fuzzy ontologies on one hand and on ontology alignment on the other hand, only few authors have addressed *fuzzy ontology alignment*. Work in this field can be classified into two families: (1) approaches extending crisp ontology alignment to deal with fuzzy ontologies and (2) approaches addressing imprecision of the matching of (crisp or fuzzy) concepts.

Based on the work on approximate concept mapping by Stuckenschmidt[47] and Akahani *et al.*[1], Xu *et al.*[54] suggested a framework for the mapping of fuzzy concepts between fuzzy ontologies. Their approach is based on finding the best approximations in an ontology for all the concepts in another ontology. The approximations (least upper approximation and greatest lower approximation) are defined by using fuzzy concept subsumption and an iterative algorithm is proposed to find a simplified least upper bound. With a similar objective, Bahri *et al.*[4] proposed a framework to define similarity relations (*More General, LessGeneral, Equivalent, Disjoint, Overlap*) among fuzzy ontology components based on their intentional definitions (i.e. a set of description logics formulas that represent the meaning of a component).

The second family of fuzzy alignment approaches is characterized by the representation of imprecision of the alignment itself, even with crisp ontologies. For instance, Ferrara *et al.*[17] propose a fuzzy approach which handles mapping imprecision and provides criteria for its validation. The principle of this approach is to interpret and translate each crisp matching result as a set of fuzzy assertions and perform fuzzy reasoning over this set. An ontology mapping approach based on fuzzy conceptual graphs and rules is proposed by Buche *et al.*[6].

To define new intra-ontology concept similarity measures, Cross *et al.*[10] model a concept as a fuzzy set of its ancestor concepts and itself. As a membership degree function, the authors use the information content (IC) of concept with respect to its ontology. IC can be measured by using external corpora (more occurrences of the concept suggest less informativeness) or by using the ontology structure – the number of ancestors and the depth of the concept in its ontology. Cross *et al.* suggest a number of intra-ontology concept similarity measures based on these fuzzy set representations. Our approach relates to this one in that we also model concepts as fuzzy sets of other concepts (in our case those of a reference ontology), and differs to it in the fuzzification method and in the fact that we operate on cross-ontology concepts.

The alignment framework that we propose, although in line with the fuzzy-based approaches, does not directly fall into either of the two families outlined above. Our model suggests that a crisp ontology can be fuzzified by the mediation of a reference knowledge body. To our knowledge, none of the existing works on fuzzy alignment is based on the use of background knowledge, which is among the principal motivations of our approach. Of course, in many cases two ontologies can be aligned directly, without taking into consideration any information related to

background knowledge. The advantage of using a reference vocabulary is that it allows for taking into account different aspects of the semantics of a concept. In contrast to many other approaches, we are able to make use of this background information, when it is given. In addition, such background knowledge is present in many situations (in our approach we use Wikipedia) and being able to exploit it in an efficient manner is advantageous. Among the original contributions of this method is the fact that we are able to apply a fuzzy framework to the specific case of instance-populated ontologies (annotating images or text). We explore the benefits of such a fuzzification in order to measure cross-ontology concept similarities as relatedness between fuzzy sets.

2.2.3. *Non-fuzzy Approaches*

Non-fuzzy approaches to handle uncertainty in the mapping process exist as well. A first idea by Gal[18] was to use the top-k matchings rather than a single best matching. Gligorov *et al.*[21] define the notion of an approximate matching by using the decomposition of a matching into a number of sub-matchings and by allowing a few of these to be unsatisfiable.

According to the underlying theory used to handle uncertainty, the rest of the non-fuzzy approaches can be classified as either based on Bayesian probability theory or on Dempster-Shafer theory of evidence[9,33,35,37,53]. Since in our approach concepts are described on the basis of similarities to reference concepts (this way defining a fuzzy membership function on concepts), there exists also a relationship to the area of topic modeling[24].

3. Background

The current section introduces notation and definitions from the field of fuzzy set theory that are necessary for the introduction of our alignment framework.

3.1. *Fuzzy Sets and Logics*

Fuzzy set theory emerged as a generalization of the classical theory of sets[55]. A fuzzy set $\mathcal{A}$ is defined on a given domain of objects X by the function

$$\mu_{\mathcal{A}} : X \longmapsto [0, 1]$$

which expresses the degree of membership of every element of X to $\mathcal{A}$ by assigning to each $x \in X$ a value from the interval $[0, 1]$. Analogously, fuzzy logics extends two-valued logics by assigning to a proposition a truth value between 0 and 1.

All crisp set and logical operations can be extended to fuzzy sets and logics. Intersection and union are defined, respectively, based on a so-called t-norm function and a t-conorm function. Crisp logical implication is extended to fuzzy logics by the help of a fuzzy implication function. We give definitions by providing examples in terms of Gödel, Łukasiewicz and product semantics, following the introduction

given by Straccia[46]. For the sake of representation within this section, we will denote $a = \mu_{\mathcal{A}}(x)$ and $b = \mu_{\mathcal{B}}(x)$. The *intersection* of two fuzzy sets $\mathcal{A}$ and $\mathcal{B}$ is given by a function $T(a, b)$, referred to as a t-norm. The Gödel t-norm is defined by $T_G(a, b) = min(a, b)$, the Łukasiewicz t-norm is given as $T_L(a, b) = max(a+b-1, 0)$ and the product t-norm – by $T_P(a, b) = a \times b$.

The *union* of two fuzzy sets $\mathcal{A}$ and $\mathcal{B}$ is given by $S(a, b)$, where S is a t-conorm. The Gödel definition is given by $S_G(a, b) = max(a, b)$, the Łukasiewicz definition is given by $S_L(a, b) = min(a+b, 1)$ and the product t-conorm is defined by $S_P(a, b) = a + b - a \times b$.

Fuzzy *implication* $\mathcal{A} \to \mathcal{B}$ is defined by by $\mu_{\mathcal{A}\to\mathcal{B}}(x) = i(\mu_{\mathcal{A}}(x), \mu_{\mathcal{B}}(x))$ where i is a function that determines the properties of the implication. Two types of fuzzy implications are commonly used: *S-implications* which extend the proposition $a \to b = \neg a \vee b$ to fuzzy logics and *R-implications (residuum-based implications)* defined as $\forall a, b \in [0, 1]$, $i(a, b) = sup\{c \in [0, 1] : T(a, c) \leq b\}$. In terms of the three considered semantics $i(a, b) = 1$ if $a \leq b$. Depending on the particular t-norm definition, the case $a > b$ is defined as follows: $i_G(a, b) = b$ (Gödel), $i_L(a, b) = 1-a+b$ (Łukasiewicz) and $i_P(a, b) = \frac{b}{a}$ (product).

In this study, we have considered the Gödel definitions of intersection, union and implication. As we shall see, this choice is justified by the properties of the Gödel implication which are used to define a fuzzy degree of subsumption preserving the crisp one. This implication is defined for two fuzzy membership functions as

$$\mu_{\mathcal{A}\to\mathcal{B}}(x) = \begin{cases} 1, & \text{if } \mu_{\mathcal{A}}(x) \leq \mu_{\mathcal{B}}(x), \\ \mu_{\mathcal{B}}(x), & \text{otherwise.} \end{cases} \tag{1}$$

Taking the infimum $\inf_{x\in X} \mu_{\mathcal{A}\to\mathcal{B}}(x)$ over all $x \in X$ in Equation (1) gives rise to the definition of a fuzzy subsumption between $\mathcal{A}$ and $\mathcal{B}$. This is used to define a fuzzy version of the ontological is_a relation in Section 5.1.

The fuzzy power set of X, denoted by $\mathcal{F}(X, [0, 1])$, is the set of all membership functions defined on X.

3.2. *Measures of Fuzzy Set Relatedness*

Let $\mathcal{A}$ and $\mathcal{B}$ be two fuzzy sets with respective membership functions $\mu_{\mathcal{A}}$ and $\mu_{\mathcal{B}}$. We consider the following measures of fuzzy set relatedness, well-known from the fuzzy ontology literature[10]. These measures are relevant to our approach and will be applied as explained later on in Section 5.2.2 for measuring the similarity between fuzzified concepts.

- Base measure:

$$\rho_{base}(\mu_{\mathcal{A}}, \mu_{\mathcal{B}}) = 1 - \max_{x\in X} |\mu_{\mathcal{A}}(x) - \mu_{\mathcal{B}}(x)|. \tag{2}$$

- Euclidean Distance ($\|x\|_2 = \left(\sum_{x\in X} |x|^2\right)^{1/2}$ is the ℓ^2-norm):

$$\rho_{diff}(\mu_{\mathcal{A}}, \mu_{\mathcal{B}}) = 1 - \|\mu_{\mathcal{A}} - \mu_{\mathcal{B}}\|_2 . \tag{3}$$

- Zadeh's partial matching index:

$$\rho_{sup\text{-}min}(\mu_{\mathcal{A}}, \mu_{\mathcal{B}}) = \sup_{x \in X} T(\mu_{\mathcal{A}}(x), \mu_{\mathcal{B}}(x)). \tag{4}$$

  In our experiments, we used a more robust variant of $\rho_{sup\text{-}min}$ that applies the average of the k largest values of $T(\mu_{\mathcal{A}}(x), \mu_{\mathcal{B}}(x))$. This measure will be called $\rho_{sup\text{-}min(k)}$.
- Jaccard coefficient:

$$\rho_{jacc}(\mu_{\mathcal{A}}, \mu_{\mathcal{B}}) = \frac{\sum_x T(\mu_{\mathcal{A}}(x), \mu_{\mathcal{B}}(x))}{\sum_x S(\mu_{\mathcal{A}}(x), \mu_{\mathcal{B}}(x))}\,. \tag{5}$$

4. A Crisp Ontology Alignment Framework

The current section proposes a generic crisp ontology alignment setting, which will be extended to a fuzzy formulation in the following sections. The introduction of the section contains standard definitions of an ontology, well known from the ontology matching literature. Further on, we provide an ontology matching algorithm. This algorithm is based on the use of a concept similarity measure and the structural properties of the ontologies. Although, to the best of our knowledge, this algorithm has not been implemented in any existing ontology matching system, the matching framework is not novel, because it relies on well-known results from the ontology matching field and graph theory. We consider an important contribution of this section the fact that we go into details regarding computation and complexity issues from a theoretical stance which gives insights on the cost and efficiency of the algorithm.

We will consider the following definition of an ontology, similar to the definition given by Stumme and Maedche[48,51].

Definition 1. (Crisp Ontology) Let

- $\mathbf{C}$ be a set of concepts[b],
- $is\_a \subseteq \mathbf{C} \times \mathbf{C}$ be a partial order on $\mathbf{C}$,
- $R = \{r_1, \ldots, r_P\}$ be a set of P relations on $\mathbf{C}$. For simplicity, we only consider binary relations.

The tuple $O = (\mathbf{C}, is\_a, R)$ is called a crisp ontology.

The partial order is_a represents the hierarchical backbone of the ontology and is an essential part of its semantics. In R, we can have additional semantic concept-level relations like meronymy, antonymy, similarity, or meta-level concept relations.

Ontologies can be used for describing data by assigning instances from a given set I to concepts (e.g., documents to a web catalogue). We give the following definition of an instance function.

[b]Mind that the use of the term "concept" is a shortcut for a "concept name" or a "concept representation" that relates to an actual concept.

Definition 2. (Instance Function) Let

- $O = (\mathbf{C}, is\_a, R)$ be an ontology,
- I be a set of instances,
- and $g : \mathbf{C} \longmapsto 2^I$ an instance function that assigns a set of instances to each concept in $\mathbf{C}$.

If

$$(A, A') \in is\_a \longleftrightarrow g(A) \subseteq g(A') \tag{6}$$

holds for all $A, A' \in \mathbf{C}$, g is called an instance function of O.

An ontology equipped with an instance function over a set of data instances is called a populated ontology. In terms of a formal semantics for ontologies, g could be seen as an interpretation for concepts if each concept is assigned the set of *all* possible instances. In practice, however, g usually assigns to each concept only a finite sample from the set of possible instances. As a consequence all computed quantities are estimates (see, e.g., Section 6).

4.1. *Aligning Crisp Ontologies*

Aligning ontologies requires a similarity measure that determines the similarity of two concepts from different ontologies. Depending on the application field and the kind of ontologies that we are working with, the concept similarity measure can be based on different characteristics, such as linguistic, structural, instance-based or other[14], as described in Section 2.1. This similarity measure may as well be a (linear) combination of a set of similarity measures. For the moment, we consider an abstract concept similarity measure given in the following definition[c].

Definition 3. (Admissible Concept Similarity Measure) Let $\mathbf{C}_1$ and $\mathbf{C}_2$ be two sets of concepts from two ontologies O_1 and O_2, respectively. We define a concept similarity measure as a mapping

$$\sigma : \mathbf{C}_1 \times \mathbf{C}_2 \cup \mathbf{C}_2 \times \mathbf{C}_1 \longmapsto [0, 1]\,.$$

We require that σ is symmetric, i.e., that $\sigma(A, B) = \sigma(B, A)$ holds for all $(A, B), (B, A) \in \mathbf{C}_1 \times \mathbf{C}_2 \cup \mathbf{C}_2 \times \mathbf{C}_1$.

The requirement that σ lies in the interval $[0, 1]$ is important for the matching algorithm that we introduce in the sequel. If a similarity function does not fulfill this requirement, it might be possible to transform it using functions like tangens hyperbolicus, which maps the range of non-negative numbers to the interval $[0, 1]$. The symmetry property is necessary at the fuzzification step in order to end up with comparable fuzzy sets. As we shall see, in the fuzzification process, direct similarities

[c] A specific similarity measure, which is based on the dot product and is suitable for text instances, can be found in Section 6.1.

between concepts from two source ontologies are replaced by similarities between these concepts and a reference concept. Note that similarity measures might also fulfill other axioms like minimality, upon which, however, our approach does not rely.

The ontology matching framework that we develop below applies the concept similarity measure for two ontologies by additionally taking into account their structures. Based on σ, the match of two ontologies[d] is thus computed by a structural procedure which takes into account the relations between intra-ontology concepts and guarantees a one-to-one matching. For two ontologies $O_1 = (\mathbf{C_1}, is\_a_1, R_1)$ and $O_2 = (\mathbf{C_2}, is\_a_2, R_2)$, we give the following definitions.

A match between two concepts A and B will be denoted as the pair (A, B). In the following, we consider 1:1 ontology alignments resulting in structures that can be considered ontologies themselves. Given two matches (A, B) and (A', B'), we would therefore like to ensure that these matches are (1) non-conflicting and additionally (2) respect the relations present in the given ontologies. Therefore, we introduce the notion of *compatibility*. For simplicity and without loss of generality, we assume that comparable relations in the ontologies are denoted by using the same index (e.g., $r_1^1 \in R_1$ is comparable to $r_1^2 \in R_2$, etc.), potentially adding empty relations if there is no counterpart in one ontology. A **comparable pair of relations** would then be $r_i^1 \in R_1$ and $r_i^2 \in R_2$ with $1 \leq i \leq P$, where $P = P_1 = P_2$ is the common cardinality of both the relations set. We define $r_0^1 = is\_a_1$ and $r_0^2 = is\_a_2$ in order to simplify notation.

Definition 4. (Compatibility of Cross-Ontology Concept Pairs) Given two ontologies, O_1 and O_2, let $A, A' \in \mathbf{C_1}$ and $B, B' \in \mathbf{C_2}$. Two pairs of concepts (A, B) and (A', B') are defined as incompatible if one of the following conditions is true:

- $A = A'$ *or* $B = B'$
- $(A, A') \in r_i^1 \wedge (B, B') \notin r_i^2$ for some $0 \leq i \leq P$
- $(A, A') \notin r_i^1 \wedge (B, B') \in r_i^2$ for some $0 \leq i \leq P$
- $(A', A) \in r_i^1 \wedge (B', B) \notin r_i^2$ for some $0 \leq i \leq P$
- $(A', A) \notin r_i^1 \wedge (B', B) \in r_i^2$ for some $0 \leq i \leq P$

Otherwise, the pairs (A, B) and (A', B') are called compatible.

An admissible ontology alignment is defined as a set of compatible concept pairs. Additionally, we introduce a mapping filter threshold θ (to be set by the user) for the minimum concept similarity that is allowable.

Definition 5. (Admissible θ-alignment) An admissible θ-alignment M is a set of cross-ontology concept pairs, which are pairwise compatible. For every concept pair (A, B) it holds that $\sigma(A, B) \geq \theta$, where θ is a similarity threshold between 0 and 1.

[d]It is mainly for the sake of simplicity that we have chosen a two-ontology presentation. The method can be applied straightforwardly for multiple ontologies.

If we consider an admissible θ-alignment $M = \{(A_1, B_1), \ldots, (A_m, B_m)\}$ then the nodes $A_1, \ldots, A_m$ define an induced substructure of O_1 that is isomorphic to the induced substructure $B_1, \ldots, B_m$. We can consider one of the substructures as the match of O_1 and O_2. Mind that we use the usual notions from graph theory, where an induced subgraph contains all the relations between the respective nodes[11]. An isomorphism is defined as a relation-preserving bijective mapping between two graphs.

In general, we do not want to find any alignment but one that is optimal with respect to the concept similarity measure.

Definition 6. (Optimal θ-alignment) Let Σ_θ be the set of all admissible θ-alignments for two ontologies O_1 and O_2. For every alignment $M \in \Sigma_\theta$, we define the criterion

$$crit(M) = \sum_{(A,B)\in M} \sigma(A, B)\,.$$

An optimal θ-alignment M^*_θ is defined as one that maximizes *crit(M)*.

Note that, in the case of strictly positive similarities, optimality implies that the set of concept alignments cannot be extended by additional pairs. It is also easy to show that $crit(M^*_\theta) \leq crit(M^*_{\theta'})$ holds for thresholds $\theta \geq \theta'$ since $\Sigma_{\theta'} \subseteq \Sigma_\theta$.

4.2. *Computing the Common Substructure of Two Ontologies*

We now have the following situation. The similarity measure σ represent the similarity of two cross ontology concepts. It may depend on terminological, structural or instance-based information or it can be a combination of several measures. The user then chooses a threshold θ that defines the minimum similarity he or she is willing to accept for a match. The match is represented by an optimal alignment M^*_θ that maximizes the sum of the node similarities. Each pair of concept pairs, (A, B) and (A', B') in M^*_θ, is required to be compatible. From the condition of optimality, it follows that M^*_θ cannot be extended by additional pairs (A'', B'') if we require strictly positive values of σ, which is the case if we choose $\theta > 0$ or if σ is defined appropriately. In the following discussion, we assume that either is the case. How do we determine an optimal alignment M^*_θ?

Firstly, we note that the base set Σ_θ of all pairs with a similarity larger than θ forms *a graph*, whose nodes correspond to cross-ontology concept pairs, which are labeled with positive numbers. We introduce an edge between two nodes (A, B) and (A', B'), if the nodes are compatible, i.e., (A, A') and (B, B') have the same relations and (A, B) and (A', B') do not share a node.

In graph theory, a clique is defined as a completely connected subgraph. Using this notion, every admissible alignment corresponds to a clique, and an optimal alignment M^*_θ corresponds to a so-called maximal clique since it cannot be extended by further nodes[16]. A clique with the maximum number of nodes is called maximum clique. Note that maximum entails maximal but not vice versa.

Our matching problem is a generalization of the so-called maximum clique problem: we have to find a maximum *weighted* clique since we are not interested in the number of nodes but their summed weight. The maximum weighted clique as a generalization of the maximum clique problem is known to be NP-complete. It was shown by Feige *et al.*[16] that for the maximum clique problem is even difficult to find approximate solutions.

The described connection to the theory of cliques hence implies that finding an optimal alignment M_θ^* is an NP-complete problem. Also, the solution might be non unique, which can be seen from simple examples. In our setting, the case of non-uniqueness (or close solutions) can only be dealt with by interaction with the user. However, it would be possible to retain alternative matches in a non-injective matching approach. In real world cases, though, we can assume that M_θ^* is frequently unique since *crit(*M_θ^**)* is a sum of potentially noisy numbers. In order to deal with the problem of NP-completeness, we propose a method which is well-known from artificial intelligence: hill-climbing search that falls into the class of greedy algorithms[28,41]. This approach was the basis for Algorithm 1 and it worked fine in our experiments although we cannot give performance guarantees. It is possible to use more elaborate algorithms[56,20].

Additional arguments to the algorithm are the concept similarity measure σ and a threshold θ that determines the minimum similarity. Note that σ may depend on the instances (as this is the case in our experiments) and in that case the algorithm needs to take as an input the two respective sets of instances of the ontologies, as well as their corresponding instance functions (i.e. I_1, I_2, g_1, g_2). The algorithm first determines the best-matching pair (A, B) according to the similarity measure σ. Whenever there is a tie, the algorithm might choose one pair at random. It then restricts the set of remaining potential matches by eliminating all incompatible pairs (A', B') from the set of remaining potential matches M, which initially takes the value $\mathbf{C_1} \times \mathbf{C_2}$.

Example: Let us consider two simple ontologies

$$O_1 = (\{A_1, A_2, A_3\}, \{(A_3, A_2), (A_2, A_1), (A_3, A_1)\}, \{\underbrace{\{\}}_{r_1^1}\}, I_1, g_1)$$

and

$$O_2 = (\{B_1, B_2, B_3\}, \{(B_2, B_1)\}, \{\underbrace{\{(B_2, B_3)\}}_{r_1^2}, I_2, g_2\})\,.$$

In the function "Match", the function"Map" is called with M being defined as the cross product of the concept sets, i.e.

$$M = \{\underbrace{(A_1, B_1)}_{\sigma(A_1,B_1)=0.1}, \underbrace{(A_1, B_2)}_{0.3}, \underbrace{(A_1, B_3)}_{0.4}, \underbrace{(A_2, B_1)}_{0.5}, \underbrace{(A_2, B_2)}_{0.8}, \underbrace{(A_2, B_3)}_{0.5},$$
$$\underbrace{(A_3, B_1)}_{0.1}, \underbrace{(A_3, B_2)}_{0.2}, \underbrace{(A_3, B_3)}_{0.7}\}$$

procedure Match(O_1, O_2, σ, θ)
// The function returns a non-extendable alignment of O_1 and O_2
begin
  return Map($\mathbf{C_1} \times \mathbf{C_2}$, O_1, O_2, σ, θ)

procedure Map($M, O_1, O_2, \sigma, \theta$)
// M is the set of concept pairs that are still possible
// σ is the concept similarity measure, θ is the similarity threshold
begin
  1. **If** $M = \emptyset$ **return** $\emptyset$.
  2. Find $(A, B) \in M$ that maximizes $\sigma(A, B)$. In case of a tie, chose one arbitrarily.
  3. **If** $\sigma(A, B) < \theta$ **then return** $\emptyset$
  4. **Compute** M' **by removing pairs** $(A', B') \in M$ **that are incompatible to** (A, B) **from** M **(see Definition 4)**
  5. **return** $\{(A, B)\} \cup Map(M', O_1, O_2, \sigma, \theta)$

Algorithm 1: A greedy algorithm for matching ontologies O_1 and O_2

The value of σ is given below the respective pair. We assume θ was set to 0.05. Since M is not empty, the algorithm picks the concept pair with the highest similarity, which is (A_2, B_2). Now the algorithm removes all incompatible pairs from M: (A_1, B_2) is incompatible since B_2 occurs in both concept pairs. (A_1, B_3) is incompatible, since there is a *is_a* relationship from A_2 to A_1, but not from B_2 to B_3. Since $(A_2, A_3) \notin r_1^1$ but $(B_2, B_3) \in r_1^2$, (A_3, B_3) is also discarded although it has the highest σ-value among the remaining pairs.

The reduced set of pairs is now

$$M' = \{\underbrace{(A_1, B_1)}_{0.1}\}$$

which becomes the value of M in the second call of "Map". Since the similarity of (A_1, B_1) is larger than θ, it is picked as the next concept pair. M' becomes empty, i.e., $\emptyset$ is returned as a result of the next call to "Map". The return values of the recursive calls to "Map" add up to

$$\{(A_2, B_2), (A_1, B_1)\}$$

which is returned by "Match". This ends the example.

The algorithm returns a set of relation-preserving 1:1 matches, which can be turned into an ontology by adding the relations that are found in O_1 (or, equivalently, those that are found in O_2). The result corresponds to a maximal θ-alignment, but not necessarily to an optimal one.

Proposition 1. *Match(O_1, O_2, σ, θ) returns a maximal θ-alignment of O_1 and O_2.*

A proof of the proposition is given in Appendix A.

5. Using a Reference Ontology for Fuzzy Ontology Alignment

The disadvantage of the crisp matching approach described in the last section lies in the fact that it only partly accounts for the imprecision of the matching result. In the matched ontology, each concept corresponds to a pair of source concepts to which a similarity value is attached by σ. However, there is no means to see what aspects of the original concepts is retained in their match. The idea that we therefore propose is to first map the source concepts into a *semantic space* that is defined by the concepts of a *reference ontology*. The use of reference concepts allows to determine what semantic information is retained in the match of a pair of source concepts. Like the source concepts, their match can be interpreted in terms of the semantic space of the reference ontology, which is shown at the end of this section.

The mapping into the semantic space of the reference ontology is modeled as a *fuzzification process* operating on the concepts of the source ontologies. The fuzzification is based on similarity scores between each source concept and all reference concepts. The obtained scores are used to construct fuzzy membership functions which represent the source concepts as fuzzy sets of reference concepts. In a second step, these membership functions are used to compute fuzzy similarities between the fuzzified source concepts. The crisp algorithm (Algorithm 1) is then applied using the fuzzy similarities for the fuzzified concepts. It outputs an ontology alignment whose concept matches are again conceived as fuzzy concepts, i.e. expressed in terms of the reference ontology.

It can be required that the reference ontology, defined by O_{ref}, has the same semantic expressiveness as the source ontologies, in the sense of number and kinds of relations and structure in general. In practice, however, we will allow that some of the defining elements of this ontology are empty, like for instance the set of relations or even the partial order. In many practical situations a set of correctly selected and weakly related terms can fulfill the role of a mediator in the concept fuzzification process. For that reason, we will refer to the reference ontology equivalently as a reference vocabulary. We denote the set of concepts of the reference ontology as X with reference concepts $x \in X$.

The particular choice of a reference vocabulary depends on the domain of application: this can be a broader domain ontology (for instance the FMA ontology in the domain of medicine or biology[39]) or a more generic knowledge source (such as populated WordNet, Wikipedia, or other). It is difficult to provide strict theoretical criteria to motivate this choice. In the experimental part of the paper, we have worked with multimedia ontologies which deal with common everyday life topics in news broadcasting, therefore the choice of Wikipedia as a reference vocabulary is justified.

Given that the reference vocabulary can be quite large, its size can be optimized by indicating what is the number and type of reference concepts that are sufficient in order to perform the fuzzification and alignment correctly. The selection of the most appropriate reference concepts can be based on various statistical heuristics.

We provide an empirical analysis of this problem in Section 6.

5.1. *Similarity-Based Concept and Ontology Fuzzification*

The fuzzification process, which we are going to describe, results in a fuzzy ontology. The general notion of a fuzzy ontology is captured by the following definition.

Definition 7. (Fuzzy Ontology) Let

- $\mathcal{C}$ be a set of (fuzzy) concepts,
- $\mathbf{is\_a} : \mathcal{C} \times \mathcal{C} \longmapsto [0,1]$ be a fuzzy *is_a* relationship,
- $\mathcal{R}$ a set of (binary) fuzzy relations on $\mathcal{C}$, i.e., $\mathcal{R}$ contains relations $r : \mathcal{C}^2 \longmapsto \{0,1\}$.

The tuple $\mathcal{O} = (\mathcal{C}, \mathbf{is\_a}, \mathcal{R})$ constitutes a fuzzy ontology.

Analogically to the definition of a crisp ontology, the instance function is extended for fuzzy ontologies in the following manner.

Definition 8. (Fuzzy Instance Function) Let

- $\mathcal{O} = (\mathcal{C}, \mathbf{is\_a}, \mathcal{R})$ be a fuzzy ontology,
- $\mathcal{I}$ a set of instances,
- and $\mu : \mathcal{C} \longmapsto \mathcal{F}(\mathcal{I}, [0,1])$ a function that assigns each concept in $\mathcal{C}$ a fuzzy membership function on $\mathcal{I}$.

If

$$\mathbf{is\_a}(\mathcal{A}, \mathcal{A}') = \inf_{i \in \mathcal{I}} \mu_{\mathcal{A} \to \mathcal{A}'}(i) \tag{7}$$

holds for all $\mathcal{A}, \mathcal{A}' \in \mathcal{C}$, the function μ is a fuzzy instance function for $\mathcal{O}$.

We will now explain how to define the membership functions for the concepts in $\mathcal{C}$ and the fuzzy **is_a**-relationship that result from fuzzifying a source ontology $O = (\mathbf{C}, is\_a, R)$ with respect to the given reference ontology.

For every concept $A \in \mathbf{C}$, we include a fuzzified counterpart $\mathcal{A} \in \mathcal{C}$ into the fuzzified ontology. For this fuzzy concept its membership function is denoted as $\mu_{\mathcal{A}}$. We now define its membership function via the similarity function σ as

$$\mu_{\mathcal{A}}(x) =_{def} \sigma(A, x) \ , \forall x \in X. \tag{8}$$

We recall that X stands for the set of concepts of the reference vocabulary. I.e., X plays the role of $\mathcal{I}$ from Definition 8. Equation (8) means that the fuzzy counterpart $\mathcal{A}$ of the concept A is defined as a fuzzy set on the set of reference concepts, with the membership values given as cross-ontology similarity values. Note that a fuzzified source concept is defined as a fuzzy set of reference concepts rather than instances. This amounts to saying that it can be seen as a fuzzy union of the reference concepts. Mind that the fuzzification depends on σ. Several examples of fuzzified concepts are given in Section 6.2.

Function Fuzzify(O, O_{ref}, σ)
Input: (A) A crisp ontology O, and O_{ref}
(B) A crisp similarity measure σ
Output: The fuzzified ontology $\mathcal{O}$ (see Definition 7) with membership functions μ
begin
1. **for** *each concept* A *in* O **do**
   1.1 **for** *each concept* x *in* O_{ref} **do**
      1.1.1 Compute $\mu_{\mathcal{A}}(x) = \sigma(A, x)$
2. **for** *each pair of fuzzified concepts* ($\mathcal{A}$, $\mathcal{A}'$) **do**
   2.1 Compute the degree of their fuzzy subsumption, i.e.,
   $\mathbf{is\_a}(\mathcal{A}, \mathcal{A}') = \inf_{x \in X} \mu_{\mathcal{A} \to \mathcal{A}'}(x)$
   2.2. If some relation $r \in R$ holds for (A, A'), then add $r(\mathcal{A}, \mathcal{A}') = 1$ to $\mathcal{O}$. Add $r(\mathcal{A}, \mathcal{A}') = 0$ otherwise
return $\mathcal{O}$

Algorithm 2: An algorithm for fuzzification of the source concepts.

Now that we have a fuzzy definition of each source concept at hand, we will see how subsumption, i.e. the (fuzzy) *is_a* relation between these concepts can be expressed. Note that the logical implication $(i \in g(A)) \to (i \in g(A'))$ holds for any instance i and concepts $A, A' \in \mathbf{C}$ such that $(A, A') \in$ *is_a*. As introduced in Section 3, the fuzzy counterpart for fuzzified concepts $\mathcal{A}$ and $\mathcal{A}'$ as given in Definition 7 is denoted by **is_a**. In order to ensure Equation (7) we set $\mathbf{is\_a}(\mathcal{A}, \mathcal{A}') =_{def} \inf_{x \in X} \mu_{\mathcal{A} \to \mathcal{A}'}(x)$ for all pairs of fuzzified concepts. This equation defines the fuzzy subsumption as a degree between 0 and 1 to which one concept is the subsumer of another.

For a fuzzy ontology $\mathcal{O}$, the function μ plays a similar role as the instance function g for the original ontology O. The original concepts were characterized by sets of instances, whereas the fuzzified concepts are described as fuzzy sets of reference concepts. It can be shown that if $\mu_{\mathcal{A} \to \mathcal{A}'}(x)$ is defined as the Gödel implication, it holds that **is_a** is a fuzzy quasi-order. It can also be shown that under other definitions of fuzzy implication (e.g., standard strict implication) **is_a** is a fuzzy partial order (a quasi-order which also fulfills antisymmetry).

Proposition 2. *If* $\mu_{\mathcal{A} \to \mathcal{A}'}(x)$ *is defined as the Gödel implication, it holds that* ***is_a*** *is a fuzzy quasi-order. This means that the following is true:*

(1) Reflexivity: For every fuzzified concept $\mathcal{A}$ *it holds that* $\boldsymbol{is\_a}(\mathcal{A}, \mathcal{A}) = 1$
(2) Transitivity: For all fuzzified concepts $\mathcal{A}, \mathcal{A}'$, *and* $\mathcal{A}''$ *it holds that*

$$\boldsymbol{is\_a}(\mathcal{A}, \mathcal{A}'') \geq \min(\boldsymbol{is\_a}(\mathcal{A}, \mathcal{A}'), \boldsymbol{is\_a}(\mathcal{A}', \mathcal{A}''))$$

A proof of the proposition is provided in Appendix B.

The last part of the fuzzified ontology that has to be defined are the membership functions for the relations in $r \in \mathcal{R}$, i.e. the value of $r(\mathcal{A}, \mathcal{A}')$. This value can, for instance, be set to 1, whenever the original relation holds between the source concepts A, A'. We define $r(\mathcal{A}, \mathcal{A}') = 0$ otherwise.

To summarize these ideas, a pseudo code of the fuzzification process of an ontology is given in Algorithm 2.

5.2. *Fuzzy Ontology Alignment*

Throughout the fuzzy alignment phase of our approach, we rely on the fuzzy set representations of the concepts of the source ontologies in order to judge on their similarity.

5.2.1. *Using the Crisp Algorithm for Fuzzy Alignment*

In order to specify the fuzzy alignment approach, we first draw the reader's attention to the fact that the fuzzification process described in the previous section results in a fuzzy ontology, which has different properties than the original crisp one. For instance, the subsumption relation holds between any two fuzzified concepts and in both ways with a certain strength between 0 and 1, forming a complete directed weighted graph.

In order to be able to use Algorithm 1 also on the fuzzified ontologies, we adjust the fuzzy ontology in the following manner. For all fuzzified concepts $\mathcal{A}$ and $\mathcal{A}'$, such that A' is the parent of A in the crisp ontology, and every reference concept x such that $\mu_{\mathcal{A}}(x) > \mu_{\mathcal{A}'}(x)$, we set $\mu_{\mathcal{A}}(x) := \mu_{\mathcal{A}'}(x)$, adjusting the scores so that the fuzzy **is_a** equals 1, whenever a crisp *is_a*-relation exists. This procedure will be denoted as "Adjust($\mathcal{O}$)" in the following.

We can now apply Algorithm 1 for matching two fuzzy ontologies. In the definition of (in-)compatibility, we only consider pairs of concepts to be connected by a (fuzzy) relation (i.e., *is_a* or a relation $r \in \mathcal{R}$), if the respective membership value corresponds to 1. Values smaller than 1 are treated as zeros (i.e., no relation exists). The described approach guarantees that the resulting fuzzy ontology is structurally simpler than the two source ontologies – the number of relations remains the same, but the number of relation tuples, i.e., pairs of concepts, for which the relation holds, is smaller than in the original ontologies and depends on the structural similarity of the source ontologies. This approach also worked well in the experiments. However, there are also more sophisticated approaches like introducing a threshold in the definition of compatibility.

5.2.2. *Similarity Measures for Fuzzified Concepts*

A second difference when using Algorithm 1 is that we can *not* use the original similarity measure, σ, in the algorithm any more. Instead, we now have to define similarity measure based on the membership functions of the fuzzy concepts to be

matched. In contrast to direct crisp matching without using a reference ontology, σ is only used in the fuzzification process.

Consider two concepts A and B from two different ontologies and their fuzzy representations $\mu_{\mathcal{A}}$ and $\mu_{\mathcal{B}}$, as defined in Section 3. In order to measure their similarity, we will use one of the similarity measures for fuzzy sets given in Equations (2)—(5), also in Section 3. This list of fuzzy set relatedness measures is not exhaustive – there are numerous other measures that can be extended for use in the proposed framework[4]. Several similarity measures defined for uncertain concepts by using rough description logics could potentially be adapted for use in our approach[15]. However, for the purposes of our study these four basic measures, often referred to in the fuzzy ontology literature[10], have been sufficient to test the efficiency of the proposed approach. We provide an evaluation of their comparative performance with respect to the precision of the achieved alignments in the experimental part of the study.

5.2.3. *Matching Fuzzy Concepts*

There is one step left to be described. The matching algorithm only outputs a set of cross-ontology concept pairs. This means that the result of the matching algorithm is not a fuzzy ontology as defined above.

One natural possibility of defining the match of two fuzzy concepts is using again the t-norm (corresponding to performing an intersection). If in the match a source concept A is matched to a source concept B, their concept match can be computed as the fuzzy set

$$\mu_{(\mathcal{A},\mathcal{B})}(x) = T(\mu_{\mathcal{A}}(x), \mu_{\mathcal{B}}(x)), \forall x \in X. \tag{9}$$

This method allows us to interpret the result of the matching algorithm again as a fuzzy ontology that in subsequent steps might be matched with other fuzzy ontologies, without having to resort to the instance sets of the source ontologies. This is one of the genuine advantages of the fuzzy approach: we start off with an instance-based approach. After the fuzzification steps, however, we do not have to use the instances any more but can resort to methods of fuzzy logical inference.

Finally, we are now able to describe how the fuzzy matching with respect to the reference ontology proceeds: see Algorithm 3. Mind that in fuzzy matching, we use the membership functions μ^1 and μ^2 to replace the original instance function parameters of "Match". For the sake of simplicity, in Equation (9) and elsewhere, we have left the superscripts out and have used the same notation for these functions, μ. Note also that we now have two similarity measures: a crisp one, σ, only used for fuzzification, and a fuzzy one, ρ, that is used in the matching algorithm. Recall that θ is a similarity threshold to be set by the user.

Function FuzzyMatch(O_1, O_2, O_{ref}, σ, ρ, θ)
begin
1. $\mathcal{O}_1$ = Adjust(Fuzzify(O_1, O_{ref}, σ))
2. $\mathcal{O}_2$ = Adjust(Fuzzify(O_2, O_{ref}, σ))
3. Compute W = Match($\mathcal{O}_1$, $\mathcal{O}_2$, ρ, θ)
4. **for** *all* $(\mathcal{A}, \mathcal{B}) \in W$ **do**
   4.1 Determine μ via $\mu_{(\mathcal{A},\mathcal{B})}(x) = T(\mu^1_{\mathcal{A}}(x), \mu^2_{\mathcal{B}}(x)), \forall x \in X$.
5. Set $\mathcal{C} = W$
6. Determine **is_a** and $\mathcal{R}$ based on μ and W
7. Set $\mathcal{O} = (\mathcal{C}, \mathbf{is\_a}, \mathcal{R})$
8. Return $\mathcal{O}$ together with μ

Algorithm 3: Summary of the fuzzy matching algorithm.

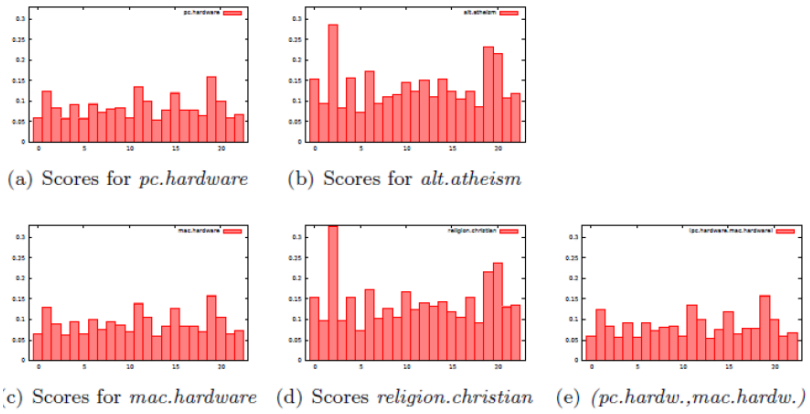
6. Experiments and Application in the Multimedia Domain

This section provides an application of the proposed approach in the textual and multimedia domains by using two alignment settings. We first describe the alignment of two collections of concepts that were obtained from the 20 Newsgroups dataset[27]. This experiment aims at showing that the transition to the fuzzy framework is successful and provides several examples of fuzzified concepts. In a second experiment, we consider the more difficult alignment of two multimedia ontologies, which are populated with images that are described by keywords. In both experiments, we use as a reference ontology a subset of Wikipedia's category system, which forms a directed acyclic graph of concepts.

6.1. *An Instance-based Similarity Measure for Crisp Concepts*

In our experiments, we focus on the case where instances of a concept are vector space representations of documents, which describe images annotated with the given concept. Let us consider the ontologies O_1 with a set of instances $g_1 : \mathbf{C_1} \longmapsto 2^{I_1}$, O_2 with a set of instances $g_2 : \mathbf{C_2} \longmapsto 2^{I_2}$ and the reference ontology O_{ref} with a set of instances $g_{ref} : X \longmapsto 2^{I_{ref}}$. In general, instances of an ontology can be, for example, n-tuples, real-valued vectors, and even whole documents, sentences within documents, or words within sentences which reference a concept, objects identified by uri's (in the context of the Semantic Web), records in a database identified by their primary keys (in the context of relational databases), etc. Our approach only requires a suitable similarity measure defined on the respective set of instances.

In the case of documents, it is possible to represent every document from a given corpus as a vector whose dimensions are different terms encountered in that corpus and the actual values are term-frequencies in the particular document in question, resulting in (possibly) sparse representations (e.g., *tf-idf*). For two ontologies to align, these representations are acquired by indexing simultaneously the two collections of documents, corresponding to each of the two ontologies which results in



(a) Scores for *pc.hardware* (b) Scores for *alt.atheism*

(c) Scores for *mac.hardware* (d) Scores *religion.christian* (e) *(pc.hardw.,mac.hardw.)*

Fig. 1. Fuzzy membership functions: Scores of some concepts w.r.t. the Inex 2007 Wikipedia Ontology. (a)–(d) represent single concept scores, while (e) represents the scores of the match of two concepts.

projecting all documents onto one single vector space.

In the current study, we relied on the straightforward idea that determining the similarity $\sigma(A, x)$ of two concepts $A \in \mathbf{C_1}$ and $x \in X$ consists in comparing their instance sets $g_1(A)$ and $g_{ref}(x)$. For doing so, we need a similarity measure for instances $\mathbf{i}^A$ and $\mathbf{i}^x$, where $\mathbf{i}^A \in g_1(A)$ and $\mathbf{i}^x \in g_{ref}(x)$. We can use, for example, the cosine $s(\mathbf{i}^A, \mathbf{i}^x) = \frac{\langle \mathbf{i}^A, \mathbf{i}^x \rangle}{\|\mathbf{i}^A\| \|\mathbf{i}^x\|}$ (i.e. the normalized scalar product). Based on this similarity measure for elements, the similarity measure for the sets can be defined by computing the similarity of the mean vectors corresponding to class prototypes, i.e.,

$$\sigma_{proto}(A, x) = s\Big(\frac{1}{|g_1(A)|} \sum_{j=1}^{|g_1(A)|} \mathbf{i}_j^A, \frac{1}{|g_{ref}(x)|} \sum_{k=1}^{|g_{ref}(x)|} \mathbf{i}_k^x\Big). \tag{10}$$

Note that (10) is consistent with the definition of a similarity measure (Def. 3) given in Section 4.1. Using this similarity measure also underlies the CAIMAN approach[26] in which concepts are assumed to be represented by their mean vector.

Other approaches of instance-based concept similarity can be employed as well like a variable selection based approach[51,52]. In general, the choice of the similarity measure is application dependent. In our experiments, it turned out that the prototype method worked best. It has the additional advantage of a low computational cost.

6.2. *Fuzzifying a Set of Textually Populated Concepts: An Example*

In what follows, we will focus on the fuzzification of single concepts and on computing concept matches. The recursive matching algorithm is the focus of the next

section.

As a reference vocabulary, we consider the 23 categories that form Wikipedia's so-called main topic classifications. For each topic category, we included a set of corresponding documents from the Inex 2007 corpus[e]. The Inex corpus contains Wikipedia pages and each page can have several categories. Note that the Wikipedia categories have Wikipedia pages (e.g., http://en.wikipedia.org/wiki/Category:Agriculture), but those do not have any content, although there is usually a Wikipedia page with a similar title as the category page (e.g., http://en.wikipedia.org/wiki/Agriculture). Only the latter are contained in the Inex corpus. However, the category structure is not represented in Inex 2007 and had to be extracted separately and combined with the Inex corpus[f]. The documents used for the respective category either directly belong to this category, or to one of its direct subcategories in the Wikipedia category tree. This way we arrived at the following 23 concepts (the numbers after the concept name correspond to the number of associated documents): *law (745), technology (293), belief (187), arts (319), society (2050), agriculture (530), social_sciences (1695), health (341), education (515), mathematics (1903), people (136), business (1202), science (547), history (445), politics (896), applied_sciences (1302), geography (164), chronology (303), environment (467), nature (234), humanities (537), language (427), culture (765).*

We have constructed two sets of classes of documents by using the 20 Newsgroup dataset, as described by Todorov et al.[50]. These can be considered as small source ontologies O_1 and O_2 (cf. Definition 1) consisting only of the following concepts and number of instances:

- O_1 = {*sci.med (990), rec.autos (990), alt.atheism (799), sport.baseball (994), pc.hardware(982)* }
- O_2 = {*sci.space (987), rec.motorcycles (993), religion.christian (997), sport.hockey (999), mac.hardware (961)* }.

In order to compute the prototype similarity, σ_{proto}, we first transformed the documents into *tf-idf* vectors. The prototype method then computes a single prototype (i.e., mean vector) for each class. For a pair of classes, their similarity corresponds to the cosine of their prototype vectors.

The diagrams in Figure 1 show the scores of some concepts (the *de facto* fuzzy sets representations) with respect to the selected categories from Inex 2007 Wikipedia. It can be seen that the membership functions of *pc.hardware* and *mac.hardware* are quite similar, as are those of *alt.atheism* and *religion.christian.* In contrast, *alt.atheism* and *religion.christian* are quite dissimilar to the hardware classes. The scores of the pair of matched concepts *(pc.hardware, mac.hardware)* are obtained by taking the minimum (see Figure 1(e)). The two religion-related con-

[e]http://www-connex.lip6.fr/ denoyer/wikipediaXML/
[f]A version of the category tree can be found here: http://wikicategory.sourceforge.net/.

O_1	O_2	ρ_{base}
sci.med	sci.space	0.781
rec.autos	rec.motorcycles	0.915
alt.atheism	religion.christian	0.924
sport.baseball	sport.hockey	0.949
pc.hardware	mac.hardware	0.963

O_1	O_2	σ_{proto}
sci.med	religion.christian	0.359
rec.autos	rec.motorcycl.	0.471
alt.atheism	religion.christian	0.537
sport.baseb.	sport.hockey	0.559
pc.hardware	mac.hardware	0.716

Fig. 2. (a) Fuzzy matches determined by largest similarities ρ_{diff}. (b) Crisp matches determined using largest similarities σ_{proto}

cepts have their highest peaks at the Wikipedia concept *belief*. For the two hardware classes, the Wikipedia concepts with the highest scores are *technology, business, and nature*, although there is no pronounced peak in the score values. This means that *pc.hardware* and *mac.hardware* cannot be characterized in a very precise manner by using the 23 concepts. In order to achieve a better separation of the source concepts, we therefore decided to add the concept *computing* to the Wikipedia ontology, giving 24 reference concepts altogether. Note that using the scores means that we reduce the concept descriptions from more than 22.220 different tokens to just 24 features.

Using the base distance ρ_{base} for selecting the best concept matches in O_1 and O_2, we arrive at the alignment in Figure 2(a), which shows the best match from O_2 for every concept in O_1. The fuzzy method is obviously able to map related yet different concepts. Even the less evident match (*sci.med*, *sci.space*) is detected.

As discussed in Section 5.2.3, it is possible to define a fuzzy membership function of the matched concepts, which is obtained as the minimum of the respective scores. The one for the match *(pc.hardware,mac.hardware)* is shown in Figure 1(e). Figure 2(b) shows the match that is found by comparing the prototypes of the respective concepts (higher values are better), i.e., the result of the crisp match between O_1 and O_2.

6.3. *Multimedia Ontology Alignment: Matching LSCOM and LabelMe*

In a second series of experiments, we align two complementary heterogeneous multimedia knowledge sources containing annotated pictures. We chose, on one hand, LSCOM[45] – an ontology dedicated to multimedia annotation. It was initially built in the framework of TRECVID[g] with the criteria of concept usefulness, concept observability and feasibility of automatic concept detection. LSCOM is populated with the development set of TRECVID 2005 videos. Since this set contains images from broadcast news videos, LSCOM is particularly adapted to annotate this kind of content, containing abstract and specific concepts (e.g., SCIENCE_TECHNOLOGY, INTERVIEW_ON_LOCATION).

[g]http://www-nlpir.nist.gov/projects/tv2005/

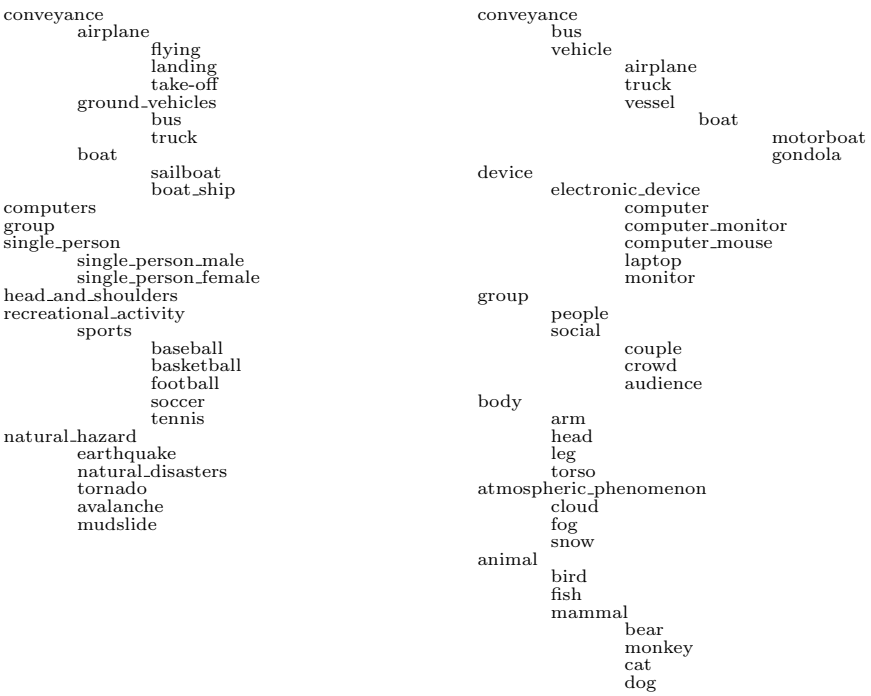


Fig. 3. Left: The part of the LSCOM ontology that was used in the experiments. Right: The subontolgy of LabelMe used in the experiments

As a second ontology, we used a hierarchical structure of synsets from WordNet[32] populated with the LabelMe dataset[40], referred to as the LabelMe ontology hereafter. This ontology has been constructed by using the LabelMe annotation tool which allows for user annotation of images. The provided keywords have then been matched to WordNet synsets and their structural relations have been extracted by following the *is_a* relations in the lexical database. In contrast to LSCOM, this ontology is very general. It is populated with photographs from daily life and contains concepts such as CAR, COMPUTER, PERSON, etc. The parts of the LSCOM and LabelMe ontologies that we used are shown in Figure 3.

Regarding the concept population with images, there have been on average several hundred of images per concept used in the experiments, but the number varies (e.g., some concepts have below hundred, others - above 500 associated images). This number depends on the availability of images that are / can be labeled by a given concept in the respective image bases. Furthermore, the documents of a given class have also been assigned to its super-classes as that was the semantics that we wanted to consider.

107_Standing One or more people standing up. 227_Bus Shots of a bus. 224_Outdoor Shots of Outdoor locations. 217_Person Shots depicting a person. The face may be partially visible. 202_Crowd Shots depicting a crowd. 181_Adult Shots showing a person over the age of 18. 104_Male_Person One or more male persons. 290_Daytime_Outdoor shots that take place outdoors during the day. 316_Group We defined a group as 3-10 people. 109_Windows An opening in the wall or roof of a building or vehicle fitted with glass or other transparent material.

Fig. 4. The LSCOM concept Bus: a visual and a textual instance.

6.3.1. *Representing an Image as a Text Document*

We acquired text representations for every instance of the two multimedia ontologies. The ontology obtained from LSCOM was initially populated with single frames of videos, each of which is annotated with a set of keywords describing the content of the picture. In order to enrich the textual description, we included the definitions of these keywords into the document describing the respective image. A challenge arises from the fact that a scene usually consists of several objects, which often are not related to the object that determines the class of the image. In such cases, the other objects in the image act as noise. An example of a visual instance of a multimedia concept together with the acquired textual description is given in Figure 4.

The second ontology was obtained from the LabelMe dataset[40], which allows users to annotate scenes with keywords describing the objects in them. In order to enrich the textual descriptions of the images, we included the WordNet definitions[32] of the keywords into the textual description of the image. However, we did no word sense disambiguation, which made the task substantially more difficult.

In order to obtain *tf-idf* representations, we performed the usual preprocessing steps for the documents from LSCOM and LabelMe, including those from Inex 2007 Wikipedia. Preprocessing included stemming, removing English stopwords, deleting tokens with less than 3 characters, and those that occur in less than 10 documents.

6.3.2. *Applying the Recursive Alignment Algorithm*

In the following, we present the result of applying the recursive Algorithm 1. The fuzzy match is based on $\rho_{sup\text{-}min(2)}$, whereas the crisp match is based on σ_{proto}. We set the similarity threshold to 0.0 resulting in largest matches.

The results by using the fuzzy and the crisp method are shown in Figures 5(a) and 5(b), respectively. One of these alignments is visualized for the fuzzy case in Figure 6. Algorithm 1 always picks the best-matching pair of concepts in each recursion. Based on its selection, it restricts the sets of remaining potential matches until the set of potential matches becomes empty. The restrictions result from the requirements that (1) only injective mappings should be computed and (2) the is_a

LSCOM	LabelMe	σ
sailboat	motorboat	0.341
bus	bus	0.273
avalanche	snow	0.258
conveyance	conveyance	0.25
truck	truck	0.24
boats	vessel	0.216
computers	device	0.169
airplane_landing	airplane	0.151
group	people	0.15
tornado	cloud	0.127
head_and_shoulder	torso	0.113
natural_hazard	atmos.	0.099
football	crowd	0.084
natural_disasters	fog	0.08
single_person	head	0.077
sports	social	0.077
basketball	couple	0.073
soccer	audience	0.062
boat_ship	gondola	0.057

(a) Fuzzy: σ (algorithm) $= \rho_{sup\text{-}min(2)}$

LSCOM	LabelMe	σ
airplane	airplane	0.347
ground_vehicles	truck	0.289
conveyance	conveyance	0.288
avalanche	snow	0.269
sailboat	motorboat	0.244
computers	comp._monitor	0.219
boats	boat	0.192
group	group	0.188
tornado	cloud	0.184
head_and_shoulder	body	0.128
natural_hazard	atmos.	0.109
single_person_male	computer	0.094
single_person_female	laptop	0.069
natural_disasters	fog	0.05
boat_ship	gondola	0.044
soccer	monitor	0.033
football	mammal	0.027
tennis	comp._mouse	0.016
baseball	fish	0.006
basketball	bird	0.002

(b) Crisp $\sigma = \sigma_{proto}$

Fig. 5. The fuzzy and crisp alignments with matches (concept pairs) as they are picked by the algorithm.

relationship is to be preserved.

The pairs in figures 5(a) and 5(b) are ordered according to their similarity value (as they are selected by the alignment algorithm). Setting the threshold θ to a value greater than 0.0 results in discarding the entries at the bottom of the lists. In both the fuzzy and the crisp case, the matches tend to get less plausible at the end of the list (lower similarity values). Setting an appropriate value of $\theta > 0.0$ will therefore result in a better alignment, since it cuts off matches with too small a similarity. Mind that the actual strengths of the **is_a** relationships in the alignment are determined based on the Gödel implication (see Equation (1)).

6.4. *Evaluation of the Alignments*

Aligning multimedia ontologies is a specific task and there does not exist a ground truth or reference alignments against which our results can be compared. Therefore, in order to evaluate our approach empirically, we have produced an evaluation benchmark by asking people to provide similarity score between target concepts from LSCOM and possible LabelMe alignments. For each LSCOM concept, a list of LabelMe concepts was presented and assessors were asked to give scores between 1 (min similarity) and 3 (max similarity) to the three concepts that they considered to be the most similar to the target one. There were 11 assessors that completed the test correctly and, in order to obtain similarities for each target concept, we selected only those similar concepts chosen by at least three evaluators in order to have a notion of consensus. For each of these, we computed the average similarity and ranked them by decreasing score to obtain a "human driven" ground truth for the alignment. In order to favour the reproducibility of this research, we make

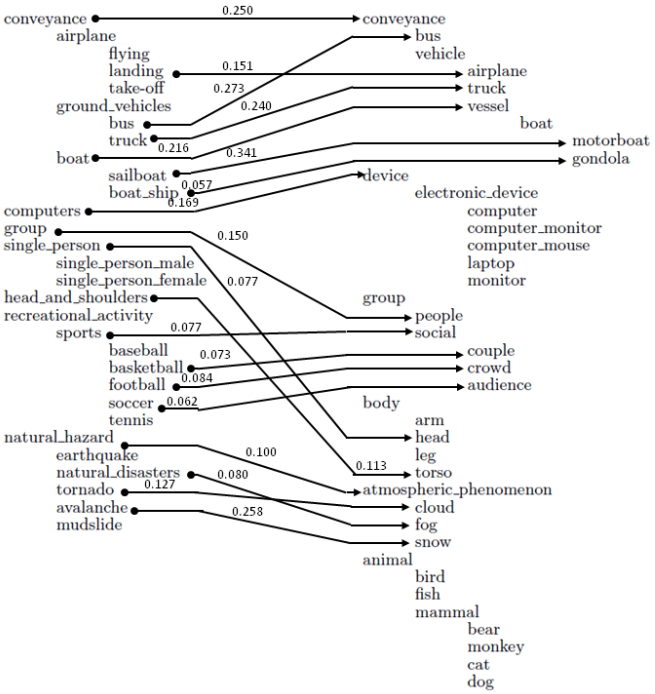


Fig. 6. The fuzzy alignment: result of applying Algorithm 1 with $\sigma = \rho_{sup\text{-}min(2)}$ and $\theta = 0.0$

this benchmark available in Figure 7. We have retained the top 3 most similar concepts for each LSCOM concept. In order to break ties, i.e., equal scores, concepts mentioned by a larger number of users were ranked higher.

The performance of the alignments is computed as the number of aligned concepts obtained with the crisp method (denoted CRISP) and with the fuzzy (denoted FUZZY) that are found in Figure 8. To have a fair comparison of recursive alignments, we retained only LSCOM concepts that had alignments with both methods - this intersection is composed of 14 concepts. For non-recursive alignments (i.e., direct alignments not using the recursive algorithm), there are 29 concepts in all and we computed the matching with the ground truth with alignment depths from 1 to 3 (i.e., number of aligned concepts considered). The results confirm that FUZZY performs better than CRISP for both alignment types. Since there are more concepts with non-recursive alignments, the difference is a bit more pronounced in this setting.

LSCOM	**LabelMe**		
	Depth 1	Depth 2	Depth 3
airplane takeoff	2.8: airplane	1.75: bird	1.7: conveyance
natural disasters	2.5: atm_phenom.	1.5: cloud	1.34: snow
single person	2.4: people	2.1: body	1.25: head
bus	3: bus	2.25: conveyance	2.0: vehicle
sailboat	2.9: boat	1.8: vehicle	1.8: vessel
baseball	2.2: group	2.13: social	2.0: crowd
natural hazard	2.3: atm_phenom.	1.6: snow	1.23: cloud
single person male	2.3: people	2.25: couple	2.0: head
truck	3: truck	2.0: vehicle	1.75: conveyance
ground vehicles	2.6: vehicle	2.1: truck	1.75: conveyance
football	2.5: crowd	2.3: body	2.0: social
airplane	3: airplane	2.1: vehicle	2.0: conveyance
computers	3: computer	2.14: comp_monitor	2.0: elec_device
airplane flying	2.9: airplane	2.0: atm_phenom.	1.75: conveyance
group	3: group	2.0: people	2.0: crowd
earthquake	2.25: people	1.8: crowd	1.67: fog
basketball	2.2: group	2.25: crowd	1.8: people
airplane landing	2.81: airplane	1.86: vehicle	1.75: conveyance
tennis	2.125: social	2.0: couple	2.0: body
boat	3: boat	2.2: motorboat	2.0: vessel
sports	2.63: social	2.0: people	2.0: group
single person female	2.4: body	2.0: people	2.0: couple
head and shoulder	2.6: head	2.0: torso	1.8: body
soccer	2.67: audience	2.4: crowd	2.25: leg
conveyance	3: conveyance	2.1: vehicle	1.75: vessel
avalanche	2.72: snow	2.2: atm_phenom.	1.67: body
tornado	2.8: atm_phenom.	1.89: cloud	1.3: snow
recr. activity	2.3: social	2.0: group	1.57: people
boat ship	2.9: boat	2.14: vessel	1.75: motorboat

Fig. 7. Human driven ground truth alignment from LSCOM to LabelMe. "Depth" corresponds to the similarity rank in the evaluation.

6.5. *Analysis of the Impact of the Choice of a Reference Vocabulary*

In the following series of experiments, we analyze the impact of the size and type of the reference vocabulary on the precision[h] of the derived alignments. In order to evaluate this quantity, we have used the ground truth alignments in Figure 7. This allows us to compute precision levels for every fuzzy similarity measure and every choice of reference vocabulary. We started by removing concepts one at a time from

[h]Precision is defined standardly as the proportion of relevant retrieved matches with respect to all found matches.

	Recursive	Non-Recursive Depth 1	Non-Recursive Depth 2	Non-Recursive Depth 3
CRISP	0.5	0.55	0.5	**0.54**
FUZZY	**0.57**	**0.62**	**0.59**	**0.54**

Fig. 8. Precisions with respect to the ground truth alignment.

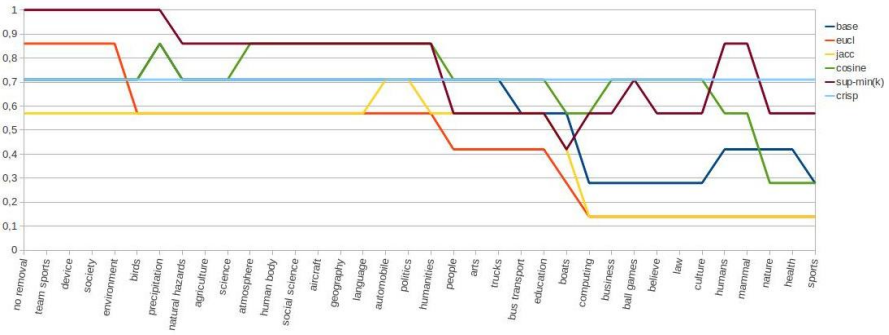


Fig. 9. Precisions of the fuzzy similarity measures with respect to different reference ontologies. The x-axis indicates the different reference ontologies used, where every ontology is characterized by the concept that has been removed at each iteration.

the reference vocabulary and observing the behavior of the precision curves at each time. The obtained results are given in Figure 9. The straight blue line indicates the precision of the direct crisp alignment (a constant line, since it describes the precision of alignments, which are independent on the reference ontology). We have made the following observations.

The curves confirm that the more concepts we have, the better the performance is (with respect to the ground truth **and** to the crisp alignment). When removing concepts, no semantic considerations were taken into account (i.e., the selection of the removed concepts and the order of their removal is arbitrary).

The precision of some measures deteriorates faster than that of other measures. This observation provides information regarding the robustness of the measures. This is an important result which helps us to justify and analyze the choice of similarity measures. In our example, we see that $\rho_{sup-min(k)}$ is more robust (in the sense of keeping an overall higher precision rate) than the other measures - an observation that is explained by the fact that it takes mean values and therefore changes slower if one concept is deleted at a time.

Finally, we could identify several "milestone" concepts, which dictate the overall behavior of the curves, i.e., concepts, for which we observe a drastic change (decline) of the precision curves.

Based on the latter observation, we repeated the same experiment, but this time starting with a vocabulary which only contains the identified milestone concepts. This time we have regrouped the concepts by following certain common sense semantic considerations. The results are given in Figure 10. The straight blue line indicates again the precision of the direct crisp alignments. The set of selected milestone concepts can be seen along the x-axis.

We now observe that if we carefully select our concepts, i.e., take only those reference concepts that are part of an optimal alignment of LSCOM and O_{ref}, and LabelMe and O_{ref}, we obtain results which are almost as good as when we use all

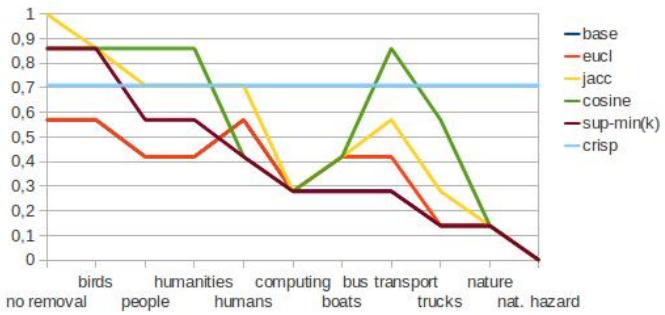


Fig. 10. Precisions of the similarity measures with respect to different reference vocabularies containing only milestone concepts (analogously to Figure 9).

reference concepts. The observation that these concepts are all part of the crisp alignments is important, since in this case their choice is not dependent on the precision curves which rely on a manually defined ground truth. We can postulate that the crisp alignment defines the precise choice of reference concepts to be used.

In conclusion, we note that both the number and type of concepts that form the reference vocabulary play a role with respect to the precision of the derived fuzzy alignments. Our results showed that, based on an optimal crisp alignment, it is possible to select a small number of "milestone" concepts for the reference ontology. In addition, some robustness properties of the fuzzy similarity measures have been revealed. Finally, it has been shown that with an appropriate selection of a reference vocabulary, the fuzzy alignment outperforms the crisp one for most of the considered fuzzy similarity measures.

6.6. *Application of the Alignments to Query Reformulation for Information Retrieval*

The LSCOM ontology was built for video news annotation purposes, while the scope of LabelMe is rather general and common-sense. Hence, these ontologies differ both in their conceptual content (number, granularity and genericity of the concepts) and in their usage. The heterogeneities that these ontologies tend to exhibit, if handled appropriately, would allow their use as complementary knowledge sources. Particularly, the collaborative search and retrieval of images across these two knowledge and data sources can be enabled by the help of an alignment. In the presented approach, such an alignment is derived from similarities on the instance level in a bottom-up manner. Indeed, the different terms from both vocabularies are projected onto a common semantic space provided by a reference vocabulary which embeds their extensional characteristics. The alignment performed in this new common semantic space helps us relate optimal query terms from both vocabularies. For example, our approach would indicate that LSCOM-images retrieved by using a query term "natural hazard" in LSCOM would be similar or related to LabelMe-images retrieved

by using the term "atmospheric phenomenon" in LabelMe.

We have evaluated the performance within an information retrieval setting of our approaches – both crisp and fuzzy, as well as both in a direct alignment and by using the recursive algorithm (Algorithm 1). We have used images from LSCOM and LabelMe annotated by the concepts of interest. In order to provide a balanced evaluation, for the recursive method we have only used the matches which are contained in the intersection of the crisp and the fuzzy approaches. The precision rates measured at 10 (P@10) and at 20 (P@20) returned results[i] are given in Figure 11.

First and foremost, we note that the alignments, both crisp and fuzzy, enrich the returned results and the overall quantity of relevant results is increased. However, there is no pronounced advantage of one of the two methods with respect to the other, even though within the ground truth evaluation provided in Section 6.4, the fuzzy method clearly outperforms the crisp one. Several observations are relevant with regard to these results and the used data.

A fact that penalizes FUZZY against CRISP is that it does worse at aligning concepts that have identical names in LSCOM and LabelMe. This is notably the case for Airplane and Truck, but also for Computer and Boat (the LabelMe concept with this name comes second for CRISP and third for FUZZY). This effect could be reduced by incorporating a test that favors the alignment of concepts with identical or very similar names in the alignment framework (outside the scope of the paper).

With respect to data properties (and particularly concept population discussed in the beginning of Section 6.3), there is a small number of concepts that have a lot of associated images for which CRISP is better than FUZZY and a larger number of concepts for which FUZZY is better than CRISP but which have only few associated images, leading to a situation analogous to that in which active minorities make the difference when facing silent majorities.

Yet another difficulty comes from the fact that, in the given setting, we have alignments that are intrinsically hard. This is the case for all sport related concepts that appear in LSCOM which are not mirrored in LabelMe. When we look at the alignments for both methods, the closest concepts are those related to people which are only remotely related to sports. Even though FUZZY is better here, this does not necessarily play in its favor, since there are only few chances to find LabelMe images tagged simultaneously with "people" and "soccer".

7. Conclusion and Future Work

In summary, we have proposed a technique for matching the concepts of a set of domain ontologies by using a fuzzy set formulation and a generic reference vocabulary as a mediator. Fuzziness helps to embed vagueness in concept representation and

[i]The choice of these precision measures is motivated by the fact that, in retrieval scenarios, people favor the top 10 to 20 images that are shown on a result page, whereas results on following pages are more seldom explored.

	CRI(1)	FUZZ(1)	CRI(2)	FUZZ(2)	CRI(3)	FUZZ(3)	CRI	FUZZ
P@10	**0.514**	0.497	**0.531**	0.500	0.462	**0.490**	**0.141**	0.117
P@20	**0.490**	0.479	**0.509**	0.502	0.453	**0.476**	0.114	**0.119**

Fig. 11. Precisions for the crisp (CRI) and fuzzy (FUZZ) alignments. The first six columns contain the results for the non-recursive alignment, the number in brackets being the number of aligned LabelMe concepts that have been used to expand the LSCOM query; the last two columns contain the results of the recursive algorithm (only one concept from LabelMe is added to an LSCOM concept).

in the matching while the use of a reference vocabulary provides uniform semantic criteria for this representation and for the comparison of any two concepts from any two source ontologies. The computation of the ontology alignment is inexpensive. As a direct consequence of the suggested fuzzy representation of the concepts, the match itself is representable as a fuzzy set of instances or reference concepts. This is particularly interesting in terms of the scalability, allowing any newly arrived concept to be directly compared to the already computed matches in a simple and inexpensive manner. The paper proposes examples of the alignment framework in the domain of multimedia, as well as an evaluation of the produced alignments in the field of query reformulation for information retrieval. An analysis of the impact of the choice of a reference vocabulary is also given.

In what follows, we will discuss some of the visible open ends of this contribution. *Data integration* is defined as the process of creating a common knowledge source which brings together several local domain ontologies[19,23], the goal being to provide uniform access to these heterogeneous data sources without loading them into a common data warehouse. In the general case, the fuzzy subsumption can be computed for any two source concepts even if they come from different ontologies. This gives rise to the construction of a novel, common fuzzy ontology which relates all the terms from two or more source ontologies describing different databases. This work is currently in progress. A preliminary description is found in the Article by Todorov, Geibel and Hudelot[49].

The proposed framework suggests the possibility of *integrating heterogeneously defined ontologies*, for example an instance-populated ontology and a purely terminological knowledge source. The fuzzification of these sources can be carried out by using similarity measures of different kinds. Alternatively, aligning ontologies which describe heterogeneous data (e.g., images on one side and text on the other) will also be tested experimentally in the future.

Handling multilingual information is becoming more and more important leading to the need of bringing multilingual semantic information and knowledge together in an explicit manner. The approach described in this paper suggests a suitable framework for aligning linguistically heterogeneous ontologies (cross-lingual ontologies). In future work, we will be investigating the possibility of redefining the alignment framework by using a multilingual reference vocabulary, instead of a

single-language one.

Finally, we note that the proposed approach is applicable to lightweight ontologies, compliant with Definition 1. Although ontologies given in RDFS can be represented using this definition, certain kinds of general OWL ontologies cannot be handled within this setting. Therefore, in the future we will be extending the proposal with elements of OWL 2, including relations and axioms between concepts and instances which is not covered by the ontology definition used in this paper.

Appendix A. Proof of Proposition 1.

Proof. We show that the function call Map(M, O_1, O_2, σ, θ) outputs a maximal θ-alignment of O_1 and O_2 for $M = \mathbf{C_1} \times \mathbf{C_2}$. However, since in the recursive call of "Map" (see step 5), the set of pairs, M, is a subset of $\mathbf{C_1} \times \mathbf{C_2}$, we have to consider the more general case $M \subseteq \mathbf{C_1} \times \mathbf{C_2}$ and show that $W = \text{Map}(M, O_1, O_2, \sigma, \theta)$ is a maximal θ-alignment **with respect to the set** M. This means we consider a slight generalization of Definition 5. Using this idea, we have to show that for all M it holds that

(1) W is an admissible alignment of O_1 and O_2,
(2) W cannot be extended by further concept pairs from M ("maximal" with respect to M), and
(3) $\sigma(A, B) \geq \theta$ for all $(A, B) \in W$.

We now consider the set size $m = |M|$ and show by complete induction on m that the output W fulfills (1)-(3) for all set sizes m.

Induction basis $m = 0$. In this case, M is empty. According to step 1 of "Map", the algorithm outputs $W = \emptyset$. Obviously, W is (1) an alignment, (2) cannot be extended with pairs from M and (3) $\sigma(A, B) \geq \theta$ holds for all $(A, B) \in W$ since there is no such pair. This proves the induction basis.

Induction Step $m > 0$. We now assume the induction hypothesis that (1) - (3) are true for all sets M for which $|M| < m$ holds. According to the steps 2 and 3 of the algorithm, we have to consider two cases:

- **Case A:** for all $(A, B) \in M$ it holds that $\sigma(A, B) < \theta$
- **Case B:** there is some $(A, B) \in M$ such that $\sigma(A, B) \geq \theta$

Case A: In this case, the algorithm returns $W = \emptyset$ in step 3 of the algorithm. This is obviously the only admissible θ-alignment of the two ontologies. $W = \emptyset$ fulfills the conditions (1) - (3).

Case B: Let us assume that (A, B) is one of the cross-ontology pairs that maximizes σ (step 3). In step 4 of the algorithm, we compute a set $M' \subset M$ that contains all pairs of M that are compatible to (A, B). In particular, M' does not contain (A, B) or any node (i.e., concept pair) that contains either A or B. In particular $|M'| < m = |M|$. The assumption hypothesis can thus be assumed to hold for M'.

We now consider step 5. Let $W' = \text{Map}(M', O_1, O_2, \sigma, \theta)$. We can now apply the induction hypothesis. This means that W' fulfills (1) - (3). The return value W is obtained by adding (A, B) to W', i.e. $W = \{(A, B)\} \cup W'$. We now have to show that W fulfills (1) - (3) on the basis that W' fulfills (1) - (3).

Condition (1) holds since W is an admissible alignment of O_1 and O_2 with respect to M if W' is an admissible alignment with respect to M'. Since $M' \subset M$ holds, W' is also an admissible alignment with respect to the larger set M. Since (A, B) is compatible to all the nodes in W', $W = \{(A, B)\} \cup W'$ is an admissible alignment with respect to M.

Condition (2): Since W' is already maximal with respect to M', W is maximal with respect to M. This is the case because if we add (A, B) to W', we cannot add any other node from $M - M'$ since all nodes in $M - M'$ are incompatible to (A, B). We also cannot add any further node from M' since W' is already maximal with respect to M'. Thus, no further node from M can be added to W'. This means that W is maximal with respect to M.

Condition (3) holds, since $\sigma(A, B) \geq \theta$ and (3) holds for W', which completes the prove of the induction step.

Note that condition (2) only states that W is maximal for the nodes in M. However, since we choose $M = \mathbf{C_1} \times \mathbf{C_2}$ in the initial call of "Map" in "Match", W is indeed a maximal alignment of O_1 and O_2 without any restriction, q.e.d.

Appendix B. Proof of Proposition 2.

Proof. 1. It holds that

$$\mathbf{is\_a}(\mathcal{A}, \mathcal{A}) = \inf_{x \in X} \mu_{\mathcal{A} \to \mathcal{A}}(x)\,.$$

By definition, the equation $\mu_{\mathcal{A} \to \mathcal{A}}(x) = 1$ is true for all x, which proves the reflexivity of **is_a**.

2. We consider the following inequation

$$\begin{aligned} \min(\mathbf{is\_a}(\mathcal{A}, \mathcal{A}'), \mathbf{is\_a}(\mathcal{A}', \mathcal{A}'')) &= \min(\inf_x \mu_{\mathcal{A} \to \mathcal{A}'}(x), \inf_y \mu_{\mathcal{A}' \to \mathcal{A}''}(y)) \\ &\leq \inf_x \min(\mu_{\mathcal{A} \to \mathcal{A}'}(x), \mu_{\mathcal{A}' \to \mathcal{A}''}(x)) \end{aligned}$$

To complete the proof, we now consider the following table for the Göodel implication :

	$\mu_{\mathcal{A} \to \mathcal{A}'}(x)$	$\mu_{\mathcal{A}' \to \mathcal{A}''}(x)$	$\mu_{\mathcal{A} \to \mathcal{A}''}(x)$
$\mu_{\mathcal{A}}(x) \leq \mu_{\mathcal{A}'}(x) \leq \mu_{\mathcal{A}''}(x)$	1	1	1
$\mu_{\mathcal{A}}(x) \leq \mu_{\mathcal{A}''}(x) \leq \mu_{\mathcal{A}'}(x)$	1	$\mu_{\mathcal{A}''}(x)$	1
$\mu_{\mathcal{A}'}(x) \leq \mu_{\mathcal{A}}(x) \leq \mu_{\mathcal{A}''}(x)$	$\mu_{\mathcal{A}'}(x)$	1	1
$\mu_{\mathcal{A}'}(x) \leq \mu_{\mathcal{A}''}(x) \leq \mu_{\mathcal{A}}(x)$	$\mu_{\mathcal{A}'}(x)$	1	$\mu_{\mathcal{A}''}(x)$
$\mu_{\mathcal{A}''}(x) \leq \mu_{\mathcal{A}}(x) \leq \mu_{\mathcal{A}'}(x)$	1	$\mu_{\mathcal{A}''}(x)$	$\mu_{\mathcal{A}''}(x)$
$\mu_{\mathcal{A}''}(x) \leq \mu_{\mathcal{A}'}(x) \leq \mu_{\mathcal{A}}(x)$	$\mu_{\mathcal{A}'}(x)$	$\mu_{\mathcal{A}''}(x)$	$\mu_{\mathcal{A}''}(x)$

By considering each row in the table, we see that it holds that

$$\min(\mu_{\mathcal{A}\to\mathcal{A}'}(x), \mu_{\mathcal{A}'\to\mathcal{A}''}(x)) \leq \mu_{\mathcal{A}\to\mathcal{A}''}(x)$$

is true for all x. Therefore we can prove the fact

$$\begin{aligned}\min(\mathbf{is\_a}(\mathcal{A},\mathcal{A}'), \mathbf{is\_a}(\mathcal{A}',\mathcal{A}'')) &\leq \inf_x \min(\mu_{\mathcal{A}\to\mathcal{A}'}(x), \mu_{\mathcal{A}'\to\mathcal{A}''}(x)) \\ &\leq \inf_x \mu_{\mathcal{A}\to\mathcal{A}''}(x) \\ &= \mathbf{is\_a}(\mathcal{A},\mathcal{A}'')\end{aligned}$$

This proves the transitivity of **is_a**. q.e.d.

References

1. J.-I. Akahani, K. Hiramatsu, and T. Satoh. Approximate query reformulation based on hierarchical ontology mapping. In *In Proc. of Intl Workshop on SWFAT*, pages 43–46, 2003.
2. Z. Aleksovski, M. Klein, W. Ten Kate, and F. Van Harmelen. Matching unstructured vocabularies using a background ontology. *Managing Knowledge in a World of Networks*, pages 182–197, 2006.
3. Z. Aleksovski, W. ten Kate, and F. van Harmelen. Exploiting the structure of background knowledge used in ontology matching. In *Ontology Matching Workshop at International Semantic Web Conference (ISWC)*, 2006.
4. A. Bahri, R. Bouaziz, and F. Gargouri. Dealing with similarity relations in fuzzy ontologies. In *Fuzzy Systems Conference, 2007. FUZZ-IEEE 2007. IEEE International*, pages 1–6. IEEE, 2007.
5. F. Bobillo. *Managing vagueness in ontologies*. PhD thesis, University of Granada, Spain, 2008.
6. P. Buche, J. Dibie-Barthélemy, and L. Ibanescu. Ontology mapping using fuzzy conceptual graphs and rules. In *ICCS Supplement*, pages 17–24, 2008.
7. S. Calegari and D. Ciucci. Fuzzy ontology, fuzzy description logics and fuzzy-owl. In F. Masulli, S. Mitra, and G. Pasi, editors, *Applications of Fuzzy Sets Theory*, volume 4578 of *LNCS*, pages 118–126. Springer Berlin / Heidelberg, 2007.
8. S. Calegari and E. Sanchez. A fuzzy ontology-approach to improve semantic information retrieval. In *Proceedings of the Third ISWC Workshop on Uncertainty Reasoning for the Semantic Web-URSW*, volume 7, page 6, 2007.
9. A. Cali and T. Lukasiewicz. Tightly integrated probabilistic description logic programs for the semantic web. In *Proceedings of the 23rd international conference on Logic programming*, pages 428–429. Springer-Verlag, 2007.
10. V. Cross and X. Yu. A fuzzy set framework for ontological similarity measures. In *WCCI 2010, FUZZ-IEEE 2010*, pages 1 – 8. IEEE Compter Society Press, 2010.
11. R. Diestel. *Graph Theory, 4th Edition*, volume 173 of *Graduate texts in mathematics*. Springer, 2012.
12. A. Doan, J. Madhavan, P. Domingos, and A. Halevy. Learning to map between ontologies on the semantic web. In *WWW'02*, pages 662–673. ACM Press, 2002.
13. D. Dubois and H. Prade. Possibility theory, probability theory and multiple-valued logics: A clarification. *Annals of mathematics and Artificial Intelligence*, 32(1-4):35–66, 2001.
14. J. Euzenat and P. Shvaiko. *Ontology Matching*. Springer-Verlag, 1 edition, 2007.
15. N. Fanizzi, C. d'Amato, F. Esposito, and T. Lukasiewicz. Representing uncertain concepts in rough description logics via contextual indiscernibility relations. In *Proc. of URSW*, volume 8, 2008.
16. U. Feige, S. Goldwasser, L. Lovasz, S. Safra, and M. Szegedy. Approximating clique is almost NP-complete. In *Foundations of Computer Science, Annual IEEE Symposium on*, volume 0, pages 2–12, Los Alamitos, CA, USA, 1991. IEEE Computer Society.

17. A. Ferrara, D. Lorusso, G. Stamou, G. Stoilos, V. Tzouvaras, and T. Venetis. Resolution of conflicts among ontology mappings: a fuzzy approach. OM'08 at ISWC, 2008.
18. A. Gal. Managing uncertainty in schema matching with top-k schema mappings. *Journal on Data Semantics VI*, pages 90–114, 2006.
19. A. Gal and P. Shvaiko. Advances in web semantics i. chapter Advances in Ontology Matching, pages 176–198. Springer-Verlag, Berlin, Heidelberg, 2009.
20. P. Geibel, K. Schädler, and F. Wysotzki. Connectionist construction of prototypes from decision trees for graph classification. *Intell. Data Anal.*, 7(2):125–140, 2003.
21. R. Gligorov, W. ten Kate, Z. Aleksovski, and F. Van Harmelen. Using google distance to weight approximate ontology matches. In *Proceedings of the 16th international conference on World Wide Web*, pages 767–776. ACM, 2007.
22. T. R. Gruber. Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In N. Guarino and R. Poli, editors, *Formal Ontology in Conceptual Analysis and Knowledge Representation*. Kluwer Academic Publishers, 1993.
23. Alon Y. Halevy, Naveen Ashish, Dina Bitton, Michael Carey, Denise Draper, Jeff Pollock, Arnon Rosenthal, and Vishal Sikka. Enterprise information integration: successes, challenges and controversies. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, SIGMOD '05, pages 778–787. ACM, 2005.
24. T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM.
25. A. Isaac, L. van der Meij, S. Schlobach, and S. Wang. An empirical study of instance-based ontology matching. *The Semantic Web*, pages 253–266, 2008.
26. M. S. Lacher and G. Groh. Facilitating the exchange of explicit knowledge through ontology mappings. In *Proceedings of the 14th FLAIRS Conf.*, pages 305–309. AAAI Press, 2001.
27. K. Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339, 1995.
28. J. Lee. *A First Course in Combinatorial Optimization*. Cambridge Texts in Applied Mathematics. Cambridge University Press, 2004.
29. J. Madhavan, P. A. Bernstein, and E. Rahm. Generic schema matching with cupid. In *The VLDB Journal*, pages 49–58, 2001.
30. D. L. McGuinness, R. Fikes, J. Rice, and S. Wilder. An environment for merging and testing large ontologies. In *Proc. 17th Intl. Conf. on Principles of Knowledge Representation and Reasoning (KR'2000)*, pages 483–493, Colorado,USA, April 2000.
31. S. Melnik, H. Garcia-Molina, and E. Rahm. Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In *Data Engineering, 2002. Proceedings. 18th International Conference on*, pages 117–128. IEEE, 2002.
32. G.A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.
33. M. Nagy, M. Vargas-Vera, and E. Motta. Dssim-ontology mapping with uncertainty. In P. Shvaiko, J. Euzenat, N. F. Noy, H. Stuckenschmidt, V. R. Benjamins, and M. Uschold, editors, *Ontology Matching Workshop at ISWC 2006*, volume 225 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2006.
34. D. Ngo, Z. Bellahsene, and K. Todorov. Opening the black box of ontology matching. In *The Semantic Web: Semantics and Big Data*, pages 16–30. Springer, 2013.
35. H. Nottelmann and U. Straccia. A probabilistic, logic-based framework for automated web directory alignment. *Soft Computing in Ontologies and Semantic Web*, pages 47–77, 2006.
36. N. Noy and M. Musen. Anchor-prompt: Using non-local context for semantic matching. In *Workshop on Ontologies and Information Sharing at IJCAI*, pages 63–70, 2001.
37. R. Pan, Z. Ding, Y. Yu, and Y. Peng. A bayesian network approach to ontology mapping. *The Semantic Web–ISWC 2005*, pages 563–577, 2005.
38. C. Reynaud and B. Safar. Exploiting wordnet as background knowledge. In *ISWC07 Ontology Matching (OM-07) Workshop*, 2007.
39. C. Rosse, J.L.V. Mejino, et al. A reference ontology for biomedical informatics: the foundational model of anatomy. *Journal of biomedical informatics*, 36(6):478–500,

2003.
40. B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1), 2008.
41. S. J. Russell and P. Norvig. *Artificial Intelligence - A Modern Approach (3. internat. ed.)*. Pearson Education, 2010.
42. E. Sanchez and T. Yamanoi. Fuzzy ontologies for the semantic web. *Flexible Query Answering Systems*, pages 691–699, 2006.
43. P. Shvaiko and J. Euzenat. Ontology matching: State of the art and future challenges. *IEEE TKDE*, 99, 2011.
44. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1349–1380, 2000.
45. J.R. Smith and S.F. Chang. Large-scale concept ontology for multimedia. *IEEE Multimedia*, 13(3):86 – 91, 2006.
46. U. Straccia. Towards a fuzzy description logic for the semantic web (preliminary report). In Asuncin Gomez-Perez and Jrme Euzenat, editors, *The Semantic Web: Research and Applications*, volume 3532 of *Lecture Notes in Computer Science*, pages 73–123. Springer Berlin / Heidelberg, 2005.
47. H. Stuckenschmidt. Approximate information filtering on the semantic web. In M. Jarke, G. Lakemeyer, and J. Koehler, editors, *KI 2002*, volume 2479 of *LNCS*, pages 195–228. Springer Berlin / Heidelberg, 2002.
48. G. Stumme and A. Maedche. Fca-merge: Bottom-up merging of ontologies. In *IJCAI*, pages 225–230, 2001.
49. K. Todorov, P. Geibel, and C. Hudelot. Building a fuzzy knowledge body for integrating domain ontologies. In F. Bobillo, R. N. Carvalho, P. da Costa, C. d'Amato, N. Fanizzi, K. B. Laskey, T. Lukasiewicz, T. Martin, and M. Nickles, editors, *URSW*, volume 778 of *CEUR Workshop Proceedings*, pages 3–14. CEUR-WS.org, 2011.
50. K. Todorov, P. Geibel, and C. Hudelot. A framework for a fuzzy matching between multiple domain ontologies. In A. König, A. Dengel, K. Hinkelmann, K. Kise, R. J. Howlett, and L. C. Jain, editors, *KES (1)*, volume 6881 of *Lecture Notes in Computer Science*, pages 538–547. Springer, 2011.
51. K. Todorov, P. Geibel, and K.-U. Kühnberger. Mining concept similarities for heterogeneous ontologies. In P. Perner, editor, *Advances in Data Mining. Applications and Theoretical Aspects*, volume 6171 of *LNCS*, pages 86–100. Springer Berlin / Heidelberg, 2010.
52. K. Todorov, N. James, and C. Hudelot. Multimedia Ontology Matching by Using Visual and Textual Modalities. *Multimedia Tools and Applications*, pages 1–25, 2011.
53. Y. Wang, W. Liu, and D. Bell. Combining uncertain outputs from multiple ontology matchers. *Scalable Uncertainty Management*, pages 201–214, 2007.
54. B. Xu, D. Kang, J. Lu, Y. Li, and J. Jiang. Mapping fuzzy concepts between fuzzy ontologies. In R. Khosla, R. J. Howlett, and L. C. Jain, editors, *Knowledge-Based Intelligent Information and Engineering Systems*, volume 3683 of *Lecture Notes in Computer Science*, pages 199–205. Springer Berlin / Heidelberg, 2005.
55. L.A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338 – 353, 1965.
56. Patric R.J. stergrd and Patric R. J. A fast algorithm for the maximum clique problem. *Discrete Appl. Math*, 120:197–207.

Undefined 1 (2016) 1–5
IOS Press

RDF Dataset Profiling - a Survey of Features, Methods, Applications and Vocabularies

Mohamed Ben Ellefi [a], Zohra Bellahsene [a], John G. Breslin [b], Elena Demidova [c], Stefan Dietze [c], Julian Szymański [d] and Konstantin Todorov [a]

[a] *LIRMM, University of Montpellier and CNRS, Montpellier, France,*
E-mail: {benellefi, bella, todorov}@lirmm.fr
[b] *ENG-3047, Engineering NUI Galway, Galway City, Ireland*
E-mail: breslin@ieee.org
[c] *L3S Research Center, Appelstr. 9a, 30167 Hannover, Germany*
E-mail: {demidova, dietze}@L3S.de
[d] *Gdańsk University of Technology, Poland*
E-mail: julian.szymanski@eti.pg.gda.pl

Abstract. The Web of Data, and in particular Linked Data, has seen tremendous growth over the past years. However, reuse and take-up of these rich data sources is often limited and focused on a few well-known and established RDF datasets. This can be partially attributed to the lack of reliable and up-to-date information about the characteristics of available datasets. While RDF datasets vary heavily with respect to the features related to quality, coverage, dynamics and currency, reliable information about such features is essential to enable dataset discovery in tasks such as entity linking, distributed query, search or question answering. Even though there exists a wealth of works contributing to the problem of dataset profiling in general, these works are spread across a wide range of communities. In this survey, we provide a first comprehensive survey of the RDF dataset profile features, methods, tools and vocabularies. We organize these building blocks of dataset profiling in a taxonomy and emphasize the links between the dataset profiling and feature extraction approaches and several application domains. The survey is aimed towards data practitioners, data providers and scientists, spanning a large range of communities and drawing from different fields such as dataset profiling, assessment, summarization and characterization. Ultimately, this work is intended to facilitate the reader to identify and locate the relevant features for building a dataset profile for intended applications together with the tools capable of extracting these features from the data.

Keywords: Linked Data assessment, RDF dataset profiling, Dataset features, Dataset profile vocabularies

1. Introduction

The Web of Data, and in particular Linked Data [10], has seen tremendous growth over the past years, leading up to the availability of a large amount of RDF datasets[1] on the Web, where a recent crawl[2] of linked datasets retrieved over 1000 datasets alone, including over 8 million explicit resources and an estimated 100 billion triples [66]. RDF datasets and their inherent subgraphs vary heavily with respect to their size, topic and domain coverage, the resource types and schemas as well as the dynamics and currency.

To this extent, the discovery of suitable RDF datasets, which satisfy specific criteria, has become a challenging problem for a variety of applications including *entity linking*, *entity retrieval*, *distributed search* and *query federation*, just to name a few examples. This prevalent problem is underlined by the strong bias towards using established and well-known reference

[1] For readability, we use the terms “RDF dataset" and “dataset" interchangeably within this survey.

[2] `http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/`

knowledge graphs such as DBpedia [4], YAGO [71] or Wikidata[3], although there exists a long tail of potentially suitable domain-specific yet under-recognized datasets.

We begin by providing definitions of several central concepts of our study. In this survey, *an RDF dataset* is defined in accordance with the dataset definition in the VoID Vocabulary[4] stating: "A dataset is a set of RDF triples that are published, maintained or aggregated by a single provider"[5]. According to VoID, this definition reflects the social dimension, such that a dataset represents a meaningful collection of triples as envisioned by its provider, such that this dataset would benefit from descriptive metadata.

A Dataset Profile Feature is a metadata element describing a certain attribute of the dataset. For instance, "dataset dynamicity" is a dataset profile feature providing information on the temporal variation of the dataset. Descriptive metadata consisting of a collection of dataset profile features constitute a dataset profile. A dataset profile is a substantial building block in facilitating effective application-oriented dataset discovery and usage.

An RDF Dataset Profile is a formal representation of a set of dataset profile features.

A dataset profile characterizes the dataset and aids dataset discovery, recommendation and comparison with regard to the represented characteristics. A dataset profile is extensible with respect to the features it contains. Usually, the relevant feature set is application-oriented and depends on the envisaged application scenarios.

A number of popular dataset registries have emerged, which tackle the problem of dataset discovery through the curation of light-weight dataset descriptions, often also exposing structured metadata according to the state-of-the-art vocabularies such as DCAT[6] or VoID. Popular examples include DataHub[7] or DataCite[8], while the LinkedUp Catalog[9] represents a domain-specific example. However, while such metadata is usually edited and curated manually, it is often sparse, not in sync with the constant evolution of the actual datasets and prone to errors.

[3] https://www.wikidata.org
[4] http://vocab.deri.ie/void
[5] See: http://www.w3.org/TR/void/#dataset
[6] http://www.w3.org/TR/vocab-dcat/
[7] http://www.datahub.io
[8] https://www.datacite.org/
[9] http://data.linkededucation.org/linkedup/catalog/

On the one hand, as the Web of Data as a whole is evolving along with the constant evolution of individual datasets, manual assessment and representation of a large variety of dataset features is neither feasible nor sustainable. On the other hand, a wide variety of competing as well as complementary approaches exist, aimed at automatic assessment and description of arbitrary datasets. This body of work is spanning several research communities and includes works in fields such as *dataset characterisation*, *data summarisation*, *dataset assessment* or *dataset profiling*. While this problem is of particular importance in the context of Linked Data, it has been identified and approached already in related fields, such as general database and data management research. Emerging from the aforementioned works, a wealth of tools, methods, vocabularies and applications for assessing, describing and profiling of datasets has become available throughout the past years, where a comprehensive overview and classification is still missing. A myriad of terms and notions does co-exist, whereas a clear distinction, classification and comparison is still required. Only recently, first efforts [24] have been made to bring together such disparate yet closely related fields.

The aim of this survey is to provide researchers, dataset providers and application developers with an overview of *dataset profiling* and closely related approaches, including *dataset profile features*, *feature extraction methods and tools*, *vocabularies* and *applications* to encourage experimentation and facilitate broader use of RDF datasets. Being the first comprehensive study in this area, we provide a thorough analysis and definition of related terms and typical dataset profiling features. Furthermore, we provide a systematic study of the available methods and tools for assessing and profiling structured datasets and survey state-of-the-art vocabularies for representing structured dataset profiles. While some of the discussed works are dedicated to profiling of graph-based RDF datasets in particular, works of relevance from other related fields are also discussed. It should be noted that the authors are aware that domain-specific approaches to profile and annotate datasets exist. However, to ensure high relevance and applicability, this survey addresses exclusively cross-domain approaches, which are agnostic to the domain of the profiled data.

In summary, in this survey we provide the following contributions:

- a taxonomy of dataset profile features, including semantic, qualitative, statistical and temporal feature categories;

- a systematic overview of dataset profile feature extraction approaches and tools discussed in the context of our dataset profile feature taxonomy;
- an overview and a classification of available vocabularies for representing dataset features and profiles;
- an illustration of the use of dataset profiles in several application scenarios.

The remainder of the survey is organized as follows: In Section 2, we present the adopted methodology to collect and organise the publications included in this survey. Next, we provide a comprehensive set of commonly investigated dataset features (Section 3), based on the existing literature in the field of dataset profiling and organize these features in a taxonomy. Then, we provide an overview of the existing approaches and tools for the automatic extraction of dataset profile features (Section 4). Following that we provide an overview of the existing RDF vocabularies for the representation of certain dataset profiles and features (Section 5). Where feasible, we also provide suggestions on the vocabulary use and offer vocabulary recommendations suitable for representing particular dataset profile features. Then, we close the circle by exemplifying subsets of features that are considered relevant in selected application scenarios in Section 6. Finally, we provide a conclusion in Section 7.

2. Survey Procedure

In this section, we present the procedure that we adopted to retrieve and filter journal articles and conference papers for this survey. The stages of the survey process are depicted in Fig. 1 and described in the following.

2.1. Terminology and Taxonomy

We began by identifying a basic terminology of dataset profile features from which we extracted potential terms that were most relevant for this systematic review, such as: profiling, dynamicity, quality, index, etc. Terms were defined and embedded into a taxonomy, which guided the overall study. The taxonomy was iteratively refined throughout the process. During the review process, we updated the taxonomy and consequently further modified the keywords by both including or excluding relevant features.

2.2. Digital Libraries (/Search Engines) Search

The extracted terms from the taxonomy were used individually and in combination to query different online databases and several search engines (cf. Fig. 1). For example, we used keywords and multiword expressions to build the following combinations: {Semantic Web, Linked Data, Linked Open Data (LOD), etc.} $\times$ {profiling, dynamicity, quality, index, etc.}.

2.3. Literature Review

Each category of the dataset profile features taxonomy covers a large range of works in the Semantic Web field and can be surveyed in a separate paper. In this article, we provide a pivotal guide for readers to obtain a global view on the various dataset profile features illustrated by examples. For this purpose, we focused our review on the existent surveys in each category of the dataset profile taxonomy, while providing some examples for: *(i)* the identification of the feature extraction methods (cf. Section 4), *(ii)* the identification of vocabularies for datset profiles representation (cf. Section 5), and *(iii)* the identification of some application-driven profiles (cf. Section 6). Of all the criteria considered, this was the one that produced the sharpest cut down on the number of the articles to be reviewed in detail.

2.4. Paper Selection and Review

By applying a careful review and paper comparison, we obtained a final list of 85 papers to be included in this survey ranging from 1996 to 2016 with about 70% of articles originating from $[2010-2016]$. The selected works are retrieved from different journals, conferences and workshops, mainly in the Semantic Web field as follows:

Journals

- Semantic Web Journal (SWJ)
- Information Processing and Management (IPM)
- ACM Computing Surveys (CSUR)
- Journal of Web Semantics (JWS)
- Australasian medical journal (AMJ)
- FnT Technology, Information and Operations Management (FnT)
- Transactions of the Association for Computational Linguistics (TACL)
- International Journal on Semantic Web and Information

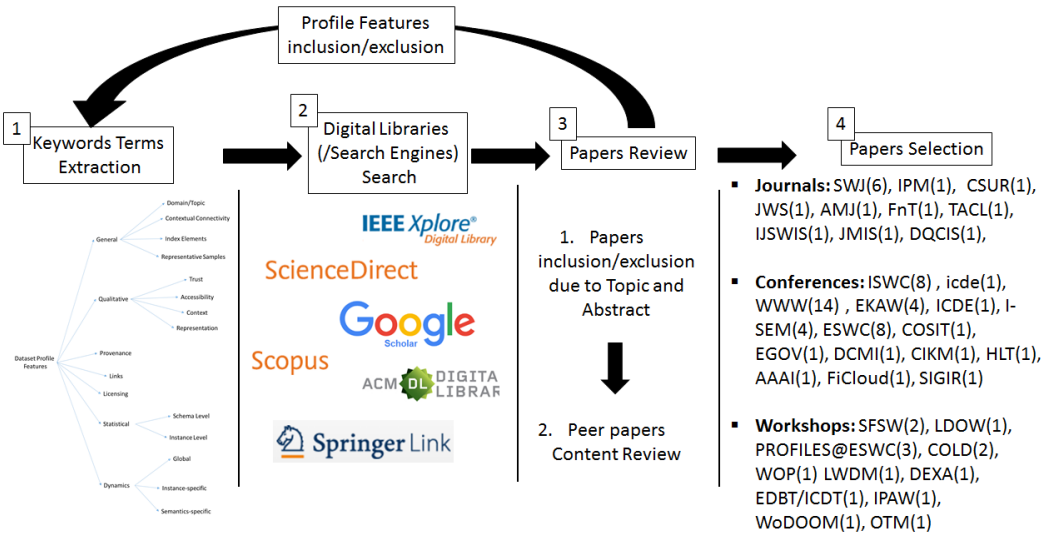


Fig. 1. Survey methodology workflow.

– Systems (IJSWIS)
– Journal of Management Information Systems (JMIS)
– Data Quality in Cooperative Information
– Systems (DQCIS)

Conferences

– International Semantic Web Conference (ISWC)
– International World Wide Web Conference(WWW)
– International Conference on Knowledge Engineering and Knowledge Management (EKAW)
– IEEE International Conference on Data Engineering (ICDE)
– I-Semantics (I-Sem)
– European Semantic Web Conference (ESWC)
– Conference on Spatial Information Theory (COSIT)
– International Conference on eDemocracy and eGovernment (EGOV)
– Dublin Core Metadata Initiative Conference (DCMI)
– Conference on Information and Knowledge Management (CIKM)
– Human Language Technology Conference (HLT)
– Association for the Advancement of Artificial Intelligence (AAAI)
– The IEEE International Conference on Future Internet of Things and Cloud (FiCloud)
– The International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)

To sum up, in this article, we intend to give the reader a bird's-eye view on the RDF datasets profiling problem (whether or not referred to explicitly by using this term) while providing some examples of worm's-eye view especially in terms of feature extraction methods, application-driven profiles and vocabularies for dataset profiles representation.

3. Dataset Profiling Features and Taxonomy

This section provides an inventory of dataset features of relevance to dataset profiling. Features are derived from the literature, in particular, from available dataset profiling methods and vocabularies presented in the following sections. Identified features are clustered and arranged in a feature taxonomy, which provides a categorisation system for the purpose of this survey. We would like to highlight that this taxonomy is extensible and provides one of several feasible ways to categorise profiling features.

In particular, based on an extensive literature overview, we propose to organise features into seven categories: *general*, *quality*, *provenance*, *licensing*, *links*, *statistical* and *dynamics*. This categorisation mirrors the profiling vocabularies distribution, as described in Section 5.

Fig. 2 depicts the resulting taxonomy including references to instances of feature extraction systems. Although we do not discuss the measurements for the different dataset features in detail within this survey, they partially follow from the definition of a particular feature (e.g. in case of statistical features) or have been

extensively discussed in the literature (e.g. qualitative features in [86]).

3.1. General Features

General features are dataset profile features carrying high-level semantic information that do not fit to any of the more specific categories defined in this survey.

1. **Domain/Topic** A domain refers to the field of life or knowledge that the dataset treats (e.g., music, people). It describes and englobes the topics covered by a dataset (e.g., life sciences or media), understood as more granular, structured metadata descriptions of a dataset, as the one found in [29].
2. **Contextual Connectivity**
   We identify two members of this group:
   (a) ***connectivity properties***, meaning the set of entities shared with other datasets, and
   (b) ***domain/topical overlap with other datasets***. Important information, especially with regard to user queries, can be made available by the overlap of the domains or topics covered by a dataset and other datasets. This overlap can be expressed, for instance, by the presence of shared topics between two datasets [79].
3. **Index Elements** Index models have been introduced in order to retrieve information from the LOD graph. An index is defined as a set of key elements (e.g., types), which are used to lookup and retrieve RDF data items. A dataset, therefore, can be inversely described by the set of index elements that are pointing to it in a given index or a set of indices. In that sense, a set of index elements is viewed as a descriptive general dataset feature. These elements can be defined at the schema level (e.g., [49]) or at the instance level (e.g., [39]).
4. **Representative Samples** This group of features is found at the schema and at the instance level. On the one hand, representative schema elements can be understood as: *(i)* the most descriptive set of types (schema concepts) [27], or *(ii)* the set of schema properties that can be used as a keys (almost keys) in instance identification. On the other hand, representative instances are understood as a group of selected data that accurately portrays the whole dataset [26].

3.2. Qualitative Features

The study of data quality has a strong and on-going tradition in the computer science community at large and particularly in the Web Data domain. According to [81], data quality is generally conceived as *fitness for use*, i.e., the capability of data to respond to the demands of a specific user given a specific use case. Data quality has multiple dimensions, and many of them cannot be evaluated in a task-independent manner.

In the context of Linked Data, Bizer *et al.* [9] classified the data quality metrics into three groups according to the type of information that is used as a quality dimension: (i) Content-based metrics – analyzing the information content or compare information with related information; (ii) Context-based metrics – employing meta-information about the information content and the circumstances in which information was claimed; and (iii) Rating-based metrics – relying on explicit ratings about information itself, information sources, or information providers. Zaveri *et al.* [86] identified further dimensions and reorganized the quality dimension into four groups: *(i) Accessibility; (ii) Intrinsic; (iii) Contextual; and (iv) Representational*. Yet another approach of metadata quality assessment can be found in [77] that monitors and assesses the quality of 82 active Open Data portals classified in six groups: *retrievability, usage, completeness, accuracy, openness and contactability*.

In this work, we collected commonly used quality features and re-ordered them in a manner that matches the global dataset profile features taxonomy that we introduce giving rise to the following groups of quality features: *(1) Trust; (2) Accessibility; (3) Context; (4) Degree of connectivity; and (5) Representation*.

1. **Trust** Trust is a major concern when dealing with Linked Data. Data trustworthiness can be expressed by the following features.
   (a) ***verifiability:*** the “degree and ease with which the information can be checked for correctness", according to [8].
   (b) ***believability:*** the “degree to which the information is accepted to be correct, true, real and credible" [64]. This can be verified by the presence of the provider/contributor in a list of trusted providers.
   (c) ***reputation:*** a judgement made by a user to determine the integrity of a source [86]. Two aspects are to take into consideration:

i. ***reputation of the data publisher:*** a score coming from a survey in a community that determines the reputation of a source; and
ii. ***reputation of the dataset:*** scoring the dataset on the basis of the references to it on the Web.

2. **Accessibility** This family of features regards various aspects of the process of accessing the data.
   (a) ***availability:*** an extent to which information is available and easily accessible or retrievable [8].
   (b) ***security:*** refers to the degree to which information is passed securely from users to the information source and back [86].
   (c) ***performance:*** the response time in query execution [86].
   (d) ***versatility of access:*** a measure of the provision of alternative access methods to a dataset [86].
3. **Representativity** The features included in this group provide information in terms of noisiness, redundancy or missing information in a given dataset.
   (a) ***completeness:*** the degree to which all required information regarding schema, properties and interlinking is present in a given dataset [86]. In the Linked Data context, the following sub-features are defined in [8]:
      i. ***schema completeness (/ontology completeness)*** – the degree to which the classes and properties of a schema are represented in the dataset.
      ii. ***property completeness*** – measure of the missing values for a specific property.
      iii. ***population completeness*** – the percentage of all real-world objects of a particular type that are represented in the dataset.
      iv. ***interlinking completeness*** – refers to the degree to which links are missing in a dataset.
   (b) ***understandability:*** refers to expression, or, as defined by [64], the extent to which data is easily comprehended.
   (c) ***accuracy / correctness:*** the equivalence between an instance value in a dataset and the actual real-world value corresponding to that instance.
   (d) ***conciseness:*** the degree of redundancy of the information contained in a dataset.
   (e) ***consistency:*** the presence of contradictory information.
   (f) ***versatility:*** whether data is available in different serialization formats, or in different formal and/or natural languages.
4. **Context/task specificity** This category comprises features that tell something about data quality with respect to a specific task.
   (a) ***relevance:*** the degree to which the data needed for a specific task is appropriate (applicable and helpful) [64], or the importance of data to the user query [8].
   (b) ***sufficiency:*** the availability of enough data for a particular task ([8] uses the term “amount-of-data").
   (c) ***timeliness:*** the availability of timely information in a dataset with regard to a given application.

3.3. Statistical Features

This group of features comprises a set of statistical features, such as size and coverage or average number of triples, property co-occurrence, etc.

1. **Schema-level** According to the schema, we can compute statistical features such as *class / properties usage count*, *class / properties usage per subject and per object* or *class / properties hierarchy depth*.
2. **Instance-level** Features at the instance level are computed according to the data instances only, i.e. *URI usage per subject (/object), triples having a resource (/blanks) as subject (/object), triples with literals, min(/max/avg.) per data type (integer / float / time, etc.), number of internal and external links, number of ingoing (/outgoing) links per instance, number of used languages per literal, classes distribution as subject (/object) per property, property co-occurrence.*

3.4. Dynamics Features

This class of features concerns the dynamicity of a dataset. In principle, every dataset feature can be dynamic, i.e. changing over time (take for example data quality). Inversely, the dynamics of a dataset can be seen as a feature of, for example, quality. For that

reason, this family of features is seen as transversal (spanning over the three groups of features described above).

1. **Global**
   (a) ***lifespan:*** measured on an entire dataset or parts of it.
   (b) ***stability:*** an aggregation measure of the dynamics of all dataset features.
   (c) ***update history:*** a feature with multiple dimensions regarding the dataset update behavior, divided into:
      i. ***frequency of change:*** the frequency of updating a dataset, regardless to the kind of update.
      ii. ***change patterns:*** the existence and kinds of categories of updates, or change behavior.
      iii. ***degree of change:*** to what extent the performed updates impact the overall state of the dataset.
      iv. ***change triggers:*** the cause or origin of the update as well as the propagation effect reinforced by the links.
2. **Instance-specific**
   (a) ***growth rate:*** the level of growth of a dataset in terms of data instances.
   (b) ***stability of URIs:*** the level of stability of URIs i.e. a URI can be moved, modified or removed.
   (c) ***stability of links:*** the level of broken links between resources, i.e. a link is considered as broken if the a target URI changes [65]. Whereas the stability of URIs is rated with respect to the source dataset, the stability of links/backlinks is rated with respect to the stability of the linked URIs in other linked datasets.
3. **Semantics-specific** [36] [25]
   (a) ***structural changes:*** evaluation of the degree of change in the structure (internal or external) of a dataset.
   (b) ***domain-dependent changes:*** this feature reflects the dynamics across different domains that impacts the data.
   (c) ***vocabulary-dependent changes:*** a measure of the dynamics of vocabulary usage.
   (d) ***vocabulary changes:*** a measure of the impact of a change in a vocabulary to the dataset that uses it.
   (e) ***stability of index models:*** the level of change in the original data after the data has been indexed.

3.5. Orthogonal Features

Here, we draw the reader's attention to the fact that some quality features may be orthogonal in the distribution of profiles features, notably to general categories. As orthogonal profile features we consider licensing, provenance and links, described as follows:

1. **Licensing** Here, we adopt the recommendation of Heath *et al.* [45]; "in order to enable information consumers to use your data under clear legal terms, each RDF document should contain a license under which the content can be used". In other words, the type of license under which a dataset is published indicates whether reproduction, distribution, modification, redistribution are permitted. This can have a direct impact on data quality, both in terms of trust and accessibility. Hence, the importance of the existence of license in both human-readable and machine-readable profiles (i.e, including the description in a license vocabulary cf. Section 5.7).
2. **Provenance** the contextual metadata that provides indicators about timelines, currency and update cycles of datasets, which are necessary to know the origin of data, trace errors and notably establish trust. Hence, the provenance is a profile feature used to determine the believability of a dataset. An example use case scenario is to determine some trust score for SPARQL query results in a data sharing triple-store with different provenance.
3. **Links** Links here is understood as the number of datasets, with which a dataset is interlinked, or as the number of triples in which either the subject or the object come from another dataset. Two datasets can be linked through: (i) explicit links when they have linked instances, for example using *owl:sameAs*[10] when sharing identical instances, and (ii) implicit links when sharing topic profiles or context profiles, where explicit links

[10] `http://www.w3.org/2002/07/owl#sameAs`

like *rdfs:seeAlso*[11] can be also used. Dataset links feature covers both schema-level and instance-level representation of links in a dataset profile.

4. Dataset Profiling and Feature Extraction Methods

The field of dataset profiling and features extraction comprises a broad range of tools that is too large to cover here. In this section, we provide examples of relevant dataset profiling approaches for each category of features, as introduced in the previous section. An overview of the dataset profile features categories and the corresponding extraction approaches is given in Fig. 2 and described below in detail.

4.1. Semantic Features Extraction

Semantic features presented in Section 3.1 include domain/topic, context, index elements and representative schema/instances. In the following we present a selection of tools that support feature extraction in this category.

FluidOps Data Portal [79] is a framework for source contextualization. It allows the users to explore the space of a given source, i.e. search and discover data sources of by topics in `http://data.fluidops.net/resource/Topics`. Here, the contextualization engine favors the discovery of relevant sources during exploration. For this, entities are extracted/clustered to give for every source a ranked list of contextualization sources. This approach is based on well-known data mining strategies and does not require schema information or data adhering to a particular form. The FluidOps Data Portal tool enables the retrieval of the "Context" features.

Linked Data Observatory [29] provides an explorative way to browse and search through existing datasets in the LOD Cloud according to the topics they cover. By deploying entity recognition, sampling and ranking techniques, the Linked Data Observatory allows to find datasets providing data for a given set of topics or to discover datasets covering similar fields. This Structured Dataset Topic Profiles are represented in RDF using the VoID vocabulary in tandem with the Vocabulary of Links (VoL) (the vocabularies will be reviewed in Section 5 in more detail). The Linked Data Observatory allows the extraction of the "Domain/Topic" dataset profile features.

voiDge is a tool that automatically generates VoID descriptions for large datasets. This tool allows users to compute various VoID information and statistics on dumps of LOD as illustrated in [14]. Additionally, the tool identifies (sub)datasets and annotates the derived subsets according to the VoID specification. The voiDge describes the "Schema/Instances" dataset profile features.

The keys discovery approaches aim at selecting the smallest set of relevant predicates representing the RDF dataset within the context of link discovery. In other words, a key represents a set of schema properties that uniquely identifies every instance of a given schema concept. We cite two main keys discovery approaches: ***(i)*** *SAKey* [73] – an approach to discover *almost keys* in datasets where erroneous data or duplicates exist. *SAKey* is an extension of *KD2R* [74], which aims to derive exact composite keys from a set of non keys discovered on RDF data sources. ***(ii)*** *ROCKER* – [70] key discovery approach that uses a refinement operator. This operator is able to detect sets of properties that describe any instance of a given class in a unique manner. Reportedly, *ROCKER* is more suited to large scale data than *SAKey*. Keys can be seen as a "Representative Schema/Instances" dataset profile feature.

RDF QTree structure [39] is an approximate multidimensional indexing structure to store descriptions of the content of RDF data sources. A Qtree is a combination of histograms and an R-tree multidimensional structure. The method identifies relevant RDF data sources for a given query that incorporates **instance-level** information by adding triples to the corresponding buckets in the QTree. The QTree structure allows the extraction of the "Index Elements" dataset profile feature.

SchemEX [49] is a stream-based indexing and schema extraction approach over Linked Data. The schema extraction abstracts RDF instances to RDF schema concepts that represent instances with the same properties. The index is each schema concept that maps to the data sources containing instances with the corresponding properties. While SchemEX provides different index structure than the QTree index, both indexing tools involve the "Index Elements" dataset profile feature in the category Semantic Features.

[11] `https://www.w3.org/TR/rdf-schema/#ch_seealso`

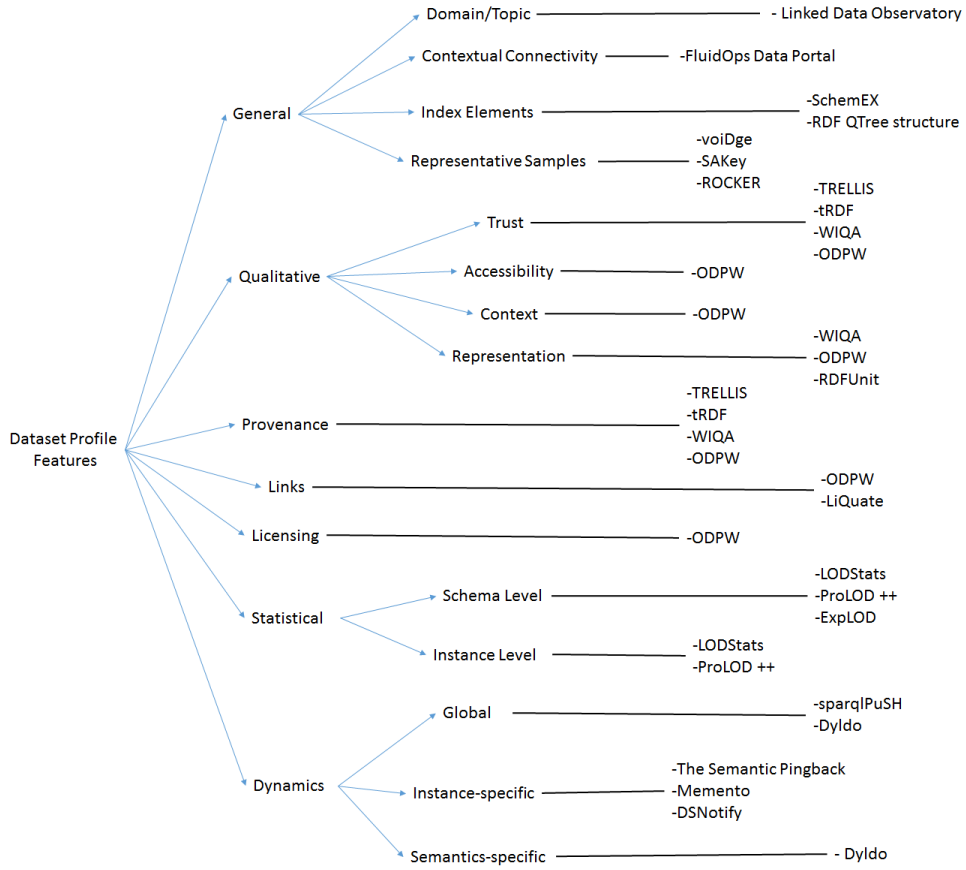


Fig. 2. A taxonomy including dataset profile features organized in general, qualitative, statistical and dynamics categories as well as links to the corresponding feature extraction systems.

4.2. *Quality Features Extraction*

As discussed in Section 3, in this survey we focus on selected groups of quality features such as trust, accessibility, context and representation, most relevant in the context of dataset profiling. In the following we discuss a selection of relevant tools for these groups. Note that a broader overview of the quality assessement approaches in the context of Linked Data in general is provided by Zaveri *et al.* [86], who conducted an extensive survey of 21 works.

TRELLIS [34] is an interactive environment that examines the degree of trust of datasets based on user annotations. The user can provide Trellis with semantic markup of annotations through the interaction with the ACE tool[12] [12]. The tool allows several users to add and store their observations and viewpoints. The annotations made by the users with ACE can be used in TRELLIS to detect conflicting information or handle incomplete information. Trellis provides description for the "Trust" feature in a dataset profile.

tRDF [40] is a framework that provides tools to represent, determine, and manage trust values that represent the trustworthiness of RDF statements and RDF graphs. It contains a query engine for tSPARQL, a trust-aware query language. *tSPARQL* is an extension of the RDF query language SPARQL in two clauses: TRUST AS clause and the ENSURE TRUST clause. The trust values are based on subjective perceptions about the query object. While TRELLIS is based on users annotation, tRDF extracts the "Trust" feature by allowing users to query the dataset and access the trust values associated to the query solutions in a declarative manner.

WIQA [9] is a set of components to evaluate the trust of a dataset using a wide range of different filtering policies based on quality indicators like prove-

[12] Annotation Canonicalization through Expression synthesis.

nance information, ratings, and background information about information providers. This framework is composed of two components: a Named Graph Store for representing information together with quality related meta-information, and an engine, which enables applications to filter information and to retrieve explanations about filtering decisions. WIQA policies are expressed using the WIQA-PL syntax, which is based on the SPARQL query language. WIQA is a generic qualitative tool which can provide description about the "Trust", "Provenance" and the "Representations" dataset profile features.

LiQuate [68] is a tool to assess the quality related to both incompleteness of links, and ambiguities among labels and links. This quality evaluation is based on queries to a Bayesian Network that models RDF data and dependencies among properties. LiQuate enables the retrieval of the "Links" dataset profile features.

RDFUnit [50] is a framework for the data quality that tests RDF knowledge based on Data Quality Test Pattern, DQTP. A pattern can be: ***(i)*** a resource of a specific type should have a certain property, ***(ii)*** a literal value should contain at most one literal for a certain language. The user can select and instantiate existing DQTPs. If the adequate test pattern for a given dataset is not available, the user has to write his own DQTPs, which can then become part of a central library to facilitate later re-use. RDFUnit provides "Representations" dataset profile features in form of DQTPs.

Open Data Portal Watch (ODPW) [77] is a publicly available dashboard component that displays quality metrics for different data portals using various views and charts. These quality metrics are grouped in six dimensions which are retrievability, usage, completeness, accuracy, openness and contactability. The openness indicator provide information to which licenses and file formats conform to the open definition. Furthermore, the watch provides a search service that retrieve the licenses for a given resource URI. ODPW involves all the quality dataset profile features besides the orthogonal features "Links", "Licensing" and the "Provenance".

4.3. Statistical Features Extraction

Statistical features discussed in Section 3.3 comprise schema-level and instance-level statistics.

LODStats [5] is a statement-stream-based tool and framework for gathering comprehensive statistics about datasets adhering RDF. The tool calculates 32 different statistical criteria on LOD such as those covered by the VoID Vocabulary. It computes descriptive statistics such as the frequencies of property usage and datatype usage, the average length of literals, or the number of namespaces appearing at the subject URI position. It is available for integration with CKAN[13] metadata repository, either as a patch or as an external web application using CKAN's API. LODStats provides descriptions for "schema-level" and "instance-level" statistical dataset profile features.

ExpLOD [48] creates usage summaries from RDF graphs including metadata about the structure of an RDF graph, such as the sets of instantiated RDF classes of a resource or the sets of used properties. This structure information is aggregated with statistics like the number of instances per class or the number of property used. ExpLOD provides description about the "schema-level" statistical features for a given dataset.

ProLOD++ [2] is an interactive user interface, which is divided into a cluster tree view and a detailed view. The cluster view enables users to explore the cluster tree and to select a cluster for further investigation for statistics. ProLOD ++ is an extension of *ProLOD* [15], which generates basic statistics. In addition to the mining and the cleansing tasks, the tool generates dataset profiling features related to key analysis, predicate and value distribution, string pattern analysis, link analysis and data type analysis. Hence, ProLOD ++ is a web-based tool, which allows to profile arbitrary LOD datasets in terms of "schema-level" and "instance-level" dataset profile features.

4.4. Temporal Features Extraction

sparqlPuSH [61] is an interface that can be plugged in any SPARQL endpoint and that broadcasts notifications to clients interested in what is happening in the store using the PubSubHubbub[14] protocol [30] i.e. $SPARQL + pubsubhubbub = sparqlPuSH$. Practically, this means that one can be notified in real-time of any change happening in a SPARQL endpoint. A resource can ping a PubSubHubbub hub when it changes, then, the notifications will be broadcasted to interested parties. sparqlPuSH consists in two steps:

[13] http://ckan.org/

[14] PubSubHubbub is a decentralized real-time web protocol that delivers data to subscribers when they become available. Parties (servers) speaking the PubSubHubbub protocol can get near-instant notifications when a topic (resource URL) they're interested in is updated.

(i) register the SPARQL queries related to the updates that must be monitored in an RDF store, *(ii)* broadcast changes when data mapped to these queries are updated in the store. sparqlPuSH extracts "global" dataset profile features in the temporal dataset profile category.

The Semantic Pingback [76] is a mechanism that allows users and publishers of RDF content, of weblog entries or of a scientific article to obtain immediate feedback when other people establish a reference to them or their work, thus facilitating social interactions. It also allows to publish backlinks automatically from the original WebID profile (or other content, e.g. status messages) to comments or references of the WebID (or other content) elsewhere on the Web, thus facilitating timeliness and coherence of datasets. It is based on the advertisement of a lightweight RPC (Remote Procedure Call) service. This system is particularly useful for detecting the stability of links/backlinks. This mechanism provides feedback about "instance-specific" features of a dataset profile.

Memento [23] is a protocol-based time travel that can be used to access archived representations of a resources identified by a given URI. The current representation of a resource is named the *Original Resource*, whereas resources that provide prior representations are named *Mementos*. This system provides relationships like the *first-memento, last-memento, next-memento* and *prev-memento*. These relationships are particulary useful for the extraction of the "instance-specific" features and in particular of the "growth rate" feature. Mementos are available both in HTML and RDF/XML.

DSNotify [65] is a link monitoring and maintenance framework, which attenuates the problem of broken links due to the URI instability. When remote resources are created, removed, changed, updated or moved, the system revises links to these resources accordingly. This system can easily be extended by implementing custom crawlers, feature extractors, and comparison heuristics. DSNotify relates to the "instance-specific" features in the dataset temporal profiling category.

The Dynamic Linked Data Observatory (Dyldo) [47], is a framework to achieve a comprehensive overview of how LOD changes and evolves on the Web. It is an observatory of the dynamicity on the Web of Data over time. The observatory provides weekly crawls of LOD data sources starting from $02/11/2008$ and contains 550K RDF/XML documents with a total of 3.3M unique subjects with 2.8M locally defined entities. The system examines, firstly, the usage of Etag and Last-Modified HTTP header fields, followed by an analysis of the various dynamic aspects of a dataset (change frequency, change volume, etc). Dyldo provides temporal dataset profile features in terms of both "global" and "semantics-specific" features.

4.5. A Note on Dataset Profiling Methods

Here, we discuss several issues regarding dataset profile extraction methods that we observed in the survey process. We begin by the most sensitive profile representations, the semantic features, which typically require domain knowledge with respect to the content of the dataset. As best practice, we recommend that the semantic category should be provided by the data domain experts (e.g. data providers or maintainers) to ensure high quality of the semantic profile. On the other hand, we consider that qualitative, statistical and temporal profile features would in general require less domain expertise and can be extracted automatically by applications in many cases. Furthermore, we observe an obvious need for more semantic profile extraction tools, notably for the "domain/topic" and "context" features, where only few approaches allow automatic extraction of such profiles features.

Further on the dynamic aspect, in order to facilitate up-to-date dataset profiles, these profiles need to be regenerated periodically, based on the dataset dynamicity. The dataset versioning/archiving also requires versioning/archiving of the corresponding dataset profiles in order to ensure coherence between the dataset snapshots and their profile versions.

Finally, we stress the fact that RDF dataset profiles need to provide representations for both human and machine reading. Hence, in Table 1, we provide an overview of the dataset profiling methods including representation formats. In other words, we check for each method if the extracted profile features are designed for human reading or machine reading. In addition the table provides links to the homepages of each extraction method.

5. Vocabularies for Representation of Dataset Profiles and Features

This section introduces vocabularies for representation of dataset profiles, ranging from general dataset metadata to vocabularies dedicated to one or more of the features introduced in Section 3. Note that general-

Method Name	H/M	Accessibility	Home Page
FluidOps Data Porta	H	O.S.	http://data.fluidops.net
Linked Data Observatory	H/M	Online	http://data-observatory.org/lod-profiles/profile-explorer/
voiDge	H/M	O.S.	https://hpi.de/naumann/projects/btc/btc-2010
SAKey	H	O.S.	https://www.lri.fr/sakey
ROCKER	H/M	O.S.	http://rocker.aksw.org/
RDF QTree structure	H/M	–	(⋆)http://swse.deri.org/index.lighttpd.html
SchemEX	H/M	–	–
TRELLIS	H	O.S.	http://www.isi.edu/ikcap/trellis
tRDF	H/M	O.S.	http://trdf.sourceforge.net/tsparql
WIQA	H/M	O.S.	http://wifo5-03.informatik.uni-mannheim.de/bizer/wiqa
LiQuate	H	Online	http://liquate.ldc.usb.ve
RDFUnit	H/M	O.S.	http://rdfunit.aksw.org
ODPW	H	Online	http://data.wu.ac.at/portalwatch
LODStats	H/M	Online	http://stats.lod2.eu/
ProLOD++	H	Online	https://www.hpi.uni-potsdam.de/naumann/sites/prolod++
PubSubHubbub	M	O.S.	https://github.com/pubsubhubbub/
sparqlPuSH	H/M	O.S.	https://code.google.com/archive/p/sparqlpush/
The Semantic Pingback	M	O.S.	https://aksw.github.io/SemanticPingback/
Memento	H/M	–	(⋆) http://mementoarchive.lanl.gov/
Dyldo	H	Online	http://swse.deri.org/dyldo
DSNotify	M	O.S.	http://www.cibiv.at/~niko/dsnotify

Table 1

Dataset profile features extraction methods: Homepages (⋆ means that the homepage was not available at the time of access); Accessibility that can be Open Source (O.S.) or Online (via SPARQL ENDOINT or via HTTP API, etc.); and Human readability (H) *vs.* machine readability (M).

purpose vocabularies such as Dublin Core[15] often provide useful terms also for dataset-specific metadata, but are not discussed in detail here to ensure sufficient focus on vocabularies of more particular relevance for RDF dataset profiling.

5.1. General Dataset Metadata Vocabularies

A range of vocabularies exist which can be used to provide more general metadata of datasets or ontologies. While the Ontology Metadata Vocabulary (OMV) [43] is aimed at providing descriptive information about ontologies - specifically their creators, contributors, reviewers, and creation/modification dates - here we focus specifically on dataset metadata vocabularies.

The Vocabulary of Interlinked Datasets (VoID) [3] provides a core vocabulary for describing datasets and their links. The schema[16] includes the classes *Dataset*, *DatasetDescription*, *LinkSet*, *TechnicalFeature*. The authors distinct *dataset* from *RDF graph*, where *dataset* refers to "meaningful collection of triples, that deal with a certain topic, originate from a certain source or process, are hosted on a certain server, or are aggregated by a certain custodian." A *LinkSet* is defined as a set of triples, where subject and object are in different datasets/namespaces. The VoID guidelines recommend additional vocabularies (DC-

[15] http://dublincore.org/documents/dces/

[16] http://vocab.deri.ie/void

Category	Datasets (Percent)
Social Web	6 (1.16)
Government	75 (40.32)
Publications	14 (13.46)
Life Sciences	29 (32.58)
User-gen. Content	6 (10.91)
Cross-domain	5 (11.36)
Media	2 (5.41)
Geographic	15 (36.59)
Total	140 (13.46)

Table 2
Adoption of VoID across LOD Datasets per Category (Source: http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/).

Terms, FOAF for general metadata and SCOVO - the Statistical Core Vocabulary[17] for statistical information. VoID is already widely used in the Web of Data, as documented by Table 2, depicting the use of VoID descriptions among the 1014 datasets and per category in the current inventory of the Web of Data[18].

The Data Catalog Vocabulary (DCAT)[19] follows a similar rationale and has been created based on a survey of government data catalogues [53]. Key classes include *Catalog*, *Dataset*, *CatalogRecord* where the latter has a similar scope as the VoID DatasetDescription, i.e., it is making the useful distinction between dataset metadata and metadata of the dataset description (the record) itself. Additional classes include *Distribution* - i.e. the instantiation of particular dataset in a specific access format (e.g., an RDF dump or a SPARQL endpoint). For categorisation of datasets, the *dcterms:subject* predicate and controlled SKOS vocabularies are recommended.

5.2. Dataset Links

Links as important features of Linked Data datasets are represented through a variety of means, covering both schema-level and entity-level links. VoID, for instance, includes specific linksets which can be instantiated to define metadata about dataset's links. *SKOS*[20], the Simple Knowledge Organization System, on the other hand provides a formal vocabulary for defining taxonomic and mapping relations among both concepts and entities and is a well used means to describe links between concepts and entities across datasets. By providing an established vocabulary for less strict relations, for instance, *broader* or *narrower*, respectively *broaderMatch* and *narrowerMatch*, it enables the representation of taxonomic relationships as well as the alignment of different schemas and knowledge bases, i.e. datasets.

A more specific approach is followed by the Vocabulary of Links (VoL)[21], which provides a general vocabulary to describe metadata about links or linksets, within or accross specific datasets. VoL was designed specifically to represent additional metadata about computed links which cannot be expressed with default RDF(S) expressions and enable a qualification of a link or linkset. This includes, for instance, the description of linking scores or linking provenance, for instance, through a specific linking method.

The Expressive and Declarative Ontology Alignment Language (EDOAL)[22] enables the representation of correspondences between entities and concepts in different ontologies beyond mere mapping relationships (equivalence, subsumption). For these reasons, EDOAL introduces formalisms for representing transformations, constructions of complex classes/entities or restrictions to constrain classes/entities. EDOAL in that sense provides the means to on-the-fly interpretation of mapping statements as part of data integration scenarios. On the other hand, in contrast to VoL, there are no means for representation of provenance of mapping statements. Next to being more comprehensive and expressive than SKOS or VoL, another major difference seems to be that the typical use case for generating EDOAL statements is the manual formalisation of mapping statements, while less expressive SKOS and VoL statements can be at least partially generated from the output of automated linking and mapping algorithms.

5.3. Dataset Quality

Early works by Supelar *et al.* in [72] define a set of knowledge quality features applicable for knowledge graphs, respectively ontologies, and a corresponding ontology. Their features are classified into *quantifiable* and *non-quantifiable* characteristics and include

[17] http://purl.org/NET/scovo
[18] http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/
[19] http://www.w3.org/TR/vocab-dcat/
[20] https://www.w3.org/TR/2009/REC-skos-reference-20090818/
[21] http://data.linkededucation.org/vol/index.htm
[22] http://alignapi.gforge.inria.fr/edoal.html

characteristics such as usability, availability, accuracy, or complexity. The suggested ontology, however, only includes a higher level taxonomy, but neither a fully fledged vocabulary for annotation nor a specific set of metrics to quantify the quantifiable metrics.

Fürber *et al.* [33] describe the DQM Ontology[23], a general vocabulary for representing data quality features, to some extent also covering statistical information, such as notions of property completeness or property uniqueness. Key concepts include:

- Data Quality Assessement as an abstract container of scores and metrics describing class/property quality aspects.
- Completeness, derived into Property Completeness - as a measure of the degree to which properties are consistently populated - and Population Completeness as the degree to which all objects of a certain reference are represented in a specific class.
- Accuracy as a notion representing the degree to which a statement captures the intended semantics and syntax (subtypes are Syntactic Accuracy and Semantic Accuracy).
- Uniqueness of properties and entities is introduced to capture the existence of duplicates.
- Timeliness captures the recency of a specific statement/entity.

In addition, the authors introduce a preliminary classification for data quality problems.

In addition, the WIQA - Web Information Quality Assessment Framework[24] describe some early work to filter content according to quality features, also introduce WIQA-PL, a vocabulary for modeling content access policies. However, the work appears to be deprecated and not maintained.

Also worth to mention is the work in [32], where authors use the SPARQL Inferencing Notation (SPIN) - a vocabulary that allows the representation of SPARQL queries - to represent data quality rules.

In addition, the Dataset Quality Vocabulary (daQ)[25] and the Data Quality Vocabulary (DQV)[26] provide complementary terms for annotating DCAT dataset descriptions with quality aspects and metrics. While both vocabularies provide a general framework for annotating quality information and metadata about associated metrics, several concerns about practical issues are raised as part of the DQV working draft documentation.

Finally, while provenance information often provides indicators about timelines, currency and update cycles of datasets, Section 5.6 introduces additional vocabularies of relevance.

5.4. Dataset Dynamics & Evolution

While there does exist a wealth of methods for assessing characteristics related to dynamics and evolution of datasets, as illustrated in earlier sections of this survey, most vocabularies in the area are dedicated to representing the actual evolution of a dataset, rather than higher level observations about dynamics.

The Dataset Dynamics group[27] for instance lists a number of vocabularies for representing dataset changeset and updates. The *Talis Changeset vocabulary*[28] provides some early, yet discontinued work on representing changeset and specific characteristics, and has a similar approach as the Delta vocabulary[29]. The *Triplify Update vocabulary*[30] provides a very simple RDF schema for capturing dataset updates where each *Update* or *UpdateSet* is annotated with provenance information about the updater and the time stamp.

In a similar direction is the recent work of Graube *et al.* [38] on *R43ples*, a revision management approach for RDF datasets using named graphs for capturing revisions and SPARQL for manipulation of the latter. Authors introduce the so-called Revision Management Ontology (RMO) based on PROV-O (cf. 5.6). While RMO implements baseline revision management notions for data graphs, it is of lesser relevance for the purpose of this section.

A more abstract approach is offered by the *Dataset Dynamics (DaDy) Vocabulary*[31], which allows the representation of more abstract dynamics-related observations for a specific dataset. It is specifically foreseen to be used in conjunction with VoID, where a *void:Dataset* is annotated with instantiations of

[23] http://semwebquality.org/dqm-vocabulary/v1/dqm

[24] http://wifo5-03.informatik.uni-mannheim.de/bizer/wiqa/#wiqapl

[25] http://purl.org/eis/vocab/daq

[26] https://www.w3.org/TR/2015/WD-vocab-dqv-20150625/

[27] http://www.w3.org/wiki/DatasetDynamics

[28] http://vocab.org/changeset/schema.html

[29] http://www.w3.org/2004/delta

[30] http://triplify.org/vocabulary/update

[31] http://vocab.deri.ie/dady

dady:UpdateDynamics. The latter captures information about the update regularity and frequency.

For capturing specific features and observation related to dynamics and evolution, beyond the ones covered by the vocabulary above, in particular the vocabularies mentioned in the following section, aimed at representing statistical dataset features, which may or may not be related to dynamics.

5.5. Statistical Dataset Metadata

A range of vocabularies exist, which partially support the representation of dataset statistics and can be used in conjunction with general dataset metadata vocabularies such as VoID or DCAT. These include, for instance, the RDF Data Cube Vocabulary[32], SDMX[33] or SCOVO[34].

The VoID guidelines, for instance, recommend the use of SCOVO to share statistical dataset features [3]. Authors foresee, on the one hand, statistics concerning the whole dataset or linkset, such as triple count, and attributing statistics to a source, to capture where a statistical datum stems from. Inline with some of the authors' concerns about the adequacy of SCOVO, it has been superseded by the Data Cube Vocabulary in the more recent past.

The RDF Data Cube vocabulary[35], currently a W3C Editors Draft developed by the Government Linked Data Working Group[36] is an RDF vocabulary for representing multi-dimensional so-called *data cubes* in RDF. The Data Cube vocabulary describes general statistical notions, such as *dimensions* or *observations*, and as such, can be perceived as a meta-level vocabulary for representing any statistical notion.

While the Data Cube vocabulary builds on SKOS, its Data Cubes approach originates from and is compatible with the cube structure underlying the SDMX (Statistical Data and Metadata eXchange)[37] information model. The latter is an ISO standard, describing an information model for exchanging statistical data and metadata which has been serialised into XML, EDI and recently, RDF. SDMX-RDF[38] can be seen as a natural predecessor of the Data Cube vocabulary which is not a one-to-one representation of SDMX but uses an SDMX subset, plus additional elements, to provide a vocabulary tailored to represent data published as RDF on the Web.

SCOVO[39], also described by Hausenblas *et al.* [44], is an earlier, native RDF vocabulary for statistical data, consisting of three main classes, *Dataset*, *Dimension*, and *Item*. While there exist efforts to merge SCOVO and SDMX-RDF [21], both approaches are superseded by the Data Cube vocabulary, which represents the state of the art in representing statistical data on the Web.

Auer *et al.* present LODStats [6], a framework for dataset analytics, which introduces a set of 32 statistical features and uses the most recommended combination of VoID and the DataCube vocabulary. Links between the Data Cube class qb:Observation and the void:Dataset class are represented using a native property (void-ext:observation). While VoID already represents properties for several statistically described objects (triples, classes, distinctSubjects, etc.), additional features were represented using void:classPartition and void:propertyPartition. While this approach combines the two state of the art vocabularies for general dataset metadata (VoID), respectively statistical data (Data Cube), it turns out to be the most future-proof approach to capture statistical dataset metadata.

5.6. Data and Dataset Provenance

A variety of definitions have been given for provenance over the past number of years. One very pragmatic definition comes from the Provenance Working Group[40] of the W3C, especially when thought of in the context of the Web: *"Provenance is defined as a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing."* On the Web, provenance can pertain to any resource found on the Web - documents, data, or datasets - but it can also be found in a resource that is used to describe the provenance of an object in the real world.

The main aim of the Provenance Working Group was to create standards that could be used to define and work with provenance data. A document from its previous incarnation as an Incubator Group states

[32] http://www.w3.org/TR/vocab-data-cube/
[33] http://sdmx.org
[34] http://vocab.deri.ie/scovo
[35] https://dvcs.w3.org/hg/gld/raw-file/default/data-cube/index.html
[36] http://www.w3.org/2011/gld/
[37] http://sdmx.org/
[38] http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/index.html
[39] http://vocab.deri.ie/scovo
[40] http://www.w3.org/TR/2013/REC-prov-dm-20130430/

the difficulties involved in such standardisation efforts: *"provenance is too broad a term for it to be possible to have one, universal definition - like other related terms such as "process", "accountability", "causality" or "identity", we can argue about their meanings forever (and philosophers have indeed debated concepts such as identity or causality for thousands of years without converging)"*.[41]

A provenance record is essentially a record of metadata that details the entities and processes that were involved in creating, modifying and delivering a resource, be it physical or digital [57]. Such records include details about when an item was created, what were the original sources of information used in its creation, what kind of evolution has the resource undergone (e.g., what were the other entities or processes that may have modified the resulting piece of information). A provenance process is defined by Moreau [56] as "the provenance of a piece of data is the process that led to that piece of data".

We will now describe some of the main provenance models used on the Web, some of which have specific applicability in terms of whole datasets.

1. **voidp** builds on and extends the aforementioned *VoID* linked dataset ontology to describe the provenance relationships of data across linked datasets. Publishers can use a lightweight set of classes and properties to describe the provenance information of data within their linked datasets using voidp. This enables users to find the right data for their tasks based not only on the types of data being sought but also on the origins of that data, e.g., "given a set of attributes and data authorship conditions, which available resources match a desired set of criteria and where can these resources be found?"
2. From the perspective of archiving and long-term preservation of data, the **Data Dictionary for Preservation Metadata (PREMIS)**[42] set of terms can be used to describe the provenance of archived, digital objects (e.g., files, bitstreams, aggregations and datasets), and therefore has applicability in our scenario. It does not provide provenance information for the descriptive metadata for those objects, and therefore one of the other vocabularies can be used for this.
3. Inspired by the notion of changesets in code or document revisions, the **Changeset Vocabulary**[43] consists of a set of terms that can be used to describe changes in the description of a resource. The primary concept is that of a ChangeSet which defines the delta (changes) between versions of a resource description.
4. The **Proof Markup Language (PML)** is used for defining and exchanging proof explanations created by various intelligent systems, including web services, machine learning components, rule engines, theorem provers and task processors. It provides terms for annotating "IdentifiedThings" such as name, description, create date and time, authors, owners, etc. IdentifiedThings are the entities used or processed in an intelligent system, of which a dataset could be one.
5. The **Semantic Web Publishing Vocabulary (SWP)** by [19] makes it possible "to represent the attitude of a legal person to an RDF graph. SWP supports two attitudes: claiming the graph is true and quoting the graph without a comment on its truth. These commitments towards the truth can be used to derive a data publisher's or a data creating entity's relation to provided or created artifacts. Furthermore, the SWP allows to describe digests and digital signatures of RDF graphs and to represent public keys."
6. The **Provenance Vocabulary**[44] was developed to describe provenance of Linked Data on the Web. It is defined as an OWL ontology and it is partitioned into a core ontology and supplementary modules.
7. The **Open Provenance Model (OPM)** is used to describe provenance histories in terms of the processes, artifacts, and agents involved in the creation and modification of a resource. The OPM model was the primary outcome of a series of Provenance Challenge workshops, and is one to which many other provenance vocabularies are mapped to. In fact, it was taken as the basis for the development of PROV-O, described below. Two variants exist, the OPM Vocabulary (OPMV)[45] as a lightweight vocabulary, and the

[41] http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/

[42] http://bit.ly/premisOntology

[43] http://purl.org/vocab/changeset

[44] http://trdf.sourceforge.net/provenance/ns.html

[45] http://purl.org/net/opmv/ns#

OPM Ontology (OPMO)[46] using more advanced OWL constructs.

8. The **PROV Ontology (PROV-O)**[47] was published as a W3C Recommendation in 2013 by the W3C Provenance Working Group to be a new standard ontology for representing provenance. This is part of a larger *PROV* Family of Documents [55] created to support "the widespread publication and use of provenance information of Web documents, data, and resources" – including a Data Model (PROV-DM) [57] and an Ontology (PROV-O) [52] – for provenance interchange on the Web. PROV defines a core data model for provenance for building representations of the entities, people and processes involved in producing a piece of data or any artifact in the world.[48]

As well as the above vocabularies that are specifically designed to facilitate provenance and related primitives, there are a number of commonly-used vocabularies and de-facto standards on the Web that also contain terms of relevance to provenance derivation and definition. These include Dublin Core (DC), Friend-of-a-Friend (FOAF), and Semantically Interlinked Online Communities (SIOC). Some of these terms were highlighted by [41], and we outline these and others below. Since a dataset can be identified by a resource, we can use many of the properties described below with full datasets as well as individual resources or pieces of data in those datasets.

– **Dublin Core**: *dcterms:contributor* and *dcterms:creator* can be used in analyses of the activity of a user in the data creation process, although the type of the user and their role may need to be further specified using other vocabularies. In our case, it could also be used to identify the creator of an entire dataset. *dc:source* describes the source from which a resource or dataset is derived, and therefore has usefulness as a provenance element. *dcterms:created* and *dcterms:modified* can be used to define both the creation of a resource or dataset and the modification of that resource or dataset respectively. *dcterms:publisher* can be used to define the provider of a particular resource or dataset, although as [41] points out the type of publisher is left ambiguous. Finally, Dublin Core also defines a *dcterms:provenance* term which can link a resource to a set of provenance change statements.
– **Friend-of-a-Friend**: *foaf:made* and its inverse functional property (IFP) *foaf:maker* can be used to link a resource or dataset to the *foaf:Agent* (person or machine) who created it. In addition, the *foaf:account* property can be used to link a *foaf:Agent* to a *foaf:OnlineAccount* or *sioc:UserAccount* which in turn can be identified as the means of creation for a resource or dataset (see below).
– **Semantically Interlinked Online Communities**: As with Dublin Core, the properties *sioc:has_creator*, *sioc:has_modifier* (and their IFPs *sioc:creator_of* and *sioc:modifier_of* respectively) can be used to refer to a resource's creators and modifiers (identified by *sioc:UserAccount*s). *sioc:has_owner* and its IFP *sioc:owner_of* indicates who has control over a resource or dataset. *sioc:ip_address* can be used to link the created data and creator if specified to an Internet address. Also, *sioc:last_activity_date* can be used to reference the last activity associated with a resource, although this may still be interpreted in different ways (modified, read, etc.). As with *dc:source*, a *sioc:sibling* can be used to define a new resource (or perhaps a dataset) that is very similar to but differs in some small manner from another one. Finally, *sioc:earlier_version*, *sioc:later_version*, *sioc:next_version* and *sioc:previous_version* can be used to connect versioned artifacts together as one would find in a provenance graph.
– In addition to the "SIOC Core" ontology terms, there are also some SIOC modules which can be used in provenance descriptions for datasets. The most relevant is probably the **SIOC Actions** [20] module, which was designed to represent how users in a community are manipulating the various digital artifacts that constitute the application supporting that community. The main terms in SIOC Actions are *sioca:Action*, *sioca:DigitalArtifact*, *sioca:byproduct*, *sioca:creates*, *sioca:deletes*, *sioca:modifies*, *sioca:object*, *sioca:product*, *sioca:source* and *sioca:uses*. These have been aligned to OPM and PROV-O in recent work by [60].

[46] http://openprovenance.org/model/opmo
[47] http://www.w3.org/TR/prov-o/
[48] http://www.w3.org/TR/2013/NOTE-prov-primer-20130430/

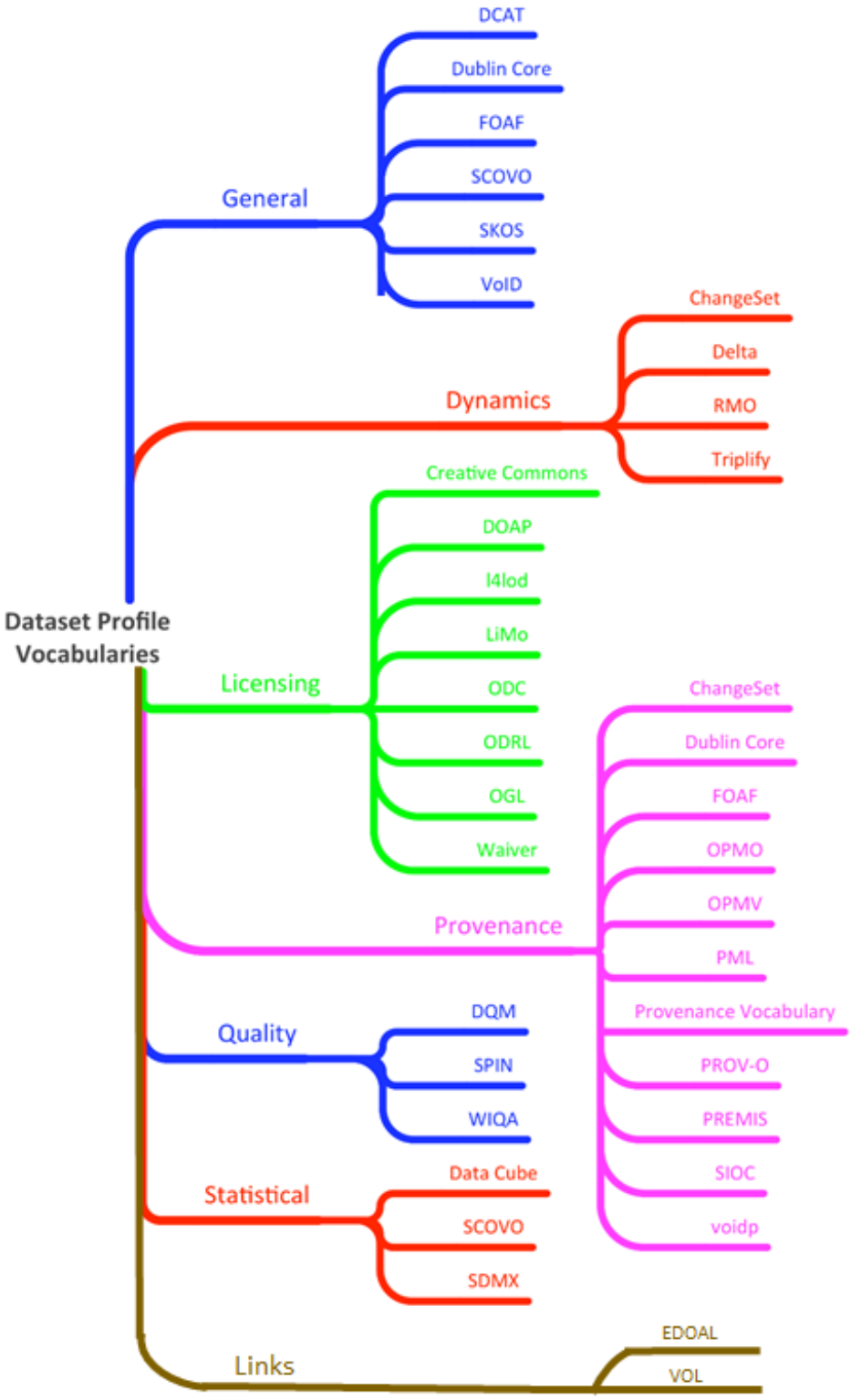


Fig. 3. Overview of relevant vocabularies as classified by type of dataset profile features.

Vocabulary Name	Type	Triples Feb. '15	Datasets Feb. '15	Triples Jan. '17	Datasets Jan. '17
Dublin Core	General, Provenance	21,397,721	154	20,056,611	**213**
FOAF	General, Provenance	3,689,178	117	3,399,261	**190**
SKOS	General	10,581,530	67	5,606,905	**108**
VoID	General	9,754	41	987	**53**
voidp	Provenance	172	21	173	16
SIOC	Provenance	148	16	**6,255**	**45**
DOAP	Licensing	306	14	53	7
Creative Commons	Licensing	16,525	12	83	**21**
Provenance Vocabulary	Provenance	84	12	61	2
Data Cube	Statistical	581,381	10	101,757	**75**
SCOVO	General, Statistical	408	9	399	1
PML	Provenance	259	8	0	0
OPMO	Provenance	63	8	4	1
SDMX	Statistical	285,904	6	90,586	**11**
OPMV	Provenance	4	2	1	1
PROV-O	Provenance	4,537	1	577	**17**
DCAT	General	8	1	**2,010**	**3**
Waiver	Licensing	1	1	0	0
Delta	Dynamics	0	0	0	0
RMO	Dynamics	0	0	0	0
Triplify	Dynamics	0	0	0	0
ChangeSet	Dynamics, Provenance	0	0	0	0
VoL	General	0	0	0	0
l4lod	Licensing	0	0	0	0
LiMo	Licensing	0	0	0	0
ODC	Licensing	0	0	0	0
ODRL	Licensing	0	0	0	0
OGL	Licensing	0	0	0	0
PREMIS	Provenance	0	0	0	0
DQM	Quality	0	0	0	0
SPIN	Quality	0	0	0	0
WIQA	Quality	0	0	0	0

Table 3
Overall usage and dataset counts for the aforementioned vocabularies, sorted by number of datasets in February 2015. Those numbers in **boldface increased in 2017. Statistics were re-checked in January 2017.**

5.7. Dataset Licensing

We will now examine what vocabularies are available to assist with licensing of data and datasets. These include RDF versions of common licensing frameworks and alignments of multiple licensing frameworks into a combined vocabulary.

– **Creative Commons (CC)**[49] is a framework that allows users to define the rights regarding how others can reuse the content that the users themselves have published. It provides various licenses to define if and how people can reuse content that has been published, if they can modify it, and if it may be used for commercial purposes. Creative Commons also allows licensing information

[49] http://creativecommons.org/licenses/by/3.0/

to be expressed in RDF using the ccREL (REL, or rights expression language) vocabulary. Many datasets in the LOD cloud are already licensed under Creative Commons, as we will see later.

- The **Open Data Commons (ODC)** license[50] was originally released by Talis in 2008 as a means to tackle the issue of Creative Commons licenses being applied to non-creative resources such as data and datasets. The ODC "Public Domain Dedication and License" was a fusion of ideas from their earlier Talis Community License and related efforts such as the provision of scientific datasets using Science Commons.
- The **Open Digital Rights Language (ODRL)** vocabulary[51] enables the fine-grained specification of licensing terms (rights, policies, etc.) in a machine-readable format. Developed by the W3C ODRL Community Group, ODRL 2.0[52] uses RDF or JSON, evolving from an earlier XML-based REL version[53].
- **Open Government License (OGL)**[54] is a license produced specifically for Crown copyright works published by the UK government and other public sector bodies. It is aligned to both CC and ODC. One of the dataset projects using OGL is the data.gov.uk service.
- The **License Model (LiMo)**[55] is an ontology for open data and dataset licensing. It links to terms from Dublin Core, VoID, CC and PROV-O, and also defines legal terms, conditions of use and distribution, and other rights. One of the main terms is *limo:LicenseModel* which is equivalent to the *cc:License* concept from Creative Commons.
- **Description of a Project (DOAP)**[56] is an RDF vocabulary that provides a common metadata modelling scheme for describing projects creating software applications, in order to provide a unified way to represent a software project no matter the source. The main class is *Project* which has properties such as its licence, the project's maintainers, the URL for subversion access, etc. Many of the concepts in DOAP could also be re-applied to datasets since they share many of the same properties.
- **Licenses for Linked Open Data (l4lod)**[57] was introduced in [37] to provide an alignment with many of the licensing vocabularies we have just described. It can be used to express a machine-readable composite license for a dataset. l4lod is composed of three deontic components (obligations, permissions and prohibitions) that can be used to reconcile a set of licenses that are associated with heterogeneous datasets whose information items have been returned together for consumption (e.g., via a single SPARQL query).

5.8. Observations

We use the LOD2 Stats service[58] to give us some context as to how often terms from these vocabularies are being used and within how many datasets. These statistics are shown in Table 3, where the type refers to the vocabulary type as per the headings above.[59] While we were unable to filter the instances of dataset profiling-specific terms from our suggested vocabularies while examining their usage statistics in LOD2, we can gain some insight into which ones may be more widely adopted by looking at the existing overall statistics and dataset usages, especially over time (i.e., from 2015 to 2016, we can see which vocabularies are consistently being used and are growing in usage). It is reasonable to assume that users will be more willing to adopt terms from widely-used vocabularies for representing dataset profiles, as long as they are fit for purpose.

251 datasets use RDF syntax, giving us an overall total. From the data in Table 3, we observe that general metadata about the datasets is readily provided, but that more specific information on provenance and statistics using specialised vocabularies is only available in somewhere around 21% (52) and 10% (25) of datasets respectively.

Another observation is that none of the quality or dynamics and evolution vocabularies appear in LOD2 Stats. That points to a significant underutilization

[50] http://opendatacommons.org/licenses/
[51] http://www.w3.org/community/odrl/two/model/
[52] http://w3.org/ns/odrl/2/
[53] http://www.w3.org/TR/odrl/
[54] http://www.nationalarchives.gov.uk/doc/open-government-licence/
[55] http://purl.org/LiMo/0.1
[56] http://usefulinc.com/ns/doap
[57] http://ns.inria.fr/l4lod/
[58] http://stats.lod2.eu/ as accessed on 2nd February 2015 and re-checked again on 19 January 2017
[59] Where multiple entries exist for a vocabulary on LOD2 Stats, we use the numbers from the largest entry rather than adding usage figures together, as modules in a vocabulary may be used together in the same dataset (e.g., DC Terms and DC Elements, or SDMX Dimension and SDMX Measure).

of terms relating to dataset quality, the evolution of a dataset, or the dynamics involved in a changing dataset. The assumption is that dataset creators are more interested in providing the datasets themselves without giving assurances to others who may want to use them about their quality or how they have changed over time.

It does not seem from Table 3 that many datasets are explicitly licensed via some machine-readable form, with just 5% (12) containing Creative Commons metadata. However, according to work by [37], 95% of the datasets in the LOD cloud[60] did indeed express licensing information via the *dcterms:license* or the *dcterms:rights* properties of Dublin Core (albeit in human-readable format). Creative Commons represented 51% of all licenses in their analysis, followed by Open Data Commons at 18%. This points to the need for more explicit license definitions in datasets, with a link to the license type and conditions and not just a simple text string in an attribute field.

6. Application-Driven Dataset Profiles

Dataset profiles are highly important for a wide variety of applications in many domains, including, for example, data linking and curation, schema inference, federated query and search, as well as question answering. In this section, we highlight important applications from these domains that use dataset profiles along with their relevant profile features. Some of these applications can use, verify and update dataset profile features (e.g., including statistical characteristics of datasets) and may in turn generate additional statistics that can become part of the dataset profile. The list of the applications and relevant features presented in this section aims to illustrate the use of dataset profiles by state-of-the-art tools and is not exhaustive.

6.1. Data Linking Applications

Data linking applications aim to annotate, disambiguate and interlink entities and events in text using Natural Language Processing (NLP) techniques and external sources including Linked Data. In this context, popular services include DBpedia Spotlight [22], Illinois Wikifier [67] as well as Babelfy [58].

Example features for data linking applications: Data linking applications typically use the semantic features discussed in Section 3.1 such as topics, domains, languages (versatility) and location coverage, as well as representative parts of schema/instances, and specifically the key candidates extracted with the key discovery approaches described in Section 4.1.

6.2. Data Curation, Cleansing and Maintenance

As linked datasets are often generated from semi-structured or unstructured sources using automated extraction approaches, these datasets vary heavily with respect to quality, currentness and completeness of the contained information [85].

A number of recent works focus on statistical methods for: (1) outlier detection to detect errors in numerical values [31], [63], [82]; (2) automatic prediction of missing types of instances [63]; and (3) the identification of incorrect links between datasets [62]. A further line of research in Linked Data quality is related to the discovery of errors in the data based on existing interlinkings (e.g., [16], [84]). Thereby some works go beyond error detection and attempt to automatically determine correct data values in case of inconsistencies [16]. As mentioned above, additional statistics generated by these approaches that can become part of the dataset profile.

Example features for error detection in numerical values: In [31] the authors detect errors in numerical values using outlier detection. To identify the properties to which numerical outlier detection can be applied, the following statistical characteristics (discussed in Section 3.3) are used: (1) total number of instances, (2) names of the properties used in the dataset, (3) frequency of usage with numerical values in the object position for each property, and (4) total number of distinct numerical values for each property.

Example features for conflict resolution in multilingual DBpedia: The features used in conflict resolution in [16] include provenance metadata at the statement, property and author levels. The temporal dataset profile includes in particular: (1) Recency of the specific statement (measured using the time of the last edit), (2) overall editing frequency of the property in the dataset, and (3) the overall number of edits performed by the specific editor.

6.3. Schema Inference

Many existing Linked Data sources do not explicitly specify schemas, or only provide incomplete specifications. However, many real-world applications (e.g.,

[60] http://lod-cloud.net/

answering queries over distributed data [11]) rely on the schema information. Recently, approaches aimed at the automatic inference of missing schema information have been developed (e.g., [63], [49]).

Example features for type inference: Statistical characteristics of datasets (see Section 3.3) play an important role in type inference applications. For example, in [63] statistics on the completeness of type statements as well as property-specific type distributions are required (i.e., the types of resources appearing in subject and object positions of each property including their frequencies).

6.4. Distributed Query Applications

The Linked Data Cloud can be queried either through direct HTTP URI lookups or using distributed SPARQL endpoints [39] that can include full-text search extensions (see e.g., [1]). Also combinations of both query paradigms are possible [42]. Typically, the first step of query answering over distributed data is the generation of ordered query plans against the mediated schema on a number of data sources [83]; In this step, dataset profiling plays an important role.

In order to guide distributed query processing, existing applications rely on indexes of varying granularity including *Schema-level Indexes* and *Data Summaries*. *Schema-level Indexes* contain information about properties and classes occurring at certain sources. *Data Summaries* use a combined description of instance- and schema-level elements to summarise the content of data sources [39]. The majority of existing federated query approaches for LOD (e.g., [42], [39], [80], [35]) aim to optimize efficient query processing and do not (yet) take the quality parameters of LOD sources into account. Therefore, existing *Data Summaries* mostly contain frequencies and interlinking statistics of varying granularity.

Example features for efficient and quality-aware query applications: The majority of existing query applications rely on semantic and statistical characteristics (see Sections 3.1 and 3.3) at the schema-level, i.e. properties and classes occurring at certain sources for effective query interpretation. In addition, applications that optimize for efficient query processing require data-level statistics (including frequency and interlinking) either on triple level or for each subject, object and predicate individually [39]. Finally, quality-aware query applications also take into account qualitative characteristics (see Section 3.2) (e.g., completeness and accuracy) at different granularity levels. This includes overall data source statistics [59], as well as property-specific [69] and type-specific statistics [83].

6.5. Information Retrieval (IR) Applications

In *IR*, Linked Data is mostly used in the context of semantic search, a typical demonstration of which can be found in [28]. The majority of semantic search applications are domain-oriented; a large number of practical cases have been shown for repositories related to biomedical sciences. For example, the concept-based search mechanism [51] allows biologists to describe the topics of interest in a search more specifically and retrieve information with higher precision (in comparison to the usage of keywords only). It should be stressed here that concept-based search requires linking to high-quality external resources (such as, e.g., UMLS [13]), which involves features related to trust, especially verifiability and believability.

Datasets providing semantic features enable us to go beyond the standard bag of words representation [75]. A wide range of methods based on linking to external, domain-oriented resources has been proposed, e.g., [67], [54], [78]. They also employ statistical features extracted from large-scale text corpora [17] and allow one to expand the user queries to increase recall [7]. In addition, geographical and temporal contexts play an increasingly important role in *IR* applications. These contexts enable the retrieval of information that is relevant with respect to the spatial [46] and temporal [18] dimensions of the query.

Example features for Information Retrieval applications: IR involves qualitative profile features related to trust (i.e., verifiability and believability) and the accessibility of data, In addition, to facilitate semantic search, *IR* implies profile features like topical domains and context.

6.6. Discussion

Overall, we observe that although existing applications make use of the whole spectrum of the dataset profile feature categories, including semantic, qualitative, statistical and temporal features discussed in this survey, the concrete set of features is application-dependent and the whole set is rarely used within any single application. Whereas some applications rely on the existing metadata, many applications choose generating dataset profile features as a part of their own processing pipelines. This can be attributed to missing dataset profile features in many cases. On the one

hand, these applications can thus directly contribute to the dataset profile generation. On the other hand, the burden to generate dataset profile features for each single application hinders usability of the datasets. Thus we think that availability of dataset profiles including a wide range of features can potentially facilitate a new generation of applications in the distributed LOD settings and enlarge the number of datasets used in real-world applications.

7. Conclusions

RDF dataset profiling is perceived as a central challenge in enabling and facilitating dataset discovery in application scenarios such as data linking, data curation, distributed query and search, just to name a few. In this survey, we provide a comprehensive overview for dataset profiling features, methods, tools, vocabularies and applications. Given the complexity of the topic, we first focused on organizing the different dataset profile features in a taxonomy. We then provided a systematic overview of a large set of approaches and tools for assessing and extracting such features from RDF datasets. We reviewed the vocabularies for representing these features, preferably as Linked Data, and finally we discussed several prominent applications of dataset profiles.

Wherever feasible, we also provided insights into the adoption and impact of the discussed works; for instance, based on the profile extraction tools distribution in the provided taxonomy, we propose that certain profiles features, notably in the semantic category, should be provided by the data domain experts to ensure high quality profiles. Another observation concerns the vocabulary usage where some features, such as the quality or the dynamicity of vocabularies do not appear in the evaluated statistics. That leads us to recommend that dataset providers need to guarantee a high confidence with respect to these profile features in order to ensure better access to their quality or how they have changed over time.

We observe that although existing applications make use of the whole spectrum of the discussed feature categories, including semantic, qualitative, statistical and temporal features, the concrete set of features is application-dependent and the whole set is rarely used within any single application. Furthemore, many applications generate dataset profile features as a part of their own processing pipelines, which can be attributed to missing dataset profiles or features in many cases. This leads us to a conclusion that a-priori availability of dataset profiles could facilitate a broader use of profiles and datasets in a variety of application domains.

Finally, we strongly recommend that dataset profiles provide representations readable for **both** humans and machines to open up the Web of Data to a wider variety of users and applications.

Given the continuous evolution and expansion of the Web of Data, we assume that the problem of dataset profiling will become an even more prominent one, and corresponding methods will form a crucial building block for enabling reuse and take-up of datasets beyond established and well-understood knowledge bases and reference graphs.

References

[1] Fedsearch: Efficiently combining structured queries and full-text search in a sparql federation. volume 8218 of *Lecture Notes in Computer Science*, pages 427–443. Springer Berlin Heidelberg, 2013.

[2] Ziawasch Abedjan, Toni Grütze, Anja Jentzsch, and Felix Naumann. Profiling and mining RDF data with prolod++. In *IEEE 30th International Conference on Data Engineering, Chicago, ICDE 2014, IL, USA, March 31 - April 4, 2014*, pages 1198–1201, 2014.

[3] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. Describing linked datasets - on the design and usage of void, the 'vocabulary of interlinked datasets'. In *WWW 2009 Workshop: Linked Data on the Web (LDOW2009)*, Madrid, Spain, 2009.

[4] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. Dbpedia: A nucleus for a web of open data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference*, pages 722–735, 2007.

[5] Sören Auer, Jan Demter, Michael Martin, and Jens Lehmann. Lodstats - an extensible framework for high-performance dataset analytics. In *Knowledge Engineering and Knowledge Management - 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012. Proceedings*, pages 353–362, 2012.

[6] SÃűren Auer, Jan Demter, Michael Martin, and Jens Lehmann. Lodstats - an extensible framework for high-performance dataset analytics. In Annette ten Teije, Johanna VÃűlker, Siegfried Handschuh, Heiner Stuckenschmidt, Mathieu d'Aquin, Andriy Nikolov, Nathalie Aussenac-Gilles, and Nathalie Hernandez, editors, *EKAW*, volume 7603 of *Lecture Notes in Computer Science*, pages 353–362. Springer, 2012.

[7] Jagdev Bhogal, Andy Macfarlane, and Peter Smith. A review of ontology based query expansion. *Information processing & management*, 43(4):866–886, 2007.

[8] Christian BIZER. *Quality-Driven Information Filtering in the Context of Web-Based Information Systems.* PhD thesis, Freie Universitat, Berlin, March 2007.

[9] Christian Bizer and Richard Cyganiak. Quality-driven information filtering using the WIQA policy framework. *J. Web Sem.*, 7(1):1–10, 2009.

[10] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.

[11] Jens Bleiholder and Felix Naumann. Data fusion. *ACM Comput. Surv.*, 41(1):1:1–1:41, January 2009.

[12] Jim Blythe and Yolanda Gil. Incremental formalization of document annotations through ontology-based paraphrasing. In *Proceedings of the 13th international conference on World Wide Web, WWW 2004, New York, NY, USA, May 17-20, 2004*, pages 455–461, 2004.

[13] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.

[14] Christoph Böhm, Johannes Lorey, and Felix Naumann. Creating void descriptions for web-scale data. *J. Web Sem.*, 9(3):339–345, 2011.

[15] Christoph Böhm, Felix Naumann, Ziawasch Abedjan, Dandy Fenz, Toni Grütze, Daniel Hefenbrock, Matthias Pohl, and David Sonnabend. Profiling linked open data with prolod. In *Workshops Proceedings of the 26th International Conference on Data Engineering, ICDE 2010, March 1-6, 2010, Long Beach, California, USA*, pages 175–178, 2010.

[16] Volha Bryl and Christian Bizer. Learning conflict resolution strategies for cross-language wikipedia data fusion. In *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume*, pages 1129–1134, 2014.

[17] Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini. *Ontology learning from text: An overview*, volume 123. 2005.

[18] Ricardo Campos, Gaël Dias, Alípio M Jorge, and Adam Jatowt. Survey of temporal information retrieval and related applications. *ACM Computing Surveys (CSUR)*, 47(2):15, 2014.

[19] Jeremy J Carroll, Christian Bizer, Pat Hayes, and Patrick Stickler. Named graphs, provenance and trust. In *Proceedings of the 14th international conference on World Wide Web*, pages 613–622. ACM, 2005.

[20] Pierre-Antoine Champin and Alexandre Passant. SIOC in Action - Representing the Dynamics of Online Communities. In *Proceedings of the 6th International Conference on Semantic Systems (I-SEMANTICS 2010)*. ACM, 2010.

[21] Richard Cyganiak, Simon Field, Arofan Gregory, Wolfgang Halb, and Jeni Tennison. Semantic statistics: Bringing together sdmx and scovo. In Christian Bizer, Tom Heath, Tim Berners-Lee, and Michael Hausenblas, editors, *LDOW*, volume 628 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2010.

[22] Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. Improving efficiency and accuracy in multilingual entity extraction. In *I-SEMANTICS 2013 - 9th International Conference on Semantic Systems, ISEM '13, Graz, Austria, September 4-6, 2013*, pages 121–124, 2013.

[23] Herbert Van de Sompel, Robert Sanderson, Michael L. Nelson, Lyudmila Balakireva, Harihar Shankar, and Scott Ainsworth. An http-based versioning mechanism for linked data. In *Proceedings of the WWW2010 Workshop on Linked Data on the Web, LDOW 2010, Raleigh, USA, April 27, 2010*, 2010.

[24] E. Demidova, S. Dietze, J. Szymanski, and J. Breslin, editors. *Proceedings of the 1st International Workshop on Dataset PROFIling & fEderated Search for Linked Data (PROFILES 2014), co-located with the 11th Extended Semantic Web Conference (ESWC 2014), Anissaras, Crete, Greece, 26 May 2014.*, volume 1151. CEUR Workshop Proceedings, 2014.

[25] Renata Queiroz Dividino, Ansgar Scherp, Gerd Gröner, and Thomas Grotton. Change-a-lod: Does the schema on the linked data cloud change or not? In *Proceedings of the Fourth International Workshop on Consuming Linked Data, COLD 2013, Sydney, Australia, October 22, 2013*, 2013.

[26] Mohamed Ben Ellefi, Zohra Bellahsene, François Scharffe, and Konstantin Todorov. Towards semantic dataset profiling. In *Proceedings of the 1st International Workshop on Dataset PROFIling & fEderated Search for Linked Data colocated with the 11th Extended Semantic Web Conference, Crete, Greece, May 26, 2014.*, 2014.

[27] Mohamed Ben Ellefi, Zohra Bellahsene, Konstantin Todorov, and Stefan Dietze. Dataset recommendation for data linking: An intensional approach. In *ESWC: European Semantic Web Conference (ESWC 2016), Crete, GrÃĺce*, number 9678, pages 36–51, 2016.

[28] Miriam Fernández, Iván Cantador, Vanesa López, David Vallet, Pablo Castells, and Enrico Motta. Semantically enhanced information retrieval: an ontology-based approach. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(4):434–452, 2011.

[29] Besnik Fetahu, Stefan Dietze, Bernardo Pereira Nunes, Marco Antonio Casanova, Davide Taibi, and Wolfgang Nejdl. A scalable approach for efficiently generating structured dataset topic profiles. In *The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings*, pages 519–534, 2014.

[30] Brad Fitzpatrick, Brett Slatkin, and Martin Atkins. Pubsubhubbub core 0.3–working draft. *Project Hosting on Google Code, available at http://pubsubhubbub.googlecode.com/svn/trunk/pubsubhubbub-core-0.3.html*, 2010.

[31] Daniel Fleischhacker, Heiko Paulheim, Volha Bryl, Johanna Völker, and Christian Bizer. Detecting errors in numerical linked data using cross-checked outlier detection. In *Semantic Web Conference (1)*, pages 357–372, 2014.

[32] Christian Fürber and Martin Hepp. Using semantic web resources for data quality management. *Management*, 6317:1–15, 1998.

[33] Christian Fürber and Martin Hepp. Towards a vocabulary for data quality management in semantic web architectures. In *Proceedings of the 1st International Workshop on Linked Web Data Management*, LWDM '11, pages 1–8, New York, NY, USA, 2011. ACM.

[34] Yolanda Gil and Varun Ratnakar. TRELLIS: an interactive tool for capturing information analysis and decision making. In *Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, 13th International Conference, EKAW 2002, Siguenza, Spain, October 1-4, 2002, Proceedings*, pages 37–42, 2002.

[35] Olaf Görlitz and Steffen Staab. Splendid: Sparql endpoint federation exploiting void descriptions. In *Proceedings of the Second International Workshop on Consuming Linked Data (COLD2011), Bonn, Germany, October 23, 2011*, 2011.

[36] Thomas Gottron and Christian Gottron. Perplexity of index models over evolving linked data. In *The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings*, pages 161–175, 2014.

[37] Guido Governatori, Antonino Rotolo, Serena Villata, and Fabien Gandon. One License to Compose Them All: A Deontic Logic Approach to Data Licensing on the Web of Data. In *Proceedings of the International Semantic Web Conference (ISWC 2013)*, 2013.

[38] Markus Graube, Stephan Hensel, and Leon Urbas. R43ples: Revisions for triples - an approach for version control in the semantic web. In *LDQ@ SEMANTICS*, 2014.

[39] Andreas Harth, Katja Hose, Marcel Karnstedt, Axel Polleres, Kai-Uwe Sattler, and Jürgen Umbrich. Data summaries for on-demand queries over linked data. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 411–420, New York, NY, USA, 2010. ACM.

[40] Olaf Hartig. Trustworthiness of data on the web. In *Proceedings of the STI Berlin & CSW PhD Workshop*, 2008.

[41] Olaf Hartig. Provenance information in the web of data. In Christian Bizer, Tom Heath, Tim Berners-Lee, and Kingsley Idehen, editors, *LDOW*, volume 538 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2009.

[42] Olaf Hartig, Christian Bizer, and Johann Christoph Freytag. Executing sparql queries over the web of linked data. In *The Semantic Web - ISWC 2009, 8th International Semantic Web Conference, ISWC 2009, Chantilly, VA, USA, October 25-29, 2009. Proceedings*, pages 293–309, 2009.

[43] Jens Hartmann, York Sure, Peter Haase, Raul Palma, and Mari del Carmen Suárez-Figueroa. OMV – Ontology Metadata Vocabulary. In Chris Welty, editor, *Ontology Patterns for the Semantic Web Workshop*, Galway, Ireland, 2005.

[44] Michael Hausenblas, Wolfgang Halb, Yves Raimond, Lee Feigenbaum, and Danny Ayers. Scovo: Using statistics on the web of data. In Lora Aroyo, Paolo Traverso, Fabio Ciravegna, Philipp Cimiano, Tom Heath, Eero HyvÃűnen, Riichiro Mizoguchi, Eyal Oren, Marta Sabou, and Elena Paslaru Bontas Simperl, editors, *ESWC*, volume 5554 of *Lecture Notes in Computer Science*, pages 708–722. Springer, 2009.

[45] Tom Heath, Michael Hausenblas, Chris Bizer, Richard Cyganiak, and Olaf Hartig. How to publish linked data on the web. In *Tutorial in the 7th International Semantic Web Conference, Karlsruhe, Germany*, 2008.

[46] Christopher B Jones, Harith Alani, and Douglas Tudhope. Geographical information retrieval with ontologies of place. In *Spatial information theory*, pages 322–335. Springer, 2001.

[47] Tobias Käfer, Ahmed Abdelrahman, Jürgen Umbrich, Patrick O'Byrne, and Aidan Hogan. Observing linked data dynamics. In *The Semantic Web: Semantics and Big Data, 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013. Proceedings*, pages 213–227, 2013.

[48] Shahan Khatchadourian and MarianoP. Consens. Explod: Summary-based exploration of interlinking and rdf usage in the linked open data cloud. In *The Semantic Web: Research and Applications*, volume 6089 of *Lecture Notes in Computer Science*, pages 272–287. Springer Berlin Heidelberg, 2010.

[49] Mathias Konrath, Thomas Gottron, Steffen Staab, and Ansgar Scherp. Schemex - efficient construction of a data catalogue by stream-based indexing of linked data. *J. Web Sem.*, 16:52–58, 2012.

[50] Dimitris Kontokostas, Patrick Westphal, Sören Auer, Sebastian Hellmann, Jens Lehmann, Roland Cornelissen, and Amrapali Zaveri. Test-driven evaluation of linked data quality. In *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014*, pages 747–758, 2014.

[51] Bevan Koopman, Peter Bruza, Laurianne Sitbon, and Michael Lawley. Towards semantic search and inference in electronic medical records: an approach using concept-based information retrieval. *The Australasian medical journal*, 5(9):482, 2012.

[52] Timothy Lebo, Satya Sahoo, and D McGuinness. PROV-O: The PROV Ontology, 2013.

[53] Fadi Maali, Richard Cyganiak, and Vassilios Peristeras. Enabling interoperability of government data catalogues. In Maria Wimmer, Jean-Loup Chappelet, Marijn Janssen, and Hans Jochen Scholl, editors, *EGOV*, volume 6228 of *Lecture Notes in Computer Science*, pages 339–350. Springer, 2010.

[54] David Milne and Ian H Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM, 2008.

[55] Paolo Missier, Khalid Belhajjame, and James Cheney. The W3C PROV family of specifications for modelling provenance metadata. In *EDBT/ICDT '13*, pages 773–776, 2013.

[56] Luc Moreau. The Foundations for Provenance on the Web. *Foundations and Trends in Web Science*, 2(2-3):99–241, 2010.

[57] Luc Moreau and Paolo Missier. PROV-DM: The PROV Data Model, 2013.

[58] Andrea Moro, Alessandro Raganato, and Roberto Navigli. Entity linking meets word sense disambiguation: a unified approach. *TACL*, 2:231–244, 2014.

[59] Felix Naumann. *Quality-driven Query Answering for Integrated Information Systems*. Springer-Verlag, Berlin, Heidelberg, 2002.

[60] Fabrizio Orlandi. *Profiling user interests on the social semantic web*. PhD thesis, National University of Ireland Galway, 2014.

[61] Alexandre Passant and Pablo N. Mendes. sparqlpush: Proactive notification of data updates in RDF stores using pubsubhubbub. In *Proceedings of the Sixth Workshop on Scripting and Development for the Semantic Web, Crete, Greece, May 31, 2010*, 2010.

[62] Heiko Paulheim. Identifying wrong links between datasets by multi-dimensional outlier detection. In *Proceedings of the Third International Workshop on Debugging Ontologies and Ontology Mappings, WoDOOM 2014, co-located with 11th Extended Semantic Web Conference (ESWC 2014), Anissaras/Hersonissou, Greece, May 26, 2014.*, pages 27–38, 2014.

[63] Heiko Paulheim and Christian Bizer. Improving the quality of linked data using statistical distributions. *Int. J. Semantic Web Inf. Syst.*, 10(2):63–86, 2014.

[64] Leo Pipino, Yang W. Lee, and Richard Y. Wang. Data quality assessment. *Commun. ACM*, 45(4):211–218, 2002.

[65] Niko Popitsch and Bernhard Haslhofer. Dsnotify: handling broken links in the web of data. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, pages 761–770, 2010.

[66] Jeffrey Pound, Peter Mika, and Hugo Zaragoza. Ad-hoc object retrieval in the web of data. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, pages 771–780, 2010.

[67] Lev-Arie Ratinov, Dan Roth, Doug Downey, and Mike Anderson. Local and global algorithms for disambiguation to wikipedia. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 1375–1384, 2011.

[68] Edna Ruckhaus, Maria-Esther Vidal, Simón Castillo, Oscar Burguillos, and Oriana Baldizan. Analyzing linked data quality with liquate. In *The Semantic Web: ESWC 2014 Satellite Events - ESWC 2014 Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers*, pages 488–493, 2014.

[69] Monica Scannapieco, Antonino Virgillito, Carlo Marchetti, Massimo Mecella, and Roberto Baldoni. The daquincis architecture: a platform for exchanging and improving data quality in cooperative information systems. *Information Systems*, 29(7):551 – 582, 2004. Data Quality in Cooperative Information Systems.

[70] Tommaso Soru, Edgard Marx, and Axel-Cyrille Ngonga Ngomo. ROCKER: A refinement operator for key discovery. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*, pages 1025–1033, 2015.

[71] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, pages 697–706, 2007.

[72] Kaustubh Supekar, Chintan Patel, and Yugyung Lee. Characterizing quality of knowledge on semantic web. In *Proceedings of AAAI Florida AI Research Symposium (FLAIRS-2004), May 1719, 2004*, 2004.

[73] Danai Symeonidou, Vincent Armant, Nathalie Pernelle, and Fatiha Saïs. Sakey: Scalable almost key discovery in RDF data. In *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, pages 33–49, 2014.

[74] Danai Symeonidou, Nathalie Pernelle, and Fatiha Saïs. KD2R: A key discovery method for semantic reference reconciliation. In *On the Move to Meaningful Internet Systems: OTM 2011 Workshops - Confederated International Workshops and Posters: EI2N+NSF ICE, ICSP+INBAST, ISDE, ORM, OTMA, SWWS+MONET+SeDeS, and VADER 2011, Hersonissos, Crete, Greece, October 17-21, 2011. Proceedings*, pages 392–401, 2011.

[75] Julian Szymański. Comparative analysis of text representation methods using classification. *Cybernetics and Systems*, 45(2):180–199, 2014.

[76] Sebastian Tramp, Philipp Frischmuth, Timofey Ermilov, Saeedeh Shekarpour, and Sören Auer. An architecture of a distributed semantic social network. *Semantic Web*, 5(1):77–95, 2014.

[77] Jürgen Umbrich, Sebastian Neumaier, and Axel Polleres. Quality assessment and evolution of open data portals. In *3rd International Conference on Future Internet of Things and Cloud, FiCloud 2015, Rome, Italy, August 24-26, 2015*, pages 404–411, 2015.

[78] Ellen M Voorhees. Using wordnet for text retrieval. *Fellbaum (Fellbaum, 1998)*, pages 285–303, 1998.

[79] Andreas Wagner, Peter Haase, Achim Rettinger, and Holger Lamm. Entity-based data source contextualization for searching the web of data. In *Proceedings of the 1st International Workshop on Dataset PROFIling & fEderated Search for Linked Data co-located with the 11th Extended Semantic Web Conference, Crete, Greece, May 26, 2014.*, 2014.

[80] Andreas Wagner, Duc Thanh Tran, Günter Ladwig, Andreas Harth, and Rudi Studer. Top-k linked data query processing. In *The Semantic Web: Research and Applications - 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012. Proceedings*, pages 56–71, 2012.

[81] Richard Y. Wang and Diane M. Strong. Beyond accuracy: What data quality means to data consumers. *J. of Management Information Systems*, 12(4):5–33, 1996.

[82] Dominik Wienand and Heiko Paulheim. Detecting incorrect numerical data in dbpedia. In *The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings*, pages 504–518, 2014.

[83] Naiem K. Yeganeh, Shazia Sadiq, and Mohamed A. Sharaf. A framework for data quality aware query systems. *Inf. Syst.*, 46:24–44, December 2014.

[84] Wancheng Yuan, Elena Demidova, Stefan Dietze, and Xuan Zhou. Analyzing relative incompleteness of movie descriptions in the web of data: A case study. In *Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference, ISWC 2014, Riva del Garda, Italy, October 21, 2014.*, pages 197–200, 2014.

[85] Amrapali Zaveri, Dimitris Kontokostas, Mohamed Ahmed Sherif, Lorenz Bühmann, Mohamed Morsey, Sören Auer, and Jens Lehmann. User-driven quality evaluation of dbpedia. In *I-SEMANTICS 2013 - 9th International Conference on Semantic Systems, ISEM '13, Graz, Austria, September 4-6, 2013*, pages 97–104, 2013.

[86] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93, 2016.

Dataset Recommendation for Data Linking: an Intensional Approach

Mohamed Ben Ellefi[1], Zohra Bellahsene[1], Stefan Dietze[2], Konstantin Todorov[1]

[1]LIRMM / University of Montpellier, France
{benellefi, bella, todorov}@lirmm.fr
[2]L3S Research Center / Leibniz University Hannover, Germany
dietze@l3s.de

Abstract. With the growing quantity and diversity of publicly available web datasets, most notably Linked Open Data, recommending datasets, which meet specific criteria, has become an increasingly important, yet challenging problem. This task is of particular interest when addressing issues such as entity retrieval, semantic search and data linking. Here, we focus on that last issue. We introduce a dataset recommendation approach to identify linking candidates based on the presence of schema overlap between datasets. While an understanding of the nature of the content of specific datasets is a crucial prerequisite, we adopt the notion of dataset profiles, where a dataset is characterized through a set of schema concept labels that best describe it and can be potentially enriched by retrieving their textual descriptions. We identify schema overlap by the help of a semantico-frequential concept similarity measure and a ranking criterium based on the *tf*idf* cosine similarity. The experiments, conducted over all available linked datasets on the Linked Open Data cloud, show that our method achieves an average precision of up to 53% for a recall of 100%. As an additional contribution, our method returns the mappings between the schema concepts across datasets – a particularly useful input for the data linking step.

1 Introduction

With the emergence of the Web of Data, in particular Linked Open Data (LOD) [1], an abundance of data has become available on the web. Dataset recommendation is becoming an increasingly important task to support challenges such as entity interlinking [2], entity retrieval or semantic search [3]. Particularly with respect to interlinking, the current topology of the LOD cloud underlines the need for practical and efficient means to recommend suitable datasets: currently, only very few, well established knowledge graphs show a high amount of inlinks, with DBpedia being the most obvious target [4], while a large amount of datasets is largely ignored.

This is due in part to the challenge to identify suitable linking candidates without prior knowledge of the available datasets and their characteristis. Linked

datasets vary significantly with respect to represented resource types, currentness, coverage of topics and domains, size, used languages, coherence, accessibility [5] or general quality aspects [6]. This heterogeneity poses significant challenges for data consumers when attempting to find useful datasets. Hence, a long tail of datasets from the LOD cloud[1] has hardly been reused and adopted, while the majority of data consumption, linking and reuse focuses on established knowledge graphs such as DBpedia [7] or YAGO [8].

In line with [9], we define dataset recommendation as the problem of computing a rank score for each of a set of datasets D_T (for Target Dataset) so that the rank score indicates the relatedness of D_T to a given dataset, D_S (for Source Dataset). The rank scores provide information of the likelihood of a D_T dataset to contain linking candidates for D_S.

We adopt the notion of a dataset profile, defined as a set of concept labels that describe the dataset. By retrieving the textual descriptions of each of these labels, we can map the label profiles to larger text documents. This representation provides richer contextual and semantic information and allows to compute efficiently and inexpensively similarities between profiles.

Although different types of links can be defined across datasets, here we focus on the identity relation given by the statement "owl:sameAs". Our working hypothesis is simple: datasets that share at least one concept, i.e., at least one pair of semantically similar concept labels, are likely to contain at least one potential pair of instances to be linked by a "owl:sameAs" statement. We base our recommendation procedure on this hypothesis and propose an approach in two steps: (1) for every D_S, we identify a cluster[2] of datasets that share schema concepts with D_S and (2) we rank the datasets in each cluster with respect to their relevance to D_S.

In step (1), we identify concept labels that are semantically similar by using a similarity measure based on the frequency of term co-occurence in a large corpus (the web) combined with a semantic distance based on WordNet without relying on string matching techniques [10]. For example, this allows to recommend to a dataset annotated by "school" one annotated by "college". In this way, we form clusters of "comparable datasets" for each source dataset. The intuition is that for a given source dataset, any of the datasets in its cluster is a potential target dataset for interlinking.

Step (2) focuses on ranking the datasets in a D_S-cluster with respect to their importance to D_S. This allows to evaluate the results in a more meaningful way and of course to provide quality results to the user. The ranking criterium should not be based on the amount of schema overlap, because potential to-link instances can be found in datasets sharing 1 class or sharing 100. Therefore, we need a similarity measure on the *profiles* of the comparable datasets. We have proceeded by building a vector model for the document representations of the profiles and computing cosine similarities.

[1] `http://datahub.io/group/lodcloud`

[2] We note that we use the term "cluster" in its general meaning, referring to a set of datasets grouped together by their similarity and not in a machine learning sense.

To evaluate the approach, we have used the current topology of the LOD as evaluation data (ED). As mentioned in the beginning, the LOD link graph is far from being complete, which complicates the interpretation of the obtained results—many false positives are in fact missing positives (missing links) from the evaluation data—a problem that we discuss in detail in the sequel. Note that as a result of the recommendation process, the user is not only given candidate datasets for linking, but also pairs of classes where to look for identical instances. This is an important advantage allowing to run more easily linking systems like SILK [11] in order to verify the quality of the recommendation and perform the acutal linking. Our experimental tests with SILK confirm the hypothesis on the incompleteness of the ED.

To sum up, the paper contains the following contributions: (1) new definitions of dataset profiles based on schema concepts, (2) a recommendation framework allowing to identify the datasets sharing schema with a given source dataset, (3) an efficient ranking criterium for these datasets, (4) an output of additional metadata such as pairs of similar concepts across source and target datasets, (5) a large range of reproducible experiments and in depth analysis with all of our results made available.

We proceed to present the theoretical grounds of our technique in Section 2. Section 3 defines the evaluation framework that has been established and reports on our experimental results. Related approaches are presented and discussed in Section 4 before we conclude in Section 5.

2 A Dataset Interlinking Recommendation Framework

Our recommendation approach relies on the notion of a dataset profile, providing comparable representations of the datasets by the help of characteristic features. In this section, we first introduce the definitions of a dataset profile that we are using in this study. Afterwards, we describe the profile-based recommendation technique that we apply.

2.1 Intensional Dataset Profiles

A dataset profile is seen as a set of dataset characteristics that allow to describe in the best possible way a dataset and that separate it maximally from other datasets. A feature-based representation of this kind allows to compute distances or measure similarities between datasets (or for that matter profiles), which unlocks the dataset recommendation procedure. These descriptive characteristics, or features, can be of various kinds (statistical, semantic, extensional, etc.). As we observe in [12], a dataset profile can be defined based on a set of types (schema concepts) names that represent the topic of the data and the covered domain. In line with that definition, we are interested here in intensional dataset characteristics in the form of a set of keywords together with their definitions that best describe a dataset.

Definition 1 (Dataset Label Profile). *The* **label profile** *of a dataset D, denoted by $\mathcal{P}_l(D)$, is defined as the set of n schema concept labels corresponding to D: $\mathcal{P}_l(D) = \{L_i\}_{i=1}^{n}$.*

Note that the representativity of the labels in $\mathcal{P}_l(D)$ with respect to D can be improved by filtering out certain types. We rely on two main heuristics: (1) remove too popular types (such as foaf:Person), (2) remove types with too few instances in a dataset. These two heuristics are based on the intuition that the probability of finding identical instances of very popular or underpopulated classes is low. We support (1) experimentally in Section 3 while we leave (2) for future work.

Each of the concept labels in $\mathcal{P}_l(D)$ can be mapped to a text document consisting of the label itself and a textual description of this label. This textual description can be the definition of the concept in its ontology, or any other external textual description of the terms composing the concept label. We define a document profile of a dataset in the following way.

Definition 2 (Dataset Document Profile). *The* **document profile** *of a dataset D, $\mathcal{P}_d(D)$, is defined as a text document constructed by the concatenation of the labels in $\mathcal{P}_l(D)$ and the textual descriptions of the labels in $\mathcal{P}_l(D)$.*

Note that there is no substantial difference between the two definitions given above. The document profile is an extended label profile, where more terms, coming from the label descriptions, are included. This allows to project the profile similarity problem onto a vector space by indexing the documents and using a term weighting scheme of some kind (e.g., *tf*idf*).

By the help of these two definitions, a profile can be constructed for any given dataset in a simple and inexpensive way, independent on its connectivity properties on the LOD. In other words, a profile can be easily computed for datasets that are already published and linked, just as for datasets that are to be published and linked, allowing to use the same representation for both kinds of datasets and thus allowing for their comparison by the help of feature-based similarity measures.

As stated in the introduction, we rely on the simple intuition that datasets with similar intension have extensional overlap. Therefore, it suffices to identify at least one pair of semantically similar types in the schema of two datasets in order to select these datasets as potential linking candidates. We are interested in the semantic similarity of concept labels in the dataset label profiles. There are many off-the-shelf similarity measures that can be applied, known from the ontology matching literature. We have focused on the well known semantic measures Wu Palmer [13] and Lin's [14], as well as the UMBC [10] measure that combines semantic distance in WordNet with frequency of occurrence and co-occurrence of terms in a large external corpus (the web). We provide the definition of that measure, since it is less well-known and it showed to perform best in our experiments. For two labels, x and y, we have

$$sim_{\text{UMBC}}(x, y) = sim_{\text{LSA}}(x, y) + 0.5e^{-\alpha D(x,y)}, \tag{1}$$

where $sim_{\text{LSA}}(x, y)$ is the Latent Semantic Analysis (LSA) [15] word similarity, which relies on the words co-occurrence in the same contexts computed in a three billion words corpus[3] of good quality English. $D(x, y)$ is the minimal WordNet [16] path length between x and y. According to [10], using $e^{-\alpha D(x,y)}$ to transform simple shortest path length has shown to be very efficient when the parameter α is set to 0.25.

With a concept label similarity measure at hand, we introduce the notion of dataset comparability, based on the existence of shared intension.

Definition 3 (Comparable Datasets). *Two datasets D' and D'' are comparable if there exists L_i and L_j such that $L_i \in \mathcal{P}_l(D')$, $L_j \in \mathcal{P}_l(D'')$ and $sim_{UMBC}(L_i, L_j) \geq \theta$, where $\theta \in [0, 1]$.*

2.2 Recommendation Process: the CCD-CosineRank approach

A dataset recommendation procedure for the linking task returns, for a given source dataset, a set of target datasets ordered by their likelihood to contain instances identical to those in the source dataset.

Let D_S be a source dataset. We introduce the notion of a *cluster of comparable datasets* related to D_S, or $CCD(D_S)$ for short, defined as the set of target datasets, denoted by D_T, that are comparable to D_S according to Def. 3. Thus, D_S is identified by its CCD and all the linking candidates D_T for this dataset are found in its cluster, following our working hypothesis.

Finally, we need a ranking function that assigns scores to the datasets in $CCD(D_S)$ with respect to D_S expressing the likelihood of a dataset in $CCD(D_S)$ to contain identical instances with those of D_S. To this end, we need a similarity measure on the dataset profiles.

We have worked with the document profiles of the datasets (Def. 2). Since datasets are represented as text documents, we can easily build a vector model by indexing the documents in the corpus formed by all datasets of interest – the ones contained in one single CCD. We use a *tf*idf* weighting scheme, which allows to compute the cosine similarity between the document vectors and thus assign a ranking score to the datasets in a CCD with respect to a given dataset from the same CCD. Note that this approach allows to consider the information of the intensional overlap between datasets prior to ranking and indexing – we are certain to work only with potential linking candidates when we rank, which improves the quality of the ranks. For a given dataset D_S, the procedure returns datasets from $CCD(D_S)$, ordered by their cosine similarity to D_S.

Finally, an important outcome of the recommendation procedure is the fact that, along with an ordered list of linking candidates, the user is provided the pairs of types of two datasets—a source and a target—where to look for identical instances. This information facilitates considerably the linking process, to be performed by an instance matching tool, such as SILK.

[3] `http://ebiquity.umbc.edu/resource/html/id/351`

2.3 Application of the Approach: an Example

We illustrate our approach by an example. We consider *education-data-gov-uk*[4] as a source dataset (D_S). The first step consists in retrieving the schema concepts from this dataset and constructing a clean label profile (we filter out noisy labels, as discussed above), as well as its corresponding document profile (Def. 1 and Def. 2, respectively). We have $\mathcal{P}_l$(*education-data-gov-uk*) = {London Borough Ward, School, Local Learning Skills Council, Adress}. We perform a semantic comparison between the labels in $\mathcal{P}_l$(*education-data-gov-uk*) and all labels in the profiles of the accessible LOD datasets. By fixing $\theta = 0.7$, we generate CCD(*education-data-gov-uk*) containing the set of comparable datasets D_T, as described in Def. 3. The second step consists of ranking the D_T datasets in CCD(*education-data-gov-uk*) by computing the cosine similarity between their document profiles and $\mathcal{P}_d$(*education-data-gov-uk*). The top 5 ranked candidate datasets to be linked with *education-data-gov-uk* are (1) *rkb-explorer-courseware*[5], (2) *rkb-explorer- courseware*[6], (3) *rkb-explorer-southampton*[7], (4) *rkb-explorer-darmstadt*[8], and (5) *oxpoints*[9].

Finally, for each of these datasets, we retrieve the pairs of shared (similar) schema concepts extracted in the comparison part:

– *education-data-gov-uk* and *statistics-data-gov-uk* share two labels "London Borough Ward" and "LocalLearningSkillsCouncil".

– *education-data-gov-uk* and *oxpoints* contain similar labels which are, respectively, "School" and "College", for the SILK results see Section 3.5.

3 Experiments and Results

We proceed to report on the experiments conducted in support of the proposed recommendation method.

3.1 Evaluation Framework

The quality of the outcome of a recommendation process can be evaluated along a number of dimensions. Ricci *et al.* [17] provide a large review of recommender systems evaluation techniques and cite three common types of experiments: (i) offline experiments, where recommendation approaches are compared without user interaction, (ii) user studies, where a small group of subjects experiments with the system and reports on the experience, and (iii) online experiments, where real user populations interact with the system.

For the task of dataset recommendation, the system suggests to the user a list of n target datasets candidates to be linked to a given source dataset. There

[4] `http://education.data.gov.uk/`
[5] `http://courseware.rkbexplorer.com/`
[6] `http://courseware.rkbexplorer.com/`
[7] `http://southampton.rkbexplorer.com/`
[8] `http://darmstadt.rkbexplorer.com/`
[9] `https://data.ox.ac.uk/sparql/`

does not exist a common evaluation framework for the datasets recommendation, thus, we evaluate our method with an offline experiment by using a pre-collected set of linked data considered as evaluation data (ED). The most straightforward, although not unproblematic (see below) choice of evaluation data for the data linking recommendation task is the existing link topology of the current version of the LOD cloud.

In our recommendation process, for a given source dataset D_S, we identify a cluster of target datasets, D_T, that we rank with respect to D_S (*cf.* Section 2.2). To evaluate the quality of the recommendation results given the ED of our choice, we compute the common evaluation measures for recommender systems, precision and recall, defined as functions of the true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) as follows:

$$Pr = \frac{TP}{TP + FP}; \quad Re = \frac{TP}{TP + FN}. \tag{2}$$

The number of potentially useful results that can be presented to the user has to be limited. Therefore, to assess the effectiveness of our approach, we rely on the measure of precision at rank k denoted by $P@k$. Complementarily, we evaluate the precision of our recommendation when the level of recall is 100% by using the mean average precision at $Recall = 1$, MAP@R, given as:

$$MAP@R = \frac{\sum_{q=1}^{\text{Total}_{D_S}} Pr@R(q)}{\text{Total}_{D_S}}, \tag{3}$$

where $R(q)$ corresponds to the rank, at which recall reaches 1 for the qth dataset and Total_{D_S} is the entire number of source datasets in the evaluation.

3.2 Experimental Setup

We started by crawling all available datasets in the LOD cloud group on the Data Hub[10] in order to extract their profiles. In this crawl, only 90 datasets were accessible via endpoints or via dump files. In the first place, for each accessible dataset, we extracted its implicit and explicit schema concepts and their labels, as described in Def. 1. The explicit schema concepts are provided by resource types, while the implicit schema concepts are provided by the definitions of a resource properties [18]. As noted in Section 2, some labels such as "Group", "Thing", "Agent", "Person" are very generic, so they are considered as noisy labels. To address this problem, we filter out schema concepts described by generic vocabularies such as VoID[11], FOAF[12] and SKOS[13]. The dataset document profiles, as defined in Def. 2, are constructed by extracting the textual descriptions

[10] `http://datahub.io/group/lodcloud`
[11] `http://rdfs.org/ns/void`
[12] `http://xmlns.com/foaf/0.1/`
[13] `http://www.w3.org/2004/02/skos/core`

of labels by querying the Linked Open Vocabularies[14] (LOV) with each of the concept labels per dataset.

To form the clusters of comparable datasets from Def. 3, we compute the semantico-frequential similarity between labels (given in eq. (1)). We apply this measure via its available web API service[15]. In addition, we tested our system with two more semantic similarity measures based on WordNet: Wu Palmer and Lin's. For this purpose, we used the 2013 version of the $WS4J$[16] java API.

The evaluation data (ED) corresponds to the outgoing and incoming links extracted from the generated VoID file using the *datahub2void* tool[17]. It is made available on `http://www.lirmm.fr/benellefi/void.ttl`. We note that out of 90 accessible datasets, only those that are linked to at least one accessible dataset in the ED are evaluated in the experiments.

3.3 Evaluation Results

We started by considering each dataset in the ED as an unlinked source (newly published) dataset D_S. Then, we ran the *CCD-CosineRank* workflow, as described in Section 2.2. The first step is to form a $CCD(D_S)$ for each D_S. The CCD construction process depends on the similarity measure on dataset profiles. Thus, we evaluated the CCD clusters in terms of recall for different levels of the threshold θ (*cf.* Def. 3) for the three similarity measures that we apply. We observed that the recall value remains 100% in the following threshold intervals per similarity measure: **Wu Palmer**: $\theta \in [0, 0.9]$; **Lin**: $\theta \in [0, 0.8]$; **UMBC**: $\theta \in [0, 0.7]$.

The CCD construction step ensures a recall of 100% for various threshold values, which will be used to evaluate the ranking step of our recommendation process by the Mean Average Precision (MAP@R) at the maximal recall level, as defined in Def. 3. The results in Fig. 1 show highest performance of the UMBC's measure with a $MAP@R \cong 53\%$ for $\theta = 0.7$, while the best MAP@R values for Wu Palmer and Lin's measures are, respectively, 50% for $\theta = 0.9$ and 51% for $\theta = 0.8$. Guided by these observations, we evaluated our ranking in terms of precision at ranks $k = \{5, 10, 15, 20\}$, as shown in Table 1. Based on these results, we choose UMBC at a threshold $\theta = 0.7$ as a default setting for *CCD-CosineRank*, since it performs best for three out of four k-values and it is more stable than the two others especially with MAP@R.

3.4 Baselines and Comparison

To the best of our knowledge, there does not exist a common benchmark for dataset interlinking recommendation. Since our method uses both label profiles and document profiles, we implemented two recommendation approaches to be

[14] `http://lov.okfn.org/dataset/lov/`
[15] `http://swoogle.umbc.edu/SimService/`
[16] `https://code.google.com/p/ws4j/`
[17] `https://github.com/lod-cloud/datahub2void`

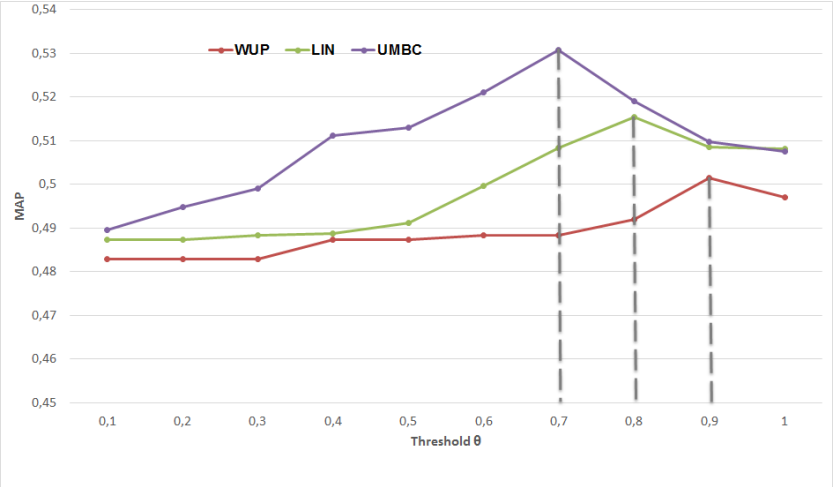


Fig. 1. The MAP@R of our recommender system by using three different similarity measures for different similarity threshold values

considered as baselines – one using document profiles only, and another one using label profiles:

***Doc-CosineRank*:** All datasets are represented by their *document profiles*, as given in Def. 2. We build a vector model by indexing the documents in the corpus formed by all available LOD datasets (no CCD clusters). We use a *tf*idf* weighting scheme, which allows us to compute the cosine similarity between the document vectors and thus assign a ranking score to each dataset in the entire corpus with respect to a given dataset D_S.

UMBCLabelRank: All datasets are represented by their *label profiles*, as given in Def. 1. For a source dataset D_S, we construct its $CCD(D_S)$ according to Def. 3 using UMBC with $\theta = 0.7$. Thus, D_S is identified by its CCD and all target datasets D_T are found in its cluster. Let AvgUMBC be a ranking function that assigns scores to each D_T in $CCD(DS)$, defined by:

$$Avg\text{UMBC}(D', D'') = \frac{\sum_{i=1}^{|\mathcal{P}_l(D')|} \sum_{j=1}^{|\mathcal{P}_l(D'')|} \max sim_{UMBC}(L_i, L_j)}{\max(|\mathcal{P}_l(D')|, |\mathcal{P}_l(D'')|)}, \quad (4)$$

where L_i in $\mathcal{P}_l(D')$ and L_j in $\mathcal{P}_l(D'')$.

Fig. 2 depicts a detailed comparison of the precisions at recall 1 obtained by the three approaches for each D_S taken as source dataset. It can be seen that the *CCD-CosineRank* approach is more stable and largely outperforms the two other approaches by an MAP@R of up to 53% as compared to 39% for *UMBCLabelRank* and 49% for *CCD-CosineRank*. However, the *UMBCLabelRank* approach produces better results than the other ones for a limited number of

Measure \ P@k	P@5	P@10	P@15	P@20
WU Palmer ($\theta = 0.9$)	0, 56	0, 52	0.53	0.51
Lin ($\theta = 0.8$)	0.57	**0.54**	**0.55**	0.51
UMBC ($\theta = 0.7$)	**0.58**	**0.54**	0.53	**0.53**

Table 1. Precision at 5, 10, 15 and 20 of the *CCD-CosineRank* approach using three different similarity measures over their best threshold values based on Fig.1

source datasets, especially in the case when D_S and D_T share a high number of identic labels in their profiles.

The performance of the *CCD-CosineRank* approach demonstrates the efficiency and the complementarity of combining in the same pipeline (i) the *semantic* similarity on labels for identifying recommendation candidates (CCD construction process) and (ii) the *frequential* document cosine similarity to rank the candidate datasets. We make all of the ranking results of the *CCD-CosineRank* approach available to the community on `http://www.lirmm.fr/benellefi/CCD-CosineRank_Result.csv`.

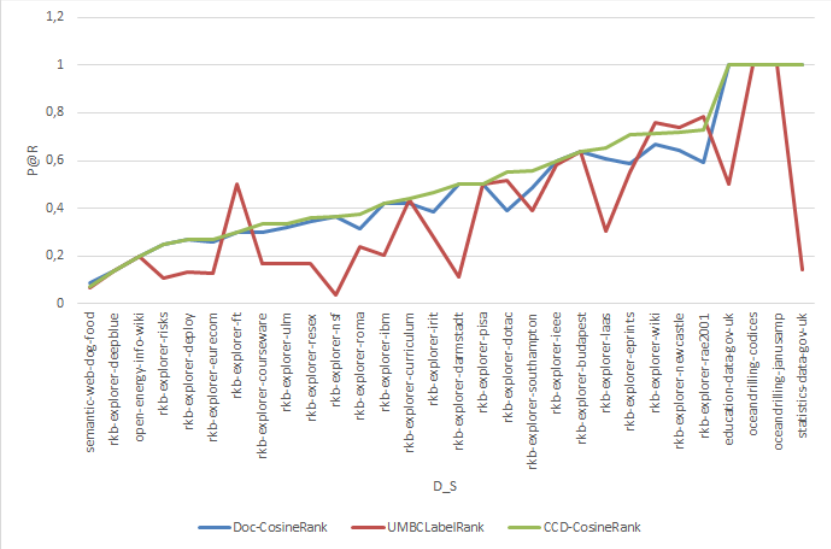


Fig. 2. Precisions at recall=1 of the *CCD-CosineRank* approach as compared to *Doc-CosineRank* and *UMBCLabelRank*

3.5 Discussion

We begin by a note on the vocabulary filtering that we perform (Section. 3.2). We underline that we have identified the types which improve/decrease the per-

formance empirically. As expected, vocabularies, which are very generic and wide-spread have a negative impact, acting like hub nodes, which dilute the results. For comparison, the results of the recommendation before removal are made available on `http://www.lirmm.fr/benellefi/RankNoFilter.csv`.

The different experiments described above show a high performance of the introduced recommendation approach with an average precision of 53% for a recall of 100%. Likewise, it may be observed that this performance is completely independent of the dataset size (number of triples) or the schema cardinality (number of schema concepts by datasets). However, we note that better performance was obtained for datasets from the *geographic* and *governmental* domains with precision and recall of 100%. Naturally, this is due to the fact that a recommender system in general and particularly our system performs better with datasets having high quality schema description and datasets reusing existing vocabularies (the case for the two domains cited above), which is considered as linked data modeling best practice. An effort has to be made for improving the quality of the published dataset [19].

We believe that our method can be given a more fair evaluation if better evaluation data in the form of ground truth are used. Indeed, our results are impacted by the problem of false positives overestimation. Since data are not collected using the recommender system under evaluation, we are forced to assume that the false positive items would have not been used even if they had been recommended, i.e., that they are uninteresting or useless to the user. This assumption is, however, generally false, for example when the set of unused items contains some interesting items that the user did not select. In our case, we are using declared links in the LOD cloud as ED, which is certain but far from being complete for it to be considered as ground truth. Thus, in the recommendation process the number of false positives tends to be overestimated, or in other words an important number of missing positives in the ED translates into false positives in the recommendation process.

To further illustrate the effect of false positives overestimation, we ran SILK as an instance matching tool to discover links between D_S and their corresponding D_Ts that have been considered as false positives in our ED. SILK takes as an input a *Link Specification Language* file, which contains the instance matching configuration. We recall that our recommendation procedure provides pairs of shared or similar types between D_S and every D_T in its corresponding CCD, which are particularly useful to configure SILK. However, all additional information, such as the datatype properties of interest, has to be given manually. This makes the process very time consuming and tedious to perform over the entire LOD. Therefore, as an illustration, we ran the instance matching tool on two flagship examples of false positive D_Ts:

Semantically Similar Labels: We choose *education-data-gov-uk*[18] as a D_S and its corresponding false positive D_T *oxpoints*[19]. The two datasets contain

[18] `http://education.data.gov.uk/`
[19] `https://data.ox.ac.uk/sparql`

in their profiles, respectively, the labels "School" and "College", detected as highly similar labels by the UMBC measure, with a score of 0.91. The instance matching gave as a result 10 accepted "owl:sameAs" links between the two datasets.

Identical Labels: We choose *rkb-explorer-unlocode*[20] as a D_S and its corresponding D_Ts, which are considered as FP: *yovisto*[21] *datos-bcn-uk*[22] *datos-bcn-cl*[23]. All 4 datasets share the label "Country" in their corresponding profiles. The instance matching process gave as a result a set of accepted "owl:sameAs" links between *rkb-explorer-unlocode* and each of the three D_T.

We provide the set of newly discovered linksets to be added to the LOD topology and we made the generated linksets and the corresponding SILK configurations available on `http://www.lirmm.fr/benellefi/Silk_Matching`.

It should be noted that the recommendation results provided by our approach may contain some broader candidate datasets with respect to the source dataset. For example, two datasets that share schema labels such as books and authors are considered as candidates even when they are from different domains like science vs. literature. This outcome can be useful for predicting links such as "rdfs:seeAlso" (rather than "owl:sameAs"). We have chosen to avoid the inclusion of instance-related information in order to keep the complexity of the system as low as possible and still provide reasonable precision by guaranteeing a 100% recall.

As a conclusion, we outline three directions of work in terms of dataset quality that can considerably facilitate the evaluation of any recommender system in that field: (1) improving descriptions and metadata; (2) improving accessibility; (3) providing a reliable ground truth and benchmark data for evaluation.

4 Related Work

With respect to finding relevant datasets on the Web, we cite briefly several studies on discovering relevant datasets for query answering. Based on well-known data mining strategies, [20] and [21] present techniques to find relevant datasets, which offer contextual information corresponding to the user queries. A feedback-based approach to incrementally identify new datasets for domain-specific linked data applications is proposed in [22]. User feedback is used as a way to assess the relevance of the candidate datasets.

In the following, we cite approaches that have been devised for the datasets interlinking candidates recommendation task and which are, therefore, directly relevant to our work. Nikolov *et al.* [23] propose a keyword-based search approach to identify candidate sources for data linking consisting of two steps: (i) searching for potentially relevant entities in other datasets using as keywords

[20] `http://unlocode.rkbexplorer.com/sparql/`
[21] `http://sparql.yovisto.com/`
[22] `http://data.open.ac.uk/query`
[23] `http://data.open.ac.cl/query`

randomly selected instances over the literals in the source dataset, and (ii) filtering out irrelevant datasets by measuring semantic concept similarities obtained by applying ontology matching techniques.

Mehdi *et al.* [24] propose a method to automatically identify relevant public SPARQL endpoints from a list of candidates. First, the process needs as input a set of domain-specific keywords, which are extracted from a local source or can be provided manually by an expert. Then, using natural languages processing techniques and queries expansion techniques, the system generates a set of queries that seek to exact literal matches between the introduced keywords and the target datasets, i.e., for each term supplied to the algorithm, the system runs a comparison to a set of eight queries: {original-case, proper-case, lower-case, upper-case} × {no-lang-tag, @en-tag}. Finally, the produced output consists of a list of potentially relevant SPARQL endpoints of datasets for linking. In addition, an interesting contribution of this technique is the bindings returned for the subject and predicate query variables, which are recorded and logged when a term match is found on some particular SPARQL endpoint. The records are useful in the linking step.

Leme *et al.* [25] present a ranking method for datasets with respect to their relevance for the interlinking task. The ranking is based on Bayesian criteria and on the popularity of the datasets, which affects the generality of the approach. The authors extend this work and overcome this drawback in [9] by exploring the correlation between different sets of features—properties, classes and vocabularies—and the links to compute new rank score functions for all the available linked datasets.

None of the studies outlined above have evaluated the ranking measure in terms of Precision/Recall, except for [9] which, according to the authors, achieves a mean average precision of around 60% and an excepted recall of 100% with rankings over all LOD datasets. However, a direct comparison to our approach seems unfair since the authors did not provide the list of the datasets and their rank performance by datasets considered as source.

In comparison to the work discussed above, our approach has the potential of overcoming a series of complexity related problems, precisely, considering the complexity to generate the matching in [23], to produce the set of domain-specific keywords as input in [24] and to explore the set of features of all the network datasets in [9]. Our recommendation results are much easier to obtain since we only manipulate the schema part of the dataset. They are also easier to interpret and apply since we automatically recommend the corresponding schema concept mappings together with the candidate datasets.

5 Conclusion and Future Work

Following the linked data best practices, metadata designers reuse and build on, instead of replicating, existing RDF schema and vocabularies. Motivated by this observation, we propose the *CCD-CosineRank* interlinking candidate dataset recommendation approach, based on concept label profiles and schema

overlap across datasets. Our approach consists of identifying clusters of comparable datasets, then, ranking the datasets in each cluster with respect to a given dataset. We discuss three different similarity measures, by which the relevance of our recommendation can be achieved. We evaluate our approach on real data coming from the LOD cloud and compare it two baseline methods. The results show that our method achieves a mean average precision of around 53% for recall of 100%, which reduces considerably the cost of dataset interlinking. In addition, as a post-processing step, our system returns sets of schema concept mappings between source and target datasets, which decreases considerably the interlinking effort and allows to verify explicitly the quality of the recommendation.

In the future, we plan to improve the evaluation framework by developing a more reliable and complete evaluation data for dataset recommendation. We plan to elaborate a ground truth based on certain parts of the LOD, possibly by using crowdsourcing techniques, in order to deal with the false positives overestimation problem. Further work should go into btaining high quality profiles, in particular by considering the population of the schema elements. We also plan to investigate the effectiveness of machine learning techniques, such as classification or clustering, for the recommendaiton task. One of the conclusions of our study shows that the recommendation approach is limited by the lack of accessibility, explicit metadata and quality descriptions of the datasets. As this can be given as an advice to data publishers, in the future, we will work on the development of recommendation methods for datasets with noisy and incomplete descriptions.

Acknowledgements

This research has been partially funded under the Datalyse project[24]-FSN-AAP Big Data n3-, by the European Commission-funded DURAARK project (FP7 Grant Agreement No. 600908) and the COST Action IC1302 (KEYSTONE).

References

1. C. Bizer, T. Heath, and T. Berners-Lee, "Linked data - the story so far," *Int. J. Semantic Web Inf. Syst.*, vol. 5, no. 3, pp. 1–22, 2009.
2. B. P. Nunes, S. Dietze, M. A. Casanova, R. Kawase, B. Fetahu, and W. Nejdl, "Combining a co-occurrence-based and a semantic measure for entity linking," in *Proc. of the 10th ESWC*, pp. 548–562, 2013.
3. R. Blanco, P. Mika, and S. Vigna, "Effective and efficient entity search in rdf data.," in *Proc. of ISWC*, vol. 7031, pp. 83–97, Springer, 2011.
4. M. Schmachtenberg, C. Bizer, and H. Paulheim, "Adoption of the linked data best practices in different topical domains," in *Proc. of ISWC*, pp. 245–260, 2014.
5. C. Buil-Aranda, A. Hogan, J. Umbrich, and P.-Y. Vandenbusshe, "Sparql web-querying infrastructure: Ready for action?," in *Proc. of the 12th ISWC*, 2013.
6. C. Guéret, P. T. Groth, C. Stadler, and J. Lehmann, "Assessing linked data mappings using network measures," in *Proc. of the 9th ESWC*, pp. 87–102, 2012.

[24] `http://www.datalyse.fr/`

7. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives, "Dbpedia: A nucleus for a web of open data," in *proc. of ISWC*, pp. 722–735, 2007.
8. F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *Proc. of WWW*, pp. 697–706, 2007.
9. G. Lopes, L. A. Paes Leme, B. Nunes, M. Casanova, and S. Dietze, "Two approaches to the dataset interlinking recommendation problem," in *Proc. of 15th on WISE 2014*, 2014.
10. L. Han, A. L. Kashyap, T. Finin, J. Mayfield, and J. Weese, "Umbc_ebiquity-core: Semantic textual similarity systems," in *Proc. of the *SEM*, Association for Computational Linguistics, 2013.
11. J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov, "Silk - A link discovery framework for the web of data," in *Proceedings of the WWW2009, LDOW*, 2009.
12. M. B. Ellefi, Z. Bellahsene, F. Scharffe, and K. Todorov, "Towards semantic dataset profiling," in *Proc. of Dataset PROFIling & fEderated Search for Linked Data Workshop co-located with the 11th ESWC*, 2014.
13. Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proc. of the 32Nd ACL*, pp. 133–138, 1994.
14. D. Lin, "An information-theoretic definition of similarity," in *Proc. of ICML*, pp. 296–304, 1998.
15. S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," *JASIS1990*, vol. 41, pp. 391–407.
16. G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, pp. 39–41, Nov. 1995.
17. F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, *Recommender systems handbook*, vol. 1. Springer, 2011.
18. T. Gottron, M. Knauf, S. Scheglmann, and A. Scherp, "A systematic investigation of explicit and implicit schema information on the linked open data cloud," in *Proc. of ESWC*, pp. 228–242, 2013.
19. M. B. Ellefi, Z. Bellahsene, and K. Todorov, "Datavore: A vocabulary recommender tool assisting linked data modeling," in *Proc.of the ISWC Posters & Demonstrations Track a track*, 2015.
20. A. Wagner, P. Haase, A. Rettinger, and H. Lamm, "Discovering related data sources in data-portals," in *Proc. of the 1st IWSS*, 2013.
21. A. Wagner, P. Haase, A. Rettinger, and H. Lamm, "Entity-based data source contextualization for searching the web of data," in *Proc. of the Dataset PROFIling & fEderated Search for Linked Data Workshop co-located with the 11th ESWC*, pp. 25–41, 2014.
22. H. R. de Oliveira, A. T. Tavares, and B. F. Lóscio, "Feedback-based data set recommendation for building linked data applications," in *Proc. of the 8th ISWC*, pp. 49–55, ACM, 2012.
23. A. Nikolov and M. d'Aquin, "Identifying relevant sources for data linking using a semantic web index," in *WWW2011, LDOW*, 2011.
24. M. Mehdi, A. Iqbal, A. Hogan, A. Hasnain, Y. Khan, S. Decker, and R. Sahay, "Discovering domain-specific public SPARQL endpoints: a life-sciences use-case," in *Proc. of the 18th IDEAS 2014*, pp. 39–45.
25. L. A. P. P. Leme, G. R. Lopes, B. P. Nunes, M. A. Casanova, and S. Dietze, "Identifying candidate datasets for data interlinking," in *Proc. of the 13th ICWE*, pp. 354–366, 2013.

Appendix C

Extended Abstract in French

Contributions Scientifiques et Projet de Recherche
-Résumé en français-

Konstantin Todorov

1. Cycle de vie des données ouvertes et liées: résumé de mes recherches

Au cours des dernières années, nous avons été témoins à un effort croissant de construction et de publication de données structurées sur le Web sous la forme de jeux de données sous la forme de graphes, également appelés graphes de connaissances, utilisant souvent le formalisme RDF et et les technologies du Web sémantique pour leur construction et leur partage. L'exemple le plus important est le projet Linked Open Data (LOD), qui regroupe actuellement des centaines de jeux de données issus de différents champs et domaines d'application, accessibles ouvertement sur le Web. De grands graphes de connaissances, tels que Google Knowledge Graph ou le Knowledge Vault [1], ont été conçus pour structurer les informations disponibles sur le Web et améliorer ainsi la recherche et l'accès à ces informations. Dans le même temps, des groupes tels que schema.org ont été mis en avant avec succès par les groupes de travail du W3C afin de fournir des moyens d'annoter le contenu Web et de faciliter la recherche d'informations pour les moteurs de recherche [2]. Nous assistons à un changement dans la façon de publier et de consommer données sur le Web, passant d'informations non structurées rendues disponibles de manière décentralisée et explorées à l'aide de requêtes ressemblant à des mots clés, vers un Web "nouvel" qui, sans remplacer le Web d'aujourd'hui, l'étend avec une couche structurée supplémentaire, ressemblant de plus en plus à une immense base de données, où les algorithmes de réponse aux questions permettent de récupérer des réponses précises aux requêtes des utilisateurs, à l'instar des bases de données relationnelles (cf. Fig. 2.1).

Au-delà des utilisateurs ordinaires du Web, des experts issus de domaines spécifiques scientifiques, culturels ou de la société en général bénéficient de ces efforts. Les chercheurs en informatique travaillent en étroite collaboration avec divers experts du domaine afin de favoriser le développement de données structurées et librement accessibles dans leurs domaines respectifs, dans le but d'améliorer et de faciliter l'accès des experts aux données, de leur permettre de découvrir de nouvelles connaissances et de les libérer du fardeau technologique dans ce processus. On citera à titre d'exemple diverses initiatives dans les domaines de la biomédecine [137], de l'agronomie[1] ou de la musique [3], notamment des projets tels que D2KAB[2], DOREMUS[3] ou ClaimsKG[4], auxquels je contribue actuellement. L'ensemble des

[1] http://agrold.southgreen.fr/agrold/
[2] http://d2kab.mystrikingly.com/
[3] https://www.doremus.org/
[4] https://data.gesis.org/claimskg/

principes qui guident le développement des connaissances structurées, leur partage et leur réutilisation ont récemment été baptisé FAIR, désignant des données *Findable, Accessible, Interoperable* et *Reusable*.

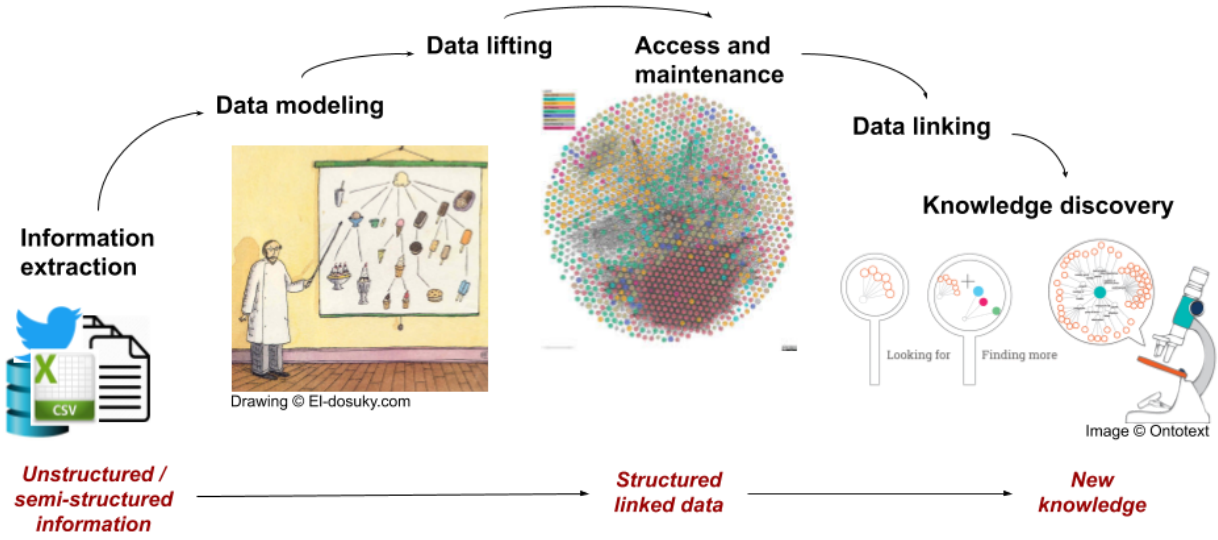


Fig. 2.1 Cycle de vie des données ouvertes et liées

Faire des données Web FAIR ouvre de nombreux défis scientifiques pour la recherche en intelligence artificielle (IA) et en informatique. Mes activités de recherche depuis le début de ma thèse sont motivées par la nécessité de répondre à certains de ces défis. Je vais donc expliquer plus en détail le cycle de vie des données (FAIR) lors de leur ouverture et de leur publication sur le Web, dans le cas particulier de l'utilisation des technologies du Web sémantique comme couche de fond. Sur la base des visions proposées dans [4, 5] ou du cycle de vie des données ouvertes liées de LOD2 présenté dans [6], je résume les principales étapes de la publication d'un nouveau jeu de données FAIR sous forme de données liées, comme indiqué ci-dessous (voir également la Fig. 2.1). :

- **Extraction d'informations.** Ce composant comprend l'extraction et la transformation d'informations (entités et relations) de la source de données brutes en données structurées par des relations, par exemple en utilisant le formalisme RDF. La source de données d'origine peut être sous forme de fichier CSV, XML, non structuré, structuré ou semi-structuré, c'est-à-dire une base de données relationnelle, un corpus de documents texte (pages Web, documents de réseau social tels que les *tweets*). Un certain nombre de techniques de traitement automatique du langage naturel (TALN) sont appliquées afin d'extraire les entités saillantes et les relations qui les relient.
- **Modélisation des données.** Le modèle, également appelé ontologie ou vocabulaire, est le socle conceptuel utilisé pour décrire et représenter les données dans un domaine d'intérêt: quels types de choses existent et comment elles sont liées les unes aux autres, quelles sont leurs propriétés. Il est essentiel que l'ontologie finale réutilise les termes de

vocabulaires pertinents existants, le cas échéant. Trouver un modèle de données approprié pour le domaine d'intérêt peut être une tâche particulièrement laborieuse et fastidieuse, car elle implique des interactions entre différentes communautés, où les informaticiens et en particulier les experts du Web sémantique travaillent en étroite collaboration avec des experts d'un domaine spécifique (par exemple, le domaine culturel, celui du patrimoine ou la sociologie, selon l'application).

- **Élévation, ou lifting, des données.** Cette étape rassemble les résultats des deux premières étapes. Cela consiste à attribuer des espaces de noms ou à donner des noms à toutes les ressources et propriétés de nos données relationnelles RDF conformément au modèle développé, à désambiguïser et relier les entités nommées et à rendre ces ressources accessibles via des identifiants uniques (URI). En d'autres termes, les entités et les relations extraites à la première étape seront désormais annotées par les noms de concepts et de propriétés de notre vocabulaire développé à la deuxième étape.
- **Accès et maintenance.** Ce composant consiste à héberger publiquement le jeu de données lié et ses métadonnées, en le rendant accessible via des endpoints SPARQL et, éventuellement, des outils 'user-friendly' orientés vers les experts non-informaticiens pour l'exploration et la recherche. Garantir la persévérance des ressources publiées est une question importante à prendre en compte.
- **Liage de données, ou 'data linking'.** Cette étape vise à libérer pleinement le potentiel du projet de données ouvertes et liées en permettant l'accès fédéré à des éléments de données répartis sur diverses sources. Cela implique le liage (automatique) de graphes de connaissances récemment publiés avec d'autres jeux de données déjà publiés en tant que données liées sur le Web en établissant des relations typées entre les ressources.
- **Découverte de nouvelles connaissances.** Enfin, le graphe de connaissances lié peut être utilisé comme connaissance de base pour la découverte de nouvelles relations, liens et nouvelles connaissances à l'aide d'algorithmes d'apprentissage automatique (prédiction de liens, regroupement ou extraction de modèles) ou d'inférence symbolique.

Ces différentes étapes ne sont pas dans une séquence stricte et n'existent pas en isolément, mais sont en cours d'enrichissement mutuel. Chacune de ces étapes pose un ensemble de défis spécifiques. Extraire des entités et des relations à partir de sources non structurées (pages Web, réseaux sociaux) ou semi-structurées (Wikipedia) afin d'alimenter et de créer des bases de connaissances et de fournir ainsi des connaissances structurées sur le Web constitue un intérêt central pour la TALN, le Web data mining et la communauté Web Sémantique au cours des dernières décennies, se concentrant sur une variété de tâches telles que la reconnaissance d'entités nommées, la liaison d'entités, l'extraction de relations ou la désambiguïsation du sens des mots. Les recherches approfondies dans ce domaine ont conduit à un très large éventail de travaux [7, 8, 9].

Le processus de modélisation des données est complexe car qu'il nécessite un effort important de la part des concepteurs de métadonnées (informaticiens) et des experts du domaine pour pouvoir, en premier lieu, élaborer un modèle conceptuel approprié du domaine d'intérêt, puis

s’attaquer aux problèmes soulevés par le besoin d’identifier les termes appropriés à partir de vocabulaires existants afin de les réutiliser conformément aux meilleures pratiques de modélisation des données ouvertes et liées [10]. Par exemple, s’il s’agit de modéliser des données sur les assertions de personnalités politiques, il peut être utile de réutiliser des propriétés de la classe ClaimReview du vocabulaire schema.org.

La découverte de liens sémantiques entre différents graphes RDF, à la fois en termes d’entités (appelé liage de données ou ‘data linking’) et de schéma (ce que nous appelons alignements d’ontologies) est un autre défi, ce qui est manuellement impossible à réaliser compte tenu de l'échelle du problème. Habituellement, parmi les différents types de liens sémantiques pouvant être établis, on s'intéresse plus particulièrement à celui l'identité. Cette relation indique que deux adresses URI différentes (concepts ontologiques ou entités de graphe de connaissance) se rapportent au même objet ou groupe d'objets réel. Par exemple, DBpedia utilise l'URI http://dbpedia.org/resource/Montpellier pour identifier la ville de Montpellier, tandis que Geonames utilise l'URI http://www.geonames.org/2992166 pour la même entité et les deux sont décrits différemment dans les deux ressources, à l'aide d'attributs différents, de labels et parfois même de langues différentes. Des techniques et des outils de couplage de données ou d'appariement ou alignement d'ontologies sont en cours de développement afin de traiter ce problème automatiquement, comme indiqué dans [11]. Cependant, ces tâches nécessitent toujours une intervention humaine, notamment dans la phase fastidieuse de configuration des outils de mise en correspondance d’instances ou dans l’identification des jeux de données candidats à lier, où la recherche de jeux de données-cibles devrait se faire presque par une recherche exhaustive de tous les jeux de données en question dans les différents catalogues. Ces derniers défis sont devant l’utilisateur même avant le début du processus de liage de données.

En raison de la quête pour automatiser autant que possible le processus de production de données liées, de nombreux graphes disponibles résultent de l'application d'outils automatiques d'extraction de connaissances et de lifting ou de liage de données. Bien que cela ait facilité la publication d'un grand nombre de jeux de données liés sur le Web, les approches automatiques ont soulevé de nombreuses questions concernant la qualité, l'actualité et l'exhaustivité des informations contenues. Par conséquent, le principal défi à l'issue de ce processus concerne l'évaluation de la qualité des données [12]. À cet égard, plusieurs problèmes se posent après la publication des données liées. D'une part, les éditeurs de données (ou plutôt les responsables de la maintenance) ont la responsabilité de garantir la maintenance continue de l'ensemble de données publié en termes de qualité, c'est-à-dire d'accès, de dynamique (de versions différentes, de miroirs) et autres. D'autre part, du point de vue des consommateurs de données liés, il est nécessaire (1) de rechercher et d'extraire des informations appropriées à partir de différents ensembles de données ouverts liés, (2) d'intégrer et de réutiliser ces connaissances, (3) d'assurer un retour constant aux responsables de la gestion des données. En ces termes, les approches permettant de créer automatiquement des profils de graphes de connaissances et de jeux de données liés apparaissent comme un effort important pour acquérir une meilleure

compréhension de la nature et du statut d'un jeu de données donné, susceptible d'améliorer la découvrabilité et la réutilisation.

Enfin, une fois les graphes de connaissances construits, de qualité attestée, publiés et liés à d'autres ressources de données - en d'autres termes, ils sont véritablement FAIR -, leur véritable potentiel au service des experts ou des scientifiques du domaine peut être dévoilé. Pour cette raison, la communauté de la connaissance et du Web doit s'assurer que ces utilisateurs non-informaticiens disposent des outils appropriés pour accéder aux données, les explorer, les analyser et les récupérer. Cela implique le développement d'environnements 'user-friendly' ne nécessitant pas la connaissance des langages de requête formels pour accéder aux données. Et finalement, grâce aux techniques de raisonnement ou d'apprentissage automatique (telles que l'exploration de motifs, la classification ou la prédiction de relations), il est possible de découvrir de nouvelles connaissances en utilisant les ressources de données liées actuellement disponibles comme connaissances de référence (background knowledge) dans ce processus [13, 14].

Comme le montre la figure 2.2, j'ai eu l'occasion de contribuer à la plupart des étapes du cycle de vie des données liées, notamment l'extraction des connaissances et l'harmonisation des données, l'intégration des données (correspondance des schémas et des instances) et la création de graphes de connaissances. Ces recherches ont été réalisées en collaboration avec un certain nombre de partenaires académiques et non-académiques du domaine de l'informatique et d'autres domaines, tels que la sociologie ou la biologie avec des applications dans les domaines du patrimoine culturel, de la génétique humaine, de la biologie végétale et des sciences sociales, dans le cadre de nombreux projets de recherche nationaux et internationaux.

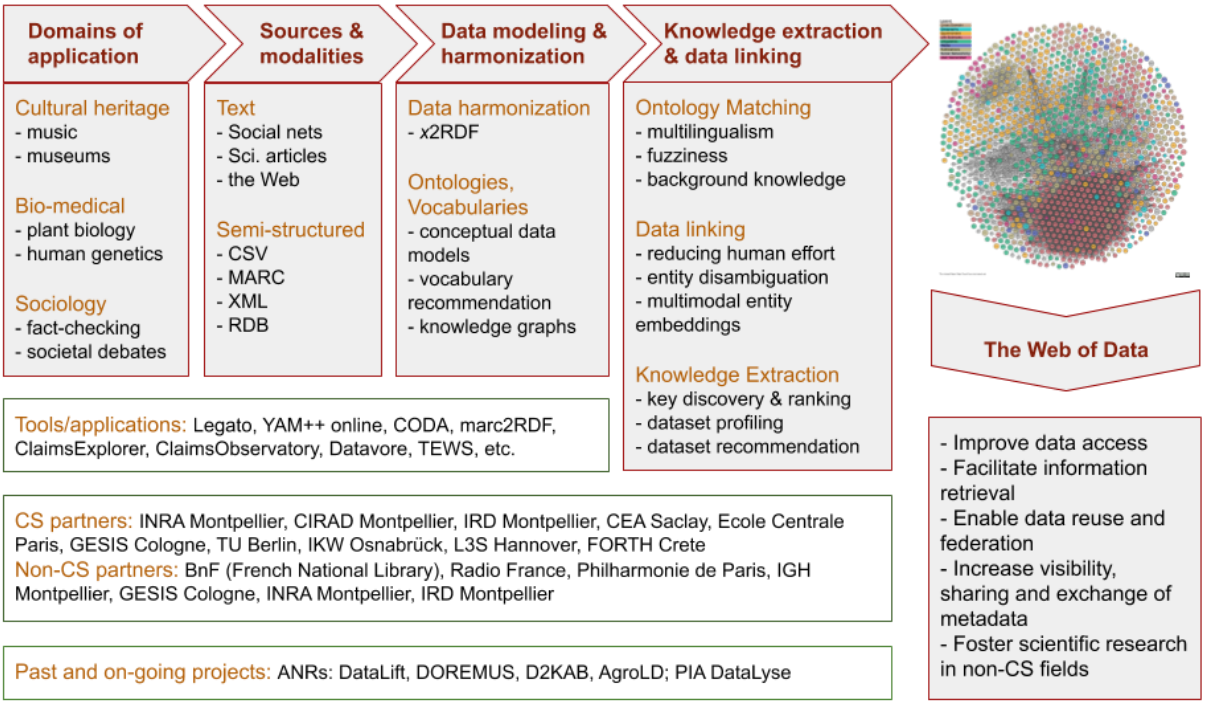


Fig. 2.2: Activités et projets de recherche, collaborations scientifiques.

2. Alignement d'ontologies

Les modèles conceptuels de données, également appelés ontologies, schémas ou vocabulaires, décrivent ce qui existe et présente un intérêt pour une application particulière dans un domaine donné, de manière partagée et formelle, permettant ainsi une compréhension commune des concepts d'importance dans ce domaine comme leurs propriétés et leurs relations. Le processus de développement de modèles de données de ce type depuis les premières années du Web sémantique a été et reste en grande partie décentralisé et biaisé par des compréhensions spécifiques individuelles ou intercommunautaires de ces domaines. Ce processus dépend également fortement de la perspective avec laquelle on considère un domaine particulier - par exemple, parler de la biologie d'un point de vue scientifique ou bibliothécaire ferait définir les concepts d'intérêt et leurs relations différemment. Ces phénomènes ont conduit à la création de vocabulaires décrivant des domaines d'intérêt identiques, très similaires ou se recoupant en utilisant différentes conceptualisations en termes de terminologie, de structure et de sémantique. Toute différence dans la description d'un ensemble de concepts donné entre deux ontologies ou plus est qualifiée d'hétérogénéité d'ontologies. Des exemples de telles hétérogénéités sont l'utilisation de différents termes pour étiqueter le même concept ou structurer les concepts différemment selon deux schémas, en termes de relations entre eux ou de jeux de propriétés les décrivant. Le domaine d'alignement des ontologies du Web sémantique, ou encore de mise en correspondance ou de réconciliation des schémas, relève le défi de fournir des approches et des outils permettant de traiter automatiquement les hétérogénéités que deux ou plusieurs ontologies manifestent en

établissant des relations typées entre des concepts ontologiques [15], avec la compréhension sous-jacente que des ontologies hétérogènes peuvent et doivent (co-)exister pacifiquement.

2.1. Alignements basés sur les instances

La contribution principale de ma thèse de doctorat consiste en le développement d'approches de correspondance d'ontologies basées sur des instances. Cela implique de s'appuyer sur des instances ontologiques pour établir des correspondances entre les classes (ou concepts) de deux ou plusieurs ontologies. Ma contribution se concentre sur le cas particulier où les concepts d'ontologies sont utilisés pour annoter des documents texte (par exemple dans le cas de répertoires Web tels que Yahoo ou Amazon). En cela, un concept est modélisé comme un ensemble d'instances (documents textes) et la relation d'équivalence entre deux concepts est établie sur la base d'une mesure de similarité dans leurs ensembles de documents correspondants. Le processus de correspondance permet d'aligner des annotations hétérogènes de contenu Web similaire.

2.2. Alignements d'ontologies pour la recherche d'information multimédia

Dans la continuité de mon travail de doctorat, au cours de la première période de ma recherche postdoctorale, j'ai appliqué et étendu les résultats obtenus sur des données multimédias, portant notamment sur l'alignement des annotations d'images afin d'améliorer la recherche d'informations multimédia. Avec mes collègues de l'Ecole Centrale Paris, j'ai proposé des méthodes permettant d'aligner les concepts multimédias issus d'ontologies d'annotation multimédia aux concepts établies de la connaissance du sens commun, par exemple, contenus dans une grande encyclopédie en ligne (Wikipédia) ou des bases de données lexicales (WordNet).

2.3. Alignements d'ontologies à l'aide de la connaissance a priori et alignements multilangues

L'utilisation des connaissances a priori dans le processus d'alignement des ontologies était l'un des défis identifiés dans le papier de revue de 2012 par Euzenat et Shvaiko [15]. En réponse à ce défi, j'ai eu l'occasion de formuler un certain nombre de propositions qui améliorent la correspondance d'ontologies et ouvrent des possibilités pour de nouvelles applications. Une des particularités des solutions proposées est qu'elles tentent de modéliser le flou et l'imprécision inhérents à tout alignement (produit par une machine ou par un humain), empruntant ainsi un appareil théorique à la théorie des ensembles flous et s'appuyant sur de la connaissance a priori (background knowledge) dans ce processus.

La vision d'un Web de données multilingue a été présentée en 2012 dans un article de Garcia et al. [16], où la création de liens entre vocabulaires multilingues a été identifiée comme l'un des défis à la réalisation de cette vision. Deux ontologies sont considérées multilingues si elles sont décrites dans deux langues naturelles différentes, alors que chacune d'elles est monolingue. La

présence d'étiquettes multilingues complique en effet le processus d'alignement, car elle ajoute une hétérogénéité supplémentaire à traiter, en plus des différences standards en termes de terminologie, de couverture ou de structure. Relevant ce défi, j'ai appliqué et étendu mes travaux sur l'appariement d'ontologies floues, à l'aide de connaissances a priori, au problème de l'alignement d'ontologies multilingues. Ce travail a été effectué à la fin de mon contrat postdoctoral et s'est poursuivi au LIRMM dans les premières années de ma nomination en tant que maître de conférences.

3. Liage de données et extraction de connaissance pour le liage de données

Le domaine de liage des données (data linking) est étroitement lié au celui d'alignement d'ontologies. Le problème est défini comme le processus d'établissement de relations typées entre entités (ou instances ou ressources) à travers des graphes de connaissance (ou ce que nous appellerons également des jeux de données ou des graphes RDF) de manière automatique. Tout comme dans le cas du problème d'appariement d'ontologies, l'observation motivante est que deux entités de deux jeux de données différents peuvent être décrites très différemment (par exemple, en utilisant des labels, des ensembles de propriétés ou des identificateurs différents) à travers de deux graphes, ce qui soulève à nouveau le problème de l'hétérogénéité. La relation d'identité entre les ressources, permettant de déclarer qu'elles se rapportent à la même entité du monde réel, présente un intérêt particulier car elle permet de créer des ponts entre des silos de données isolés, permettant ainsi des requêtes fédérées et améliorant le partage des données.

Outre une contribution traitant directement du problème de liage de données défini ci-dessus par le développement d'un nouveau système de liaison de données appelé *Legato*, j'ai pu contribuer par des travaux sur trois problèmes annexes: (1) le profilage de graphes de connaissances, (2) la recommandation de jeux de données pour le liage et (3) la découverte et le ranking des clés RDF. Celles-ci comprennent un ensemble de techniques se déroulant avant le processus de liaison de données proprement dit, permettant de découvrir de nouvelles connaissances sur les jeux de données en question, qui visent à faciliter le choix des candidats à la liaison ou la configuration d'un outil de liaison de données générique.

4. Modélisation de la connaissance et des données et construction de graphes de connaissances

Le but de la communauté scientifique en Web sémantique est de structurer les données et les connaissances sur le Web et de les mettre à la disposition à de vastes communautés de consommateurs de données. Les graphes de connaissances, terme inventé par Google en 2012, désignent les artefacts de données décrivant de manière structurée les entités présentant un intérêt dans un domaine donné ainsi que leurs relations. Ils peuvent être considérés comme des «produits» finaux de ce long pipeline que l'on a nommé "cycle de vie des données" au début. Les graphes de connaissances, en particulier lorsqu'ils sont librement accessibles et construits conformément aux principes de FAIR, sont des outils puissants permettant la

réutilisation et la fédération des données, améliorant ainsi la recherche d'informations et facilitant la recherche et la découverte de connaissances dans divers domaines de la vie, notamment la science, le patrimoine culturel ou l'éducation. Leur construction implique l'orchestration de plusieurs tâches d'ingénierie de la connaissance, allant de la collecte et de l'harmonisation des données à l'extraction de connaissances, à la modélisation et à la mise en relation des données, à la fourniture d'outils d'accès et d'exploration de données sur mesure. experts en informatique. Ce processus peut dépendre en grande partie du domaine d'intérêt et nécessite donc souvent l'élaboration d'approches spécifiques mieux adaptées aux données et au domaine concerné. En outre, la création de tels graphes de connaissances pose un certain nombre de défis scientifiques liés notamment à la modélisation des données et à l'extraction des connaissances. Dans mes travaux de recherche, j'ai pu contribué à la construction de grands graphes de connaissances dans deux domaines différents. (1) Le graphe de connaissances DOREMUS d'œuvres musicales liées [3] regroupe des métadonnées sur la musique classique provenant de trois grandes institutions culturelles françaises. Ce travail a principalement été réalisé en collaboration avec Manel Achichi, mon doctorant, et EURECOM (Raphael Troncy et son étudiant, Pasquale Lisena). (2) Mes travaux récents m'ont amené à la la construction de ClaimsKG, un graphe de connaissances de revendications (claims) vérifiées, facilitant les requêtes structurées sur leurs valeurs de vérité, auteurs, dates, revues journalistiques et autres types de métadonnées et fournissant des données de vérité de terrain à l'appui des recherches sur l'analyse des débats de société sur le Web [17]. Ce dernier projet est le résultat d'une collaboration continue avec des collègues de l'institut de sciences sociales computationnel GESIS en Allemagne.

5. Résumés de mes projets de recherche

Je présente deux axes de recherche, dans lesquels je compte orienter mes efforts dans un avenir proche en termes de réseaux de collaborations, d'opportunités de financement et de supervision des étudiants de doctorat et de masters. Les deux axes sont très larges et comportent chacun plusieurs défis de recherche plus spécifiques que je résume plus en détail dans le manuscrit.

Le premier axe est un suivi direct de mes contributions de longue date dans le domaine de la liaison de données et de l'intégration des connaissances que j'avais déjà commencées au cours de ma thèse de doctorat, mais propose une nouvelle façon d'aborder ce problème qui pourrait potentiellement dépasser les limites des approches actuelles. Notez qu'un article décrivant cet axe de recherche a récemment été publié dans le track «Outrageous Ideas» de l'International Semantic Web Conference 2019 [18][5]. Ce track de la conférence fournit un forum d'idées visionnaires, de défis à long terme et d'opportunités pour le domaine du Web sémantique[6]. Cet axe de recherche comprend une famille de défis que je traite plus en détail dans le document principal. En résumé, au lieu d'essayer d'adapter une solution générique à un problème de liage

[5] Le papier a gagné le premier prix de la société CCC Blue Skies Ideas (1000$) à la conférence.
[6] https://iswc2019.semanticweb.org/call-for-outrageous-papers/

ou à un type de jeux de données, je suggère de permettre une meilleure compréhension des données sous-jacentes avant d'appliquer une solution ciblée mieux adaptée aux jeux de données en question. Je m'appuie sur le principe que l'analyse en profondeur de grandes quantités de données liées permettra d'isoler un nombre limité de problèmes de liage de données identifiables que les modèles d'apprentissage automatique basés sur des profils de jeux de données aideront à détecter automatiquement. Le moment est opportun pour adopter cette approche pour des raisons, d'une part, de la disponibilité importante et croissante de données liées dans un nombre toujours plus grand de domaines et, d'autre part, l'existence d'une pléthore d'outils de liage de données, fruit de décennies de recherche et de pratique. La canalisation de ces efforts décentralisés favorisera et facilitera l'application de technologies de données liées dans des domaines encore plus variés et libérera l'expert du domaine du fardeau technologique.

Le deuxième axe consiste en l'élaboration de thèmes de recherche plus récents sur lesquels je n'ai commencé à travailler que depuis deux ans. Il comprend la combinaison d'approches sémantiques pour la représentation conceptuelle de connaissances et de données avec des techniques d'apprentissage automatique afin de permettre une meilleure compréhension des assertions et des informations associées aux assertions sur le Web, allant au-delà du paradigme actuel de recherche sur le Web, de fourniture et de consommation de données. Bien que l'apprentissage automatique ait été le sujet de mes études de master et sous-tend en partie les contributions de ma thèse de doctorat, il peut être considéré à la fois comme un nouvel axe thématique de recherche et comme un retour à mon domaine d'origine. Le champ d'application des recherches envisagées le long de cet axe s'enracine dans le contexte plus large de l'analyse des débats de société sur le Web et s'inspire d'une récente collaboration en cours avec des spécialistes des sciences sociales informatique de Cologne (Allemagne).

Le Web a évolué pour devenir une plate-forme omniprésente où chacun a la possibilité d'être un éditeur, d'exprimer des opinions et d'interagir avec les autres. Cela facilite potentiellement la construction de connaissances de manière démocratique et ascendante et crée des possibilités sans précédent pour mieux comprendre les débats de société sur des sujets controversés. Dans le même temps, des conditions sont créées pour manipuler le discours public en diffusant des informations erronées ou biaisées, offrant ainsi un grand potentiel d'influence sur la société [19]. Les conséquences (à long terme) ne peuvent pas encore être évaluées dans leur intégralité, mais des effets néfastes sur le discours public démocratique sont largement assumés [20]. Par conséquent, il existe un besoin croissant de méthodes et de jeux de données pouvant faciliter l'analyse des débats de société pour les scientifiques des communautés à la fois internes et externes à l'informatique et aux sciences sociales (informatiques). Les études actuelles souffrent souvent d'une généralisabilité et d'une représentativité limitées en raison d'un ensemble restreint et disparate de sources analysées, de sujets, de types de supports ou de délais [21]. Parallèlement, il n'existe pas de jeux de données bien structurés et accessibles au public décrivant des informations sur les débats de société antérieurs ou en cours, qui pourraient être exploités dans les recherches actuelles et futures. Les graphes de connaissances (KG), tels que DBpedia ou Wikidata, sont largement utilisés pour des tâches

telles que l'annotation par entité de documents en ligne ou la recherche sur le Web, capturant des informations factuelles sans garder trace de la diversité, de la connexion ou de l'évolution temporelle des points de vue, aspects (sous-thèmes), les revendications et les sources associées à des entités et des sujets particuliers. Pouvoir intégrer des données polyvalentes provenant à la fois du Web, mais aussi d'archives scientifiques, de bibliothèques et d'instituts de recensement dans des ressources de connaissances sémantiquement structurées et librement accessibles permettra aux analyses basées sur les données de sujets controversés d'améliorer les pratiques démocratiques de nos sociétés. Compte tenu de la grande importance sociétale de ces questions, la fourniture d'outils automatisés d'extraction et d'un accès structuré aux informations pertinentes constitue une tâche urgente pour les communautés d'informatiques et d'intelligence artificielle (IA). En cela, les recherches dans cette direction permettront d'aller au-delà du stockage et de l'interopérabilité d'informations factuelles pour aller vers une meilleure compréhension et modélisation des assertions.

Le manuscrit développe mon projet de recherche dans cet axe selon deux directions étroitement liées:

1. Fournir des méthodes et des outils open source pour la construction de grands graphes de connaissances sur des sujets controversés, des revendications, sources et autres métadonnées connexes, en améliorant et en allant au-delà des travaux en cours.
2. Développer des approches pour la modélisation informatique et l'analyse des revendications (claims), leur véracité et leur relations.

Mon hypothèse est que ces approches contribueront au final à l'enrichissement du Web de manière à donner aux revendications et aux informations relatives aux revendications une forme compréhensible par la machine permettant aux moteurs de recherche d'interpréter les revendications en relation avec des controverses, de la même manière que le Web actuel, grâce aux graphes de connaissances, permet de comprendre et interpréter les faits.

Références

[1] Xin Dong et al. “Knowledge vault: A web-scale approach to probabilistic knowledge fusion”. In: ACM SIGKDD. ACM. 2014, pp. 601–610.
[2] Ran Yua et al. “KnowMore-Knowledge Base Augmentation with Structured Web Markup”. In: Semantic Web J., IOS Press (2017).
[3] Manel Achichi et al. “DOREMUS: A graph of linked musical works”. In: ISWC. Springer. 2018, pp. 3–19.
[4] Bernadette Hyland, BV Terrazas, and S Capadisli. “Cookbook for Open Government Linked Data”. In: W3C, W3C Task Force-Government Linked Data Group (2011).
[5] Michael Hausenblas; Richard Cygankiak. “Linked Data Life cycles”. In: formerly at http://linked-data-life-cycles.info/
[6] Sören Auer and Jens Lehmann. “Creating knowledge out of interlinked data”. In: Semantic Web 1.1, 2 (2010), pp. 97–104.

[7] Aldo Gangemi. “A comparison of knowledge extraction tools for the semantic web”. In: Extended semantic web conference. Springer. 2013, pp. 351–366.
[8] Jose L Martinez-Rodriguez, Aidan Hogan, and Ivan Lopez-Arevalo. “Information extraction meets the semantic web: a survey”. In: SemanticWeb Preprint (2018), pp. 1–81.
[9] GerhardWeikum, Johannes Hoffart, and Fabian Suchanek. “Knowledge harvesting: achievements and challenges”. In: Computing and Software Science. Springer, 2019, pp. 217–235.
[10] Christian Bizer, Tom Heath, and Tim Berners-Lee. “Linked data: The story so far”. In: Semantic services, interoperability and web applications: emerging concepts. IGI Global, 2011, pp. 205–227.
[11] Markus Nentwig et al. “A survey of current Link Discovery frameworks”. In: Semantic Web (2015), pp. 1–18.
[12] Mohamed Ben Ellefi et al. “RDF dataset profiling–a survey of features, methods, vocabularies and applications”. In: Semantic Web (2018).
[13] Sahar Vahdati et al. “Unveiling scholarly communities over knowledge graphs”. In: International Conference on Theory and Practice of Digital Libraries. Springer. 2018, pp. 103–115.
[14] Ignacio Traverso-Ribón and Maria-Esther Vidal. “GARUM: a semantic similarity measure based on machine learning and entity characteristics”. In: International Conference on Database and Expert Systems Applications. Springer 2018, pp. 169–183.
[15] Pavel Shvaiko and Jérôme Euzenat. “Ontology matching: state of the art and future challenges”. In: IEEE Transactions on knowledge and data engineering 25.1 (2013), pp. 158–176
[16] Jorge Gracia et al. “Challenges for the multilingual web of data”. In: Web Semantics: Science, Services and Agents on theWorldWideWeb 11 (2012), pp. 63– 71.
[17] Andon Tchechmedjiev et al. “ClaimsKG -AKnowledge Graph of Fact-checked Claims”. In: ISWC. 2019.
[18] Konstantin Todorov. “Datasets First! A Bottom-up Data Linking Paradigm”. In: Procs of the ISWC 2019 Satellite Tracks (Outrageous Ideas). 2019, pp. 338–342.
[19] Soroush Vosoughi, Deb Roy, and Sinan Aral. “The spread of true and false news online”. In: Science 359.6380 (2018), pp. 1146–1151.
[20] Hunt Allcott and Matthew Gentzkow. “Social media and fake news in the 2016 election”. In: Journal of Economic Perspectives 31.2 (2017), pp. 211–36.
[21] Gerret von Nordheim, Karin Boczek, and Lars Koppers. “Sourcing the Sources”. In: Digital Journalism. 2018, pp. 807–828.