Investigating One Million XRefs in Thirthy Ontologies from the OBO World

Amir Laadhar, Elcio Abrahão, Clement Jonquet

Investigating One Million XRefs in Thirthy Ontologies from the OBO World

Amir LAADHAR [a], Elcio ABRAHÃO [a], Clement JONQUET [a]
[a] *Laboratory of Informatics, Robotics and Microelectronics of Montpellier (LIRMM), University of Montpellier, France*

Abstract. The more ontologies are produced, the more need to identify mappings between them becomes important. Several practices and tools have been developed to support ontology alignment, but still, it remains a challenge. In the OBO world, ontology developers use cross reference annotations (formally using the `oboInOwl:hasDbXRef` property) to link a term to an external entity, including another term in another ontology (i.e., a mapping). These XRefs contains information of high value, because they were manually declared/verified by domain experts when the ontology was built. However, extracting and exploiting them remains a challenge for users due to their heterogeneous and chaotic descriptions. In this paper, we report on analysis of approximately 1 million XRefs in 30 ontologies from the OBO world. Our results show that 10.7% of these XRefs are ontology mappings, but confirm that semantically-ambiguous usage of the cross-reference property which make it impractical or even impossible to reuse. We describe and quantify several issues identified (e.g., different kind of XRefs, miscellaneous formatting, ambiguous targets), and discuss some way to mitigate them especially by using more relevant specific properties taken from standard semantic Web metadata vocabularies.

Keywords. OBO Foundry, OBO ontologies, biological and biomedical ontologies, cross-references, OBO-XRefs, ontology alignment, data linking, data curation.

1. Introduction

Semantic interoperability in biological and biomedical applications has been influenced by the use of ontologies. Many domain-specific ontologies are developed by subject matter experts. Driven by several use cases in agronomy, we have designed AgroPortal, a repository of ontologies and vocabularies for agronomy and related domains (agriculture, plant science, food, and biodiversity – `http://agroportal.lirmm.fr`) [5,6]. As of May 2020, AgroPortal includes 120 ontologies, thesaurus, and vocabularies encoded in different formats. AgroPortal host 22 OBO Foundry ontologies [13] such as the Gene Ontology, Protein Ontology, Plant Ontology, Environment Ontology, Agronomy Ontology, or FoodOn Ontology. In addition, AgroPortal hosts ontologies not referenced in the OBO Foundry but developed with the OBO format or methodology such as the Livestock Product Trait Ontology and the Soy Ontology or some crop-specific trait dictionaries in the Crop Ontology project. The more ontologies are added to AgroPortal, the more the need to identify mappings between them becomes important. Today, we want to create in AgroPortal a mapping repository to store and merge any kind of mappings between

the ontologies stored in the portal. Therefore, we have investigated methods to extract mappings explicitly declared by developers inside the ontology source files [1].

The OBO Foundry [13] is an initiative to create a set of well-defined reference ontologies and to enforce a set of design principles for ontologies to be orthogonal, interoperable and avoid unnecessary duplication of work. There exist several ways of identifying equivalent or similar terms from other ontologies. In the OBO world,[1] ontology developers employ a cross-reference mechanism (later abbreviated XRef) to link ontology terms to external entities including terms in different ontologies. Formally, this is done using the XRef property in the OBO format or the `oboInOwl:hasDbXRef` property in OWL and referencing an external term with the `prefix:id` notation. This notation combines a prefix and an id (e.g., `GO:0005623`. Indeed, each term has a unique identifier that consists of a prefix (the ontology IDSpace) concatenated with the id for that term within the ontology. Based on the OBO format specification: "The IDSpace (...) should ideally be registered on the GO xrefs page or with OBO". This was supposed to facilitate the identification of terms throughout the whole OBO world. However, we have identified an ambiguous and chaotic usage of the XRef mechanism by ontology developers which makes potentially very valuable information, impractical or even impossible to reuse. Chriss Mungall has stated in a blog post [2] that different representations of mappings serve different purposes. He states that erroneous usage of XRefs can lead to the propagation of erroneous information across multiple ontologies, which could have bad consequences. For instance, in addition to external ontology terms, we found uses of the XRef property to reference databases or elements inside databases, Web pages, citations, or even the curator of a term. We have found spelling mistakes, different prefixes for the same term (e.g., NCBITAXON, NCBI), and different ways to identify a term (e.g., `FMA:31396`, `FMA:Cartilage_of_inferior`), including different way of using the `prefix:id` notation (with / or _ instead of :).[3]

To extract ontology mappings from XRefs, we had to disambiguate and curate these misleading XRefs using a semi-automated methodology. We have developed the Ontology Mapping Harvesting Tool (OMHT), a tool to extract mappings from inside ontology source files and reify them into specific objects with metadata and provenance information. OMHT can semi-automatically curate ambiguous IDs and prefixes existing in the XRef properties. In this paper, we present an analysis of the use of XRefs in a corpus of 30 OBO and OWL ontologies related to the agri-food domain hosted in AgroPortal in March 2020. We illustrate discrepancies found in the ways XRefs are used, we describe and quantify several issues identified and discuss some way to mitigate them especially by using more relevant specific properties taken from standard semantic Web metadata vocabularies and unambiguous identifiers. The main contributions of this work are:

- An openly available tool (OMHT) to extract and disambiguate XRefs;
- An analysis of the use of 1 million XRefs across a subset of 30 ontologies;
- A curated dataset of 551,957 XRefs from ontology terms to other entities;
- Recommendations for substituting XRefs by better alternative representations when the actual use of XRefs is semantically ambiguous.

[1] We use this expression to englobe the OBO Foundry and Library ontologies as well as ontologies designed with the same practices or format but not formally declared in the Foundry (`www.obofoundry.org`).

[2] `https://bit.ly/2DtNMgq`

[3] The OBO Foundry guidelines name this notation "CURIE" (`http://www.obofoundry.org/id-policy.html`), but we use here the expression `prefix:id` to include valid CURIES and alternative uses.

The rest of this paper is organized as follows: the next section presents related work. In Section 3, we describe the methodology to retrieve and curate XRefs. In Section 4, we present an analysis of the XRefs curated and non-curated. In Section 5, we discuss the results and we give some recommendations. Finally, we conclude and present some perspectives in Section 6.

2. Related work

Mappings ontologies are necessary when working with multiple ontologies. These mappings can be used for a variety of reasons, such as data integration, decision support, semantic search, data annotation, and reasoning [3]. For instance, Mungall et al. 2016 [12] makes extensive use of cross-reference mappings to build a unified representation of disease. Although the semantics of the XRef property is not explicitly defined by the OBO format we supposed that it has been frequently used to indicate equivalence/similarity between terms [7]. The reference specification (syntax and semantics) of the OBO format[4] requires the target term to be declared using the `prefix:id` notation. All XRefs should follow the OBO Foundry ID policy i.e., all ID-spaces must be registered and approved, and the entire ID should consist of the ID-space followed by ':' followed by a zero-padded numeric local identifier (a minimum of 7 digits is recommended). Plus, the ontology header specifies how to consider XRefs (e.g., `treat-xrefs-as-equivalent`, `treat-xrefs-as-is_a`). Nevertheless these guidelines, the document itself acknowledged that XRefs frequently do not conform to the canonical `prefixed:id` pattern.

Challenges remain when integrating term identifiers from multiple sources due to a lack of consistency in how identifiers are reported [11]. Plenty of work has been published about the automatic generation of mappings between ontologies [3]. However, to the best of our knowledge, ontology mappings analysis is quite rare. Mapping analysis can lead to discover interesting insights related to the nature of the aligned ontologies and the quality of mappings. In 2009, Amir et al. studied term overlap between biomedical ontologies in the NCBO BioPortal [14] using LOOM, a lexical matching system to find terms with similar labels between ontologies [4]. In 2015, Kamdar et al. [9] investigated term reuse and overlap in 377 ontologies from BioPortal [14]. More recently, Kamdar et al. investigated term reuse and overlap in 509 ontologies from BioPortal. Kamdar et al. concluded that despite best practices recommendations biomedical ontologies were still far from achieving ideal term reuse, beyond the upper level and popular ontologies [10].

The only previous work we have identified on exploiting XRefs is the one done to build the Ontology XRef Service (OxO) at the European Bioinformatics Institute [7]. Like us here, OxO is interested in extracting ontology mappings from XRefs in 104 ontologies. They harmonize the identifiers to provide an integrated resource of mappings and build a nice visualization (`www.ebi.ac.uk/spot/oxo`). To disambiguate the prefix of XRefs' target data sources, OxO automatically relies on Identifiers.org [8], the OBO library, and prefixcommons.org. In their study, OxO identified 1.4 million XRefs, where the prefixes are identified automatically. There remain around 80 identifier prefixes that are not automatically identified. OxO curates only mappings to ontologies and does not deal with web pages or RDF datasets. In our work, we propose a similar semi-automatic approach to identify prefixes to different kinds of targets including semantic resources

[4] `https://owlcollab.github.io/oboformat/doc/GO.format.obo-1_4.html`

and web pages; then we manually curate and complete the prefixes and the identifier patterns as described after.

To extract and curate mappings from ontologies (including XRefs), we have not found any reusable tool except the oxo-loader which is limited for use with EBI Ontology Lookup Service. Therefore, we have developed a more generic tool (OMHT) that deals directly with an ontology source file accessed in AgroPortal or NCBO BioPortal.

3. Methodology: XRefs extraction and curation

In this section, we describe the methodology for extracting a dataset of XRefs from ontologies stored in AgroPortal using a semi-automatic curation of the XRefs. Each XRef references a source term (in the ontology processed by OMHT) and a target entity. Some ontologies formally import other ontologies using the construct `owl:imports`; in these cases, we did not process the imports of external ontologies nor include their XRefs in our dataset. The XRefs extraction using OMHT mainly follows four steps (in the following sections, we detail step 2 and 3):

1. Access ontologies RDF/XML serialization via AgroPortal REST API;
2. Identify XRefs in the ontologies looking up `oboInOwl:hasDbXRef` occurences;
3. Curate ambiguous XRefs using a manually built target curation file;
4. Store the final curated XRefs as reified objects (using JSON format used for mappings in AgroPortal) and generate statistics about curated and non-curated XRefs.

OMHT parses the RDF/XML files of the ontologies to identify XRefs. Once an XRef is identified locating the property `oboInOwl:hasDbXRef`, the script extracts the source term URI from the current ontology, and tries to explicitly generate an URI for the target ontology and target term (or in our case the target entity if not an ontology term). The target term or entity can be identified based on the following patterns:

- A `prefix:id` pair: It is common to find XRefs that describe the target term using a `prefix:id` notation, with various separators (e.g., `AGRO:4356` or `GRO_000023` or `CHEBI/0686857547`). In these cases, the OMHT identifies the prefix as the target ontology identifier and the id as the target term identifier. By extension, in this study the prefix was also considered as a database identifier and the id as an identifier of an entity in this database.
- An URI or URL candidate: OMHT parses the string trying to identify patterns in the target term or entity to extract the ontology/database and term/entity URIs. If founded, the URIs are stored as this in the XRef extracted. Otherwise, the URIs are saved for manual curation.
- Invalid target: If the script founds a string content in another format than the two described above they are discarded. These invalid XRefs are stored in a curation file to manually identify the proper target ontology/database and term/entity.

When an unambiguous ontology and ontology term cannot be directly identified as target, OMHT relies on a curation file which groups invalid pair of `prefix:id`, invalid URL/URI and non recognized targets and map them to a proper target. Being grouped and sorted by the number of occurrences allow a curator to disambiguate targets by

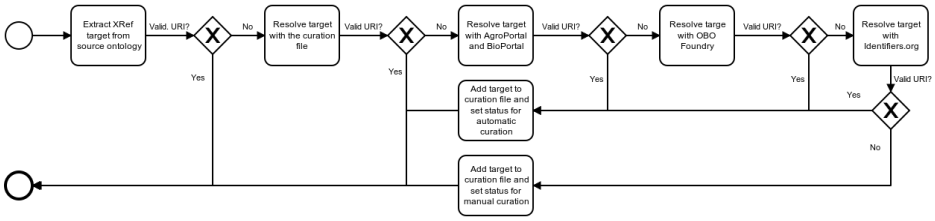

Figure 1. Semi-automatic curation process to identify valid targets from XRefs.

batches. The file used for curation is publicly available with OMHT source code.[5] Figure 1 depicts the semi-automatic curation process: OMHT searches for the target ontologies automatically using Web services from AgroPortal [6], BioPortal [14], OBO Foundry [13], and Identifiers.org [8]. If the prefix of the target ontology is found, the information is stored in the curation file to facilitate/accelerate the curation work. If not, a specific status informs that this need curation. Additional information for each target ontology is provided to facilitate the curation process, as the number of occurrences for that target, the acronyms of the source ontologies where that target was founded, and syntax example of targets. The external curation file is maintained by our research group and uses iteratively by OMHT: once a target has been manually disambiguate once, future executions of OMHT will consider the target extracted valid. Therefore, manual curation help to replace invalid targets by valid URIs for ontology and ontology term. OMHT has been primarily designed to extract ontology mappings, but in this study, we used it to extract also annotations or cross-references to other entities contained in XRefs. The curator (final author) manually tagged each target with its type:

- Y: the target is a valid ontology (and a valid identifier for the ontology is provided);
- RDF: the target is an RDF dataset (in which URIs can be used) but not an ontology;
- URL: the target is a valid reference to a database or database element accessible by a URL.

Once the curation process is finished, we can generate statistics to get insights from the usage of the extracted XRefs.

4. Results and examples

In this section, we present an analysis of the XRefs extracted from our corpus of ontologies using OMHT. We classify and quantify the number of curated XRefs based on the type of target, that can refer either to ontology terms, RDF datasets, or Web pages. Then, we analyze the non-curated XRefs. As of May 2020, AgroPortal hosts 122 semantic resources related to agri-food, plant sciences, and ecology-biodiversity. We examined a set of 97 ontologies, which includes 88 ontologies in the OWL format (e.g., ENVO, GO, SO, PO) and 9 ontologies in the OBO format (e.g., TRIPHASE, OntoBiotope, Wheat Phenotype). Within these 97 ontologies, 30 contains XRefs (17 ontologies from the OBO library, 13 others). Using the OMHT tool, we identified 951,957 XRefs in the ontology source files. In Figure 2, we depict a nested pie chart that shows the number of curated and non-curated XRefs as well as their repartition. The percentage of curated XRefs is 58.3%, and there are 25 ontologies with at least one curated XRef.

[5] `https://github.com/agroportal/ontology_mapping_harvester`

In this corpus, we were not able to identify the target of 398,247 XRefs (after called non-curated). The dataset of curated XRefs in openly available in the OntoPortal JSON format at `https://zenodo.org/record/3842750`.

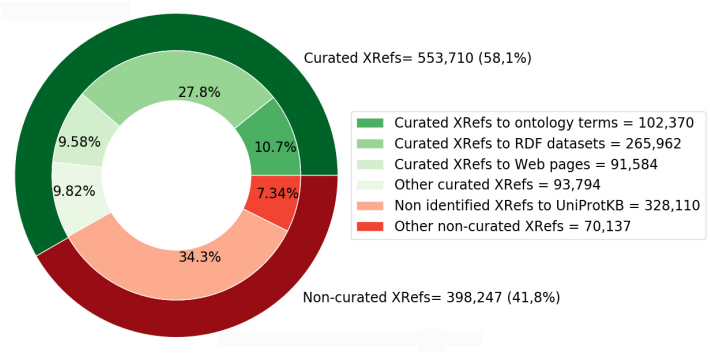

Figure 2. Number and repartition of curated and non-curated XRefs.

4.1. Curated XRefs analysis

With OMHT's semi-automatic approach, we were able to identify a total of 553,710 curated XRefs, around 58% of the XRefs found. As illustrated in Figure 2, the majority of the curated targets reference external RDF datasets (27,8%) and only 10,7% of the XRefs are explicit mappings between ontology terms. This first result is important as it tells us two things: (i) XRefs are mainly not ontology mappings (89,3% are something else) and (ii) 10,7% of XRefs as mappings between ontology terms is already a quite significant number of mappings that could be very useful for the community.

4.1.1. Curated XRefs to ontology terms

OMHT successfully extracted a total of 102,370 XRefs (10,7%) mappings to ontology terms. We classify curated XRefs mappings to ontology terms as follows:

- Internal mapping: when the target ontology is actually also hosted in AgroPortal;
- Inter-portal mapping: when the target ontology is stored in another portal of the OntoPortal family (mostly the NBCO BioPortal in our case);
- External mapping: when the target ontology is simply identified by its URI and not a local identifier in AgroPortal or BioPortal.

The majority of mappings to ontology terms (68,400 mappings in 26 ontologies) are internal mappings. There are 9,417 inter-portal mappings and 24,553 external mappings. This observation states the good interconnection between AgroPortal ontologies, and was for us a proof of coherence of the domain covered. A large part of the mappings (58,529) is between the NCBI Taxonomy and Gramene Taxonomy. The rest of the top-aligned ontologies are described in Table 1. In the following, is an example of an XRef found in the Plant Trait Ontology (TO) which captures a mapping to another term in the Plant Ontology (PO):

```
<owl:Class rdf:about="http://purl.obolibrary.org/obo/TO_0000017">
   <oboInOwl:hasDbXRef rdf:datatype="xsd:string">PO:0025131
   </oboInOwl:hasDbXref> </owl:Class>
```

The mapping extracted by OMHT, after the curation step, which disambiguates/validates the PO prefix, represented in OntoPortal JSON format is:

```
{"creator":     "http://data.agroportal.lirmm.fr/user/mappingadmin",
 "relation":    ["http://www.geneontology.org/formats/oboinowl#hasdbxref"],
 "source":      "http://data.agroportal.lirmm.fr/ontologies/TO",
 "comment":     "Generated using OMHT v.1.3 - 07/04/2020 14:42 CEDT",
 "classes": {   "http://purl.obolibrary.org/obo/PO_0025131": "PO",
                "http://purl.obolibrary.org/obo/TO_0000017": "TO"}}
```

In this representation, ontologies are identified by their acronyms in AgroPortal (a locally unique identifier), but it would be straightforward to use the URI of the ontologies instead. In this example, and very often, identifying the target ontology was straightforward as XRefs formally uses OBO libraries defined prefixes. Other examples required some human investigation. In the following, are examples of miscellaneous ontology targets found in the corpus:

- From GR-TAX, TO, EO and PECO ontologies, OMHT extracted 58,795 XRefs to `ncbi_taxid`. Manual curation resolved this prefix to the NCBI Taxonomy. However, CL and EO use another prefix `ncbitaxon` and Food ontology uses the full URI: `http://purl.obolibrary.org/obo/ncbitaxon`. This mix of prefixes referencing the same ontology increased the complexity of the curation process.
- From Food ontology, OMHT extracted 9,368 XRefs to textttsubset_siren. Manual curation of this prefix resolved this prefix to the LANGUAL thesaurus. The identification of this prefix was significantly longer.

Table 1. Principal targets cross-referenced by ontologies using XRefs.

Source ontology	**Target ontology or semantic resource**	**Number of mappings**
FOODON	LANGUAL	12,075
CL	FMA	5,985
CL	EMAPA	2,461
PECO	CHEBI	1,155
Source ontology	**RDF dataset**	**Number of XRefs**
GO	Enzime Portal (`http://identifiers.org/ec-code`)	6,495
GO	Rhea (`http://rdf.rhea-db.org`)	3,157
PECO	UniProtKB (`www.uniprot.org`)	183
TO	UniProtKB (`www.uniprot.org`)	103
Source ontology	**Web site or databases**	**Number of XRefs**
PR	PANTHER (`www.pantherdb.org`)	11,814
GO	Reactome (`https://reactome.org`)	5,588
PR	EcoCyc E. coli Database (`https://ecocyc.org`)	4,415
FOODON	Encyclopedia of Life (`http://eol.org`)	2,820
Source ontology	**Other curated XRefs**	**Number of XRefs**
GO	goc	58,529
CL	goc	5,072
PO	poc	2,611
TO	poc	2,608

4.1.2. Curated XRefs to RDF datasets

OMHT successfully extracted a total of 265,962 XRefs (27,8%) to RDF dataset elements (i.e., an entity in an RDF dataset identified by a specific URI) like DBpedia or UniProtKB. We count a total of 9 curated target RDF datasets. The top cross-referenced RDF dataset is UniProtKB, with 221,462 cross-references from the Protein Ontology (PR). We also found 33,639 cross-references from the same ontology to the OMA orthology database. The rest of the most cross-referenced RDF datasets are described in Table 1. In

the following, is an example of an XRef found in PR which captures a cross-reference to an entry in the UniProtKB dataset:

```
<owl:Class rdf:about="http://purl.obolibrary.org/obo/PR_A0A0B4K7J2">
   <oboInOwl:hasDbXref rdf:datatype="xsd:string"> UniProtKB:A0A0B4K7J2
   </oboInOwl:hasDbXref> </owl:Class>
```

Identifying the target (`http://purl.uniprot.org/uniprot/`) was not straightforward and OMHT relied on Identifiers.org to automatically find a candidate target which was then manually validated. The URI pattern had also to be validated to be sure to non-ambiguously identify the right cross-referenced entity (in this case `http://purl.uniprot.org/uniprot/A0A0B4K7J2`). After is a series of examples of miscellaneous RDF dataset targets found in the corpus, often as the form of base URI because in these cases the `prefix:id` notation was not much used: DBPedia (`http://dbpedia.org/resource/`), Deutsche National Bibliothek (`http://d-nb.info/gnd/`), UMTHES environmental thesaurus (`http://data.uba.de/umt/`).

4.1.3. Curated XRefs to Web pages

OMHT successfully extracted a total of 91,584 XRefs (9,58%) to Web pages (i.e., a Web document identified by a specific URL) like a Wikipedia entry or an ISBN database entry. We count a total of 20 curated Web site or database targets. The top cross-referenced Web site is Reactome, with 13,130 Web links from PR. The rest of the most cross-referenced web sites and databases are described in Table 1. For instance, we found 1,761 cross-references from the Environment Ontology (ENVO) to Wikipedia pages:

```
<owl:Class rdf:about="http://purl.obolibrary.org/obo/ENVO_00000020">
    <oboInOwl:hasDbXref rdf:datatype="xsd:string">
        https://en.wikipedia.org/wiki/Lake
    </oboInOwl:hasDbXref> </owl:Class>
```

For this example, there is no use of the `prefix:id` notation. The `oboInOwl:hasDbXref` property directly contains the URL to the Web page. As another example, we identified in the Gene Ontology (GO) 1,761 references to books identified with an ISBN. For example:

```
<owl:Class rdf:about="http://purl.obolibrary.org/obo/GO_0000003">
   <oboInOwl:hasDbXref rdf:datatype="xsd:string"> ISBN:0198506732
   </oboInOwl:hasDbXref> </owl:Class>
```

In this example, there is no direct usage of the URL as illustrated in ENVO before. Here, the GO developers employed the XRef property to annotate terms with related books and extrapolated the `prefix:id` notation to reference the books. These types of target had of course to be manually identified. When possible, the curator also identified a possible specific Web page for the corresponding targeted element and specified the pattern for OMHT to generate proper cross-reference URLs. In the previous example, we were able to identify `ISBNdb.com` database and the specific URL: `https://isbndb.com/book/0198506732`. A similar situation was found with PubMed citations references, where the '`pmid`' or '`pubmed`' prefixes were used (28,735 Xrefs in 19 ontologies) and we transformed them to the proper URL in PubMed (e.g., `https://pubmed.ncbi.nlm.nih.gov/22654893`).

4.1.4. Other curated XRefs

In complement, OMHT extracted a total of 93,794 other XRefs (9,82%) in 165 targets, which were not classified as ontology term, RDF datasets or Web pages. In these targets, we surprisingly discovered 83,592 XRefs documenting the curators of ontologies terms (e.g., `goc`, `poc`) i.e., the person considered as the creator or validator of a term. We identified 9 such targets in 19 ontologies. Principal ontologies exposing this issue is described in Table 1. For example, the Gene Ontology (GO) contains 58,066 XRefs where the `prefix:id` notation is non consistently extrapolated to identify one curator (e.g., `GOC:jl`), a group (e.g., `GOC:go_curators`) or even a method (e.g., `GOC:isa_complete`). The term 'reproduction"(`GO_0000003`) for instance, contains the three situations:[6]

```
<owl:Class rdf:about="http://purl.obolibrary.org/obo/GO_0001558">
   <oboInOwl:hasDbXref rdf:datatype="xsd:string""> GOC:go_curators
   </oboInOwl:hasDbXref> </owl:Class>
```

Identifying curators out of XRefs was not easy, and might not be error-proof; plus in this case, we were not able to assign the identified targets an unambiguous identifier as we did before with URIs or URLs. We have reported the problem to the OBO Foundry for follow-up[7]. Other examples of curated XRefs include go_ref, which documents the author of terms via GitHub web pages. In the following, examples of other curated XRefs:

- From GO, OMHT identified 3,885 XRefs to "`go_ref`" (e.g., `go_ref:0000022`);
- From PECO and TO ontology, OMHT identified 666 Xrefs to the prefix "`kegg_compound`" with no id values;
- In a similar way, from PECO ontology, OMHT identified 534 Xrefs to "`iupac`".

4.2. Non-curated XRefs Analysis

Among the non-curated XRefs, we count 328,110 (34,3%) occurrences (82% of the total of non-curated XRefs) from the Protein Ontology to the UniProt Knowledge Base (UniProtKB), which is an external resource for protein-related information. We were not able to identify frequent target entities such as '`PRO:DNx`', '`PRO:JR`'.

```
<owl:Class rdf:about="http://purl.obolibrary.org/obo/PR_000035415">
    <oboInOwl:hasDbXref rdf:datatype="xsd:string"> PRO:DNx
</oboInOwl:hasDbXref></owl:Class>
```

For the rest of the non-curated cross-references (70,137), we were not able to identify the nature of the ambiguous targets (over a thousand) in reasonable curation time. For the majority of targets found by OMHT on our corpus, half of them appear only one or two times. We have discarded the less frequent occurrences and focused on the most frequent ones. In the following, examples of non identified frequent targets that follow the `prefix:id` notation include:

[6] We have identified this issue in 2006 terms of this ontology.

[7] `https://github.com/OBOFoundry/OBOFoundry.github.io/issues/826`

- From TO, PO, FLOPO and PPO ontologies, OMHT identified 3,581 XRefs to "po_git" (e.g., po_git:511)[8];
- From PR and CL ontologies, OMHT identified 2,899 XRefs to IUPHARobj (e.g., IUPHARobj:250);
- From TRIPHASE thesaurus, OMHT identified 2,655 to "TyDI_semClass" (e.g., TyDI_semClass:4928444).

As a more detailed example, we found in the OntoBiotope ontology and the TRIPHASE thesaurus, 6100 XRefs prefixed by 'tydi' (e.g., tydi:4927819). OMHT was not able to find this prefix in Identifiers.org nor the OBO foundry or an ontology repository. We understood from exchanging with the developers of these semantic resources that TyDI was a tool to assist in terminology extraction and these XRefs were capturing some kind of design/provenance information. Of course, we were not able to contact the developers of each ontology in which such a situation of ambiguous XRefs targets occurred. Among the non-curated XRefs targets are many occurrences of specific URLs that appear only once or twice in the corpus of XRefs. Although they probably represent valid XRefs to Web pages, there are too many of them to be manually curated.

5. Discussion and recommendations

Our results show XRefs are largely employed by the community but hardly reusable without heavy curation. With OMHT we were able to curate 58.3% of the identified XRefs which means more than 40% of XRefs in our corpus are unusable, modulo a significantly longer time for curation (e.g., to contact ontology developer one by one). We believe there are two major ways of enhancing the value and reusability of XRefs:

- Using more relevant specific properties taken from standard semantic Web metadata vocabularies to replace the tote-bag oboInOwl:hasDbXrefs property. For instance, using the pav:curatedBy property to annotate a term with a curator or skos:exactMatch to map to another ontology term.
- Using URIs or permanent URLs as much as possible to identify external entities. For instance, using the ORCID or DOI permanent URL rather than the name of the researcher or a miscellaneous reference to a digital object.

Searching for the right prefixes and ids is a complex and time-consuming task, which affects any automation. We believe our study show the limits of the prefix:id notation to which the use of unambiguous and permanent identifiers shall be preferred. Identifiers.org helped in the process of resolving prefixes, but still sometimes ambiguity remains and XRefs cannot be resolved. And getting the right prefix is only half of it.

OMHT extracted 102,370 mappings to ontology terms and 265,962 XRefs to RDF dataset elements. In both cases, URIs always exists for the targets. Sometimes URIs are used; why not systematically use them rather than an ambiguous prefix:id form? OMHT extracted 91,584 XRefs to specific Web URLs either directly (e.g., Wikipedia pages) or indirectly (e.g., ISBN or PMID). However, ontology developers are not following any clear guidelines to link their term to external "online entities". In some cases, the prefix:id solution works fine (e.g,. ISBN:0198506732), assuming the prefix can be resolvable , but still this will require to always rely on an external resolving service, whereas in many cases permanent URLs are already supported by the providers (e.g.,

[8]Those ones were assumed to reference a GitHub tracker issue maybe were the term was debated. We have not investigated them further yet

https://isbndb.com/book/[ID], https://doi.org/[ID] or https://orcid.org/[ID]). We recommend the use of specific annotation properties such as `rdfs:seeAlso` to link to a related Web page or `dct:bibliographicCitation` to annotate a term with a citation. Using specific annotation properties will result in a richer semantic integration.

Finally, in the case of XRefs documenting the curators of ontology terms, one may prefer a property from the Provenance Ontology (a W3C Recommendation) or the Provenance, Authoring and Versioning vocabulary that are recognized metadata standards: `pav:curatedBy`, `pav:authoredBy`, `pav:createdBy`, `prov:wasAttributedTo`, `pav:contributedBy`. The use of these properties will help humans and machines to better identify and credit ontology contributors. In Table 2, we propose some ways to mitigate the use of ambiguous XRefs by employing more relevant specific properties taken from standard semantic Web metadata vocabularies.

Table 2. Property and value recommendations to mitigate the cases of ambiguous XRefs.

Situation	Property recommendation	Value recommendation
Mapping to another term	Use `owl:sameAs` or `owl:equivalentClass` or SKOS mapping properties (`skos:exactMatch`, `skos:closeMatch`) depending on the semantics.	Use the URI of the term in the target ontology.
XRef to RDF dataset element	Use `owl:sameAs` if the entities are semantically equivalent. Prefer a more precise property to describe the exact relation between the ontology term and the element (`dct:isVersionOf`, `pav:derivedFrom`, etc.).	Use the URI of the element in the target RDF dataset.
Link to Web page	Use `rdfs:seeAlso` or `dcat:landingPage` depending on the meaning of the link.	Use directly the URL of the page as identifier.
Link to citations	Use `dct:bibliographicCitation`	Use the DOI or ISBN permanent URLs
Annotating curator of a term	Use a property from PROV-O or PAV such as `pav:curatedBy`, `pav:authoredBy`, `pav:createdBy`, `prov:wasAttributedTo`, `pav:contributedBy`	Use ORCID permanent URLs or IDs specific to a system but somehow resolvable (e.g., GitHub accounts).
Link to database as a whole	Use a specific property to describe the relation between the term and the database. (e.g., `schema:includedInDataCatalog` or `dct:source`, `foaf:page`).	Use the best known identifiers for the database if available or directly the database URL.

6. Conclusion

Mappings between ontology terms and cross-references between a term and an external database element or Web page are a very useful means for data integration and interoperability. These links are the backbone of the Linked Open Data vision [2]. However, today, in the OBO world, the use of XRefs prevent the valorization of the information originally captured by ontology developers; XRefs are poorly represented and semantically ambiguous. Miscellaneous uses of XRefs range from cross-references to database elements or the databases as a whole, curators of a term, bibliographic citations, and in some cases, mappings to related ontology terms.

In this paper, we have studied the use of XRefs in a corpus of 30 ontologies with the original motivation to extract ontology mappings in the AgroPortal ontology repository. We have designed and developed OMHT, a tool to semi-automatically extract and harmonize XRefs from AgroPortal ontologies. Nevertheless, out of almost one million extracted XRefs, we have identified only 10,7% of them representing the mapping between terms which confirms that XRefs are primarily something other than inter-ontology mappings. When ontology developers use XRefs to reuse a term from another ontology, OMHT resolves the targets quite easily. However, for the other 89,3%, XRefs are quite ambiguous to be resolved by OMHT. Our study quantifies the different cases of uses of XRefs while establishing the mandatory curation needed to reuse them. Even with manual curation, we were unable to give sense to a bit more than 40% of the XRefs in our corpus, which demonstrate the loss of information and efforts originally invested

by ontology developers. We have discussed recommendations on means to mitigate this situation, mostly by using more relevant metadata properties and identifiers.

To get more insights about the use of the XRefs and refine our recommendations, as future work, we plan to analyse the usage of XRefs in a bigger corpus of ontologies from the "OBO world". We will however, make a distinction between the ontologies formally declared in the OBO library (`www.obofoundry.org`) an the others. This will help the OBO community to identify the issues and possibly move forward. Our work could also be extended to study other ways of capturing ontology cross-references (not only XRefs).

Acknowledgements

This work was achieved with support of the AGRO Labex (ANR-10-LABX-0001), the NUMEV Labex (grant ANR-10-LABX-20) and the Data to Knowledge in Agronomy and Biodiversity (D2KAB – `www.d2kab.org`) project that received funding from the French National Research Agency (grant ANR-18-CE23-0017).

References

[1] Amir Laadhar, Elcio Abrahão, C.J.: Systematic analysis of term reuse, term overlap and mapping extracted across agronomy and biodiversity semantic resources. In: 22nd International Conference on Knowledge Engineering and Knowledge Management, EKAW'20 (UNDER SUBMISSION) (2020)

[2] Di Noia, T., Mirizzi, R., Ostuni, V.C., Romito, D., Zanker, M.: Linked open data to support content-based recommender systems. In: 8th International Conference on Semantic Systems. pp. 1–8 (2012)

[3] Euzenat, J., Shvaiko, P.: Ontology matching, Second edition. Springer-Verlag (2013)

[4] Ghazvinian, A., Noy, N.F., Musen, M.A.: Creating Mappings For Ontologies in Biomedicine: Simple Methods Work. In: American Medical Informatics Association Annual Symposium, AMIA'09. pp. 198–202. Washington DC, USA (Nov 2009)

[5] Jonquet, C., Toulet, A., Arnaud, E., Aubin, S., Dzalé-Yeumo, E., Emonet, V., Graybeal, J., Musen, M.A., Pommier, C., Larmande, P.: Reusing the NCBO BioPortal technology for agronomy to build AgroPortal. In: Jaiswal, P., Hoehndorf, R. (eds.) 7th International Conference on Biomedical Ontologies, ICBO'16, Demo Session. CEUR Workshop Proceedings, vol. 1747, p. 3. Corvallis, Oregon, USA (August 2016)

[6] Jonquet, C., Toulet, A., Arnaud, E., Aubin, S., Yeumo, E.D., Emonet, V., Graybeal, J., Laporte, M.A., Musen, M.A., Pesce, V., Larmande, P.: AgroPortal: a vocabulary and ontology repository for agronomy. Computers and Electronics in Agriculture 144, 126–143 (Jan 2018)

[7] Jupp, S., Liener, T., Sarntivijai, S., Vrousgou, O., Burdett, T., Parkinson, H.: OxO — A Gravy of Ontology Mapping Extracts. In: 8th International Conference on Biomedical Ontology, ICBO'17. CEUR Workshop Proceedings, vol. 2137, p. 2. Newcastle, UK (Sep 2017)

[8] Juty, N., Novère, N.L., Laibe, C.: Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. Nucleic Acids Research 40(D1), 580–586 (2011)

[9] Kamdar, M.R., Tudorache, T., Musen, M.A.: Investigating term reuse and overlap in biomedical ontologies. International Conference on Biomedical Ontology (2015)

[10] Kamdar, M.R., Tudorache, T., Musen, M.A.: A systematic analysis of term reuse and term overlap across biomedical ontologies. Semantic web 8(6), 853–871 (2017)

[11] McMurry, J.A., Juty, N., Blomberg, N., Burdett, T., Conlin, T., Conte, N., Courtot, M., Deck, J., Dumontier, M., Fellows, D.K., et al.: Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. PLoS biology 15(6) (2017)

[12] Mungall, C.J., Koehler, S., Robinson, P., Holmes, I., Haendel, M.: k-boom: A bayesian approach to ontology structure inference, with applications in do construction. BioRxiv p. 048843 (2016)

[13] Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., et al.: The obo foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology 25(11), 1251–1255 (2007)

[14] Whetzel, P.L., Shah, N.H., Noy, N.F., Dai, B., Dorf, M., Griffith, N.B., Jonquet, C., Youn, C.H., Coulet, A., Callendar, C., Rubin, D.L., Smith, B., Storey, M.A., Chute, C.G., Musen, M.A.: BioPortal: Ontologies and Integrated Data Resources at the Click of a Mouse. In: Bio-Ontologies: Knowledge in Biology, SIG, Poster session, ISMBECCB'09. Stockholm, Sweden (Jul 2009)