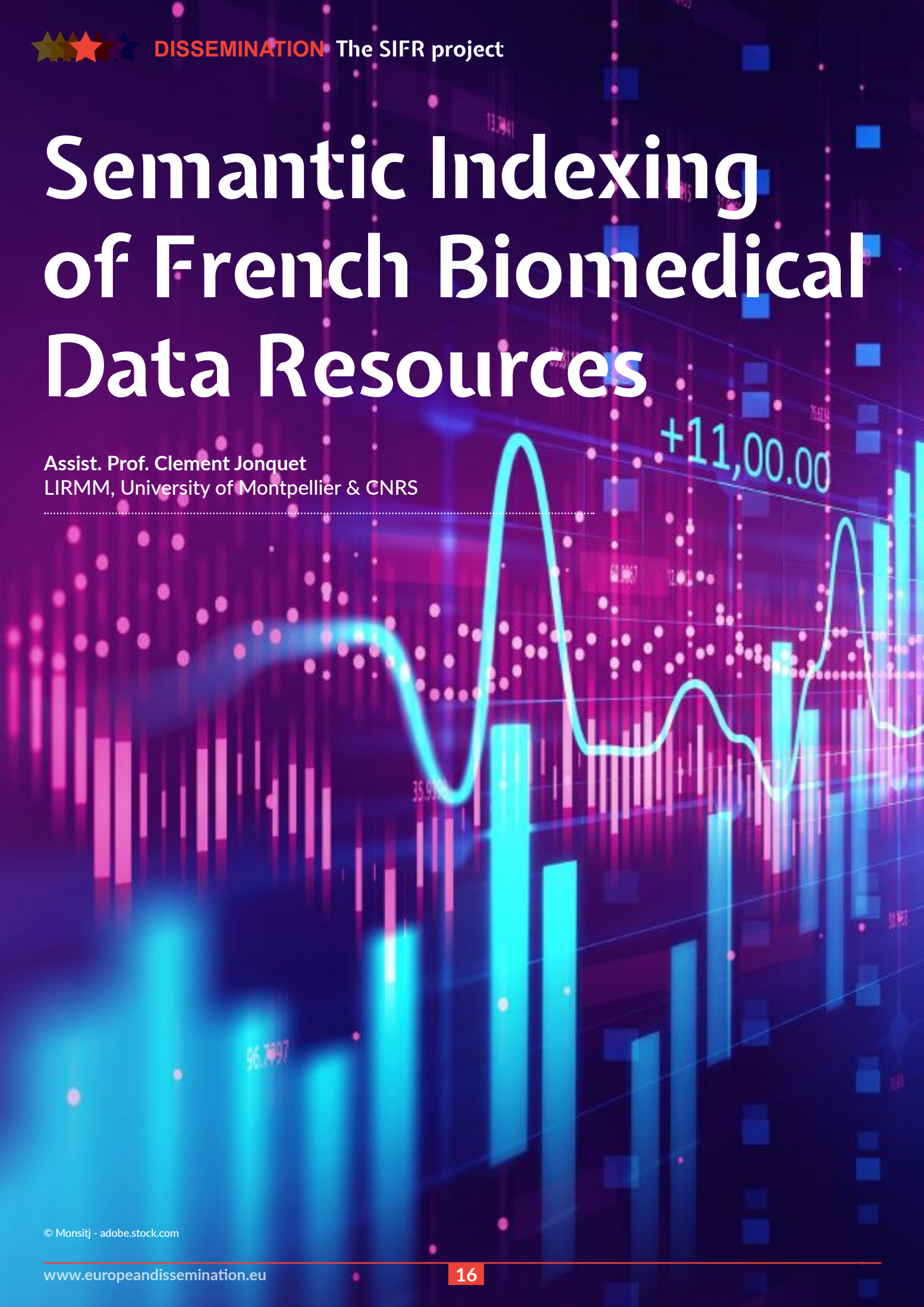
Semantic Indexing of French Biomedical Data Resources

Clement Jonquet

Semantic Indexing of French Biomedical Data Resources

Assist. Prof. Clement Jonquet
LIRMM, University of Montpellier & CNRS



The volume of data in biomedicine is constantly increasing. Despite a large adoption of English in science, a significant quantity of these data uses the French language. Biomedical data integration and semantic interoperability are necessary to enable new scientific discoveries that could be made by merging different available data.

The community has turned to ontologies and terminologies to design semantic indexes of data that leverage the medical knowledge. However, besides the existence of various English tools, there are considerably fewer ontologies available in French and there is a strong lack of related tools and services to exploit them. This lack does not match the huge amount of biomedical data produced in French, especially in the clinical world, e.g., electronic health records. The Semantic Indexing of French Biomedical Data Resources (SIFR) project investigates the scientific and technical challenges in building ontology-based services to leverage biomedical ontologies and terminologies in indexing, mining and retrieval of French biomedical data (Figure 1). Within the project, we work on several research questions from semantic indexing, text mining, terminology extraction, ontology enrichment, disambiguation, ontology alignment, multilingualism in ontologies and semantic annotation in order to offer the community with services and applications capable of leveraging the use of biomedical ontologies in their data workflows.

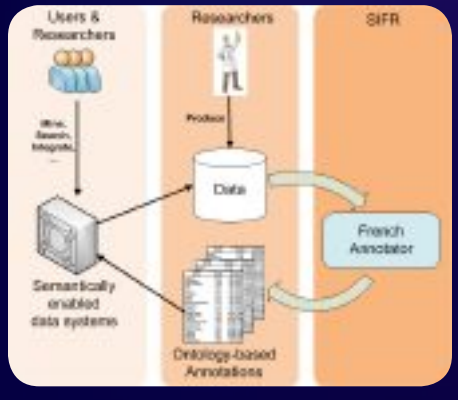


Figure 1. Using the SIFR 'French' Annotator to semantically index biomedical data

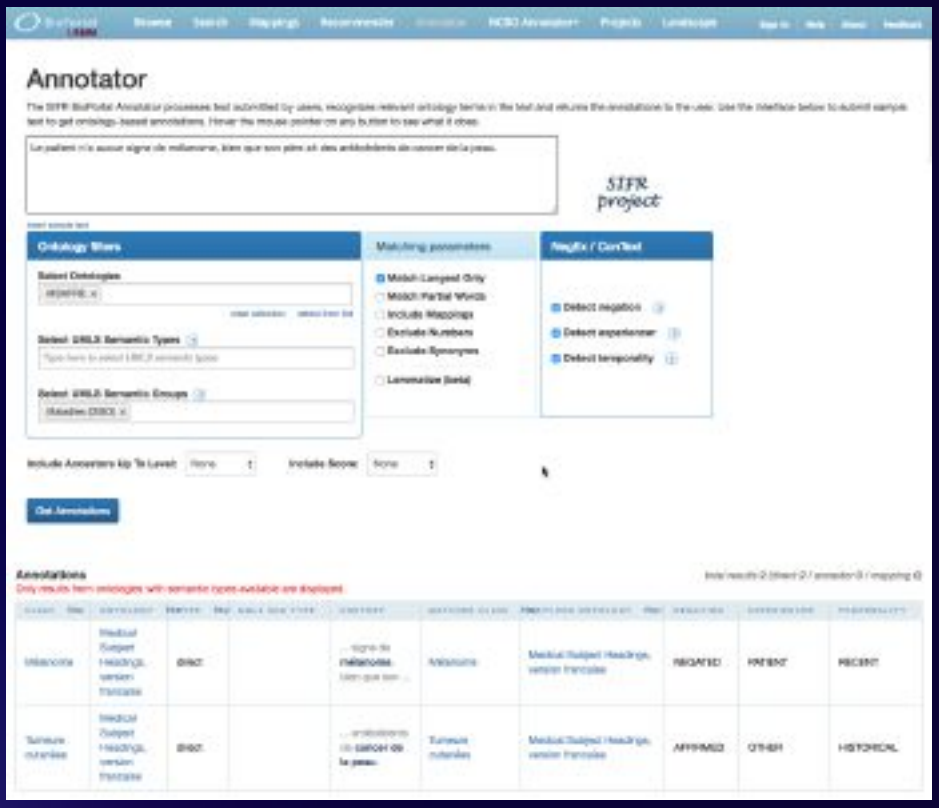


Figure 2. SIFR Annotator interface in the SIFR BioPortal

Design and implementation of the SIFR Annotator

Researchers have called out the need for automated semantic annotation methods and for leveraging natural language processing tools in the curation process. Within SIFR, we build an ontology-based indexing workflow, i.e., SIFR Annotator, similar to that which exists for English resources but dedicated and specialised for French.

The SIFR Annotator (http://bioportal.lirmm.fr/annotator) is a publicly accessible ontology-based annotation tool to process biomedical text data in French. It is a web service that for a given piece of text will return biomedical ontology concepts directly mentioned in the text or semantically expanded. The service overpasses the limits of keyword-based approaches and uses the semantics that the ontologies encode, such as different properties of classes, the class hierarchies, and mappings, in order to improve the annotation experience for our users. The interface of the SIFR Annotator is illustrated in Figure 2. A complete description and evaluation of the tool is available in a 30–page journal published in BMC Bioinformatics (Tchechmedjiev et al., 2018a).

The Annotator is available within the SIFR BioPortal (http://bioportal.lirmm.fr) (Jonquet et al., 2016) a repository of 30 French biomedical ontologies/terminologies which reuses the US National Center for Biomedical Ontology BioPortal technology (Noy et al., 2009), developed at Stanford University. Ontologies have been offered by the CISMEF group from Rouen University Hospital, or taken from the UMLS Meta-thesaurus – a database of medical terminologies maintained by the US National Library of Medicine – or directly uploaded by users. The repository facilitates use and fostering of ontologies by offering a set of services (search, mappings, metadata, versioning, visualisation, recommendation), including for annotation purposes.

In building the SIFR BioPortal and Annotator our vision was to embrace semantic web standards and promote openness and easy

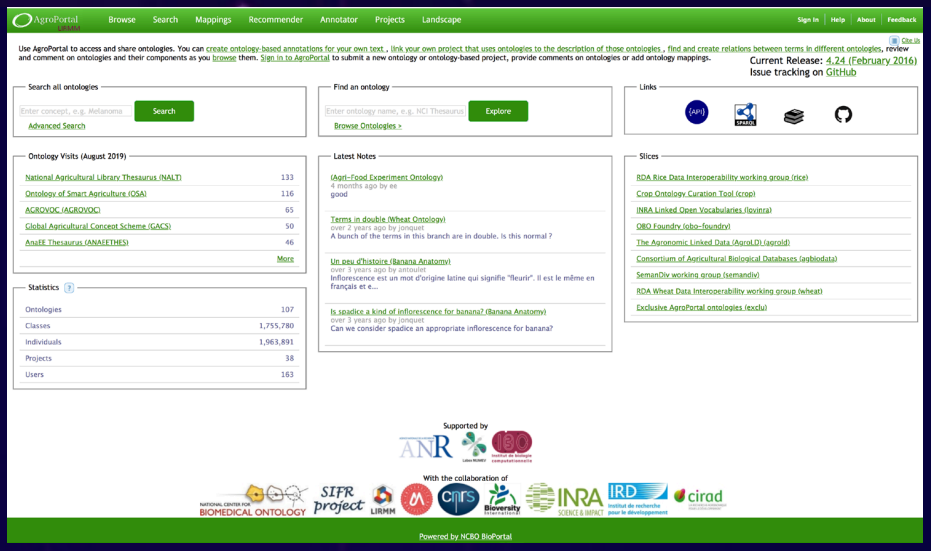


Figure 3. AgroPortal homepage

access. There was no such service publicly available in France (for French data). In addition to the SIFR Annotator and in collaboration with Stanford, we developed a proxy web service for the NCBO Annotator to process English data and offer new features that have been investigated and implemented within SIFR. We have then implemented enhanced functionalities for annotating and indexing free text such as: scoring, detection of context (negation, experiencer, temporality), new output formats and coarse-grained concept recognition. This work is published in Bioinformatics (Tchechmedjiev et al., 2018b).

Semantic Indexing of French Electronic Health Records

In the context of the ANR funded PractiKPharma partner project (Practice-based evidences for actioning Knowledge in Pharmacogenomics – http://practikpharma.loria.fr), the SIFR team investigated the processing of French/English clinical data and developed specific features for the SIFR Annotator to address the needs of our partner, the European Hospital G. Pompidou. When annotating clinical text, the context of the annotated clinical conditions is crucial: distinguishing between affirmed and negated conditions, e.g., 'no sign of cancer'; whether a condition pertains to the patient or to others, e.g., family members; or temporality (is a condition recent or historical). In a national publication (Abdaoui et al., 2017) (extended revised version under submission to Journal of Biomedical Informatics), we present French ConText: an adaptation and enrichment of NegEx/ConText (Harkema et al., 2009) to the French language. We integrated French ConText in SIFR Annotator, and thanks to the proxy architecture plugged the original ConText (for English) in the NCBO Annotator (Tchechmedjiev et at., 2018b). We offer now, both for English and French, a unique open ontology-based annotation service that both recognise ontology concepts and contextualise them allowing non-natural-language-processing experts to both annotate and contextualise medical conditions in clinical notes.

Experiments in agronomy with AgroPortal

Many vocabularies and ontologies are produced to represent and annotate agronomic data. However, those ontologies are spread out over the web (or even unshared), in many different formats and types, of different size, with different structures and from overlapping domains. Therefore, there is a need for a common platform to receive and host them, align them, and enable their use in agro-informatics applications. The AgroPortal project (http://agroportal.lirmm.fr) (Jonquet et at., 2018), is a community effort started by the Montpellier scientific community (LIRMM, IRD, CIRAD, INRA, Bioversity International) and in partnership with Stanford to build an ontology repository for agronomy and related domains. Our goal is to facilitate the adoption of metadata and semantics to facilitate open science and produce Findable Accessible Interoperable and Reusable data. Within SIFR, we are reusing the scientific outcomes and experience of the biomedical domain in the context of agronomy, plant sciences, food and biodiversity. By enabling straightforward use of ontologies, we expect data managers and researchers to focus on their tasks, without requiring them to deal with the complex engineering work needed for ontology management. The repository (see Figure 3) currently hosts 108 vocabularies or ontologies with more than 2/3 of them not present in any similar ontology repository and 11 private ontologies. We have identified 80 other candidate ontologies that will be loaded in the future to complement this valuable resource. The platform already has more than 170 registered users and some vocabularies are visited more than 100 times per month.

Concluding remarks

SIFR offers the French biomedical community, e.g., clinicians, health professionals and researchers, highly valuable ontology-based services that will enhance their data production and consumption workflows. By evaluating and comparing the SIFR Annotator to state-of-the-art results (Tchechmedjiev et al., 2018a), we showed the web service performs comparably to other knowledge-based annotation approaches in recognising entities in biomedical text and reach state-of-the-art levels in clinical context detection (negation, experiencer, temporality). Additionally, the SIFR Annotator is the first openly accessible web tool to annotate and contextualise French biomedical text with ontology concepts leveraging a dictionary currently made of 30 terminologies and ontologies and 380K concepts. SIFR BioPortal has become the largest generic and open (with publicly access resources, code and related data) French-language biomedical ontology

and terminology repository in France. In turn, SIFR Annotator is today the richest French language open annotator web service (competing annotators are either not available or closed-source online services). We are currently developing several partnerships in France to use SIFR Annotator within hospitals (CHRU Nancy, G. Pompidou European Hospital in Paris) or for large-scale annotation efforts, e.g., to annotate the corpus of course of the French national medicine curriculum in the SIDES 3.0 project.

We are also transferring our results in the agronomic domain by kicking-off the AgroPortal project. AgroPortal will be a core component of a new ANR funded project starting mid-2019 called D2KAB (Data To Knowledge in Agronomy and Biodiversity – www.d2kab.org). D2KAB's primary objective is to create a framework to turn agronomy and biodiversity data into semantically described, interoperable, actionable, open knowledge, along with investigating scientific methods and tools to exploit this knowledge for applications in science and agriculture. Agronomy/agriculture and biodiversity face several major societal, economical, and environmental challenges; a semantic data science approach will help to address these. We shall provide the means (ontologies and linked open data) for agronomy and biodiversity to embrace the semantic web to produce and exploit FAIR data. This new 3.1M€ (950K€ of ANR support) project gathers ten French partners for four years. Each of the project driving scenarios (food packaging, agro-agri linked data, wheat phenotype, ecosystems and plant biogeography) will have a significant impact and produce concrete outcomes for agronomy and biodiversity scientific communities and socio-economic actors in agriculture.

References

1. Tchechmedjiev, A., Abdaoui, A., Emonet, V., Zevio, S., Jonquet, C.: SIFR Annotator: Ontology-Based Semantic Annotation of French Biomedical Text and Clinical Notes. BMC Bioinformatics. 19, 405–431 (2018).
2. Jonquet, C., Annane, A., Bouarech, K., Emonet, V., Melzi, S.: SIFR BioPortal : Un portail ouvert et générique d'ontologies et de terminologies biomédicales françaises au service de l'annotation sémantique. In: 16th Journées Francophones d'Informatique Médicale, JFIM'16. p. 16. , Genève, Suisse (2016).
3. Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N.B., Jonquet, C., Rubin, D.L., Storey, M.-A., Chute, C.G., Musen, M.A.: BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research. 37, 170–173 (2009).
4. Tchechmedjiev, A., Abdaoui, A., Emonet, V., Melzi, S., Jonnagaddala, J., Jonquet, C.: Enhanced Functionalities for Annotating and Indexing Clinical Text with the NCBO Annotator+. Bioinformatics. 34, 1962–1965 (2018).
5. Abdaoui, A., Tchechmedjiev, A., Digan, W., Bringay, S., Jonquet, C.: French ConText: Détecter la négation, la temporalité et le sujet dans les textes cliniques Français. In: 4ème Symposium sur l'Ingénierie de l'Information Médicale, SIIM'17. p. 10. , Toulouse, France (2017).
6. Harkema, H., Dowling, J.N., Thornblade, T., Chapman, W.W.: ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. Journal of biomedical informatics. 42, 839–51 (2009).
7. Jonquet, C., Toulet, A., Arnaud, E., Aubin, S., Dzalé Yeumo, E., 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, (2018).

SIFR project

SUMMARY

The SIFR project investigates scientific and technical challenges in building ontology-based services to leverage biomedical ontologies and terminologies in indexing, mining and retrieval of biomedical data. Our main goal is to enable straightforward use of ontologies freeing health researchers to deal with knowledge engineering issues and focus on biological/medical challenges. An ontology-based indexing workflow was built to annotate French biomedical data. The project also abstracted and generalised results by offering a vocabulary and ontology repository for agronomy and related domains called AgroPortal. Our goal is to encourage the adoption of metadata and semantics to facilitate open science and produce FAIR data.

PROJECT PARTNERS:

SIFR (2013-2019) was a collaborative action between LIRMM (University of Montpellier) and BMIR (Stanford University). Partners include researchers from TETIS (CIRAD & IRSTEA) and CHU Rouen.

PROJECT LEAD PROFILE:

Dr. C. Jonquet is the Principal Investigator (PI) of the SIFR project and an associate professor (HDR) at the University of Montpellier. He has 12 years experience in ontologies and semantic Web research applied to biomedicine and agronomy. He works on the design and development of ontology repositories and ontology-based services, especially semantic annotation as the project PI.

CONTACT DETAILS

Clement Jonquet

+33 467 14 97 43

jonquet@lirmm.fr

www.lirmm.fr/sifr

FUNDING

This project has received funding from the French National Research Agency (ANR) under the Young Researcher program grant ANR-12-JS02-01001 as well as by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 701771 (Marie Skłodowska-Curie Individual Fellowship).