Leveraging Concepts in Open Access Publications

Andrea Bertino, Luca Foppiano, Laurent Romary, Pierre Mounier

Leveraging Concepts in Open Access Publications

Andrea Bertino,[1] *Luca Foppiano, Laurent Romary, Pierre Mounier*
Göttingen State and University Library, Göttingen, Germany; ALMAnaCH, Inria, Paris, France;
OpenEdition, EHESS, Paris, France.

ABSTRACT
Aim: This paper addresses the integration of a Named Entity Recognition and Disambiguation (NERD) service within a group of open access (OA) publishing digital platforms and considers its potential impact on both research and scholarly publishing. This application, called *entity-fishing*, was initially developed by Inria in the context of the EU FP7 project CENDARI (Lopez et al., 2014) and provides automatic entity recognition and disambiguation against Wikipedia and Wikidata. Distributed with an open-source licence, it was deployed as a web service in the DARIAH infrastructure hosted at the French HumaNum.
Methods: In this paper, we focus on the specific issues related to its integration on five OA platforms specialized in the publication of scholarly monographs in social sciences and humanities as part of the work carried out within the EU H2020 project HIRMEOS (*High Integration of Research Monographs in the European Open Science infrastructure*).
Results and Discussion: In the following sections, we give a brief overview of the current status and evolution of OA publications and how HIRMEOS aims to contribute to this. We then give a comprehensive description of the entity-fishing service, focusing on its concrete applications in real use cases together with some further possible ideas on how to exploit the generated annotations.
Conclusions: We show that *entity-fishing* annotations can improve both research and publishing process. *Entity-fishing annotations* can be used to achieve a better and quicker understanding of the specific and disciplinary language of certain monographs and so encourage non-specialists to use them. In addition, a systematic implementation of the entity-fishing service can be used by publishers to generate thematic indexes within book collections to allow better cross-linking and query functions.

Keywords: Named Entity Recognition and Disambiguation (NERD), Entity-Fishing, Open Access, Monographs, Digital Publishing Platforms

1 Challenges and Perspectives for OA Digital Monographs

The publication of scholarly monographs in OA is of great benefit to researchers, both as authors and as readers: it increases the visibility of their publications, ensures a wider dissemination of research results in international and interdisciplinary contexts, enables added value such as comprehensive indexing and also allows the innovative design of traditional formats. However, adopting open access publishing models for scholarly monographs is still slower than for scholarly essays (Eve, 2015). In order to simplify the integration of monographs in the Open Access universe, several infrastructures and services have been developed in recent years: OAPEN (Open Access Publishing in European Networks, 2010); OpenEdition Books (2012); Open Monograph Press (2012); Ubiquity Press (2012); The Directory of Open Access Books (DOAB, 2012); Knowledge Unlatched (2012); and the JSTOR Open Access Book Programme (2016). Nevertheless, the landscape of academic publishing has so far remained highly fragmented along various national, linguistic and subject-specific lines, particularly in SSH

[1] bertino@sub.uni-goettingen.de

(OPERAS Consortium, 2017). The current uncoordinated situation represents a major obstacle to the optimal dissemination of research results of the SSH disciplines and their impact on the structures of open science.

2 The H2020 HIRMEOS Project

The HIRMEOS project focuses on the monograph as a significant mode of scholarly communication in SSH and tackles the main obstacles to the full integration of five large-scale platforms supporting open access content. The main objective of HIRMEOS is to optimise five OA digital platforms for the publication of monographs from the SSH and to ensure their interoperability. An integrated publishing system would support scientific work by fostering basic research activities - the so-called scholarly primitives - i.e. writing, finding, annotating, referencing, assessing, exemplifying, presenting, as well as elementary activities in the digital field such as searching in browsers, connecting digital texts, collecting data, scanning and creating standards of data handling (data practices) (Palmer et al., 2009). HIRMEOS intends to transform collections of passive documents into corpora of enriched texts. More specifically, the participating platforms will be enhanced with services that enable identification, authentication and interoperability (via DOI, ORCID, FundRef), the annotation of monographs, the gathering of usage and alternative metric data, as well as using tools - like *entity-fishing* - that enrich the text with linked data.

3 Entity Resolution Service

With the digital information explosion, over the last few decades the extraction and resolution of entities has been studied extensively (Milne et al., 2007; Cucerzan, 2007) and has become a crucial task in large-scale text mining activities. Entity extraction and resolution is the task of determining the identity of entities mentioned in a text against a knowledge base representing the reality of the domain under consideration. This could be the recognition of generic Named Entities suitable in general purpose subjects, like person name, location, organisation name and so on, but also the resolution of specialist entities in different domains. *Entity-fishing* addresses these needs and provides a generic service for entity extraction and disambiguation (NERD) against Wikidata, supporting possible further adaptations for applications to specialist domains. This allows it to be independent of a particular framework and usage scenario for maximum reuse. *Entity-fishing* API allows the processing of different input (raw or partially annotated texts, PDF, search query), different languages and different formats. *Entity-fishing* employs supervised Machine Learning algorithms for both the recognition and the disambiguation tasks using training data generated from Wikipedia article structures (Milne and Witten, 2008).

4 *Entity-fishing* Integration: Applications, Use Cases and beyond

The integration of the service during the HIRMEOS project was supervised and measured by different levels of increasing complexity (from the access to the API to the creation of new services using the generated data). After having successfully completed the basic integration to the API, each partner could choose to use the annotations based on their own needs and practices. The results have been summarised as a set of use cases, providing an initial feedback on common needs among publishing platforms.

One of the most implemented use cases was the enhanced facet search based on entities extracted from the library content. Targeting Named Entity of Person and Location, the users were able to further restrict their search to some content-based information. This functionality required processing the entire collection and indexing all the generated annotations. A variant of this idea was the automatic generation of a word cloud at the repository level; in this way the users were able to access the most important concepts present in the monographs hosted

at the digital library. An interesting evolution which could improve the search quality would be the generation of the word cloud at the search or book level, the cloud being updated at each query as the user narrows down the number of results. The annotations could in this way help to achieve better and quicker understanding of the specific and disciplinary language of certain monographs and so encourage non-specialists to use them. In this way, exploiting NERD annotations would foster interdisciplinary research.

Other partners worked on a different approach, more focused on enhancing the visualisation of the monograph, by supporting annotation generated by entity-fishing to the monographs' landing page in order to automatically annotate content and seamlessly visualise it to the users. This visualisation could be further enhanced by having a slider at the side of the page that allows users to reduce or increase the detail of the annotations based on occurrence or type of entity. Finally, another interesting aspect pursued was the possibility to group books by their content extracted entities. Linking related books is one of the keys to boosting dissemination. Classic recommendation systems attempt to suggest additional articles or products the user might be interested in by exploiting the user's purchase history or the navigation. This could have a big impact on the dissemination in open access monograph catalogues. A simpler approach (which would not require collecting any user data) would be to process the generated annotations with clustering techniques. This would enable more cluster configurations, exploiting different aspects (domains, similarity, frequency, etc) of the collection annotations.

There are indeed many more ideas that could be implemented to improve the user experience in terms of navigation or search. The search box could be improved by adding an additional layer dealing with disambiguation in the query in order to expand it to match the correct concept among several ambiguous possibilities. *Entity-fishing* could be also integrated in the process of feature generation as recommended by the CORE (https://core.ac.uk/), extracting concepts to be used as keywords for linking research outputs together. Another challenging idea is the possibility to extract all events and temporal expressions in order to build a timeline visualisation graph at the collection or book level. These use cases are applicable not only to other publishing platforms in SSH but potentially to any open access repository.

References

1. Lopez P, Meyer A, Romary L.: CENDARI Virtual Research Environment & Named Entity Recognition techniques. Grenzen überschreiten – Digitale Geisteswissenschaft heute und morgen, 2014.
2. Eve MP: Open access publishing and scholarly communications in non-scientific disciplines. Online Information Review, 39(5) 2015, 717–732.
3. OPERAS Consortium. Operas design study, October 2017.
4. Palmer CL, Teffeau LC , Pirmann CM: Scholarly information practices in the online environment: Themes from the literature and implications for library service development, 2009
5. Milne DN, Witten IH, and Nichols DN: Extracting corpus specific knowledge bases from wikipedia, 2007.
6. Cucerzan S: Large-scale named entity disambiguation based on wikipedia data. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007.
7. Milne D, Witten IH: Learning to link with wikipedia. In Proceedings of the 17th ACM conference on Information and knowledge management, 2008, 509–518.