Large-scale Semantic Annotation for Improved Content Discovery in a Digital Library: A Case Study on Érudit

Fabrizio Gotti, Philippe Langlais, Vincent Letard

RALI – Université de Montréal

fabrizio.gotti@umontreal.ca, philippe.langlais@umontreal.ca, letardvi@iro.umontreal.ca

Scientific report – August 2022

Abstract

In this work, we describe through a case study how two natural language processing techniques, end-to-end entity linking and Open Information Extraction (OIE), can be combined to perform semantic annotation of scholarly documents, with a view to improving the discoverability of their content. Our case study is carried out via the implementation of this hybrid approach within a fully realized prototype, Allium, whose primary goal is to facilitate content discovery and navigation by a human user within the Érudit digital library. Érudit comprises over 150 scholarly journals and 38 cultural publications in social sciences and humanities, from Québec and Canada. We start by showing that entity linking allows for a solid foothold in Linked Open Data (LOD), even if it lacks in recall, especially regarding relations between entities. We then show the potential of OIE for content discovery, in part because it can complement the information gleaned from LOD. We propose and implement methods of integrating these elements to a full-fledged publication platform. Finally, we perform a system-oriented assessment and a user-oriented (human) evaluation confirming that LOD and OIE annotations are compatible and complementary for content discovery.

Keywords: digital libraries, semantic web, natural language processing, semantic annotation, open information extraction, exploratory search

1 Introduction

1.1 Exploratory search: at the heart of scholarly work

Scholarly work involves intellectual activities ranging from the mundane lookup of facts to the convoluted collaboration effort between multiple researchers. Studies have proposed typologies for these activities. James Unsworth (2000) describes scholarly primitives as "basic functions common to scholarly activity across disciplines, over time, and independent of theoretical orientation". They include discovering, annotating, and comparing. Palmer, Teffeau, and Pirmann (2009) draw on this and propose five *core activities*: searching, collecting, reading, writing, and collaborating. Anderson, Blanke, and Dunn (2010) explore the nature of scholarly work in the digital age, and offer their typology, which is then applied (Blanke & Hedges, 2013) to devising new infrastructure elements to help researchers in their work. Among these *methodological commons*, we contend that exploration and comparison occupy central roles, because they can promise, among other desirable features, *exploratory search*.

Exploratory search (Marchionini, 2006) constitutes an iterative and interactive process by which a user gradually uncovers networks of related concepts and online communities that, in turn, participate in knowledge acquisition, assessment and synthesis. This fertile "adventure in a new world of information riches", as Marchionini puts it, can be facilitated by dedicated user interfaces.

However, in part because of the ever-increasing volumes of scholarly texts available, content exploration and comparison are challenging when pursued without some form of assistance. Therefore, researchers need an environment supporting the integration of heterogeneous data sources (local or dispersed), that is scalable, on-demand, and equipped with discovery tools (Blanke & Hedges, 2013).

Computer systems have long been used to provide such discovery tools. Typically, search engines play a vital role (Hyvönen, 2012). While successful, these techniques often rely on a shallow "reading" of textual character strings, without leveraging their semantic content (Cornolti, Ferragina, & Ciaramita, 2013; Gagnon, 2013). This impedes discovery and comparison. Firstly, string-matching techniques can miss or confound mentions of entities (e.g. the phrase `the Bard` refers to William Shakespeare or, perplexingly for computers, to Robert Burns, depending on the context). Secondly, shallow approaches struggle at finding *commonality* between mentions (be they in the same text or distributed across databases), because these tools are not designed to interpret unstructured text to find well-defined, canonicalized entities, and even less so to bridge the gap between distinct knowledge sources. This latter weakness undermines the comparison primitive. Thirdly, traditional tools handling surface forms of text are not tailored to reveal *relations* between entities, which link entities together in a given document repository or point to other entities in external repositories.

More advanced Natural Language Processing (NLP) techniques address some of these difficulties, and allow mining unstructured texts for entities and their relations. Notably, NLP offers *end-to-end entity linking*, the process of spotting and disambiguating entity mentions in text to their counterpart in a given knowledge repository. This repository is often part of the Semantic Web (Berners-Lee, Hendler, & Lassila, 2001), and contains candidate entities as well as the relations they entertain

between them. *Open Information Extraction* (OIE) is a more recent research area that seeks to acquire shallow semantic representation elements directly from unstructured (free-form) texts (Del Corro & Gemulla, 2013; Yates et al., 2007). Typically, this extracted knowledge takes the form of relational triples like (`Einstein, was born in, Germany`) composed of two arguments flanking a relation. Traditional OIE does not attempt to link these arguments to external repositories.

1.2 Combining Entity Linking and Open Information Extraction to Enhance Content Discovery and Exploration

Together, entity linking and OIE offer fertile *semantic annotations* for a corpus, or *semantic enhancements* (Shotton, Portwin, Klyne, & Miles, 2009) of a scholarly document. The extracted and properly structured knowledge (concepts and their interconnections) can be cross-referenced, compared, and aggregated, in order to improve information discovery and to help a corpus reach its full documentary potential. Shotton (2009) describes the potential of such approaches in semantic publishing: they enhance the meaning of an article, facilitate its automated discovery and link it to relevant articles.

Contrarily to other studies focused on a given field, like (Shotton et al., 2009) and (Seringhaus & Gerstein, 2007) for biomedical texts, our methods strive to find semantic enhancements for collections in any scientific field. Our work is akin to (Marchand, Gagnon, & Zouaq, 2020), who also used OIE tools to semantically enrich French cultural documents, although their study leverages a corpus ten times smaller, and studies relations and entities in real estate.

In this paper, we describe a case study of a large-scale semantic annotation of a digital library, with the intention of facilitating the exploration, discovery and comparison of content. To the best of our knowledge, this is the first large-scale case study on the relevance and usability of computer-generated semantic annotations for the exploration of a large digital library. In Section 2, we present the Érudit digital library, whose documents we annotated by combining the fields of entity linking (Section 3) and Open Information Extraction (Section 4). Section 5 presents the prototype Allium, which grafts itself onto Érudit to reap the fruits of semantic annotation. We provide both a system-oriented and user-oriented (human) evaluations of Allium in Section 6, then conclude in Section 7.

2 The Érudit Digital Library

Érudit's mission is to promote and disseminate research and creation results. Their digital library comprises 150 scholarly journals and 38 cultural journals in the humanities, social sciences, and arts and letters from Québec and Canada, as well as a few articles in natural science. Érudit's web platform, at www.erudit.org, combines a digitized collection and an e-journal. It offers open access and closed access content. For our research, we used a snapshot of the complete collection as of March 2017, which counts 175,000 texts.

Document types can be scholarly articles (51%), short reviews (literary criticism, essays on works of art, etc.; 31%), notes (1%), or other publications (e.g. biographical notes, recent publications, etc.; 16%). The collection contains 91% French texts, 7% English publications, and 2% of documents in

other languages. We have focused exclusively on the French part of the collection (159k documents). On average, a document in the collection counts 3000 words, or 169 sentences.

3 Semantic Web annotations: LOD entities and relations found in Érudit

Since we wish to find exploratory links between documents and knowledge bases, we need to find *anchors* for these links, i.e. we would like to identify elements in a text that are worthy of being either origins or destinations for such links. Among anchor candidates, *entities* in the text (e.g. `Barack Obama` or `linguistic norm`) are interesting, because they constitute topics the reader has before their eyes and because they are by definition (Hoffart et al., 2011) registered in *external* knowledge bases.

One such external resource is the Linked Data, part of the technologies of the Semantic Web. First outlined by Tim Berners-Lee (Berners-Lee, 2006), the Linked Data initiative provides principles to encode and disseminate information about concepts and their relations. A data element within the Linked Data often takes the form of a semantic triple subject-relationship-object, e.g. (`Barack Obama`, `was born in`, `1961`). Here, we focus on Linked Open Data (LOD), i.e. Linked Data triples released under an open license. The LOD holds more than a thousand datasets, with tens of billions of triples, interlinked by hundreds of billions of relations. These entities and relations are well suited to guide the identification of entities and relations, and to lay the groundwork for a rich and principled way to offer discovery anchors and links in a corpus.

3.1 Entities

Often, the first step towards interconnecting a corpus to the Linked Open Data is to perform *end-to-end entity linking* (Demartini, Difallah, & Cudré-Mauroux, 2012). This two-step process first finds *mentions*, i.e. spans of text containing entities, and then disambiguates these to the correct entity in a given knowledge base. We picked the LOD dataset *DBpedia* for this. DBpedia is an initiative that aims to automatically extract structured content from the information created in Wikipedia. DBpedia has emerged as one of the central interlinking hubs in the LOD, both by its size and breadth of topics (Bizer et al., 2009). It offers immediate exploration targets for the reader, since numerous entities in DBpedia are described by the Wikipedia articles they were extracted from. DBpedia is also valuable because many end-to-end entity linking systems and algorithms rely on its entities.

We performed end-to-end entity linking offline, using DBpedia Spotlight (Daiber, Jakob, Hokamp, & Mendes, 2013; Mendes, Jakob, García-Silva, & Bizer, 2011) with the freely available French extraction models. Figure 1 shows a sample of DBpedia Spotlight's output.[1] DBpedia Spotlight identifies DBpedia entities within text, e.g. the entity URI http://fr.dbpedia.org/resource/Canada for `Canada` or http://fr.dbpedia.org/page/Disque_compact for `compact disc`. DBpedia Spotlight is, to our

[1] We will be illustrating this work with examples taken from the Érudit article available here: https://id.erudit.org/iderudit/1009894ar. We have translated into English most of the examples cited. The reader should bear in mind that all cited text elements were originally in French.

knowledge, the only mature, end-to-end entity linking software freely available that can handle French text. As we will show shortly, its performance is also satisfactory for our needs.

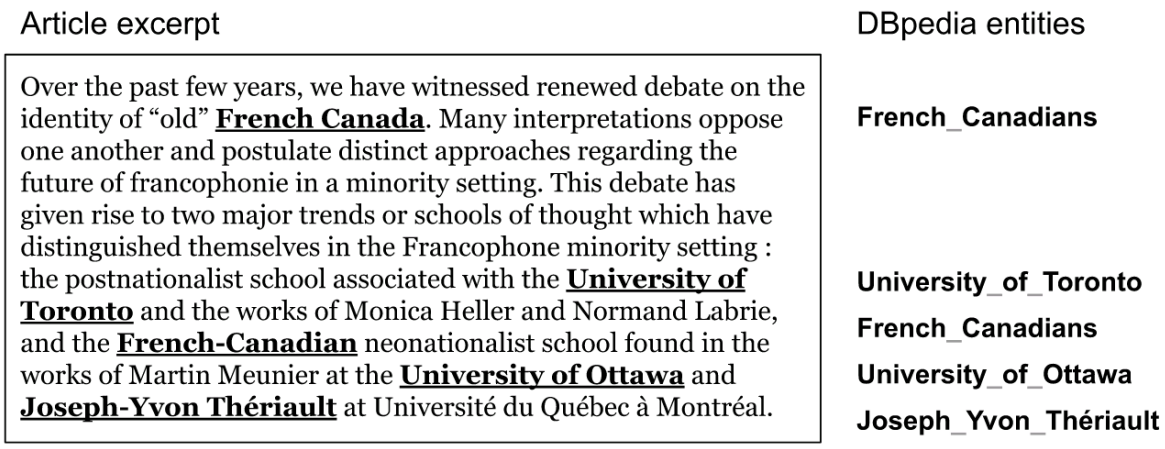

Figure 1

Sample output of DBpedia Spotlight on an excerpt of the running example article. Entity mentions are underlined and in bold in the text (left). On the right, we show the corresponding DBpedia entity. Note that different mentions (`French Canada` and `French-Canadian`) can disambiguate to the same entity.

When using DBpedia Spotlight, we use a (empirically determined) confidence threshold of 0.4. This confidence parameter ranges from 0 (high recall, low precision) to 1 (low recall, high precision). It considers factors such as topical pertinence and context ambiguity.

On average, a document contains 114 mentions, one every 24.9 words. Longer documents are naturally more likely to contain more entities. Many of these mentions refer to the same entity within a document, since its topics are recurrent, e.g. the mention `multiculturalism` may occur 10 times and `multicultural` 5 times within an article, but they all refer to the same entity. We observe 53 distinct entities on average per document, with a Zipfian distribution, i.e. a handful of entities that occur very often, along with many entity hapaxes. Over the whole collection, the entities that are present in the most documents are—unsurprisingly given the corpus—`Québec` (52.6% of documents), `Montréal` (37.5%), and `Canada` (35.1%). The first non-proper-named entity is `Philosophy`, at rank 10, occurring in 12.9% of documents.

We manually evaluated a random sample of 500 linked entities for precision and recall. Precision measures whether the entity link predicted is correct for the mention identified. We observed an 82.6% precision figure, which is quite high for this task. Most errors occur when under-specified mentions create an ambiguity when linking, e.g. the mention `Saint-Lambert` can refer to more than 30 places or persons. Interestingly, 43.8% of entities are not proper-named entities but include `wine`, `philosophy`, and `artificial intelligence`. Precision is similar for proper-named entities (84.1%) and non-proper-named ones (80.8%).

Recall indicates the percentage of entity occurrences actually present in the text that DBpedia Spotlight correctly spots. Recall is much more difficult to assess due to its subjective nature. For every verb, common noun or proper noun, there is very often a corresponding DBpedia entity. This means that an overzealous system would analyze the first sentence in Figure 1 and link the words `year`, `witnessed`, `debate`, and `identity` to corresponding entities. We therefore had to determine if a missed entity would have contributed to the goals of content discovery, a delicate task. We opted to reject ubiquitous common nouns (`year`, `witness`, `debate`), but to keep concepts central to the semantics of the text (e.g. `identity`). We manually annotated 50 randomly selected sentences and observed that about half of the potential entities in the text were missed by the system, largely because they refer to concepts not covered by DBpedia. For instance, researcher `Martin Meunier` or the scholarly "`Heller-Labrie theory`" are not covered in (French) Wikipedia, but are semantically important. Other missed entities belong to the aforementioned key common nouns.

3.2 Relations

Once LOD entities are extracted, we can use them to query billions of facts scattered in interconnected datasets, that link entities via predetermined relations, e.g. `x hasBeenAwarded y`. For a given entity, numerous facts are typically available. These queries are usually submitted in the dedicated language SPARQL against a server on the Web called an *endpoint*. The French DBpedia's SPARQL endpoint responds in a few milliseconds.

On average, when querying French DBpedia with an entity, we find it is involved in 43 distinct relations, for a total of 352 instances of those relations (the same relation can be employed multiple times). The most frequent relations are http://www.w3.org/2002/07/owl#sameAs (7.1% of all relations), indicating an equivalence relationship between two resources (for instance equivalent entities in different datasets); and http://dbpedia.org/ontology/country (6.6%), which indicates the country an entity is located in. On average, a given entity is linked to 252 other distinct entities within French DBpedia alone, via its relations. These other 252 entities do not necessarily appear on Érudit.

Generally speaking, the relations found in the major LOD dataset repositories are accurate, because they are derived from human-curated databases. Once again, recall is very difficult to assess, as the exhaustive list of relations a given entity entertains in real life is impossible to produce realistically. It is likely very low when compared against such a completely theoretical ground truth.

This does not mean that the information gleaned from DBpedia is scant: once we gain a foothold within the LOD, the volume of available information is overwhelming for humans, and one must choose what to use. In our case, useful discovery paths are those that (1) shed light on the text being read, and (2) provide useful ways of discovering and exploring relevant content. Furthermore, both there should be a preference for content in the source language, French here. French DBpedia is therefore a reasonable choice. Moreover, DBpedia entities are tightly linked to Wikipedia articles. The encyclopedia provides a way to query its API with a DBpedia entity, yielding definitions, explanatory snippets, and images useful to the reader. We added yet another LOD dataset to DBpedia, the repository of the Bibliothèque Nationale de France (BNF). Table 1 illustrates the relations found in these two LOD repositories.

Source	Triples
French DBpedia	<dbpedia-fr:Joseph_Yvon_Thériault, dbpedia-owl:birthPlace, dbpedia-fr:Caraquet> *Joseph Yvon Thériault was born in Caraquet.* <dbpedia-fr:Joseph_Yvon_Thériault, dbpedia-owl:education, dbpedia-fr:Université_d'Ottawa> *Joseph Yvon Thériault graduated from the University of Ottawa.* <dbpedia-fr:Université_de_Toronto, dbpedia-owl:motto, "Velut arbor aevo"> *University of Toronto's motto is "Verlut arbor aevo".*
BNF	<Joseph_Yvon_Thériault, marcrel:aut, cb34916973t> *Joseph Yvon Thériault authored "La Société civile ou la Chimère insaisissable : essai de sociologie politique".* <university_of_toronto, bnf-onto:firstYear, 1827> *The University of Toronto was founded in 1827.* <university_of_toronto, marcrel:edt, cb33223380k> *The University of Toronto is editor of "Taxation of the forest industries in Ontario".*

Table 1
Examples of relations found in French DBpedia and BNF open datasets.

For 15.4% of DBpedia entities extracted in Érudit, DBpedia offers a counterpart in the BNF dataset, through a dedicated equivalence relation. Most authors and locations are well covered by the BNF, less so for common concepts. Additional information is available through the BNF catalog's SPARQL endpoint.[2] The BNF pages are rich with content, chiefly complete bibliographies for authors, sometimes with links to the relevant documents and catalog entries.

4 Open information extraction on the collection

We have seen that the relations available from the LOD (and DBpedia in particular) are determined a priori: knowledge repositories usually define relation templates (`x hasBeenAwarded y`) before they are populated. This is often carried out by international standardization bodies or by field-specific experts (Hyvönen, 2012). While potentially numerous, the relations that are part of this schema constitute a closed list that typically excludes ad hoc, expressive inter-argument links like "differs from" or "is content to unmask", that are typically found in unstructured text.

To capture these less traditional semantic elements, we propose to use Open Information Extraction (OIE). Our goal remains the same: create a way to explore how different entities and concepts interact within a single document, and to leverage these elements in the corpus as a whole.

4.1 Filtering OIE extraction output

OIE systems are numerous for English, and rely on various strategies to extract structured tuples, typically following the tripartite schema (*argument 1*, *relation*, *argument 2*), where *relation* is a relation phrase linking the two arguments (we also call them *concepts*) to one another. These three

[2] See https://data.bnf.fr/current/sparql.html, yielding https://data.bnf.fr/ark:/12148/cb12000599q for Joseph Yvon Thériault for instance.

elements are most often excised verbatim from the original text and rendered into a triple, e.g. (`Thériault`, `recognizes`, `French Canada's modernity`).

The OpenIE project[3] is a very mature OIE tool relying on an assemblage of different studies (Christensen, Soderland, & Etzioni, 2011; Pal, 2016; Saha, 2018; Saha, Pal, & Mausam, 2017). Most OIE extractors are language-dependent, and process English only. For French, we relied on an adaptation of the extractor ReVerb (Fader, Soderland, & Etzioni, 2011) to French (Gotti & Langlais, 2016). We call *FReVerb* this adaptation henceforth.

We ran FReVerb on the entire French Érudit corpus, yielding 28.0M triples. The extraction process is extremely noisy (Léchelle, 2019): It produces a lot of nonsensical or nebulous triples (see Table 2 for a sample of FReVerb's output). We kept only the triples whose arguments were both noun phrases, and we relied on a score *S* to rank triples according to their perceived quality. To compute *S* for a triple (arg_1, rel, arg_2), we calculate (for each argument *arg*) the count of *arg* in the document, multiplied by another factor measuring the degree of specificity of *arg* to the current document. We add the results for both arguments to get *S*. This score, based on tf-idf (Manning, Raghavan, & Schütze, 2018), conveys the quality of OIE extractions, because erroneous extractions usually have lower frequencies, and therefore lower scores.

Before filtering	**After filtering and ranking (top triples)**
(it, **continues to have**, a meaning)	(Thériault, **observes in**, French-speaking minority)
(the goal of this text, **is**, to present the debate)	(Thériault, **recognizes**, French Canada's modernity)
(the authors, **see**, two tendencies)	(Thériault, **differs from**, the Heller-Labrie theory)
(two schools, **clash**, in a minority setting)	(Thériault, **defines**, the vital intention)
(the first tendency, **associates**, French Canada)	(Thériault, **identifies**, two schools of thought)
(we, **also find**, this tendency)	(Labrie, **is not about**, explaining difficulties) †
(Thériault, **organizes**, the researchers' work)	(one intention, **is**, French Canada's)
… 352 in total, average score S *= 61.0*	*… 40 in total, average score* S *= 174.9*

Table 2

OIE results on the running example article, before (left column) and after filtering and ranking the OIE triples in decreasing order of score (right column). The triple marked with † was erroneously extracted from "`Their controversy with Heller and Labrie is not about explaining difficulties suffered by French Canadians…`".

To further refine the quality of the extractions yielded, we retain only the 8 best-scoring arguments per document, and keep only the 5 best triples involving each top-scoring argument. Therefore, per document, we obtain *at most* 8 × 5 = 40 triples, and 40 × 2 = 80 arguments. These thresholds were obtained empirically in an attempt to clean up the OIE output, and remain subjective.

[3] https://github.com/dair-iitd/OpenIE-standalone

Filtering yields 3.0 million triples (10.6% of the original count), a reasonable yield for a ReVerb-based tool. ReVerb's authors report a 0.23% yield when processing a large corpus from the Web and filtering the triples for quality.[4] For our corpus, 3.0 million triples correspond to 18.8 triples per document. On average, we found 22.8 distinct arguments per document.

The most frequent relation phrases are the plain `be`, `have`, and `do`, but there are also more meaningful and frequent relations, like `constitute`, `become`, and `present`. The most frequently occurring arguments are `Québec`, `women`, `the reader`, `Montréal`, and `the world`. It is noteworthy that some of these frequent arguments are identical to frequently occurring entities extracted by DBpedia Spotlight, described in the previous section.

4.2 Complementarity of OIE and entity linking

To get a sense of the complementarity between the LOD and OIE results, we examined the data available from each source for the entity "Joseph Yvon Thériault", the Canadian sociologist. DBpedia holds 103 facts about the scientist, 93% of which are accurate (for instance, his date of birth). A subset is illustrated in Table 1. In parallel, we found 59 OIE triples mentioning Thériault either as the first or second argument over the whole Érudit collection. About 83% of them were accurate and informative. We show a subset of these in Table 2 (right-hand column). Interestingly, there is almost no overlap between the two data sources for this example: a single fact is mentioned by both the LOD and OIE results, namely the fact that Thériault was born in New Brunswick, Canada.

We also conducted a corpus-wide analysis of the complementarity of all LOD and OIE entities extracted, in an automated manner. For each document, we measured the overlap between DBpedia and FReVerb entities (after filtering as explained in Section 4.1). Two entities are considered the same when they have the same surface or are extracted from the same text excerpt. This is a necessary approximation, given the volume of data. Nevertheless, on average, we observe that there are only 1.8 entities that are common to DBpedia and FReVerb per document. Thus, since we saw above that FReVerb finds 22.8 concepts per document, then FReVerb has the potential to add 22.8 – 1.8 = 21.0 unseen concepts for any given document. For our running example document, FReVerb finds (among others) the concepts "Canada", "Joseph Yvon Thériault", "Normand Labrie", "Heller-Labrie theory", "intention", "French-speaking world", "the birth of a new identity", and "French Canadian modernity". Only the first two perfectly match DBpedia Spotlight entities.

In the whole Érudit collection, we find 284k *distinct* DBpedia entities, and 1.98M *distinct* concept labels, out of which 1.69M (85.3%) are not found by DBpedia Spotlight. We sorted these 1.69M non-overlapping concepts in decreasing order of frequency to examine them. The most frequent are "the author", "students", and "movie". These concepts are present in French DBpedia (e.g. http://fr.dbpedia.org/resource/Auteur for author), but they are very rarely *extracted* by DBpedia Spotlight. It therefore seems that these common, yet non-overlapping entities are more indicative of the idiosyncrasies of the LOD extraction tool we used than they are of actual gaps in DBpedia itself. We must go much further down our reverse-sorted list of non-overlapping concepts to find truly

[4] See http://reverb.cs.washington.edu/README_data.txt.

original entities (i.e. missing entirely from French DBpedia), for instance FReVerb's "learner" (as in "English learner") at rank 411, or "single-parent family" at rank 1037.

A large-scale comparison between relations found by LOD and OIE tools is more difficult. Finding overlaps between relations like DBpedia's `isBirthPlaceOf` and FReVerb's relation phrase "`was born in`" is complex to produce automatically. We resorted to a manual evaluation of a random sample of 100 high-quality FReVerb triples from our corpus. To the best of our knowledge, we found a modest 6% perfect overlap with DBpedia, e.g. FReVerb's "be abrogated in" is equivalent in meaning to DBpedia's http://fr.dbpedia.org/property/abrogation.

Overall, it appears that entities and relations yielded by OIE bring a significant novelty compared to those identified within the LOD, and that the two avenues can be complementary when it comes to linking entities with relations, and to linking together the documents that mention these entities. This is important because, if we are to provide ways of exploring a large corpus, the more tightly linked are its documents, the better the chance of finding relevant, exploratory paths between them.

5 Putting it all together

In this section, we propose a way to make the *reader* benefit from the semantic annotations produced earlier. We have annotations that link passages in Érudit documents to French DBpedia entities, and that identify OIE triples. On average, a document contains 18.8 OIE triples and 114 DBpedia entities. The latter allow us to find in the LOD the following additional elements for an entity:

- A text snippet and picture explaining the concept, through Wikipedia's API.
- A link to the relevant page on the Bibliothèque Nationale de France's (BNF) database, when available.
- A latitude and longitude for geographical concepts, allowing positioning on a map.
- For historical events, start dates and end dates.

To propose these elements to the reader, we built a prototype called Allium that offers OIE and entity linking results to the user as additional features (akin to a "plugin") within the existing Érudit web pages. These new features call for a trade-off between the desire to leverage numerous semantic annotations and the need for a workable interface. As shown in Figure 2, Allium attempts to bring together these annotations in order to make an Érudit document part of a network of discoverable resources. A video of Allium in action for the running example is available online.[5]

[5] https://youtu.be/UnEWdTBIkUI

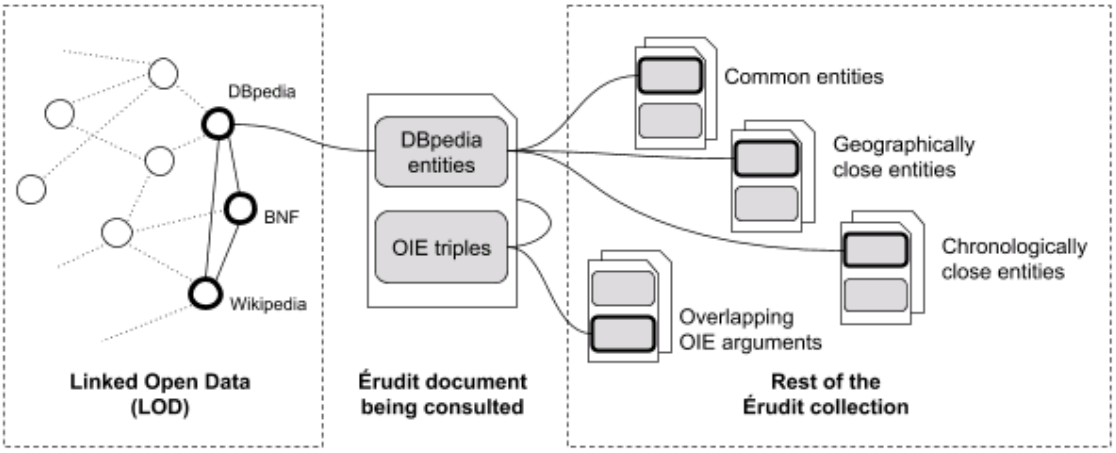

Figure 2
Allium gathers the annotations from DBpedia entities and OIE triples for the document consulted by the reader, and proposes links to relevant elements, in the LOD or within the Érudit collection. Related documents share some form of similarity with the document consulted. Chronological relatedness is proposed here, but Allium does not implement it. OIE triples can also help to explore the consulted document's content itself.

5.1 Exploring LOD entities

To present entity linking annotations, Allium anchors each entity in the text excerpt where it was detected, and provides the reader with a link to relevant elements, the first link in the exploration chain: it can lead to information about the entity, gleaned from the LOD, but also to other documents within and outside Érudit itself.

We have used *call-outs* to render this information. Allium's call-outs are panels containing data (mostly textual) related to the document's content. In their review of the topic, Wolfe and Neuwirth (2001) point out that call-outs are useful to highlight topics and important passages, and to supplement the text in a dynamic setting, which is precisely the case for Allium.

Pop-up call-outs The user can click on an unobtrusive link in the text to reveal information about the underlying DBpedia entity in a pop-up call-out (Figure 3).

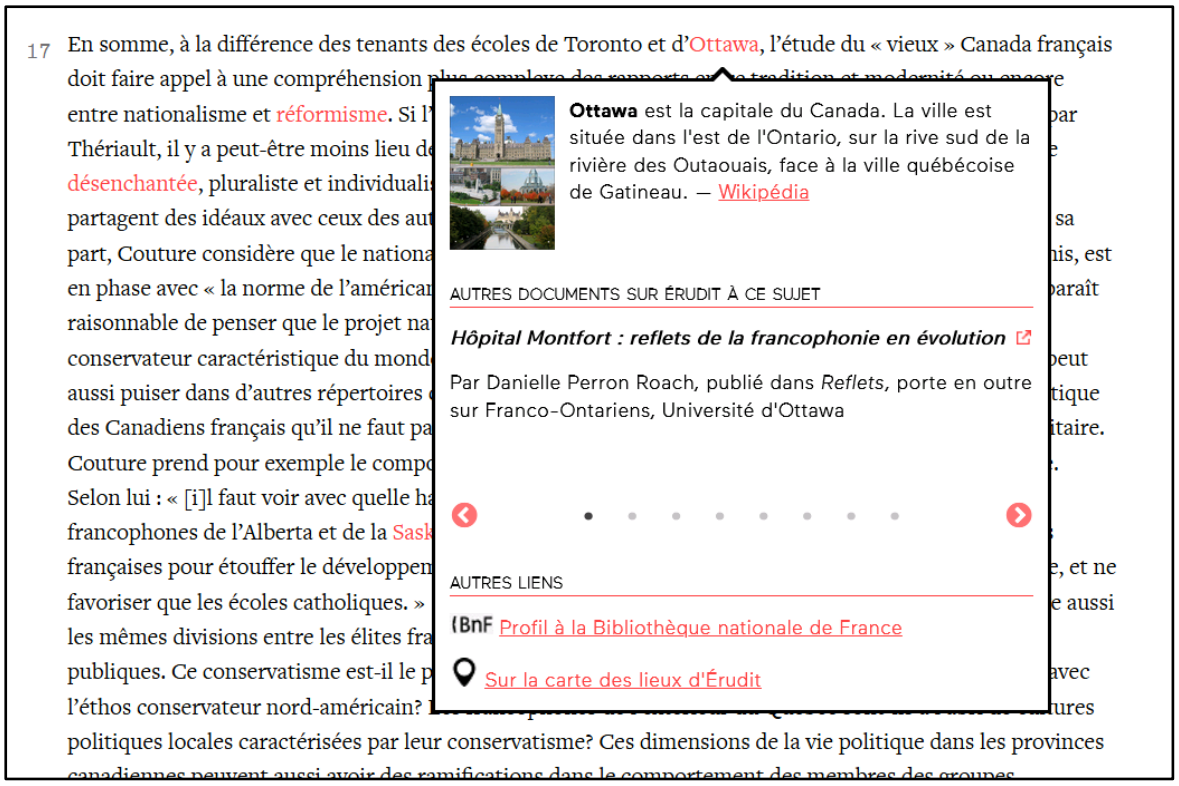

Figure 3

A pop-up call-out showing information relevant to a DBpedia entity identified in the text. The top part is a snippet from Wikipedia; the middle section proposes related documents within the Érudit collection in a "carousel" the user can flip through. At the bottom, our Allium prototype shows links to external resources, here the BNF and a geographical map of concepts (shown in Figure 5).

Picking the right entities to add call-outs to is essential to avoid overwhelming the reader, since we observed that a potential link can be made every 25 words on average, which is excessive. Therefore, we filtered out DBpedia entities whose corresponding Wikipedia pages have more than 3000 in-links, an empirical threshold. This muted trivial entities (e.g. Canada, Water). Moreover, we never added the same entity link more than once within a given paragraph. However, reasoning that a user can start their reading at any place in the document, we did repeat identical links when they did not appear in the same paragraph. We reached an average density of 1 entity call-out every 60 words, a 58% decrease.

Margin call-outs Margin call-outs are horizontally aligned with the entity which they explain (see Figure 4). They offer a form of gloss for a subset of salient entities in a text. This option spares the user from having to click on entity links to start exploring. The call-outs also lay out the topology of the document's contents (Duchastel & Chen, 1980), offering the reader a way to navigate to regions of interest. We use a weighing heuristic to determine salient concepts among those identified in the whole text. The weight W of a given entity e is a score from 0 to 100 empirically set to

$$W = 60 \times \text{freq}(e) / \text{freq\_max} + 40 \times \text{e\_intitle}$$

where freq(*e*) is the frequency of the entity *e* in the document, freq_max is the number of occurrences of the most frequent entity in the document and e_intitle is equal to 1 if the entity is found in the document title, 0 otherwise.

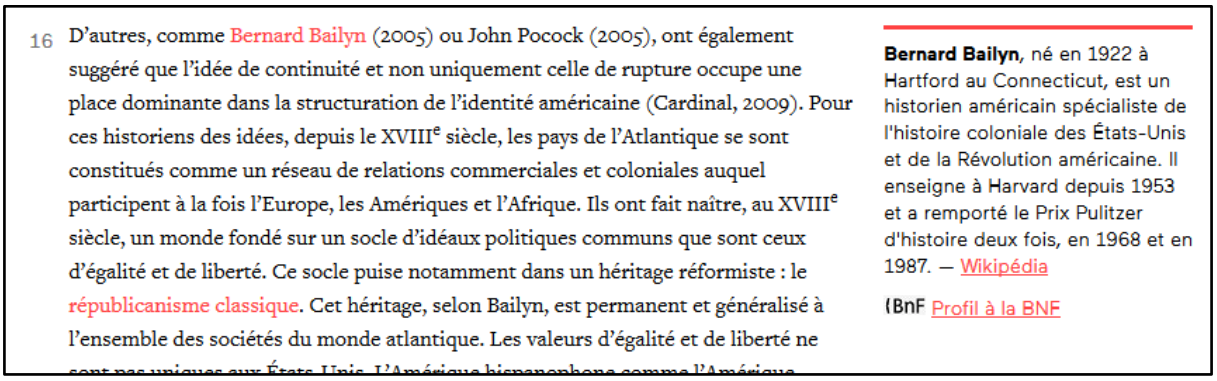

Figure 4

Margin call-out for the entity http://fr.dbpedia.org/resource/Bernard_Bailyn (Bernard Bailyn). A short gloss taken from Wikipedia is proposed, along with other links.

5.2 Using LOD entities for document discovery

We use the LOD entities in our semantic annotations to display a list of documents related to the one being consulted (see middle section of the call-out in Figure 3). We define a relatedness score rel_{doc} between documents based on the entities they share, a fast and sensible heuristic.

$$rel_{doc} = \sum_{e \in E} \begin{cases} 1 \text{ if } doc \text{ contains } e \\ 0 \quad \text{otherwise} \end{cases}$$

In this formula, E is the set of the 20 most frequent entities in the document being consulted by the reader, and *doc* is the related document to be scored. In Figure 3, the related article shown has three common entities: `Ottawa` (the entity the user clicked on), as well as `Franco-Ontarians` and the `University of Ottawa`. We only show the top 8 related documents. Each related document is accompanied with the labels of entities common with the current document, so that the user can see which entities they have in common. Other relevance scores could have been used.

Since most landmarks and geographical entities are associated with a longitude and latitude in DBpedia, Allium can build a "Map of Érudit", which users can navigate to find documents that mention the locations they want to explore (shown in Figure 5). A location is associated with 43.0 documents on average and there are 10.8 distinct geographical concepts per article on average. The most frequent locations are `Québec` (in 84k articles), `Montréal` (in 60k articles), `Canada` (56k articles), and `France` (52k articles). There are 18k hapaxes, e.g. `11 Downing Street` appears once.

Figure 5

Part of a geographical "Map of Érudit" allowing the exploration of neighboring documents. Whenever the user clicks on a position marker, the documents mentioning the geographical landmark or location are displayed in the right-hand list in alphabetical order. Clicking on a title leads to the corresponding article's page.

Similarly to geospatial maps, we attempted to position articles on a timeline, through the chronological entities spotted in their text. On average, an article has 2.6 distinct time entities, with the most popular being 19th century (12k articles). Nonetheless, since articles placed on a timeline tend to form unwieldy clusters of thousands of articles for a given time period, we have not yet managed to devise a workable interface for the end-user.

5.3 OIE for content exploration

Using Open Information Extraction (OIE) tuples in a workable user interface is understudied. We contend that they are best used to show (1) salient concepts covered by the triples' arguments and (2) how these concepts interact within a document. Proposing them in the document text itself is confusing for the reader, therefore we resorted to adding a dedicated pane (Figure 6) housing a document's triples. In the left-hand section, important concepts captured as arguments are listed. When the user selects a concept, the list of triples involving the concept are shown on the right-hand section. Each OIE triple is rendered in its original sentence, with the arguments and relation clearly identified. A discreet link leading to the original sentence in the text is also provided.

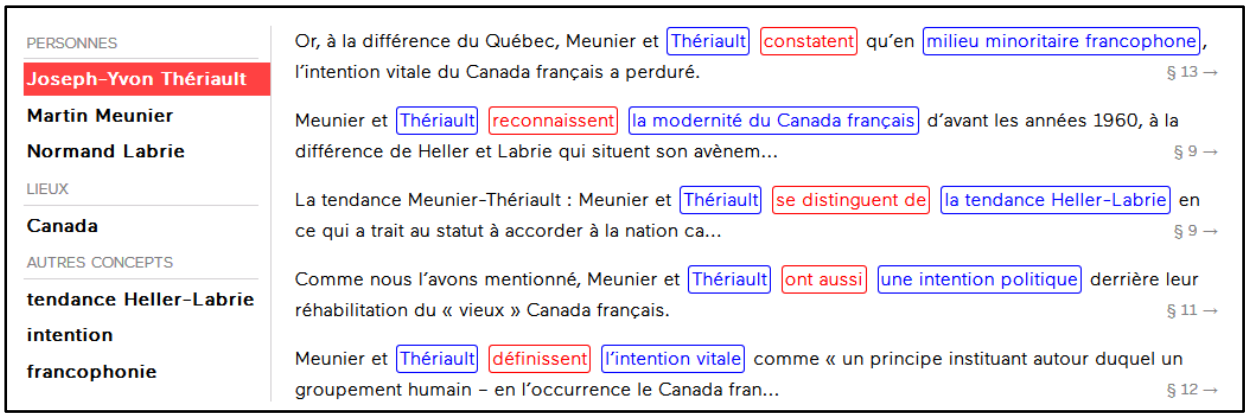

Figure 6
Open Information Extraction results panel shown to the user, at the very top of an Érudit document page. When a concept is selected on the left, its corresponding triples are shown on the right, in the context from which they were extracted. Arguments and relations are clearly identified.

In a way, triples act as a sort of distillation of a document's content, broken down into its fundamental elements. They become tools of single-document exploration. Strictly speaking, the text snippets displayed are not necessary to show these interconnections, and we could have proposed a graph of concepts divorced from its source text. However, the understandability of OIE triples usually deteriorates without context. Nevertheless, it could be interesting to provide such a graph.

6 Evaluations and results

6.1 The Érudit collection connection graph

Discoverability and exploration are abstract ideas. Nonetheless, Allium's performance is amenable to a formal system-oriented analysis. We can reframe the Érudit collection augmented with Allium as a mathematical *graph*, then measure the latter's density, that is, the number of connections between documents.

In this graph, each vertex represents a document. An edge is added between two vertices when the corresponding documents are linked. We can build two graphs. The **exhaustive** graph adds edges whenever two documents have a common OIE concept or DBpedia entity. The smaller **Allium** graph links document A to document B *only* when Allium proposes to the reader a link from document A to document B. This latter configuration is more realistic. We do not include links derived from the "Map of Érudit". For both exhaustive and Allium graphs, we can further measure the links contributed respectively by DBpedia entities and OIE concepts. The results are shown in Figure 7.

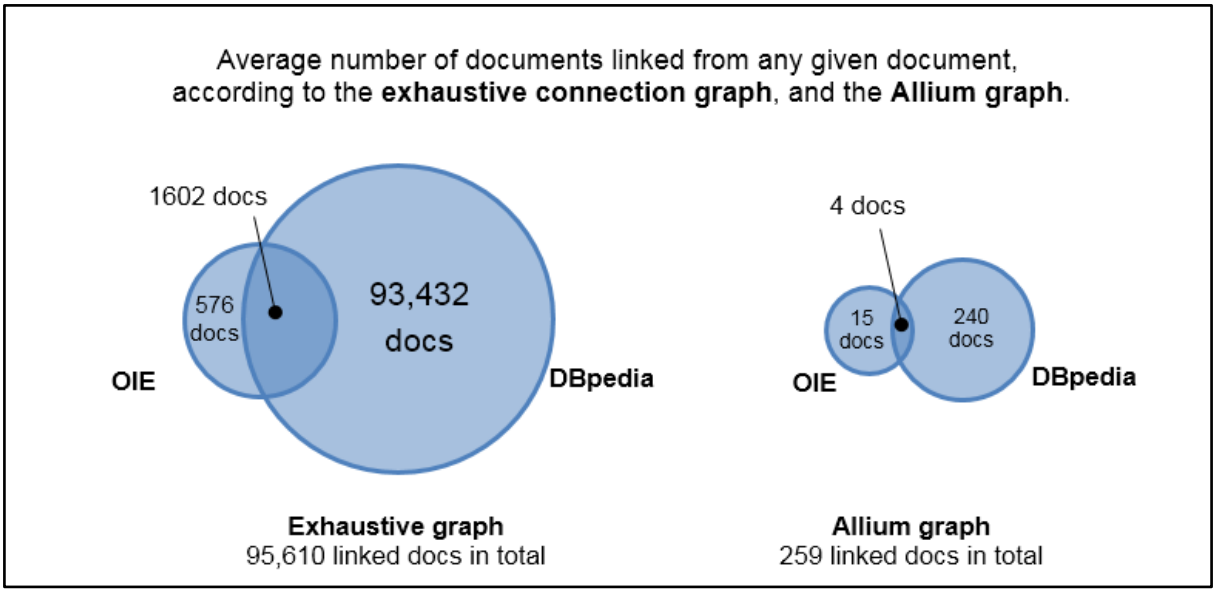

Figure 7

Average number of documents linked by semantic annotations within any given document, in the exhaustive connection graph (left) and in the graph made available by Allium (right) to a human reader.

On average, the exhaustive graph links a document to 95,610 other documents, far more than Allium, with 259 linked documents on average. This is expected, because Allium was designed to retain only a manageable number of links for a human. Additionally, OIE's links are far fewer than DBpedia's, due to the lower density of semantic annotations from OIE. In Figure 7, at the intersection of OIE's and DBpedia's circles lie the documents that are linked by *both* types of semantic annotations. In the exhaustive graph, this intersection is relatively large, but for Allium, this intersection is modest. This attests to the complementary nature of both types of annotations. The documents at this intersection are "doubly" interesting to the reader, and could be suggested more prominently.

We further observed that 0.4% of documents are isolated vertices in the exhaustive graph. These unreachable "island documents" are extremely short (e.g. https://id.erudit.org/iderudit/6026ac counts 6 words) or have no body text and consist rather in lists of references (for instance, a "Book received" section). These documents may require a special treatment to connect them to others, but this problem remains to be addressed.

6.2 Human evaluation

In this section, we propose a user-oriented evaluation. Digital library evaluation is quite complex, in part because evaluation protocols have not kept up with developments in the field (Zhang, 2010) and because of the multiplicity of evaluation criteria. In our case, the evaluation's goal is to assess whether subjects perceived that content discovery was facilitated by Allium.

6.2.1 Experimental design

We recruited 10 participants; four had PhDs, six had MScs. Five had backgrounds in the humanities and the other five were in the field of computer science. They are considered the typical end-users of digital libraries. Because of the COVID-19 pandemic, we explained the protocol by video conference, and then asked the participants to enter their results in a Google Sheets document in the cloud. We designed a two-part experiment, corresponding to two information discovery contexts.

In the first part, participants were asked to consult an article, and then to discover other articles within Érudit that pertained to two entities relevant to the article being read. For instance, while reading an article on French Canada, a user would be asked to find articles that pertain to both French Canadians and Québec. Participants were instructed to use Allium's pop-up call-outs (Figure 3). We measure the correctness of the task. On a five-point Likert scale, we ask the users to evaluate the panels' usefulness and ergonomics. To measure recall on the named entities, we ask users to select an arbitrary paragraph of the test article and to indicate which additional entities (if any) would have been helpful to them. Precision is also measured, by assessing the correctness of the entity linking process, for all the entities identified by Allium in the same paragraph. We also asked participants their opinions about margin call-outs, specifically if they help understand and explore the article.

In the second part, OIE concepts are evaluated. For a given article, we ask users to read all OIE concepts shown in the OIE results panel (Figure 6), and indicate whether they are relevant to the document, and if the triples proposed reveal useful, salient passages in the article's text. For each OIE concept, we then randomly select one article among those flagged by Allium as relevant to the concept, and ask the user if they find the article pertinent.

Finally, post-search, we asked users if they felt that the named entities and OIE concepts were complementary, and we collected the participants' comments about the process.

6.2.2 Results and interpretation

The precision of the article-locating task was 95%, indicating that almost all articles discovered by users do pertain to the entities requested. The error rate of 5% is due to a few users being unable to find an article. Likert-scale results are shown in Table 3. Named entity call-outs added by Allium are perceived as beneficial, as shown by the positive responses to questions 1 and 2. These results are encouraging, as they substantiate usefulness and usability.

Question	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
1. Do named entities pop-out call-outs facilitate exploration of relevant documents?	60%	30%	10%	0%	0%
2. Are named entities call-outs easy to use?	50%	50%	0%	0%	0%
3. Do margin call-outs help the exploration of relevant documents?	50%	20%	10%	20%	0%
4. Do margin call-outs help illuminate the article's layout and themes?	30%	40%	20%	10%	0%
5. Do margin call-outs help you understand the article?	40%	20%	30%	10%	0%

Table 3

Likert-scale evaluation of call-outs proposed by Allium (n = 10 participants).

Named entity precision and recall Named entity precision was a respectable 94.0%, i.e. 80 out of the 85 entities evaluated were found to be accurately identified and linked. Analysis reveals that most errors are due to ambiguous entities or improper mentions, which is consistent with Section 3.1. This figure of 94.0% is higher than that of 82.6% we observed in Section 2.2.1, possibly due to a smaller sample (85 versus 500 entities), and/or more leniency on the test subjects' part. In any case, it is good.

For recall, 69% of relevant entities had been identified by Allium, according to participants' analysis of the paragraph they were asked to inspect. On average, this means that readers would have wanted 45% more entities to guide their work. We carefully examined these missing entities, 37 in total. Persons accounted for 40.5% of these, concepts (e.g. `military regime`) for 37.8%, and the rest were various historical events, political parties, etc. About 89% of missing entities are actually present in DBpedia, so it is not the latter that is lacunary, but rather the entity linking process that is tricky. Indeed, most of the difficult mentions are derivations of the relevant entity: e.g. the mention `technicist` for the entity `technicity` or `Concertation for democracy` for `Coalition of Parties for Democracy`. Such mentions are bound to confuse even the best entity-linking systems.

Importantly, user-reported recall indicates that participants wished there were more links within the text being read. We limited the density of links to an entity call-out every 60 words on average (Section **Erreur ! Source du renvoi introuvable.**), but participants found acceptable a density of one entity every 40 words. Ultimately, we feel that participants tend to desire pop-out entity call-outs for all obscure entities they encounter when reading, which is reasonable, but unfortunately highly dependent on the reader.

Margin call-outs The potential of margin call-outs for discovery was received positively by only 70% of participants; 30% of them were neutral or negative in their perception (Table 3, question 3). This is somewhat regrettable, as we counted on them to lead the reader to external documents, e.g. Wikipedia or the BNF. Participants had the same rather tepid opinion about the ability of margin call-

outs to illuminate the article's layout and theme (Table 3, question 4) or its meaning (question 5). Participants opined that margin glosses have potential but they are not implemented satisfactorily. Some users wanted more of them, even though our current implementation leaves little room for this in the margin. Others indicated that some margin call-outs could have appeared earlier, or later. While it is undeniably desirable to provide call-outs for newly introduced concepts before going any further in a scholarly article, this would result in a very high concentration of such definitions at the very beginning of the article. This is difficult to implement because entities would overlap graphically. Such difficulties cast doubt on the practicality of these margin elements: there is too little space to accommodate so many desiderata. One non-trivial possibility would be to limit margin call-outs' scope, e.g. by providing only definitions, only biographical notices, etc.

OIE concepts were deemed relevant and salient to the article being read in 71.3% of cases (n = 80). The main problem identified by test subjects was saliency. Relevant but generic topics like `democracy` or `music` were deemed too broad and thus uninteresting. OIE concepts were considered useful for intra-document exploration in 67.5% of the cases. Here, the main problem is the target of the OIE link within the document, i.e., users would have preferred a more definitory target text, or a text where a clearer relationship exists between the OIE concept clicked and the main points made by the article consulted. A user suggested that the target text should always include a clickable concept mention with an associated pop-out panel. As for their potential for discovery, encouragingly, 81.7% of concepts were found to be fruitful jump points to other documents within Érudit. Further analysis suggests that, again, generic concepts rarely offer satisfying document targets. Fortunately, too-broad concepts can be filtered based on their count. It then remains to find the satisfactory frequency threshold.

OIE concepts and named entities The post-search questionnaire indicated that 70% of users found OIE concepts and named entities to be overlapping but complementary, 20% found them completely different in nature and necessary each in their own way, and only 10% found them uselessly redundant. In other words, 90% of participants found that traditional named entities and OIE concepts are complementary when attempting to discover content relevant to their research. This is compatible with the system-oriented evaluation proposed in the previous section, and encouraging.

7 Conclusion

In this work, we set out to enrich Érudit, a French digital library of scholarly documents, with semantic annotations, with a view to help the reader discover and explore relevant content. We used two distinct natural language processing strategies to carry this out: entity linking and open information extraction (OIE). We respectively employed DBpedia Spotlight and ReVerb for French, and then materialized these annotations within a prototype, Allium. It is noteworthy that we did not make any assumptions on the topics of the documents we annotated, and propose methods that could readily be applied to other collections. The only constraint is the language of the corpora: entity linking and OIE tools are typically available in English and a few other languages only.

This case study showed that producing semantic annotations is feasible, and it also demonstrates the difficulties of this initiative. A significant challenge is both recall and precision for the annotations produced. For entity linking, precision is very good, even if the tool we used had its idiosyncrasies.

Recall is lower, which can be problematic: Even if it is impossible to present *all* the entities detected to the user, a good recall means nevertheless that any given document will be tightly linked to other relevant counterparts within a collection. For OIE, precision is much lower, and great care must be taken to filter out spurious triples and concepts. Here, we propose simple strategies to achieve this, but more sophisticated classifiers could have been devised. It is also highly likely that different LOD and OIE tools will generate problems specific to them. Here, we have used tools for French text, but it remains to be seen how their English counterpart would behave in a similar experimental setting.

Manual examinations have shown that LOD and OIE annotations present some overlap, but not much, and that they do play complementary roles. This complementarity is not entirely surprising, as one of the initial goals set out for Open Information Extraction by its inventors was to handle previously unseen ways of expressing knowledge and cast it in structured tuples.

The Allium prototype attempts to leverage semantic annotations to facilitate navigation and content discovery. LOD entities lend themselves well to their integration in a web platform, but OIE triples decidedly less so. While we propose a way to show these triples to the reader, divorced from the origin text itself, we feel our solution does not fully do them justice. Particularly, it remains to be seen how their relations could be aggregated and used to offer a complete network of concepts and links at a larger scale than that of a single document. For instance, extracted relations like “endorse”, “give his support to”, etc. could serve to paint a complete picture of agreements between philosophers identified in Érudit.

A system-oriented evaluation revealed that DBpedia Spotlight and ReVerb cooperate well to link documents together within Allium. A user-oriented evaluation with 10 participants tended to corroborate this, and showed that readers felt both type of annotations helped them discover Érudit. The evaluation also highlighted (many) remaining challenges. The most recurring one is choosing *which annotations* are necessary when reading and discovering. Generally, test subjects did want more of them, which is non-trivial. For one, all users do not want the same entities, and not in the same order. Moreover, it is ergonomically garish to fill a limited space with too many call-outs, especially in the text’s margin. Entities that are too generic should be avoided. Nonetheless, we find encouraging this enthusiasm for more semantic annotations. Finally, the complementary natures of named entity recognition and open information extraction when discovering a large digital library is quite heartening.

Acknowledgments

This work is part of the CO.SHS project (co-shs.ca) and benefited from the financial support of the Canada Foundation for Innovation (Cyberinfrastructure Initiative – Challenge 1). We are grateful to our friends at Érudit for many insightful comments. Thanks to Simon van Bellen who reviewed an earlier version of this work. We also wish to thank all the contributors who made DBpedia possible.

References

1. Anderson, S., Blanke, T., & Dunn, S. (2010). Methodological commons: Arts and humanities e-Science fundamentals. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1925), 3779–3796.
2. Berners-Lee, T. (2006). *Linked data-design issues*. Retrieved from http://www.w3.org/DesignIssues/LinkedData.html
3. Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web: A new form of Web that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*.
4. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). DBpedia-A crystallization point for the Web of Data. *Journal of Web Semantics*, *7*(3), 154–165.
5. Blanke, T., & Hedges, M. (2013). Scholarly primitives: Building institutional infrastructure for humanities e-Science. *Future Generation Computer Systems*, *29*(2), 654–661.
6. Christensen, J., Soderland, S., & Etzioni, O. (2011). An analysis of open information extraction based on semantic role labeling. *Proceedings of the Sixth International Conference on Knowledge Capture*, 113–120.
7. Cornolti, M., Ferragina, P., & Ciaramita, M. (2013). A framework for benchmarking entity-annotation systems. *Proceedings of the 22nd International Conference on World Wide Web*, 249–260.
8. Daiber, J., Jakob, M., Hokamp, C., & Mendes, P. N. (2013). Improving efficiency and accuracy in multilingual entity extraction. *Proceedings of the 9th International Conference on Semantic Systems*, 121–124.
9. Del Corro, L., & Gemulla, R. (2013). ClausIE: Clause-based Open Information Extraction. *Proceedings of the 22nd International Conference on World Wide Web*, 355–366. New York, NY, USA: ACM. https://doi.org/10.1145/2488388.2488420
10. Demartini, G., Difallah, D. E., & Cudré-Mauroux, P. (2012). ZenCrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. *Proceedings of the 21st International Conference on World Wide Web*, 469–478.
11. Duchastel, P., & Chen, Y.-P. (1980). The use of marginal notes in text to assist learning. *Educational Technology*, *20*(11), 41–45.
12. Fader, A., Soderland, S., & Etzioni, O. (2011). Identifying Relations for Open Information Extraction. *Empirical Methods in Natural Language Processing*, 1535–1545.
13. Gagnon, M. (2013). Les bibliothèques numériques sont-elles solubles dans le Web sémantique? *Documentation et Bibliothèques*, *59*(3), 161–168.
14. Gotti, F., & Langlais, P. (2016). Harnessing Open Information Extraction for Entity Classification in a French Corpus.
15. Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., … Weikum, G. (2011). Robust disambiguation of named entities in text. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 782–792.
16. Hyvönen, E. (2012). Publishing and using cultural heritage linked data on the semantic web. *Synthesis Lectures on the Semantic Web: Theory and Technology*, *2*(1), 1–159.

17. Léchelle, W. (2019). *Protocoles d'évaluation pour l'extraction d'information libre*. Retrieved from https://papyrus.bib.umontreal.ca/xmlui/handle/1866/22659
18. Manning, C. D., Raghavan, P., & Schütze, H. (2018). *Introduction to information retrieval*. Cambridge: Cambridge University Press.
19. Marchand, E., Gagnon, M., & Zouaq, A. (2020). Extraction of a knowledge graph from French cultural heritage documents. *ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium*, 23–35. Springer.
20. Marchionini, G. (2006). Exploratory search: From finding to understanding. *Communications of the ACM*, *49*(4), 41–46.
21. Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011). DBpedia spotlight: Shedding light on the web of documents. *Proceedings of the 7th International Conference on Semantic Systems*, 1–8.
22. Pal, H. (2016). Demonyms and compound relational nouns in nominal open IE. *Proceedings of the 5th Workshop on Automated Knowledge Base Construction*, 35–39.
23. Palmer, C. L., Teffeau, L. C., & Pirmann, C. M. (2009). Scholarly information practices in the online environment. *Report Commissioned by OCLC Research. Published Online at: Www. Oclc. Org/Programs/Publications/Reports/2009-02. Pdf*.
24. Saha, S. (2018). Open information extraction from conjunctive sentences. *Proceedings of the 27th International Conference on Computational Linguistics*, 2288–2299.
25. Saha, S., Pal, H., & Mausam. (2017). Bootstrapping for Numerical Open IE. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 317–323. Vancouver, Canada: Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-2050
26. Seringhaus, M. R., & Gerstein, M. B. (2007). Publishing perishing? Towards tomorrow's information architecture. *BMC Bioinformatics*, *8*(1), 17.
27. Shotton, D. (2009). Semantic publishing: The coming revolution in scientific journal publishing. *Learned Publishing*, *22*(2), 85–94.
28. Shotton, D., Portwin, K., Klyne, G., & Miles, A. (2009). Adventures in semantic publishing: Exemplar semantic enhancements of a research article. *PLoS Comput Biol*, *5*(4), e1000361.
29. Unsworth, J. (2000). Scholarly primitives: What methods do humanities researchers have in common, and how might our tools reflect this. *Symposium on Humanities Computing: Formal Methods, Experimental Practice. King's College, London*, *13*, 5–00.
30. Wolfe, J. L., & Neuwirth, C. M. (2001). From the margins to the center: The future of annotation. *Journal of Business and Technical Communication*, *15*(3), 333–371.
31. Yates, A., Banko, M., Broadhead, M., Cafarella, M. J., Etzioni, O., & Soderland, S. (2007). Textrunner: Open information extraction on the web. *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 25–26.
32. Zhang, Y. (2010). Developing a Holistic Model for Digital Library Evaluation. *Journal of American Society for Information Science and Technology*, *61*(1), 88–110.