Build a corpus of scientific articles with semantic representation

Jean-Claude Moissinac

LTCI, Telecom ParisTech, Universite Paris-Saclay, 75013, Paris, France

Abstract. As part of the SemBib project, we undertook a semantic representation of the scientific production of Telecom Paristech. Beyond the internal objectives, this enriched corpus is a source of experimentation and a teaching resource. This work is based on the use of text mining methods to build graphs of knowledge, and then on the production of analyzes from these graphs. The main proposal is the disjoint graph production methodology, with clearly identified roles, to allow for differentiated uses, and in particular the comparison between graph production and exploitation methods. This article is above all a methodological proposition for the organization of semantic representation of publications, relying on methods of text mining. The proposed method facilitates progressive enrichment approaches to representations with evaluation possibilities at each step.

Keywords: semantic, publication, LOD, SPARQL

1 Introduction

The SemBib project is an initiative within Telecom ParisTech to build and operate a knowledge graph on our scientific publications. Faced with large warehouses of bibliographic references, we consider that a federation of projects similar to SemBib makes sense. In particular, an institution is better able to work on the quality of its own corpus and to tend towards a significant representation of its own production, which large warehouses can not provide. We therefore argue for a federation of local repositories of semantically related scientific articles [7].

SemBib also serves as a basis for experimenting with semantic representation and graph exploration. We think it is useful to rely on this type of representation to realize and extend the many bibliometric approaches developed over the past decades. Some locate in 1950, others in 1926, still others before the basics of bibliometrics and scientometry. Of course, we take into account this long history, but we are mainly interested in the services that can be rendered by researchers semantic representation methods applied to scientific production, rapidly growing, often exceeding the individual capabilities of exploration. Finally, SemBib is a good case study to discover the use of graphs of knowledge, motivating for our students, who regularly choose to carry out projects on these data.

In this article, we present the constitution of the corpus in section 2 and the specific problems raised. In Section 3, we illustrate uses of this representation and

present methods for enriching the corpus. In Section 4, we present SemBib-based realizations and sketch methods that intersect knowledge graphs and learning methods. Finally, in conclusion, we present our perspectives on the publication of this data and methodological advice to interconnect similar graphs.

2 Build the corpus

2.1 Context

Many initiatives aim to improve the pathways in the body of knowledge of scientific publications. Some apply to give an analytical vision of a set of quotes. For example, the work of [9] to associate skills with people by analyzing their publications. Others, for example, help find relevant documents on a given subject such as [8]. The most successful work we have identified on this subject is that of [10].

Massive access to bibliographic data is also offered by some major SEO systems (Google Scholar [1], Microsoft Academic Graph [2] (accessed 1/12/2017), for example). See [7] for a more complete analysis of bibliographic data sources.

However, all these solutions prove to give a very partial vision of our scientific production and are not open to the production of analyzes and exploitation extending those already proposed. As shown by citet Larsen: 2010: Scientometrics: 20700371, the number of scientific publications is growing rapidly; It is therefore becoming increasingly important to equip users of these publications so that they can use them effectively.

2.2 Reference data

SemBib relies on the existence of an already constituted database which lists the main meta-data on Telecom ParisTech's publications [3]. There are similar bases in many institutions. The approaches presented here make it possible to enrich and interconnect such bases, in a decentralized approach to bibliographic information. This approach is part of a movement that aims to promote the interconnection of scientific publications repositories exploiting enrichments enabled by the semantic web [1].

This database identifies 11311 [4] documents published since 1969. This database constitutes an essential reference, at least for the titles, the year and the authors of each publication, as well as the attribution to Telecom ParisTech . Other meta-data are unequally filled in: unique DOI identifier, URL to the complete document, keywords ...

Of the 11311 documents referenced:

– 1394 associates keywords proposed by an author, that is 12 %,

[1] https://scholar.google.fr/ (accessed 1/12/2017)
[2] https://academic.microsoft.com
[3] http://biblio.telecom-paristech.fr/cgi-bin/selectform.cgi (accessed 1/12/2017)
[4] to 20/11/2017

- 1048 associates a URL, supposed to designate a source of the complete document (9 %); however, in many cases this URL is out of date, or points to a page that offers a redirect to a sold version or hard-to-access document by program,
- 1178 associates a DOI identifier, that is 10 %; the notations used here, informed by the authors, have a great variability, for example they present themselves with or without doi: prefix.

In addition, we noted a great variability in the strings used to designate the publication channels (conferences or journals).

These characteristics make this base little suitable for a thorough exploitation, but it makes it possible to know the entities - authors and publications - on which we must collect information and to estimate if collected data must be associated with one of our publications . Moreover, from this base, our knowledge of many of the entities it mentions allows us to create reference datasets, constituting 'ground truths'.

Large warehouses of bibliographic data exist elsewhere. Unfortunately, they give a very truncated view of our production, in particular because they are not able to resolve the changes in the name of our institution and their usual variants. Moreover these bases do not have information on internal structures of research: projects, departments and groups of research ...

Documents known from large warehouses use many different titles at the level of the affiliation of authors, which makes uncertain research on these warehouses of all publications attributable to Telecom ParisTech. For example, HAL lists only 3001 publications [5] on the 11311 known from our database. They also often use multiple variants for author names.

2.3 Functions sought and problems posed

The objective is first to enrich the semantic descriptions of each author of each of our publications and a description of each publication and publication channel.

For that, we have to make representation model choices. Our hypothesis is that advances in the semantic web are able to give us new ways to efficiently exploit the data we collect, in order to provide research and analysis functions.

Given the weaknesses of the basic data that we have, we must first find other sources of data and consolidate the whole by cross-linking data sets. Once the data has been collected, and in particular the published documents, we have to analyze them to produce semantic representations (assuming that these representations provide a good basis for very diverse analyzes)

This assumes in particular:

- to validate the relevance of documents and data collected
- to remove ambiguities about affiliations, author names, titles, publication channels

[5] at 8/1/2018

– to model the different types of data for their integration in one or more semantic graphs in the spirit of LOD (Linked Open Data)

These problems show why the constitution of a qualified corpus can be complex, the interest to equip its build and finally, the interest that the collected data can have for some thematic works, and text processing with the help of graphs of knowledge.

2.4 Choice of representation

Our choices are driven by the main choice: to rely on semantic web technologies. For the data from our reference database, it seemed useless to create strong links with the rest of Linked Open Data. So we created our own vocabulary which is a direct projection of the data of the reference database, except for some data for which an obvious choice was available: foaf for the names of persons, dc (Dublin Core) for some information on the documents .

In a second step, after studying several specialized ontologies, we selected the SPAR ontology family for its coverage of a large set of concepts related to bibliographic representation. We have represented with SPAR concepts and properties some values associated with our publications (for example with the URI fabio:ResearchPaper). Finally for the concepts that make up our domain vocabularies (see below) we searched the links with schema.org, DBPedia, Wikidata and we will integrate Wordnet.

2.5 Gather documents and perform basic treatments

A Python crawler made it possible to systematically search documents referenced in our database, when they were not available on our site. We were able to collect about 5000 documents out of the 11311 referenced as a publication of Telecom Paristech. As producers of these documents, we have the right to hold a copy for our treatments; an analysis is underway to determine which ones will be made publicly available in respect of the rights of publishers. Most documents are in PDF format.

The crawler retrieves documents searched on the Web from the title. Then, it must be verified that the document obtained corresponds to the document sought. We were inspired by the strategy of [10]: we check that the title is very similar, that the number of authors is identical, that the author names are very similar and finally that the date of publication is the same (if available); documents that do not meet these conditions are kept to supply our database, but not associated with the authors or Telecom ParisTech. This simple approach allowed us to be very precise (98% of a sample of 100 verified papers was correct) The recall is worse since out of 11311 listed publications, we only recovered 4939 (43

For each document:

– a semantic representation of the metadata has been realized by relying on SPAR ontologies (bibo, fabio) and ontologies commonly used for documents (Dublin Core, schema, foaf ...);

- the extraction of a structured representation was performed using the tool GROBID [4] which produces a TEI representation [6] of the article; it allows us to have access to different parts of the text, keeping structural and rhetorical information about each portion of text; a graph for each article is under design based on the DoCO ontology[7] [[2]];
- the keywords indicated by the authors in our database or extracted from the documents were added in a graph constituting a basic vocabulary of the domain; we will talk below about the graphs of concepts constructed from words; several graphs have been constructed;
- the summary was taken from this last representation to be added in the semantic graph;
- the frequency of the terms of each document has been calculated and stored in an associated file; we will see later if it is necessary to put this information in a graph of knowledge;
- The most common terms in each document, excluding empty or blacklisted terms, have been added to another chart.

Graphs For semantic representation, our approach is to split the data into several graphs that refer to each other; for example, there is a graph for people whose URIs are used in the publication graph, which also refers to the graph on publication channels. This approach transposes the SoC design principles - Separation of Concerns - applied in software design; it helps to facilitate the disjoint evolution of graphs with different functions. The references use the principles of the semantic web: each reference is designated by a URI.

The main graphs are:

- a graph for people (authors), comprising 7478 people (it includes co-authors external to Telecom ParisTech) described by 50411 facts;
- a graph for publications, including 11311 publications described by 195857 facts (RDF triplets);
- a graph for publication channels, comprising 4407 channels described by 6599 facts;
- a graph for the concepts of the domain, comprising 15964 entities described by 43878 facts

The sizes of the graphs are constantly evolving as successive enrichments are made to the different graphs by integrating data from new sources. For example, we can see that the graph on the publication channels contains barely more than one fact per channel. This is the next graph that we will significantly enrich with WikiCFP and the conference site pages referenced in our reference database.

The choice is made to add entities to the graph of concepts whenever a new concept seems useful - for example on criteria related to the TfIdf on the corpus. We limit this graph to concepts isolated from each other, possibly

[6] http://www.tei-c.org/
[7] http://www.sparontologies.net/ontologies/doco/source.html (accessed 24/11/2017)

linked to external entities, like DBPedia entities. Relationships between concepts are described in a separate graph -or different versions of such graph- to test different linking strategies and compare their results. We will see in section 3 the consequences of the general approach that leads to working with a set of disjoint graphs.

As another example, we need to link publications with concepts. We have a graph for concepts and a graph for publications. Then, we can build a graph which contains only links between concepts and publications. So, neither the concepts graph, neither the publications graph are modified. We can then try different strategy to associate concepts and graphs and compare them or with a ground truth.

The division into disjointed graphs also facilitates the constitution of graphs that will serve as "ground truth"for various operations, in particular concerning data collection methods and methods of associating themes with other entities (authors, publications, conferences).

Concepts graphs We have chosen to create several graphs of concepts, which will differ mainly by their method of construction, but possibly also by their methods of exploitation. As seen above, the keywords indicated by the authors were put in a first graph. However, we have seen the low use of these words in our database: a tenth of the publications have associated keywords, less than half of the authors fill in the keywords always or sometimes. Also, we added to this graph the keywords taken from the articles by our TEI extraction with GROBID. For each retained word, after normalization, a URI in our namespace was created. This graph is our first domain vocabulary.

A second graph was fed by a selection of words obtained by calculating the TfIdf coefficients for all the words encountered in a sub-part of the corpus covering 5 recent years (about 4000 references, but only about 1200 complete texts at the time of realization) and keeping the words appearing as the most significant. This is a second vocabulary of the domain.

A third graph was build by eliminating the hollow words and blacklisted words from the full text obtained through the extraction with GROBID of the complete texts from a sub-part of the corpus covering 5 recent years.

We believe that it is useful to build these graphs, and perhaps others later, in order to easily compare different approaches. A systematic comparison of these three graphs remains to be established. The exploitation of the current SemBib data was done with the first graph. We will see later that the concepts retained can constitute a basis for an analysis of the texts based on vectors of concepts instead of vectors of words.

The constitution of these graphs linked to external semantic graphs constitutes a basis of work for analysing texts by relying on semantic representations and not only on sets of words or linguistic structures. This extends the steps we presented in [11] for the coupling of structural and semantic representations of texts. Companion graphs are in progress to associate their concepts with reference graph concepts.

Dissemination channel and affiliations graph: associated treatments Whether in our reference database or in external sources, many variants of character strings are used to designate the same series of conferences-acronyms, acronyms with year, expanded name, ...- or the same conference. For example, the conferences in the series where we publish the most, ICASSP, are referenced in our sources under 31 different denominations. It is the same for the designation of Telecom ParisTech, for which we have met 53 different designations.

We worked on the construction of a graph whose main nodes are unique URIs associated with each publication channel, eg conference, each conference series, each affiliation. Each conference is associated with a set of names and, if necessary, a series of conferences. In the publication graph, each publication is associated with a publication channel and one or more affiliations. To bring back to a single URI, each designation pertaining to a single organization, we used a semi-interactive method, which finds its place when the number of elements to be treated is limited.

For example, for Telecom ParisTech, we have:

- manually created a first list of known representations
- added a standardized version of these representations: all set to lowercase, accented characters replaced by their unaccented equivalent ...
- searched in all affiliations found in our sources those that contain one of the known basic representations or a similar representation
- checked in the list found the new acceptable designations that are then added to the list of known designations
- re-iterated this process until no new acceptable designation is found

This treatment has two consequences:

- all the publications concerned by these designations could be attached to a single organization designated by its URI;
- in future publications, affiliations can be compared to this reference list to increase the chances of attaching the publications of the authors of Telecom ParisTech to this institution

The same process was applied to conferences for which the initial list was reduced to the acronym of the conference, possibly supplemented by a form of the expanded name. The methods thus implemented for identification, disambiguation and coherence are applicable to other designations.

3 Explore a bibliographic graph

3.1 SPARQL Access Point

A usual access mode on fact graphs is to establish queries on this graph via a SPARQL access point. The graphs presented previously can be exploited via a SPARQL access point. SPARQL is a query language on sets of facts described by RDF triplets and gathered in graphs (see example below). Several

implementations were assured concerning the database and the access point SPARQL (Triple Store): with Virtuoso, with Jena-Fuseki, with ARC2.

The advantage of having distributed the data in several graphs is related to the limitation of the size of each graph, both in number of nodes and number of relations, but, above all, to separate the facts to organize them. With ARC2, this requires writing queries knowing when to differentiate graphs where to look for certain data. This complicates the writing of requests (see example below). This is the only possibility with ARC2. Options of Virtuoso and Jena-Fuseki make it possible to establish queries that consider as a default graph the merger of all the graphs, which greatly simplifies their writing. [3] show that the performance of different access points varies according to the complex characteristics of the graphs actually targeted. We therefore did not rely on performance considerations measured by benchmarks for our access point choices, but on technical considerations related to hosting.

Tests on queries involving several graphs simultaneously were made with Jena-Fuseki. Surprisingly, they show better performance on merged graphs than with a query specifying graphs where to look for data.

For example, the query [8]:

```
select distinct ?title {?if ieee:title ?title.
                       ?s dcterms:title ?title}
```

which explores the default pseudo-graph to give 824 titles runs 100 times in 19s, while the query

```
select distinct ?title {
graph sb:ieee { ?if ieee:title ?title}.
graph tpt:library {?s dcterms:title ?title}}
```

which distinguishes the data sources of each graph, gives the same result in 24s.

Example: authors sharing keywords The following query:

```
select distinct ?c1 ?c2 {
graph tpt:biblio {
   ?s1 a fabio:ResearchPaper;
       dcterms:creator ?c1;
      schema:keyword ?k1;
      schema:keyword ?k2.
   filter (?k1! =?k2).
   ?s2 a fabio:ResearchPaper;
         dcterms:creator ?c2;
      schema:keyword ?k1;
      schema:keyword ?k2.
```

[8] to simplify, here and in the following, the prefixes have been omitted

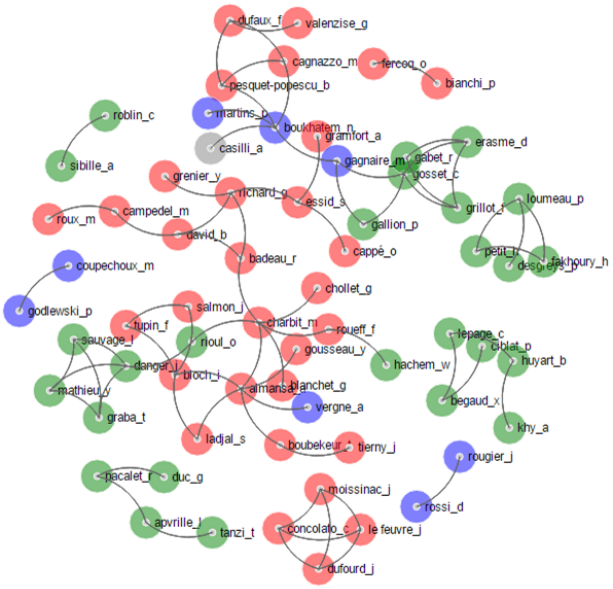

Fig. 1. Network of researchers sharing keywords

```
    filter (?k1!=?k2).
    filter (?s1!=?s2).
    filter (?c1!=?c2).
  }
}
}
```

lets you select two authors (?c1 and ?c2) from two different publications (?s1 and ?s2) that share two keywords.

A similar request, covering only the permanent employees of Telecom ParisTech, made it possible to establish the graphical representation of the Figure 1. It reveals communities that do not appear directly in the structural organization of Telecom ParisTech.

SPARQL access points provide results in JSON format, which is the preferred format for representing data on the Web. This format is very suitable for data consumption by the D3 library, with which Figure 1 has been produced. D3 makes it possible to produce graphics for the Web with SVG technology.

Important progress remains to be made to improve the exploitation of semantic graphs. their adoption, especially on the Web, is still very limited compared to

the potential of these representations, especially driven by thematic operations - music, events ... - carried by major search engines (cf schema.org). Users, be they developers, with well thought-out APIs, or end users wishing to explore the data, need to be helped to take advantage of these representations. Facilities must be made available to developers to produce interfaces integrating in depth the possible enrichments through the Semantic Web.

We explored the use of:

- SemanticForms [9] to create forms for creating, editing, and enriching facts in a semantic graph,
- Uduvudu which proposes a model of development of Web interfaces based on queries on semantic graphs; the model allows a good separation of skills between the semantic data specialist, the application designer and the interface developer; the portal referred to in the next section is based on Uduvudu [5].

Ideally, in our approach, we need to have a SPARQL endpoint which enables the configuration of the default graph by choosing several graphs and aggregating them in it. An efficient solution for that is to be found.

3.2 Access to data linked to the SemBib portal

A SemBib data access portal has been implemented [10]. It is evolving rapidly to integrate the possible enrichments thanks to the graphs generated.

For an author, we will have HTML pages like:

`http://givingsense.eu/sembib/onto/persons/David_Bertrand`

this same access makes it possible to obtain information in JSON format easily exploitable by software in all languages, with the following similar address:

`http://givingsense.eu/sembib/onto/persons/David_Bertrand.json`

The same principles are applied to publications. Thus, without the need for SPARQL queries, it is also possible to access part of the data programmatically. In addition, the generated pages embed the corresponding RDFa data inside the page. This makes this data directly usable by search engines and other SEOs, which helps to improve the visibility of our work.

Exploring and enriching the concept graph A common approach is to associate concepts with authors, publications and publication channels. This association is a basis for mapping between (groups of) authors, or between a (group of) author and a publication or publication channel. In this section, we discuss the association of concepts with different entities.

Only one tenth of the publications have keywords associated with the authors when they register their publications in our database. Less than half of the authors always provide or sometimes key words. Only 39 keywords are used more than 5 times in the database. This relative weakness of our base has prompted

[9] https://github.com/jmvanel/semantic_forms (accessed 8/1/2018)
[10] http://givingsense.eu/sembib/

us to collect many more values - keywords, concepts, themes - directly from the content of articles (see above concept graphs).

Our main current work focuses on the enrichment and exploitation of the graph of concepts generated from the vocabularies encountered in articles (keywords, distinctive words obtained with Tf-Idf ...). In the future, other classical methods of text mining can be integrated to couple and enrich them by semantic graph methods. The idea is to use external graphs -DBpedia, Wikidata, Wordnet- to establish relationships between the concepts we use and exploit these relationships to better interpret the data.

Companion graphs of those previously mentioned are being compiled. It contains owl:sameAs, dc:subject and skos:broader links between entities of the different graphs or to external graphs such as DBPedia, Wikidata, ... owl: sameAs allows to indicate that two URIs are considered as designating the same entity- and skos:broader allows us to specify a hierarchical relationship between entities. We isolate these associations from the main concept graph in order to work on different association strategies.

The first step is to try to associate each concept of our graph with at least one concept of one of the external graphs. Several approaches have been implemented for this association. For example, using the DBPedia Spotlight service or the DBPedia Lookup service.

Then we need to establish links between concepts of our graph. The approach that seems most promising is that proposed by [6]. The principle is to establish a representation of each concept of our graph from existing relationships in DBPedia, then to establish an evaluation of the similarity between these representations and, finally, to establish a relation between the most similar concepts.

We have also begun to evaluate methods of learning based on words vectors, but transposed on concept vectors, exploiting the possibility of grouping concepts from pre-established similarities [6]. Finally, the association between entities of SemBib based on the concepts associated with each of these entities can meet many needs, see [10].

A rigorous evaluation of these different possibilities has not yet been carried out. The structuring of the data we have adopted lends itself well to the implementation of very different approaches. This is particularly useful for proposing student projects addressing both text mining and RDF semantic representations.

Achievements based on SemBib The availability of these structured data graphs of knowledge has already made many achievements. We have already mentioned the search for authors sharing keywords, we can also mention as examples:

- search for publication channels where Telecom ParisTech researchers are the most active;
- Telecom ParisTech researchers who have already published in a common channel,
- a graph of co-authors
- search for articles similar to a given article (student project)

4 Conclusion and Outlook

We have seen a general methodological approach for testing different approaches to the description of bibliographic entities by association with concepts. Our approach consists in producing datasets in the form of disjoint graphs in their subjects and their implementation, but interconnected by relations. We also discussed methods of ambiguity removal on the characteristics of bibliographic entities. An important next step is the RDF documentation of this dataset with DCAT, and then the publication of this data. Indeed, we are convinced that such a dataset with qualified data can be a valuable material for scientometric work and the search of scientific documents. Another result is the availability as linked data of a semantic description of all the scientific publications of Telecom Paristech.

References

1. Capadisli, S., Guy, A., Verborgh, R., Lange, C., Auer, S., Berners-Lee, T.: Decentralised Authoring, Annotations and Notifications for a Read-Write Web with dokieli, pp. 469–481. Springer International Publishing, Cham (2017), https://doi.org/10.1007/978-3-319-60131-1_33
2. Constantin, A., Peroni, S., Pettifer, S., Shotton, D., Vitali, F.: The document components ontology (DoCO). Semantic Web 7(2), 167–181 (2016), http://dx.doi.org/10.3233/sw-150177
3. Duan, S., Kementsietsidis, A., Srinivas, K., Udrea, O.: Apples and oranges: a comparison of RDF benchmarks and real RDF datasets. In: SIGMOD Conference. pp. 145–156. ACM (2011)
4. Lopez, P.: Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In: Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries. pp. 473–474. ECDL'09, Springer-Verlag, Berlin, Heidelberg (2009), http://dl.acm.org/citation.cfm?id=1812799.1812875
5. Luggen, M., Gschwend, A., Anrig, B., Cudré-Mauroux, P.: Uduvudu: a graph-aware and adaptive ui engine for linked data. In: LDOW@WWW (2015)
6. Mirizzi, R., Noia, T.D., Sciascio, E.D., Ragone, A.: Using DBpedia for searching related terms in the IT domain. Tech. rep., Politecnico di Bari, Via Orabona, 4, 70125 Bari, Italy (2012)
7. Moissinac, J.C.: Pour une fédération de dépôts locaux d'articles scientifiques sémantiquement reliés. In: ToTh (2017)
8. Rizzo, G., Tomassetti Federico, Vetrò, A., Ardito, L., Torchiano, M., Morisio Maurizio, Troncy, R.: Semantic enrichment for recommendation of primary studies in a systematic literature review. Digital Scholarship in the Humanities, Oxford University Press, 13 August 2015 (08 2015), http://www.eurecom.fr/publication/4675
9. Sateli, B., Löffler, F., König-Ries, B., Witte, R.: Semantic user profiles: Learning scholars' competences by analyzing their publications. In: Semantics, Analytics, Visualisation: Enhancing Scholarly Data (SAVE-SD 2016). Springer, Springer (04/2016 2016)

10. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer: extraction and mining of academic social networks. ACM Knowledge Discovery and Data Mining pp. 990–998 (2008)
11. Vincent, G., Moissinac, J.C., Luc, A.: Automated generation of a "lossless semantic" eBook. In: 17ème Colloque International sur le Document Numérique (CIDE 17), le livre post-numérique : historique, mutations et perspectives. Fès, Morocco (Nov 2014), https://hal.archives-ouvertes.fr/hal-01097869

In not differently specified, all links were last followed on January 12, 2018.