On the Automated Generation of Scholarly Publishing Linked Datasets: The Case of CEUR-WS Proceedings

Francesco Ronzano(✉), Beatriz Fisas, Gerard Casamayor del Bosque, and Horacio Saggion

TALN Research Group, Universitat Pompeu Fabra,
C/Tanger 122, 08018 Barcelona, Spain
{francesco.ronzano,beatriz.fisas,gerard.casamayor,
horacio.saggion}@upf.edu

Abstract. The availability of highly-informative semantic descriptions of scholarly publishing contents enables an easier sharing and reuse of research findings as well as a better assessment of the quality of scientific productions. In the context of the ESWC2015 Semantic Publishing Challenge, we present a system that automatically generates rich RDF datasets from CEUR-WS workshop proceedings and exposes them as Linked Data. Web pages of proceedings and textual contents of papers are analyzed through proper text processing pipelines. Semantic annotations are added by a set of SVM classifiers and refined by heuristics, gazetteers and rule-based grammars. Web services are exploited to link annotations to external datasets like DBpedia, CrossRef, FundRef and Bibsonomy. Finally, the data is modelled and published as an RDF graph.

Keywords: Semantic Web · Information extraction · Scholarly publishing · Open Linked Data

1 Extract and Semantically Model Scholarly Publishing Contents

During the last few years several approaches have been proposed to turn on-line information into Linked Datasets, dealing with contents coming from a huge variety of domains and ranging from structured to semi-structured and unstructured sources. Proper languages [3] and tools [4,5] to *map a relational database schema to ontologies and automate the generation of RDF triples from it* have been developed [2]. *Semantic annotation and generation of RDF graphs from textual contents* have also been deeply investigated. In this context, information extraction techniques and tools are widely exploited to mine concepts and relations from texts, ranging from the identification of shallow linguistic patterns

The work described in this paper has been funded by the European Project Dr. Inventor (FP7-ICT-2013.8.1 - Grant no: 611383).

F. Gandon et al. (Eds.): SemWebEval 2015, CCIS 548, pp. 177–188, 2015.
DOI: 10.1007/978-3-319-25518-7_15

typical of open-domain approaches [18] to methodologies that strongly rely on semantic knowledge models like ontologies [19,20]. On-line tools and Web services to extract Named Entities from documents and disambiguate them by associating proper URIs are currently extensively available. Systems like NERD [6] and the RDFa Content Editor [7] compare many of these tools and mix their output. Current approaches to create RDF graphs by processing unstructured texts often rely on deep parsing and semantic annotation of textual contents to support the generation of RDF triples. Examples of this kind of systems are LODifier [8] and the text analysis pipeline presented by [9].

In such a context of extensive creation and exploitation of semantic data, scholarly publishing represents a knowledge domain that would strongly benefit from an enhanced structuring, interlinking and semantic modeling of its contents [10]. This goal represents the core objective of **semantic publishing** [1,11]. Semantic Web technologies are an enabling factor towards this vision [12]. They provide the means to structure and semantically enrich scientific publications so as to support the generation of Linked Data from them [13,14], thus fostering the reproducibility and reusability of their outcomes [21]. Recently, a few scientific publication repositories including DBLP[1], ACM[2] and IEEE[3] have been also published as Open Linked Data. In general, however, they expose only basic bibliographic information that is too generic to properly support the diffusion and the assessment of the quality of scientific publications.

With the purpose of experimenting with new approaches to generate rich and highly descriptive scholarly publishing Open Linked Datasets, in the context of the ESWC2015 Semantic Publishing Challenge (2015 SemPub Challenge), in this paper we present a system that automatically analyses the contents of the workshop proceedings of CEUR-WS Web portal, both Web pages and PDF papers, and exports them as an RDF graph. Our system extends our approach to the 2014 SemPub Challenge [22] by dealing with the new information extraction and data modeling needs identified by the 2015 SemPub Challenge. In particular, the 2015 SemPub Challenge proposes two different tasks focused on the extraction of information respectively from CEUR-WS Web pages (SemPub Task 1) and from the content of PDF papers published by CEUR-WS (SemPub Task 2). In Sect. 2 we introduce our system motivating our information extraction approach to both Tasks. Section 3 provides a detailed description of all the data processing phases that characterize our system. In Sect. 4 we explain how we semantically model the information extracted from workshop proceedings as an RDF graph by reusing and extending existing ontologies. Section 5 discusses the evaluation of the RDF datasets generated. In Sect. 6 we analyze the lessons we learned when building our system and outline future work.

[1] http://dblp.l3s.de/d2r/.
[2] http://acm.rkbexplorer.com/.
[3] http://ieee.rkbexplorer.com/.

2 Turning On-line Workshop Proceedings into RDF Graphs: Overall Approach

The ultimate goal of the data processing pipelines we developed is to generate rich semantic descriptions of scientific workshops and conferences. In particular, we mined CEUR-WS on-line workshop proceedings to semantically model *detailed descriptive information of each workshop* from Web pages and *data concerning authors, affiliations, cited papers and mentions of funding bodies and ontologies* from PDF papers[4]. In this way we can easily relate and aggregate information across multiple workshops in order to track their evolution and experiment with new metrics to evaluate them.

CEUR-WS on-line workshop proceedings are organized into volumes; at time of writing there are 1343 published volumes. Each volume contains the proceedings of one or more workshops that are usually co-located at the same conference. Each volume is described by an HTML page including links to the PDF documents of the papers presented at the workshop. Microformats[5] and RDFa[6] annotations are available for some of these HTML documents, and missing in others.

In the context of the 2015 SemPub Challenge, we rely on the following considerations to properly process workshop proceedings:

- Since 2010, 20 microformat classes (CEURVOLEDITOR, CEURTITLE, CEURAUTHORS, etc.) have been adopted to annotate HTML pages detailing the contents of each proceeding volume. The occurrences of each class provide a set of examples of relevant kinds of information required to be extracted by SemPub Task 1. **This data can be exploited to train an automatic text annotation system** in order to add these annotations to proceedings where they are not present.
- Several scholarly publishing resources accessible on-line refer and partially replicate CEUR-WS contents in a structured or semi-structured format. Among them there are Bibsonomy[7], DBLP[8] , Wiki CFP[9], the CrossRef Database[10], and FundRef[11]. These resources can be exploited **to support the information extraction process and to make the RDF contents generated by our system strongly linked with related datasets**. In this context, links to DBpedia[12] can also be established by means of SPARQL

[4] For a detailed description of how workshop related data are modeled as an RDF graph, refer to Sect. 4.
[5] A semantic markup approach that conveys metadata and other attributes in Web pages by existing HTML/XHTML tags.
[6] A semantic markup useful to embed RDF triples within XHTML documents.
[7] http://www.bibsonomy.org/.
[8] http://dblp.uni-trier.de/.
[9] http://www.wikicfp.com/cfp/.
[10] http://crossref.org/.
[11] http://www.crossref.org/fundref/.
[12] http://dbpedia.org/.

queries or by relying on more complex Semantic Web Named Entities disambiguation tools like DBpedia Spotlight[13] [17].

On the basis of the previous considerations, we have designed and implemented two data processing pipelines that respectively convert CEUR-WS proceeding volumes and PDF papers into rich RDF datasets.

3 Data Analysis Pipelines

In this Section we describe in detail the data processing pipelines that mine respectively the Web pages of CEUR-WS proceedings (SemPub Task 1, Subsect. 3.1) and the contents of PDF papers (SemPub Task 2, Subsect. 3.2).

3.1 Task 1: Processing CEUR-WS HTML Contents

We mine the information contained in each on-line proceeding by relying on an extended version of the processing pipeline we introduced in the 2014 SemPub Challenge [22]. In particular, we increase the number of external datasets and Web services exploited to support information extraction. We also refine the heuristics useful to validate, sanitize and normalize the data extracted. We keep out from this pipeline the parts that are devoted to process the contents of PDF papers from CEUR-WS proceedings. These components, properly extended, have been integrated in the PDF processing pipeline exploited in the context of SemPub Task 2 (see Subsect. 3.2). Figure 1 outlines the high level architecture of our system. This pipeline is implemented by relying on the GATE Text Engineering Framework[14] [15], and complemented by external tools and interactions with on-line Web services and knowledge repositories. We functionally describe each pipeline component hereafter.

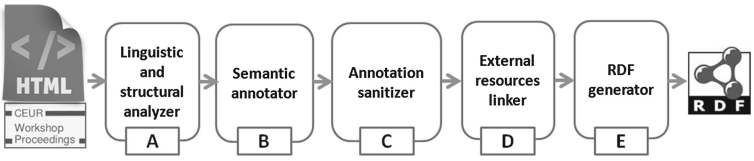


Fig. 1. Task 1: CEUR-WS Proceeding data processing pipeline

(T1.A) Linguistic and Structural Analyzer. Given a set of CEUR-WS proceeding Web pages, their contents are retrieved and characterized by means of linguistic and structural features, useful to support the execution of the following processing steps. In particular, the textual contents of each proceeding are properly split into lines containing homogeneous information by relying on

[13] http://spotlight.dbpedia.org/.
[14] https://gate.ac.uk/.

both HTML markup and custom heuristics. Linguistic analysis is performed in order to tokenize and POS-tag these texts exploiting the information exaction framework ANNIE[15]. Occurrences of paper titles and authors names, acronyms of conferences and workshops, names of institutions, cities and states are pointed out by means of a set of gazetteers; they rely on lists of expressions compiled by crawling WikiCFP, processing the XML dump of DBLP and parsing European Projects information retrieved from the European Union Open Data Portal[16]. Text tokens that denote common names related to research institution (like 'department', 'institute', etc.) or refer to ordinal numbers are also properly spotted.

(T1.B) Semantic Annotator. This component automatically adds semantic annotations to the textual contents of proceedings without semantic markups (volumes up to 558). To this purpose we exploited a set of chunk-based and sentence-based Support Vector Machine (SVM) classifiers [16]. We trained these classifiers over the CEUR-WS microformat annotations existing in proceedings volumes from 559 to 1343. We considered the 14 most frequent microformat classes adopted by CEUR-WS (CEURTITLE, CEURAUTHORS, ect.), thus compiling 14 training corpora. Each corpus includes all the CEUR-WS volumes available on-line that are annotated with the corresponding microformat class. Since we want to model the affiliation of workshop editors and there is no CEUR-WS microformat class for it, we introduced an additional dedicated annotation type, CEURAFFILIATION. We created a training corpus by randomly choosing 75 proceedings that were manually annotated with editor affiliations, thus generating 256 training examples. The first three columns of Table 1 show, for each type of annotation, the number of proceeding volumes where such annotation is present and the total number of annotation examples that are available.

The features added to the textual contents of each proceeding by the *Linguistic and structural analyzer* are exploited to characterize textual chunks and sentences so as to enable their automatic classification. For each annotation type we trained a chunk-based and a sentence-based SVM classifier to automatically perform the annotation task. We chose to automatically annotate proceedings that do not have or include incomplete microformat annotations by exploiting the classifier that better performs for each annotation type (best F1 score, see Table 1).

In general, token-based classifiers perform better with annotation types covering a small number of consecutive tokens that are characterized by a highly distinctive set of features and can be easily discriminated from preceding and following sets of tokens. On the contrary, sentence-based classifiers obtain better results with classes that can be better characterized by sentence level features rather than token level ones.

(T1.C) Annotation Sanitizer. A set of heuristics are applied to fix cases when the annotation borders are incorrectly identified or to delete annotations

[15] http://gate.ac.uk/sale/tao/splitch6.html.
[16] https://open-data.europa.eu/en/data.

Table 1. For each annotation type, number of proceeding volumes including such annotations, number of training examples, precision, recall and F1 score (10-fold cross validation) of token-based and sentence-based SVM classifiers; in bold the classifier chosen to be applied in our system - (*) = manual annotation

Annotation type	Num. Proc.	Num. Examp.	Prec/Rec/F1 (Token)	Prec/Rec/F1 (Sent.)
CEURVOLACRONYM	429	429	0.995/0.980/**0.987**	0.953/0.975/0.963
CEURURN	785	785	1.000/1.000/**1.000**	0.988/1.000/0.994
CEURLOCTIME	785	785	0.973/0.920/0.945	0.966/0.986/**0.975**
CEURVOLTITLE	782	782	0.981/0.909/**0.942**	0.759/0.732/0.745
CEURPUBDATE	581	581	1.000/0.926/0.961	0.997/1.000/**0.999**
CEURVOLEDITOR	785	2901	0.832/0.570/0.676	0.951/0.957/**0.954**
CEURVOLNR	786	786	1.000/0.998/**0.999**	0.998/1.000/0.999
CEURTITLE	784	12807	0.641/0.328/0.434	0.951/0.994/**0.972**
CEURAUTHORS	777	777	0.673/0.376/0.482	0.936/0.982/**0.958**
CEURFULLTITLE	778	778	0.854/0.710/0.775	0.992/0.918/**0.953**
CEURPUBYEAR	777	777	0.998/0.998/0.998	0.998/1.000/**0.999**
CEURPAGES	522	7387	0.983/0.985/**0.984**	0.964/0.987/0.975
CEURSESSION	463	1740	0.930/0.871/0.899	0.876/0.940/**0.906**
CEURCOLOCATED	242	242	0.927/0.928/0.924	0.945/0.975/**0.958**
CEURAFFILIATION (*)	75	256	0.841/0.601/0.699	0.938/0.972/**0.953**

that are not compliant with the normal sequence of annotations of a proceeding (e.g. editor affiliations annotated after the list of paper titles and authors). In addition, links between pairs of related annotations are created (e.g. authors and papers by considering the sequence of annotations or editors and affiliations by means of their markups).

(T1.D) External Resources Linker. This component extends annotations with information retrieved from external resources. Bibsonomy REST API are exploited to link CEURTITLEs to Bibsonomy entries and import the related BibTeX meta-data, if any. DBpedia Spotlight Web Service is exploited to identify DBpedia URIs of occurrences of States, Cities and Organizations in CEURLOCTIMEs and CEURAFFILIATIONs.

(T1.E) RDF Generator. All the information gathered by the previous processing steps is aggregated and normalized so as to generate a highly-informative Open Linked Dataset. The contents of each proceeding are modelled by reusing and extending widespread semantic publishing ontologies. Section 4 provides further details about RDF data modelling.

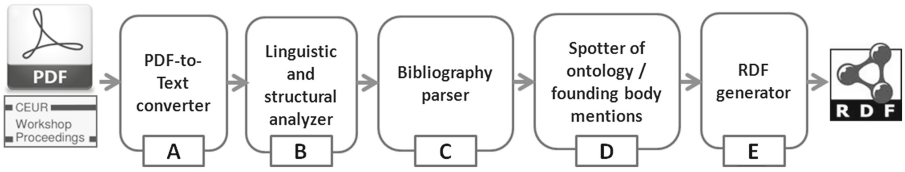


Fig. 2. Task 2: PDF papers data processing pipeline

3.2 Task 2: Mining PDF Papers

In order to extract information from PDF papers as required by SemPub Task 2, we set up a dedicated text analysis pipeline that takes as input one or more PDF papers published by CEUR-WS proceedings and generates an RDF graph. As in SemPub Task 1, the pipeline is based on the GATE Text Engineering Framework. This pipeline takes advantage of part of the external tools and online Web services and knowledge repositories exploited in Task 1. The high level architecture of the pipeline is outlined in Fig. 2. We functionally describe its components.

(T2.A) PDF to Text Converter. We rely on two different PDF-to-text conversion tools: the Web service **PDFX**[17] and the command line utility **Poppler**[18]. Even if the following text analysis phases are mainly based on the textual conversion generated by PDFX, we exploit the output of Poppler to complement it since Poppler preserves information concerning the layout of the original PDF paper. We use this information to support the identification of authors names and to match authors and affiliations in paper headers. PDFX is a PDF-to-text conversion Web service that implements a rule-based iterative PDF analyzer. The style and layout of PDF documents are exploited by PDFX to extract basic meta-data and structural / rhetorical segmentation.

(T2.B) Linguistic and Structural Analyzer. In a similar way to Task 1, the textual contents of each paper are split into lines, tokenized and POS-tagged thanks to the information exaction framework ANNIE[19]. The same gazetteer lists referenced in Task 1 are exploited to point out occurrences of authors names as well as names of institutions, cities and states. Information retrieved from the European Union Open Data Portal and the FundRef founding agencies database is exploited to spot full names, identification numbers and acronyms of European Projects as well as full and abbreviated names of funding agencies. Text tokens that denote common names related to research institution (like 'department', 'institute', etc.) are also identified.

[17] http://pdfx.cs.man.ac.uk/.
[18] http://poppler.freedesktop.org/.
[19] http://gate.ac.uk/sale/tao/splitch6.html.

(T2.C) Bibliography Parser. PDFX spots each bibliographic entry present at the end of the paper. We parse this text by aggregating the results of three on-line services:

- *CrossRef API*[20]: to match free-form citations to DOIs;
- *Bibsonomy API*[21]: to retrieve the BibTeX record of the cited paper;
- *Freecite on-line citation parser*[22]: to identify the constituent elements of a bibliographic entry (author, title, year, journal name, etc.) by applying a sequence tagging algorithm over its tokens.

We enrich each bibliographic entry by merging the processing output of these three services. This information is properly exploited in order to generate the RDF triples modelling the bibliography of the paper.

(T2.D) Spotter of Ontology and Founding Body mentions. This component implements a set of JAPE grammars[23] useful to spot mentions of *ontologies* and *founding bodies* (EU projects, grants, founding agencies). Mention spotting relies on a set of textual patterns that match the annotations produced by the Linguistic and structural analyzer. JAPE grammars have been created by manually analysing the context of occurrences of mentions of *ontologies* and *founding bodies* in the papers of the training set of SemPub Task 2. When mentions of *founding bodies* are matched to entries of the lists of FundRef founding agencies or European Projects, we can enrich such mentions with meta-data like the FundRef URI of the founding agency. These meta-data will contribute to generate a richer RDF graph.

(T2.E) RDF Generator. All the paper-related information gathered by the previous processing steps is aggregated and normalized so as to generate a highly-informative Open Linked Dataset. We exploit widespread semantic publishing ontologies to model the contents of each paper. Section 4 provides further details about RDF data modeling.

4 Modeling Workshop Data as an RDF Graph

In order to properly model the information concerning workshop proceedings and papers we exploited and extended widespread semantic publishing ontologies. In particular, we relied on:

- the *Semantic Web for Research Communities Ontology* (prefix swrc) that is useful to shape many relevant domain concept and relationships;

[20] http://search.crossref.org/help/api.
[21] http://www.bibsonomy.org/help/doc/api.html.
[22] http://freecite.library.brown.edu/.
[23] https://gate.ac.uk/sale/tao/splitch8.html.

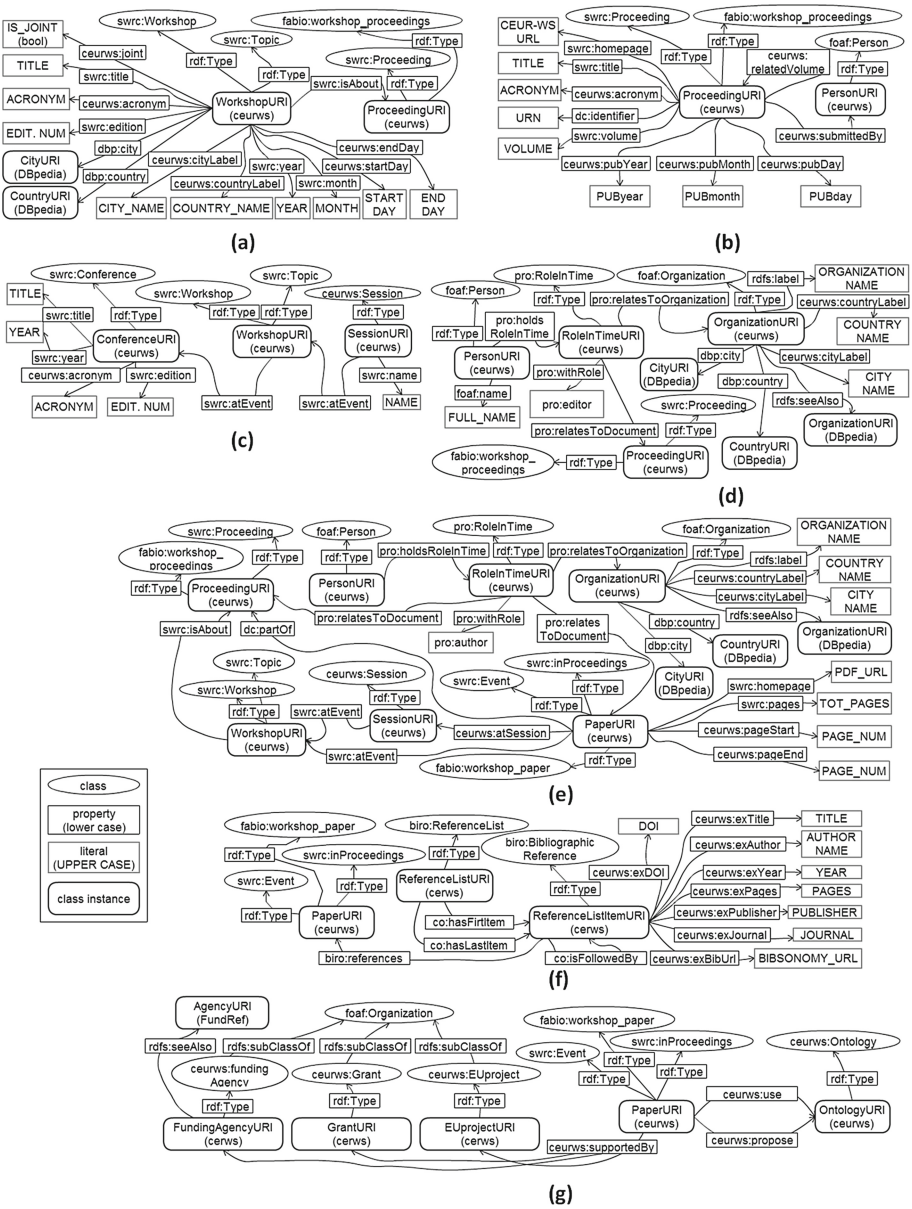


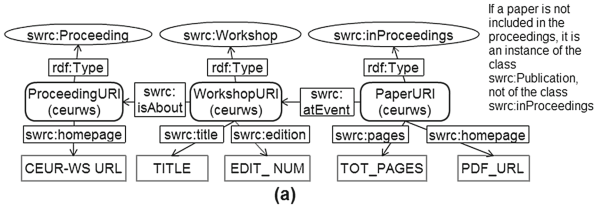
Fig. 3. RDF data models of workshops (a), proceedings (b), conferences (c), editors (d), papers and authors (e)

– the *Bibliographic Reference Ontology* (prefix biro) that is useful to model the bibliographic information of a paper;
– the *FRBR-aligned Bibliographic Ontology* (prefix fabio) that is useful to better characterize bibliographic records of scholarly endeavours like papers and proceedings;
– the *Publishing Role Ontology* (prefix pro) that is useful to model the roles of researchers as editors of workshops and authors of papers.

From the classes and the properties modeled by these ontologies, we have reused and derived - in the ceur-ws namespace - sub-classes and sub-properties: the RDF Datasets we generate from CEUR-WS Proceedings include the related T-BOX axioms. Figure 3 visually represents our data modeling approach.

5 Evaluating Workshop Linked Datasets by SPARQL Queries

The evaluation procedure of the 2015 SemPub Challenge consisted of a set of 20 queries expressed in natural language, each one of them aggregating data of a workshop or serving as an indicator of its quality (e.g. list the full names of all authors who have (co-)authored a paper in workshop W). Participants had to rewrite these queries as SPARQL queries so that the organizers could run them against the participants RDF dataset and evaluate the results. In Fig. 4 we provide an example of a query and its SPARQL formulation for our dataset model.




```
SELECT ?procURL (Count(?procURL) AS ?np)
       (Avg(xsd:int(?numPages)) AS ?al)
WHERE {
     ?proceedings a swrc:Proceedings ;
          swrc:homepage ?procURL ;
          swrc:isAbout ?workshop .
     ?workshop a swrc:Workshop;
          swrc:atEvent ?paper .
     ?paper a swrc:InProceedings ;
          swrc:pages ?numpages
} GROUP BY ?procURL
```

(b)

Fig. 4. (a) RDF model of papers presented at a workshop, included in a proceeding volume; (b) SPARQL query for **Numbers of papers** (?np) and **Average length of papers** (?al)

We covered 15 out of the 20 SPARQL queries proposed by the SemPub Task 1 and the 10 SPARQL queries proposed by the SemPub Task 2. Our system has large margin of improvement with respect to the extraction of the information required in the context of the challenge from the Web pages of CEUR-WS Proceedings (SemPub Task 1). The performance of our pipeline improves when it deals with the extraction of authors' names, country and affiliation and the analysis of bibliographic entries from PDF papers (SemPub Task 2).

6 Conclusions and Future Work

We described a system that extracts structured information from CEUR-WS on-line proceedings by parsing both Web pages and PDF papers and modeling their contents as Linked Datasets.

Our system design has been motivated by the need of flexibility and robustness in the face of different ways in which information is written, structured or annotated in the input dataset. Despite that, we found that **customized and often laborious information extraction and post processing steps are essential to correctly deal with borderline information structures that are difficult to generalize**, like unusual markups, infrequent ways to link authors and affiliations, etc.

In general, we hope that the increasing availability of structured and rich scientific publishing Linked Datasets will enable larger communities to easily discover and reuse research outcomes as well as to propose and test new metrics to better understand and evaluate research outputs. In this context we believe that, in parallel to the investigation of approaches to automate the creation of semantic datasets by mining partially structured inputs, it is also essential to push scientific communities towards standardized, shared and opened procedures to expose their outcomes in a structured way.

References

1. Shotton, D.: Semantic publishing: the coming revolution in scientific journal publishing. Learned Publishing **22**(2), 85–94 (2009)
2. Spanos, D.E., Stavrou, P., Mitrou, N.: Bringing relational databases into the semantic web: a survey. Semant. Web **3**(2), 169–209 (2012). IOS Press
3. World Wide Web Consortium: R2RML: RDB to RDF mapping language. W3C Recommendation (2012)
4. Bizer, C., Cyganiak, R.: D2r server-publishing relational databases on the semantic web. In: Poster at the 5th International Semantic Web Conference (2006)
5. Knoblock, C.A., Szekely, P., Ambite, J.L., Goel, A., Gupta, S., Lerman, K., Muslea, M., Taheriyan, M., Mallick, P.: Semi-automatically mapping structured sources into the semantic web. In: Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V. (eds.) ESWC 2012. LNCS, vol. 7295, pp. 375–390. Springer, Heidelberg (2012)
6. Rizzo, G., Troncy, R., Hellmann, S., Bruemmer, M.: NERD meets NIF: lifting NLP extraction results to the linked data cloud. In: Proceedings of the Linked Data on the Web Workshop (2012)
7. Khalili, A., Auer, S., Hladky, D.: The RDFa content editor - from WYSIWYG to WYSIWYM. In: Proceedings of the IEEE Computer Software and Applications Conference, COMPSAC (2012)
8. Augenstein, I., Padó, S., Rudolph, S.: LODifier: generating linked data from unstructured text. In: Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V. (eds.) ESWC 2012. LNCS, vol. 7295, pp. 210–224. Springer, Heidelberg (2012)
9. Exner, P., Nugues, P.: Entity extraction: from unstructured text to DBpedia RDF triples. In: Proceedings of the Web of Linked Entities Workshop, WoLE (2012)

10. Stegmaier, F., et al.: Unleashing semantics of research data. In: Rabl, T., Poess, M., Baru, C., Jacobsen, H.-A. (eds.) WBDB 2012. LNCS, vol. 8163, pp. 103–112. Springer, Heidelberg (2014)
11. Eefke, S., Van Der Graaf, M.: Journal article mining: the scholarly publishers' perspective. Learned Publishing **25**(1), 35–46 (2012)
12. Bizer, C.: Linking data and publications expert report, global research data infrastructure of European Union (2012)
13. Ciancarini, P., Di Iorio, A., Nuzzolese, A.G., Peroni, S., Vitali, F.: Semantic annotation of scholarly documents and citations. In: Baldoni, M., Baroglio, C., Boella, G., Micalizio, R. (eds.) AI*IA 2013. LNCS, vol. 8249, pp. 336–347. Springer, Heidelberg (2013)
14. Attwood, T.K., Kell, D.B., McDermott, P., Marsh, J., Pettifer, S.R., Thorne, D.: Utopia documents: linking scholarly literature with research data. Bioinformatics **26**(18), 568–574 (2010)
15. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: a framework and graphical development environment for robust NLP tools and applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics, ACL (2002)
16. Li, Y., Bontcheva, K., Cunningham, H.: Adapting SVM for Data sparseness and imbalance: a case study on information extraction. Nat. Lang. Eng. **15**, 241–271 (2009). Cambridge University Press
17. Mendes, P.N., Jakob, M., Garca-Silva, A., Bizer, C.: DBpedia spotlight: shedding light on the web of documents. In: Proceedings of the 7th International Conference on Semantic Systems, pp. 1–8. ACM (2011)
18. Etzioni, O., Banko, M., Soderland, S., Weld, D.S.: Open Information Extraction from the Web. Communications of the ACM - Surviving the data deluge **51**(12), 68–74 (2008)
19. Wimalasuriya, D.C., Dou, D.: Ontology-based Information Extraction: An Introduction and a Survey of Current Approaches. Journal of Information Science **36**(3), 306–323 (2010)
20. Saggion, H., Funk, A., Maynard, D., Bontcheva, K.: Ontology-based information extraction for business intelligence. In: Aberer, K., et al. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 843–856. Springer, Heidelberg (2007)
21. Bechhofer, S., et al.: Why linked data is not enough for scientists. Future Gener. Comput. Syst. Spec. Sect. Recent Adv. e-Sci. **29**(2), 599–611 (2013). Elsevier
22. Ronzano, F., del Bosque, G.C., Saggion, H.: Semantify CEUR-WS proceedings: towards the automatic generation of highly descriptive scholarly publishing linked datasets. In: Presutti, V., et al. (eds.) SemWebEval 2014. CCIS, vol. 475, pp. 83–88. Springer, Heidelberg (2014)