Semantic Publishing Challenge – Assessing the Quality of Scientific Output in its Ecosystem

Anastasia Dimou[1], Angelo Di Iorio[2], Christoph Lange[3,4], and Sahar Vahdati[3]

[1] Ghent University - iMinds, Belgium anastasia.dimou@ugent.be
[2] Università di Bologna, Italy diiorio@cs.unibo.it
[3] University of Bonn, Germany vahdati@uni-bonn.de
[4] Fraunhofer IAIS, Germany math.semantic.web@gmail.com

Abstract. The Semantic Publishing Challenge aims to involve participants in extracting data from heterogeneous sources on scholarly publications, and producing Linked Data which can be exploited by the community itself. The 2014 edition was the first attempt to organize a challenge to enable the assessment of the quality of scientific output. The 2015 edition was more explicit regarding the potential techniques, i.e., information extraction and interlinking. The current 2016 edition focuses on the multiple dimensions of scientific quality and the great potential impact of producing Linked Data for this purpose. In this paper, we discuss the overall structure of the Semantic Publishing Challenge, as it is for the 2016 edition, as well as the submitted solutions and their evaluation.

Keywords: Linked Data, Information Extraction, Challenge

1 Introduction

Changes in technology, tools, funding and social aspects raise new challenges in scholarly publishing. On the other hand, a growing amount of research on publishing and consuming Linked Data, i.e., data represented and made available in a way that maximizes reusability, has facilitated Semantic Web adoption. In this context, the idea of *Semantic Publishing* emerged. Semantic Publishing is "the enhancement of scholarly publications by the use of modern Web standards to improve interactivity, openness and usability, including the use of ontologies to encode rich semantics in the form of machine-readable RDF metadata" [15,16]. It is expected that the semantic publishing advent will foster advanced services for scholars and non-expert users, such as semantic search, identification of research trends, discovery of connections between research works, people and institutions, and so on. However, to achieve such services, richer datasets represented as Linked Data are required. Nevertheless, even though several tools have been developed to generate Linked Data about scholarly publications, the procedure is still cumbersome and most available datasets still have some limitations.

The Semantic Publishing Challenge series aims at investigating novel approaches for improving scholarly publishing using Linked Data technology [3,7].

The first editions focused on extracting information both from non-semantic as well as semantic data sources: these sources range from proceeding volumes and their papers (as published in PDF) to semantically enhanced datasets (with information derived from different data sources).

In this paper we present the 2016 edition of the Semantic Publishing Challenge, giving an overview of the tasks, together with a short introduction of the proposed solutions and their final results.

The remaining of the paper is structured as follows. Section 2 introduces the Challenge, Section 3 presents the different tasks, Section 4 the training and evaluation datasets considered to accomplish the tasks and Section 5 the queries to be addressed. Section 6 presents the different solutions which were submitted and Section 7 their evaluation and the corresponding results.

2 Semantic Publishing Challenge

In 2014, we considered defining a challenge about Semantic Publishing which would act as an enabler for evaluating different solutions producing corresponding Linked Data [7]. The tasks are defined in a way that such solutions can generate richer Linked Data for Semantic Publishing compared to existing datasets. Datasets existing at that time focused on basic bibliographic metadata, or domain specific data, rendering more advanced applications, especially in respect to assessing the scientific output, difficult, if not impossible.

The 2014 edition was designed to produce an initial dataset which would be useful for future challenges and the community could experiment on it. However, the two information extraction tasks with an objective evaluation had received few submissions. Thus, an open task with a subjective evaluation was also introduced. For the 2015 edition, one of the 2014 edition's tasks, **Task 1**, remained the same because the results were encouraging and it was intended to give another opportunity and incentive to the 2014 edition's participants to improve their tools and participate again, though without excluding new participants.

The other task with an objective evaluation, namely **Task 2**, which was focused on extracting information from the papers' full text, remained also the same, but the underlying data source was changed. In 2014, papers encoded in XML JATS[5], a language for encoding journal articles derived from the NLM Archiving and Interchange DTD, and its TaxPub extension for taxonomic treatments, served as the data source. In 2015, the same data source as for Task 1 was considered as the data source for Task 2: the CEUR-WS.org open access computer science workshop proceedings were considered as the input dataset for Task 2 too, aiming to foster synergies between the two tasks and to encourage participants to compete in both tasks. For this edition in 2016, both **Task 1** and **Task 2** remained the same.

Nevertheless, in 2015 there was only one team competing for both tasks (cf. the overview of the 2015 challenge [3]). Therefore, aligning Task 1 and 2 became a priority for the 2016 edition which was expected to be achieved relying on

[5] http://jats.nlm.nih.gov/

Task 3. Task 3 was radically changed from the 2014 edition which was an open task with a subjective evaluation to the 2015 edition whose Task 3 was formed in a way that allowed an objective evaluation. For 2015, it aimed at interlinking CEUR-WS.org dataset with other existing Linked Data. For the 2016 edition, Task 3 was designed to be more focused on promoting synergies and aligning Task 1 and 2. The dataset of the 2015 winning solutions both for Task 1 and 2 were considered as the data sources to be aligned, while aligning the resulting dataset with external Linked Data were of subsequent priority.

3 Tasks and Motivation

In this section, we outline the different tasks defined, or how existing were adjusted for the 2016 edition of the challenge, and we describe the underlying motivation for defining or modifying each one of them.

3.1 Task 1

Task 1 was designed to assess the ability to extract data from a full body of HTML documents. Task 1 is an extension of the 2015 edition. All quality indicators from the previous edition are reconsidered, some are defined more precisely, while one was completely new.

To be more precise, the input dataset for Task 1 consists of HTML documents at different levels of encoding quality and semantics. Therefore, Task 1 mainly requires to employ information extraction and semantic annotation techniques. Participants are asked to extract information from a set of HTML tables of contents published in the CEUR-WS.org workshop proceedings. The extracted information enables describing data which might act as means for assessing the quality of these workshops, for instance by measuring growth, longevity, etc.

Motivation. Common questions related to the quality of a scientific venue include whether a researcher should submit a paper to it or accept an invitation to its programme committee, or whether a publisher should publish its proceedings, and whether a company should sponsor it [2]. Being aware of the quality of an event helps to assess the quality of the papers accepted there.

3.2 Task 2

Task 2 was designed to assess the ability to extract data from the papers full text, namely their PDF corpus. It follows the last two editions' Tasks 2, which were focused on extracting information from citations in the first place, as well as affiliations and funding since the 2015 edition. For the 2016 third edition, the aforementioned are still in scope (apart from citations), but extracting information regarding the internal structure, e.g., tables and figures, comes also in context.

To be more precise, the input dataset for Task 2 consists of PDF documents of papers published with CEUR-WS.org. Therefore, Task 2 mainly requires PDF mining techniques and some natural language processing. The extracted information describes the organisation of the paper and provides a deeper understanding of the context in which it was written. The extracted information should describe, on the one hand, the internal structure of sections, tables, figures and, on the other hand, the authors' affiliations and research institutions, and funding sources.

Motivation. Scientific papers are not isolated units. Common questions related to the quality of a scientific contributions include factors that directly or indirectly contribute to the origin and development of a paper include citations, affiliations, funding agencies or even the venue where the paper was presented. The internal organisation and the structural components of a paper are also good indicators of its quality and potential impact.

3.3 Task 3

Task 3 was designed to assess the ability to generate cross-datasets links. It follows the previous, 2015 edition. However, the 2016 edition narrows down the task's scope to a smaller number of external datasets, whereas cross-task links between the previous edition's Task 1 and 2 datasets are now also explored.

To be more precise, the input dataset for Task 3 consists of Linked Datasets. Therefore, Task 3 mainly requires entity interlinking techniques and some natural language processing. Participants are asked to interlink the CEUR-WS.org dataset with relevant datasets already existing in the Linked Open Data cloud. In particular, they are expected to interlink persons, papers, events, organisations and publications. All these entities are identified, disambiguated and interlinked to their correspondences in other datasets.

Motivation. Scientific papers and venues are not isolated units and should not be considered separately from each other. They belong in a broader context of scientific contributions, offering complementary information when it is associated with prior existing information, else it remains incomplete.

4 Input Dataset

In this section, we describe the input dataset considered for the different tasks. A summary of statistics related to the training and evaluation dataset for each task is summarized in Table 1.

4.1 Task 1 Dataset

To support the evolution of extraction tools, the 2016 training dataset is largely the same as the union of the 2015 training and evaluation dataset, with a few additions. To be more precise, the Task 1 training dataset consists of:

Table 1. Training and evaluation datasets

	Training Dataset	Evaluation Dataset
Task 1		
Workshops	118	50
Task 2		
Papers	45	40
Task 3		
Datasets	5	5

- one HTML index page linking to all CEUR-WS.org workshop proceedings volumes[6] (invalid but still uniformly structured HTML 4);
- the volumes' HTML tables of contents[7], which link to the individual workshop papers. Their format is largely uniform but has changed over time. In more details, the training dataset consists of:
  - valid HTML5 pages with microformats and sometimes RDFa,
  - valid and invalid HTML 4.01 with or without microformats

4.2 Task 2 Dataset

To support the evolution of extraction tools, but also to align with Task 1, the 2016 training dataset is largely the same as the one of the 2015 edition, taken from some of the workshops which are also analyzed in Task 1[8]. The selected papers use different formats and styles (ACM, LNCS, IEEE) and different rules for bibliographic references, headers, affiliations and acknowledgments.

The training and evaluation datasets were totally disjoint (differently form previous the past editions) and shared the same internal structure, with the same distribution of styles. Papers were clustered according to their similarities and randomly selected within each cluster.

4.3 Task 3 Dataset

To align with Task 1 and Task 2, Task 3 considered the previous 2015 edition output from Task 1 and 2, besides the external datasets that already exist in the Linked Data cloud and were considered also in the 2015 edition. In total 5 different dataset were considered for the training and evaluation dataset. First of all, the CEUR-WS.org proceedings dataset as it was formed by the solutions that performed best for Task 1 [8][9] and Task 2 [17][10] in 2015. Then, the COL-

[6] http://ceur-ws.org/
[7] https://github.com/ceurws/lod/wiki/SemPub16_Task1
[8] https://github.com/ceurws/lod/wiki/SemPub16_Task2
[9] http://rml.io/data/SPC2016/CEUR-WS/CEUR-WStask1.rdf.gz
[10] http://rml.io/data/SPC2016/CEUR-WS/CEUR-WStask2.rdf.gz

INDA[11], the DBLP[12] and the Springer LD[13] datasets were also considered as input datasets.

5 Queries

In this section, we describe the queries which the different solutions should be able to answer. Based on the results of those queries, the solutions were evaluated for their capacity to address the different tasks.

5.1 Task 1

The submitted solutions are required to produce a dataset for Task 1 against which the following queries can be answered, roughly ordered by increasing difficulty:

- **Q1.1**: List the full names of all editors of the proceedings of workshop W
- **Q1.2**: Count the number of papers in workshop W
- **Q1.3**: List the full names of all authors who have (co-)authored a paper in workshop W
- **Q1.4**: Identify the full names of those chairs of workshop W who are affiliated in the same country in which the workshop took place
- **Q1.5**: Compute the average length (in numbers of pages) of a paper in workshop W
- **Q1.6**: Find out whether the proceedings of workshop W were published on CEUR-WS.org before the workshop took place
- **Q1.7**: Identify all editions that the workshop series titled T has published with CEUR-WS.org
- **Q1.8**: Identify the full names of those chairs of the workshop series titled T that have so far been a chair in every edition of the workshop that was published with CEUR-WS.org
- **Q1.9**: Identify all CEUR-WS.org proceedings volumes in which papers of workshops of conference C in year Y were published
- **Q1.10**: Identify those papers of workshop W that were (co-)authored by at least one chair of the workshop
- **Q1.11**: List the full names of all authors of invited papers in workshop W
- **Q1.12**: Determine the number of editions that the workshop series titled T has had, regardless of whether published with CEUR-WS.org
- **Q1.13**: Determine the title (without year) that workshop W had in its first edition
- **Q1.14**: Of the workshops of conference C in year Y, identify those that did not publish with CEUR-WS.org in the following year (and that therefore probably no longer took place)

[11] http://www.colinda.org/
[12] http://dblp.l3s.de/dblp++.php
[13] http://lod.springer.com/

- **Q1.15**: Identify the papers of the workshop titled T (which was published in a joint volume V with other workshops)
- **Q1.16**: List the full names of all editors of the proceedings of the workshop titled T (which was published in a joint volume V with other workshops)
- **Q1.17**: Of the workshops that had editions at conference C both in year Y and $Y + 1$, identify the workshop(s) with the biggest percentage of growth
- **Q1.18**: Return the acronyms of those workshops of conference C in year Y whose previous edition was co-located with a different conference series.
- **Q1.19**: Of the workshop series titled T, identify those editions that took place more than two months later/earlier than the previous edition that was published with CEUR-WS.org
- **Q1.20**: Identify the affiliations and countries of all editors of the proceedings of workshop W. Use DBpedia resources for the countries.
- **Q1.21**: Identify the full names of those authors of papers in the workshop series titled T that have so far been a (co-)author of a paper in every edition of the workshop that was published with CEUR-WS.org

5.2 Task 2

The submitted solutions are required to produce a dataset for Task 2, against which the following queries can be answered:

- **Q 2.1** Affiliations in a paper:
  Identify the affiliations of the authors of the paper X.
- **Q 2.2** Countries in affiliations:
  Identify the countries of the affiliations of the authors in the paper X.
- **Q 2.3** Supplementary material:
  Identify the supplementary material(s) for the paper X.
- **Q 2.4** Sections:
  Identify the titles of the first-level sections of the paper X.
- **Q 2.5** Tables:
  Identify the captions of the tables in the paper X
- **Q 2.6** Figures:
  Identify the captions of the figures in the paper X.
- **Q 2.7** Funding agencies:
  Identify the funding agencies that funded the research presented in the paper X (or part of it).
- **Q 2.8** EU projects:
  Identify the EU project(s) that supported the research presented in the paper X (or part of it).

5.3 Task 3

The submitted solutions are required to produce a dataset for Task 3 answering the following queries, roughly ordered by increasing difficulty:

- **Q 3.1** Same person – Multiple URIs:
  Identify and interlink same entities that represent the same editor and/or author but appear with different URIs within the CEUR dataset of Task 1.
- **Q 3.2** Same conference – Multiple URIs:
  Identify and interlink same entities that represent the same conference but appear with different URIs within the CEUR dataset of Task 1.
- **Q 3.3** Same cited paper – Multiple URIs:
  Identify and interlink same entities that represent the same cited paper but appear with different URIs within the CEUR dataset of Task 2.
- **Q 3.4** Same people – Different URIs in CEUR-WS subsets:
  Identify and interlink same entities that represent the same editor and/or author but appear with different URIs within the CEUR dataset of Task 1 and Task 2.
- **Q 3.5** Same workshops in the CEUR-WS and COLINDA datasets:
  Identify and interlink same entities that represent the same workshop but appear with different URIs within the CEUR dataset of Task 1 and the COLINDA dataset.
- **Q 3.6** Same workshops in the CEUR-WS and DBLP datasets:
  Identify and interlink same entities that represent the same workshop but appear with different URIs within the CEUR dataset of Task 1 and the DBLP dataset.
- **Q 3.7** Same people in the CEUR-WS and DBLP datasets:
  Identify and interlink same entities that represent the same person but appear with different URIs within the CEUR dataset of Task 1 and the DBLP dataset.
- **Q 3.8** Cited papers in CEUR dataset presented at conferences described in Springer dataset:
  Identify and interlink same entities that represent the same conference but appear with different URIs within the CEUR dataset of Task 2 and the Springer dataset.

6 Solutions

Five solutions were submitted and accepted for Task 2, while there were no solutions at all submitted neither for Task 1 nor for Task 3.

6.1 Solution 1

Solution 1 by Ahmad et al. [1] proposed a heuristic-based approach that uses a fruitful combination of tag-based and plain-text-based information extraction techniques which is not frequently encountered in bibliography. Their approach identifies patterns and rules from integrated formats which are stored in knowledge bases. The PDF extraction occurs using the PDFX library[14], while the

[14] `http://pdfx.cs.man.ac.uk`

PDFbox Java library[15] is considered to extract the supportive material links because the former extracts them as plain text, whereas the later as links. Besides the PDF parsers (both PDF-to-XML and PDF-to-Text), the entire solution is modular consisting additionally of the following modules: content pre-processing, rule identifier, information extraction and triplification. The information extraction module, in its own turn, consists of the following sub-extractors: (i) authors extractor, (ii) section heading extractor, (iii) table extractor, (iv) figure extractor, (v) supplementary material extractor, and (vi) funding extractor.

6.2 Solution 2

Solution 2 by Klampfl and Kern [6] extended their approach for the 2015 edition [5]. They implemented a processing pipeline that analyzes a PDF document structure incorporating a diverse set of machine learning techniques, unsupervised to extract text blocks and supervised to classify blocks into different meta-data categories. Heuristics are applied to detect the reference section and sequence classification to categorize the tokens of individual references strings. Last, Named Entity Recognition (NER) is used to extract references to grants, funding agencies and EU projects. In 2016, they changed or improved some parts of their solution. They employed different processing steps of their tool which were not used in the previous edition. To be more precise, the current solution processes section headings, hierarchy and captions, but it also introduces novel aspects for extracting links from supplementary material. Its modular structure allows separate training of its parts relying on different datasets.

6.3 Solution 3

Solution 3 by Nuzzolese et al. [10] relied on the Article Content Miner (ACM) which extends the Metadata And Citations Jailbreaker (MACJa – IPA) [9], namely their approach which was submitted to the 2015 edition of the challenge. The tool integrates (i) the PDFMiner, a Python library[16], to extra the information from PDF; (ii) hybrid techniques based on Natural Language Processing (NLP), for instance Combinatory Categorial Grammar, Discourse Representation Theory (DRT), or Linguistic Frames and heuristics that exploit existing tools lexical resources and gazetteers to generate representation structures according to the DRT; (iii) FRED[17], a novel machine reader that produces RDF/OWL ontologies having classes depending on the lexicon used in the text; and (iv) modules to query external services to enhance and validate data.

6.4 Solution 4

Solution 4 by Sateli and Witte [14] relied on LODeXporter[18], a system composed from two modules: (i) a text mining pipeline based on the GATE framework to

[15] https://pdfbox.apache.org/
[16] https://github.com/euske/pdfminer/
[17] http://wit.istc.cnr.it/stlab-tools/fred
[18] http://www.semanticsoftware.info/lodexporter

extract structural (syntactic processing) and semantic entities (semantic processing), leveraging existing NER tools; and (ii) a LOD exporter, to translate the document annotations into RDF according to custom rules. The text preprocessing occurs relying on PDFX to transform the PDF documents into XML documents[19], which are subsequently used by the GATE framework. The GATE framework then tokenises and lemmatises the text, detects sentence boundaries and performs gazetteering on the text, while the DBpedia Spotlight service is used for entity tagging. They also relied on their solution for the 2015 edition of the challenge [13]. In 2016, the PDF extraction tool used was changed and a number of additional or new conditional heuristics were added.

6.5 Solution 5

Solution 5 by Ramesh et al. [11] proposed an approach based on a three-level Conditional Random Fields (CRF) supervised learning approach. Their approach follows the same feature list as [5]. However, they extract PDF to an XML document that conforms to NLM JATS DTD, and generate RDF using an XSLT transformation tool dedicated for JATS. The Apache PDFBox library[20] is used to extract a stream of characters, their bounding boxes and information about their fonts. The aforementioned are fed into the three-level CRF model, namely (i) formatting, (ii) vocabulary, (iii) heuristic and (iv) language modeling features.

7 Evaluation

The different solutions which were submitted, were evaluated using the SemPub Evaluator[21] on a set of forty papers, as described in Section 4, and relying on a set of eight queries, as described in Section 5. The overall evaluation results for each solution are summarized in Table 2. The table presents the precision, recall and F-score for each solution, and the same values for those who participated in the 2015 edition as well. The best performing tool is the one with the highest F-score. For the 2016 edition of the challenge, the Solution 1 by Ahmad et al. was the winner of the best performing tool award.

Unfortunately, the best performing tool of the 2015 edition did not participate again for the 2016 edition. Nevertheless, the most innovative solution for the 2015 edition, namely current Solution 4 by Sateli and Witte, participated again and it was ranked second. For the 2016 edition, Solution 2 by Klampfl and Kern won the most innovative solution award.

8 Conclusions

The 2016 edition of the Semantic Publishing Challenge was built on top of the previous ones. This continuity was crucial for the success of the event: the

[19] http://pdfx.cs.man.ac.uk
[20] https://pdfbox.apache.org/
[21] https://github.com/angelobo/SemPubEvaluator

Table 2. Precision, recall and F-score for each Task 2 solution (2016 and 2015, where available).

	Authors	Precision	Recall	F-score	Precision 2015	Recall 2015	F-score 2015
#1	Ahmad et al.	0.775	0.778	0.771	–	–	–
#4	Sateli et al.	0.640	0.629	0.632	0.3	0.252	0.247
#2	Klampfl et al.	0.593	0.606	0.592	0.388	0.285	0.292
#3	Nuzzolese et al.	0.412	0.43	0.416	0.274	0.251	0.257
#5	Ramesh et al.	0.393	0.428	0.389	–	–	–

participation and the quality of the output encourage us to organise further editions in the future. Note in fact that all solutions which participated in the 2015 edition showed significant improvement in respect to their precision, recall and F-score, while new solutions proposed equally competitive approaches.

There is still room for improvements though. In fact, this year we also took the opportunity to review our experience in organising the challenge and we investigated in more detail both the overall organisation (tasks, datasets, evaluation procedures, etc.) and the results produced by the participants (approaches, tools, adopted vocabularies, etc.). More details can be found in [18].

Our conclusion is that challenges are very good enablers for producing Linked Data and helping the community to refine practices, datasets and tools.

Acknowledgements

Part of this research has been funded by the European Union under grant agreement no. 643410 (OpenAIRE2020).

References

1. R. Ahmad, M. T. Afzal, and M. A. Qadir. Information Extraction for PDF Sources based on Rule-based System using Integrated Formats. In Sack et al. [12]. Accepted for publication.
2. V. Bryl, A. Birukou, K. Eckert, and M. Kessler. What's in the proceedings? combining publisher's and researcher's perspectives. In A. García Castro, C. Lange, P. Lord, and R. Stevens, editors, *4th Workshop on Semantic Publishing (SePublica)*, number 1155 in CEUR Workshop Proceedings, Aachen, 2014.
3. A. Di Iorio, C. Lange, A. Dimou, and S. Vahdati. Semantic publishing challenge – assessing the quality of scientific output by information extraction and interlinking. In Gandon et al. [4], pages 65–80.
4. F. Gandon, E. Cabrio, M. Stankovic, and A. Zimmermann, editors. *Semantic Web Evaluation Challenges: Second SemWebEval Challenge at ESWC 2015, Portorož, Slovenia, May 31–June 4, 2015, Revised Selected Papers*, number 548 in Communications in Computer and Information Science, Cham, 2015. Springer International Publishing.

5. S. Klampfl and R. Kern. Machine learning techniques for automatically extracting contextual information from scientific publications. In Gandon et al. [4], pages 105–116.
6. S. Klampfl and R. Kern. Reconstructing the Logical Structure of a Scientific Publication using Machine Learning. In Sack et al. [12]. Accepted for publication.
7. C. Lange and A. Di Iorio. Semantic publishing challenge – assessing the quality of scientific output. In V. Presutti, M. Stankovic, E. Cambria, I. Cantador, A. Di Iorio, T. Di Noia, C. Lange, D. Reforgiato Recupero, and A. Tordai, editors, *Semantic Web Evaluation Challenge: SemWebEval 2014 at ESWC 2014, Anissaras, Crete, Greece, May 25–29, 2014, Revised Selected Papers*, number 457 in Communications in Computer and Information Science, pages 61–76, Cham, 2014. Springer International Publishing.
8. M. Milicka and R. Burget. Information extraction from web sources based on multi-aspect content analysis. In Gandon et al. [4].
9. A. G. Nuzzolese, S. Peroni, and D. R. Recupero. MACJa: Metadata and citations jailbreaker. In Gandon et al. [4], pages 117–128.
10. A. G. Nuzzolese, S. Peroni, and D. R. Recupero. ACM: Article Content Miner for Assessing the Quality of Scientific Output. In Sack et al. [12]. Accepted for publication.
11. S. H. Ramesh, A. Dhar, R. R. Kumar, V. Anjaly, K. Sarath, J. Pearce, and K. Sundaresan. Automatically Identify and Label Sections in Scientific Journals using Conditional Random Fields. In Sack et al. [12]. Accepted for publication.
12. H. Sack, S. Dietze, A. Tordai, and C. Lange, editors. *The Semantic Web: ESWC 2016 Challenges, Anissaras, Crete, Greece, May 29–June 2, 2016, Revised Selected Papers*, Communications in Computer and Information Science, Cham, 2016. Springer International Publishing. Accepted for publication.
13. B. Sateli and R. Witte. Automatic construction of a semantic knowledge base from CEUR workshop proceedings. In Gandon et al. [4], pages 129–141.
14. B. Sateli and R. Witte. An Automatic Workflow for the Formalization of Scholarly Articles' Structural and Semantic Elements. In Sack et al. [12]. Accepted for publication.
15. D. Shotton. Semantic publishing: the coming revolution in scientific journal publishing. *Learned Publishing*, 22(2):85–94, 2009.
16. D. Shotton, K. Portwin, G. Klyne, and A. Miles. Adventures in semantic publishing: Exemplar semantic enhancements of a research article. *PLoS Computational Biology*, 5(4), 2009.
17. D. Tkaczyk and L. Bolikowski. Extracting contextual information from scientific literature using CERMINE system. In Gandon et al. [4].
18. S. Vahdati, A. Dimou, C. Lange, and A. Di Iorio. Semantic publishing challenge: Bootstrapping a value chain for scientific data. In A. Gonzalez-Beltran, F. Osborne, and S. Peroni, editors, *Semantics, Analytics, Visualisation: Enhancing Scholarly Data*, Lecture Notes in Computer Science, Heidelberg, 2016. Springer.