Semantic Publishing Challenge: Bootstrapping a Value Chain for Scientific Data

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

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

Abstract. The objective of the Semantic Publishing (SemPub) challenge series is to bootstrap a value chain for scientific data to enable services, such as assessing the quality of scientific output with respect to novel metrics. The key idea was to involve participants in extracting data from heterogeneous resources and producing datasets on scholarly publications, which can be exploited by the community itself. Differently from other challenges in the semantic publishing domain, whose focus is on *exploiting* semantically enriched data, SemPub focuses on *producing* Linked Open Datasets. The goal of this paper is to review both (i) the overall organization of the Challenge, and (ii) the results that the participants have produced in the first two challenges of 2014 and 2015 – in terms of data, ontological models and tools – in order to better shape future editions of the challenge, and to better serve the needs of the semantic publishing community.

1 Introduction

Semantic publishing – defined as the *use of Semantic Web technologies to make scholarly publications and data easier to discover, browse and interact with* [15] – is a lively research area in which a big number of projects and events have emerged and showcase the potential of Linked Data technology. Extracting, annotating and sharing scientific data (by which, here, we mean standalone research datasets, data inside documents, as well as metadata about datasets and documents), up to building new research on them, will lead to a *data value chain* producing value for the scientific community [10].

Bootstrapping and enabling such value chains is not easy. A solution that has proved to be successful in other communities is to run *challenges*, i.e. competitions in which participants are asked to complete tasks and have their results ranked, often in objective way, to determine the winner. Even a number of *projects* have been launched to accelerate this process, for instance LinkedUp[5]

[5] http://linkedup-project.eu/

or Apps for Europe[6]. The success of the LAK[7] or Linked Up[8] Challenges is worth mentioning here. However, these challenges focus on *exploiting* scholarly linked data for different purposes (for instance, to monitor progress) but less on actually *producing* such datasets.

To this end, we started a series of Semantic Publishing Challenges (SemPub), aiming at the production of datasets on scholarly publications. To the best of our knowledge, this was the first challenge of its kind. Now, in 2016, we are running the 3rd edition of SemPub and we believe it is good time to review the challenge and share some lessons learned with the community. On the other hand, community feedback can help us shape the future of SemPub. Continuous refinement is in fact a key aspect of our vision.

Section 2 of this paper reviews the background of SemPub, its history, structure and evaluation methods. Section 3 presents lessons learned from the challenge organization and Section 4 from the overall approaches of the submitted solutions.Section 5 concludes and provides an outlook to how future SemPub challenges will take these lessons into account.

2 History of the SemPub Challenge

We draw a brief history of the SemPub Challenge to provide the necessary background for the following discussion. More detailed reports have been published separately for the 2014 [8] and 2015 [1] challenges.

We started in 2014, reasoning about a challenge in the semantic publishing domain that could be measured in an objective way. This was difficult because of the tension between finding appealing and novel tasks and measuring them. We thus asked participants to extract data from scholarly papers and to produce an RDF dataset that could be used to answer some relevant queries: concretely, queries about the *quality* of scientific output. We were aware of other topics of interest for the community – nanopublications, research objects, etc. – but focused on papers only to bootstrap the initiative and to start collaboratively producing initial data.

We designed different tasks, sharing the same organization, rules and evaluation procedure. For each task, we published a set of queries in natural language and asked participants to translate them into SPARQL and to submit a dataset on top of which these queries would run. In line with the general rules for the new Semantic Web Evaluation Challenge track at ESWC, we also published a training dataset (TD) on which the participants could test and train their extraction tools. A few days before the submission deadline, we published an evaluation dataset (ED): the input for the final evaluation.

The evaluation consisted of comparing the output of these queries, given as CSV, against a gold standard and measuring precision and recall. All three

[6] `http://www.appsforeurope.eu/`

[7] Learning Analytics and Knowledge; see `http://meco.l3s.uni-hannover.de:9080/wp2/?page_id=18`

[8] `http://linkedup-challenge.org/`

editions used the same evaluation procedure, but the tasks were refined over time. Table 1 summarizes all tasks, their data sources and queries.

Table 1. Description, source and format of the tasks in SemPub editions (2014–2016).

	2014	2015	2016
Task1	Extracting data on workshops' quality indicators; Source: CEUR-WS.org Format: HTML+PDF	Format: HTML	Format: HTML
Task2	Extracting data on citations Source: PubMed Format: XML	Extracting data on affiliations, citations, funding Source: CEUR-WS.org Format: PDF	Extracting data on affiliations, internal structure, fundings Source: CEUR-WS.org Format: PDF
Task3	Open tasks: showcase semantic publishing applications	Interlinking Sources: CEUR-WS.org, Colinda, DBLP, Lancet, Semantic Web Dog Food (SWDF), Springer LD)	Interlinking Sources: CEUR-WS, Colinda, DBLP, Springer LD.

Two tasks have been defined at the very beginning (see [8] for full details and statistics):

- **Task 1**: participants were asked to extract information from selected CEUR-WS.org[9] workshop proceedings volumes (HTML tables of content using different levels of semantic markup, plus PDF full text) to enable the computation of indicators for the workshops' quality assessment. They were asked to answer 20 different queries.
- **Task 2**: participants were asked to extract data about citations, to enable precise assessment of linking, sharing and evaluating research through citations. The dataset included a set of XML-encoded research papers, taken from PubMedCentral and Pensoft Open Access archives, and heterogeneous in terms of internal structure, styles and numbers. Both dataset and queries were completely disjoint from Task 1.

Having called for submissions, we received feedback from the community that mere information extraction, even if motivated by quality assessment, was not the most exciting task related to the future of scholarly publishing, as it assumed a traditional publishing model. Furthermore, to address the primary target of the challenge, i.e. "publishing" rather than just "metadata extraction", we widened the scope by adding an *open task*, whose participants were asked to showcase data-driven applications that would eventually support publishing. We received a good number of submissions; winners were selected by a jury.

In 2015 we were asked to include only tasks that could be evaluated in a fully objective manner, and thus discarded the open task. We reduced the distance between Tasks 1 and Task 2 by using the same dataset for both. We transformed Task 2 into a PDF mining task and thus moved all PDF-related queries there.

[9] `http://ceur-ws.org/`

The rationale was to differentiate tasks on the basis of the competencies and tools required to solve them, but to make tasks interplay on the same dataset.

CEUR-WS.org data has become the central focus of the whole Challenge, for two reasons: on the one hand, the data provider (CEUR-WS.org) takes advantage of a broader community that builds on its data, which, before the SemPub Challenges, had not been available as linked data. On the other hand, data consumers gain the opportunity to assess the quality of scientific venues by taking a deeper look into their history, as well as the quality of the publications. While Task 1 queries remained largely stable from 2014 to 2015, the queries for Task 2 changed, mainly because the setting was completely new (PDF rather than XML sources), and we wanted to explore participants' interest and available solutions. We asked them to extract data not only on citations but also on affiliations and fundings.

In 2015 we added a new Task 3, focusing on interlinking the dataset the winners of the first Challenge had extracted from a single source to further relevant datasets. Participants had to make such links explicit and exploit them to answer comprehensive queries about events and persons. CEUR-WS.org on its own provides incomplete information about conferences and persons, which can be complemented by interlinking with other datasets to broaden the context and to allow for more reliable conclusions about the quality of scientific events and the qualification of researchers.

Continuity is the key aspect of 2016 edition. The tasks are unchanged (allowing previous participants to use and refine their tools), except for details: Task 2, in particular, is extended to structural components of papers and does not include citations anymore.

3 Lessons learned from the Challenge organization

In this section we discuss lessons learned from our experience in organizing the challenge. Our goal is to distill some generic guidelines that could be applied to similar events, starting from the identification of critical issues, errors and strengths of our initiative. The primary focus of the paper is on (even unexpected) aspects that emerged while running the challenge. This section will also present the lessons learned by looking at the solutions and data produced by the participants. We have grouped the lessons in four categories for clarity, even though there is some overlap between them.

3.1 Lessons learned on defining tasks

The definition of the tasks is the most critical part of organizing a challenge. In our case, it was difficult to define appealing tasks that bridge the gap between building up initial datasets and exploring possibilities for innovative semantic publishing. As discussed in Section 2, we refined the tasks over the years according to the participants' and organizers' feedback. Overall, we think that tasks could have been improved in some parts – and undeniably other interesting ones

could have been defined – but they were successful. There are other less evident issues which are worth discussing here.

L1.1. Continuity: allow users to re-submit the improved version of their tool over different editions. One of the goals of the first edition of the challenge was also to explore the interest of the participants. Exploiting such feedback and creating a direct link between different editions is a success key factor. In 2015, in fact, the Challenge was re-organized aiming to commit participants to re-submit overall improved versions of their first year submissions. Results were very good, as the majority of first year's participants competed for the second year too. Continuity is also a key aspect of SemPub2016, whose tasks are the same as last year's edition, allowing participants to reuse their tools to adapt to the new call after some tuning.

L1.2. Split tasks with a clear distinction of the competencies required to complete them. One of the main problems we faced was that some tasks were too difficult. In particular the Task 2 – extraction from XML and PDF – showed unexpectedly low performance. The main reason, in our opinion, is that the task was actually composed of two sub-tasks that required very different tools and technologies: some queries required participants to basically map data from XML/PDF to RDF, while the others required additional processing on the content. Some people were discouraged to participate as they only felt competitive for the one and not for the other. Our initial goal was to explore a larger amount of information and to give participants more options but, in retrospect, such heterogeneity was a limitation. A sharper distinction between tasks would have been more appropriate. In particular, it is important to separate tasks on plain data extraction from those on natural language processing and semantic analysis.

L1.3. Involve participants in advance in the task definition. Though we collected some feedback when designing the tasks, we noticed that such preliminary phase was not given enough relevance. The participants' early feedback can help to identify practical needs of researchers and to shape tasks. Talking with participants, in fact, we envisioned alternative tasks, such as finding high-profile venues for publishing a work, summarizing publications, or helping early career researchers to find relevant papers. Proposing tasks emerged from the community can be a winning incentive to participate.

3.2 Lessons learned on building input datasets

The continuity between tasks (L1.1) can be applied to the datasets as well:

L2.1. Use the same data source for multiple editions. We noticed benefits of using the same data sources across multiple editions of the Challenge. From the task 1 of the 2014 edition, in fact, we obtained an RDF dataset that served as the foundation to build the same task in 2015 and 2016. Participants were able to reuse their existing tools and to extend the previously-created knowledge-bases with limited effort. For the other tasks, which were not equally stable, we had to *rebuild the competition* every year without being able to exploit the past experience.

L2.2. Design all three tasks around the same dataset. Similarly, it is valuable to use the same dataset for multiple tasks. First of all, for the participants: they could extend their existing tools to compete for different tasks, with a quite limited effort. This also opens new perspectives for future collaboration: participants' work could be extended and integrated in a shared effort for producing useful data. It is also worth highlighting the importance of such uniformity for the organizers. It reduces the time needed to prepare and validate data, as well as the risk of errors and imperfections. Last but not least, it enables designing interconnected tasks and producing richer output.

L2.3. Provide an exhaustive description of the expected output on the training dataset. An aspect that we underestimated in the first editions of the Challenge was the description of the training dataset. While we completely listed all papers we did not provide enough information on the expected output: we went into details for the most relevant and critical examples but we did not provide the exact expected output for all papers in the training dataset. Such information should instead be provided as it impacts directly the quality of the submissions and help participants to refine their tools.

3.3 Lessons learned on evaluating results

All three editions of the Challenge shared the same evaluation procedure (see Section 2 for more details). The workflow presented some weaknesses, especially in the first two years, which we subsequently addressed. Three main guidelines can be derived from these issues.

L3.1. Consider all papers in the final evaluation. Even though we asked participants to run their tools on the whole evaluation dataset, we considered only some exemplary papers for the final evaluation. These papers have been randomly selected from clusters representing different cases, which participants were required to address. Since these papers were representative of these cases we received a fair indication of the capabilities of each tool. On the other hand, some participants were penalized as their tool could have worked well on other values, which were not taken into account for the evaluation. In the third edition, we will radically increase the coverage of the evaluation queries and their number in order to assure that greatest part of the dataset (or the whole dataset) is covered.

L3.2. Make evaluation tool available during the training phase. The evaluation was totally transparent and all participants received detailed feedback about their scores, together with links to the open source tool used for the final evaluation. However we were able to release the tool only after the Challenge. It is instead more helpful to make it available during the training phase, as participants can refine their tool and improve the overall quality of the output. Such an approach reduces the (negative) impact of output imperfections. Though the content under evaluation was normalized and minor differences were not considered as errors, some imperfections were not expected and were not handled in advance. Some participants, for instance, produced CSV files with columns in a different order or with minor differences in the IRI structure. These all could

have been avoided if participants received feedback during the training phase, with the evaluation tool available as a downloadable stand-alone application or as a service.

L3.3. Use disjoint training and evaluation datasets. A 2015 participant raised the issue that we underestimated when designing the evaluation process: the evaluation dataset was a superset of the training one. This resulted in some over-training of the tools, and caused imbalance in the evaluation. It is more appropriate to use completely disjoint datasets, a solution we are implementing for the last edition.

3.4 Lessons learned on expected output and organizational aspects

Further suggestions can also be derived from the Challenge's organizational aspects, in particular regarding the expected outcome:

L4.1. Define clearly the license of produced output. Some attention should be given to the licensing of the output produced by the participants. We did not explicitly say which license they should use: we just required them to use an open license on data (at least permissive as the source of data) and we encouraged open-source licenses on the tools (but not mandatory). Most of the participants did not declare which exact license applies to their data. This is an obstacle for the reusability: especially when data come from heterogeneous sources and are heterogeneous in content and format, as in the case of CEUR-WS papers, it is very important to provide an explicit representation of the licensing information.

L4.2. Define clearly how the output of the challenge will be used. The previous observation can be generalized into a wider guideline about reusability. It is in fact critical to state how the results of the challenge will be eventually used, in order to encourage and motivate participants. The basic idea of the Challenge was to identify the best performing tool on a limited number of papers and to use the winning tool – or a refined version – to extract the same data on the whole CEUR-WS corpus [10]. The production of the CEUR-WS Linked Open Dataset was actually slower than expected and we are finalizing it in these days. This is a critical issue: participants' work should not target the challenge only, but it should produce an output that is directly reusable by the community.

L4.3. Study conflicts and synergies with other events. The last guideline is not surprising and was confirmed by our experience as well. In 2015, in fact, we introduced a task on interlinking. The community has been studying interlinking for many years and a lot of research groups could have participated in the task (and produced very good results). However we did not receive enough submissions. One of the issues – not the only one, communication might be another – is the conflict with events like OAEI (Ontology Alignment Evaluation Initiative). Even though Task 3 of SemPub2015 did not intend to cover the specialized scope of OAEI, but rather put the interlinking task in a certain use case

[10] At least, on the subset of CEUR-WS.org whose license scheme allowed us to republish metadata.

scope that merely serves in aligning the tasks output among each other and with the rest LOD cloud. The study of overlapping and similar events should always be kept in mind. Not only to identify potential conflicts but also to generate interest: the fact that the SePublica workshop was at ESWC 2014, for instance, was positive since we had fruitful discussions with the participants and the two events could benefit each other.

4 Lessons learned from submitted solutions

In this section we discuss lessons learned from the participants' solution. We start with an overview of the solutions; next, we group the lessons into four categories: lessons on submitted tools, used ontologies, submitted data and evaluation process; even though there is some overlap between these aspects.

4.1 Solutions by Task

Task 1 solutions – 2015 and 2014 There were four different solutions proposed to address Task 1 in 2014 and 2015. Three participated in both editions, whereas the fourth solution participated only in the second edition.

Solution 1.1. [5] [6] presented a case-specific crawling based approach for addressing Task 1. It relies on an extensible template-dependent crawler that uses sets of special predefined templates based on XPath and regular expressions to extract the content from HTML and to convert it in RDF. The RDF is then processed to merge resources using fuzzy-matching. The use of the crawler turns the system tolerant to invalid HTML pages. This solution improved its precision in 2015 as well the richness of the data model.

Solution 1.2. [3] [2] exploited a generic tool for generating RDF data from heterogeneous data. It uses the RDF Mapping Language (RML)[11] to define how data extracted from CEUR-WS Web pages should be semantically annotated. RML extends R2RML to express mapping rules from heterogeneous data to RDF. CSS3 selectors are considered to extract the data from the HTML pages. The RML mapping rules are parsed and executed by the RML Processor[12]. In 2015 the solution reconsidered its data model and was extended to validate both the mapping documents and the final RDF, resulting in an overall improved quality dataset.

Solution 1.3. [12] [13] designed a case-specific solution based on a linguistic and structural analyzer. It uses a pipeline based on the GATE Text Engineering Framework. To produce annotations, it relies on chunk-based and sentence-based support vector machine (SVM) classifiers which are trained using the CEUR-WS proceedings with microformat annotations. The annotation sanitizer has a set of heuristics which are applied to fix imperfections and interlink annotations. The produced dataset is also extended with information retrieved from external resources.

[11] `http://rml.io`

[12] `https://github.com/RMLio/RML-Mapper`

Solution 1.4. [9] presented an application of the FITLayout framework [13] This solution participated in the Semantic Publishing Challenge only in 2015. It combines different page analysis methods, i.e. layout analysis and visual and textual feature classification to analyze the rendered pages, rather than their code. The solution is quite generic but needs to be domain/case-specific at certain phases (model building step).

All solutions are summarised in Table 2, which also add details about the languages and technologies exploited by the participants.

Table 2. HTML-code-based and content-based solutions for Task 1.

	Solution 1	**Solution 2**	**Solution 3**	**Solution 4**
Primary Method	Crawling	Generic solution for abstracted mappings	Linguistic and structural analysis	Visual layout Multi-aspect content analysis
Case-specific	YES	NO	YES	NO
Template-based	YES	YES	NO	~NO
Implementation basis	–	RML tools	–	FITLayout
Implementation Language	Python	Java	–	Java/HTML
Mappings/Rules	XPath (embedded in the code)	–	RML/CSS (abstracted from the code)	Hard coded
RegEx	YES	YES	–	YES

Task 2 solutions – 2015 Solution 2.1. CERMINE [16] is an open source system for extracting structured metadata and references from scientific publications published as PDF files. It has a loosely captured architecture and a modular workflow which is based on supervised and unsupervised machine-learning techniques which simplifies the system's adoption to new document layouts and styles.

Solution 2.2. [4] implemented a processing pipeline that analyzes the structure of a PDF document incorporating a diverse set of machine learning techniques, unsupervised to extract text blocks and supervised to classify blocks into different meta-data categories. Heuristic are applied to detect the reference section and sequence classification to categorize the tokens of individual references strings. Finally, named entity recognition (NER) are used to extract references to grants, funding agencies and EU projects.

Solution 2.3. [11] presented Metadata And Citations Jailbreaker (MACJa – IPA), a tool that integrates hybrid techniques based on Natural Language Processing (NLP) and incorporating FRED, a novel machine reader. It also includes modules to query external services to enhance and validate data.

[13] `http://www.fit.vutbr.cz/\char`~burgetr/FITLayout/`

Solution 2.4. [14] presented a system composed by two modules: a text mining pipeline based on the GATE framework to extract structural and semantic entities, leveraging also existing NER tools, and a LOD exporter, to translate the document annotations into RDF according to custom rules.

Solution 2.5. [7] relies on a rule-based and pattern matching approach, implemented in Python and some external services for improving the quality of the results (for instance, DBLP for validating author's data). It also relies on an external tool to extract the plain text from PDFs.

Solution 2.6. [12] extended their framework used for Task 1 (and indicated as Solution 1.3 before) to extract data from PDF as well. Their pipeline includes text processing and entity recognition modules and employs external services for mining PDF articles. Table 3 represents tools and its components:

Table 3. Tools and its components.

	Solution 2.1	**Solution 2.2**	**Solution 2.3**	**Solution 2.4**	**Solution 2.5**	**Solution 2.6**
Implementation language	Java/HTML	Java	Java/Python	Java	Python/HTML	Java
Implementation on basis	CERMINE[14]	code-annotator[15]	–	–	–	–
Components	LibSVM[16], GRMM[17], Mallet[18]	OpenNLP[19], GATE[20], ParsCit[21], crfsuite[22]	FRED[23] CrossRef API[24], FreeCite[25], Stanford CoreNLP[26], NLTK[27], (WordNet[28], BabelNet[29])	GATE[30]	Grab spider[31], Beautiful-Soup[32]	GATE[33], Poppler[34], CrossRef API[35], FreeCite[36], Bibsonomy API[37]
PDF/character extraction	iText[38]	Apache PDFBox[39]	PDFMiner[40]	Xpdf[41]	PDFMiner	PDFX[42]
Open Source	NO	YES/broken	NO	NO	YES	NO
Intermediate representation	XML/TrueViz, NLM JATS/XML	–	JSON	–	TXT, HTML	–
Ontologies reused[43]	–	dul, dbpedia, schema	SPAR ontologies	http://lod.semanticsoftware.info/pro/, deo, sro	bibo, foaf, dc, swrc, dbpedia, arpfo	prism, pro, fabio, pro

4.2 Lessons learned from the tools

L5.1. There are both generic and ad hoc solutions. All solutions were methodologically different among each other. For Task 1, for instance, two solutions (1.1 and 1.3) primarily consisted of a tool developed specific to this task, whereas the other two solutions wrote task-specific templates in the otherwise

generic implementations (adaptive to other domains). In the later case, Solution 1.2 abstracted the case-specific aspects from the implementation, whereas Solution 1.4 kept them inline with the implementation. It becomes, therefore, clear that there are alternative approaches which can be used to produce RDF datasets.

L5.2. There are HTML code and content-based approaches to information extraction. Even though solutions were methodologically different, two main approaches for dealing with the HTML pages prevailed: HTML-code-based and content-based.

4.3 Lessons learned from models and ontologies

L6.1. All solutions used almost the same data model (Task 1). All solutions of Task 1 tend to converge regarding the model of the data. The same occurs but on a higher level in the case of Task 2. In particular for Task 1, Solution 1.4 domain modeling was inspired by the model used in Solution 1.2, with some simplifications. Note also that Solution 1.2 was the winner solution in 2014. Based on the aforementioned, we observe a trend of converging regarding the model the CEUR-WS data set should have, as most of the solutions converge on the main identified concepts in the data (Conference, Workshop, Proceedings, Paper and Person).

L6.2. All solutions used almost the same vocabularies for the same data (Task 1). There is a wide range of vocabularies and ontologies that can be used to annotate scholarly data. However, most of the solutions preferred to (re)use almost the same existing ontologies and vocabularies (see Table 4 for Task 1). This is a good evidence that the spirit of vocabulary reuse gains traction. However, it is interesting that different solutions used the same ontologies to annotate the same data differently (see L6.3).

L6.3. Different solutions used different annotations (Task 1). Even though all solutions used almost the same vocabularies, not all solutions used the same classes to annotate same entities. To be more precise, all solutions only converged on annotating persons using *foaf:Person*. For the other main concepts the situation was heterogeneous, as reported in Table 5.

L6.4. Different solutions used different vocabularies for the same data (Task 2). In contrast to Task 1 solutions which all intuitively converged on using the same vocabularies and ontologies, Task 2 solutions use relatively different vocabularies and ontologies, but again pre-existing ones, to annotate same entities appearing in the same data. However, most of the Task 2 solutions use sub-ontologies of the family of SPAR ontologies. It is interesting to observe if the Task 2 solutions of 2016 will converge towards using same ontologies, being inspired one from the other, or if solutions will keep using different vocabularies.

4.4 Lessons learned from submitted RDF

L7.1. Overall dataset improved over successive challenges.

Table 4. Vocabularies for the same data for Task 1.

Vocabulary[45]	Solution 1	Solution 2	Solution 3	Solution 4
bibo			–	
biro	–	–		–
co	–	–		–
dbpedia		Java		
dc		Java		
dcterms			–	
event	–		–	–
fabio	–			–
foaf				
frbr	–	–		–
pro	–	–		–
skos		–	–	–
swc		–	–	
swrc				
timeline		–	–	
others/custom	–	–		

Table 5. Vocabularies for different annotations for Task 1.

Task 1 / 2015	Solution 1	Solution 2	Solution 3	Solution 4
Person	foaf:Person	foaf:Person	foaf:Person	foaf:Person
Paper	bibo:Article	swrc:InProceedings	swrc:Publication	swc:Paper
Conference	swc:OrganizedEvent	bibo:Conference	swrc:Conference	swc:ConferenceEvent
Proceeding	bibo:Proceeding	bibo:Proceedings	swrc:Proceedings	swc:Proceedings
Proceeding	bibo:Workshop	bibo:Workshop	swrc:Workshop	swc:section

From the first edition to the second edition of the Semantic Publishing Challenge, we expected that participants who re-submit their solutions would have improve the overall dataset, rather than optimize it for answering the queries. All three solutions of Task 1 both in 2014 and 2015 edition modified the way they represented their model in 2014 for their submissions in 2015 which resulted in corresponding improvements to the overall dataset. Although this happens to a certain extend and indeed the results were more satisfying, we still see that there is room for overall improvement.

L7.2. Participants preferred custom solutions. Custom solutions for a particular task, such as publishing CEUR-WS.org proceedings, may obviously result in more accurate output in terms of answers to queries, however they lack repurposeability, as they cannot be reused for other input data. Moreover, despite the fact that there are generic tools for extracting RDF datasets, challenge participants preferred to develop custom solutions. This can be interpreted as a lack of alternatives of HTML specific tools to address the task.

L7.3. Striking differences in coverage. We further observed that solutions rarely agree upon the extracted information. Overall, we observe significant differences in respect to the number of identified entities per category. The results for Task 1 are summarized in the Table 6.

Table 6. The number of instances produced for each class (for Task 1).

>–	**Solution 1.1**	**Solution 1.2**	**Solution 1.3**	**Solution 1.4**
>**#Conferences**	46	46	51	47
>**#Workshops**	252	1393	127	198
>**#Proceedings**	243	1392	202	1353
>**#Papers**	3801	2452	720	2470
>**#Persons**	6700	6414	3402	11034

Let us consider the proceedings for example. Apparently, Solution 1.1 and Solution 1.3 used the individual pages to identify the proceedings, whereas Solution 2 and Solution 4 used the index page to identify the proceedings, this is the reason that there is so big difference in the numbers. The number of identified papers is also significantly different among the different solutions, but in the case of persons we observe the most variation in terms of numbers. However, the more the solutions improve, the more we expect to find solutions that converge at least regarding the number of retrieved and/or distinctly identified entities.

L7.4. RDF datasets differed significantly w.r.t. statistics. Produced datasets were also very heterogeneous in term of size, number of triples, entities and so on. Table 7 summarizes the statistics for Task 1.

Table 7. Statistics about the produced dataset (for Task 1).

>–	**Solution 1.1**	**Solution 1.2**	**Solution 1.3**	**Solution 1.4**
>**dataset size**	9.6M	6.6M	3.8M	5.1M
>**#triples**	177,752	95,015	62,231	79,444
>**#entities**	11,208	11,719	11,589	19,090
>**#properties**	46	23	42	23
>**#classes**	10	10	10	6

Note that the size of the largest dataset is almost double the size (9.6M) of the smallest (5.1M). Similarly, the largest dataset in terms of triples (~180,000), contains three times more triples compared to the smallest (~63,000) to model the same data set. Solution 4 is the only one which required significant larger number of entities to represent the same data. Considering that Solution 4 presents

a very large number of persons, the correspondingly high number of entities is not so surprising.

L7.5. No provenance or other metadata. Unfortunately, no team intuitively provided any provenance or other metadata information. In particular licensing metadata information is of crucial importance for subsequent use of datasets.

4.5 Lessons learned from the solutions with respect to the evaluation

L8.1. Performance ranking of the tools evolved but not as expected. In 2015 the performance ranking of the three tools evolved from 2014 has not changed but their performance has improved except for Kolchin et al., who improved precision but not recall. Disregarding the two queries that were new in 2015, the tool by Kolchin et al., which had won the best performance award in 2014, performs almost as well as Milicka's/Burget's.

L8.2. New and legacy solutions were both valuable. Task 1 participants both in 2014 and 2015 and they all had an improved version of different aspects of their solution which resulted in correspondingly improved versions of the final dataset. The new solution which introduced a fundamentally new approach and participated in Task 1 achieved equally good results as the best solution of 2014. In conclusion, legacy solutions might be able to improve and bring stable and good results, however there is still room for improvement and mainly for fundamentally new ideas that surpass problems that legacy solutions can not deal with.

L8.3. Newly introduced approaches have equal chances in winning the challenge. The winners of Task1 in 2014 participated in 2015 with an improved version of their tool but they did not win. The 2015 winner was a new tool with a brand new approach. The winners were not the same in the two versions of the challenge, creativity won.

5 Conclusions

One of the objectives of the SemPub Challenge series is to produce Linked Data that contribute to improving the scholarly communication. This includes supporting researchers finding relevant and high-quality papers by exploiting the information available in these datasets. Semantic Web technologies in this context do not only solve isolated problems, but generates further value in that data can be shared, linked to each other, and reasoned on.

The goal of this work is to shade light on the first editions of the Challenge and to distill some lessons learned from our experience. In particular, we were interested in both organizational aspects and evidences from the solutions proposed by the participants. Our conclusion is that we are moving in the right direction but the goal has not been fully reached yet. There are several positive aspects, among which the high participation and the quality of the produced

results. The possibility of sharing knowledge and solutions among participants was another key factor of SemPub. The Challenge allowed us to share experience on semantifying scholarly data, using some ontological models, refining and extending existing datasets. On the other hand, our analysis showed that some other aspects have to be necessarily improved. In particular, we have to make the produced output well integrated in the Linked Open Data ecosystem and exploited by the community.

The next step in fact is to investigate what are the services that we can build on top of the produced data and how they can be offered. Some natural (and challenging) questions arise: what services can already be delivered based on the data we currently have? How do we need to extend these data to provide novel services? What would be the interface of such services look like? Which functionalities should be implemented first? The challenge will also be to turn all these questions into new material for a new Challenge, even better if measurable in an objective way.

Acknowledgments: We would like to thank our peer reviewers for their constructive feedback. This work has been partially funded by the European Commission under grant agreement no. 643410.

References

1. Di Iorio, A., Lange, C., Dimou, A., & Vahdati, S. Semantic Publishing Challenge – Assessing the Quality of Scientific Output by Information Extraction and Interlinking. SemWebEval. CCIS 548. Springer, 2015.
2. Dimou, A., Vander Sande, M., Colpaert, P., De Vocht, L., Verborgh, R., Mannens, E., Van de Walle, R.: Extraction and semantic annotation of workshop proceedings in HTML using RML. SemWebEval. CCIS 475. Springer 2014
3. Heyvaert, P., et al.: Semantically annotating CEUR-WS workshop proceedings with RML. SemWebEval. CCIS 548. Springer, 2015
4. Klampfl, S., Kern, R. Machine Learning Techniques for Automatically Extracting Contextual Information from Scientific Publications. SemWebEval. CCIS 548. Springer, 2015.
5. Kolchin, M., et al.: CEUR-WS-LOD: conversion of CEUR-WS workshops to linked data. SemWebEval. CCIS 548. Springer, 2015
6. Kolchin, M., Kozlov, F.: A template-based information extraction from web sites with unstable markup. SemWebEval. CCIS 475. Springer 2014
7. Kovriguina, L. et al. Metadata Extraction From Conference Proceedings Using Template-Based Approach. SemWebEval. CCIS 548. Springer, 2015.
8. Lange, C., & Di Iorio, A., Semantic publishing challenge – assessing the quality of scientific output. SemWebEval. CCIS 475. Springer, 2014.
9. Milicka, M., Burget, R.: Information extraction from web sources based on multi-aspect content analysis. SemWebEval. CCIS 548. Springer, 2015.
10. Miller, G. & Mork, P., (2013). From Data to Decisions: A Value Chain for Big Data, Spart IT, ITPro.
11. Nuzzolese, A. G., Peroni, S., Reforgiato Recupero, D. MACJa: Metadata And Citations Jailbreaker. SemWebEval. CCIS 548. Springer, 2015.

12. Ronzano, F. et al. On the automated generation of scholarly publishing Linked Datasets: the case of CEUR-WS Proceedings. SemWebEval. CCIS 548. Springer, 2015.
13. Ronzano, F., Casamayor Del Bosque, G., Saggion, H. Semantify CEUR-WS Proceedings: towards the automatic generation of highly descriptive scholarly publishing Linked Datasets. SemWebEval. CCIS 475. Springer, 2014.
14. Sateli, B., Witte, R. Automatic Construction of a Semantic Knowledge Base from CEUR Workshop Proceedings. SemWebEval. CCIS 548. Springer, 2015.
15. Shotton, D. (2013). Publishing: Open citations. Nature, 502(7471).
16. Tkaczyk, D., Bolikowski, L. Extracting contextual information from scientific literature using CERMINE system. SemWebEval. CCIS 548. Springer, 2015.