SAVE-SD 2016 Workshop

Semantics, Analytics, Visualisation: Enhancing Scholarly Data

Alejandra Gonzalez-Beltran[1], Francesco Osborne[2], and Silvio Peroni[3]

[1] Oxford e-Research Centre, University of Oxford, Oxford, UK
[2] Data Science Group, Knowledge Media Institute,
The Open University, Milton Keynes, UK
[3] Digital And Semantic Publishing Laboratory,
Department of Computer Science and Engineering,
University of Bologna, Bologna, Italy
alejandra.gonzalezbeltran@oerc.ox.ac.uk, francesco.osborne@open.ac.uk,
silvio.peroni@unibo.it

Abstract. The SAVE-SD 2016 workshop took place in Montreal, Canada, on 11th April 2016, co-located with the World Wide Web conference. Here, we provide an overview of the SAVE-SD workshop series and 2016 workshop edition, as well as introduce the manuscripts that were accepted as part of the proceedings.
RASH:
https://w3id.org/people/essepuntato/paper/pref-savesd2016.html

Keywords: SAVE-SD 2016, WWW 2016, analytics, scholarly data, semantics, visualisation

1 SAVE-SD Workshops

Supporting new forms of scholarly data publication and analysis is a crucial task for researchers, publishers and companies working in the innovation space. Currently, most research papers are poorly annotated and published as PDF, which makes it hard to extract information from the full text. This hinders discoverability, reproducibility, and reuse of research data and findings. In addition, the content of most papers is only described with simple metadata provided as a set of keywords or topic categories. A more structured and semantic rich representation of the research outcomes could bring significant advantages to various areas: linking more effectively research and industry, supporting researchers' work, fostering cross-pollination of ideas and methods across different areas, driving research policies, and acting as a source of information for a variety of applications.

This topic has attracted high priority attention from both the industrial and the academic worlds. Funding agencies and publishers are now supporting open and accessible data publication – in fact it is one of the major themes in the EU

Horizon2020 program[4]. International community forums are also working in this direction. FORCE11[5], a community of scholars, librarians, archivists, publishers and research funders, have the aim of facilitating knowledge creation and sharing. Similarly, the Research Data Alliance (RDA)[6], a community including to 4300 members from 111 countries (as of September 2016)[7], is working since 2013 on building the social and technical infrastructure to enable open data sharing.

At the same time, companies are becoming increasingly active in providing novel and more efficient ways to share and analyse research knowledge. Journals such as Springer Nature Scientific Data[8] and Biomed Central Gigascience[9], among others, offer incentives for data publication and sharing. Thomson Reuters[10] and Elsevier[11] offer access to large datasets of scholarly data as a service to university and companies. Google Scholar[12] allows users to browse the large repository of paper indexed by Google. Microsoft Academic Search[13] offers a system for browsing research data and the Microsoft Academic Search Graph, a structured dataset containing metadata on research publications, authors, venues, organizations, and topics. Repositories supporting data publication and preservation (e.g., Zenodo[14], Dryad[15] and Figshare[16]) allow to make publicly available both documents and datasets in a citable, shareable and discoverable manner. Research social networks (ResearchGate[17], Academia.edu[18]) allow research to share and discuss they work with colleagues from all over the word. Altmetrics[19] and ImpactStory[20] offer a service based on the analysis of social network for computing alternative metrics with the aim of assessing academic performance. Finally, a number of companies in the field of innovation brokering and "horizon scanning" (Idex Labs[21], Linknovate[22]) constantly analyse the research landscape for finding relevant expert and informing the strategies of client companies. Hence, the interest from the business world presents an unprecedented opportunity for rapidly transforming academic knowledge into

[4] http://ec.europa.eu/research/press/2016/pdf/opendata-infographic_072016.pdf
[5] http://www.force11.org/
[6] https://www.rd-alliance.org/
[7] https://rd-alliance.org/node/51727
[8] http://nature.com/sdata/
[9] http://gigascience.biomedcentral.com/
[10] http://wokinfo.com/
[11] http://www.scopus.com/
[12] http://scholar.google.com/
[13] http://academic.research.microsoft.com/
[14] https://zenodo.org/
[15] https://datadryad.org/
[16] http://figshare.com/
[17] http://researchgate.net/
[18] https://www.academia.edu/
[19] http://altmetrics.org/
[20] https://impactstory.org/
[21] https://www.idexx.com/
[22] http://www.linknovate.com/

practice and achieving the data-driven science and innovation that is promised by the Big Data era.

With respect to academia, several conferences related to the World Wide Web and the Semantic Web offered a number of workshops on related topics – such as:

1. Sepublica 2011-2016 on semantic publishing at European Semantic Web Conference,
2. BigScholar 2014-2016 at the International World Wide Web Conference (on exploration and management of the Web of Scholars),
3. Linked Science 2011-2016 at the International Semantic Web Conference (on the use of Semantic Web technologies for integrating scientific data) and
4. the previous edition of this workshop, SAVE-SD 2015 at the International World Wide Web Conference.

We have also observed active participation in challenges such as the ESWC Semantic Publishing Challenge 2015[23] and 2016[24].

However, despite the rapid developments in this area, there is still a need for further dialogue between academia and industry, as well as other stakeholders working towards the vision of enhanced research data. In particular, the exchange of knowledge between the communities of *scholarlydata representation*, *research dataanalytics* and *human computer interaction* is still lacking. Indeed, research data need to be first annotated and enhanced semantically, then analysed, indexed, classified and enriched, and finally the resulting structured information should be conveyed to different kind of stakeholders in a user friendly and intuitive manner. Starting a dialogue between the experts in areas such as Knowledge Engineering, Semantic Web, Natural Language Processing (NLP), Scholarly Communication, Bibliometrics, Human-Computer Interaction, Information Visualisation is thus vital for realizing a comprehensive workflow for sharing scientific knowledge.

The aim of the SAVE-SD workshops is to offer a forum to bring together researchers, publishers and other companies, to discuss the present scenarios concerning the production and use of scholarly data, and to strategise future research and industrial directions. We believe that the combination of different expertise and perspective could be a fertile ground for the creation of innovative and scalable solutions for sharing, reusing and processing research knowledge.

In particular, The SAVE-SD Workshops focus on the following topics:

1. *semantics* of scholarly data, i.e. how to categorise, connect, integrate and represent scholarly data and its provenance information semantically, in order to foster data sharing, interoperability, reusability and reproducibility;
2. *analytics* on scholarly data, i.e. designing and implementing novel and scalable algorithms for knowledge extraction with the aim of understanding research dynamics, forecasting research trends, fostering connections between

[23] `https://github.com/ceurws/lod/wiki/SemPub2015`
[24] `https://github.com/ceurws/lod/wiki/SemPub2016`

groups of researchers, informing research policies, analysing and interlinking experiments and deriving new knowledge;
3. *visualisation* of and interaction with scholarly data, i.e. providing novel user interfaces and applications for navigating and making sense of scholarly data and highlighting their patterns and peculiarities.

This article introduces SAVE-SD 2016 proceedings, which corresponds to the second edition of the workshop. A selection of the papers from the first edition was published in the PeerJ Computer Science Journal[25].

2 SAVE-SD 2016: the second edition

As the workshop co-chairs, we now give an overview of the 2016 edition of SAVE-SD.

The SAVE-SD 2016 workshop [4] took place in Montreal, Canada, on 11th April 2016, co-located with the World Wide Web conference (WWW 2016)[26]. It was attended by about 50 people. The workshop received a total of 16 submissions from authors of 15 countries in three continents: Europe, Asia, Americas. Table 1 shows the number of papers submitted, accepted and the acceptance rate per paper type. The reason why the number of accepted poster and demo papers are higher than the submitted ones is that a number of rejected full papers were actually accepted as poster and demo papers.

Table 1. Statistics of SAVE-SD 2016 submitted/accepted papers.

Type	Number Submitted	Number Accepted	Acceptance Rate
Full papers	11	6	54.5%
Position papers	1	2	100.0%
Poster/Demo papers	4	6	100.0%

The workshop opened by the keynote of Alex Wade, Director of Scholarly Communications at Microsoft Research, who currently works on Microsoft Academic. In particular, the talk presented the novel Microsoft Academic Graph[27], a novel entity graph of research publications, authors, venues, organizations, and topics which are now driving new features in Bing, Cortana, and Microsoft Academic.

2.1 Programme Committees

As SAVE-SD aims to address the gap between the theoretical/academic and practical/industrial aspects of scholarly data, the review process ought to con-

[25] https://peerj.com/collections/24-save-sd-2015/
[26] http://www2016.ca/
[27] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

sider both perspectives. Thus, SAVE-SD has three different programme committees (PCs), whose members are listed below:

1. an Industrial PC, who mainly evaluates the submissions from an industrial perspective – by assessing how much the theories/applications described in the papers do/may influence (positively or negatively) the publishing and technological domain and whether they could be concretely adopted by publishers and scholarly data providers;
2. an Academic PC, who evaluate the papers mainly from an academic perspective – by assessing the quality of the research described in such papers.
3. a Senior PC, whose members act as meta-reviewers and have the crucial role of balancing the scores provided by the reviews from the other two PCs;

We, the workshop chairs, are very grateful to all the members of the three PCs for their high quality reviews and constructive feedback, which improved significantly the quality of the papers contained in these proceedings.

Senior Program Committee

- Aldo Gangemi (Université de Paris 13, France, and CNR, Italy)
- Daniel Schwabe (Pontifical Catholic University of Rio de Janeiro, Brazil)
- Enrico Motta (KMi, The Open University, UK)
- Fabio Vitali (University of Bologna, Italy)
- Ivan Herman (Digital Publishing Lead, W3C)
- Pascal Hitzler (Wright State University, USA)
- Simone Teufel (University of Cambridge, UK)
- Susanna Assunta Sansone (University of Oxford and NPG, UK)
- Timothy W. Clark (Harvard University, USA)

Industrial Program Committee

- Alex Wade (Microsoft Research)
- Aliaksandr Birukou (Springer)
- Anita de Waard (Elsevier)
- Anna Tordai (Elsevier)
- Christoph Lange (CEUR-WS.org)
- Eric Prud'hommeaux (W3C)
- Kris Jack (Mendeley)
- Laurel L. Haak (ORCID)
- Lyubomir Penev (Pensoft Publishers)
- Maarten Fröhlich (IOS Press)
- Michele Pasin (Springer Nature)
- Patricia Feeney (CrossRef)
- Paul Groth (Elsevier Labs, The Netherlands)
- Petr Knoth (Mendeley)
- Scott Edmunds (GigaScience and BioMed Central)
- Thomas Ingraham (F1000Research)

Academic Program Committee

- Alexander García Castro (Florida State University, USA)
- Andrea Bonaccorsi (University of Pisa, Italy)
- Andrea Giovanni Nuzzolese (ISTC-CNR Rome, Italy)
- Angelo Di Iorio (University of Bologna, Italy)
- Anna Lisa Gentile (University of Mannheim, Germany)
- Asunción Gómez Pérez (UPM, Spain)
- Bahar Sateli (Concordia University, Canada)
- Daniel Garijo (UPM, Spain)
- Davide Buscaldi (Université de Paris 13, France)
- Eamonn Maguire (CERN, Switzerland)
- Francesco Poggi (University of Bologna, Italy)
- Francesco Ronzano (Universitat Pompeu Fabra, Spain)
- Ilaria Tiddi (KMi, The Open University, UK)
- Jodi Schneider (University of Pittsburgh, USA)
- Leyla Jael García Castro (University of Munich, Germany)
- Mathieu d'Aquin (KMi, The Open University, UK)
- Oscar Corcho (UPM, Spain)
- Paolo Ciancarini (University of Bologna, Italy)
- Paolo Ciccarese (Harvard University, USA)
- Philippe Rocca-Serra (University of Oxford, UK)
- Rinke Hoekstra (VU Amsterdam, The Netherlands)
- Rob Davey (Genome Analysis Centre, UK)
- Stefan Dietze (L3S Research Center, Germany)
- Steffen Lohmann (University of Stuttgart, Germany)
- Steve Pettifer (University of Manchester, UK)
- Tom Heath (Open Data Institute, UK)
- Tomi Kauppinen (Aalto University, Finland, and University of Münster, Germany)
- Tudor Groza (Garvan Institute of Medical Research, Australia)

2.2 Sponsors

We are grateful to our sponsors:

1. Springer Nature[28], who provided a 150 euros voucher to buy Springer Nature's products for the best RASH paper award;
2. Pensoft[29], who hosted a free-of-charge special collection for selected position/poster/demo papers in the Research Idea and Outcomes (RIO) Journal;
3. GigaScience[30], who provided cool "Data is coming" t-shirts for the workshops attendees.

[28] `http://www.springernature.com/`
[29] `http://pensoft.net/`
[30] `http://gigascience.biomedcentral.com/`

3 SAVE-SD advocating enhanced papers

The adoption of Web-based formats in scientific literature is an important step towards the complex and exciting vision of Semantic Publishing. The goal is to unlock the knowledge hidden in other formats. For this reason SAVE-SD is actively encouraging author to submit their research papers in HTML-based formats.

In particular, SAVE-SD explicitly offers support for submission in the Research Articles in Simplified HTML[31] (RASH) format[32]. RASH is a markup language that restricts the use of HTML to 32 elements [3]. This solution allows authors to include semantic relationships in their work either by associating RDFa annotations or by inserting plain Turtle, RDF/XML or JSON-LD content in a `script` element. To encourage submission in RASH the organisers introduced a special award for the best submission in RASH, whose criteria are indicated in the next section.

SAVE-SD 2015 was the first workshop to accept RASH papers. In the first edition it received 6 out of 23 submissions in RASH and after the review process an additional author chose to prepare the camera ready paper in RASH. Today, together with other HTML-based formats, RASH is accepted by the main Semantic Web conferences (ISWC, ESWC, EKAW) and by a number of related workshops and challenges[33].

SAVE-SD 2016 received 6 out of 16 submissions in RASH from 14 authors from Italy, Sweden, Greece, Germany, Belgium, and United States. In total, 5 out of the 14 accepted papers were in RASH, including two full papers, one position paper, and two demo papers.

4 Selection of best paper and best RASH paper

SAVE-SD 2016 awarded two prizes: one for best paper and another for best RASH paper. The latter was sponsored by Springer Nature.

The criteria to select the best paper award considered the reviewers scores and selected the paper with the best score. The best paper award was given to:

> Salatino, A., Motta, E. (2016). Detection of Embryonic Research Topics by Analysing Semantic Topic Networks [11]

The best RASH paper award is given to the paper that makes best use of the RASH format. This is chosen by an automatic score system which rates all the RASH submissions considering:

1. the quality of the markup (i.e., how many errors the document has compared with the RASH grammar),

[31] `https://github.com/essepuntato/rash`

[32] `https://github.com/essepuntato/rash`

[33] `https://github.com/essepuntato/rash/#venues-that-have-adopted-rash-as-submission-format`

2. the quality of HTML (i.e., how many errors the document has compared with HTML5),
3. the number of RDF statements defined,
4. the number of RDF links to LOD datasets.

The best RASH paper of SAVE-SD 2016 was:

Philipson, J. (2016). Citation functions for knowledge export - a question of relevance, or, can CiTO do the trick? [9]

5 Short overview of the papers

The articles of this edition focus on the extraction of semantic information from research papers and their use for characterizing citations and analysing research topics and trends. In this section, we will summarize their contents and contributions of all the papers.

We believe that this collection of papers constitutes an excellent example of current research in the field of Semantic Publishing, addressing both theoretical and practice issues. We are thus confident that these proceedings will be helpful to researchers and companies interested in this growing research area.

Although some papers contain to a degree all the components suggested by the SAVE-SD workshop, i.e., semantics, analytics and visualization, we can classify them in two main categories: the ones that address the extraction of semantic information from full-text or pre-existent datasets and the ones that focus on exploiting semantic techniques for fostering the analysis of citations, researchers and topics. The first category includes two full papers, one position paper and four poster papers.

In recent years there has been a number of efforts to extract scientific artefacts (e.g., genes [1], chemical components [2]) and epistemological concepts (e.g., hypothesis, motivation, experiments) [6] [5] from research publications. The following five papers are dedicated to this intriguing task.

The paper by Ronzano and Saggion introduces a platform to represent as RDF several aspects of scientific publications, using techniques such as rhetorical sentence classification and text summarisation. The research publications are analysed by relying on the Text Mining Framework developed in the context of the European Project Dr. Inventor [10]. In line with the SAVE-SD vision, this framework also offers a number of relevant web visualizations[34] for exploring the produced RDF dataset.

Gábor et al. propose a method for automatically extracting semantic relations from articles in the science/engineering domain. Their approach allows to identify the entities and concepts that describe a scientific field (e.g., methods, problems) and the semantic relations between them (e.g., tackle, develop). The proposed workflow combines natural language processing techniques with statistical term extractors and external ontological resources.

[34] `http://backingdata.org/dri/viz/`

Marsi and Øzturk introduce a framework for finding events in natural science literature, such as the increase/decrease of variables. The resulting knowledge base enables semantic search for events and variables, which can be used to assess possible correlations – e.g., the increase in the sea level vs the decrease of the ice sheet. The system offers also a user interface to browse the events and visualize their type, frequency and relation strength.

Similarly, Sateli and Witte's demo paper describes a workflow for converting research paper in a Linked Open Data compliant knowledge base. Their solution includes a NLP pipeline for tokenisation, sentence splitting, part-of-speech (POS) tagging, stemming, and verb group analysis; the Rhetector component for automatically detecting rhetorical entities; LODtagger for linking Dbpedia entities to the paper; and LODeXporter for generating the output RDF.

Finally, the poster paper by Alexiou et al. presents the OpenAIRE LOD services, the RDF version of the well known Open Access Infrastructure for Research in Europe dataset[35], which includes publications and datasets from more than 100,000 research projects. In particular, the poster describes the scalable workflow used for the RDFization process of such a huge database.

The complex research entities extracted by these approaches have the potential to revolutionize the way we analyse scientific literature. However, key phrases are still the most common means to represent the content of articles for the benefits of users, search engine and recommendation systems. For this reason, the paper by Daudaravicius introduces a new statistical approach for extracting key phrases from scientific journals in the fields of astrophysics, mathematics, physics, and computer science. Their method uses the additive smoothing of TF-IDF for improving the quality of key phrases derived from large sample of papers.

The task of extracting semantic information from scholarly papers was addressed since 2015 by the Semantic Publishing Challenge (SemPub) at the Extended Semantic Web Conference. SemPub created a framework for comparing in an objective way a number of systems in the semantic publishing domain and encouraged researcher to produce and make available a number of relevant Linked Data dataset. The paper of Vahdati et al. examines the overall organization of the Challenge and the results produced in the 2015 and 2016 editions. It also analyses the different system proposed for the different tasks and discusses a number of good lessons learned by the organizers.

Scholarly metadata can also be found on the web, in formats such as as RDFa[36], Microdata[37] and Microformats[38]. However, it is not always easy to recover and aggregate this data. The position paper by Sahoo et al. contributes to this challenge by presenting an analysis on Web Data Commons (WDC) dataset[39] with the aim of identifying frequent types and terms, the key providers

[35] https://www.openaire.eu/
[36] http://www.w3.org/TR/xhtml-rdfa-primer/
[37] http://www.w3.org/TR/microdata
[38] http://microformats.org/
[39] http://webdatacommons.org/

of bibliographic markup and the most common errors. The findings include the prevalence of statements describing authors, publishers and keywords and the fact that Springer.org appears to be the most active data provider by a large margin in the sample under analysis.

The second category of papers addresses the use of semantic technologies for citation and topic analysis and is composed by two full papers, a position paper and a poster paper.

The position paper by Philipson addresses the use of citation functions for promoting knowledge export and discusses the use of the Citation Typing Ontology (CiTO) [8] for this task. Indeed, while in many contexts different kinds of citations are treated as equal, they can be radically different according to their semantics and rhetorical context. The paper examines in particular cross-disciplinary citation functions, such as "comparison", "evidence", "force", "method" and "result". It concludes that currently CiTO is not specific enough to capture the subtle differences between some of citation functions and suggest that a combination of citation functions and subject headings, extracted from both citing and cited entities might offer even better prospects for knowledge export.

The rest of the papers highlight the advantage of a semantic characterization of research topics [7] for describing researcher and analysing the evolution of research trends.

Sateli et al. propose a novel method for automatically creating authors' profiles according to their topics of interest. Indeed, a number of scholarly applications build on a representation of researchers in term of their competence, for supporting services such as expert search and paper recommendation. The automatically extraction of this profile from the full-text of research papers is performed by means of a text mining pipeline, which detects relevant topics as grounded named entities from DBpedia. Interestingly, the evaluation showed that the topic extracted within specific rhetorical zones are more representative of the author's competences.

The paper by Salatino and Motta, which won the best paper award, focuses on the detection of embryonic research topics that can be used for anticipating future research trends. It theorizes that the appearance of novel research areas is anticipated by specific dynamics between existing ones and suggests a method based on the analysis of 3-cliques for detecting these dynamics. The paper presents an experiment on a sample of 3 million research papers which confirms the hypothesis. The main finding is that the pace of collaboration in the subgraphs of topics that will give rise to a new research area is significantly higher than the one in the control group. This knowledge could foster a variety of methods for trend detection which currently focus on topics already associated with a label or a substantial number of documents.

Portenoy and West poster paper addresses a similar issue, proposing a new kind of visualization for representing the evolution of a topic and its influence on other fields, according to the citations graph. Their application exploits hierarchical clustering techniques to partition the citation graph into clusters repre-

senting fields and subfields. A demo of this visualization is publicly available at `http://scholar.eigenfactor.org/fields`.

6 Journal issues for extended papers and posters

The authors of full papers were invited to submit an extended version of their work to a special issue that will be published as part of the PeerJ Computer Science. The authors of position, demo, and poster papers of the workshop were invited to submit an extended version of their works to a special issue that will be published as part of the Research Ideas and Outcomes (RIO) Journal.

The reader will be able to find further information of such extended papers at `http://cs.unibo.it/save-sd/2016/`.

References

1. Carpenter, B.: LingPipe for 99.99% recall of gene mentions. In Proceedings of the Second BioCreative Challenge Evaluation Workshop (Vol. 23, pp. 307-309). (2007)
2. Corbett, P., Copestake, A.: Cascaded classifiers for confidence-based chemical named entity recognition. BMC bioinformatics, 9(11), p.1. (2008)
3. Di Iorio, A., Nuzzolese, A.G., Osborne, F., Peroni, S., Poggi, F., Smith, M., Vitali, F. and Zhao, J.: The RASH Framework: enabling HTML+ RDF submissions in scholarly venues. In 14th International Semantic Web Conference. (2015)
4. Gonzalez-Beltran, A., Osborne, F., and Peroni, S.: SAVE-SD 2016: Second Workshop on Semantics, Analytics and Visualisation: Enhancing Scholarly Data. In Proceedings of the 25th International Conference Companion on World Wide Web (WWW '16 Companion) (pp. 1043-1044). (2016)
5. Groza, T.: Using Typed Dependencies to Study and Recognise Conceptualisation Zones in Biomedical Literature. PloS one, 8(11), p.e79570. (2013)
6. Liakata, M., Teufel, S., Siddharthan, A., Batchelor, C.R.: Corpora for the Conceptualisation and Zoning of Scientific Papers. In Proceedings of the 2010 International Conference on Language Resources and Evaluation. (2010)
7. Osborne, F., and Motta, E.: Klink-2: integrating multiple web sources to generate semantic topic networks. In International Semantic Web Conference (pp. 408-424). Springer International Publishing. (2015)
8. Peroni, S., and Shotton, D.: FaBiO and CiTO: ontologies for describing bibliographic resources and citations. Web Semantics: Science, Services and Agents on the World Wide Web, 17, pp.33-43. (2012)
9. Philipson, J.: Citation functions for knowledge export - a question of relevance, or, can CiTO do the trick? In Proceeding of the SAVE-SD 2016 Workshop. (2016)
10. Ronzano, F., and Saggion, H.: Dr. Inventor Framework: Extracting Structured Information from Scientific Publications. Discovery Science (pp. 209-220). Springer International Publishing. (2015)
11. Salatino, A. A. and Motta, E.: Detection of Embryonic Research Topics by Analysing Semantic Topic Networks. In Proceeding of the SAVE-SD 2016 Workshop. (2016)