Report on the First Workshop on Linking and Contextualizing Publications and Datasets

Paolo Manghi
ISTI, Consiglio Nazionale delle Ricerche, Italy
paolo.manghi@isti.cnr.it

Lukasz Bolikowski
ICM, University of Warsaw, Poland
l.bolikowski@icm.edu.pl

Nikos Houssos
National Hellenic Research Foundation, Greece
nhoussos@ekt.gr

Jochen Schirrwagen
Bielefeld University Library, Germany
jochen.schirrwagen@uni-bielefeld.de

1. INTRODUCTION

Contemporary scholarly communication is undergoing a paradigm shift, which in some ways echoes the one from the start of the Digital Era, when publications moved to a digital form. There are multiple reasons for this change, and three prominent ones are: (*i*) emergence of data-intensive science (Jim Gray's Fourth Paradigm), (*ii*) evolving reading patterns in modern science, and (*iii*) increasing heterogeneity of research communication practices (and technologies).

Motivated by e-Science methodologies and data-intensive science, contemporary scientists are increasingly embracing new data-centric ways of conceptualizing, organizing and carrying out their research activities. Such paradigm shift strongly affects the way scholarly communication is conducted, promoting datasets as first class citizen of the scientific dissemination. Scientific communities are eagerly investigating and devising solutions for scientists to publish their raw and secondary datasets – e.g. sensor data, tables, charts, questionnaires – to enable: (*i*) discovery and re-use of datasets and (*ii*) rewarding the scientists who produced the datasets after often meticulous and time-consuming efforts. Data publishing is still not a reality in many communities, while for others it has already solidified into procedures and policies.

Due to the ability to have immediate Web access to all published material, be them publications or datasets, scientists are today faced with a daily wave of new potentially relevant research results. Several solutions have been devised to go beyond the simple digital article and facilitate the identification of relevant and quality material. Approaches aim at enriching publications with semantic tags, quality evaluations, feedbacks, pointers to authority files (for example persistent identifiers of authors, affiliation, and funding) or links to other research material. Such trends find their motivations not only from the need of scientists to share a richer perspective of research outcome, but also from traditional and novel needs of research organisations and funding agencies to: (*i*) measure research impact in order to assess and reward their initiatives, e.g. research outcome must be linked to affiliations, authorships, and grants, and (*ii*) guarantee the results of public research is made available as interlinked and contextualized Open Access material, e.g. research datasets are interlinked to related publications and made available via online data repositories and publication repositories. The most prominent example of such requirements is provided by the European Commission with the Open Access mandates for publications and data in Horizon2020.

Finally, researchers rely on different technologies and systems to deposit and preserve their research outcome and their contextual information. Datasets and publications are kept into data centres and institutional and thematic repositories together with descriptive metadata. Contextual information is scattered into other systems, for example CRIS systems for funding schemes and affiliation, national and international initiatives and registries, such as ORCID and VIAF for authors and notable people in general. The construction of *Modern Scholarly Communication Systems* capable of collecting and assembling such information in a meaningful way has opened up several research challenges in the fields of Digital Library, e-Science, and e-Research.

Solving the above challenges would foster multidisciplinarity, generate novel research opportunities, and endorse quality research. To this aim, sectors of scholarly communication and digital libraries are investigating solutions for "interlinking" and "contextualizing" datasets and scientific publications. Such solutions span from publishing methodologies, processes, policies, to technical aspects involving data modelling, systems, architectures, and applications. The goal of the first workshop on Linking and Contextualizing Publications and Datasets (LCPD)[1] was to provide researchers and practitioners in the fields of Digital Library, e-Science, and e-Research with a forum where they can constructively explore foundational, organizational and systemic challenges in contexts having publishing, interlinking, preservation, discovery, access, and reuse of publications and datasets.

2. WORKSHOP PRESENTATIONS

All submitted contributions were peer reviewed by three of the seventeen members of the Program Committee and ten were accepted. The workshop structure comprised an invited speakers session followed by the presentation of the ten contributions. Each session is introduced as a separate subsection.

2.1 Keynotes

The workshop had two invited talks respectively covering the aspects of dataset creation and publishing and how Linked Data can be exploited as a mean to leverage scientific publishing.
Sarah Callaghan in her talk entitled "Datasets: from creation to publication" made a clear statement of how research data is becoming central to many scientific disciplines, for which repeatability, verification and transparency of experiments is the key to reward and enable good research. As a consequence, research data should be published in ways like data journals are supporting, e.g. peer review of data stored in data repositories and cited in scientific literature. Access to the data is important to understand and validate conclusions made in research papers. This requires a change in the current scholarly communication practices [4], but also in the culture of scientists, who must ensure that their research stories are transparent, their outcomes viable to sharing, in order to make their research used and trusted by others.
Sören Auer in his talk entitled "How can Linked Data facilitate scientific publishing and knowledge exchange?" suggested the possibility of integrating the Linked Data approach to traditional scientific literature by using tools that would allow researchers to embed structure and semantics to their articles in order to represent them as conceptual RDF graphs, to be then processed by applications more sophisticated than traditional viewers. Examples are document ontologies (e.g. identifying sections, paragraphs, figures, sentence), rhetorical ontologies (e.g. claim, explanation, argument), or semantic annotations (e.g. concepts, external links). Semantic annotation [5] may help finding related work, gaining reputation on social networks, improve visualization, engage researchers with games, and be implemented by researchers in a distributed fashion. In the long term, such practices would increase the number of citations and provide evidence to achieve research funding.

2.2 Papers Presentation

The presentations from the ten contributions were organized in different sessions, which covered the areas of dataset contextualisation, interlinking datasets and publications, representing and visualizing datasets, and issues regarding packaging datasets and metadata for datasets.

Dataset Contextualization. With respect to dataset contextualisation, Łukasz Bolikowski presented the paper "Tagging Scientific Publications using Wikipedia and Natural Language Processing Tools". The authors propose and evaluate the effectiveness of two methods for contextualizing scientific publications by tagging them with labels reflecting their content. Labels correspond to Wikipedia pages or to noun phrases obtained with NLP tools and can be used as new forms of document representations, for example to enable machine learning tasks, such as document similarity, clustering, topic modelling.

Next, Jon Blower presented the paper "Understanding Climate Data through Commentary Metadata: the CHARMe project". CHARMe is an EC funded project, which aims to link climate datasets with publications, user feedback and other items of commentary metadata. The resulting information system will help users learn from previous community experience and select datasets that best suit their needs, as well as providing direct traceability between conclusions and the data that supported them. Although the project focuses on climate science, the technologies and concepts are very general and could be applied to other fields.

Interlinking Publications and Datasets. Exploring the areas of interlinking research outcomes, Mark Depauw and Tom Gheldof presented the paper "Trismegistos. An interdisciplinary Platform for Ancient World Texts and Related Information". Trismegistos is a metadata platform for the study of texts (e.g. inscriptions) from the Ancient World (roughly 800 BC – AD 800) whose descriptions are kept in several databases worldwide. Its aim is to correlate, i.e. interlink, these digital descriptions to surmount

[1] *LCPD2013's web site,* `http://lcpd2013.research-infrastructures.eu`

barriers of language, discipline and geography.

Next, Paolo Manghi presented the work "Preliminary Analysis of Data Sources Interlinking - Data Searchery: a case study". Data Searchery is a lightweight configurable tool supporting researchers at real-time relating publications and/or research data across online data sources by identifying relationships between their metadata descriptions.

Finally, Nuno Lopez reported on the paper "Linked Logainm: Enhancing Library Metadata using Linked Data of Irish Place Names". Linked Logainm is a newly created Linked Data version of Logainm.ie, an online database holding the authoritative hierarchical list of Irish and English language place names in Ireland. The authors demonstrate the benefit of Linked Data to the library community using Linked Logainm to enhance the Longfield Map collection, a set of digitised 18th-19th century maps held by the National Library of Ireland.

Dataset representation and visualization. With regard to dataset representation and visualisation aspects, Marcin Skulimowski presented the paper titled "From Linked Data to Concept Networks". The author proposes the usage of concept networks for scientific publications making use of RDF links between relevant entities from publications. In particular, a web tool supporting creation of concept networks for quantum mechanics was presented.

Subsequently, András Micsik reported on the work "LODmilla: shared visualization of Linked Open Data". LODmilla is a browser embedding the basic commodity features for generic LOD browsing including views, graph manipulation, searching, etc. Users can navigate and explore multiple LOD datasets and they can also save LOD views and share them with other users.

The session concluded with Harry Dimitropoulos presenting the paper "Content visualization of scientific corpora using an extensible relational database implementation". The authors describe a method for the supervised classification and visualization of collections of scientific publications. By integrating a text classification module, which leads to class probability estimation, along with a dimensionality reduction technique, which represents each class in the 2-D space, any collection of unlabelled documents can be intelligibly visualized.

Metadata and packaging of datasets and publications. In this session Nikos Houssos, on behalf of Anna Clements, presented the work "CERIF for Datasets (C4D)". The aim of CERIF for Datasets (C4D) is to develop a framework for incorporating metadata into CERIF (the Common European Research Information Format) such that research organisations and researchers can better discover and make use of existing and future research datasets, wherever they may be held.

The session ended with Catherine Jones presenting the paper "Investigations as research objects within facilities science". The authors explore the notion of data publication in the context of large-scale scientific facilities and propose to publish an investigation, a more complete record of the experiment, including its parameters and context details. In particular they relate this investigation to the emerging concept of a research object [2].

3. WORKSHOP DISCUSSION

The concluding brainstorming session brought up the following main relevant considerations and future issues with respect to publications and datasets.

Publications: beyond traditional papers. Research is focusing on novel forms of publication (e.g. enhanced publications [1]), which are interpreting interlinking and contextualisation of publications and datasets in different, often discipline-specific, scenarios. "Modern" publishing should try to preserve the narrative spirit of literature while integrating procedural aspects of the underlying research (e.g. experiments, workflows), other material, and post-publication information (e.g. semantic tags), etc. Areas of interest identified in the discussion regard annotations, experiments, and linked open data.

Annotation of publications for human and machine interpretation: this area addresses "horizontal" alternatives to the traditional "vertical" reading of paper. Tools, models, and practices are conceived in order to describe the structure of papers (e.g. ordering and interlinking of concepts, rhetorical and reading structure), to enrich papers with semantics (e.g. LOD), and to interlink papers with other research outcome via semantic relationships. Such techniques improve publication discovery capabilities, enable navigation by concepts and improve the ability to interpret research outcomes.

Enabling experiment repeatability and reproducibility: this area focuses on the realization of new digital publications, including a narrative part describing the research and the components necessary to execute the experiment underlying such research. Several solutions exist (e.g. research objects [2]), more or less specific to a different execution environment, but a lot of work has to be done, especially on embedding the life-cycle of such publications within the practices of research infrastructures.

Linked Open Data: this area investigates the potential benefits of adopting LOD techniques in a cross-disciplinary scenario, where common dataset and publication standards cannot be adopted. While it is evident how LOD may fit with the needs of modern scholarly communication when applied within the semantic boundaries of one discipline, it is less clear on how existing experiences could help at leveraging multi-disciplinarity across several domains.

While traditionally an article was the "unit" of publication, contemporary scholarly communication is more fine-grained. Proposed solutions must tackle the heterogeneity of scientific disciplines where the notion of experiment as well as that of datasets changes considerably, e.g. different concept networks, ontologies, metadata descriptions, digital encodings of research datasets. This inherent complexity introduces discipline-oriented practices, standards, and technologies, which can sometime hardly be reused to serve different disciplines. On the other hand, it also leverages effective quality measurement and reuse of scientific results, and introduces a level of granularity which facilitates the identification of cross-links between disciplines.

Datasets: "Publishing without datasets is publishing without evidence, is not research but advertising" (Graham Steel). A crucial challenge in the years to come is addressing the cultural barriers behind data publishing and data citation. These practices are often disregarded by researchers once their results have been published. Research communities should take a rigorous approach and identify the optimal procedures to ensure publications and datasets are published, interlinked and properly described to maximise their discovery and re-usability [3]. In order to achieve real benefits, such procedures should be standardised and imposed to the scientists. Some of the aspects to be studied are: (*i*) guidance about data citation as good research practice, possibly involving publishers and data centres, (*ii*) reward and give visibility to re-use of datasets (e.g. measuring dataset downloads, surveys about the re-use of data), (*iii*) changing the scientific reward paradigm to include dataset production.

On the other hand, data publishing and citation would lose all their appeal without enabling proper data evaluation mechanisms and processes capable of supporting dataset quality control. An area worth investigations is that of annotation of datasets for human and machinery interpretation, adopting techniques which are similar to those used for publications. Datasets could be enriched during their life-cycle (from the collection of raw data to the generation of intermediate secondary datasets) by collecting provenance and lineage information or other application-specific information. Automating dataset peer-review is another challenge. In most scenarios (e.g. data papers) data to be published is quality-checked by humans who verify if datasets are well described and complete, e.g. checking usage of standard formats. This approach cannot scale for large datasets (i.e. big data) or for datasets with non-legible formats (time-series). In such scenarios, dataset quality in publishing workflows should be demanded to proper machinery, e.g. research infrastructure tools used to run scientific experiments.

4. ACKNOWLEDGMENTS

Our sincere gratitude goes to all program committee members, invited speakers, authors, and participants, whose enthusiasm and vision constituted the soul of this workshop and of the future research in the field. The event was funded by the OpenAIREplus project (grant agreement: 283595, Call: FP7-INFRA-2011-2) and the EuroCRIS Consortium.

5. REFERENCES

[1] Bardi, A., and Manghi, P. Enhanced publications: Data models and information systems. *LIBER Quarterly 23*, 4 (2014).

[2] Bechhofer, S., Roure, D. D., Gamble, M., Goble, C., and Buchan, I. Research Objects: Towards Exchange and Reuse of Digital Knowledge. In *The Future of the Web for Collaborative Science (FWCS 2010)* (February 2010). Co-located with WWW'10 Event Dates: April 2010.

[3] Callaghan, S., Murphy, F., Tedds, J., Allan, R., Kunze, J., Lawrence, R., Mayernik, M. S., and Whyte, A. Processes and procedures for data publication: A case study in the geosciences. *IJDC 8*, 1 (2013), 193–203.

[4] CODATA-ICSTI Task Group on Data Citation Standards and Practices. Out of Cite out of Mind: The Current State of Practice, Policy and Technology for the Citation of Data. *Data Science Journal 12* (2013).

[5] Ngomo, A.-C. N., and Auer, S. LIMES: a time-efficient approach for large-scale link discovery on the web of data. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three* (2011), AAAI Press, pp. 2312–2317.