Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 38 (2014) 22 – 27

Procedia
Computer Science

10th Italian Research Conference on Digital Libraries, IRCDL 2014

The Future of Digital Scholarship

Costantino Thanos*

Institute of Information Science and Technologies (ISTI), Italian National Research Council (CNR), Via G. Moruzzi 1, 56124 Pisa, Italy

Abstract

This paper advocates that connectivity is the technological foundation of digital scholarship and argues that the characteristics of modern science, i.e. data-centric, multidisciplinary, open, network-centric and heavily dependent on internet technologies entail the creation of a linked, semantically enhanced scholarly record composed of interconnected discipline-specific literature and scientific, social, and humanities data spaces. The changing scenario of the scholarly record is illustrated by describing the principal transformations now being enabled by advanced linking and semantic technologies. The main functionality of a cyberscholarship infrastructure is described, i.e. the ability to effectively and efficiently support a linking environment.

Keywords: digital scholarship; scholarly record; digital scholarship infrastructure

1. Introduction

A new paradigm is emerging in science characterized by: (1) the availability of publicly network-accessible vast volumes of curated scientific data, i.e. *data intensive science*; (2) the drawing on multiple scientific disciplines in order to find solutions to difficult problems on the basis of a new understanding of complex situations, i.e. *multidisciplinary science*; (3) the sharing of results, ideas, methods and data between scientists and the public much earlier and much more extensively than previously, i.e. *openness*; (4) the increasingly global collaborations – of scientists and of shared resources – enabled by the interconnection of the components of science ecosystems (digital libraries, data centers, institutional repositories, etc.) distributed worldwide and overcoming language, policy, social, barriers, i.e. *globalism of science*; and (5) scientific investigations enabled by computer science technologies, computing infrastructures and internet, i.e. *e-Science*.

* Corresponding author. Tel.: +39 050 621 2910; fax: +39 050 315 2810.
E-mail address: costantino.thanos@isti.cnr.it

doi:10.1016/j.procs.2014.10.005

These characteristics of modern science entail that the publication and consumption of scientific information should undergo radical changes:

- Scientific data should become a *first class citizen* of scientific communication; as such it must have its own identity and be integrated with scientific publications in order to support repeatability, reproducibility and reanalysis.
- Scientific data should be *discoverable*, i.e. scholars must be able to find data that supports scientific research quickly and accurately, *understandable* to those scrutinizing them and *assessable* enabling potential users to evaluate them.
- Scientific data and publications have to cross disciplinary boundaries; therefore, in order to maintain the interpretative context they must be semantically enhanced, i.e. semantic mark-up of textual terms with links to ontologies/terminologies/vocabularies, interactive figures, etc. Semantic services will help readers to find actionable data, interpret information and extract knowledge.
- Modern scientific communication should be characterized by modularity: it should allow for non linear and haphazard reading. A scientific article will be composed of a certain number of "*information modules*" meaningfully connected by relationships.

All these changes will have an important transformational impact on the scholarly record. Scholarly record is taken to mean the aggregation of scientific journals, gray literature and conference presentations plus the underlying data and other evidence to support the published findings. At present, this record is poorly connected and this constitutes a major obstacle to a full engagement by scholars who need to be able to move from paper to paper, from hypotheses to evidence, or between paper and underlying data efficiently. Modern science requires the establishment of a linked scholarly record in order to effectively support scholarly inquiry.

2. The Changing Scenario of the Scholarly Record

We have identified the following main transformations that depend on linking technologies.

2.1. Scientific literature is becoming increasingly network-centric: Linked Scientific Literature Spaces

We foresee that future scientific literature will be increasingly network-centric. By this we mean that the linear form of the scientific article will be overcome and it will be presented as a network of modules connected by relations. These modules can contain organizational information concerning the structural aspects of an article as well as scientific discourse information concerning hypotheses made by the author of an article, evidence for the hypotheses, underlying datasets, findings, pointers to future research, etc. The conceptual relations between modules are materialized by explicitly labeled links. We can have different types of links: *semantic links*, *rhetorical links* and *pragmatic links*.

A generalization of the network-centrality of scientific information leads to the creation of a **linked scientific literature space** of disciplinary or interdisciplinary scope. A scientific contribution thus becomes a rigorously connected, substantiated node or region in a linked scientific literature space.

Linked literature spaces allow scholars to surf literature, enabling them to find and access not only a specific article but also a whole range of information contained in diverse articles. An important consequence of the creation of linked literature spaces is that information seeking becomes a horizontal rather than a vertical form of behavior. This enables researchers to find prevailing opinion more easily. Linked scientific literature spaces can thus hasten scientific consensus. A linked literature space also facilitates the identification of the lineage of an idea or concept.

2.2. Scientific information is becoming increasingly data-centric: Linked Scientific Data Spaces

New high-throughput scientific instruments, telescopes, satellites, accelerators, supercomputers, sensor networks and running simulations are generating massive amounts of data. In a data-dominated science, there are rapidly-expanding demands of "data-everywhere"**.** The most acute challenge stems from research teams relying on a large number of diverse and interrelated datasets but having no way to manage their scientific data spaces in a principled fashion. Therefore, there is a need for mechanisms and approaches that allow the linking of datasets produced by diverse research teams. Linking a dataset refers to the capability of linking it to other external datasets, and in turn being linked from external datasets. A generalization of the linking data concept leads to the creation of **linked scientific data spaces** of disciplinary or interdisciplinary scope. The concept of scientific data spaces responds to the increasing demands by researchers for data-everywhere. It will enable researchers/search engines to start browsing in one dataset and then to navigate to related datasets.

Not only can the scientific disciplines benefit from the application of the linked data paradigm but also the Humanities. In fact, advanced information technologies have led to the generation of very large volumes of linguistic data, i.e. text and speech corpora, dictionaries, lexicons, language descriptions, etc. The challenge is now to interlink this wealth of data accumulated in more than half a century of computational linguistics, of empirical, corpus-based study of language and of computational lexicography in all its heterogeneity. By linking this data, automated analyses of literary works can be carried out much more effectively. Not only researchers in the Humanities will benefit from linking linguistic data, but also the general public can be offered extended search possibilities.

A generalization of the linking language data and humanities data concept leads to the creation of **linguistic linked data spaces as well as Humanities linked data spaces**.

2.3. Scientific information as a result of Multidisciplinary Research

Scientific information is increasingly the result of Multidisciplinary Research: Linking Discipline-specific Literature Spaces and/or Discipline-specific Data Spaces with Supporting Discipline-specific Ontologies/ Terminologies/Vocabularies.

Scientific communication across disciplinary boundaries needs semantic enhancements to make the text intelligible to a broad audience composed of specialists in different scientific disciplines. This need motivated the current development of semantic publishing. Semantic publishing is taken to mean the enhancement of the meaning of an online research article by automatically disambiguating and semantically defining specialist terms. This can be achieved by linking to discipline-specific ontologies and standard terminology repositories, by linking to other information sources of relevance to the article, and by direct linking to all of the article's cited references. An additional semantic enhancement can be obtained by intelligently linking a scientific text to third-party commentaries, archived talks and Web sites.

Ontologies are also considered as a suitable formal tool for sophisticated data access. Data collections should be connected to appropriate ontologies.

In the context of a networked scholarly world, domain-specific ontologies are not standalone artifacts. They relate to each other in ways that can affect their meaning, and are distributed in a network of interlinked semantic resources, reflecting their dynamics, modularity and contextual dependencies. The alignment of domain-specific ontologies is crucial for data reusability. It is achieved through a set of mapping rules that specify a correspondence between various entities, such as objects, concepts, relations, and instances. Several concept and relation constructors are offered to construct complex expressions to be used in mappings.

2.4. Scientific information is increasingly based on the integration between text and data: Linking Literature Spaces with Data Spaces

In a data-dominated science, research data are increasingly being regarded as first class citizens of scientific communication, with their own identity and metadata. Modern science requires integrated support to the whole research data life cycle. Scientists need to publish their raw data sets, experimental details, analytical methods and visualizations, in addition to traditional scholarly publications. This record of the complete scientific discovery process will enable scientists to examine the underlying data while reading a paper. They could redo analyses and reproduce and verify results. Or they could examine all the literature about this data.

In summary, linking publications to the underlying data can produce significant benefits:

- it can help the data to be better discoverable;
- it can help the data to be better interpretable;

- it can provide the author with better credit for the data;
- and conversely it can add depth to the article and facilitate better understanding.

Future scholarly cyberinfrastructures should enable the unification of all scientific data with all literature to create a world in which data and literature interoperate. This vision implies the capability to link literature spaces with data spaces.

Linking Literature spaces with Data Spaces will increase scientific "information velocity" and will enhance scientific productivity as it will improve data availability, discoverability, interpretability and re-usability.

3. Digital Scholarship Infrastructure

By *Digital Scholarship Infrastructure* we mean an enabling framework for data, information and knowledge discovery, advanced literature analyses, and new scholar reading and learning practices based on linking and semantic technologies.

In particular, future *Digital Scholarship Infrastructures* should support:

- a *Scholarly Linking Environment* that:
    - provides a core set of *linking* services that create discipline-specific linked literature spaces and discipline-specific linked data spaces, connect literature spaces with data spaces, and build connections between diverse discipline-specific literature spaces;
    - supports the creation, operation and maintenance of a core set of *linkers*. A linker is a software module that exploits encoded knowledge and metadata information about certain datasets or articles in order to build a relation between modules and/or datasets. Different types of linkers should be supported in order to implement the different types of relations between article modules and datasets. Linkers that connect modules related by a causality relationship, by a similarity relationship, by an "aboutness" relationship, or by a generic relationship; linkers that connect an article with the underlying data set; linkers that connect a dataset with the supported articles, etc.;
    - provides a core set of *intermediary services* that make the holdings of discipline-specific repositories and data centers, data archives, research digital libraries and publisher's repositories discoverable, understandable, and (re)usable;
    - supports the creation, operation and maintenance of *mediators*. A mediator is a software module that exploits encoded knowledge and metadata information about certain datasets or articles in order to implement an intermediary service. A core set of mediators should include: data discovery mediators, article module discovery mediators, mapping mediators, matching mediators, consistency checking mediators, data integration mediators, etc.;

- a *Scholarly Reading and/or Learning Environment* that:
    - supports the creation, operation, and maintenance of a core set of "*scholarly workflows*." Scholars should be enabled to describe an "*abstract workflow*" by specifying a number of abstract tasks. These tasks include identity resolution, text analysis, literature analysis, lineage analysis, reproducibility of work, repeatability of experiments, etc. The abstract workflow or workflow template is mapped into a *concrete workflow* using mappings that, for each task, specify a linker or a mediator, or a service to be used for its implementation. An abstract workflow is an acyclic graph in which the nodes are tasks and the edges present links that connect the output of a given task to the input of another task, specifying that the artifacts produced by the former are used by the latter. The instantiation of a workflow, that is, the mapping of an abstract workflow into a concrete workflow results in a *"scholarly reading/learning pattern*". By scholarly reading/learning pattern we mean a set of meaningfully linked article modules and data sets that support a scholarly activity (reading/learning/research). In essence, scholarly reading/learning patterns draw paths within the linked scholarly record;
    - supports the creation and maintenance of reading and learning profiles in order to enable the creation of "personalized reading/learning patterns";
    - enables scholars, readers and learners to find the scientific information they are looking for and

correctly interpret it by allowing them to surf the linked scholarly record following suitable scholarly patterns.

In addition, cyberscholarship infrastructures should:

- maintain article module metadata registries;
- maintain link metadata registries;
- maintain data dictionaries, discipline-specific ontologies and terminologies.

4. Concluding Remarks

Jim Gray's vision of a world in which all scientific literature and all scientific data are online and interoperating is rapidly becoming a reality. Building linked discipline-specific scientific records will provide crucial support to modern science by producing radical changes in the scientific method, greatly contributing to educating young scholars, revolutionizing scientific publication, and increasing the productivity of scientists. It will produce a shift in scientific practice from advances based on the traditional scientific method to advances being driven by patterns of data. New insights will arise from connections and correlations found between diverse types of information resources. Another shift will be produced in scholarly information seeking behavior; moving rapidly through the linked scholarly record and identifying relevant information on the move will become a fundamental activity.

To make this happen, advances in many other technologies are needed. The further development of all these technologies will accelerate the building of digital scholarship infrastructures that will provide advanced services and tools to support the scholarly lifecycle.

Acknowledgements

This paper has been inspired by a number of papers addressing some of the issues discussed; the references to these papers are reported below. The author is grateful to the authors of these papers.

References

1. Altman M, King G. A Proposed Standard for the Scholarly Citation of Quantitative Data. *D-Lib Magazine* March/April 2007;13(3/4).
2. Bechhofer S, De Roure D, Gamble M, Goble C, Buchan I. Research Objects: Towards Exchange and Reuse of Digital Knowledge. In: *Proc. The Future of the Web for Collaborative Science (FWCS 2010)*. Raleigh, NC, USA; 2010.
3. Belhajjame K, Corcho O, Garijo D, Zhao J, Missier P, Newman DR, Palma R, Bechhofer S, Garcia-Cuesta E, Gómez-Pérez JM, Klyne G, Page K, Roos M, Ruiz JE, Soiland-Reyes S, Verdes-Montenegro L, De Roure D, Goble CA. Workflow-Centric Research Objects: A First Class Citizen in the Scholarly Discourse. In: *Proceedings of the ESWC2012 Workshop on the Future of Scholarly Communication in the Semantic Web (SePublica2012)*. Heraklion, Greece, 2012.
4. Bizer C, Heath T, Berners-Lee T. Linked Data – The Story So Far. *International Journal on Semantic Web and Information Systems* 2009;**5**(3):1-22.
5. Bizer C. Interlinking Scientific Data on a Global Scale. *Data Science Journal* 2013;**12**:6-12.
6. Bizer C. Evolving the Web into a Global Data Space (Keynote talk). In: *28th British National Conference on Databases (BNCOD2011)*. Manchester, UK; 2011.
7. Borgman CL. Data, disciplines, and scholarly publishing. Learned Publishing 2008;**21**:29-38.
8. Bourne P, Clark T, Dale R, de Waard A, Herman I, Hovy E, Shotton D, editors. Improving The Future of Research Communications and e-Scholarship (Dagstuhl Perspectives Workshop 11331). *Dagstuhl Manifestos* 2011;**1**(1):41-60.
9. Bourne P. Will a biological database be different from a biological journal? *PLoS Comp Biol* 2005;**1**(3):e34.
10. Buckingham Shun SJ, Uren V, Gangmin L, Sereno B, Mancini C. Modeling Naturalistic Argumentation in Research Literatures: Representation and Interaction Design Issues. *International Journal of Intelligent Systems* 2007;**22**(1):17-47.
11. Buckingham Shun SJ. *Net-Centric Scholarly Discourse?* Available at the URL: http://slidesha.re/qvoqoU
12. Chiarcos C, Nordhoff S, Hellmann S, editors. *Linked Data in Linguistics*. Heidelberg: Springer Verlag 2012.
13. Castelli D, Manghi P, Thanos C. A Vision towards Scientific Communication Infrastructures. *International Journal on Digital Libraries* 2013;**13**(3-4):155-169.
14. Decker S. *From Linked Data to Networked Knowledge*. URL: http://videolectures.net/eswc2013_decker_networked_knowledge/
15. de Waard A. From Proteins to Fairytales: Directions in Semantic Publishing. *IEEE Intelligent Systems* March/April 2010;**25**(2):83-88.

16. de Waard A, Buckingham Shum SJ, Carusi A, Park J, Samwald M, and Sándor A. Hypotheses, evidence and relationships: The HypER approach for representing scientific knowledge claims. In: *Proceedings 8th International Semantic Web Conference, Workshop on Semantic Web Applications in Scientific Discourse*. Berlin: Springer Verlag 2009.
17. Evans J. Electronic Publication and the Narrowing of Science and Scholarship. *Science* 2008;**321**(5887):395-399.
18. Fink JL, Fernicola P, Chandran R, Parastatidis S, Wade A, Naim O, Quinn GB, Bourne PE. Word add-in for ontology recognition: semantic enrichment of scientific literature. *BMC Bioinformatics* 2010;**11**:103-110.
19. Franklin M, Halevy A, Maier D. From Databases to Dataspaces: A New Abstraction for Information Management. *SIGMOD Record* 2005;**34**:27-33.
20. Goble C, De Roure D. The Impact of Workflows on Data-centric Research. In 24.
21. Gray J, Szalay A, Thakar A, Stoughton C, van de Berg J. Online Scientific Data: Curation, Publication and Archiving. Technical Report MSR-TR- 2002-74. Redmond, WA: Microsoft Research 2002.
22. Gruber T. Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In: Guarino N, Poli R, editors. *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Kluwer Academic Publishers. Substantial revision of paper presented at the International Workshop on Formal Ontology, March, 1993, Padua, Italy. Available as Technical Report KSL 93-04, Knowledge Systems Laboratory, Stanford University.
23. Harmsze F. *A Modular Structure for Scientific Articles in an Electronic Environment (PhD thesis)*. University of Amsterdam, 2000.
24. Hey T, Tansley S, Tolle K, editors. *The Fourth Paradigm: Data Intensive Scientific Discovery*. Redmond, WA: Microsoft Research 2009.
25. Hunter J. Scientific Models – A user-oriented Approach to the Integration of Scientific Data and Digital Libraries. In: *Proceedings VALA 2006*. Melbourne, 2006.
26. Ikeda R, Widom J. Panda: A System for Provenance and Data. *IEEE Data Engineering Bulletin* (Special Issue on Data Provenance) 2010;**33**(3):42-49.
27. Kircz JG, Harmsze F. Modular Scenarios in the Electronic Age. In *Conferentie Informatiewetenschap,* 2000. p. 31-43.
28. Kircz JG. New Practices for Electronic Publishing - New Forms of the Scientific Paper. Learned Publishing 2002;**15**:27-32.
29. Lagoze C, Van de Sompel. H. *The OAI Protocol for Object Reuse and Exchange*. URL: http://www.openarchives.org/ore/
30. Lynch CA. Jim Gray's Fourth Paradigm and the Construction of the Scientific Record. In: Hey T, Tansley S, Tolle K, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft Research; 2009;177-183.
31. Mackenzie Owen JS. *The Scientific Article in the Age of Digitization (PhD thesis)*. University of Amsterdam, 2005.
32. McPherson T. Scaling Vectors: Thoughts on the Future of Scholarly Communication. *Journal of Electronic Publishing* 2010;13(2).
33. Moreau L, Freire J, Futrelle J, McGrath RE, Myers J, Paulson PR. The Open Provenance Model: An Overview. In: Freire J, Koop D, Moreau L, editors. *Provenance and Annotation of Data and Processes, Second International Provenance and Annotation Workshop (IPAW 2008)*. Revised Selected Papers. Springer Verlag 2008:323-326.
34. Nicholas D, Huntington P, Jamali HR, Rowlands I, Dobrowolski T, Tenopir C. Viewing and reading behaviour in a virtual environment: The full-text download and what can be read into it. *Aslib Proceedings* 2008;**60**(3):185-198.
35. Nicholas D, Huntington P, Jamali HR, Dobrowolski T. Characterizing and Evaluating Information Seeking Behavior in a Digital Environment: Spotlight on the "Bouncer". *Information Processing and Management* 2007;**43**(4):1085-1102.
36. Paskin N. Digital Object Identifier for Scientific Data. In: *19th International CODATA Conference*. Berlin, 2004.
37. Poggi A, Lembo D, Calvanese D, De Giacomo G, Lenzerini M, Rosati R. Linking data to ontologies. In: Spaccapietra S, editor, *Journal on data semantics X*, Berlin:Springer-Verlag 2008;133-173.
38. Renear A, Palmer C. Strategic Reading, Ontologies, and the Future of Scientific Publishing. *Science* 2009;**325**:828-832.
39. Seringhaus T, Gerstein M. Publishing Perishing? Towards Tomorrow's Information. *BMC Bioinformatics* 2007;8:17.
40. Shotton D. Semantic Publishing: The Coming Revolution in Scientific Journal Publishing. Learned Publishing 2009;**22**(2):85-94.
41. Simon B, Miklos Z, Nejdl W, Sintek M, Salvachua J. Smart Space for Learning: A Mediation Infrastructure for Learning Services. In: *Proceedings of the 12th World Wide Web Conference* 2003.
42. Strang T, Linnhoff-Poppien C. A Context Modeling Survey. In: *Workshop on Advanced Context Modeling, Reasoning and Management associated with the Sixth International Conference on Ubiquitous Computing (UbiComp 2004).*
43. Taylor IJ, Deelman E, Gannon D, Shields MS, editors. *Workflows for e-Science: Scientific Workflows for Grids*. Springer, 2007.
44. Tenopir C, King D. Electronic Journals and Changes in Scholarly Article Seeking and Reading Patterns. *D-Lib Magazine* 2008;**14**(11/12).
45. Thearling K. *An Introduction to Data Mining*. URL: http://www.thearling.com/dmintro/dmintro_2.htm
46. Woutersen-Windhouwer S, Brandsma R, Verhaar P, Hogenaar A, Hoogerwerf M, Doorenbosch P, Durr E, Ludwig J, Schmidt B, Sierman B. In: Vernooy-Gerritsen M, editor, *Enhanced Publications*. Amsterdam: Amsterdam University Press; 2009.