A Vision and Agenda for Theory Provenance in Scientific Publishing

Ian Wood[1], J. Walter Larson[1,2,3], and Henry Gardner[1]

[1] Department of Computer Science, The Australian National University
Canberra ACT 0200 Australia
[2] Computation Institute, University of Chicago, Chicago, IL USA
[3] Mathematics and Computer Science Division, Argonne National Laboratory,
Argonne, IL 60439, USA

Abstract. Primary motivations for effective data and process provenance in science are to facilitate validation and reproduction of experiments and to assist in the interpretation of data-analysis outcomes. Central to both these aims is an understanding of the ideas and hypotheses that the data supports, and how those ideas fit into the wider scientific context. Such knowledge consists of the collection of relevant previous ideas and experiments from the body of scientific knowledge, or, more specifically, how those ideas and hypotheses evolved, the steps in that evolution, and the experiments and results used to support those steps. This information we term the provenance of ideas or *theory provenance*. We propose an integrated approach to scientific knowledge management, combining data, process and theory provenance, providing full transparency for effective verification and review.

Keywords: provenance, theory provenance, knowledge representation, scientific publishing, grid, semantic grid, semantic citation, semantic network.

1 Introduction

Data provenance has been described as a record of the computational steps that transform raw experimental data into that which is published [1]. Data provenance provides transparency in data acquisition and processing, allowing those who use the data to determine its validity and to verify its accuracy [2]. It helps identify the significance and meaning of derived data, which can be obscured by complex automated workflows, not only from those reading published work, but also from those who created the data [3]. Two surveys of data provenance practices in eScience have been compiled which report that, though provenance issues are being addressed, there is still much work to be done, in particular on standards to allow the portability of provenance metadata [4,5]. Zhao et al. recognised the need to identify theoretical context within data provenance [6].

The provenance of *data* and process provides, in essence, a history of how the data was produced and manipulated. The provenance of ideas provides a history

L. Chen et al. (Eds.): DASFAA 2009, LNCS 5667, pp. 112–121, 2009.

of how *ideas* evolve and how they relate to preceding enquiries. Data provenance provides the *concrete history* of the development of the data, whereas theory provenance provides the *abstract history* of the ideas that relate to the data. The provenance of ideas provides *context* for the data, helping us interpret its meaning and to understand the evolution of the experimental techniques used. Figure 1 illustrates this relationship during the development of new theories.

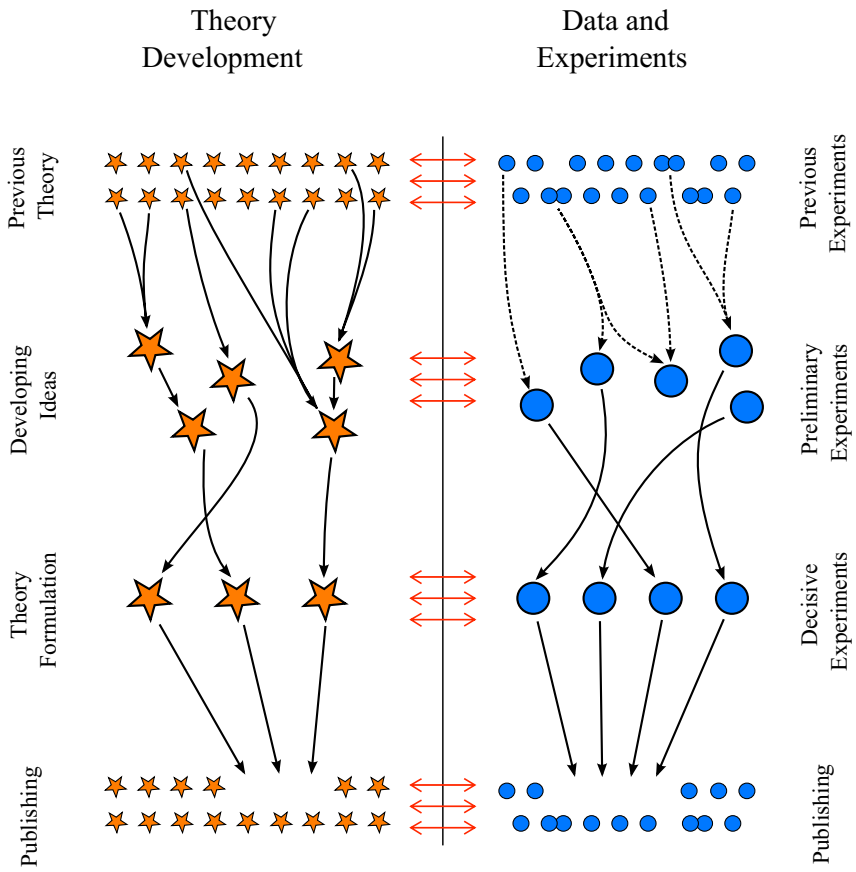

Fig. 1. Science Lifecycle

Our vision is that all the steps in this figure can be represented as a semantic network and exposed to diverse automated analyses, thus improving knowledge utility and refining knowledge services such as knowledge discovery, validation and attribution.

There have been concerns expressed in the scientific community about the lack of provenance in published work. For example, new algorithms published in computational science often lack sufficient detail to reproduce the published

results [7]. Furthermore, most data provenance approaches tag data transformations at the level of "name of software package," "version number," "platform and build configuration on which the code was executed," but not at the level of "algorithm(s) implemented in the code," "underlying assumptions under which the code is valid," etc... These latter aspects speak to how much we can *trust* the results of a computational workflow. We believe these concerns are legitimate and require a solution that in fact complements data provenance. Specifically, we identify this missing element of scientific process provenance as the provenance of the ideas implemented by the data transformations applied.

To track the evolution of ideas, we must delve into records of their development and presentation. These records could, for example, take the form of laboratory notes and records of collaborative events [8,9,10] or published scientific literature. The provenance of ideas is exposed when these records contain references indicating the relationships between ideas, such as "extends", "depends on" or perhaps "refutes". We refer to such references as *semantic citations*, extending current citation techniques. Zhuge has made a study of the types of relationships that might exist [11]. The resultant semantic web of knowledge, or *knowledge grid* [11] could be readily analysed to reveal the provenance of an idea.

In Section 2 we further develop the idea of theory provenance in the context of scientific publishing and outline some initial requirements and implementation paths. In Section 3 we discuss theory provenance with refined knowledge representation and automated reasoning, introducing existing scientific knowledge representation initiatives. We conclude with a discussion of the need for knowledge management standards for metadata in scientific publishing and outline the core aspects that such a standard should include.

2 Theory Provenance and Scientific Publishing

Theory provenance provides a history of how ideas evolve and how they relate to those that precede them. In this section we outline how theory provenance can be integrated with and facilitated by existing e-science and knowledge management technologies. We suggest an incremental approach with a highly flexible standard for knowledge representation and linking and discuss research directions for easy implementation in publishing and knowledge development contexts.

2.1 Scientific Publishing

In order to access the provenance of ideas in published literature, we must delve into the body of published science and identify which previous results relate to the ideas under consideration, and the manner of that relationship. Advances in scientific publishing have greatly simplified that task. The majority of scientific journals and conferences (possibly all) now publish their material in digital form. Repositories of scientific articles often have associated knowledge management services such as keyword searches and subject categorisation (eg: Springer, IEEE and ACM) and third party search and categorisation services such as

CiteSeer [12], Google Scholar [13] and the ISI Web of Knowledge [14] provide one-stop portals to much of the worlds published science.

Despite these advances, our scientific publishing techniques can be said to have the spirit of being "on paper", albeit digital paper, largely failing to utilise many of the powerful knowledge management techniques that are available [15]. The underlying format remains solely text based with unsophisticated citation techniques and little scope for directly referencing supporting data[1]. Though keyword indexing and full text searches provide some ability to link related articles, they currently fall well short of tracking ideas through the body of scientific literature. A key concept needed here is the idea of semantic citation.

2.2 Semantic Citation

Citations are the vehicle for capturing provenance in scientific publishing, but the current citation techniques are coarse. Without reading the text (a task difficult for machines and onerous for humans), a citation tells you nothing about which specific concepts are related, and nothing about the relationship between those concepts or results. Currently, this information can only be obtained by reading both papers and considering the context in which the citations appear.

A citation could contain information about the nature of the relationship between publications. Further, if the key concepts and arguments in a paper were available in a machine readable form (see Section 3), a citation could indicate which specific concepts are related. For example, it could indicate that a concept in the new paper assumes the validity or truth of one in the earlier one, or conversely that the new concept contradicts the earlier concept, or it could simply indicate that the new concept is distinct from or is a refinement or subconcept of the earlier one. One important semantic role is an indication that the cited concept is the same as the other. We will refer to citations with semantic information about their relationship as *semantic citations*. A related idea was presented by Carr et. al. in [16]. They present a service that semantically links documents that contain similar concepts, utilising existing document metadata. In essence, they are creating something similar to semantic citations between existing documents on the web.

In practical terms, given a format for representing scientific knowledge, semantic citations should be straightforward to implement. Analogous to URI's (Universal Resource Indicators [17]) and DOI's (Digital Object Identifiers [18]), elements of represented knowledge (data, theories, entities etc..) could be given unique identifiers which could be quoted in the citing document or it's metadata. The imposition on scientists to annotate their citations would not be significantly greater than the current citation model. In addition, modern data mining techniques could be applied to existing publications to identify the semantic role of citations. To the best of our knowledge this has not been attempted.

[1] This publishing model was first used in 1665 when the first editions of "Journal des sçavans" and "Philosophical Transactions of the Royal Society" appeared and has not changed significantly in the ensuing 350 years.

2.3 Provenance and the Development of Ideas

The process of developing new ideas frequently entails a collection of notes, experimental results and other records that can be semantically linked in a similar way to published results. Electronic laboratory notebooks and other collaborative tools (see, for example, [8,9,10]) incorporate knowledge management services for annotating and organising records of experiments, meetings and other collaborative events. Used appropriately, these tools could track the provenance of ideas as they develop during scientific collaborations.

A simple flexible framework for representing semantic citations and knowledge could be implemented for such systems. Scientists could add citations to published papers to these notes as they work. This information would then be readily available when authoring a new paper, and the represented knowledge could be incorporated into the paper, providing a semantic representation of the published work with little extra effort.

2.4 Trust and Validity

We have discussed theory provenance both as ideas evolve during the development of new hypotheses and theorems and within the body of peer reviewed, published science. It would be useful to give such contexts different levels of trust of validity. Other levels may be desirable as well - for example pre-prints that have not yet undergone peer review, but which the authors consider to be of publishable quality. There is scope for adapted peer review structures, utilising the opinions of a wider community of scientists with relevant expertise in a similar way to collaborative tagging.

A standard for theory provenance should include scope for levels of verification, validity and trust.

3 Granular Representation and Automated Reasoning

A scientific publication often contains several key concepts, experimental techniques and other elements. To maximise the effectiveness of semantic citations, these sub-concepts could also be represented in a machine readable way. A citation could then point to and from specific semantic elements.

The granularity of the represented concepts could, in principle, be very fine, including individual steps in the flow of logic within a publication. This could lead to automatic or semi-automatic verification of the logical conclusions presented.

Compiled libraries of formally represented mathematics and their attendant theorem provers/checkers such as MIZAR [19] and IsarMathLib [20] faithfully represent theory dependencies and supporting arguments and as such they provide a substantial step toward granular theory provenance for science.

3.1 Knowledge Representation in Science

As the quantity and complexity of scientific data and knowledge has increased, new technologies have been developed to organise and effectively utilise it. In

many areas of science, substantial knowledge bases have been created or are in the process of creation. These knowledge bases are primarily in the form of description logic ontologies. Their application has led to sophisticated data retrieval and resource management systems, and reference ontologies [21,22,23,24]. Another application of ontologies in science is data integration—well developed ontologies, made in collaboration with relevant expert communities, serve as a standard form of annotation and allow diverse data formats to be utilised interoperably. There are several projects working on these issues [25,26]. Numerous platforms and methodologies for ontology construction and maintenance have been developed [22,27,28,29].

Significant work in ontology development for science has been in association with the construction of semantic grids. The term *Semantic Grid* was coined by De Roure, Jennings and Shadbolt to describe "the application of Semantic Web technologies both on and in the Grid" [30]. The effectiveness and efficiency of Grid services is substantially enhanced by this approach, particularly when the Grid contains large and complex resources [31,32]. This can also be seen in research on workflow automation [33] and resource discovery [34,35]. Virtual observatories [36,37], though they do not claim to be semantic grids, satisfy Foster's grid criteria [38] and apply Semantic Web technologies.

Semantic Grids may be the natural platform for our vision of semantic publishing. Grids federate resources to create virtual organisations; thus they may be employed by a group of scientists to define scope for their fields of study. Grids control access to resources, allowing differing levels of authorisation; thus they allow some users to read and write resources (such as ontologies or other semantic descriptive data), while others may merely read. Semantic Grids provide a framework for semantic annotation of resources and services for workflow automation and resource discovery. Grid protocols use open standards, reducing barriers to integration of knowledge repositories.

Specialised scientific markup languages have been driven by the need to extend HTML to perform typesetting of technical information such as mathematical formulae, by the need for standard information and data exchange formats, and the need for standard formats for automated processing. Numerous markup languages supporting science and eResearch exist or are under development[39]. In general, these are **not** description logic based, and many are too expressive for effective automated reasoning, as we shall see in the following section.

For theory provenance at any level to become an effective tool for science, there is a need for flexible standards for representing scientific concepts and the links between them. Ideally, such standards should be able to incorporate widely differing representation formats for scientific knowledge and information (such as the various markup languages and DL ontologies above) as well as sophisticated conceptual links such as those used in the mathematical libraries mentioned above.

3.2 Description Logics

The study of knowledge representation for artificial intelligence led to questions about the tractability and computational complexity of different systems.

Theoretical attempts by logicians to answer these questions led to a deeper understanding of the tractability of automated reasoning and the development of a family of languages for representing knowledge. These languages are called *description logics* (DLs).

Seminal work on the computational complexity of DLs was done by Hector J. Levesque and Ronald J. Brachman [40] who recognized that there is a trade-off between the expressive power of a language for knowledge representation and the difficulty of reasoning with the resultant knowledge bases. A fundamental result is that languages entailing first order logic result in potentially unbounded reasoning operations. In a sufficiently expressive system, there will *always* be questions for which an automatic reasoner will *never* find an answer. Mathematics with the real numbers is such a system.

Description logics form the basis of many automated reasoning applications and systems today. In particular the Web Ontology Language (OWL—the W3C standard ontology language for the Semantic Web [41]), is based on a description logic. Semantic Web technologies utilising automated reasoning have been effectively applied to enhance qualities of service and efficiencies in semantic grids and other science applications.

In the previous section we mentioned the need to accommodate diverse standards for representing scientific knowledge. In order to gain the indexing and search efficiencies and other services that Semantic Web technologies provide, translations from or approximations of these standards to the languages of the Semantic Web would be needed.

3.3 Reasoning with Complex Knowledge

We might hope that automated reasoning techniques could be applied to a body of finely-represented scientific knowledge to obtain new results that are implied by some combination of known results, but that have not, as yet, been recognised. This would be possible if the knowledge can be faithfully represented by appropriate description logics, however in many areas of science core results are expressed in mathematics. We can devise formalised representation systems for such knowledge, but, as we saw in the previous section automated reasoning based on logic is unreliable with such highly expressive systems. This does not mean, however, that automated reasoning is not useful for scientific applications. Instead, we observe that in mathematics, automated theorem proving systems have successfully aided researchers to find mathematical proofs that had not previously been known [42]. These systems often require (sometimes substantial) human intervention. Cotton has recently reviewed the state of automated and semi-automated reasoning for mathematics with results that are encouraging [43]. However much work would have to be done before these techniques can provide substantial support for mathematical representations of scientific knowledge.

4 Conclusions

We have defined two key concepts that we believe are necessary to extend data provenance to become scientific process provenance: *Theory provenance* is the

provenance of the ideas and reasoning behind scientific results or algorithms implemented in software that performs scientific data processing. *Semantic citation* is the addition of attributes to a citation to describe its semantic role. Taken together, these two concepts are at the centre of a vision for semantic publication that will (1) enable the integration of data provenance into a scientific argument, and (2) provide a fuller identification of the underlying assumptions and applicability of transformations used in scientific computational work flows. We have described in detail the relationships between theory provenance and semantic citation and a set of knowledge representation technologies that we believe can be employed to implement our vision for semantic publication. We have identified semantic grids as a promising platform for implementing our vision.

The ideas we have presented here would have profound implications to science publishing, however uptake of the ideas would require substantial changes to the science publishing infrastructure as well as to the publishing habits of scientists. It is unlikely that such changes will be adopted by the wider scientific community without effective demonstration of theory provenance. Clearly, the vision will also not be realised unless these changes can be implemented without undue burden on working scientists, and automation in the form of collaborative technologies (see, for example, [8,9,10]) are likely applicable, or may be adapted to allow automatic implementation of semantic citation.

Note that there are other implications of our proposal outside of provenance. A network of finely-represented semantically linked knowledge would be open to many forms of automated analysis, as well as innovative applications of Web 2.0 technologies and other technologies of the future.

These ideas are also applicable beyond science. A promising area for future investigation is the applicability of the techniques described in this paper to public policy formulation, and objective measures of how well a given public policy aligns with supporting domain research—for example policies under discussion to respond to anthropogenically-generated global warming.

Acknowledgements. We would like to acknowledge Dr. Peter Baumgartner, Dr. Ian Barnes, Dr. Roger Clarke, Dr. Scott Sanner, Dr. Catherine Legg, Dr. Jason Grossman and Dr Tom Worthington for their discussions and expert advice.

References

1. Foster, I., Kesselman, C. (eds.): The Grid 2: Blueprint for a New Computing Infrastructure. The Morgan Kaufmann Series in Computer Architecture and Design. Morgan Kaufmann, San Francisco (2003)
2. Moreau, L., et al.: Special Issue: The First Provenance Challenge. Concurrency and Computation: Practice and Experience 20(5), 409–418 (2008)
3. Miles, S., Deelman, E., Groth, P., Vahi, K., Mehta, G., Moreau, L.: Connecting scientific data to scientific experiments with provenance. In: IEEE International Conference on e-Science and Grid Computing, December 2007, pp. 179–186 (2007)
4. Simmhan, Y.L., Plale, B., Gannon, D.: A survey of data provenance in e-science. SIGMOD Rec. 34(3), 31–36 (2005)
5. Bose, R., Frew, J.: Lineage retrieval for scientific data processing: a survey. ACM Comput. Surv. 37(1), 1–28 (2005)

6. Zhao, J., Goble, C., Stevens, R., Bechhofer, S.: Semantically linking and browsing provenance logs for E-science. In: Bouzeghoub, M., Goble, C.A., Kashyap, V., Spaccapietra, S. (eds.) ICSNW 2004. LNCS, vol. 3226, pp. 158–176. Springer, Heidelberg (2004)
7. Quirk, J.: Computational science, same old silence, same old mistakes, something more is needed .... In: Adaptive Mesh Refinement - Theory and Applications. Lecture Notes in Computational Science and Engineering, vol. 41, pp. 3–28. Springer, Heidelberg (2005)
8. Shum, S., De Roure, D., Eisenstadt, M., Shadbolt, N., Tate, A.: CoAKTinG: Collaborative Advanced Knowledge Technologies in the Grid. In: 2nd Workshop Advanced Collaborative Environments, http://www.aktors.org/coakting/
9. Myers, J., Mendoza, E., Hoopes, B.: A Collaborative Electronic Notebook. In: Proceedings of the IASTED International Conference on Internet and Multimedia Systems and Applications (IMSA 2001), August 2001, pp. 13–16. ACTA Press (2001)
10. Myers, J.D., Chappell, A., Elder, M., Geist, A., Schwidder, J.: Re-integrating the research record. Computing in Science and Engineering 5(3), 44–50 (2003)
11. Zhuge, H.: The Knowledge Grid. World Scientific, Singapore (2004)
12. CiteSeer: http://citeseer.ist.psu.edu/ or http://citeseersx.ist.psu.edu/
13. GoogleScholar: http://scholar.google.com/
14. ISI: Isi web of knowledge, http://apps.isiknowledge.com/
15. de Waard, A.: Science publishing and the semantic web, or: Why are you reading this on paper. In: European Conference on the Semantic Web (2005)
16. Carr, L., Hall, W., Bechhofer, S., Goble, C.: Conceptual linking: ontology-based open hypermedia. In: WWW 2001: Proceedings of the 10th international conference on World Wide Web, pp. 334–342. ACM, New York (2001)
17. W3C: Naming and addressing: Uris, urls, ..., http://www.w3.org/Addressing/
18. DOI: The digital object identifier system, http://www.doi.org/
19. MIZAR: The mizar project for formalized representation of mathematics, http://www.mizar.org/
20. IsarMathLib: Library of formalized mathematics for isabelle/isar (zf logic), http://savannah.nongnu.org/projects/isarmathlib, See also [IsarMathLib]
21. Stevens, R., Goble, C., Bechhofer, S.: Ontology-based knowledge representation for bioinformatics. Briefings in Bioinformatics 1(4), 398–414 (2000)
22. Stevens, R.D., Robinson, A.J., Goble, C.A.: myGrid: personalised bioinformatics on the information grid. Bioinformatics 19(suppl. 1), i302–i304 (2003)
23. EMBL-EBI: Biological ontology databases. European Bioinformatics Institute, an Outstation of the European Molecular Biology Laboratory, http://www.ebi.ac.uk/Databases/ontology.html
24. Hu, X., Lin, T., Song, I., Lin, X., Yoo, I., Lechner, M., Song, M.: Ontology-Based Scalable and Portable Information Extraction System to Extract Biological Knowledge from Huge Collection of Biomedical Web Documents. In: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 77–83. IEEE Computer Society, Washington (2004)
25. Fox, P., McGuinness, D., Raskin, R., Sinha, A.: Semantically-Enabled Scientific Data Integration. In: US Geological Survey Scientific Investigations Report, vol. 5201 (2006), http://sesdi.hao.ucar.edu/
26. Zhang, X., Hu, C., Zhao, Q., Zhao, C.: Semantic data integration in materials science based on semantic model. In: IEEE International Conference on e-Science and Grid Computing, December 2007, pp. 320–327 (2007)

27. Bao, J., Hu, Z., Caragea, D., Reecy, J., Honavar, V.: A tool for collaborative construction of large biological ontologies. In: 17th International Conference on Database and Expert Systems Applications, 2006. DEXA 2006, September 2006, pp. 191–195 (2006)
28. Lin, H.N., Tseng, S.S., Weng, J.F., Lin, H.Y., Su, J.M.: An iterative, collaborative ontology construction scheme. In: Second International Conference on Innovative Computing, Information and Control, 2007. ICICIC 2007, September 2007, p. 150 (2007)
29. Xexeo, G., de Souza, J., Vivacqua, A., Miranda, B., Braga, B., Almentero, B., D' Almeida Jr., J.N., Castilho, R.: Peer-to-peer collaborative editing of ontologies. In: The 8th International Conference on Computer Supported Cooperative Work in Design, 2004. Proceedings, May 2004, vol. 2, pp. 186–190 (2004)
30. Roure, D., Jennings, N., Shadbolt, N.: Research agenda for the semantic grid: a future escience infrastructure, vol. 9. National e-Science Centre, Edinburgh (2001)
31. Roure, D.D., Jennings, N.R., Shadbolt, N.R.: The semantic grid: A future e-science infrastructure. In: Berman, F., Fox, G., Hey, A.J.G. (eds.) Grid Computing, pp. 437–470. Wiley, Chichester (2003)
32. Goble, C.: Putting semantics into e-science and grids. In: First International Conference on e-Science and Grid Computing, 2005, December 2005, p. 1 (2005)
33. Siddiqui, M., Villazon, A., Fahringer, T.: Semantic-based on-demand synthesis of grid activities for automatic workflow generation. In: IEEE International Conference on e-Science and Grid Computing, December 2007, pp. 43–50 (2007)
34. Somasundaram, T., Balachandar, R., Kandasamy, V., Buyya, R., Raman, R., Mohanram, N., Varun, S.: Semantic-based grid resource discovery and its integration with the grid service broker. In: International Conference on Advanced Computing and Communications, 2006. ADCOM 2006, December 2006, pp. 84–89 (2006)
35. Andronico, G., Barbera, R., Falzone, A.: Grid portal based data management for lattice qcd. In: 13th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, 2004. WET ICE 2004, pp. 347–351. IEEE, Los Alamitos (2004)
36. Berriman, B., Kirkpatrick, D., Hanisch, R., Szalay, A., Williams, R.: Large Telescopes and Virtual Observatory: Visions for the Future. In: 25th meeting of the IAU, Joint Discussion, vol. 8, p. 17 (2003)
37. Fox, P., McGuinness, D.L., Middleton, D., Cinquini, L., Darnell, J.A., Garcia, J., West, P., Benedict, J., Solomon, S.: Semantically-enabled large-scale science data repositories. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 792–805. Springer, Heidelberg (2006)
38. Foster, I.: What is the Grid? A Three Point Checklist. Grid Today 1(6), 22–25 (2002)
39. Bartolo, L.M., Cole, T.W., Giersch, S., Wright, M.: NSF/NSDL Workshop on Scientific Markup Languages. D-Lib Magazine 1(11) (2005)
40. Levesque, H., Brachman, R.: Expressiveness and tractability in knowledge representation and reasoning 1. Computational Intelligence 3(1), 78–93 (1987)
41. W3C: Web ontology language, `http://www.w3.org/TR/owl-features/`
42. Mccune, W.: Solution of the robbins problem. Journal of Automated Reasoning 19(3), 263–276 (1997)
43. Colton, S.: Computational discovery in pure mathematics. In: Džeroski, S., Todorovski, L. (eds.) Computational Discovery 2007. LNCS (LNAI), vol. 4660, pp. 175–201. Springer, Heidelberg (2007)