# Ontology Use for Semantic e-Science

Editors: Krzysztof Janowicz, Pennsylvania State University, USA, and Pascal Hitzler, Wright State University, USA.
Solicited Reviews: Michel Dumontier, Carleton University, Canada, and Manfred Hauswirth, DERI, National University of Ireland, Ireland.

Boyan Brodaric[a,*] and Mark Gahegan[b]
[a]*Geological Survey of Canada, 234B – 615 Booth St., Ottawa, Ontario, Canada, K1A 0E9*
[b]*Centre for eResearch and School of Environment,University of Auckland, Human Sciences Building, 10 Symonds St., Auckland, New Zealand*

**Abstract.** Ontologies are being widely used in online science activities, or e-Science, most notably in roles related to managing and integrating data resources and workflows. We suggest this use has focused on enabling e-science infrastructures to operate more efficiently, but has had less emphasis on scientific knowledge innovation. A greater focus on online innovation can be achieved through more explicit representation of scientific artifacts such as theories and models, and more online tools to enable scientists to directly generate and test such representations. This should lead to routine use of ontologies by scientists, and foster new and potentially different scientific results to help usher in next generation e-Science.

Keywords: ontology use, e-Science, semantic web

## 1. Introduction

Before the onset of scientific computing, the data, methods and theory used for science were often kept close together, in the head and notebook of the researcher. Development of computational infrastructure over the last 50 years has allowed first data and next methods to move far from their scientific creators. Many research communities are now congregating around online infrastructures that contain shared repositories of primarily data and methods. Such infrastructures are being used for the discovery, retrieval, and integration of online scientific resources, mainly scientific databases, and increasingly also to capture and describe scientific instruments, software, workflows, and experiments. These infrastructures and associated activities collectively comprise e-Science [6]. The number of e-Science initiatives is vast. Some examples are:

- The Geosciences Network: *GEON* (www.geongrid.org) [10]
- Cancer Biomedical Informatics Grid: *caBIG* (https://cabig.nci.nih.gov/) [13]
- Global Ocean Observing Initiative: *GOOS* (http://www.ioc-goos.org/) [2]

These and similar efforts are realizing important scientific benefits, which we claim can be largely attributed to three factors: improvements in resource quantity, improvements in representation, and improvements in communication:

(1) **Improvements in resource *quantity*** are realized by leveraging and integrating greater numbers of relevant online resources. New results ensue when more and bigger online assets are brought to bear on a problem, for example, such as when distributed computing is running remote applications, often automated and in parallel, over networks of massive databases or sensors. Data that is often expensive to capture or create is then more likely to see secondary use. The same goes for methods and other e-resources.

(2) **Improvements in *representation*** are realized by recording more complete and complex expressions of scientific knowledge as well as related research activities. More reliable results then accrue from far better levels of repeatability and explanation,

*Corresponding author. E-mail: brodaric@nrcan.gc.ca.

because online environments host machine-processable representations of many (ideally all) aspects of scientific investigation, and these can be accessed by greater numbers of scientists.

(3) **Improvements in *communication*** are realized by facilitating deeper and more frequent online collaboration between scientists. The enhanced connectivity of online science environments then increases the exchange of ideas.

Ontologies are already playing a pivotal role in these areas. For example, in virtual observatories [8] ontologies are: (1) being used to annotate the structure and content of scientific databases and workflows to make them interoperable, (2) helping guide the structure and content of scientific workflow provenance to illuminate scientific reasoning [14], and more generally they are (3) facilitating scientific discourse by providing content and context for online dialog in virtual communities.

However, these improvements are mainly impacting the online use of scientific data and methods, while the surrounding knowledge, the theory, assumptions, reasoning and other context, have largely been left behind. This is highlighted by the position of ontologies in the infrastructures where they are frequently shuffled to the background. Indeed, ontologies are rarely used directly by scientists, despite the potential for them to help represent knowledge that might otherwise seem to be absent. Instead, they are more often directly used by computers to enable automated components of the infrastructure to work properly. This raises outstanding questions about how effectively ontologies are being used to innovate knowledge from their background position in the infrastructure.

We suggest that ontologies are underutilized in the development of new scientific knowledge in each of the three aspects above. This is largely due to the fact that—for the most part—ontologies are being treated as engineering artifacts required to execute tasks more efficiently, rather than knowledge artifacts that, for example, help to describe some gap in scientific theory or flaw in the reasoning. Indeed, we claim e-Science ontology use is at present largely motivated by operational efficiency, with downstream impacts on scientific knowledge development minimized at present, and significantly below their potential. A contrasting vision prioritizes knowledge innovation in which scientists use ontologies both to express hypotheses, theories and models, and also to generate and test them [4, 18]. In this aspirational vision, scientists use ontologies directly as part of routine scientific investigation because the e-Science environments are designed to facilitate this. Such direct scientist interaction with the ontology-enabled knowledge, i.e. 'in-silico' semantic science, should then help revitalize online scientific methodology by helping generate richer insights, and improving our ability to repeat, report, and validate scientific findings.

## 2. Resource Quantity

The focus on operational efficiency is best exemplified by the quantity aspect (described in 1 above), in which significantly more online resources can be marshaled and then applied to some task. This usually involves ontology-enabled semantic interoperability to connect greater volumes of data, software, instruments, and computing resources. The associated ontologies typically consist of application ontologies that describe particular resources, or a slightly more general domain ontology that spans the application ontologies and serves as a unifying conceptualization for the system [16, 21]. However, neither of these ontology types typically encapsulates broad domain knowledge, as each tends to include only those concepts needed to enable the interoperability of specific resources. These ontologies are seldom even seen by scientists and they mainly remain part of black-box components that allow the system to automatically handle greater volumes of resources than could perhaps be handled manually. Even when the ontologies are seen by scientists, for example in query interfaces used to search distributed databases, the focus is on efficient retrieval of resources. Knowledge innovation is thus tied to insights gleaned by scientists from greater and faster resource retrieval and integration, rather than deeper understanding of the resources. Arguably, this often does not involve the application of new online scientific methods, but rather the mirroring of manual methods within the online environment, such that conventional lines of reasoning are carried out online by scientists. While this is certainly leading to new scientific results, there remains the real possibility that dramatic new insights might be achieved with complementary lines of investigation that involve increased use of machine techniques related to learning, analogical and abductive reasoning, data mining, and so on, that are starting to be applied to scientific discovery [5, 13].

The contrasting vision would thus leverage ontologies and more automated methods to facilitate the

proposal of new hypotheses as well as offer mechanisms to test their validity. The role of the scientist is not diminished but the system plays a greater and more direct role in knowledge innovation as some tasks are automated, like resource comparison and evaluation, and as new avenues of investigation are recommended to the scientist. Ontologies also play a greater role, because as authoritative representations of domain knowledge they become key expressions of the inputs and outputs of the research, which causes them to be consulted and updated regularly. The ontologies then constitute a far richer knowledge repository for a domain, and consist of theories, models, methods, and other artifacts of scientific work. This is a significant advance on present ontology contents, which primarily contain scientific categories such as 'granite', 'mass', 'temperature', and 'melanoma'. It is also a significant shift in ontology use as ontologies would be deployed directly by scientists, as well as machines, in all stages of knowledge discovery.

## 3. Representation

The representation aspect is probably best exemplified by the role of ontologies in online scientific provenance [11, 19], where ontologies are used to represent many aspects of scientific investigation. Scientific provenance refers to the historical context surrounding some scientific activity or result, and typically involves a description of the methods and applications used, the processes and reasoning steps carried out by a scientist for some purpose, and the old as well as new states of knowledge and data [20]. It is most widely encountered in established scientific workflow environments, such as myExperiment [7], which orchestrate scientific processing and from which the provenance elements can be readily obtained. While traditional best practices would necessarily have such process information recorded manually, online environments allow this to happen transparently by recording each operation as it occurs, and also recording it more finely so that each step can be captured, repeated and questioned.

Ontologies are widely used in provenance systems. They serve as common conceptualizations in the query interface for viewing and querying provenance, and for semantic interoperability across various data and provenance stores [11]. They are also used to annotate metadata associated with components in a scientific workflow [10, 11], and underpin trust systems that evaluate the quality and reliability of a scientific resource [2]. However, as with the quantity factor, such ontology use primarily has an efficiency imperative: more often than not the ontologies are used to describe low-level system resources such as a web service interface or a specific data product, rather than scientific objectives such as the hypothesis being tested or the reasoning used. The ontologies are thus primarily used to make the provenance system work, but how this affects knowledge generation is left to the scientist to determine. Our vision of provenance extends these notions to include ontologies of scientific method and reasoning, such that online processing steps can be understood in terms of scientific objectives, for instance to verify a result or evaluate a hypothesis. A particular workflow could thus be described in terms of system operations as well as scientific reasoning steps, so that scientists could interact with the workflow in terms of scientific goals as well as system mechanics. This necessarily involves a conceptualization of the general science knowledge cycle as well as effective interfaces and functions to operate over it.

## 4. Communication

The communication factor is best exemplified by online scientific collaboratories in which scientists utilize multi-media and social networking resources to work together on common tasks [17]. The general intent is scientific progress through increased scientific interaction, with a particular focus on augmented and clearer online discourse. Tools to represent and search scientific discourses are usually coupled to literature repositories or other resources, which provide subject matter for the discourse. Ontologies are used to represent concepts inherent in the discourse, including discourse concepts and scientific domain concepts, and these are often realized as annotations to papers in the literature repositories. The emphasis, though, is on the nature of the rhetoric surrounding some knowledge [3, 14] and on the validity of a given line of reasoning typically within a descriptive logic, with far less focus on the representation and evolution of higher-order scientific concepts such as theories and models. Again, this can be largely viewed as a gain in efficiency in that scientific statements indexed against rhetorical or basic domain concepts can be more readily found, likely in semantically annotated repositories [15], and more scientists are able to collaborate more often. It can also be

viewed as a marginal gain in knowledge interpretation as the knowledge is parsed into relatively simple structures, which nevertheless are critically evaluated such that inconsistencies in the reasoning are identified and conceptual gaps are highlighted. Critical evaluation might include the proposal of new hypotheses, but discourse systems on their own do not enable those hypotheses to be tested in a scientific sense, against data using established methods; at least not without being coupled to additional resources such as workflows, databases, instruments, and so forth. Our vision would see that coupling take place, such that dynamic hypothesis generation and testing could occur on deeper knowledge structures during online scientific discourse, where it could be tracked as well as evaluated for trust and eventual re-use.

## 5. Challenges

The vision of a scientific semantic web, in which ontologies drive science knowledge discovery, comes with many significant challenges related to capturing, designing, and using ontologies:

(1) **Ontology capture**: although some domains such as biomedical are routinely evolving ontologies, the vast bulk of science knowledge exists in growing literature repositories from which ontologies are absent and must be captured. At present, existing automated and semi-automated techniques for ontology extraction are limited to the capture of relatively simple science concepts and shallow structures explicit in the text, such as domain terms and large rhetorical blocks. A serious challenge is the capture of complex concepts and deep structures often implicit in the text, such as theories and lines of reasoning, and automation of this capture to deal with the large volume of source material. However, it is likely that techniques to capture knowledge as it develops within workflows will be more effective than those geared towards extraction form texts produced after some experiment has been completed, because the former contains more sources of context and more opportunities for direct interaction with the researchers. The design, management, and interoperability of related science knowledge repositories is a related concern.

(2) **Ontology design**: challenges for ontology design include the development of guidelines, design patterns, and formal methods for the construction and evaluation of ontologies within and across science domains. At present, general ontology engineering approaches are being successfully adapted to help domain ontology construction, but largely without recourse to general knowledge elements common to science. The main hurdle to overcome involves tuning these established techniques specifically to science domains, taking into account commonalities and differences. This further requires careful work to build general ontologies of science, which should lead to more consistent and coherent ontologies within domains, and facilitate connectivity across domains by providing a unifying upper-level of generic concepts such as theory, data, model, induction, method, experiment, and so on. Significant challenges also abound concerning how scientists might collaborate on the development of these elements, such as theories that are shared, overlapping, or in conflict, and how online resources can aid in the resolution of knowledge disputes.

(3) **Ontology use**: perhaps the holy grail of semantic e-Science is the quest for online (semi-) automated knowledge discovery. This requires a combination of human and computer methods to analyze and compare ontology-driven knowledge elements, such as theories and models, and to propose and test knowledge gaps. Coordinating deductive, inductive, and abductive reasoning in workflows operating over distributed online resources is an important part of this challenge. The development of user-friendly query and browsing interfaces attuned to a large-scale science knowledge framework is another significant challenge that must be overcome to ensure the framework will be usable.

## 6. Conclusions

Our conception of semantic e-Science amalgamates the enhanced visions discussed above. It includes semantic repositories of knowledge in which ontologies are a base representation for scientific concepts, theories, models, methods, and other science knowledge elements. These are coupled to workflow operations driven by scientific objectives and methods, and to scientific provenance described in terms of scientific reasoning steps. Finally, scientific collaboratories enable community discourse to occur over any of the previously mentioned components, to evaluate them for quality, trust, veracity and re-usability. In such an online environment scientists would focus on knowledge innovation, in as trans-

parent a way as possible, harnessing both efficiency and innovation objectives for next generation science.

*Acknowledgements*

We gratefully acknowledge of the support of the Geological Survey of Canada and the Ministry of Research, Science and Technology, New Zealand, and kindly thank the editors and reviewers for their suggestions which led to improvements in the manuscript: Krzyszof Janowicz, Pascal Hitzler, Michel Dumontier, Manfred Hauswirth.

## References

[1] Alverson, K. (2008). Filling the gaps in GOOS. The Journal of Ocean Technology, 3(3):19-23.

[2] Artz, D., Gil, Y. (2007). A Survey of Trust in Computer Science and the Semantic Web. Journal of Web Semantics, 5(2):58-71.

[3] Blake C. (2010). Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. Journal of Biomedical Informatics, 43(2):173-89.

[4] Brodaric, B., Reitsma, F. Qiang, Y. (2008) SKIing with DOLCE: toward an e-Science Knowledge Infrastructure. In: Eschenbach, C., Gruninger, M., (Eds.) Formal Ontology in Information Systems, Proceedings of the Fifth International Conference (FOIS08), IOS Press, 208-219.

[5] Colton, S., Steel, G. (1999). Artificial Intelligence and Scientific Creativity. Artificial Intelligence and the Study of Behaviour Quarterly, Vol. 102, 1999.

[6] De Roure, D., Gil, Y., Hendler, J. (2004). E-Science. IEEE Intelligent Systems, 19(1) :24-25.

[7] De Roure, D., Goble, C., Stevens, R. (2009). The design and realisation of the myExperiment Virtual Research Environment for social sharing of workflows. Future Generation Computer Systems, 25(5):561-567.

[8] Fox, P., McGuinness, D. L., Cinquini, L., West, P., Garcia, J., Benedict, J. L., Middleton, D. (2009). Ontology-supported scientific data frameworks: The Virtual Solar-Terrestrial Observatory experience. Computers and Geosciences, 35(4): 724-738.

[9] Gahegan, M., Luo, J., Weaver S.D., Pike, W., Banchuen, T. (2009). Connecting GEON: Making sense of the myriad resources, researchers and concepts that comprise a geoscience cyberinfrastructure. Computers and Geosciences, 35(4):836–854.

[10] Gil, Y. (2009), From Data to Knowledge to Discoveries: Scientific Workflows and Artificial Intelligence. Scientific Programming, 17(3).

[11] Golbeck, J., Hendler, J. (2008). A semantic web approach to the provenance challenge. Concurrency Computation Practice and Experience, 20(5):431-439.

[12] Hede K . (2010). In silico research: pushing it into the mainstream. Journal of the National Cancer Institute, 102(4):217-219.

[13] Langley, P. (2000). The computational support of scientific discovery. International Journal of Human-Computer Studies, 53(3):393-410.

[14] Li, G., Uren, V., Motta, E., Shum, S.B., Domingue, J. (2002). ClaiMaker: Weaving a Semantic Web of Research Papers. Lecture Notes In Computer Science, Vol. 2342, Proceedings of the First International Semantic Web Conference on The Semantic Web, pp. 436 – 441.

[15] Novacek, V., Groza, T., Handschuh, S. Decker, S. (2010). CORAAL -- Dive into Publications, Bathe in the Knowledge. Journal of Web Semantics, 8(2-3):176-181.

[16] Noy, N.F. (2004) Semantic integration: A survey of ontology-based approaches. SIGMOD Record, 33(4):65-70.

[17] Olson, G.M., Zimmerman, A., Bos, N. (Eds.) (2008). Scientific Collaboration on the Internet. MIT Press, Cambridge, MA, 432 pp.

[18] Poole, D., Smyth, C., Sharma, R. (2009). Ontology Design for Scientific Theories that Make Probabilistic Predictions. IEEE Intelligent Systems, 24(1):27-.36.

[19] Sahoo, S.S., Sheth, A., Henson, C. (2008) Semantic provenance for eScience: Managing the deluge of scientific data. IEEE Internet Computing, 12(4):46-54.

[20] Simmhan, Y.L., Plale, B., Gannon, D. A Survey of Data Provenance in e-Science. SIGMOD Record, 34(3):31-36.

[21] Wache, H., Vogele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., Hubner, S. (2001). Ontology-based integration of information - a survey of existing approaches. In: Proceedings of the IJCAI'01: 17th International Joint Conferences on Artificial Intelligence. Seattle, WA, pp. 108–117.