2010 Sixth IEEE International Conference on e–Science

A Semantic eScience Platform for Chemistry

Mark Borkum*, Carl Lagoze†, Jeremy Frey* and Simon Coles*
*School of Chemistry, University of Southampton, Southampton, UK
†Information Science, Cornell University, Ithaca, NY USA

***Abstract*—The oreChem project, funded by Microsoft Research, is investigating the design and deployment of a semantic-based eScience infrastructure for chemistry. The results of the project include the creation of an ontology that provides the basis for describing the entities and relationships for a scientific experiment, and the implementation of a workflow to combine extracted and in situ information from multiple sources, which provides a framework for computational enhancement of the data and querying thereof.**

I. Introduction

Advances in computing, sensor, and network technologies are changing the nature of science and scholarship by putting an ever-increasing quantity and diversity of data in the hands of researchers. This has been called the "Fourth Paradigm" of science [1]. The opportunities offered by this *data-intensive science* are huge, providing new capacities to address complex, long-term, large-scale, and multi-disciplinary challenges such as understanding and modelling climate change.

A survey across multiple scientific disciplines reveals the scope of this transformation and its impact on the manner in which research is undertaken. Notable examples include the virtual observatories in astronomy [2, 3], new techniques for the synthesis and visualisation of ecological observational data [4], and innovative efforts to improve preventative health care and enhance epidemiological knowledge by mining health records [5].

Notably, scientific fields differ in their level of receptiveness to these innovations [6, 7]. For example, various fields of physics, mathematics, and quantitative biology were enthusiastic adopters of even the earliest manifestations of change [8]. However, for a variety of reasons, which are summarised in this paper, chemistry in general [and especially experimental (in contrast to theoretical) chemistry] has been comparatively slow in its uptake of these recent innovations in data sharing and data publishing. The oreChem project, funded by Microsoft Research and the context for the work described in this paper, was created as a vehicle for exploring the mechanisms that might jump-start the adoption of these innovations in chemistry scholarship. It is a collaboration between chemistry scholars and computer and information scientists to develop and deploy the infrastructure, services, and applications that are necessary to enable new models for research and dissemination of the scholarly results of chemistry research. Work undertaken within oreChem includes the definition of various ontologies to represent entities of chemistry scholarship and the development and refinement of methods for retrospective extraction of semantic information from chemistry publications [9].

This paper describes one thread of work within the oreChem project; to facilitate the capture and dissemination of the methodology (planned method) of scientific experiments such that all derived and reported results (data products) can be discovered and reused in the correct context.

The key implication of this approach is that it enables both the informal inspection and formal replication of the methodology by independent parties, which in turn facilitates the systematic verification and validation of the derived and reported results [10]. This work is of vital importance, as it is the provenance information (also referred to as the lineage, pedigree or audit trail) of the derived results that ultimately determines the level of trust that can be invested in the reported results.

It is important to note that within the context of a scientific experiment, provenance information constitutes a record of *both* the methodology that was applied and the results that were obtained. Hence, there are two distinct types of provenance [11]: prospective and retrospective. Prospective provenance captures the structure of the methodology, which corresponds to the sequence of abstract operations (also referred to as activities or stages) that *will be realised* in order to generate one or many data products (or abstract classes of data product). Retrospective provenance captures the relationships that exist between each realisation and the original methodology, which corresponds to the sequence of concrete operations that *resulted in the realisation* of one or many data products, i.e., the causality.

We argue that there exists a cost/value proposition that must be evaluated for each data product, e.g., the costs associated with obtaining, understanding and verifying the provenance information; and the relative value of incorporating the knowledge into a new methodology. Providing a complete representation of the methodology along with an accurate description of each realisation (or enactment of the methodology), constitutes a significant reduction in these costs and hence increases the overall value of the data products. Furthermore, the presence of fixed representation of the methodology would allow for all results to be reported, including intermediates (which are often as important as the final results) and data products that are 'erroneous' for logical or socio-political reasons.

The remainder of this paper is structured as follows. The next section briefly summarises some aspects of the context of the technical work that is reported in the main body of this paper. Section III describes the novel ontology-based, prove-

978-0-7695-4290-4/10 $26.00 © 2010 IEEE
DOI 10.1109/eScience.2010.18

nance model that has been developed by the oreChem project. Section IV then includes a description of the workflow of our technical work, which integrates and enriches crystallography data from a variety of sources. The paper closes with Section V, enumerating some related work, Section VI, describing future work, and Section VII, which summarises future plans.

II. Technology in Context

While there are numerous examples of the adoption of innovative research methods and communication mechanisms throughout science and scholarship, these initiatives have failed to reach critical mass and have not become a core component of the chemistry research process. Outside of specific sub-fields, such as cheminformatics and crystallography, few chemists seem to perceive these developments as opportunities to enhance their own research and communication practices [12]. We might naively assume that adoption of new models of research and communication in chemistry is merely delayed in comparison to other disciplines. However, such simplistic assumptions of 'technical determinism' or inevitability have fallen out of favour with those who study scientific communications systems [13]. Instead, current thinking on the transformation of scholarly practices emphasises their historical contingency and notes that they are social [as much as technical] arrangements, where the social and technical aspects mutually shape one another [14–16].

A full discussion of this issue is beyond the scope of this paper. The interested reader is directed to a white paper and/or corresponding summary article [17, 18] that were the results of an NSF funded workshop held in 2009, which was led by and included several members of the oreChem project.

However, some of the results of that workshop and its full investigation are germane to the technical results that are the core of this paper. Furthermore, they are instructive to both eScience and cyberinfrastructure efforts in general, which to their peril have at times failed to recognise the socio-technical complexity of this type of work.

The problem in chemistry is not one of general resistance to technical innovation. Historically, the chemistry community and professional societies have been receptive to selected research and communication innovations. In fact, as early as pre-Web times in the 1980s, the American Chemical Society (ACS) was one of the first publishers to experiment with electronic versions of research articles [19]. Furthermore, there have recently been a number of Web-based innovations and experiments that enhance the collection, communication, and management of chemical information. These include efforts to leverage Semantic Web technologies in order to enable large-scale data mining, and to support drug discovery [20, 21], prototypes of electronic laboratory notebooks by proponents of open notebook science [22, 23], and initiatives to promote data publishing and access [24].

However, the pioneers of these efforts (and our own efforts in oreChem) have faced major hurdles to the widespread adoption of these techniques due to aspects of the chemistry research culture, the economics of chemistry data, and the proprietary regimes in which most chemical data reside, which restrict access to and the integration of chemical information on the Web.

The last point is particularly pertinent to our work within oreChem. An overwhelmingly large amount of chemistry data and publications are held by the world's largest scientific society, the ACS. Despite the fact that it is a non-profit organisation, at times very much behaves like a commercial entity. Particularly with regard to issues such as open access to the information services that it develops and offers.

The two most notable examples of this are its control over the Chemical Abstracts Service (CAS) and the proprietary nature of the dominant chemical identification system (the CAS number). It could be argued that the proprietary policy towards large-scale use of the identifier system undermines widespread experimentation and innovation by third parties that rely on the accurate integration of chemical information, such as the work undertaken by the oreChem project.

III. Data Model and Ontology

A core component of the oreChem infrastructure is a provenance model, the oreChem Core Ontology (CO). The model defines entities and relationships that describe both prospective and retrospective provenance, i.e., the description of the methodology of a scientific experiment; the realisation of methodologies; and the causality of data products. As a result, the CO facilitates queries over both the specification and outcome of methodologies.

The entities and relationships of the CO are depicted in Figure 1. At the centre of the model is the *plan stage* entity, which is an abstract description of an event that will occur during the realisation of a methodology. Within the context of a scientific experiment, a *plan stage* could represent the act of making an observation or taking a measurement; the process of synthesising a new chemical substance; or the requirement to invoke a specific software application or Web service. The *plan object* entity represents an artefact (or abstract class of artefacts) that will be consumed or generated during the realisation of a *plan stage*. Within the context of a scientific experiment, a *plan object* could represent an individual measurement; a specimen of a chemical substance; or a data-file with a specific quality (such as conformance to a content type or the presence of an XPath). Finally, the *plan* entity type is an aggregation of *plan stage* and *plan object* entity types, which are referenced using a 'contains' relationship.

The *plan stage* and *plan object* entity types are linked by four relationships: requires, emits, follows and derives. The 'requires' and 'emits' relationships assert pre- and post-conditions for the *plan stage*, i.e., that the realisation of the *plan stage* is not valid unless a specified *plan object* has been realised (as either input or output).

The 'follows' relationship asserts that the realisation of a *plan stage* must not occur unless a specified *plan stage* has been realised, and is inferred when two *plan stages* are linked to the same *plan object* using a pair of 'requires' and 'emits'

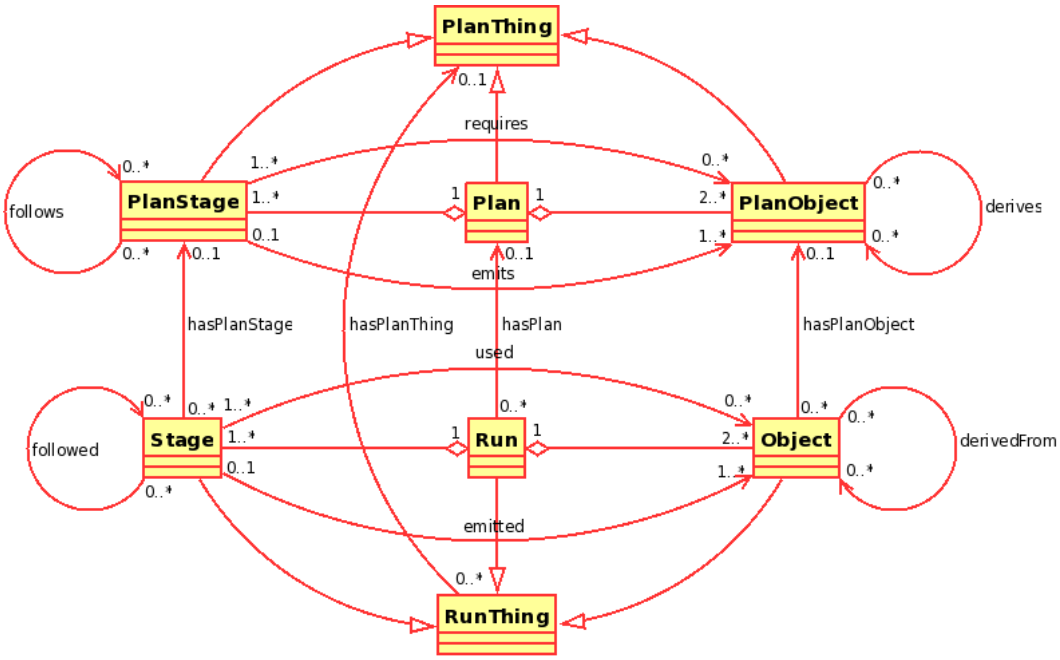


Fig. 1. UML class diagram for the oreChem Core Ontology (CO).

assertions. The 'follows' relationship can also be asserted explicitly in order to preserve the 'logical' order of events, e.g., to preserve the concept of a sequence of events.

The 'derives' relationship asserts that the realisation of a *plan object* will, at some point in the future, effect the realisation of a specified *plan object*, and is inferred when the requirements of a *plan stage* are distinct from the emissions. The 'derives' relationship can also be asserted explicitly, e.g., to describe compositions of physical objects. Hence, the *plan stage* entity type is a reification of the 'derives' relationship.

It is important to note that there is no logical equivalence between the realisation of a *plan object* and the notion of an information resource, i.e., while all realisations of the *plan object* entity type are themselves information resources, not all information resources should be represented in the methodology as a *plan object*. For example, within the context of a methodology that describes an experiment in physical chemistry, it would be reasonable to specialise the *plan object* entity type to describe a specimen of a specific chemical substance. However, it would not be beneficial to define an entity type that describes the abstract notion of the chemical substance itself. This is because, within the context of physical chemistry, the specimen is the first-order entity type and the chemical substance is the second-order entity type.

Within the context of a scientific experiment, a methodology is realised when it is enacted. In the CO, each 'enactment' is described by an instance of the *run* entity type, which references exactly one *plan* using a functional object property. Hence, a chain of retrospective provenance is established between the *run* and the *plan* such that for any enactment, it is possible to obtain the original methodology.

The *run* entity type is an aggregation of the *stage* and *object* entity types, which are themselves realisations of the *plan stage* and *plan object* entity types. The structure of a *run* mirrors that of a *plan*. Moreover, instances of the *stage* and *object* entity types reference exactly one instance of *plan stage* and *plan object* using functional object properties. Hence, a chain of retrospective provenance is established between each 'run thing' and 'plan thing', i.e., for any 'run thing', it is possible to obtain the original 'plan thing'. Furthermore, for any 'plan thing', it is possible to obtain a set of realised 'run things'.

Within the context of a *plan*, we label a *run* as 'satisfied' if there exists at least one realisation of every *plan stage* and *plan object*. The main advantage of this approach is it allows for the generation of arbitrarily complex aggregations, without placing a restriction on the number of times that a 'plan thing' can be realised. Furthermore, it allows for each *run* to be individually identified. Hence, the CO can be used to describe iterative processes. This is particularly useful with regard to the scientific method, which requires that one demonstrates repetition, e.g., when taking measurements or making observations.

The life cycle for 'run things' is depicted in Figure 2. The CO defines five time-stamp properties that can be asserted by 'run things', which are summarised in Table I: created, ready, started, finished, and destroyed.

The 'destroyed' time-stamp is provided in order to represent the concept of annihilation, i.e., when the physical artefact that is described by an information resource *ceases to exist*. The semantics are such that an *object* may not be used after the 'destroyed' time-stamp has been asserted. This is particularly

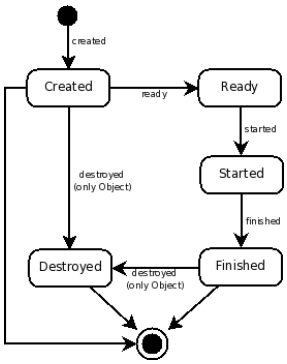


Fig. 2. Life cycle for *run*, *stage* and *object* entity types with respect to the assertion of oreChem time-stamp properties.

TABLE I
LIST OF TIME-STAMP PROPERTIES.

Time-stamp	Description
created	when resource was realised
ready	when resource was initialised (or 'ready for use')
started	when resource was first used
finished	when resource was last used
destroyed	when the artefact described by an *object* was annihilated (or 'ceased to exist')

relevant to the field of chemistry. For example, consider a *plan stage* that represents the classical aqueous acid-base reaction (acid + base → salt + water). In this scenario, the *plan stage* requires two specimens (an acid and a base) and emits a third specimen (a solution of a dissolved salt). While it is possible for 'future' *plan stages* to reference the information resources that describe the inputs, i.e., the instances of the 'specimen' entity type, it would be impossible to reuse the physical artefacts themselves, i.e., as the constituents of another chemical reaction, as they have ceased to exist.

IV. APPLICATION OF THE ORECHEM ONTOLOGY

In order to assess the effectiveness of our approach, we are implementing a workflow that integrates and enriches crystallography resources from multiple data sources, including: eCrystals[1] and CrystalEye[2].

In what follows, we describe the workflow we developed for integrating and enriching crystallography resources. In Section IV-B, we demonstrate the query capabilities that are provided by our data. Finally, in Section IV-C, we present an example of a visualisation technique that can be applied to our provenance data.

A. Crystallography Workflow

eCrystals is an institutional repository for crystal structures that are generated by the Southampton Chemical Crystallography Group and the EPSRC UK National Crystallography Service (NCS). *CrystalEye* is an aggregator for crystallography resources that is developed at the University of Cambridge.

The workflow is depicted in Figure 3. In the laboratory, scientists apply X-ray spectroscopy techniques to a specimen of an 'unknown' chemical substance, and 'raw' data is collected. The raw data is processed according to the NCS Crystal Structure Determination Workflow (CSDW) [24], and deposited into the eCrystals repository as a new record.

The CSDW is encoded as a *plan* and exposed as a machine-readable resource[3]. Each record in the eCrystals repository is described by a *run*, which references the *plan*. The *runs* are aggregated by a publicly accessible Web feed[4].

CrystalEye subscribes to the Web feeds of eCrystals and many open access journals (including Acta Crystallographica, the American Chemical Society, and the Royal Society of Chemistry). When a new resource is discovered, i.e., when a new entry is published to a Web feed, CrystalEye attempts to obtain, enrich and curate all available crystallographic data by consolidating it to a single Chemical Markup Language (CML) document [25]. The "complete" CML documents that are obtained for each journal are published in separate Web feeds, e.g., the feed of enriched eCrystals resources[5].

Finally, the *Computation* component subscribes to all available CrystalEye Web feeds. When a new resource is discovered, the Computation component further enriches the available crystallographic data using computational chemistry software – Gausian09[6]. The results of the computation are deposited into a triple-store.

B. Querying oreChem

The relationships provided by the ontology (see Section III) can be used to construct high-level provenance queries. Figure 4 illustrates an example of a query that operates over the retrospective oreChem provenance data in the eCrystals repository. The query is specified in SPARQL as follows:

```
PREFIX orechem:
  <http://www.openarchives.org/2010/05/24-orechem-core-ns#>
PREFIX ecrystals:
  <http://ecrystals.chem.soton.ac.uk/plan.rdf#>
SELECT ?run ?raw ?derived ?reported
WHERE {
  ?run a orechem:Run ;
       orechem:hasPlan ecrystals:Ecrystals ;
       orechem:containsObject ?raw ;
       orechem:containsObject ?derived ;
       orechem:containsObject ?reported .
  ?raw a orechem:File ;
       orechem:hasPlanObject ecrystals:HKL .
  ?derived a orechem:File ;
       orechem:derivedFrom ?raw .
  ?reported a orechem:File ;
       orechem:hasPlanObject ecrystals:CIF ;
       orechem:derivedFrom ?derived .
}
```

The query returns a set of 4-tuples of references to an oreChem *run* and three data files, which constitute the raw,

[1]http://ecrystals.chem.soton.ac.uk
[2]http://wwmm.ch.cam.ac.uk/crystaleye/
[3]http://ecrystals.chem.soton.ac.uk/plan.rdf
[4]http://ecrystals.chem.soton.ac.uk/cgi/latest_tool?output=Atom
[5]http://wwmm.ch.cam.ac.uk/crystaleye/summary/soton/ecrystals/2010/08-06/
[6]http://www.gaussian.com/g_prod/g09.htm

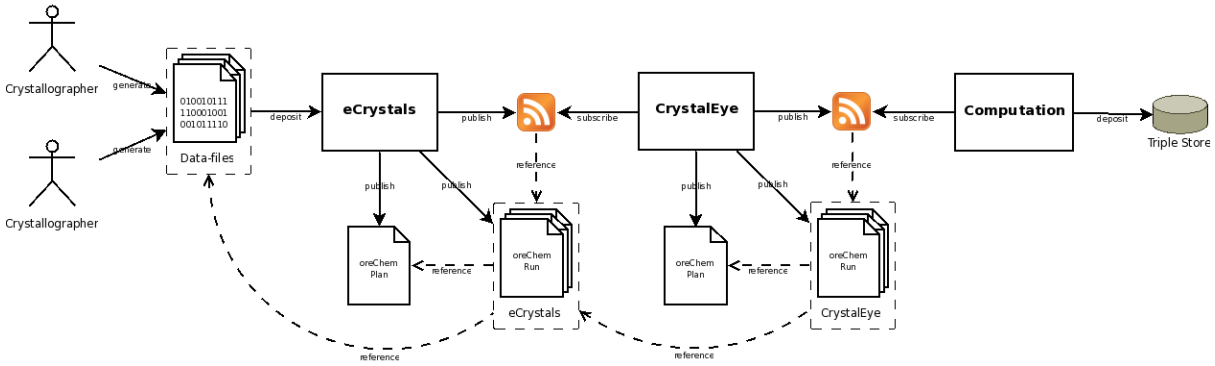


Fig. 3. Overview of the oreChem workflow. Scientists in the laboratory deposit raw and derived data into the eCrystals repository, which exposes a Web feed of oreChem *run* resources. The feed is monitored by CrystalEye, which republishes the derived data (in a separate Web feed) as a Chemical Markup Language (CML) document. The CML documents are enriched by the Computation component, which calculates computational chemistry information using Gaussian09 and deposits the results in to a triple-store.

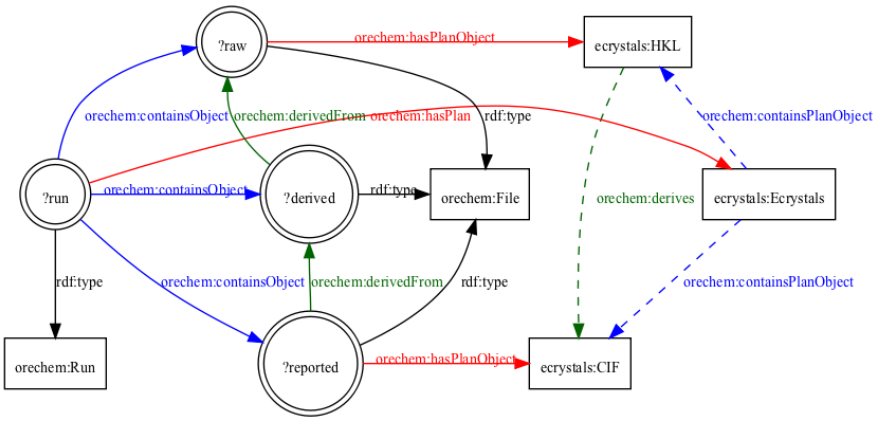


Fig. 4. Example provenance query for oreChem data in the eCrystals repository. The query returns a set of 4-tuples of references to an oreChem *run* and three data files, which constitute the 'raw', 'derived' and 'reported' data. The context of the derived data is specified by two transitive closures. The search space is restricted by references to two *object types* that are defined by the eCrystals *plan*. Dashed edges are asserted by or inferred from the *plan* (and are included for illustrative purposes).

derived and reported data. In the eCrystals repository, the 'raw' crystallography data is given by a *.hkl* file, which contains the diffraction data for a single crystal. The 'reported' crystal structure is given by a Crystallographic Information File (*.cif* file), which is the International Union of Crystallography (IUCr) standard format for representing crystallography data. The presence of the 'derived' data file is inferred by the use of two transitive closure operations, i.e., assertions of the 'derived from' relationship. The search space is restricted by the assertion of the 'has plan object' relationships.

C. Visualising oreChem

The provenance data is highly amenable to visualisation. We have implemented a plug-in for the eCrystals repository that exposes the retrospective provenance data for each record as an interactive graph.

The 'complete' graph, which displays all retrospective provenance data for a single enactment of the eCrystals *plan*, is given in Figure 5(a), and is composed of two types of node: rectangles and ellipses. Rectangles correspond to the execution of software applications, and are instances of the *stage* entity type. Ellipses correspond to data files, and are instances of the *object* entity type. Dashed and solid edges correspond to assertions of the 'used' and 'emitted' relationships respectively.

By applying inference rules to the assertions, two new graphs are obtained: a graph of the causal relationships between *stages* (assertions of the 'followed' relationship; depicted in Figure 5(b)), and a graph of the causal relationships between *objects* (assertions of the 'derived from' relationship; depicted in Figure 5(c)).

V. Discussion and Related Work

The goal of the Open Provenance Model (OPM) is to provide a standard model for communicating provenance information [26]. The OPM is defined by a core vocabulary and a set of rules, which define the inferences that can be made from provenance graphs. Many of the concepts that are modelled by the OPM can also be modelled by oreChem, e.g.,

the OPM "process" and "artefact" entity types are equivalent to the oreChem *stage* and *object* entity types. However, unlike the OPM, the oreChem Core Ontology does not contain an entity type to model the notion of an "agent" (the entity who enacts the scientific experiment, and consumes and generates resources). Both OPM and oreChem support the description of any "thing", whether it is a physical, digital or logical resource. However, the OPM does not model the concept of 'object annihilation' (see Section III).

The Scientific Workflow Provenance Data Model (SWPDM) aims to integrate the provenance information of computational workflows [27]. The SWPDM is defined by an upper ontology (an abstract model of a workflow) that is specialised by each workflow system. The advantage of this approach is that, in order to achieve interoperability, only one mapping is required between each workflow system and SWPDM.

VI. Future Work

There are several areas of work that we plan to explore in the future. An important goal of this work has been to provide a flexible and extensible data model for the description of methodologies, however, we have only explored those that are applicable to the fields of computational and physical chemistry (with particular regard to crystallography). In the future, we would like to explore the description of scientific experiments in other data-driven fields, e.g., ecology, social science, and high-energy physics. An important aspect of this work is the formulation of constructs for the description of 'aggregate' research objects, i.e., artefacts that are themselves composed of other artefacts. Specifically, we plan to specialise our ontology in order to describe the structure and usage of data-sets, which are themselves composed of multiple data-points. Other directions that we plan to pursue are the exploration of semantics for the description of observations and measurements, and indicative conditional branch statements, i.e., logical operations that act upon oreChem entities and govern the control flow of a methodology.

In many ways, this work resembles the concept of a computational workflow that is found in Taverna Workbench[7], Kepler[8] or VisTrails[9]. We believe that a workflow is an example of a *plan* that is enacted entirely *in silico*, i.e., that the description of a workflow is semantically-contained within the oreChem data model. In the future, we would like to expand upon this idea by specialising our ontology in order to provide an interoperability framework for computational workflow systems.

VII. Conclusion

Advances in eScience and cyberinfrastructure can provide researchers with new tools for undertaking, disseminating, manipulating, and understanding the research process and the products thereof. Arguably, among the most interesting applications of these tools is their potential to reveal the process and provenance by which results were derived. Visibility and transparency of process and provenance facilitates validity testing, repeatability, and full comprehension of research results.

In this paper we have described results of the oreChem project - an ontology-based, provenance model. By semantically distinguishing between the notion of a plan and an enactment of a plan, while simultaneously establishing clear mappings between the two, the model permits machine-readable representations of methodologies, results, and derivations. We have developed prototype implementations of this model using data from the open access eCrystals repository and shown its utility for provenance-focused queries. Our future work plan includes exploring the utility of this model in disciplines outside of chemistry and understanding and codifying the relationship of this model to other process and data models under development for eScience.

[7]http://www.taverna.org.uk

[8]https://kepler-project.org

[9]http://www.vistrails.org

Acknowledgment

The authors would like to thank Microsoft Research for their generous funding of this project. The strong support of Tony Hey, Lee Dirks, Alex Wade, and Savas Parastatidis is also greatly appredicated. Additional support for related activities was provided by the National Science Foundation through grant IIS-738543 SGER.

The authors would also like to those who have contributed to this project (in alphabetical order): Nico Adams, William Brouwer, Rameswara Sashi Kiran Challa, Nick Day, Jim Downing, C. Lee Giles, Na Li, Prasenjit Mitra, Karl Mueller, Peter Murray-Rust, Marlon Pierce, Joe Townsend, and Theresa Velden.

The content of this paper does not constitute the views of Microsoft Research or the National Science Foundation.

References

[1] "The Fourth Paradigm", Microsoft External Research, Redmond, WA USA, 2009.

[2] P.J. Quinn, D.G. Barnes, I. Csabai, C. Cui, F. Genova, B. Hanisch, A. Kembhavi, S.C. Kim, A. Lawrence, O. Malkov, et al., "The International Virtual Observatory Alliance: recent technical developments and the road ahead" in *Proceedings of SPIE*, vol. 5493, pp. 137, 2004.

[3] A. Szalay and J. Gray, "The world-wide telescope", *Science*, vol. 293, no. 5537, pp. 2037, 2001.

[4] W.M. Hochachka, R. Caruana, D. Fink, A.R.T. Munson, M. Riedewald, D. Sorokina, and S. Kelling, "Data-mining discovery of pattern and process in ecological systems", *Information*, vol. 71, no. 7, 2007.

[5] T.D. Wang, C. Plaisant, A.J. Quinn, R. Stanchak, S. Murphy, and B. Shneiderman, "Aligning temporal data by sentinel events: discovering patterns in electronic health records", in *Proceeding of the 26th annual SIGCHI conference on Human Factors in Computing Systems*, pp. 457–466, 2008.

[6] B. Cronin, "Scholarly communication and epistemic cultures", *New Review of Academic Librarianship*, vol. 9, no. 1, pp. 1–24, 2003.

[7] J. Fry and S. Talja, "The intellectual and social organization of academic fields and the shaping of digital resources", vol. 33, no. 2, pp. 115, 2007.

[8] K. Gunnarsdottir, "Scientific journal publications: On the role of electronic preprint exchange in the distribution of scientific

literature", *Social Studies of Science*, vol. 35, no. 4, pp. 549, 2005.

[9] N. Li, L. Zhu, P. Mitra, K. Mueller, E. Poweleit, and C.L. Giles, "oreChem ChemXSeer: a semantic digital library for chemistry," *International Conference on Digital Libraries*, pp. 245-254, 2010.

[10] J.R. Helliwell, P.R. Strickland, and B. McMahon, "The role of quality in providing seamless access to information and data in e-science; the experience gained in crystallography", *Information Systems and Use*, vol. 26, no. 1, pp. 45–55, 2006.

[11] B. Clifford, I. Foster, M. Hategan, T. Stef-Praun, M. Wilde, and Y. Zhao, "Tracking provenance in a virtual data grid", *Concurrency and Computation: Practice and Experience*, vol. 20, no. 5, pp. 565–575, 2008.

[12] M. Todd, "Open access and open source in chemistry", *Chemistry Central Journal*, vol. 1, no. 1, Feb. 2007.

[13] N.A. Van House, B. Cronin, "Science and Technology Studies and Information Studies", *Annual review of Information Science and Technology*, vol. 38, no. 1, pp. 3–86, 2004.

[14] P.N. Edwards, S.J. Jackson, G.C. Bowker, and C.P. Knobel, "Understanding infrastructure: Dynamics, tensions, and design", *First Monday*, vol. 12, no. 6, 2007.

[15] R. Kling, G. McKim, and A. King, "A bit more to it: Scholarly communication forums as socio-technical interaction networks", *Journal of the American Society for Information Science and Technology*, vol. 54, no. 1, pp. 46–67, 2003.

[16] C.P. Lee, P. Dourish, and G. Mark, "The human infrastructure of cyberinfrastructure" in *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*, 2006.

[17] T. Velden and C. Lagoze, "White Paper: The Value of New Scientific Communication Models for Chemistry". Cornell University, Ithaca, NY USA, 2009.

[18] T. Velden and C. Lagoze, "Communicating Chemistry", *Nature Chemistry*, vol. 1, no. 1, pp. 673–678, 2009.

[19] L. Garson, "Communicating original research in chemistry and related sciences", *Acc. Chem. Res*, vol. 37, no. 3, pp. 141–148, 2004.

[20] E. Neumann, "Finding the critical path: applying the Semantic Web to drug discovery and development", *Drug Discovery World*, pp. 25–33, 2005.

[21] I.V. Tetko, "Computing chemistry on the web", *Drug Discovery Today*, vol. 10, no. 22, pp. 1497–1500, 2005.

[22] J. Bradley, "Enhancing Scientific Communication through Open Notebook Science", Drexel University Library, Philadelphia, PA USA, 2005.

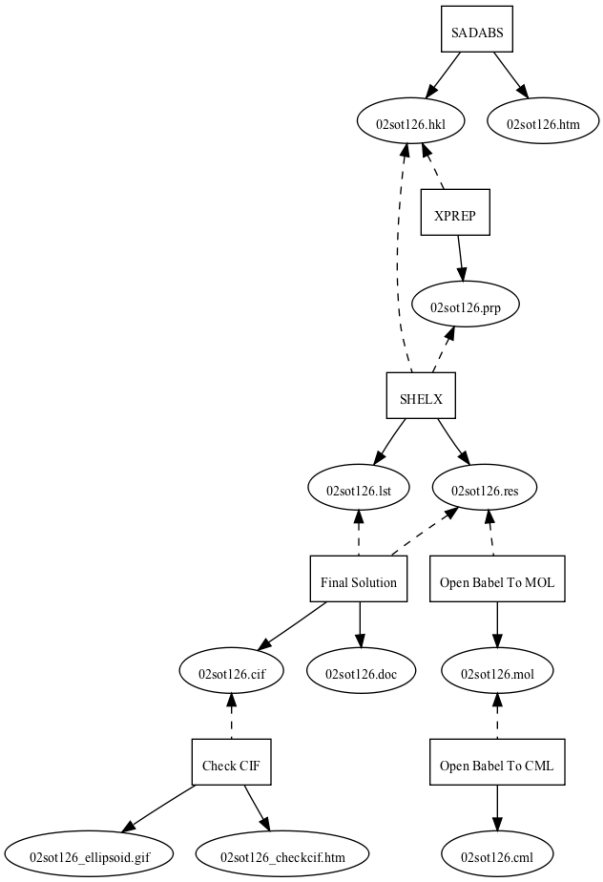
[23] G. Hughes, H.R. Mills, D. De Roure, J.G. Frey, L. Moreau, m.c. schraefel, G. Smith, and E. Zaluska, "The semantic smart laboratory: a system for supporting the chemical eScientist", *Organic & Biomolecular Chemistry*, vol. 2, no. 22, pp. 3284–3293, 2004.

[24] S.J. Coles, J.G. Frey, M.B. Hursthouse, M.E. Light, A.J. Milsted, L.A. Carr, D. De Roure, C.J. Gutteridge, H.R. Mills, K.E. Meacham, M. Surridge, E. Lyon, R. Heery, M. Duke, and M. Day, "An e-science environment for service crystallography from submission to dissemination", *Journal of Chemical Information and Modelling*, vol. 46, no. 3, pp. 1006–1016, 2006.

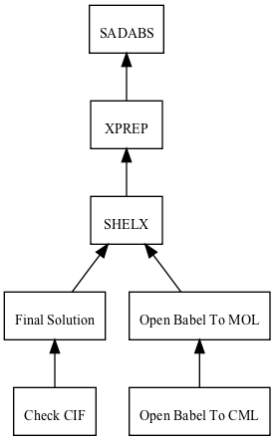
[25] P. Murray-Rust, "Chemical markup language", *World Wide Web Journal*, vol. 2, no. 4, pp. 135–147, 1997.

[26] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. Stephan, and J. Van den Bussche, "The open provenance model core specification (v1.1)", *Future Generation Computer Systems*, 2010 [Online]. Available: http://eprints.ecs.soton.ac.uk/21449/.

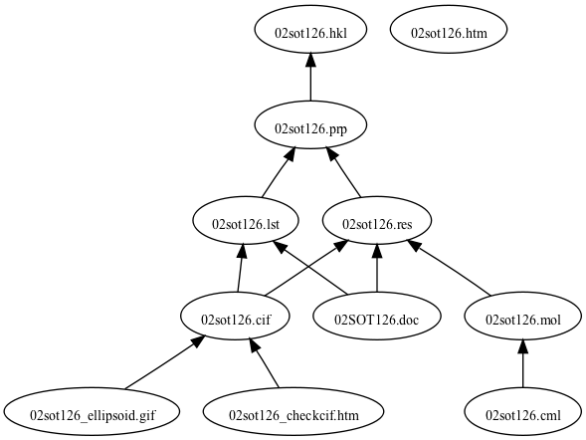
[27] T. Ellqvist, D. Koop, J. Freire, C. Silva, and L. Strömbäck, "Using Mediation to Achieve Provenance Interoperability", *Congress on Services - I*, pp. 291-298, 2009.



(a) Rectangles correspond to the execution of software applications, and are instances of the *stage* entity type. Ellipses correspond to data files, and are instances of the *object* entity type. Dashed and solid edges correspond to assertions of the 'used' and 'emitted' relationships respectively.



(b) Edges correspond to inferred assertions of the 'followed' relationship.



(c) Edges correspond to inferred assertions of the 'derived from' relationship.

Fig. 5. Retrospective provenance data for a record in the eCrystals repository for crystal structures: http://ecrystals.chem.soton.ac.uk/29/