# The oreChem Project: Integrating Chemistry Scholarship with the Semantic Web

**Carl Lagoze**
Information Science, Cornell University
lagoze@cs.cornell.edu

The oreChem project, funded by Microsoft, is a collaboration[1] between chemistry scholars and information scientists to develop and deploy the infrastructure, services, and applications to enable new models for research and dissemination of scholarly materials in the chemistry community. Although the focus of the project is chemistry, the work is being undertaken with an attention to general cyber infrastructure for eScience, thereby enabling the linkages among disciplines that are required to solve today's key scientific challenges such as global warming. A key aspect of this work, and a core aim of this project, is the design and implementation of an interoperability infrastructure that will allow chemistry scholars to share, reuse, manipulate, and enhance data that are located in repositories, databases, and Web services distributed across the network.

The foundations of this planned infrastructure are the specifications developed as part of the Open Archives Initiative-Object Reuse and Exchange (OAI-ORE) [9] effort. These specifications provide a data model [8] and set of serialization syntaxes [10-12] for describing and identifying *aggregations* of Web resources and describing the relationships among the resources that are constituents of aggregations. The OAI-ORE specifications are firmly grounded in the Web architecture [6] and in the principles of the semantic web [4, 7] and the Linked Data Effort [3]. The relevant connections of the OAI-ORE specifications to mainstream Web and Semantic Web architecture include:

- All aspects of data model are expressed in terms of resources, representations, URIs, and triples.
- The fundamental entity in the data model, the *Aggregation,* is a resource without a Representation (a "non-document" resource). This paradigm is similar to the manner in which real-world entities or concepts are included in the Web via the mechanisms proposed by the Linked Data Effort [3],
- The description of an Aggregation, a *Resource Map*, is a separate Resource, which is accessible via the URI of the Aggregation using the mechanisms defined for *Cool URIs* [15].
- The result of an HTTP access of a Resource Map URI is a serialization of the triples describing the Aggregation. This serialization may be in any of the OAI-ORE serialization syntaxes: RDF/XML [2], RDFa [1], and Atom [14] (triples can be extracted from this via an OAI-ORE defined GRDDL-compliant XSLT script).

---

[1] Collaborators in the oreChem Project are University of Cambridge (Peter Murray Rust, Jim Downing), Cornell University (Carl Lagoze, Theresa Velden), University of Indiana (Geoffrey Fox, Marlon Pierce), Penn State University (C. Lee Giles, Prasenjit Mitra, Karl Mueller), PuBChem (Steve Bryant), and University of Southampton (Jeremy Frey, Simon Coles).

Our initial work in the oreChem Project is the design of a graph-based object model that specializes the core OAI-ORE data model for the chemistry domain. This model builds on the centrality of the molecule, or chemical compound, in the record of chemistry scholarship. In the nature of a relational database key, a molecule or compound, identified in a universal manner [13], forms the central hub for linkages to other entities such as investigations, experiments, scholars, and processes related to that molecule. We are then using this model to design interfaces and APIs to exchange molecular information and their relationships among distributed repositories, services, and agents.

We are demonstrating this infrastructure by adapting a number of existing chemistry data repositories[2] to the APIs and models. We are also further populating these repositories by developing and refining automated techniques for retrospectively extracting chemical information and interlinking chemical data from existing chemistry research corpora. Following this we will develop and deploy a number of tools, such as chemical structure searching, over the repositories that have been adapted to the infrastructure. In the latter stages of the project, we will extend the retrospective data extraction techniques with active "in the lab" capture of chemistry data, and the addition of that "in-process" data to the knowledge network defined by the infrastructure data model.

Ultimately, we envision that this common data model, interchange protocols, and suite of data extraction and data capture tools will enable an eChemistry Web – a semantic graph with embedded subgraphs representing molecules which are then interrelated to publications that refer to them, experiments that work with them, the context of these experiments, the researchers working with these molecules, annotations about publications and experiments, and the like. A particularly interesting aspect of this semantic graph is the manner in which it mixes data, publication artifacts, and people – providing an information-rich social network built around the notion of object-centered sociality [5]. In the latter phases of the project we hope to build innovative analysis tools that will extract new "scientometric" information and knowledge from the eChemistry Web.

Our work in the oreChem Project and, in particular, our design of the interoperability infrastructure, is being undertaken with the recognition that chemistry, like any scholarly discipline, is not an island, but has complex linkages to scholarship in other disciplines and into related activities such as education, and in fact to the general network-based information environment. By basing our work on OAI-ORE, we hope that the interoperability paradigm designed for oreChem will coexist with similar work in other disciplines and in fact with the general Web information space and its ubiquitous search tools, services, and applications.

---

[2] These repositories include CrystalEye, 100,000 molecules and 100,000 fragments from crystal structures with full crystallographic details and with 3D coordinates; SPECTRaT, open theses with molecules; Pub3D, MMFF94-optimized 3D structures for PubChem compounds; Chem$_X$Seer, an integrated digital library and database allowing for intelligent search of documents in the chemistry domain and data obtained from chemical kinetics; eCrystals, high level crystal structures and processed x-ray diffraction data; and R4L, experimental spectroscopic and analytical chemical data.

**References**


1. Adida, B., Birbeck, M., McCarron, S. and Pemberton, S. *RDFa in XHTML: Syntax and Processing. A collection of attributes and processing rules for extending XHTML to support RDF*. W3C, 2008. http://www.w3.org/TR/2008/PR-rdfa-syntax-20080904/.
2. Beckett, D. and McBride, B. *RDF/XML Syntax Specification (Revised)*. W3C, 2004, http://www.w3.org/TR/rdf-syntax-grammar/.
3. Bizer, C., Cyganiak, R. and Heath, T. *How to Publish Linked Data on the Web*. Free University of Berlin, 2007.
4. Brickley, D. and Guha, R.V. *RDF Vocabulary Description Language 1.0: RDF Schema*. McBride, B. W3C, 2004. http://www.w3.org/TR/rdf-schema/.
5. Engestrom, J. *Why some social network services work and others don't — Or: the case for object-centered sociality* http://www.zengestrom.com/blog/2005/04/why_some_social.html. *Last Accessed on May 24, 2007.*
6. Jacobs, I. and Walsh, N. *Architecture of the World Wide Web*. W3C, 2004. http://www.w3.org/TR/2004/PR-webarch-20041105/.
7. Klyne, G. and Carroll, J.J. *Resource Description Framework (RDF): Concepts and Abstract Syntax*. W3C, 2004. http://www.w3.org/TR/rdf-concepts/.
8. Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R. and Warner, S. *Abstract Data Model*. Open Archives Initiative, 2008. http://www.openarchives.org/ore/1.0/datamodel.
9. Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R. and Warner, S. *ORE Specification and User Guide - Table of Contents*. Open Archives Initiative, 2008. http://www.openarchives.org/ore/1.0/toc.
10. Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R. and Warner, S. *Resource Map Implementation in Atom*. Open Archives Initiative, 2008. http://www.openarchives.org/ore/1.0/atom.
11. Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R. and Warner, S. *Resource Map Implementation in RDF/XML*. Open Archives Initiative, 2008. http://www.openarchives.org/ore/1.0/rdfxml.
12. Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R. and Warner, S. *Resource Map Implementation in RDFa*. Open Archives Initiative, 2008. http://www.openarchives.org/ore/1.0/rdfa.
13. McNaught, A. *The IUPAC International Chemical Identifier:InCh*, 2007, http://www.iupac.org/publications/ci/2006/2806/4_tools.html. *Last Accessed on September 18, 2007.*
14. Nottingham, M. and Sayre, R. *The Atom Syndication Format*. Network Working Group, Internet Engineering Task Force, 2005. http://tools.ietf.org/html/rfc4287.
15. Sauermann, L., Cyganiak, R. and Volkel, M. *Cool URIs for the Semantic Web*. World Wide Web sConsortium, 2007. http://www.w3.org/TR/cooluris/.