Proposal to Microsoft Corporation for the eChemistry Project

Principle Investigator: Carl Lagoze (Cornell University, lagoze@cs.cornell.edu)

Project Term: Two years – January 1, 2008 – December 31, 2009

Requested Budget: $999, 649 - $499,699 year 1, $499,999 year 2 (contingent on y1 results)

Table of Contents

Project Summary

We request $999,649 from Microsoft Corporation for a two-year project to develop and deploy the infrastructure, services, and applications that will enable new models for research and dissemination of scholarly materials in the Chemistry community – hereby referred to as *eChemistry*. A key aspect of these new models, and a core aim of this proposed project, is the design and implementation of an *interoperability infrastructure,* that will allow scholars to share, reuse, manipulate, and enhance data[1] that is located in repositories, databases, and web servers distributed across the network.

We envision that the eChemistry infrastructure will enable a number of innovations including new forms of publication that are more information-rich and interactive than static textual documents (Murray-Rust & Rzepa, 2004), applications that capture chemistry scholarship as it develops "in the lab", and environments that combine "Web 2.0" social networking tools such as blogs, wikis, and annotation tools with traditional artifacts of research. The social network-based tools will allow scholarly communities to collaboratively and openly build new information context and information resources from existing scholarship.

Our broader goal is to use Chemistry as an exemplar discipline, and via this project play a role in influencing the manner in which eScience, eScholarship, and cyberinfrastructure evolve across disciplines. We argue that the particular nature of Chemistry – its current reliance on traditional scholarly communication practices, its breadth and distribution over several sub-disciplines, its relationships with disciplines (e.g., physics) that have radically different scholarly communication "cultures", and the presence of strong advocates for change (most of whom are involved in this project) – makes it a particularly useful exemplar. We imagine that the results of this project, both technical and "cultural", will apply across the broader spectrum of disciplines and will permit us to leverage these results in future projects and funding initiatives.

In addition to this we foresee particular benefit to Microsoft via their funding of this project. In a general sense, this benefit will derive from the public perception of Microsoft as a strong advocate in this rapidly changing area, and in particular as a supporter of open standards to accomplish this. This aligns with existing activities and funding by Microsoft, and its sponsorship of conferences and publications in this area. But, we envision a more direct corporate benefit for Microsoft. Tools and applications marketed by Microsoft, such as the Office suite, are already an important component of scholarship. Recently, Microsoft has become more active at the product level in Web 2.0 social network technology. This project will provide the opportunity for integration of these existing and developing products into cutting edge standards and practices embedded in the transformation of the scholarly practices of a discipline. We hope to work closely with Microsoft researchers and product teams to provide this venue of application development and demonstration. Several intersections between this

[1] We use the term *data* in the broadest sense – including traditional text-based information resources plus images, binary data produced by experiments, and software such as computer-based simulations.

project and Microsoft products are mentioned in the body of this proposal. These are highlighted in **bold**. These are also summarized in a section at the end of the proposal.

The structure of the proposed project is as follows. Our initial effort will be the design of a graph-based object model that will form the core of the interoperability infrastructure. This model will build on the centrality of the *molecule*, chemical compound, or chemical investigation/experiment/process in the record of chemistry scholarship. The proposed object model will have facilities to link these molecular sub-graphs to the other entities –papers, researchers, experiments – in the eChemistry context. We then propose to design interfaces and APIs to exchange these models among distributed repositories, services, and agents. Following this model and infrastructure design we will demonstrate the infrastructure by adapting a number of existing chemistry data repositories to the APIs and models defined by the infrastructure and further populating these repositories by developing and refining automated techniques for retrospectively extracting chemical information and inter-linking chemical data from existing chemistry research corpora. We then plan to develop and deploy a number of tools, such as chemical structure searching, over the repositories that participate in the infrastructure. In latter stages of the project, we plan to 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 network provided by the infrastructure.

Ultimately, we envision that this common model, interchange protocols, and suite of data extraction and data capture tools will enable an *eChemistry web* – a semantic graph with embedded sub-graphs representing molecules, which are then inter-related to publications that refer to them, experiments that work with them, the contexts of these experiments, the researchers working with these compounds, annotations about these papers and experiments, and the like. In latter phases of the project we hope to build innovative analysis tools that will extract new information and knowledge from the eChemistry web.

The proposed interoperability infrastructure will be designed with the recognition that Chemistry, like any scholarly discipline, is not an island, but has complex linkages to scholarship in other disciplines and to related activities such as education, and in fact to the general network-based information environment. Thus, the interoperability paradigm proposed for eChemistry should generalize across multiple disciplines and information contexts, allowing reuse of scholarly artifacts across discipline boundaries. This is the vision that is proposed within cyberinfrastructure (National Science Foundation Cyberinfrastructure Panel, 2007), and other related eScholarship efforts.

With that context in mind, we plan to base the modeling aspect of our work with those developed within the Open Archives Initiative Object Reuse and Exchange Project (OAI-ORE)[2], which is developing standardized, interoperable, and machine-readable mechanisms to express information about compound information objects on the web (Warner et al., 2007). The OAI-ORE standards will make it possible for web clients, agents, and applications to reconstruct the boundaries of compound objects,

[2] http://openarchives.org/ore.

the relationships among their internal components, and their relationships to other resources in the web information space. This will provide the foundation for the development of value-added services for analysis, reuse, and re-composition of compound objects, especially in the areas of e-Science, e-Scholarship, and scholarly communication, which are the target areas of OAI-ORE.

OAI-ORE is currently funded for two years (September 2007 – August 2009) by a grant to Cornell University from the Andrew W. Mellon Foundation. The co-directors of OAI-ORE are Carl Lagoze of Cornell University, the principal investigator of this proposal, and Herbert Van de Sompel of the Los Alamos Research Library who will advise this project. OAI-ORE work is also funded by a separate grant from Microsoft, so-called *ORE Acceleration*, with the goal of producing an alpha specification of ORE interoperability models and protocols by end of September 2007 for use by this proposed project.

We plan to deploy the core ORE models as the basis for representing chemical compounds in the eChemistry context, with specific attention to selected sub-disciplines of Chemistry - crystallography, spectroscopy, computational chemistry, environmental chemistry, and chemical informatics and then demonstrate the utility of these models across distributed repositories, services, and applications.

Project Teams and Personnel

Institution	Personnel	Project Focus
Cambridge	**Peter Murray Rust** - Reader in Molecular Informatics at the University of Cambridge and Senior Research Fellow of Churchill College	• Retrospective Data Extraction • Searching and Indexing • Data Models/Ontologies • Tools and Applications
Cornell	**Carl Lagoze –** Senior Research Associate in Information Science. **Theresa Velden –** Ph.D. student in Information Science.	• Data Models • Interoperability infrastructure • Project Management • Publicity and outreach
Indiana	**Geoffrey Fox –** Professor of Informatics, Indiana University School of Informatics. **Marlon Pierce –** Assistant Director of Community Grids Laboratory of the Pervasive Technology Laboratories.	• Infrastructure Integration • Trust and Provenance • Tools and Applications • Scalable Data Services
LANL[3]	**Herbert Van de Sompel** - team leader of the Digital Library Research and Prototyping Team at the Research Library	• Data Models • Interoperability infrastructure
PuBChem[3]	**Steve Bryant** - Senior Investigator, in the Computational Biology Branch of the National Center for Biotechnology Information at NIH.	• Chemical Structure Archive • Results of Experimental Biological Activity Testing • Cross References to BioMedical Databases
Penn State	**C. Lee Giles** - David Reese Professor at the College of Information Sciences and Technology. **Prasenjit Mitra** –Assistant Professor, School of Information Sciences and Technology. **Karl Mueller** - Associate Professor of Chemistry.	• Retrospective Data Extraction • Searching and Indexing • Analysis
Southampton	**Jeremy Frey -**Professor in Chemistry. **Simon Coles –** Manager of the UK National Crystallography Service.	• Prospective & Retrospective Data Provision • Tools and Applications • In-process capture of data • Data Linking – in analysis & pub.

[3] LANL and PuBChem are not submitting budgets and will serve largely advisory and data storage roles.

Summary of Project Components

Specification of the core (discipline-independent) OAI-ORE data model

Members of the OAI-ORE technical committee and representatives of the semantic web community will produce an alpha specification of the core ORE model by September 30, 2007. This effort is supported by separate *ORE Acceleration* support from Microsoft to the OAI-ORE team, funding a face-to-face meeting and intensive team collaboration during the August/September 2007 time frame. This alpha model will provide the foundation for bootstrapping the eChemistry project. It is expected, however, that the model will be further refined throughout the course of the eChemistry project.

Specialization of the core ORE model for eChemistry

As decided at the April 2007 eChemistry meeting at Microsoft, the core of the eChemistry data layer will be a common model for representing molecular structures. This data model will build on the Core ORE data model and the work will involve extending the ORE ontology with eChemistry-specific ontology constructs. We will leverage constructs from the CML effort by project member Peter Murray-Rust (Murray-Rust, Mitchell, & Rzepa, 2005). We plan to formalize the CML dialect that the project will use (i.e. a 50% subset of the CML schema) and will also support common de-facto standards for chemical structure representation such as SMILES and SDF.

As described earlier, we envision that these molecular data models will be hubs in semantic networks, with linkages to papers, experiments, researchers, annotations, and the like. Enabling such rich knowledge networks will involve further ontology development for expressing the "external" linkages of molecules to other entities (e.g., papers, research groups, etc.). These semantic links have been explored in related semantic web work by project member Jeremy Frey (Coles, 2006; Taylor, Essex et al., 2006; Taylor, Gledhill et al., 2006), and we expect to leverage that work.

We note that modeling the relationships among diverse entities in this manner requires mechanisms for identification. Our decisions regarding identification will leverage a number of existing efforts including the linked data notions circulating in the W3C (Bizer, Cyganiak, & Heath, 2007), the use of DOI(s) for identifying publications (Paskin & Rust, 1999), the use of the International Chemical Identifier (InCHI) to identify and describe chemical structures[4], and the various efforts to develop and deploy a scholarly author identification mechanism (Dervos, 2006).

Specification of interfaces and APIs for access to eChemistry information

We plan to expose the functionality of the proposed infrastructure through service-oriented distributed computing principles (Booth, 2004). These systems are based on programming interfaces and associated over-the-wire message formats (typically in XML). Modern scalable systems are based on the W3C standards WSDL and SOAP, although the even simpler REST-style services (such as the Atom

[4] http://www.inchi.info/index.html

Publishing Protocol and XML Syndication Format) have widespread adoption in the Web 2.0 community. JavaScript Object Notation (JSON) and Microformats (standard extensions to XHTML) are also important message encoding mechanisms, so it is likely that our services will need to be multilingual. Web services are combined into composite applications as either workflows (Fox & Gannon, 2006) or mash-ups.

The Indiana University Project team will work with the other team members to design the service infrastructure, message exchange patterns, remote operations (such as put, retrieve, and discover), and message formats that will be necessary to implement the proposed work. The Indiana University team members will also support the implementation of the service infrastructure in collaboration with other team members. Extending current collaboration with the Murray-Rust group on OSCAR3 services, IU will work with the Cambridge and Penn State teams to adapt their document mining and related tools into online, programmable Web services. Various aspects of security (see below) will be addressed.

In parallel, the ORE team, led by Lagoze and Van de Sompel will develop interfaces and APIs for the exposure and transmission of ORE-based models. In earlier work (Bekaert et al., 2006) we explored the basic API functionality including *obtain, harvest,* and *register* transactions. In more recent work (Warner et al., 2007) we have explored the use of lightweight transmission vehicles such as Atom. The ORE team will collaborate with the UI team to integrate ORE interfaces with the broader service-oriented infrastructure.

Integration of Wikipedia MSDS ining chemistry research data

We will integrate a s into the proposed infrastructure, providing the foundation for new forms of information aggregations and mash-ups. At this point we plan to integrate the following (with the expectation that more will be added later as the project progresses):

Cambridge

- CrystalEye, 100,000 molecules and 100,000 fragments from crystal structures with full crystallographic details and with 3D coordinates. All single molecules and all fragments have an InChI and this can be used to link directly to Pubchem. MOPAC and GAMESS calculations are linked to many structures. We plan to forward to PubChem InChIs for any molecules not in stored there already.
- SPECTRaT, Open theses with molecules - number unknown but hopefully in the 1000-10000 range

Indiana University

- Pub3D: This service provides access to MMFF94 optimized 3D structures for PubChem compounds. Structures are returned in SD format and can be accessed by CID or by SMARTS patterns.
- PubDock: Provides methods to get the docked structures (for a given target) for PubChem compounds based on CID, sorted score values or by SMARTS patterns. Ligands are returned in

Add on PubChem

SDF format. Currently only the ligand structures are accessible and the actual score values are coming soon. This database is a useful initial filter in workflows for more sophisticated docking and scoring.

- PubChem Derived properties: Get calculated properties (SLogP and SMRef) given a compound ID. Can also search via exact values and ranges (but this is very slow at the moment).
- PubChem Structure: Provides methods to get Pubchem Compound information
- PubChem Synonyms: Provides methods to get synonyms given a compound or substance ID.
- Varuna; These services allow the submission and querying of Jaguar Quantum Mechanics and Molecular Mechanics data in our Varuna Database.
- Several additional services (such as online R statistical packages, OSCAR text analysis, tabular data operations, toxicity calculations) are also available.
- In addition, we collaborate with the developers of the CIMA crystallography grid project[5]. This project will provide access to data and metadata repositories for crystallography labs in the US, UK, and Australia. We have collaborated with CIMA on numerous research problems, including fine-grained data access control.

Penn State

$Chem_{x}Seer$ (chemxseer.ist.psu.edu), an integrated digital library and database allowing for intelligent search of documents in the chemistry domain and data obtained from chemical kinetics. This system is being developed with funding (2005-2010) from the National Science Foundation. Includes the following data sources:

- Chemical Entity Search: This tool identifies chemical formulae and chemical names, disambiguates the terms from other general terms, and tags them. Novel similarity scores, ranking functions and search methods are used to enable searching for chemical entities.
- TableSeer: This tool automatically identifies tables in digital documents and extracts the contents in the cells of the tables. The contents are stored in a queryable table in a database. TableSeer extracts table metadata, and uses a novel ranking function to search for tables relevant to user queries.
- Databases: This data repository contains experimental data obtained from various sources. Our tools can process, store and link data in multiple formats, e.g., Excel, XML, Gaussian, and Charmm. A metadata ad-on can help annotate the data and link multiple datasets. The metadata is then used to link the data to published articles allow the end-user to search for relevant data.

[5] http://156.56.94.164:8080/gridsphere/gridsphere.

PubChem

PubChem will support the project by providing open access to structures/properties/identifiers within its archive. PubChem will also support linkages to bioactivity summaries within PubChem and to other biomedical databases cross-referenced to PubChem, such as PubMed biomedical article abstracts. Where the developers of other repositories wish to cross-reference with PubChem, PubChem will host (via deposited chemical substance records) links to the repositories hosted by other project participants.

Southampton

The collection of repositories at Southampton covers the data curation aspects of the whole scientific research lifecycle, exemplified by the following individual repositories:

- eCrystals, contains both the high level crystal structures and processed x-ray diffraction data. Contents will be linked to the raw image data in homecomb store.
- R4L repository, contains experimental data (spectroscopic, analytical chemical data)
- SHG experiment database together with laboratory environmental database,
- Semantically rich Electronic Laboratory Notebook records for synthetic chemistry and crystallography
- SHG Experiment Blogjet.
- Southampton and Stockholm surface tension data, and Sum Frequency.
- Southampton e-print system that contains talks, papers.
- Molecular Simulation (raw) data.

The use of ORE will enable us to link the content between these repositories and other external data repositories with molecules, experiment or researcher view.

Development of trust and provenance mechanisms

For adoption by chemists, crystallographers, and the chemical informatics community, the proposed infrastructure must protect the integrity, pedigree and privacy of both the data and the metadata. We envision three basic levels of privacy within our system: public, private, and protected. Public data is published information available without restriction (such as available from PubChem). Private data and metadata must be accessible only by the direct owners of the intellectual property. Protected data represents an intermediate stage, in which data is shared with collaborators or reviewers.

Much work in Cyberinfrastructure security has taken place, and standard techniques such as Public Key Infrastructure can provide the core security capabilities (authentication, privacy, integrity). The gaps in these systems include identity federation, ease of usage, and assurance. Work that must be done for this project includes

- Identity representation: we must be able to represent users' identities across virtual organizations. OpenID[6] is one candidate Web standard.
- Security federation: we will need to provide methods for federating multiple services' security requirements. Related work here includes Shibboleth[7].
- Ease of usage: our security system must be simple to use and administer.
- Access control: users own and control their data and have the ability to set the permissions. We will examine existing systems such as PERMIS (Chadwick, 2003; Yin, 2006)
- Auditing: we must go beyond authentication-only security systems to provide the user with information on the access (and attempted access) to his/her data. This will provide an additional level of security to detect system wide break-ins, unauthorized usage, etc.
- Authenticated reviewing, tagging, and commenting: the next generation of Cyberinfrastructure will need to follow Web 2.0 models of commenting and review. For example, protected data under review may need to be accessible by authenticated but anonymous reviewers with read but not write/edit privileges. Similarly, open data may be reviewed and rated by the general community. These may be anonymous or not, and in addition to authentication, reputation will need to be tracked.

Retrospective data extraction, storage, and access

As noted earlier we plan to populate repositories in the infrastructure via retrospective extraction of data from existing sources. The Penn State team therefore will develop tools to analyze multidimensional maps of spectroscopy data to detect interesting patterns. This type of analysis is of particular interest in environmental, analytical, and physical chemistry. As an example, imaging spectroscopy can be used to scan the structure of chemical elements on material surfaces. The data contains the intensity at a given wavelength at a two-dimensional position on the surface of interest. Elements have their unique and known spectral profiles, and these profiles can be obtained using a variety of spectroscopic and imaging techniques. We will use data mining methods to determine the elements present in a material and the composition of the material. Extracting a spectral profile is a difficult task because wavelengths of neighboring chemical elements may combine or overlap at a point.

For the first year, we propose to utilize principal component analysis (PCA) or other machine learning methods to detect patterns in the high dimension data. It will divide the data into principal components representing the uncorrelated chemical elements. We will explore with machine learning methods, like neural networks, to use training data to identify features in the spectra that correlate to the absence or presence of chemical elements. For the second year, these learned models can then be used on test data to classify the spectra and further refine the methods plus investigate extensions to other chemical

[6] http://openid.net/

[7] http://shibboleth.internet2.edu/

domains. Data will be extracted and available for the appropriate repositories, including storage of relevant data in the Chem$_X$Seer system at Penn State.

Tools for discovery of molecular structures

The Penn State group will extend their chemical entity search engine to handle structural search queries. When using a chemical structure database, end users want to query the database to find ``similar'' chemical compounds to the query molecule structure. The end-users' notion of similarity is heavily influenced by domain knowledge that most end-users in the scientific community possess. Yet, such domain knowledge is not present in the database. Notwithstanding novel work on indexing and querying graph databases (Cheng, Ng, & Lu, 2007; Xifeng Yan, 2005), a similarity ranking function that computes the similarity between chemical compounds only on the basis of their structures cannot return results that satisfy end-users' intuition. If there is an exact structure-based match or an exact substructure match (where the query is a substructure of a chemical structure in the database), existing techniques can efficiently answer the query. However, as discussed above, the interesting case happens when there is no exact match and a search for the ``nearest'' chemical structure present in the database has to be executed

Most of these structure queries are about structures of chemical compounds; the determination of crystallographic structure of inorganic compounds is beyond the scope of this work. Based on our interactions with chemists, we believe that if a chemist was manually answering a chemical structure query, they first recognize the backbone of an organic compound (e.g., linear, branched or cyclic) --- GString (Jiang, 2007) utilizes this information except that they refer to branched as ``star'', while we prefer to use the term ``branched'' as used by chemists) and the functional groups present in the chemical structure (GString does not consider this). Generally, the same functional group is expected to undergo similar chemical reaction(s) regardless of the overall size of the molecule, as long as the functional group is accessible (which we often refer to as “reactive”, in relation to, for example, reactive surface area in the environment). After determining the backbone, the chemists look for those functional groups in a chemical database and identify chemical structures with similar functional groups as similar to the chemical structure query.

In attempting to improve chemical structure queries, our first observation is that we must consider the functional groups both in the query and in the data in the database. Beyond the matching of linear paths, branches, and ring structures, a good similarity function must consider the similarity of the functional groups. The main features of our solution are as follows:

- While inserting a chemical structure in a database, extract the backbone and functional groups of the structure.
- Index the chemical structure (using both backbone and functional group information).
- When a query is posed to the system, extract the backbone and functional groups of the structure.

For the first year, we will design an algorithm for the first item above. Automatically segmenting a chemical structure to determine the backbone and the functional groups is a challenging but tractable

problem. We intend to utilize a rule-based and a machine learning based strategy. The rule-based component uses rules to determine the backbone and looks up common functional groups from tables of most common functional groups to match the chemical structure. For the second year we will extend GString to not only index the linear, branch, and "star" structures in the backbone but extend the string representation of the chemical structure used in GString to include the location and type of functional groups. The physical index used in GString is a suffix tree structure; experimentation will reveal if this structure is suitable or needs alteration to address more complex domain-knowledge issues.

For the second year we will extend GString to not only index the linear, branch, and star structures in the backbone but extend the string representation of the chemical structure used in GString to include the location and type of functional groups.

We will provide all extracted data as databases for inclusion in the appropriate chemical information repositories.

Other Tools and Applications

The Southampton group will demonstrate the utility of the infrastructure in a number of experiments. One proposed experiment is an SHG study on crown ethers which would link, crystal structures, solid forms, SHG data, surface tension data, simulation data, talks, papers etc sourced from various repositories. We will then generate a series of different views on the 'Crown ether experiment' –

- based on the molecule and comparison with similar molecules,
- one on an experiment that has multiple types of spectra and links to databases and blogs,
- one that has the project as a whole and links in more documents and presentations

Other similar cross-experimental studies are planned. We will demonstrate how we can link to data contained within one of the ORE-based repositories and use the data for analysis and comparison. Within Word, PPT, Excel and programs like the statistical package R, we aim to link to a data source within a repository, ensure that data link is maintained and can be globally resolved, and then the resulting analysis description, together with the data form part of a new object. We can investigate if the new object should have all the data from the 'linked to' object or only the data actually used, but with a link to the wider source for future investigations.

We would like to investigate to tools that we could use to 'bring data in' from a repository to ensure the metadata links are maintained and do not have to be added or reconstructed again later, in order to facilitate Smart Papers.

Community Acceptance & Involvement

Building on the community acceptance of the "eCrystals Repository" project for crystallographic data, the Southampton group will guide the development and deployment of the repository infrastructure such that it is an evolutionary development of current working & publication practices, demonstrating

how the new paradigms of data dissemination can work with the established needs for recognition, validation, long term curation, and industrial involvement.

- Separation of repository structure from services that use the data
- Understanding of local & public use of the same repository structures
- Understanding of the needs of embargo and publication
- The need to provide a adequate data report not simply placing “data on the web”
- The needs for persistent identifiers that are understood by the community
- The importance of data citation
- The importance of dealing with differing working practices between laboratories

Analysis of eChemistry information

Research in environmental chemistry is becoming increasingly collaborative and multidisciplinary in scope and approach. For example, within the Penn State Center for Environmental Kinetics Analysis (CEKA; see www.ceka.psu.edu), researchers are taking a multidisciplinary approach to linking kinetic information in environmental chemistry across spatial and temporal scales. A main goal of such research is to integrate experimental, analytical, and simulation results performed on systems from molecular to field scales in order to approximate the complex physical, chemical, and biological interactions controlling the fate and transport of contaminants better. New scientific questions can be generated when users have access to a broad spectrum of related results. As connections are made among field observations, experimental kinetics, spectroscopic analyses, and model predictions, gaps in the information web will become apparent. Approaches to filling these gaps can then be addressed by the collaborative team. An easily queried, intelligent eChemistry environment will provide access to critically relevant data for a diverse community of users, enabling these users to achieve higher order scientific goals. In short, data collection and synthesis will lead to better science and improved education of scientists. In the second and third years of the proposed work, we intend to integrate new eScience tools into current studies of environmental kinetics at Penn State University, and air/water interfacial studies at Southampton. We envision this project leveraging international collaborations where research proposals will be submitted simultaneously to the NSF and EPSRC.

Annual Project Goals and Deliverables

Year 1: September 1, 2007 – August 31, 2008

University of Cambridge

1. Robustify CML and CML dictionary ontologies in collaboration with the project so that everyone can use this as a standard vocabulary.

- Q1 - publish revised and frozen CML schema 2.6
- Q2-4 publish enhanced documentation for CML. Identify any (project-independent) revisions or upgrades to CML

2. Expose CrystalEye with static RDF consistent with the ontologies

- Q2: revision of ontology based on project feedback
- Q1-4: Continued releases of crystal Eye due to increased content

Cornell University

1. Publish alpha specification the of core ORE model (Q1)

The core data model, which is intended to be non-discipline specific includes notions of:

- Resource aggregation: a named set of identified resources
- Resource map: a serialized machine readable description of a resource aggregation

The model also includes a base vocabulary with notions of hasPart, hasView, hasVersion, plus resource types incorporated from the DCMI type vocabulary

Deliverable: Publically available specifications available via the OAI-ORE web site.

2. Extension of model for eChemistry semantics (Q2)

The core data model for ORE is non-discipline specific, but is intended to be extendable to the needs and vocabularies of multiple application contexts. Our goal in this aspect of the work is to extend the vocabulary in two dimensions

- Specific to the notions shared across scholarly communication and eScience. This includes ontological concepts such as *journal, paper, issue, dataset, citation,* etc. and relationships therein.
- Specific to the notions in eChemistry, as a specialized scholarly eScience application. As described earlier, our proposed work builds around the notion of the chemical compound as the fundamental unit of chemistry research. Examples of types and relationships in this context are *experiment, laboratory, instrument,* and the like.

We plan to work with members of the project team with chemistry vocabulary experience – specifically Frey and Murray-Rust to develop these vocabularies, formally specify them and demonstrate their use in the context of the ORE data model.

Deliverable: Publically available specifications and schema (RDFs) available via the OAI-ORE web site

3. Definition of transport alternatives for ORE models (Q3)

We are investigating various technologies for the serialization and transmission of instances of the ORE (core and extended) data model. These include the following:

- ATOM: This publishing protocol is well-adaptable to the needs of ORE, since it already includes the notion of a "compound object". In addition, ATOM and feed syndication is well-deployed technology, and leveraging it would benefit the utility of ORE to a wide variety of communities. We plan to publish as specification of how to specialize ATOM for ORE purposes. Our intention is to remain compliant with the ATOM specification (Nottingham & Sayre, 2005)
- RDF/XML: The ORE data is naturally congruent with the RDF data model, since our model of a resource aggregation is graph-based. We intend to publish a specification that maps the ORE data model to the RDF model, and develop an RDFs schema that specifies the RDF/XML mapping of that same model.
- OAI-PMH: This metadata harvesting format is widespread and an expression of ORE within it would be accessible to a variety of client communities that already have experience with PMH. We intend to publish a specification of the deployment of the ORE data model within OAI-PMH.

Note that the expressive power of these serialization/transport alternatives is not equivalent – some may only be able to represent limited aspects of the ORE model.

Deliverable**:** Publically available specifications and schema (RDFs, xml schema) available via the OAI-ORE web site.

4. Specification of ORE interface abstractions (Q3)

Our motivation is to make ORE and the eChemistry tools available to a wide variety of communities with varying levels of expertise. This is the motivation for a variety of serialization formats of the model, each with different levels of expressiveness, and subsequent ease of adoption. In this same spirit we plan varying levels of API or protocol based access to the operational semantics of the model. In this spirit, we have in the NSF-funded Pathways project (related to this effort) described three interface astractions:

- ***Obtain:*** Access an instance of the model representing a specific resource aggregation.
- ***Harvest:*** Batch access to multiple instances of the model.
- ***Register:*** Request deposit of an instance of the model in a hosting service (**originating from an authoring client such as Office**, or a repository such as DSpace or Fedora).

Inspired by work on the ATOM publishing protocol (Gregorio & Hora, 2007), we add the following two operational primitives:

- ***Delete*:** Remove a resource that is an aggregated resource
- ***Edit:*** Modify an instance of an aggregated resource.

Our goal in this work is to specify the operational semantics of these ORE interfaces independent of an implementation choice. This specification will describe access management issues related to each interface abstraction as well. This specification will then form the basis of later interface specification documents (e.g., ATOM publishing protocol, RESTful transactions, web service calls, etc.) with varying levels of expressiveness.

Deliverable**:** Publically available specifications on the OAI-ORE web site

5. Evaluate coordination of ORE technology with web services/grid environment (Q4)

This is follow-on to the work described above. Our plans within the core ORE work are to author implementation specifications based on REST and ATOM publishing protocol. Our eChemistry work mandates an implementation within the context of the grid architecture, specifically the web services expression of that (Tuecke et al., 2003). This implementation will incorporate notions of security, robustness, privacy, and other factors not included in the simpler implementations. We plan an exploratory paper that will outline the issues in transitioning ORE to this more complex implementation environment. The paper will provide the basis for related work in the second year of the project.

Deliverable**:** White paper, distributed internally to project group

6. Project management, technology coordination, and reporting (ongoing)

The Cornell group will have primary responsibility for project coordination, management, and reporting.

Deliverable**:** Quarterly reports

Indiana University

1. ORE exchange format (Q1)

Basic ORE metadata exchange format implemented in multiple format bindings: XML, Atom, JSON, Microformats. We will participate in the graph design process led by Cornell and will lead the multilingual support efforts.

Deliverables: technical specification of Atom, JSON, and Microformat bindings.

2. Message Integrity and Privacy (Q1)

Associated message integrity and privacy guarantees will be examined (CMU, IU, …): we will define extension points in the metadata that can be used to convey cryptographic signatures and message digest (one-way hash) information that can be used to verify the metadata's authenticity and integrity. We will also design the architectural security that will ensure secure message transmission.

Deliverables*:* technical specifications of ORE message extensions to address security.

3. Protocol Integration with Access Control (Q2)

Basic protocols/APIs for metadata publishing, searching, and retrieval defined and implemented with associated authentication and access control mechanisms (CMU, IU, …). We will assist Cornell in the definition and implementation of REST-style services based on best available practices, such as the Atom Publishing Protocol and the XML Syndication Format. IU's particular focus will be to a) implement multilingual services, and b) specify the extensions of these services to capture the strong authentication and access controls (read/write access to metadata) required by the eChemistry community. **We will implement prototypes using appropriate MS .NET tools.**

Deliverables*:* technical specification of protocols (with Cornell), software reference implementation of REST-style service.

4. Scaling (Q2)

Investigation of scaling issues in IU's PubChem-derived chemical databases (PubDock and Pub3D). **We will convert existing database infrastructure to use MS SQL Server and will work with the SQL Server team on scaling and performance. This will be critical in order to support stateless Web 2.0 services: query times must be kept to 10's of seconds.**

Deliverables: Databases implemented in MS SQL Server.

5. Service Integration (Q3)

Selected services (PubChem prototype, crystallography databases, biochemical databases, image analysis databases) adapted to support protocol and message formats (1 and 2 above). These will be determined at the project kickoff meeting and will be drawn (initially, for prototyping) from our NIH CICC collection of services.

Deliverables*:* demonstration

6. Use Case Demonstration (Q4)

Initial use cases for crystallography, biochemical/chemical informatics implemented and demonstrated. Use cases will be defined at the kickoff meeting and will be drawn initially from our CICC collection of services and from our CIMA collaboration.
Deliverables*:* demonstration

7. Social Networks (Q4)

Sample Web 2.0 social network and user client interfaces to ORE-described data products. **Will be based on appropriate MS Web 2.0 tools and frameworks such as Silverlight and Popfly**. Will be used to demonstrate basic capabilities of the system: access controls, publishing, searching.

Deliverables*:* Delivered web capability; will be used to support Q3 and Q4.1 demonstration deliverables.

Penn State University

1. Metadata for Chemical Kinetics Data via Excel

Environmental scientists enter the majority of the data on chemical kinetics using Microsoft Excel. This data will then be submitted to our database via a website. A moderator approves the acceptance of the data for depositing in a database. **At the server-side, we will employ a MS Windows-based machine to process the submitted Excel data.**

Although Excel is an excellent tool for chemists to use to enter their data, to the best of our knowledge, it does not have any standards on annotating the data with metadata that is useful for interpreting the data. **In this work, we will design and implement a macro using Excel to capture the metadata.** There will be several types of metadata that will be captured:

- file-level metadata that pertains to all worksheets within the file,
- worksheet-level metadata that provides metadata for each column in the worksheet,
- crosslinking metadata that provides information about the links between the different worksheets and between different parts of the worksheets, e.g., data in a row in worksheet 1 may be related to data in row 3 in worksheet 2.

The tasks involved in creating this macro involves (a) standardizing the metadata that is suitable for chemists and environmental kinetics researchers, (b) reusing existing standards like Dublin Core, ORE, and CML to make the metadata conformant to one or more of these standards. The exact choice of constructs and standards is an item of research and depends upon the types of data and metadata that the chemist wants to capture, (c) creating the client-side macro, (d) creating a server-side processing module that can extract these metadata and populate a database.

For successful deployment of the macro and for enabling its wide-spread use and acceptance, we would require to enable ability to reuse existing metadata from existing datafiles with minimum effort, because many datafiles may be generated from the same experiment or the same study. Consequently, the end-user does not want to reenter the same metadata fields for data from the same study, but maybe simply alter the few fields that are different between the metadata sheets. Reusability, may seem a simple requirement, but is crucial to improving user satisfaction. User satisfaction is ultimately one of the major factors that determines whether a tool will be used and will be successful or not.

We would like to explore with the MS Excel group if the strategy indicated above using macros is the most suitable method for embedding required metadata for chemical kinetics data. If there is interest, we would like to discuss alternative solutions and if there is interest in providing some of this

functionality in base Excel products without requiring our user-defined macros, we would welcome that.

- Quarter 1: Design: Requirements analysis and identification of the metadata required for chemical kinetics data
- Quarter 2: Implementation: Complete the metadata standard specification and implement macro that conforms to the standard.
- Quarter 3: Evaluation & Enhancement Request Identification: Evaluate the macro using real users and identify areas of improvement, e.g., functionality required for ease of use, reuse capabilities, etc.
- Quarter 4: Refinement: Complete the refinement of the macro using feedback from Q3. **Design and implement server-side module to process the Excel datasheets with macro to transfer data to a database.**

Deliverable**: Easy-to-use macro for MS Excel with simple documentation that chemical kinetics specialists can use to enter metadata for data.**

2. Search Engine for Chemical Formula Search

We seek to build a search engine to enable search for chemical formulae, chemical compounds and other chemical entities. Specifically, we seek to enable retrieval using (possibly partial) chemical structures as keywords. Different chemists use different terminologies. Subgroups of chemists like chemical kinetics specialists and environmental geochemists use different terms for the same concept and sometimes under-specify their queries. In order to enable search using diverse vocabularies, we intend to explore query expansion techniques. We plan to Extend chemical entity search engine to handle structural search queries, initially designing algorithms to extract backbone and functional groups from structures.

Our evaluation metrics will be measure the new search algorithm against Gstring.

- Quarter 1: Design a structural matching algorithm based on identifying functional groups and backbones of chemicals
- Quarter 2: Design and implement backbone and functional group extraction and identification algorithm
- Quarter 3: Design and implement ranking functions for similarity measure for structural search.
- Quarter 4: Evaluate the algorithm ; write paper and publish; identify opportunities for improvements

Deliverable: In general, we intend to make the software that enables chemical entity extraction from text available publicly. Our software will be able to segment large compound chemical names and index them to enable efficient search. These indexes can be combined with indexes of regular search engines to enable search for chemical entities. Our novel query models and ranking functions help return (chemically) relevant documents to users' search queries. **This software can be easily integrated with Microsoft Live Search if desired to improve searching for chemical formulae and chemical names using that search engine.**

We will request that query logs from MS Live Search be made available to us. The query logs will enhance and tune our query expansion algorithms in conformance with the Microsoft customers' behaviors. We believe some query logs are available to other Microsoft awardees (of research grants) and the same will be valuable for our research and make sure that our work is of potential use for Microsoft's search engine.

We list possibilities of collaboration with Microsoft product groups because we envision a lot of technology transfer and collaboration with the product groups but are unsure about what can be made available by Microsoft. **If we are allowed access to Microsoft Live Search Engine code or we get cooperation from the Microsoft Live Search group, we believe our enhancements can be added on to the Microsoft Live Search Engine easily.**

University of Southamption

1. Repository and Data Model Design (Q1-Q2)

Survey & extend repository architectures. We have identified a need to build a repository for Molecular Dynamics simulation data using Microsoft SQL-Server back end database to contain typical MD runs of several GB each.

Survey & extend existing XML compatible data descriptions for experiments. Compare existing XML data schemas (e.g. e-Bank) with Office XML and combine to facilitate smooth data and metadata sharing between Office and the Repositories

Create necessary experimental data types to include surface tension, SHG, SFG. Achieve this with as close match as possible to Office XML

Deliverables: Design documents for next quarter implementation.

2. Repository Population (Q2-Q3)

Populate Laboratory Experimental Repositories Build Office (e.g. Word, Excel, IE etc) interface to the repository to extract as much metadata as possible from the Office and Windows properties descriptions to minimize data input on ingest to the repository and enable ingest from what ever Windows Office tool the researcher would normally use to view the data. Develop Ingest tool on the Office toolbar.

- R4L spectral data (crown ether example)
- eCrystal
- SHG
- Surface Tension
- Lab environment
- Auto-generated data to repository

- SFG Repository (Stockholm)

Deliverables: Loaded repositories for subsequent interoperability integration.

3. Repository integration with ORE (Q3-Q4)

Make populated repositories compatible with ORE, based on ORE specs developed in Q1 and Q2.

- Create high level descriptions to facilitate conversion of repository to ORE
- **Convert repositories and Office interface to ORE**
- DOI linkages to ORE objects

Deliverables: Interoperable repositories.

Year 2: September 1, 2008 – August 31, 2009

NOTE: Year 2 plans are less well-defined than year 1 plans for some participants. Refunding for year 2 will be contingent on refinement of year 2 work plans.

University of Cambridge

1. Robustify CML and CML dictionary ontologies in collaboration with the project so that everyone can use this as a standard vocabulary.

- Q1-Q3: publish enhanced documentation for CML. Identify any (project-independent) revisions or upgrades to CML

2. Expose CrystalEye with static RDF consistent with the ontologies

- Q1-Q3: Continued releases of crystal Eye due to increased content

3. Understand the criteria for distributed chemical search and substructure search

- Q2: coordinate installation of prototype search tools in selected project repositories
- Q3: gather feedback on search tools and processes. (PP)21 publish draft report, protocol and experimental toolkit for chemical search

4. Evaluate the issues in scaling RDF across the repositories

- Q2: report on interoperability of CML-RDF with non-chemical RDF in repositories
- Q4: publish final recommendation for CML-RDF

5. Additional activities

- Add SPECTRaT repository
- Evaluate distributed search issues

Cornell University

1. Refinement of models in response to year 1 results (Q1)

As described in the previous year we will deliver and maintain on the ORE web site publically available specifications of both a data model, and core and eChemistry specific vocabularies. Like all specification development, we expect this to be an iterative process, with refinements and corrections growing out of implementation experience. In fact, a major goal of the entire project is this proof-of-context and verification of the specs by implementation practice. We plan, however, to stabilize the specifications developed in the previous year's work by 3rd quarter 2008 (the beginning of the second year of the project).

Deliverable: Publically available specifications available via the OAI-ORE web site

2. Integration of ORE models into IU web services and grid work (Q2)

As noted earlier, our schedule for the 1st year of the project includes a white paper on integration of ORE protocols and models into a XML services/grid framework. In collaboration with our IU partners, we will build on this as follows:

- Write and publish an implementation guideline document describing the use and deployment of ORE in a grid framework.
- Prototype and deploy proofs of concept of this integration.

Deliverable: Publically available specifications/implementation guides posted on the ORE web site.

3. Integration of ORE into .NET framework (Q3)

An earlier deliverable specifies the integration of ORE into a general web services framework. This naturally provides the foundation for understanding and prototyping ORE standards in .NET, a web services based distributed framework. This deliverable has two parts:

- Publish an internal white paper, developed in collaboration with Microsoft, that describes the functionality and requirements (at a relatively high level) of such a deployment.
- Prototype plugins with such functionality.

Deliverable: Internal white paper and possible prototypes

4. Integration of ORE models with Office tools (Q4)

We hope that widespread deployment of ORE-based standards will lead to "killer-app" eScience user applications that will, for example, allow researchers to author, cite, visualize, decompose, share, and analyze compound documents, of the sort we model in ORE. **One can imagine "Office for Scientists" – a set of highly refined applications that provide a powerful interface to eScience activities. This deliverable has two parts:**

- **Publish an internal white paper, developed in collaboration with Microsoft, which describes the functionality and requirements (at a relatively high level) of such an application.**
- **If possible, work with Office application group to prototype or develop plugins with such functionality.**

Deliverable: Internal white paper and possible prototypes

5. Investigate in-the-lab semantic network creation and smart papers (Q4)

Over the long-term we would like to see a suite of rich in-the-lab tools an infrastructure to capture research results as they develop. This is a key part of the vision of a eChemistry semantic knowledge network. Jeremy Frey's group at Southampton has done preliminary investigation in this area and we would like to see this as a major focus in the third year of the project, if it is funded. In this deliverable we will deliver a white paper that sets the stage for this additional year's work.

Deliverable**:** Internal white paper

6. Project management, technology coordination, and reporting (ongoing)

The Cornell group will have primary responsibility for project coordination, management, and reporting.

Deliverable**:** Quarterly reports

Indiana University

1. Enhanced Protocol and Message Format (Q1)

Complete next major iteration of protocols and message formats, based on Year 1. We will document "lessons learned" from our Year 1 design and implementation efforts and improve our protocol and message formats accordingly.

Deliverables*:* updated specifications of Year 1/Q1 documents.

2. Additional Service Integration (Q2)

Additional team services are ORE-enabled: in collaboration with partner institutions, we will provide documented libraries, toolkits, and human resources to implement additional ORE-based services. Candidate services include OSCAR3 and CrystalEye (from Cambridge), ChemXSeer (Penn State), and Southampton data services.

Deliverables*:* demonstration of modified services.

3. Web 2.0 Interfaces (Q2)

Sample Web 2.0 **user and social network interfaces (based for example on MS Web 2.0 tools including Silverlight) for aggregating, tagging, commenting, and reviewing ORE-defined data products**. This will extend the deliverables of Year 1/Q2.2.

Deliverables*:* Enhanced Year 1/Q2.2 Web site; software used to build site will also be provided.

4. Advanced Security Design (Q3)

Advanced security policy design and initial implementation completed. This will allow examine the problems of inter-service federation and trust (building on authentication and access control mechanisms defined in Year 1). The key problem will be to federate (in a scalable fashion) the individual authentication mechanisms of isolated services.

Deliverables*:* design document, demonstration of integration with selected services.

5. Structure and Docking Services (Q4)

Completion of scalable IU chemical structure and docking services.

Deliverables*:* demonstration and documentation of federated database system with optimized queries, capable of handling 100 million (or more) structures obtained from NIH PubChem. Make available through Web 2.0 style interfaces (see Q 2.2 deliverables).

6. Use Cases (Q4)

Intermediate/advanced use cases implemented and demonstrated. These will feature Year 2/Q2 services.

Deliverables: demonstration.

Penn State University

- For the retrospective data analysis work, classify spectra and further refine methods. Data will be extracted and available for the appropriate repositories, including storage of relevant data in the ChemXSeer system at Penn State.
- Extend our search capabilities not only to index the linear, branched, and other structures in the backbones of molecules, but extend the string representation of the chemical structure to include the location and type of functional groups. Provide all extracted data as databases for inclusion in the appropriate chemical information repositories.
- Understated how spectral classification and pattern detection techniques will be extended to other chemical domains; whether search strategies can address more complex domain-knowledge issues.

University of Southampton

1. Repositories and Services (Q1)

- Test ORE interfaces and linking of data, i.e. Link repository content by molecule, linking of molecule to PubChem, and with Cambridge computational and structural data, and ensure these links can be readily expressed and manipulated from within Word and IE as the front-end to the user.
- Link with Pen State Search Services
- Link repository content by experiment and by project, person

2. Application Interfaces (Q2-Q3)

- **Deployment of Microsoft's 'Live Clip Board' to maintain the metadata link as the data is copied to and from the repository to the office application**. Provide tools for 'time line' views to provide provenance.
    - **Office ribbon extensions to accommodate this functionality.**
    - **Integration with Silverlight, Sharepoint**
- Link services and triple stores (semantic ELN & Laboratory Blog) to ORE repositories

- Ensure link of data repositories with Publication Repositories works smoothly
- Create “Time Line” views of data in Repositories – The Provenance GUI

3. Smart Paper Demo (Q4)

- Produce a Smart Paper Demo (hand crafted?)

Year 3: September 1, 2009 – August 31, 2010

Contingent on the success of the two-year project, we hope to extend our work into a third year. In this year we hope to examine two main areas:

- In-process, in-the-lab capture – The eChemistry web that we envision must include more data than that which can be captured through retrospective analysis. The creation of a truly novel eChemistry environment involves recording the process and preliminary products of research as they appear in the lab. We plan to extend the work of the Southampton group in this area, building on the infrastructure created in earlier stages of the project.

- Analysis and knowledge extraction/creation – Recent work on analysis of web data and historical work on citation/bibliometric analysis demonstrates the rich information that can be mined from large-scale graph-based information systems. A new field called Scientometrics is taking shape that extends bibliometrics to semantic heterogeneous graphs. We plan to developed advanced analysis tools that exploit the expressiveness of the eChemistry web that we will develop in earlier stages of the project.

Project Management and Communication

The Cornell University team will have responsibility for overall project management. This involves ensuring the project schedules and deadlines are met, communicating project changes to Microsoft, and submitting reports.

The team will communicate mainly by electronic means: email and blogs. The budget includes funding for three face-to-face meetings among senior personnel (one representative from each team). The first meeting will be a kick-off meeting to set initial project work, at this point scheduled for September 2007.

Integration with Microsoft Products and Applications

As stated earlier, funding of this project provides two distinct benefits to Microsoft. The first is "good will" and public perception due to involvement in a high-profile effort that has the potential to influence and democratize science and scholarship. One of the participants in this project – Open Archives Initiative Object Reuse and Exchange – is already partially funded by Microsoft and has, even in its initial pre-release stages, attracted considerable international attention for its role in open access scholarship. In addition, project participants Murray-Rust and Frey, regarded as innovators in the Chemistry community, have received support from Microsoft for work that this project will extend. This proposed project will provide the opportunity to leverage those already successful investments.

Parallel to this benefit, there is the more tangible benefit to Microsoft's product line and future application development. This project will provide Microsoft researchers and product teams with unique access to leading researchers in the area of digital scholarship, thereby providing a testbed for product innovation in this area. We note that this is not a fringe application. As noted by a number of U.S. cyberinfrastructure and international eScience/eScholarship reports (Atkins et al., 2003; National Science Foundation Cyberinfrastructure Panel, 2007), this new form of scholarship marks a fundamental change in the way that research is undertaken and disseminated. Many predict that these changes will integrate the process and products (textual, visual, data and compute-centric) of scholarship into the web 2.0 social network that has been rapidly developing over the past several years (Ginsparg, 2007; Lynch, 2007). The result will position scholarship as one part of the emerging "digital lifestyle" in which computer and networks are increasingly embedded in daily entertainment, learning, communication, and business activities. Microsoft has been a major player in this area ever since its first release of Internet Explorer. Recent products such as Silverlight, Popfly, and the Windows Live online environment demonstrate an increased focus in this area. This project will demonstrate how these new products, and others, improve science and scholarship, and help integrate them into the broader digital environment.

The members of the project team are committed to working closely with Microsoft on these product related areas. Our goal as researchers and practitioners in this field is to facilitate wide-spread deployment of the products of our work. Because of Microsoft's position as the largest vendor of software in the world, with almost ubiquitous desktop products like Office, it is uniquely positioned to push-start the innovations we will investigate in this project. Our success in producing the deliverables linked to these products depends in many cases on our ability to liaison with Microsoft product teams. We believe these liaisons will prove beneficial to both parties. Furthermore, as the project progresses we are willing to modify plans in order to meet contingencies or new points of possible interest that arise from these liaison relationships.

The remainder of this section summarizes the Microsoft applications with which we plan to experiment and develop extensions or plug-ins to. Please refer back to project plan for more details on our proposed work with these applications.

NET

We propose deliverables that demonstrate the utility and value-add of Microsofts's .NET framework for eScience and eScholarship. We envision OAI-ORE compound object format as the medium of exchange of information about eScience objects among distributed Web Services. The security enhancements provided by .NET over the industry-standard web services layer are particularly essential in the project's eScience framework, especially in chemistry where the value of certain research products must be protected.

The Indiana team, in collaboration with the Cornell team, will implement ORE compound object protocol and data transfer standards within .NET. This will demonstrate the utility of .NET for exchange of compound eScience data while respecting the security and privacy concerns that are essential to scholarship.

SQL Server

This is a data-base centric project, and as such experimentation with Microsoft's SQL server is an integral component. The Indiana team will investigate scaling issues in their PubChem-derived chemical databases (PubDock and Pub3D). They plan to test response to query times. The Southampton team will test the SQL server in the context of Molecular Dynamics simulation data and its adaptability to simulation runs of several GB each run. In addition, we are interesting in the transactional functionality of the SQL server vis-à-vis robust transfer and deposit of multi-component eScience objects – i.e., the types of content modeled by OAI-ORE. We anticipate that the results of this project will both lead to improvements the SQL server product and demonstrate the utility of that product in the eScience domain.

Office

We propose two areas of work that will impact Microsoft's Office suite:

1. *Word and Excel as eScience authoring applications*: We include deliverables that will produce plug-ins to both Word and Excel that will allow scientists to use them as the entry point for eScience data. The Penn State group will experiment with Excel as a data entry tool and, in particular, experiment with Excel macros for capturing metadata. The Cornell group and the Southampton group plan, in the second year of the project, plan a more extensive liaison with the Office applications group to explore and possibly prototype an "Office for Scientists", with which scientists will use this near-ubiquitous desktop application as a tool for creating new forms of scientific publications, that combine data, imagery, text, and the like, and then upload those "publications" via OAI-ORE to repositories. Clearly, these will not produce finished applications. But, we expect to produce valuable results that will make it possible for Microsoft to evaluate the feasibility of these applications, and consider the benefits of producing marketable applications.

2. *Integration of data and experimental xml formats with Office XML*: XML is increasingly the *ligua franca* for marking up digital media. In the realm of data, there are a number of emerging

standards for encoding experimental metadata and data in XML. At the product level, Microsoft has demonstrated a commitment to open XML standards by basing Office 2007 on an open XML standard. We propose deliverables that will produce mappings between these XML formats, making it possible to ingest and manipulate experimental data in the Office context and then subsequently export it. Mechanisms that allow the inclusion of scientific data in Office (in particular Word) documents are vital as more and more scientists work with published results that are data-inclusive, rather than being simply data-descriptive.

Internet Explorer

Browser functionality and ease of use has improved dramatically over the last decade. One area that is currently unexploited are tools that make it possible to manipulate aggregations of web resources. The description of these aggregations is the goal of the ORE standards, which then provide the foundation for a number of enhanced browser-based tools. For example, we imagine browser extensions that will leverage ORE descriptions to provide advanced services such as print, visualize, and summarize over ORE-described compound objects. The Cornell team will use student research teams to experiment with ORE and IE as a means of enhancing the browser experience with such services and feed the results back to Microsoft. We anticipate that, as ORE standards are more widely deployed, these extensions will be a significant value-add for IE.

Silverlight

The OAI-ORE standards facilitate the production of more complex and information rich digital content. Tools that help users manipulate and understand this content are essential if this content is to be useful. Microsoft's new Silverlight provides an excellent environment for cross-browser navigation tools. The Indiana team will experiment with SIlverlight as the basis of knowledge navigation tools.

Popfly

Our proposed work on this project is partially motivated by the now familiar "mashup" metaphor. We would like to provide scholars with the infrastructure to easily combine digital content from multiple sources into new information units. Microsoft's new Popfly product, in combination with Silverlight, makes it possible to build web applications that allow such mashups. The Cornell team will work with Jane Hunter's group at the University of Queensland, Australia, who have already created a prototype ORE-based eScience authoring environment (Cheung, Hunter, Lashtabeg, & Drennan, 2007), and implement this prototype using Silverlight and Popfly.

Live Search

Existing text-based web search applications represent only the initial stages in networked information discovery. Search engines need to move beyond text and generalized discovery into the realm of specialized domain-specific and personalized discovery (Lagoze & Singhal, 2005). The Penn State team plans to work with the Windows Live Search product group to integrate chemical search results into the Live environment. While these investigations will focus on Chemistry, we argue that the "beyond text" focus will apply to other search domains where finding structured objects is the focus of search.

Sharepoint

We intend to provide the infrastructure for rich collaborations between members of the Chemistry research community. This collaboration should involve popular social network tools such as blogs and IM, with exchange of data-rich research products. The Southampton team will experiment with Sharepoint services as the vehicle for this scientific collaboration. They will investigate how ORE standards integrate into the Sharepoint fabric, and enable sharing of rich, compound scientific content.

References

Atkins, D.E., Droegemeier, K.K., Feldman, S.I., Garcia-Molina, H., Klein, M.L., Messerschmitt, D.G., et al. (2003). Revolutionizing Science and Engineering Through Cyberinfrastructure: National Science Foundation Blue-Ribbon Panel on Cyberinfrastructure.

Bekaert, J., Liu, X., Van de Sompel, H., Lagoze, C., Payette, S., & Warner, S. (2006, June). In Pathways Core: A Content Model for Cross-Repository Services. Paper presented at the Joint Conference on Digital Libraries, Chapel Hill, NC. ACM/IEEE.

Bizer, C., Cyganiak, R., & Heath, T. (2007). How to Publish Linked Data on the Web: Free University of Berlin.

Booth, D. (2004). Web Services Architecture: W3C.

Chadwick, D.W., & Otenko, A. (2003). The PERMIS X.509 role based privilege management infrastructure. Future Generation Comp. Syst., 19(2), 277-289.

Cheng, J., Ng, W., & Lu, A. (2007). Fg-index: towards verification-free query processing on graph databases, SIGMOD: ACM.

Cheung, K., Hunter, J., Lashtabeg, A., & Drennan, J. (2007). SCOPE - A Scientific Compound Object Publishing and Editing System, 3rd International Digital Curation Conference. Washington, D.C.

Coles, S.J., Frey, Jeremy G., Hursthouse, Michel B., Light, Mark E., Milsted, Andrew J., Carr, Leslie A., De Roure, David, Gutteridge, Christopher J., Mills, Hogo R., Meacham, Ken E., Surridge, Michael, Lyon, Elizabeth, Heery, Rachel, Duke, Monica and Day, Michael. (2006). An e-science environment for service crystallography from submission to dissemination. . Journal of Chemical Information and Modeling, 46(3).

Dervos, D.A., Samaras, N., Evangelidis, G., Hyvärinen, J., & Asmanidis, Y. (2006). The Universal Author Identifier System (UAI_Sys), 1st International Scientific Conference, eRA: The Contribution of Information Technology in Science, Economy, Society and Education. Tripolis, Greece.

Fox, G.C., & Gannon, D. (2006). Special Issue: Workflow in Grid Systems. Concurrency and Computation: Practice and Experience, 18(10), 1009-1019.

Ginsparg, P. (2007). Next-Generation Implications of Open Access. CTWatch Quarterly, 3(3).

Gregorio, J., & Hora, d. (2007). The Atom Publishing Protocol: IETF.

Jiang, H., Wang, H., Yu, P., & Zhou, S. (2007). GString: A Novel Approach for Efficient Search in Graph Databases, ICDE.

Lagoze, C., & Singhal, A. (2005). Information Discovery: Needles and Haystacks. IEEE Internet Computing, 2005(May/June).

Lynch, C.A. (2007). The Shape of the Scientific Article in The Developing Cyberinfrastructure. CTWatch Quarterly, 3(3).

Murray-Rust, P., Mitchell, J.C., & Rzepa, H.S. (2005). Chemistry in Bioinformatics. BMC Bioinformatics, 6(141).

Murray-Rust, P., & Rzepa, H.S. (2004). Towards the Chemical Semantic Web. An introduction to RSS. Internet Journal of Chemistry, 6(4).

National Science Foundation Cyberinfrastructure Panel. (2007). Cyberinfrastructure Vision for 21st Century Discovery. Washington, D.C.: National Science Foundation.

Nottingham, M., & Sayre, R. (2005). The Atom Syndication Format (Request for Comments No. 4287): Network Working Group, Internet Engineering Task Force.

Paskin, N., & Rust, G. (1999). The Digital Object Identifier Initiative: Metadata Implications (No. Version 3): International DOI Foundation.

Taylor, K.R., Essex, J.W., Frey, J., Mills, H.R., Hughes, G., & Zaluska, E. (2006). The semantic grid and chemistry: experiences with CombeChem. Journal of Web Semantics, 4(2).

Taylor, K.R., Gledhill, R.J., Essex, J.W., Frey, J.G., Harris, S., & De Roure, D.C. (2006). Bringing chemical data onto the semantic web. Journal of Chemical Information and Modeling, 46(3).

Tuecke, S., Czajkowski, K., Foster, I., Frey, J., Graham, S., Kesselman, C., et al. (2003). Open Grid Services Infrastructure (OGSI): Version 1.0 (No. draft-ggf-ogsi-gridservice-33): Global Grid Forum.

Warner, S., Bekaert, J., Lagoze, C., Liu, X., Payette, S., & Van de Sompel, H. (2007). Pathways: Augmenting interoperability across scholarly repositories. International Journal on Digital Libraries special issue on Digital Libraries and eScience, forthcoming.

Xifeng Yan, P.S.Y., Jiawei Han. (2005). Substructure similarity search in graph databases,, SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of Data. Baltimore: ACM Press.

Yin, H., Brenes-Barahona, S., McMullen, D.F., Pierce, M., Huffman, K., & Fox, G. (2006). A PERMIS-based Authorization Solution between Portlets and Back-end Web Services, Second International Workshop on Grid Computing Environments GCE06 at SC06. Tampa, FL.