SemanticEye: A Semantic Web Application to Rationalize and Enhance Chemical Electronic Publishing

Omer Casher

Clinical Imaging Centre, GlaxoSmithKline, Harlow CM19 5AW, U.K.

Henry S. Rzepa*

Department of Chemistry, Imperial College, London SW7 2AY, U.K.

Received April 13, 2006

SemanticEye, an ontology with associated tools, improves the classification and open accessibility of chemical information in electronic publishing. In a manner analogous to digital music management, RDF metadata encoded as Adobe XMP can be extracted from a variety of document formats, such as PDF, and managed in an RDF repository called Sesame. Users upload electronic documents containing XMP to a central server by "dropping" them into WebDAV folders. The documents can then be navigated in a Web browser via their metadata, and multiple documents containing identical metadata can then be aggregated. SemanticEye does not actually store any documents. By including unique identifiers within the XMP, such as the DOI, associated documents can be retrieved from the Web with the help of resolving agents. The power of this metadata driven approach is illustrated by including, within the XMP, InChI identifiers for molecular structures and finding relationships between articles based on their InChIs. SemanticEye will become increasingly more comprehensive as usage becomes more widespread. Furthermore, following the Semantic Web architecture enables the reuse of open software tools, provides a "semantically intuitive" alternative to search engines, and fosters a greater sense of trust in Web-based scientific information.

1. INTRODUCTION

In its 17-year history, the Web has revolutionized the dissemination of scientific journals. Chemists read about 40% more journals now than 20 years ago primarily because of the Internet-based delivery mechanism.[1] The Web presented new opportunities to enhance publishing, and next generation scientific electronic publishing models have been emerging. Early chemistry exemplars, such as CLIC[2] and the *Internet Journal of Chemistry*,[3] delivered highly structured enhanced chemical electronic journal models with data-centric entry points such as structure-based searching. Despite the added value, the enhanced chemical electronic journal has yet to be realized, and most electronic media are nothing more than digital replicas of their paper-based counterparts.

The widespread adoption of the digital replica model can be attributed to the dynamics of the Web's growth among the chemical research community. However, according to Goodman's historic account,[4] the Web is now starting to achieve saturation coverage, and from hereon further growth will not be attained by increased use but by applying new ideas. This represents an opportune time to re-explore the enhanced electronic journal model for chemical publishing.

We believe that the next generation of scientific electronic publishing will move toward a "deconstructed journal" model,[5] aligned to the positive aspects of the Web while addressing diverse stakeholder interests. The perspectives of the individual stakeholders are examined here.

1.1. Publisher. Although not significantly cheaper than paper publishing, the existing electronic publishing model is appealing to publishers because of diminishing Internet infrastructure costs, an established charging model, and the transfer of printing costs to the reader. Furthermore, the deliverable to the reader is a printed document which preserves the publisher's look and feel while remaining true to the structure of the scientific journal. Although value-added components would improve the reader's experience, arguably they would not generate the necessary revenue to cover the ensuing support costs. Currently, publishers would prefer to maintain their established revenue-generating infrastructures and leave the management of nonrevenue-generating components to others. For the enhanced chemical electronic journal to become viable, publishers may expect this model to persist.

Pressure from some quarters to make articles freely available has resulted in increasing attention being given to Open Access journals,[6] a radically different publishing paradigm. A high value Open Access constituent is the preprint repository which, although popular in areas with a preprint culture such as physics,[7] has still not caught on in chemistry. Until now, there has been no hard evidence that the availability of preprints undermines a journal's viability, and any concerns by the publishers or readers might be unfounded. However because authors generally want their published work to be widely read, an alternative preprint repository model, better aligned to the publishing patterns of the chemistry community, may well emerge in the future. We recognize that for such a model to succeed, any cultural barriers will need to be well understood and addressed.

1.2. Author. For the enhanced electronic chemical journal to achieve critical mass, authors would be tasked to provide,

* Corresponding author e-mail: rzepa@imperial.ac.uk.

10.1021/ci060139e CCC: $33.50
Published on Web 11/01/2006

along with the electronic manuscript, data results and some mandatory metadata. Manual metadata entry is particularly bothersome and would be subjective and error prone without validation. Publishers assume that authors will always try to avoid such activities and that imposing them would incite the scientist to publish elsewhere.

In reality the author has a proclivity for publishing in as respected a journal as possible. If a highly respected journal mandates inclusion of results data and metadata, the burden of complying will be absorbed by the author(s), albeit begrudgingly. A greater concern to most chemists is that the sharing of experimental data to scientists outside of the research team risks misinterpretation and reuse out of context. Providing interpretation/annotation to such data which conveys the intended context is an additional burden, but one which the provision of good communication software tools and training can help address.

1.3. Reviewer. There will always be an elementary need for good peer review. However, the steady rise of research papers is placing an increased burden on the available pool of reviewers. As a result, the peer review process is going through a period of profound change, and new publishing trends and alternative models are emerging.[8]

1.4. Archivist. Whereas the long-term archival of printed journals has traditionally been a library service, digital replicas are archived by the publishers, who charge for such access. Any experimental data included with manuscripts must frequently be archived by the authors themselves, as it is not financially advantageous for publishers to do this. We predict[9] that this role will shortly be fulfilled by emerging "institutional repositories", which are central electronic archives set up in universities and centers of excellence. Authors would upload the experimental data to one of these facilities and provide links to the data from the manuscripts. The growth of institutional repositories has been largely driven by projects such as DSpace[10] at MIT. This system provides tools to capture, store, index, preserve, and redistribute an organization's electronic research material. Its uptake at academic institutions is growing.

1.5. Reader. King and Tenopir, who carried out a study on reading patterns for scientific journals over a 25-year period, presented a report on journal articles read by chemists between the years 2000 and 2003.[11] Three trends here are noteworthy:

(1) 23% have been sourced through online article searching, thereby replacing interrogation of printed indexing services, which was 35% in 1977.

(2) 51% have been located by browsing journals, virtually unchanged from 1977 (49%).

(3) 14% have been located by word of mouth, compared to 3% in 1977.

The fact that journal browsing consistently accounts for about half of the articles read by chemists reveals that the context and provenance which a journal provides is still very important to chemists. Furthermore, a journal's readership helps shape a research community to which the reader would feel an association. However, electronic distribution of digital journal replicas has transmuted the concept of an article from a "collective within a community" to an "independent singular" which is discovered with the help of a search engine. This would explain why just over twice as many articles read by chemists are discovered from browsing compared to online searching.

Hence, despite better communication, electronic publishing has yet to address the reader's need for attaining context and community among multiple electronic journal articles. This untapped opportunity is the focus of our study. Although our proposed solution is reader-oriented, we recognize that its value will be ameliorated if the needs of the other stakeholders are addressed by other solutions.

2. DEFRAGMENTING ELECTRONIC PUBLISHING

The fragmentation of the Web in general and of electronic publishing in particular is not caused by technology. The Web indeed relies on technology but is not driven by it. The problem is mainly cultural, and for publishers the priority is to manage and safeguard their own content. All publishers, at least in the near term, would not be expected to attain a level of standardization that would enable the Web to be fully exploited. Thus, defragmenting the existing situation must be carried out external to the journal, and we propose the following approach to achieve it.

(1) Create semantic relationships between electronic articles to establish context and community of importance to readers. These relationships should not be restricted to journals or publishers.

(2) Broaden these semantic relationships to disclose other resources on the Web, not necessarily within the publishing domain.

We believe that the more developed these relationships are, the more effective the Web will be as a medium for journal dissemination. Creation of the relationships can be carried out by one or more individuals, supported by an organization or a community. Alternatively they can develop organically as a collaborative effort. This latter option is unbounded and self-sustaining, provided the underlying framework would be engaging enough to encourage participation and that the main constituent of this framework is metadata.

2.1. Metadata Approach. Although there is no shortage of metadata models and management techniques in the literature, no single approach has been embraced by the electronic publishing domain. Instead a metadata analogy to the digital replica model has persisted whereby electronic journal metadata follow archaic paper-based journal classification methods. Metadata management is the most challenging aspect of content management, and getting it wrong has expensive ramifications down the line. Several fundamental guiding principles should mitigate the risks however.

(1) Anticipate the reader's working patterns and expectations. A solution which is cumbersome and confusing is potentially counterproductive.

(2) Keep the solution simple and applicable to as many scientific journals as possible. It is too expensive to maintain multiple journal specific software applications. Ideally any approach should not be disruptive to the publishers' core business.

(3) Use open standards wherever possible.

(4) Understand the metadata lifecycle and how it flows from capture through to archival as it has implications on the selection of standards and software. Once the metadata

has been modeled, the ensuing information and software architecture should ideally be a reuse of existing services. The metadata model must be simplified to minimize software complexity and maximize reusability across multiple domains. A journal driven approach risks locking down the software to a narrow subject domain. A technology driven approach compromises the flow of metadata and has a high risk of failure.

(5) Central resolving agencies for key metadata identifiers should be harnessed wherever possible. A resolving agency is a Web service that will validate an identifier, provided as an HTTP request, and redirect the request to the object to which the identifier points, if it exists.

(6) A carrot and stick approach is required for authors to capture the metadata and include it with the manuscript. Any solution will fail if this task is optional. Publishers might need to assume at least some of the responsibility for entering the metadata even though the subject matter expertise resides with the authors.

For scientific electronic publishing to truly prosper, a metadata framework attuned to the virtues of the Web is required and, fortunately, one already exists.

2.2. The Semantic Web. Tim Berners-Lee's vision of the Semantic Web is to "assist the evolution of human knowledge as a whole".[12] Because this vision also reflects scholarly publishing, applying Semantic Web principles for electronic publishing should be a matter of course. However, the Semantic Web is no magic bullet, and realizing this vision, if it is indeed possible, requires some work. Nevertheless, now would be a good time to evaluate it as an enabler of electronic publishing for the following reasons:

(1) The Semantic Web principle has been subject to theoretical scrutiny for more than 5 years.

(2) Adoption of its standards among the scientific community, particularly in bioinformatics, has been increasing.

(3) Supporting technology solutions are maturing.

From a technical standpoint, the Semantic Web is a knowledge management framework for navigation and discovery of distributed resources on the Web. Its utility in chemistry has been rigorously studied.[13] The fundamental difference between the Web and the Semantic Web is that the Web is designed for human dissemination and consumption, whereas the Semantic Web also permits the participation of software agents to assist with the dissemination. Agents are not expected to predict semantic meaning from existing Web content using clever algorithms. Instead, content owners invest some effort lacing the content with suitable metadata, with the payback being that more readers and agents would be able to locate the content and understand it. For electronic publishing, this process can be assisted with specialist software tools.

2.2.1. The Semantic Web Architecture. The Semantic Architecture is not universally agreed and is currently under discussion among a community of developers, researchers and interested parties. The best known representation is Berners-Lee's Semantic Web stack[14] (Figure 1). It is based on a hierarchy of technologies and standards, each of which exploits the features and extends the capabilities of the layers below.

The URI and Unicode layers ensure the use of international character sets and provide a means of uniquely identifying resources. The XML and Namespace layers signify the XML

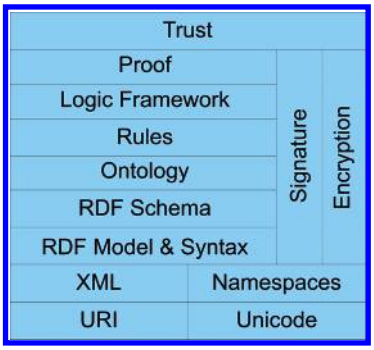

Figure 1. Semantic Web layers according to Berners-Lee (ref 12).

underpinning of the RDF layers, which are discussed in more detail in the next section. The Ontology layer represents the evolution of RDF vocabularies containing relations between the different concepts. The Rules layer provides methods of drawing inferences, expressing constraints, specifying policies, reacting to events, or transforming data of ontologies with the help of a rules language which could ultimately form part of the ontology language. The remaining layers are still under academic investigation. The Logic Framework will enable the integration of rule-based systems. The Proof layer will provide methodologies for Semantic Agents to generate justifications of results. Electronic signatures and document encryption add to the level of confidence. Once the proofs are believed by the users, trust, has been achieved.

2.2.2. The Resource Description Framework. The Semantic Web metadata standard is the Resource Description Framework (RDF),[15] a vocabulary for constructing relationships based, not on name-value pairs but, on triples. An RDF triple describes how a *subject* Web resource relates to an *object* Web resource via a *predicate* attribute. This *object* resource can, in turn, be the *subject* resource in another triple, thereby building up a semantic map with the potential to grow *ad infinitum*. Names are prefixed by namespaces to ensure their uniqueness.

RDF triples are reconciled in N-Triple notation, which applications normally read to construct the semantic map. N-Triples can also be repurposed into XML, thereby exploiting the utility of XML as an interchange format. Because RDF triples can be represented as XML in multiple ways, RDF does not lend itself to XML Schema representations, which restrict XML vocabularies to a single data structure. The RDF Schema is more appropriate for defining the data specific to an RDF vocabulary without imposing rigid XML structure representations. Moreover, unlike with XML Schemas, multiple RDF Schemas can be combined without any repurposing of data structures.

A drawback of the Semantic Web is a misunderstanding of its goals. Many content creators feel that converting at least a portion of their existing content into RDF without any preparatory business modeling makes this content Semantic Web enabled nonetheless. We feel that this approach is counterproductive in the long run. The starting point for a Semantic Web solution should be a semantic model from which the RDF vocabulary and supporting technologies would be derived.

2.3. Adapting a Successful Semantic Model. Two of the Web's greatest success stories, Google.com and Amazon.com, use semantic models,[16] though neither is based on the Semantic Web. Google presumably derives a semantic

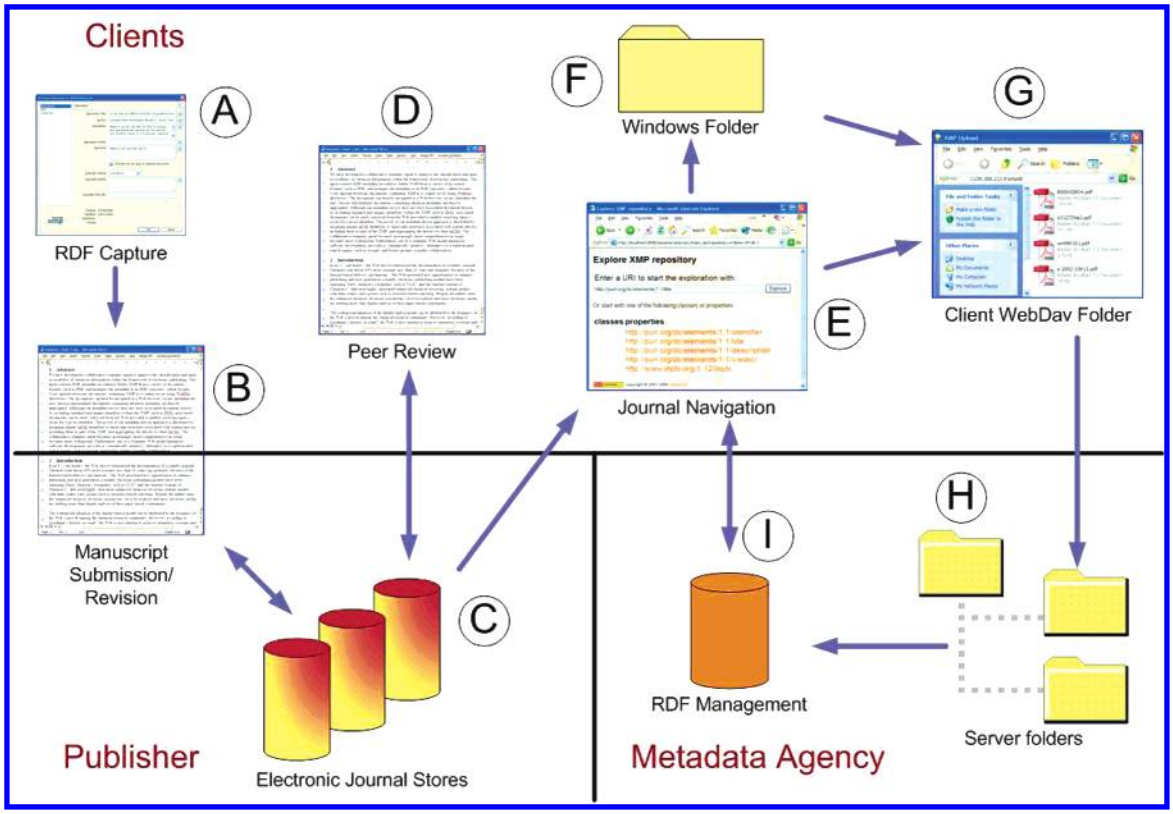

Figure 2. Overall dataflow of the RDF metadata lifecycle.

model of the entire Web and relies on hyperlinks between pages for relationships. This permits it to perform more intelligent queries than its competitors, even though it does not mandate a metadata vocabulary. Amazon.com does not derive a semantic model but adds to its product database a semantic layer containing customer buying habits. This allows targeted and personalized customer experiences, the effectiveness of which vastly exceeds untargeted content such as banner advertisements.

A third successful semantic model is used for the management of digital music files. Here, if a music CD is inserted in the CD-ROM drive, or several music files are dragged and dropped in a suitable window, for example an iTunes window, metadata about the music is retrieved from a Web agency. The success of music metadata management has been compared with the failings of scholarly publishing,[17] and some interesting findings were uncovered.

(1) Digital music metadata is standardized and moves with the asset, while journal metadata is neither standardized nor embedded.

(2) Digital music metadata lookup services are collaborative and automate metadata retrieval for digital music. Journal metadata lookup services are not collaborative.

(3) Music metadata was initially developed for the personal management of a growing library of music and later used for information retrieval. Journal metadata was developed for information retrieval with little focus on personal information management.

In this article, we coin the term **SemanticEye** as a model for rectifying electronic journal shortcomings by adapting the digital music semantic model to chemical electronic publishing. However, unlike the three semantic models described above, we use the Semantic Web model as our starting point.

3. ARCHITECTURE

Our target architecture is an adaptation of the digital music semantic model to create a framework for navigating electronic journal articles in multiple sources, whereby context and community are intrinsic and the reader experience is as intuitive as possible. This section describes our approach in realizing the target architecture.

3.1. Dataflow. Because SemanticEye is metadata driven, our Semantic Web model can be landscaped with the help of a dataflow diagram (Figure 2). The various functions of the dataflow can be summarized as follows:

(1) RDF metadata is captured by the reader using appropriate tools (A).

(2) RDF metadata is embedded in a manuscript (B) either by the reader, the publisher, or both.

(3) The manuscript is submitted to the publisher (C).

(4) Following peer review (D) and possible revisions (B) the manuscript is published and made available on the Internet (C).

(5) Journal articles are retrieved through searching or browsing using a Web browser (E).

(6) Articles can be saved either in a folder on the user's desktop (F) or in a WebDAV folder (G).

(7) Articles saved in a folder on the user's desktop (F) can be dragged and dropped into a WebDAV folder (G).

(8) Articles are uploaded to a folder (H) at the Metadata Agency.

(9) RDF metadata is extracted from the documents and stored in an RDF repository (I), thereby creating semantic associations between the contents of the repository.

(10) A representation of the RDF can be navigated by the reader in a Web browser (E). Journals and supporting data

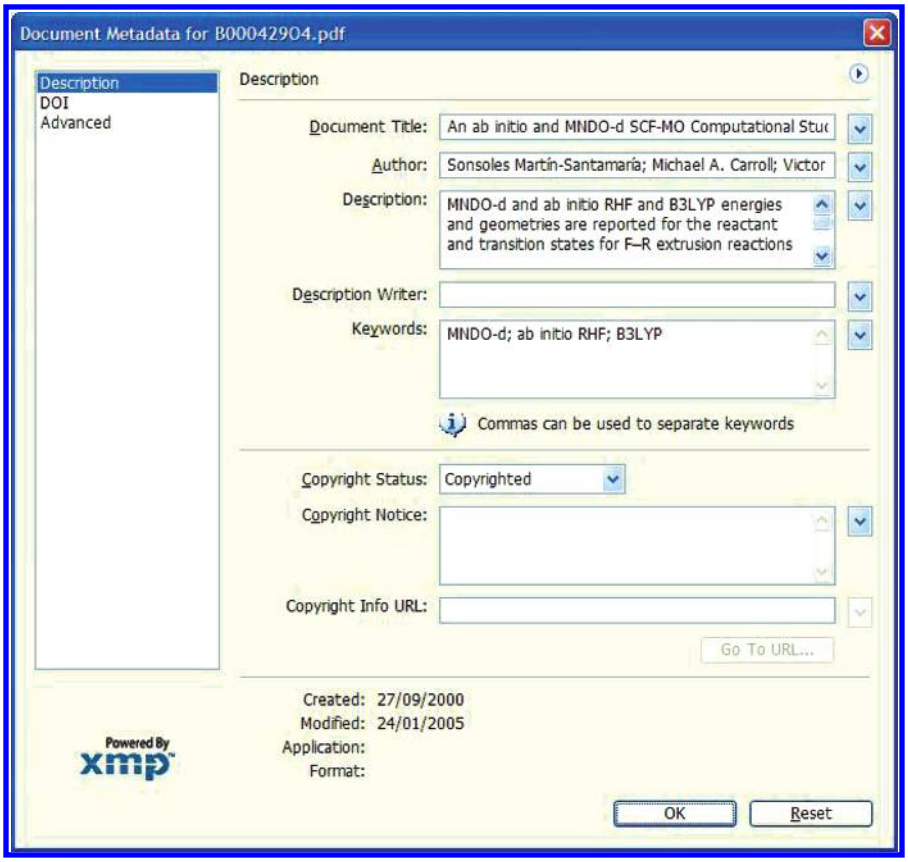

Figure 3. Acrobat Professional XMP Tool.

of interest can retrieved from the Internet using appropriate resolving agencies.

(11) RDF in the repository (I) can be edited via the browser (E).

Achieving the data flow within a Semantic Web framework required deployment of several relatively novel yet powerful standards-based technologies.

3.2. Extensible Metadata Platform − XMP. XMP,[18] Adobe's RDF standard designed for the management and storage of documents, is a natural starting point for our metadata application. As with music files, XMP is embedded in a document and travels with the document when its location changes. Documents can be in any format, and Adobe has published a specification for embedding XMP in some of the most common including HTML, Word, SVG, JPEG, TIFF, GIF, and PDF. Adobe's successes with making PDF and SVG into global standards signify the long-term viability of XMP provided it is straightforward to use and reuse. Adobe supports XMP editing within its document authoring software suite. Adobe also supports an XMP-API for XMP handling within applications.

To date, the uptake of XMP has been limited, which can be explained by the slow growth of RDF as a whole as well as by a number of caveats: XMP does not have an associated RDF Schema. Only a succinct specification, which includes the Dublin Core RDF Schema, is handled the by Adobe software. Most XMP developers would want to extend the vocabulary with their own RDF Schemas or external Schemas, such as PRISM,[19] an RDF Schema for publishing. However, in contrast to polymorphism of RDF, XMP extensions require a restricted RDF serialization to enable correct handling by Adobe software.

XMP capture is achieved by inputting metadata using a form tool (Figure 3) which is available as an "Advanced Document Metadata" feature in Acrobat Professional, Adobe's PDF authoring application. For capturing user-defined metadata, the form tool can be enhanced by editing a panel description file. Alternatively, XMP can be edited outside of the software and imported, an approach which can potentially be automated using chemical data capture techniques. Sustainable metadata capture processes would normally require multiple software tools which can interoperate. The importance of open standards such as XML for software interoperability is exemplified in an Open XML architecture proof of concept.[20]

Adobe's XMP support includes a useful mouse-over feature (Figure 4), whereby positioning the mouse cursor over a PDF document icon reveals a number of the XMP attributes.

3.3. The Document Object Identifier (DOI). The DOI is an internationally recognized open standard and a mainstay of the science, technology, and medical (STM) journal publishing industry. It is a unique identifier referring to an Internet-based resource, at whatever level of granularity is important to the resource's owner, therefore allowing resources to be cross-linked and discovered more easily.

The DOI syntax comprises the following:

<Directory Code>. <Registrant Code> /
< DOI Suffix String >

There is no limit on the lengths of the components.

DOIs are maintained by a central resolving agency which routes a resource's DOI to its actual location. This resolving

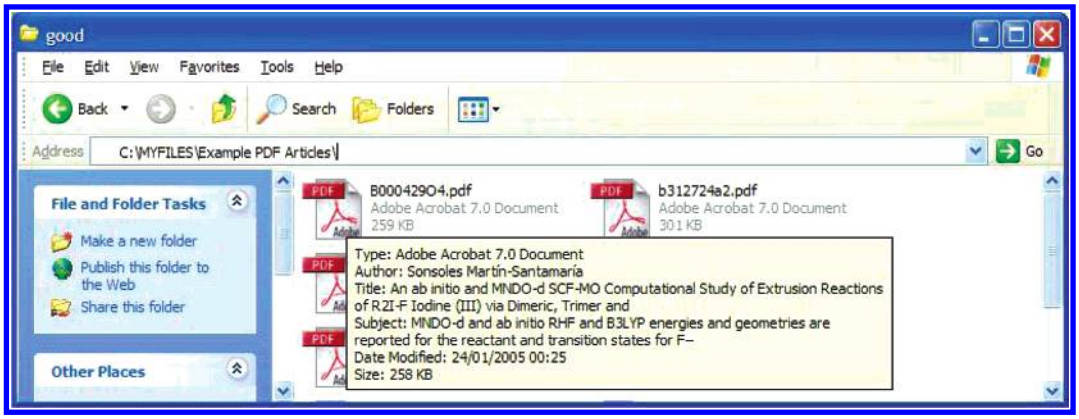

Figure 4. The mouse-over feature for Acrobat.

agency is analogous to the Domain Name Service, or DNS, for resolving networked host names. A DOI is resolved within URL syntax by prefixing a proxy server, http://dx.doi.org, to the DOI. The DOI Handbook,[21] considered the authoritative source of information on the DOI, provides the DOI namespace:

xmlns:doi="http://www.doi.org/2004/DOISchema"

Use of a DOI resolving agency has three noteworthy advantages:

1. Link Maintenance. Unlike a URL which refers to a resource's physical location and has a (usually undefined) shelf life, a DOI link in principle never changes. Any change to the resource's location must propagate to the resolving agency. A MultiLink feature permits multiple links, which are defined by the resource owner, to be established from a DOI link. This provides an enriched user experience over a URL, which can only point to a single physical location.

2. Access Control. The resolving agency has an OpenURL resolver which permits only those resources for which the user has the right access level to be fetched. In the absence of this facility, the user would need to log on to the system which manages the resource.

3. Discoverability. DOI registration identifies content that has been deemed to be of value and whose validity and currency are actively managed. Hence the use of DOIs for electronic publications potentially improves Google search rankings provided the DOIs resolve to the actual document and not to a DOI "landing page" (see section 4.3.3).

The power of the DOI framework for Internet-based content management coupled with its widespread adoption by publishers has led to an increased uptake outside of scientific publishing, particularly by commercial content companies.

3.4. The IUPAC International Chemical Identifier (InChI). The chemical structure of a compound is its true identifier, and before the recently released International Chemical Identifier (InChI), no satisfactory or open means of serializing it existed. The InChI is the result of an IUPAC project[22] to establish a unique label for chemical substances. It is not designed to be human readable, but freely available supporting algorithms were developed to generate (and with version 1.1 of the software release parse) it.

Whereas the DOI is globally unique and applicable to any resource, the InChI is specific to and constructed from a molecular compound. Hence if two compounds are identical, at least in terms of properties such as connection table and stereochemistry, their InChIs will also be identical. The InChI is also nonproprietary, unlike e.g. the SMILES descriptor, and it relies on a single algorithm to establish a unique canonical label independent from the way in which it was drawn.

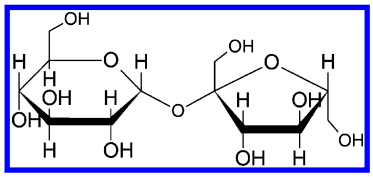

Figure 5. Constitution of sucrose.

The InChI has been adopted by both public and commercial chemistry databases, and the pharmaceutical and chemical industries have been monitoring its development with interest. Chemical structure drawing packages such as ACD/ChemSketch support it. An example of an InChI (InChI = 1/C12H22O11/c13-1-4-6(16)8(18)9(19)11(21-4)23-12(3-15)10(20)7(17)5(2-14)22-12/h4-11,13-20H,1-3H2/t4-,5-,6-,7-,8+,9-,10-,11-,12+/m1/s1) is shown for sucrose (Figure 5).

3.5. WebDAV for File Handling. WebDAV, an IETF standard, stands for Web-based Distributed Authoring and Versioning. Its aim is to make the Web a more writable medium, in line with Tim Berners-Lee's original vision of the Web. The WebDAV framework for manipulating documents on a remote Web server is supported by most operating systems. Files and folders on a remote WebDAV server appear as if they are stored locally. Although its focus has been for general Web-based file management, its versioning aspects have yet to be realized. WebDAV is ideally suited for the file upload capabilities which we require. No bespoke software needs to be written, and its support for drag and drop in particular is very useful.

3.6. RDF Repository − Sesame. A method for managing the XMP was needed and preferably one where Open Source tools can be reused with minimal development. Sesame,[23] an Open Source Java-based framework for storing, querying and reasoning with RDF and RDF Schema met this requirement. Its extensive features include querying in SeRQL, RDQL, and RQL, parsing and writing RDF in several serialization syntaxes, and support for MySQL, PostgreSQL, Oracle, SQL server, and in-memory databases. It can be deployed as an RDF database or as an RDF Java library for embedding in applications.

```
<rdf:RDF xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#' xmlns:iX='http://ns.adobe.com/iX/1.0/'>
 <rdf:Description rdf:about='uuid:d9f55125-8c6d-44f4-b6af-9a45efc4f438'
  xmlns:pdf='http://ns.adobe.com/pdf/1.3/'>
  <pdf:Producer>Acrobat 3.0 Import Plug-in</pdf:Producer>
 </rdf:Description>
 <rdf:Description rdf:about='uuid:d9f55125-8c6d-44f4-b6af-9a45efc4f438'
  xmlns:xap='http://ns.adobe.com/xap/1.0/'>
  <xap:ModifyDate>2005-04-13T14:19:09+01:00</xap:ModifyDate>
  <xap:CreateDate>2003-01-30T10:07:04Z</xap:CreateDate>
  <xap:CreatorTool>Acrobat 3.0 Capture Plug-in</xap:CreatorTool>
  <xap:MetadataDate>2005-04-13T14:19:09+01:00</xap:MetadataDate>
 </rdf:Description>
 <rdf:Description rdf:about='uuid:d9f55125-8c6d-44f4-b6af-9a45efc4f438'
  xmlns:xapMM='http://ns.adobe.com/xap/1.0/mm/'>
  <xapMM:DocumentID>uuid:846a0003-4bf4-49dd-bc03-1f0d8cd3f538</xapMM:DocumentID>
  <xapMM:InstanceID>uuid:4397af35-0086-424d-bab3-7787bca39069</xapMM:InstanceID>
 </rdf:Description>
 <rdf:Description rdf:about='uuid:d9f55125-8c6d-44f4-b6af-9a45efc4f438'
  xmlns:dc='http://purl.org/dc/elements/1.1/'>
  <dc:format>application/pdf</dc:format>
  <dc:title>
   <rdf:Alt>
    <rdf:li xml:lang='x-default'>Electrolytic partial fluorination of organic compounds.
Electrosynthesis of novel hypervalent iodobenzene chlorofluoride derivatives and its application
indirect anodic gem-difluorination</rdf:li>
   </rdf:Alt>
  </dc:title>
  <dc:description>
   <rdf:Alt>
    <rdf:li xml:lang='x-default'>Electrosynthesis of novel hypervalent iodobenzene chlorofluorides
successfully performed for the first time and it was demonstrated that p-methoxyiodobenzene chlorofluo
could be used as a mediator for indirect anodic gem-difluorinafion of dithioa</rdf:li>
   </rdf:Alt>
  </dc:description>
  <dc:identifier>
   <rdf:Alt>
    <rdf:li xml:lang='x-default'>10.1016/0040-4039(96)00951-3</rdf:li>
   </rdf:Alt>
  </dc:identifier>
  <dc:creator>
   <rdf:Seq>
    <rdf:li>Toshiyasu Fujita</rdf:li>
    <rdf:li>Toshio Fuchigami*</rdf:li>
   </rdf:Seq>
  </dc:creator>
 </rdf:Description>
 <rdf:Description rdf:about='uuid:d9f55125-8c6d-44f4-b6af-9a45efc4f438'
  xmlns:rsc_1_1='http://www.inchi.org/1.12/'>
  <rsc_1_1:inchi>
   <rdf:Seq>
    <rdf:li>1.12Beta/C6H5Cl2I/c7-9(8)6-4-2-1-3-5-6/h1-5H
  </rdf:li>
     </rdf:Seq>
    </rsc_1_1:inchi>
   </rdf:Description>
   <rdf:Description rdf:about='uuid:d9f55125-8c6d-44f4-b6af-9a45efc4
    xmlns:UniqueIdentifier='http://www.doi.org/2004/DOISchema/'>
    <UniqueIdentifier:doi>
     <rdf:Alt>
      <rdf:li>10.1016/0040-4039(96)00951-3</rdf:li>
     </rdf:Alt>
  </UniqueIdentifier:doi>
   </rdf:Description>
 </rdf:RDF>
```

Figure 6. Example of XMP metadata created for a journal article.

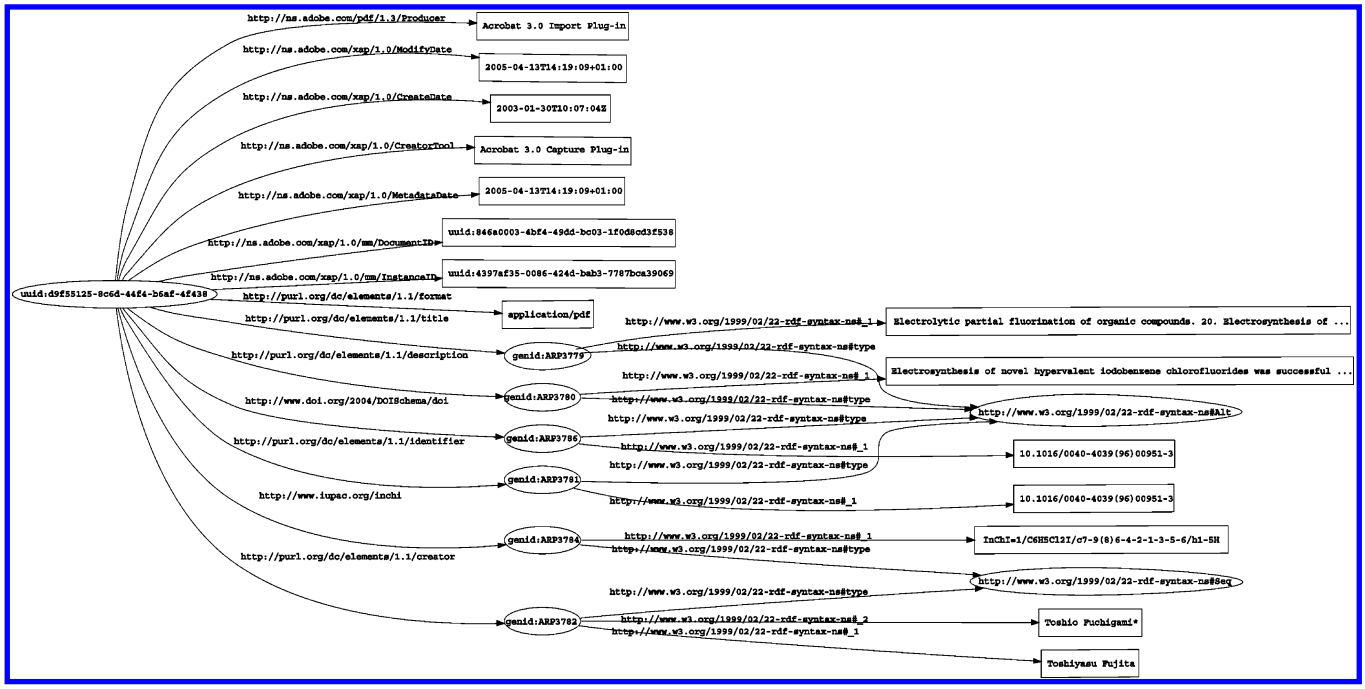

Figure 7. Graphical representation of an XMP example.

One of the main attractions of Sesame is its service-oriented architecture. The Storage and Inference Layer (SAIL) abstracts the storage device used. Between the Client and the SAIL are the services which implement specific functions. These include protocol handlers to deal with HTTP or SOAP requests. The Access API service provides functionality for client applications, either locally or remote. The Repository API enables querying, storing, and extracting RDF in different serialization formats. The Graph API provides a model for representing RDF graphs, enabling the user to perform fine-grained RDF manipulation. Sesame can be deployed as a remote database server or embedded within a desktop application.

3.7. Implementation of the Architecture. *3.7.1. Sample Articles.* The goal of this study is to establish critical RDF-related issues that would need to be resolved before future larger scale studies can be undertaken. Thus for testing purposes a relatively small number of target chemical journal articles sufficed. Each article would require XMP vocabulary populated with Title, Author, Keywords, Abstract, DOI, and InChI. The first four are covered in the standard XMP set, while DOI and InChI required extensions to the XMP. [The XMP property "Description" is a synonym for the term Abstract used in most scientific publications.] Because no electronic articles containing XMP exist, the samples were enriched with XMP using the metadata tool provided in Adobe Acrobat Professional. We deliberately excluded publisher-centric specific metadata such as the journal titles and the names of publishers from the XMP. Instead, the DOI is used to resolve to publishers' Web sites, effectively rendering these Web sites an extension of our Semantic Web model.

Of the eight XMP enriched electronic articles which were prepared, two of them[24,25] have a common author. They are also two of the rare articles which have supporting structural coordinate information from which corresponding InChIs were determined. Iodobenzene dichloride established the commonality of the other six articles.[26−31] They were sourced using a SciFinder[32] substructure search of the molecule. ACD/ChemSketch was used to create its InChI representation. DOIs for the eight articles were easily located on the Web.

3.7.2. Creation and Capture of XMP Metadata. An automated XMP capture tool is not needed for a relatively small sampling of articles but would be a prerequisite for larger scale studies. We initially tried to semiautomate this process with the help of the IsaViz RDF Editor. However, Acrobat Professional could not understand the RDF serialization of the XMP that IsaViz exported. Consequently an Acrobat compliant version of XMP, an example of which is shown in Figure 6, had to be created by manual RDF editing. The XMP was then inserted into the article via the Acrobat Professional metadata tool (Figure 3). A graphical representation (map) of the XMP we use is shown in Figure 7.

3.8. Process Flow in SemanticEye. Figure 8 depicts the overall process flow for SemanticEye. This model expands on components F, G, and H of the dataflow model (Figure 2). Swim lane notation is used to indicate the level in the three-tier architecture within which each step of the process is occurring. For this study, the server components and the Sesame database are installed on the same computer, but they can also run on separate computers if necessary. The following sections elaborate on the steps.

The downstream RDF database which the users navigate will be referred to as the Chemical Journal Ontology. Its RDF structure is simpler to navigate than that of the XMP (Figure 7 vs Figure 9).

3.9. System Architecture. Figure 10 depicts the three-tier system architecture of SemanticEye. As the client tier is totally reliant on Web standards, no client development was carried out. For clarity the various components were grouped within subsystems that are distinguished by color. All of the components within the middle and data tiers have been installed on the same Pentium 4 server running Windows XP Server 2003. The following sections describe the components in more detail.

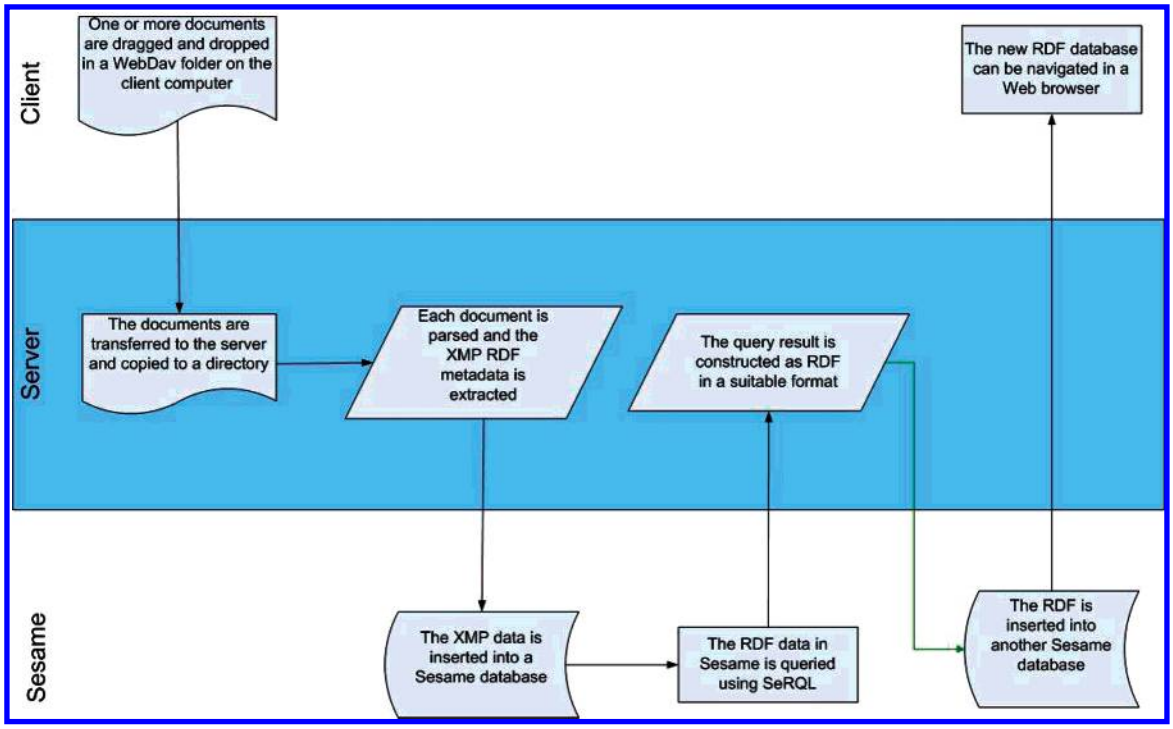

Figure 8. Steps to transfer metadata from the articles to the RDF repository.

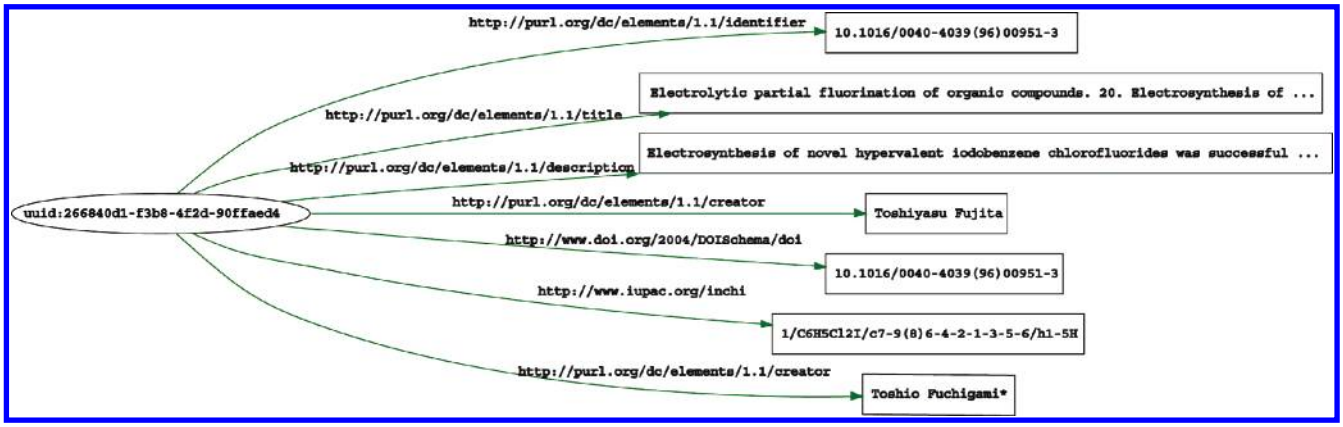

Figure 9. Graphical representation of the RDF structure of the Chemical Journal Ontology.

3.9.1. RDF Management (Figure 10, Orange). Apache Tomcat is the preferred servlet container. It is Open Source and extensively tested with Sesame. Installation and servlet deployment in our environment was relatively straightforward. For RDF triple storage we installed the Open Source relational database, MySQL, on the same server. If necessary, Tomcat and MySQL could be installed on separate platforms. The installation and configuration of the entire Sesame environment was relatively straightforward.

A database within Sesame was created to archive XMP triples. Navigating this database using Sesame's standard browser interface was counterintuitive however. The journal predicates, such as dc:title, do not reference their corresponding objects. Instead, they referenced meaningless "blank" resources which, via meaningless predicates, reference the correct objects.

To remedy this problem, the XMP was repurposed in a separate RDF Ontology where journal predicates directly reference their corresponding literals. Journal metadata could now be navigated in a manner that would not be confusing to the users and (importantly) without the need to customize the Sesame client.

3.9.2. XMP Service Suite (Figure 10, Blue). The XMP Service Suite was a new development effort for this project. The Suite consists of four modules:

(1) SemanticEye Controller. This module, written in Visual C#, runs as a Windows service and monitors a server side target folder for changes. When one or more documents containing XMP are uploaded to a target folder the Controller will instance the XMP Extractor, Sesame XMP Import Client, the Sesame SeRQL Client, and again the Sesame XMP Import Client in sequence.

(2) XMP Extractor. This module, written in Visual C++, is an implementation of the Adobe XMP SDK. It inputs a file path as an argument, extracts the XMP if it exists, and writes the XMP to a file. It outputs the Document ID which is the main subject within the XMP.

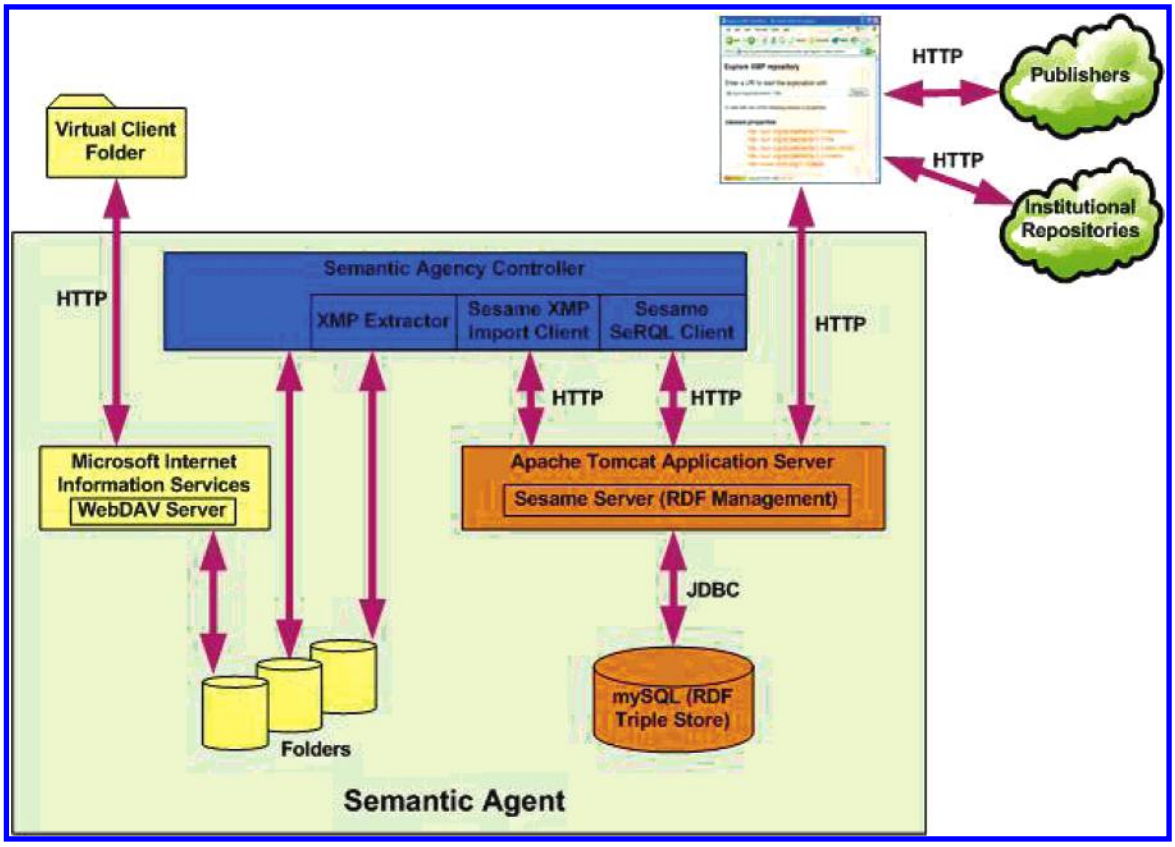

Figure 10. System architecture of SemanticEye.

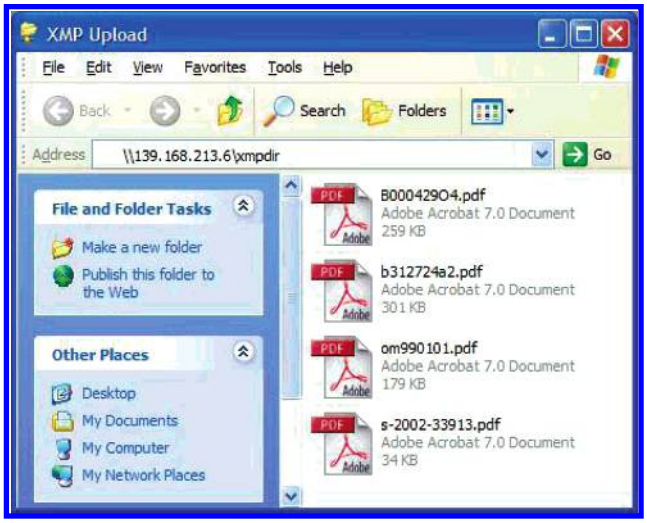

Figure 11. WebDAV folder as it appears on the Windows XP client.

(3) Sesame XMP Import Client. This module, written in Java, will insert an RDF file such as the XMP outputted by XMP Extractor into a specified RDF triple store database within Sesame.

(4) Sesame SeRQL Client. This module, written in Java, will run a specified SeRQL query on a Sesame database and output the result as RDF.

3.9.3. XMP Uploader (Figure 10, Yellow). To enable the sharing of server directories on the client via WebDAV, the Microsoft Internet Information Services (IIS) was installed on the server, and a WebDAV folder was created. This folder was shared on a Windows XP client (Figure 11) via the "Add Network Place" Wizard.

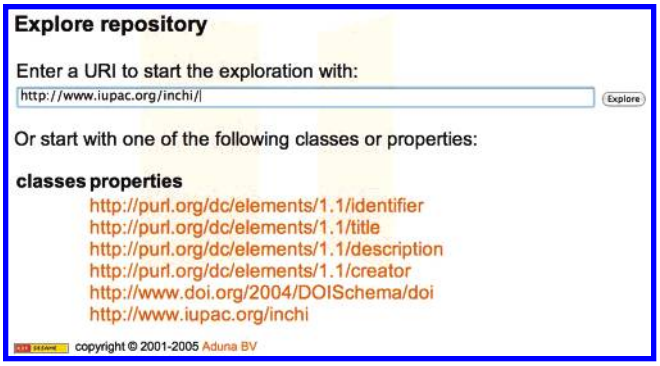

Figure 12. Entry point of the SemanticEye Ontology.

3.10. Sesame Security. In Sesame, a flexible security service interface provides basic access control as well as user and group management. An interesting feature is its support for the import and export of the security setup as RDF. Using its API, custom modules can be created for specific user or group manipulation.

3.11. Navigation of Journals. The figures in this section illustrate a typical workflow that a reader might follow when navigating the SemanticEye ontology. The ontology is essentially an RDF repository with a small sampling of journal metadata and is not restricted to a single journal or publisher. A reader would locate articles of interest simply by navigating whatever metadata the RDF model (Figure 9) contains.

Figure 12 shows the entry page which lists all of the properties, or RDF predicates, in the ontology. An obvious aspect of this example is that each predicate is a fully qualified Uniform Resource Identifier or URI. This would be confusing to users who are not familiar with the concepts

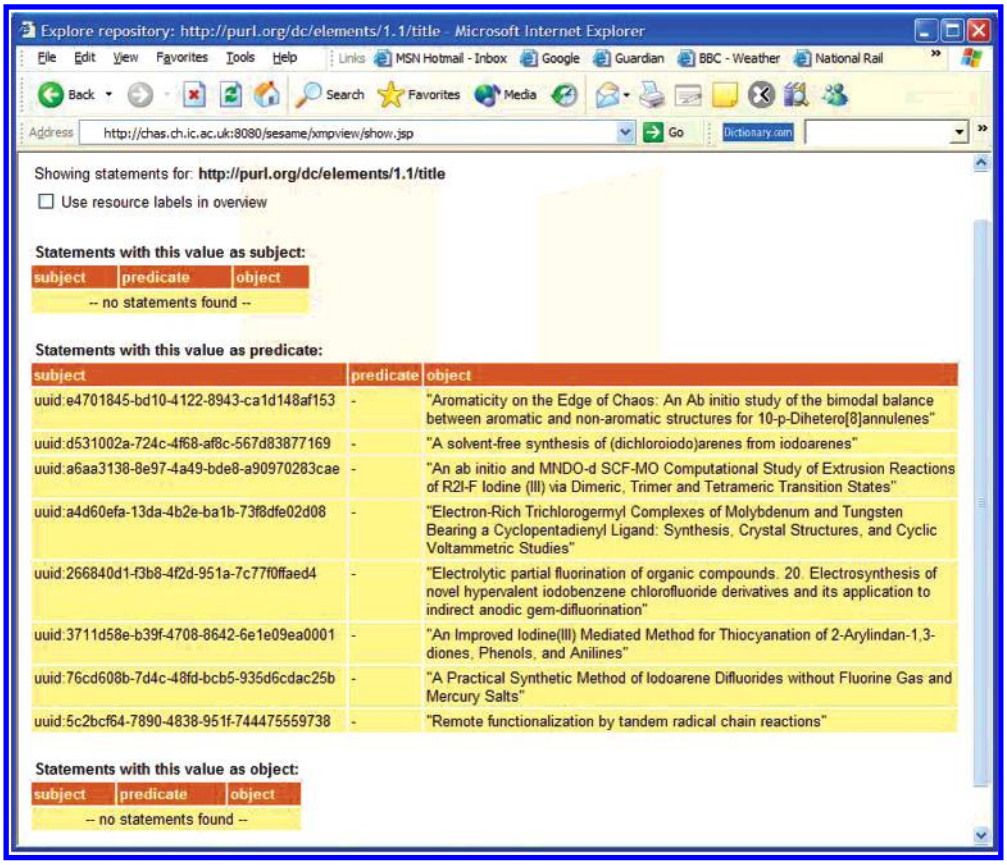

Figure 13. The title predicate is selected to reveal the eight titles in the Ontology.

of namespaces. These users might assume that the predicates, the URIs of which would appear to contain Web addresses of organizations external to SemanticEye, resolve to these Web addresses. However, they actually resolve to corresponding objects within the SemanticEye ontology. For example, selecting http://purl.org/dc/elements/1.1/title would take the user to a list of the eight titles that currently exist in the SemanticEye ontology (Figure 13). The user would not be taken to a site called "purl.org" and retrieve information about titles.

Continuing in the same vein, selecting one of the titles returns all of the metadata associated with it (Figure 14). Selecting any one of the metadata properties will return a list of all of the articles which contain the property. For example if the InChI is selected, all of the articles which contain an identical InChI are listed (Figure 15). A "Get Article" link follows each article entry. This link is the DOI for the article prepended with the URL of the DOI resolving agency, dx.doi.org. By selecting it, the article is fetched from the Internet (Figure 16).

4. RESULTS AND DISCUSSION

SemanticEye adapts the digital music semantic model in a Semantic Web framework. Context and community are applied to chemical electronic journal articles through a small ontology. Building it was relatively straightforward with no insurmountable technical hurdles. Significantly, no RDF Schema or specialist toolkits needed to be built. Instead a variety of Open products were pooled together: (1) Adobe XMP RDF vocabulary, (2)Adobe XMP SDK toolkit, (3) InChI, (4) DOI, (5) WebDAV, (6) Sesame RDF repository, and (7) Sesame SeRQL RDF query language.

SemanticEye is currently a proof-of-concept, the building of which uncovered a number of strengths and weaknesses in the Semantic Web concept which are elaborated in the following sections.

4.1. Metadata Management. The Semantic Web is not immune to the prevailing metadata management challenges. One of the grand challenges is deciding what approach to take for storing the metadata. There are essentially two approaches here:

(1) Embed the metadata within the content as is done with digital music files. The advantage here is ease of implementation and robustness of metadata. There is also no risk of the document and metadata becoming dissociated. The disadvantage is the difficulty in managing embedded metadata and the reliance on clever search algorithms for navigating content.

(2) Manage the metadata separately from the content in a metadata repository. Here metadata and content can have their own respective lifecycles and be managed separately. Content management is improved, metadata manipulation is more flexible, and there is no need for clever search algorithms for content navigation. The disadvantage is that a mechanism is needed to ensure that the document and its metadata remain associated.

SemanticEye uses both approaches. Metadata are embedded in electronic journal articles and are also centrally managed in a separate repository. The linking between the metadata repository and the source documents is achieved via the DOI.

4.2. The SemanticEye Ontology. Because the centrally managed electronic journal metadata constitutes a molecule (InChI) centric knowledge bank, we will brand it as the

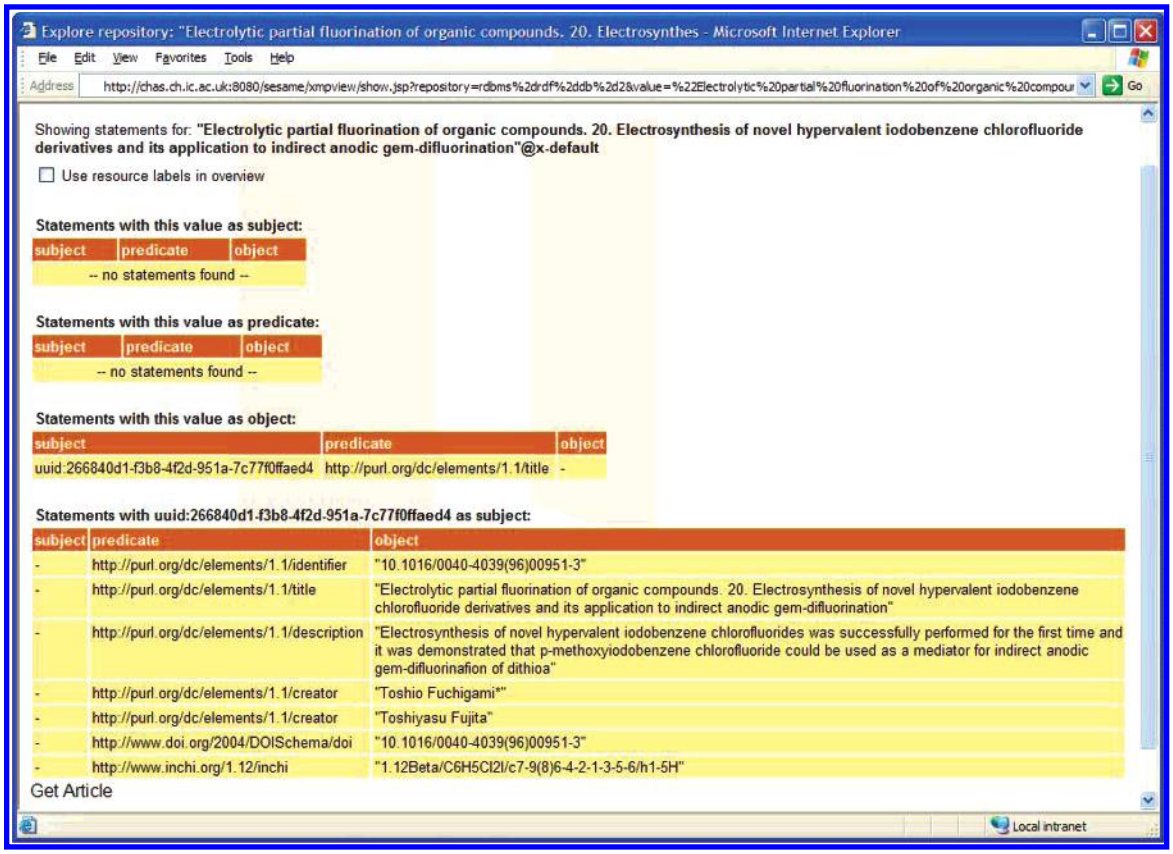

Figure 14. All of the metadata are shown for a Document selected by title.

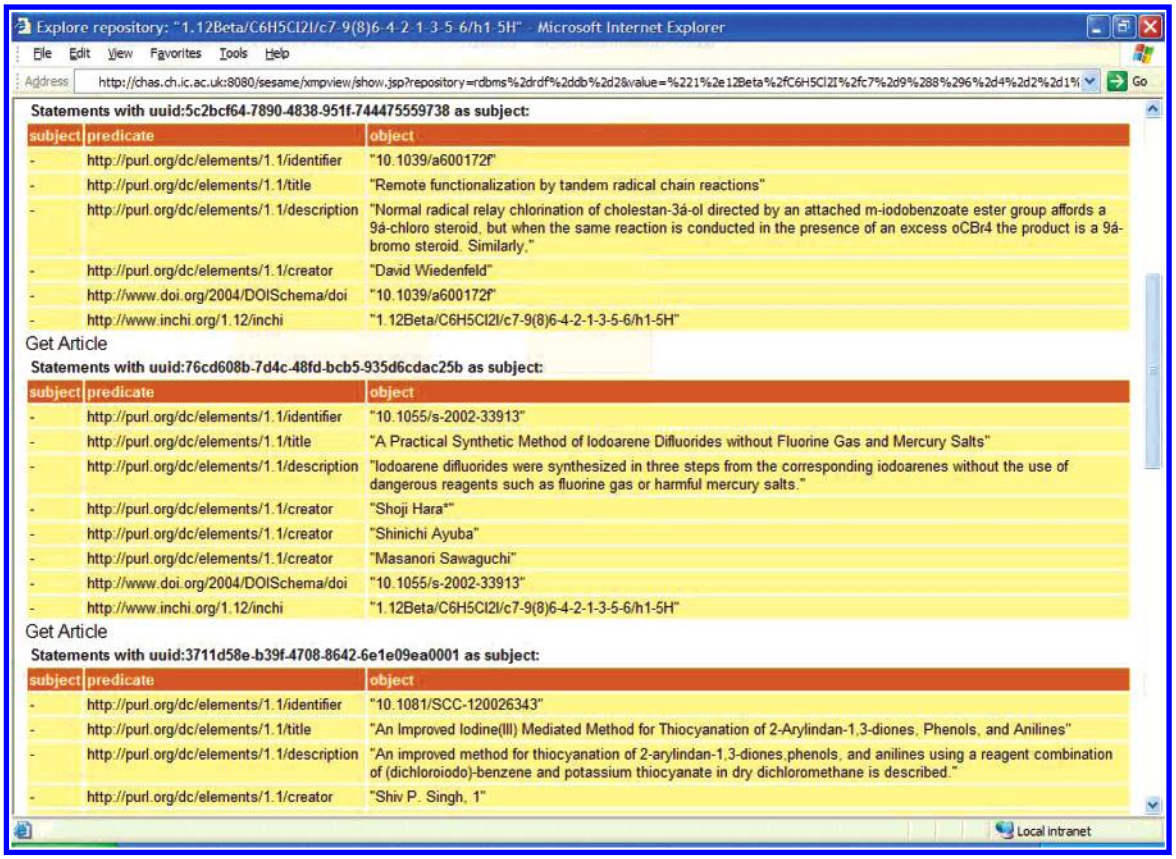

Figure 15. The InChI is selected, and all of the articles which contain this InChI are listed.

SemanticEye Ontology and position it at the Ontology level of the Berners-Lee architecture (Figure 1). In the pursuit of user uptake, we expect it to continually grow and to be continually optimized throughout its lifecycle. Fortunately, it is compact and therefore easier to maintain than bio-ontologies. More importantly, a small ontology is more

Figure 16. The DOI is resolved by selecting "Get Article".

Figure 17. Data flow to create the Chemical Journal Ontology.

readily accepted by a broad user community. A large ontology would require considerably more time to overcome the technical (as well as political) barriers in order to achieve the same level of user acceptance.

As discussed in section 3.8, the automated process of creating the SemanticEye Ontology follows these steps:

(1) Insert the XMP input into the XMP Triple Store.

(2) Run an SeRQL query to export appropriate RDF from the XMP triple store.

(3) Insert the exported RDF into the downstream SemanticEye Ontology.

From Sesame's perspective, the XMP Triple Store and the SemanticEye Ontology are identical RDF triple stores distinguished only by the different RDF vocabularies that are inserted into them. The data flow (Figure 17) exemplifies the importance of a powerful RDF query language, such as SeRQL, to construct an optimized ontology.

We anticipate that in the future SemanticEye will need to be refined in order to support new or changing scientific workflows. SemanticEye must however remain chemical electronic publishing centric. In those cases where enhancing the SemanticEye Ontology would be out of scope, new ontologies could be created to meet the needs of particular scientific research functions.

The Algorithms Directory, maintained by the Blue Obelisk Movement,[33] is arguably compatible with the SemanticEye Ontology. The Algorithms Directory is essentially an XML document within which many chemoinformatics algorithms are listed along with references to original literature. However, because it is algorithm centric, it requires its own

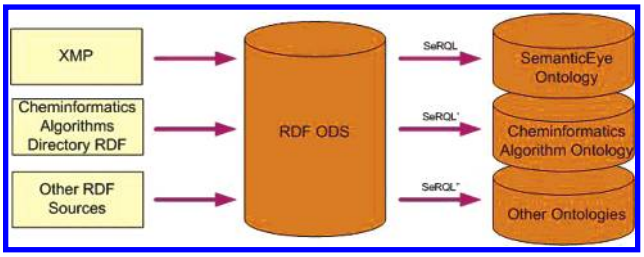

Figure 18. Future data flow.

ontology. Supporting this scenario within the SemanticEye framework would be trivial provided the Algorithms Directory XML is converted into RDF.

In theory, any RDF vocabulary can be loaded into an RDF database. Without doing any additional technical modifications, the XMP Triple Store would evolve into a central RDF Operational Data Store (ODS) that would take as input a variety of RDF schema and not be restricted to XMP. Each downstream ontology would have an SeRQL query associated with it. Figure 18 depicts how the data flow is likely to evolve.

4.3. Validation and Trust in SemanticEye. As mentioned in section 1.5, readers using SemanticEye would locate articles of interest not through a search query but by browsing a collection of articles in a single journal. The SemanticEye Ontology must attain this user experience or, better yet, improve upon it to build up readership. We have demonstrated that the Semantic Web can provide the required technology framework. However, over the years publishers have mastered the editorial process of organizing articles into a collective whole, and readers *trust* this process. SemanticEye would need to attain this same level of trust which, in the Semantic Web framework (Figure 1), means striving for the top level of the Berners-Lee architecture now.

We determined that attaining trust requires, for the metadata and their relationships, a validation process which

meets the following requirements:

(1) Scientific publishers must mandate the inclusion of metadata within all manuscripts.

(2) The capture of metadata must be automated and validated.

(3) A means of uniquely identifying authors is needed.

(4) The DOI must always resolve to its corresponding electronic publication.

(5) The InChI must resolve to an authoritative resource.

Without the first requirement being met, a system such as SemanticEye could never progress beyond proof of concept status. Publishers must be convinced of the value that this activity brings to their readership and, at least as importantly, to themselves. The following sections discuss how all of the other requirements can be met.

4.3.1. Metadata Capture. The growth of the Semantic Web will be fostered by each repository developer and content creator, who would be responsible for tagging up the content with metadata. SemanticEye would not be viable without this process. However, it has long been known that such an approach is problematic. Content creators tend to tag information resources with an implicit sense of how they themselves would use the resources and would not necessarily appreciate the importance of unique identifiers and metadata validation for the benefit of the wider community. A solution to this problem is the provision of a suitable tool to assist the content owner. In the case of SemanticEye, this tool would need to create the XMP metadata enhanced with InChI and DOI objects. Because most manuscripts are created with Microsoft Word, we investigated how Word handles metadata. We found that the metadata does not map to Adobe XMP counterparts or maps incorrectly. Extending Adobe's XMP capture tool (Figure 3) has been ruled out as it would require the purchase of one or more Adobe authoring applications. A freely available XMP authoring tool is needed. This tool requires enough functionality to extract most, if not all, of the required metadata from the manuscript and create the InChIs from molecular structure information.

The Experimental Data Checker[34] (OSCAR), a collaborative effort between Cambridge University and the RSC, follows a guiding principle that experimental molecular data is published in a consistent manner and does not vary much between journals. Regular expression parsing is performed on the manuscript, patterns and phrases in free text are identified, and some checks are then run to test the data for consistency. It can find molecular names and associate them with structures. The reported success rate of 92% is impressive. Enhancing the Experimental Data Checker for XMP capture will therefore be part of a follow up investigation.

As to who would ultimately do the tagging, a trained manuscript editor is obviously better qualified than the manuscript's author. If the manuscript has been submitted to a publisher, the manuscript editor would validate and ensure that the metadata is handled consistently in all the manuscripts. For manuscripts submitted to an Open Access journal where editorial control is minimal, the burden of tagging would most likely be borne by the manuscript author.

4.3.2. Authorship and Digital Signatures. It is common practice to locate articles associated with one or more authors. Authorship substantiates the subject matter and helps determine the reader's level of trust in an article. Validating the author within metadata is particularly challenging however. There is no rule explicitly stating how an author's name should be represented in a journal. John Paul Gardner, John P. Gardner, and J. P. Gardner are equally valid. A more rigorous metadata approach could be applied whereby first, middle, and last names each had their own metadata. This does not completely solve the problem because different authors can have the same name.

The only valid solution is for each author to have a unique and unchanging identifier. There are a number of ways in which an author can be uniquely identified, a national insurance number being one possibility. However adopting any existing unique identifier for electronic publishing has sociological barriers that might be impossible to surmount.

The Digital Signature is intrinsic to the Semantic Web, and its importance has been expounded in the literature. An article describing its use in chemistry has been published.[35] A Digital Signature is unique to an author and certifies that the article comes from a trusted source. Hence, if all authors were to "digitally sign" their manuscripts, the problem of uniquely identifying authors becomes a technical rather than sociological challenge. Surprisingly however, very few authors currently sign their manuscripts. This is undoubtedly going to change in the face of the mounting security concerns of the Internet. We need to somehow catalyze this change.

Digital signatures come in three types: enveloped, enveloping, and detached. An enveloped signature is embedded within the document. An enveloping signature embeds the document within it. A detached signature is separate from the document being signed.

The XML Signature, or XMLDsig,[36] is a detached signature vocabulary. Although XMLDsig is the preferred Digital Signature type for the Semantic Web, PDF only supports the enveloped type and does not store Digital Signature information within XMP. Further investigation is needed to reconcile this disparity in order to include digital signatures either within the XMP or within separate Named Graphs.[37] At the same time publishing scientists need to be educated on the importance of the Digital Signature in order to help increase its uptake.

4.3.3. The Validation of DOIs. SemanticEye does not manage any documents. Instead, it manages a collection of DOIs which, if the requestor has the appropriate security credentials, should resolve to documents managed elsewhere. DOIs provide document persistence which not only is vital for metadata management but also opens up automated information analysis opportunities. As discussed previously, the consistent structure of a chemical journal article makes the extraction of useful chemical information from it straightforward. Being able to automatically fetch articles would permit information analysis on the article by other agents.

However, the manner in which many publishers manage DOIs provides a serious shortcoming. DOIs often resolve to intermediate "landing pages" where some user action is required for the desired article to be fetched. An example of such an action would be selecting between an HTML or PDF rendition of the article. An agent would not be able to perform this action without policies which the agent would have to understand and to which all of the publishers can adhere. Ideally the DOI authority would apply governance to ensure validation of the DOIs.

4.3.4. An InChI Agent. As the InChI is the key object for establishing context and community within the SemanticEye ontology, SemanticEye can be considered as an InChI agent that can classify scholarly publications via molecular structures. Being an InChI agent should not imply that SemanticEye will incorporate any structure analysis algorithms. SemanticEye's scope is electronic publishing, and it treats the InChI purely as a unique molecule identifier. Of course, other agents are not precluded from using the documents fetched via the InChI, or the InChI itself, for analysis purposes.

Currently there is no means of resolving an InChI to corresponding molecular structure data within an appropriate collection such as PubChem. As the uptake of the InChI increases, we anticipate that such a facility will become available and linking SemanticEye to it should be straightforward.

4.4. Extension of SemanticEye with Other Frameworks. Currently SemanticEye can only be navigated by a person. Its functionality is not exposed through a Web Service in a manner that would enable navigation by a Web agent. Although the Sesame architecture supports the development of Web Services, user uptake and feedback of SemanticEye is needed before such a Web Service can be built. Nevertheless it would be worthwhile understanding the existing applications that could potentially integrate to a SemanticEye Web agent. We have identified three other Web-based frameworks which potentially contain related information.

4.4.1. The Open Archives Initiative. The goal of the Open Archives Initiative (OAI)[38] is to deliver an interoperability framework for institutional repositories. Although closely related to the Open Access[6] movement, the OAI technology and standards are much more broadly applicable than scholarly publishing alone. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) defines a mechanism for content owners to expose their metadata. OAI-PMH mandates that individual archives map their metadata to the Dublin Core metadata standard which is supported within XMP. URNs are assigned to OAI-Identifiers which can be resolved by an agent similar to the DOI resolving agent.

The OAI framework is appropriate for data that supports the results of journal articles but would not be included with them. The OAI would therefore be a natural extension to SemanticEye. Although few chemical journal articles currently reference supporting data, the increased use of institutional archives is likely to change this. Authors would need to store supporting data in institutional archives and to provide OAI-Identifiers for the data within the manuscript.

4.4.2. The Semantic Grid. The Grid is a framework of standards and technologies for connecting Internet- and Intranet-based resources. It has two principle categories. The Compute Grid enables applications requiring computational horsepower to run on multiple computers which would otherwise be idle. The Data Grid enables resources from multiple organizations to be drawn together into specialist "virtual organizations". Although the Grid is still largely within the auspices of academic research, the number of commercial Grid products is steadily growing.

Metadata management is a main constituent of the Data Grid. One research initiative called the Semantic Grid[39] attempts to address this aspect with the help of Semantic Web technologies. A related publication[40] presents an integration example where a DNA Grid and a Protein Grid are linked via Semantic Web repositories. SemanticEye should exploit the Semantic Grid if the integration to the scientific resources which it exposes is straightforward.

4.4.3. The Wiki. *""RDF and the Wiki-principle are a perfect match. The takeoff of the semantic web is slowed by the need for trust. Anybody can write information on the web, but it is hard to see which information is indeed correct. Wikipedia has shown that the collaborative editing of articles leads to better quality and less disinformation.""*[41]

Several projects are currently investigating the intersection between the Semantic Web and the Wiki.[42] Mediawiki RDF[43] and Platypus Wiki[44] enhance the Wiki to handle Semantic Web vocabularies. OntoWiki[45] uses standard Wiki technology as a platform for community driven ontology building and maintenance. SemanticEye could be a beneficiary of these projects. If the SemanticEye ontology were repurposed as a Wiki, its metadata could be edited by users, and it could potentially align to related Wikis such a Chemistry Wiki.[46] Further investigation here is required.

5. CONCLUSION

One of the mains goals of the SemanticEye project is to help initiate the inclusion of metadata within scientific electronic publications by exemplifying the payback to the research community. We realize however that adding metadata to manuscripts would always be considered a low priority by the majority of scientists irrespective of any future payback. The success of SemanticEye therefore hinges on the proactive involvement of publishers.

By restricting SemanticEye's scope to electronic publishing we are admittedly focusing on only a small portion of content on the Web. However it is a high value portion with a well understood readership. An unfocused exploratory project is not only likely to lead to failure but is also likely to add to the confusion and scepticism surrounding the Semantic Web.[47] It is this confusion which explains why uptake of the Semantic Web has been slow and restricted to a few vertical domains. Hence from our experience we would propose the following steps for any new Semantic Web undertaking:

(1) Clearly identify an unsolved problem and focus exclusively on it.

(2) Model a solution on a small scale using existing Semantic Web technologies.

(3) Establish the correct process and policies for the solution.

(4) Scale-up the solution for the target audience.

In the case of SemanticEye, the unsolved problem it is trying to address is the lack of context and community of an electronic journal compared to its hard copy equivalent. We modeled the solution by introducing two metadata objects, InChI and DOI, into existing Semantic Web technologies and tested our model with a small sample set of articles. A scalable process has been established, but before moving on to the final step we need to implement metadata validation.

A fundamental sociological problem must be also addressed. That is, a rigorous enforcement of policies is required by Web agents to avoid the incorrect use by subscribing resources. Otherwise, the agents might ultimately

be usable only by humans and not by programs which would negate the principle goal of the Semantic Web.

Acronyms Used within this Paper. RDF − Resource Description Framework, XMP − Extensible Metadata Platform, SeRQL − Sesame RDF Query Language, WebDAV − Web-based Distributed Authoring and Versioning, InChI − International Chemical Identifier, and DOI − Document Object Identifier.

REFERENCES AND NOTES

(1) Tenopir, C.; King, D. W. Precis of 'Towards Electronic journals: realities for Scientists, Librarians, and Publishers', 2000. Psycoloquy. http://psycprints.ecs.soton.ac.uk/archive/00000084/#html (accessed Jan 23, 2006).

(2) James, D.; Whitaker, B. J.; Hildyard, C.; Rzepa, H. S.; Casher, O.; Goodman, J. M.; Riddick, D.; Murray-Rust, P. The Case for Content Integrity in Electronic Chemistry Journals: The CLIC Project. *New. Rev. Inf. Networking* **1995**, *1*, 61−70.

(3) Bachrach, S. M.; Burleigh, D. C.; Krassivine, A. Designing the Next-Generation Chemistry Journal: The Internet Journal of Chemistry. *Issues Sci. Tech. Librarian* [Online] **1998**, *17*, Article 1. http://www.library.ucsb.edu/istl/98-winter/article1.html (accessed Jan 28, 2006).

(4) Goodman, J. M. Chemistry on the world-wide-web: a ten year experiment. *Org. Biomol. Chem.* **2004**, *2*, 3222−3225. DOI: 10.1039/b409956g.

(5) Smith, J. W. T. The Deconstructed Journal − a New Model for Academic Publishing. *Learned Publishing* **1999**, *12*, 79−91.

(6) Budapest Open Access Initiative. http://www.soros.org/openaccess (accessed Feb 2, 2006).

(7) lanl.arXiv.org e-Print archive mirror. http://xxx.lanl.gov/ (accessed Feb 2, 2006).

(8) Weller, A. C. Can editorial peer review survive in a digital environment? In *ACS Division of Chemical Information (CINF)*, Proceedings of the 228th ACS National Meeting and Exposition, Philadelphia, PA, 2004. http://acscinf.org/docs/meetings/228nm/presentations/228nm14.pdf (accessed Jan 16, 2006).

(9) Murray-Rust, P.; Rzepa, H. S.; Stewart, J. J. P.; Zhang, Y. A global resource for computational chemistry. *J. Mol. Model*. **2005**, *11*, 532−541. DOI: 10.1007/s00894-005-0278-1.

(10) DSpace Federation. http://www.dspace.org/ (accessed Feb 5 2006).

(11) King, D. W.; Tenopir, C. 25 Year trends in information seeking and reading patterns of chemists. In *ACS Division of Chemical Information (CINF)*, Proceedings of the 228th ACS National Meeting and Exposition, Philadelphia, PA, 2004. http://acscinf.org/docs/meetings/228nm/presentations/228nm05.pdf (accessed Jan 16, 2006).

(12) Berners-Lee, T.; Hendler, J.; Lassila, O. The Semantic Web. *Sci. Am.* **2001,** *5*, 34−43.

(13) Taylor, K. R.; Gledhill, R. J.; Essex, J. W.; Frey, J. G. Bringing Chemical Data onto the Semantic Web. *J. Chem. Inf. Model.* **2006**, *46*, 939−952. DOI: 10.1021/ci050378m.

(14) Berners-Lee, T. The Semantic Web and Challenges, 2003. World Wide Web Consortium. http://www.w3.org/2003/Talks/01-sweb-tbl/ (accessed Feb 21, 2006).

(15) Lassila, O.; Swick, R. Resource Description Framework (RDF) Model and Syntax Specification, 1999. World Wide Web Consortium. http://www.w3.org/TR/PR-rdf-syntax (accessed Feb 27, 2006).

(16) Brennan, K.; Petrosillo, S. Demystifying the Semantic Web: Is Migration Right for You?, 2003. Creative Behavior. http://www.creativebehavior.com/index.php?PID=87 (accessed Mar 5, 2006).

(17) Howison, J.; Goodrum, A. Why can't I manage academic papers such as MP3s? The evolution and intent of Metadata standards. In Proceedings of Colleges, Code and Copyright, ACRL Publications in Librarianship no. 57, College Park, MD, 2004; ACRL: Chicago, 2005.

(18) Adobe XMP: Adding Intelligence to Media. http://www.adobe.com/products/xmp/ (accessed 28 May, 2004).

(19) PRISM: Publishing Requirements for Industry Standard Metadata. http://www.prismstandard.org/ (accessed Feb 19, 2006).

(20) Murray-Rust, P.; Mitchell, J. B. O.; Rzepa, H. S. Communication and re-use of chemical information in bioscience. *BMC Bioinformatics* [Online] **2005**, *6*, Article 180. http://www.ch.ic.ac.uk/rzepa/bmc2/ (accessed Dec 11, 2005). DOI: 10.1186/1471-2105-6-180.

(21) The Digital Object Identifier System. http://www.doi.org/hb.html (accessed Feb 10, 2006). DOI: 10.1000/182.

(22) IUPAC. http://www.iupac.org/projects/2000/2000-025-1-800.html (accessed Feb 12, 2006).

(23) Broekstra, J.; Kampman, A.; van Harmelen, F. Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. In *Lecture Notes in Computer Science*, Proceedings of the First International Semantic Web Conference, Sardinia, Italy, 2002; Horrocks, I., Hendler, J. A., Eds.; Springer: Berlin, 2002; pp 54−68.

(24) Martín-Santamaría, S.; Carroll, M. A.; Pike, V. W.; Rzepa, H. S.; Widdowson, D. A. Fluoridation of Heteroaromatic Iodonium Salts: Experimental Evidence Supporting Theoretical Prediction of the Selectivity of the Process. *J. Chem. Soc., Perkin Trans. 2* **2000**, *2*, 2158−2161. DOI: 10.1039/b000429o.

(25) Rzepa, H. S.; Sanderson, N. Aromaticity on the Edge of Chaos: An Ab initio study of the bimodal balance between aromatic and non-aromatic structures for 10-π-Dihetero[8]annulenes. *Phys. Chem. Chem. Phys*. **2004**, *6*, 310−313. DOI: 10.1039/b312724a.

(26) Wiedenfeld, D. Remote functionalization by tandem radical chain reactions. *J. Chem. Soc., Perkin Trans. 1* **1997**, *3*, 339−347. DOI: 10.1039/a600172f.

(27) Zielinska, A.; Skulski, L. A solvent-free synthesis of (dichloroiodo)-arenes from iodoarenes. *Tetrahedron Lett.* **2004**, *45*, 1087−1089. DOI: 10.1016/j.tetlet.2003.11.071.

(28) Fujita, T.; Fuchigami, T. Electrolytic Partial Fluorination of Organic Compounds. 20.1 Electrosynthesis of Novel Hypervalent Iodobenzene Chlorofluoride Derivatives and Its Application to Indirect Anodic gem-Difluorination. *Tetrahedron Lett.* **1996**, *37*, 4725−4728. DOI: 10.1016/0040−4039(96)00951−3.

(29) Prakash, O.; Kaur, H.; Pundeer, R.; Dhillon, R. S.; Singh, S. P. An Improved Iodine(III) Mediated Method for Thiocyanation of 2-Arylindan-1,3-diones, Phenols, and Anilines. *Synth. Comm*. **2003**, *33*, 4037−4042. DOI: 10.1081/SCC-120026343.

(30) Filippou, A. C.; Winter, J. G.; Kociok-Köhn, G.; Troll, C.; Hinz, I.; Electron-Rich Trichlorogermyl Complexes of Molybdenum and Tungsten Bearing a Cyclopentadienyl Ligand: Synthesis, Crystal Structures, and Cyclic Voltammetric Studies. *Organometallics* **1999**, *18*, 2649−2659. DOI: 10.1021/om990101+.

(31) Sawaguchi, M.; Ayuba, S.; Hara, S. A Practical Synthetic Method of Iodoarene Difluorides without Fluorine Gas and Mercury Salts. *Synthesis* **2002**, *13*, 1802−1803. DOI: 10.1055/s-2002−33913.

(32) SciFinder. http://www.cas.org/SCIFINDER/scicover2.html (accessed Mar 8, 2006).

(33) Guha, R.; Howard, M. T.; Hutchison, G. R.; Murray-Rust, P.; Rzepa, H. S.; Steinbeck, C.; Wegner, J.; Willighagen, E. L. The Blue Obelisk − Interoperability in Chemical Informatics. *J. Chem. Inf. Model.* **2006**, *46*, 991−998. DOI: 10.1021/ci050400b.

(34) Townsend, J. A.; Adams, S. E.; Waudby, C. A.; de Souza, V. K.; Goodman, J. M.; Murray-Rust, P. Chemical documents: machine understanding and automated information extraction. *Org. Biomol. Chem.* **2004**, *2*, 3294−3300. DOI: 10.1039/b411033a.

(35) Gkoutos, G. V.; Murray-Rust, P.; Rzepa, H. S.; Wright, M. Chemical Markup, XML and the World-Wide Web. Part III: Towards a signed semantic Chemical Web of Trust. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1124−1130. DOI: 10.1021/ci000406v.

(36) IETF/W3C XML-DSig Working Group. http://www.w3.org/Signature/.

(37) Carroll, J. J.; Bizer, C.; Hayes, P.; Stickler, P. Named Graphs, Provenance and Trust. In *International World Wide Web Conference*, Proceedings of the 14th International Conference on World Wide Web, Chiba, Japan, 2005; Ellis, A., Hagin, T., Eds.; ACM Press: New York, U.S.A., 2005; pp 613−622. DOI: 10.1145/1060745.1060835.

(38) Open Archives Initiative. http://www.openarchives.org/ (accessed Mar 25, 2006).

(39) De Roure, D. Semantic Grid. http://www.semanticgrid.org/ (accessed Mar 14, 2006).

(40) Brown, M. C. What is the semantic grid?, 2005. IBM developerWorks.

(41) http://www-128.ibm.com/developerworks/grid/library/gr-semgrid/ (accessed Mar 14, 2006).

(42) Wikimedia Meta-Wiki − RDF Metadata. http://meta.wikimedia.org/wiki/RDF_metadata (accessed Mar 29, 2006).

(43) Wikipedia. http://en.wikipedia.org/wiki/WikiWiki (accessed Mar 29, 2006).

(44) Wikimedia Meta-Wiki − RDF. http://meta.wikimedia.org/wiki/RDF (accessed Mar 29, 2006).

(45) Campanini, S. E.; Castagna, P.; Tazzoli, R. Platypus Wiki: A Semantic Wiki Wiki Web. Presented at the 1st Italian Semantic Web Workshop. [Online], Ancona, Italy, 2004; Semantic Web Applications and Perspectives (SWAP). http://semanticweb.deit.univpm.it/swap2004/ (accessed Mar 27, 2006).

(46) Hepp, M.; Bachlechner, D.; Siorpaes, K. OntoWiki: Community-driven Ontology Engineering and Ontology Usage based on Wikis. In *International Symposium On Wikis*, Proceedings of the 2005 International Symposium on Wikis, San Diego, CA, 2005; Riehle, D., Ed.; ACM Press: New York, 2005.

(47) Chemistry − Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Chemistry (accessed Mar 30, 2006).

(48) Marshall, C. C.; Shipman, F. M. Which Semantic Web? In *Conference on Hypertext and Hypermedia*, Proceedings of the 14th ACM Conference on Hypertext and Hypermedia, Nottingham, U.K., 2003; Ashman, H., Brailsford, T., Carr, L., Hardman, L., Eds.; ACM Press: New York, 2003; pp 57−66. DOI: 10.1145/900051.900063.

CI060139E