

ChemSem: An Extensible and Scalable RSS-Based Seminar Alerting System for Scientific Collaboration

Henry S. Rzepa,* Andrew Wheat, and Mark J. Williamson

Department of Chemistry, Imperial College of Science, London SW7 2AY, United Kingdom

Received September 19, 2005

A seminar announcement system based on the extensive use of XML-based data structures, CML/MathML for carrying more domain-specific molecular content, and open source software components is described. The output is a resource description framework (RDF) site summary (RSS) feed, which potentially carries many advantages over conventional announcement mechanisms, including the ability to aggregate and then sort multiple and diverse RSS feeds on the basis of declared metadata and to feed into RDF-based mechanisms for establishing links between different subject areas.

1. INTRODUCTION

The 10 years since the introduction of Web-based processes have had an enormous impact on almost all aspects of scientific communication and collaboration. The list includes examples such as on-line journals and database searching, electronic conferences, grid-based computing and e-science, e-mail discussion lists, blogs, and wikis. Unfortunately, this proliferation has not been matched by the development of metadata descriptions designed to introduce context, relevancy, and semantic relationships to all this information and data. This vacuum has to some extent been filled by the introduction of increasingly sophisticated search engines such as Google, in which the ranking of hits resulting from any given (text-based) search term appears to be mediated on the basis of incoming links from other pages, rather than replying pre-eminently on the presence of any discretely declared metadata on which to base the rankings (this may also be due to the historical misuse of metadata to achieve artificially enhanced site visibility). Resolving any particular item from, for example, a Google search can be something of a lottery, because the retrieved document may be cast in a number of forms (HTML, Acrobat, Powerpoint, a raster image), and there is no certainty of finding any particular type of data in such documents. By data, we mean reusable components which can, if necessary, be further analyzed or processed using, for example, more specialized software. In this article, we take as our example *chemical data*. Goodman¹ has reviewed the 10 year history of chemistry on the Web and has concluded that such reusability, or as he puts it, “the power of the internet to exploit molecules”, has yet to be fully realized.

It is in this context that we started to investigate an alternative to the classical “Web page” for providing reusable data and information for use by both a human reader and potentially automated software agents. RSS 1.0 (RDF site summary; RDF = resource description framework)² is a metadata-driven technology to which a reader (or software) can subscribe; thereafter, information from the subscribed “feed” can be delivered (“pushed”) to an RSS client on a regular basis. At the client end, it can then be suitably sorted and filtered according to their needs. The use of RSS differs from the manner in which most, for example, Google-like searches are delivered in one key regard; the latter is most

frequently an on-demand initiated user action (information pull), whereas RSS is more commonly automatically driven (“pushed”) by the initial subscription. Because RSS is XML-based, it can be readily extended to handle structured and domain-specific information and, in some measure, directly addresses Goodman’s concern regarding the exploitability of chemical and molecular information. We have shown² how the addition of Chemical Markup Language (CML) components to a conventional RSS feed can achieve precisely this, illustrating it with the construction of a simple chemical inventory system we called ChemStock.³ It struck us that such a system was in fact ideally suited to one of the few remaining areas of scientific communication largely untouched by the Web-based information revolution of the past 10 years, that of the scientific seminar.

Scientific seminars in fact arguably predate even journals as mechanisms for communicating the latest advances in a topic; most early journals indeed originated from a need to document discourses delivered to an audience. At a typical large scientific institution, some 50–100 talks may be delivered in a single week on a vast variety of topics, and the multidisciplinary nature of modern research could mean that a significant number of these might have some item of interest to many chemists. Typically, however, advance notice of such seminars is disseminated by a variety of (arguably) unsuited mechanisms. These include small (A4-sized) printed fliers attached to prominent thoroughfares in a building (notice boards, entrance doors, etc.), postings by e-mail to small groups of “interested individuals”, and only rarely via mechanisms that reach beyond the walls of an average department into other departments, or in exceptional cases, into other institutes. Where seminar programs are posted to, for example, a Web page, search engines may only index the contents *after* the seminar has taken place, and the need to visit many such sites means the average reader checks only a few key lists. There really are no effective mechanisms to aggregate and scale such information beyond a department or institution and no standard mechanisms for adding useful metadata beyond simply the title of the talk, the speaker, and the date and location of the presentation. Such data is essential to enable connections between diverse seminars to be made, the interdisciplinary cross-fertilization so much a part of modern science. Overtly chemical information (e.g.,

Chart 1

```

<Directory "/disk1/www/htdocs/pss/">
  Options FollowSymLinks
  AuthType Basic
  AuthName "Seminar Admin"
  AllowOverride None
  Order deny,allow
  Deny from all
  Allow from ic.ac.uk
  AuthLDAPEnabled on
  AuthLDAPStartTLS on
  AuthLDAPURL
ldap://unixldap.cc.ic.ac.uk/ou=everyone,dc=ic,dc=ac,dc=uk?uid?sub?(objectClass=posixAc
count)
  AuthLDAPAuthoritative on
  require valid-user
</Directory>

```

a formal definition of the molecular structure of any species which might be the topic of, say, a talk with chemical content) is almost entirely absent, so again, molecular connections between different seminars will not be enabled. Part of the issue here, of course, is the evanescent nature of such presentations. In some sciences, information presented in seminars is deliberately short-term and incomplete, being intended merely to stimulate discussions rather than representing a longer-term record. The converse is true in other areas, where seminars are at least as highly regarded as formal publications. Because of these often subtle social issues, the task of distributing seminar meta-information has not hitherto been exposed to technological enhancement in the way that journal publishing has.

We decided, therefore, to address some of these issues by designing a database-driven system for creating a distributed mechanism for creating seminar content and metadata descriptions within a group, a department, a faculty, or an institute, and with output that takes the form of an RSS-based newsfeed. Such newsfeeds could be aggregated on as large a scale as might be necessary and could provide much semantically rich reusable data and information, so completely lacking in more conventional mechanisms. Here, we describe the essential features of such a system which we call **ChemSem**, which we offer to the community as open source.

2. THE DATABASE AND ADMINISTRATION INTERFACE

We have previously described in some detail for ChemStock³ how a MySQL database can be constructed, together with a PHP-based data entry and administration system (PHP = Hypertext Preprocessor). Here, we describe only the specific features of the ChemSem system. As before, it is based on open source components, which include (1) the Apache Web server,⁴ configured to support modules including (2) the PHP system for building an interface to the database, augmented using the PEAR package library toolset⁵ and the SMARTY template engine;⁶ (3) the MySQL relational database;⁷ and (4) the lightweight directory application protocol (LDAP).⁸

The introductory screen to the system is designed to be hierarchical, with information presented according to the access rights of the user. Thus, a casual visitor with no login access is presented with read-only information, whereas at the other extreme, the SuperUser administrator has global permissions, which enable them to (1) create new groups which have a discrete seminar program, (2) create lists of venues for seminars on the basis of the entered groups, (3)

and specify users who will administer these groups, granting them appropriate permissions. Each administration page is itself created from PHP templates and MySQL tables, using a URL of the type https://site_name/chemsem/index.php?page=admin.

Secure access to these administration entries is achieved via an LDAP-based authentication, configured within Apache Web server using the entry given in Chart 1 to force authentication and comparison of the PHP variable `PHP_AUTH_USER` with a fixed list defined in a PHP configuration file.

The list of administrative users is derived from the institutional list of controlled users and, again, authenticated using LDAP. The SuperUser has the permissions to control the activities of the various administrators, including group and global permissions, and general permissions such as the ability to add and edit (but not delete!) venues and departments. The latter can be defined via a simple URL pointing to, for example, a map. More innovatively, it could be a pointer to an XML-based description of a location, using, for example, Keyhole Markup Language⁹ (for use with a photographic map-based system such as Google Earth). In general, the PHP/MySQL query only displays on the screen those fields any given user has permission to view and edit.

3. CREATING SEMINAR ENTRIES IN THE DATABASE.

Once the system is configured as above, the more routine operation of adding periodic seminar entries can proceed (Figure 1); access to these pages is always via a secure login and LDAP-based authentication. This authentication establishes the permissions any given user has for access to the system, and the displayed page reflects this. In the example shown, the logged-in user only has permissions to add seminars to six groups.

The "add a seminar" page can be directly bookmarked or invoked via a URL of the type <https://site/chemsem/index.php?page=add>. Fields such as the venue are limited to the enumerated list created earlier. Other controlled values (data types) include the date and time. Optional fields include such items such as the speaker's e-mail address, their institutional address, and their "home page" URL. This latter could either be a pointer to a conventional unstructured Web document, written in HTML, or it could be a pointer to a more specific personal description using the so-called Friend-of-a-friend (FOAF) XML descriptor.¹⁰ This latter allows individuals to define the context of their (scientific) interests more precisely and to establish formally expressed links both to colla-

you are logged in as:
rzepa

logout

view a seminar
search the seminars

RSS Feeds
selective RSS feed generator
seminars (RSS)
recent additions (RSS)
recent changes (RSS)

Seminar Administration
add a seminar
edit seminars
cancel seminars
delete seminars

Site Administration
add/edit user permissions
add/edit venues
add/edit groups

Other
submit a bug/suggestion

Edit the fields below, a red star indicates a mandatory field.

* Seminar title

* Seminar Date 08 January 2006

* Seminar time 17 00

* Seminar Venue Central building: Pippard

* Speaker's name

* Group
✓ Computational, Theoretical and Structural Chemistry
Electronic Materials
Catalysis and Advanced Materials
Chemical Synthesis
Chemistry
Statistics

Speaker's email

Speaker's URL/FOAF

Speaker's address

Seminar keywords

Seminar description/abstract

Key PRISM/Digital Object Identifier(s) associated with topic

XML (CML, MathML, SVG ...)

submit seminar reset

Figure 1. Page allowing creation of seminar entries to one of a list of enumerated groups, showing the additional structured information that can be associated with the seminar.

borators, students, and other colleagues and to other research topics they may wish the world in general to know about.

The last four fields shown are specifically for the purpose of inclusion in the RSS feed; two of these are specific XML-based fields which, if completed, will result in additional domain-specific entries. Thus, PRISM¹¹ or a digital object identifier would allow the citation of any key publications the speaker wishes to expose the audience to prior to giving the talk. The XML field allows formally structured subject-specific content to be added. Thus, if the talk has an explicit molecular topic, a CML² description of a molecule (or group of molecules) can be pasted in, which in turn could be viewed using RSS clients supporting these components. Currently, we envision CML, MathML,¹² and SVG (scalable vector graphics) for inclusion, but in principle, any XML language could be inserted, provided it is appropriately namespaced. If the reader's RSS client does not support these components, they will simply be ignored. Their purpose is to allow aggregation and RDF-aware software agents to produce, for example, concept graphs¹³ illustrating the semantic links between diverse seminars and, hence, in principle, allow the construction of an overview of seminar-based science at an institute (or higher aggregate) to be produced.

Once created, existing seminars can be edited; again, only fields the user has permission to alter are available, and the appropriate values are retrieved from the database for modification. Similar pages are available for the cancellation of a given seminar (in which case the seminar details are retained, but it becomes marked as canceled). Certain users only have the privileges to delete a seminar if it has been added in error.

The advantages of such a database-driven system is that all the editors and collators of institutionally organized sem-

inar lists have access to a common database-driven system with a common data schema, with controlled values for many of the fields (date, venue, etc.), which reduces the possibility of error. The system also encourages and raises awareness of the usefulness of information not normally associated with seminar announcements, including contact details for speakers, maps for precise venue locations, identifiers for key bibliographic information, and the possibility of including highly structured metadata for the talk such as machine processable molecular information.

4. VIEWING INFORMATION IN THE DATABASE

4.1. Generating the Feed. At this point, the database is ready to be queried. Because an individual user may be interested in any combination of the available group seminar programs, the system will provide a customized feed to accomplish any of the following: (1) only the recent additions or recent changes to seminars or (2) a feed of any combination (one or more) of the seminar programs (including the complete set).

The first of these can be accomplished with a database query of the following types: *http://server_address/index.php?page=RSSFeeds&type=additions* and *http://server_address/index.php?page=RSSFeeds&type=changes*. If a selection of programs is required, a separate page allows the query to be constructed (Figure 2).

4.2. Viewing the Feed in a Web Browser. If the query URL is then presented to the seminar database, an RSS 1.0 document is generated in response. This document includes the declaration shown in Chart 2 at the top. This will allow a sensible presentation style to be added so that if the issuing program is a Web browser, the resulting display is "human readable". Some Web browsers such as Firefox can display

you are logged in as:
rzepa [logout](#)

[view a seminar](#)
[search the seminars](#)

RSS Feeds
[selective RSS feed generator](#)
[seminars \(RSS\)](#)
[recent additions \(RSS\)](#)
[recent changes \(RSS\)](#)

Seminar Administration
[add a seminar](#)
[edit seminars](#)
[cancel seminars](#)
[delete seminars](#)

Site Administration
[add/edit user permissions](#)
[add/edit venues](#)
[add/edit groups](#)

Your customized feed is shown below. Either Copy/paste it into your RSS browser,
<http://www.ch.ic.ac.uk/pss/index.php?page=RSSFeeds&type=next&group=21>
 Or if your Web browser supports the feed: protocol, [click here](#):

Select the type of feed
☐ recent additions
☐ recent changes
☒ next 15 to happen

Select the group programmes.
 For multiple selections, press the CTRL key whilst selecting

Computational, Theoretical and Structural Chemistry
 Catalysis and Advanced Materials
 Electronic Materials
 Interfacial and Analytical Science
 Chemical Synthesis
 Applied Mathematics: Seminars
 Applied Mathematics: Dynamical Systems
 Applied Mathematics: Mathematical Biology
 Pure Mathematics: London Geometry and Topology
 Pure Mathematics: London Number Theory

Figure 2. User page for creating a customized feed from the database. Readers are encouraged to try entering, e.g., the URL <http://www.ch.ic.ac.uk/pss/index.php?page=RSSFeeds&type=next&group=21> into an RSS or Web browser.

Physical Sciences Seminar RSS feed

Physical Sciences Seminar RSS feed
<http://www.ch.ic.ac.uk/pss/>

MathML

<http://www.ch.ic.ac.uk/pss/index.php?page=view&action=view&ID=63>
 MathML 2004-11-15T00:13:50-00:00 mjlw99

$$\hat{\text{E}}\text{il}_Y(Z, z, \tau) := \int_Y \left(\prod_i \frac{\theta\left(\frac{y_i}{2\pi i}\right) \theta\left(\frac{y_i}{2\pi i} - z\right) \theta'(0)}{\theta(-z) \theta\left(\frac{y_i}{2\pi i}\right)} \right) \times \left(\prod_k \frac{\theta\left(\frac{e_k}{2\pi i} - (\alpha_k + 1)z\right) \theta(-z)}{\theta\left(\frac{e_k}{2\pi i} - z\right) \theta(-(\alpha_k + 1)z)} \right)$$

Foliacenes

<http://www.ch.ic.ac.uk/pss/index.php?page=view&action=view&ID=56>
 Ab initio calculations predict that the cyclic trefoilenes 2 can be stabilized by formation of a complex 4 with early transition metals. The metal atom within the complex is nested within the carbon ring and is considerably closer to the ring centroid than in traditional metallocene complexes. Stabilization is explained by a unique form of 16-electron delocalization involving the metal atom for which we suggest the name foliate-aromaticity. The aromaticity of various polyfoliate systems such as 9 suggests this 16-electron motif is more robust than Clar-like aromatic 6 pi-sextets. The open hemisphere of the metal in such foliacene complexes is predicted to coordinate a variety of ligands.
 Foliacenes 2004-11-09T00:16:00-00:00 rzepa

SVG in Chemistry

<http://www.ch.ic.ac.uk/pss/index.php?page=view&action=view&ID=71>
 SVG in Chemistry
 SVG in Chemistry 2004-12-03T00:09:30-00:00 rzepa

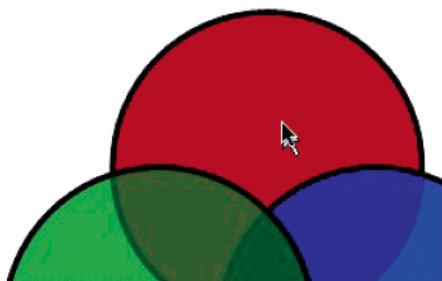


Figure 3. Seminar entry viewed in a SVG- and MathML-aware RSS client (SVG- and MathML-enabled FireFox).

Chart 2

```
<?xml version="1.0" encoding="iso-8859-1"?>
<?xml-stylesheet href="http://www.w3.org/2000/08/w3c-synd/style.css" type="text/css"?>
```

not only the styled feed using the above stylesheet but also specific XML components such as, for example, any ad-

ditional SVG or MathML objects. Figure 3 shows how such objects can be displayed using such a browser.

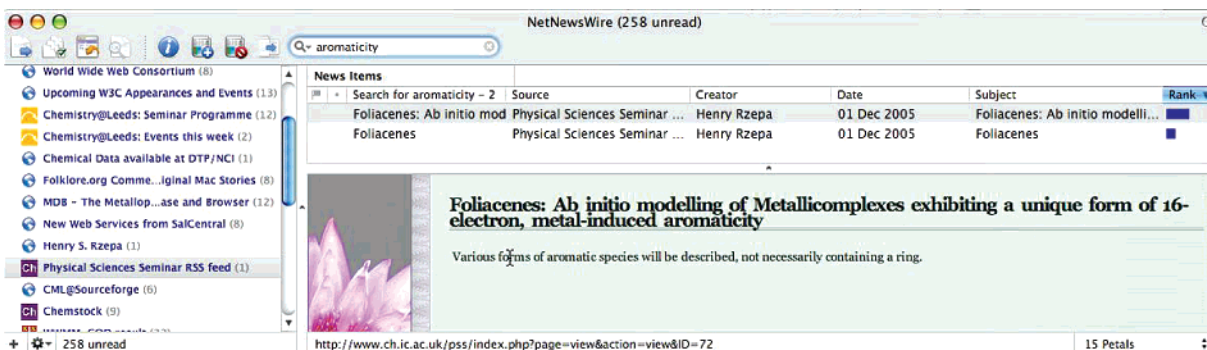


Figure 4. Seminar entry viewed in an RSS client. Note how the search term “aromaticity” is being used as a content filter to show only items relating to this subject.

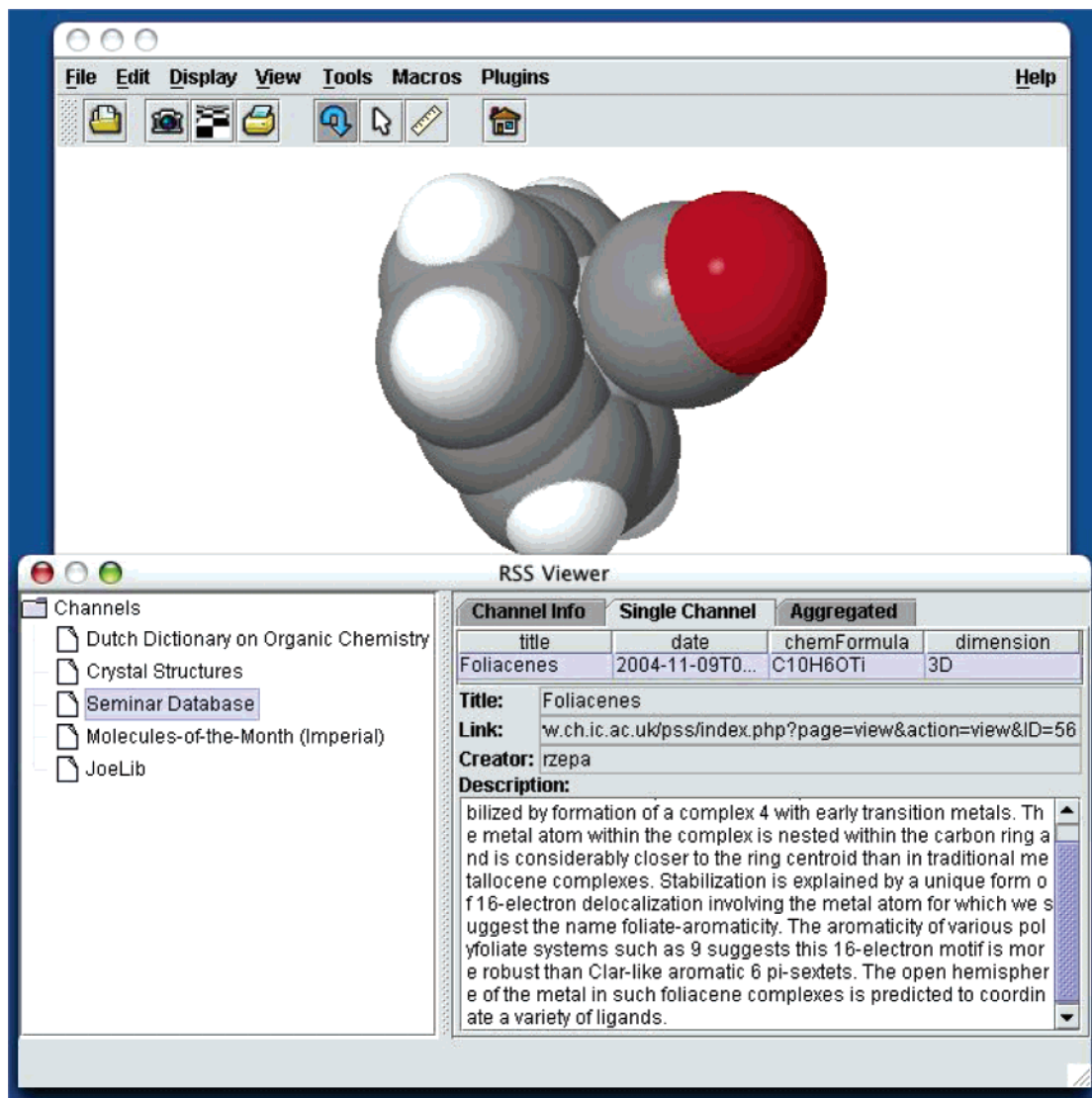


Figure 5. Seminar entry viewed in a molecule (CML-aware) RSS client. The molecular content of the entry is displayed in a 3D molecular viewer (Jmol), which can be used for further chemical manipulation. A 2D editor (JChemPaint) is also available for this purpose. For further information about Jmol and JChemPaint, see ref 2.

4.3. Viewing the Feed in a Generic RSS Client. These queries, however, are really designed to be sent by an RSS client. The advantages of doing so are that the RSS client can repeat the query at regular intervals (typically from every 10 min to 2–3 h). The client can also further process the result once it is returned, such as, for example, sorting it according to the entry metadata such as title, data, creator, subject, or in some RSS clients, even by a prespecified search term (Figure 4).

4.4. Viewing the Feed in a Chemical RSS Client. In principle, any program can be made “RSS-aware”, and the example (Figure 5) shows how a molecular display program can be enhanced with this feature. Note in particular how the molecular component can be viewed (as a rotatable object in this case) or indeed reused (as an entry to, e.g., a 3D modeling program). The entries can even be sorted using metadata derived from the CML component, such as the molecular formula, and further, a filter can be preapplied to

such data to display only a subset of the items of particular interest to the reader.

4.5. Printed Notices. The principle function of the system described here is to produce a machine-processable information system capable of aggregation, sorting, and repurposing. One such purpose might be the production of a printed notice to be placed on notice boards and the like. Because the output of the system is formal XML, the application of a suitable style sheet to reformat the information (e.g., the XSL-FOP or formatting objects processor¹⁴) is straightforward (although not currently implemented in the system).

4.5. Software-Based RSS Uses. All of the above applications illustrate how a human would use the system; the prime purpose after all is to alert them to attend a presentation given by another human! However, the design of the system is such that it need not be a human who would be invoking such actions; any RSS-aware program could equally well (and indeed tirelessly) monitor potentially a much larger selection of subscribed feeds. The software could then add some form of “added value” to the collection (for example, if many of the feeds contain molecular structures, it could identify common substructures) and then even re-export the result again as an RSS feed of its own. It seems perfectly possible to imagine a scenario where essentially all the scientific seminars and talks given on a global basis could be monitored in this sense, in a semantically aware and critical manner, and where the human can, if they wish, only respond to a small selection of the items. They could indeed derive some (albeit different) benefit from not actually being able to attend the presentation because of physical distance or other reason.

Although we defer a major discussion of RDF to a separate article,¹³ the RSS feed generated above does contain RDF-based metadata declarations relating to the seminars. RDF is essentially a mechanism for defining a relationship between an information subject and a second entity referred to as the object, on the basis of a well-defined assertion or predicate. The predicates themselves can be regarded as representing a commonly shared meaning or understanding of the relationship between the subject and the object.^{15,16} A predefined set of definitions for predicates (an ontology) can be created using an RDF schema, each predicate being globally and uniquely identified using a uniform resource identifier. Collections of these in turn would allow formal inferences, or comparisons, to be made about objects from disparate sources. When RDF is aggregated into a large database (more formally called a triplet store), these links could be used to automatically infer connections between diverse concepts and subjects. A human, who tends to operate on a much smaller and parochial scale, might not have the patience to establish such connections. By focusing on seminars, one can take advantage of the (often unique) expertise of an invited speaker to establish a high-quality set of connections within a particular subject. The longer-term objective is to construct a formal database of connections in the context of formal ontologies linking related subjects. Ultimately, if the mechanism succeeds, then links will be created between information in a more systematic manner than that dependent on rather more unpredictable and evanescent human activity alone.

5. CONCLUSIONS

The announcement of a scientific seminar or presentation represents an almost perfect example of metadata, that being information about a data content object (the seminar itself). This metadata has both generic components (title, speaker, date, and venue) and the potential for a much richer description (relationships to bibliographic information such as associated published article) along with subject-specific items (for a chemistry talk, relationships to unique molecules and their properties). Exposing such metadata using RDF, we suggest, is an integral part of the currently ambitious attempt to transform the Web into a semantically structured fabric of information and knowledge. The RSS/RDF system we describe here provides a controlled interface for organizations to create useful nodes in this fabric and to simultaneously give an interested human access to an alerting mechanism which they can control to deliver the information they wish to receive. The RSS mechanism provides for automatable mechanisms with which to aggregate, sort, and add value to seminar information, which could be integrated with other envisaged components of the semantic Web (journals, dictionaries and ontologies,¹⁵ and e-science resources) to create a far richer information environment for the scientist.

Supporting Information Available: Illustrative screen shots of various administrative pages associated with the operation of the system. This material is available free of charge via the Internet at <http://pubs.acs.org>.

REFERENCES AND NOTES

- (1) Goodman, J. M. Chemistry on the World-Wide-Web: A Ten Year Experiment. *Org. Biomol. Chem.* **2004**, *2*, 3222–3225. DOI: 10.1039/b409956g.
- (2) Murray-Rust P.; Rzepa, H. S. Chemical Markup, XML, and the World Wide Web. 4. CML Schema. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 757–772. DOI: 10.1021/ci0256541. Murray-Rust, P.; Rzepa, H. S. Towards the Chemical Semantic Web. An Introduction to RSS. *Internet J. Chem.* **2003**, *6*, article 4. Murray-Rust, P.; Rzepa, H. S.; Williamson, M. J.; Willighagen, E. L. Chemical Markup, XML and the Worldwide Web. Part 5. Applications of Chemical Metadata in RSS Aggregators. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 462–469. DOI: 10.1021/ci034244p.
- (3) Rzepa, H. S.; Williamson, M. J. Chemstock: A Web-Based Chemical Inventory System Built from OpenSource Software Components. *Internet J. Chem.* **2002**, *5*, article 6.
- (4) Apache, see <http://httpd.apache.org/> (accessed Mar 2006).
- (5) PHP, see <http://www.php.net/> (accessed Mar 2006). PEAR, see <http://pear.php.net/> (accessed Mar 2006).
- (6) SMARTY, see <http://smarty.php.net/> (accessed Mar 2006).
- (7) MySQL, see <http://www.mysql.com/> (accessed Mar 2006).
- (8) LDAP, see <http://www.openldap.org/> (accessed Mar 2006).
- (9) Keyhole Markup Language (KML), see <http://www.keyhole.com/kml/docs/webhelp/> (accessed Mar 2006).
- (10) FOAF, see <http://www.foaf-project.org/> (accessed Mar 2006).
- (11) PRISM, see <http://www.prismstandard.org/> (accessed Mar 2006).
- (12) MathML, see <http://www.w3.org/Math/> (accessed Mar 2006).
- (13) RDF, see <http://www.w3.org/RDF/> (accessed Mar 2006). See also Casher, C.; Rzepa, H. S. *J. Chem. Inf. Model.* To be submitted.
- (14) XSL-FOP, see <http://xmlgraphics.apache.org/fop/> (accessed Mar 2006).
- (15) Wang, X.; Gorlitsky, R.; Almeida, J. S. From XML to RDF: How Semantic Web Technologies Will Change the Design of ‘Omic’ Standards. *Nat. Biotechnol.* **2005**, *23*, 1099–1103. DOI: 10.1038/nbt1139.
- (16) Hughes, G.; Mills, H.; De Roure, D.; Frey, J. G.; Moreau, L.; Schraefel, M. C.; Smith G.; Zaluska, E. The Semantic Smart Laboratory: A System for Supporting the Chemical eScientist. *Org. Biomol. Chem.* **2004**, *2*, 3284. DOI: 10.1039/B410075A.
- (17) Murray-Rust, P.; Rzepa, H. S.; Tyrrell, S. M.; Zhang, Y. Representation and Use of Chemistry in the Global Electronic Age. *Org. Biomol. Chem.* **2004**, *2*, 3192–3203. DOI: 10.1039/B410732B.