INTERNET BASED MOLECULAR COLLABORATIVE AND PUBLISHING TOOLS

OMER CASHER

A Thesis submitted in fulfilment of requirements for the degree

of Doctor of Philosophy of Imperial College London and

for the Diploma of Membership of Imperial College London

Department of Chemistry

Imperial College London

July 2010

Abstract

The scientific electronic publishing model has hitherto been an Internet based delivery of electronic articles that are essentially replicas of their paper counterparts. They contain little in the way of added semantics that may better expose the science, assist the peer review process and facilitate follow on collaborations, even though the enabling technologies have been around for some time and are mature. This thesis will examine the evolution of chemical electronic publishing over the past 15 years. It will illustrate, which the help of two frameworks, how publishers should be exploiting technologies to improve the semantics of chemical journal articles, namely their value added features and relationships with other chemical resources on the Web.

The first framework is an early exemplar of structured and scalable electronic publishing where a Web content management system and a molecular database are integrated. It employs a test bed of articles from several RSC journals and supporting molecular coordinate and connectivity information. The value of converting 3D molecular expressions in chemical file formats, such as the MOL file, into more generic 3D graphics formats, such as Web3D, is assessed. This exemplar highlights the use of metadata management for bidirectional hyperlink maintenance in electronic publishing.

The second framework repurposes this metadata management concept into a Semantic Web application called SemanticEye. SemanticEye demonstrates how relationships between chemical electronic articles and other chemical resources are established. It adapts the successful semantic model used for digital music metadata management by popular applications such as iTunes. Globally unique identifiers enable relationships to be established between articles and other resources on the Web and SemanticEye implements two: the Document Object Identifier (DOI) for articles and the IUPAC International Chemical Identifier (InChI) for molecules. SemanticEye's potential as a framework for seeding collaborations between researchers, who have hitherto never met, is explored using FOAF, the friend-of-a-friend Semantic Web standard for social networks.

Declaration of Originality

I declare that the thesis entitled “Internet based Molecular Collaborative and Publishing Tools” and the work presented in the same reflect my own original research. Furthermore, I confirm that:

- This work was undertaken mainly while in candidature for a research degree at Imperial College London.
- All of this work was undertaken completely external to my employer, GlaxoSmithKline.
- I have tried to make it clear from the context whether a result is claimed as original.
- When presenting a result which is not original, the name of the source and a reference to the earliest known proof is provided.
- Where I have consulted others for help, this is always clearly attributed.
- If work is presented that was done by myself jointly with others, I have made explicitly clear exactly what my personal contribution was.

Signed:

Date:

To Gapsun and Rezin

Acknowledgements

The research and this thesis have been carried out whilst I have been in full time employment over a number of years. It would certainly have been impossible without the support, motivation, inspiration, and vision of my supervisor Professor Henry Rzepa. I especially value our brainstorming sessions where new and sometimes completely unexpected ideas emerged.

I am thankful for the various forms of feedback and advice that I received from my colleagues at Imperial College, namely Chris Leach, Stuart Green, Sonsoles Martên-Santamarêa, George Gkoutos, Carlota Conesa, and Humberto Chaves. I am especially thankful to Guillem Suñer and Chris Page for all their input and support, and for the many coffee, food, and pub sessions that we are still sharing to this day. Last but not least, Saeed Sannuga, who shares my passion for 3D molecular animation and home cooking, always provided me with a place to stay whenever needed.

My extensive interactions over the years with my collaborators at the RSC have been invaluable in my research, particularly in gaining unique insight into the scientific publishing world. In particular, I would like to acknowledge David James, Jonathan Griffiths, Becky Allen, Dave Riddick, Richard Kidd and Colin Bachelor. I would also like to acknowledge the excellent input that I recently received from Tony Hammond, Alf Eaton and Laura Croft at Nature Publishing and thank them for our recent meeting.

I would also like to thank my work colleagues both at GSK and MSI for their support and encouragement, in particular Steve Herbert, Anthony Richards, Osman Güner, Vin Cunningham, Roger Gunn, Jonathan Rosalki, Paul Matthews, Tom Flores and Colin Wood.

The support of my loved ones cannot be forgotten. My dear wife Gapsun has always been there with me providing support and “encouraging” me to complete this thesis. My mother’s belief in me, even during the more difficult times is greatly appreciated. She has provided me with support, both moral and financial. I surely would not have completed this thesis without my wife and mother and I dedicate it to the two of them.

Contents

ABSTRACT **2**

DECLARATION OF ORIGINALITY **3**

ACKNOWLEDGEMENTS **5**

CONTENTS **6**

LIST OF FIGURES **11**

PREAMBLE **13**

1. INTRODUCTION **17**

1.1 THE NINE ARTICLES OF ROBERT BOYLE 17

1.2 CHEMICAL ELECTRONIC PUBLISHING (1976-1994) 18

1.3 A UNIQUE PERSPECTIVE (1995-2010) 19

2. CHEMICAL JOURNAL WEB DOCUMENT MANAGEMENT **24**

2.1 THE CLIC PROJECT 24

2.1.1 Journal Presentation 25

2.1.1.1 SGML 25

2.1.2 HTML Creation 26

2.2 WEB DOCUMENT MANAGEMENT APPROACH 27

2.2.1 Metadata Management 27

2.2.1.1 Dublin Core 27

2.2.2 Hyperwave 28

2.2.2.1 URL Integrity 30

2.2.2.2 Hyperwave URL Maintenance 30

2.2.2.3 Offline Journal Production 31

2.3 CLIC ELECTRONIC JOURNAL PRODUCTION AT THE RSC 31

2.3.1 Manuscript Preparation 31

2.3.2 CLIC article Navigation: Standard Web Server 32

2.4 CLIC ELECTRONIC JOURNAL PRODUCTION WITH HYPERWAVE 34

2.4.1 Manuscript Preparation 34

2.4.2 Link Maintenance for CLIC articles 35

2.4.3 CLIC article Navigation: Hyperwave 36

2.5 OPEN HYPERMEDIA AND COLLABORATION 38

2.6 CURRENT SCIENTIFIC PUBLISHING PROCESS 39

2.7 XML STANDARDS 39

2.7.1 XML Syntax 40

2.7.2 XML Validation 41

2.7.2.1 Document Type Definition (DTD) 41
2.7.2.2 XML Schema 42
2.7.3 XML Namespace 42
2.7.4 Manipulating XML with DOM and SAX 43
2.7.5 XPATH 43
2.7.6 Extensible Stylesheet Language (XSL) 43
2.7.7 XML Linking Language (XLink) 44
2.7.8 XML query language (XQuery) 44
2.8 XML DATABASES 45
2.8.1 MarkLogic 45
2.9 SUMMARY 46

3. MOLECULAR SUPPORTING DATA MANAGEMENT 48

3.1 THE ELUSIVE ENHANCED CHEMICAL ARTICLE 48
3.1.1 Publisher 48
3.1.2 Reviewer 49
3.1.3 Archivist 49
3.1.4 Reader 49
3.1.5 Author 50
3.2 AN OPTIMAL APPROACH 50
3.2.1 XML for Molecular Supporting Data 51
3.2.2 3D Graphics Formats for Presentation 51
3.2.3 Molecular Repository Approach 53
3.2.3.1 Archive Molecular Data with the Article 53
3.2.3.2 Archive Molecular Data in an External Archive 54
3.2.3.3 Archive Molecular Data in a Molecule Database 55
3.3 MANAGEMENT OF MOLECULAR DATA IN THE CLIC PROJECT 56
3.3.1 Management of URLs from Article to Structure 58
3.3.2 Structure Retrieval 58
3.4 CHEMSPIDER 58
3.5 SUMMARY 59

4. CHEMICAL JOURNALS AND THE SEMANTIC WEB 60

4.1 DRIVERS FOR IMPROVED CHEMICAL ELECTRONIC PUBLISHING 60
4.1.1 The Hype Cycle 60
4.1.2 The Need for Context and Community 61
4.1.3 Defrosting the Chemical Journal 62
4.2 IMPROVING THE METADATA 63
4.2.1 Fundamental Guiding Principles 63

4.2.2 Adapting a Successful Semantic Model 64
4.3 THE SEMANTIC WEB 65
4.3.1 The Semantic Web Architecture 66
4.3.2 The Resource Description Framework 67
4.3.2.1 Web Ontology Language (OWL) 67
4.3.2.2 Semantic Reasoning 68
4.3.3 Semantic Searching 68
4.3.3.1 The SPARQL Query Standard 69
4.3.4 RDF Triple Stores 69
4.4 SEMANTICEYE TARGET ARCHITECTURE 70
4.4.1 Dataflow 70
4.4.2 The Document Object Identifier (DOI) 71
4.4.3 The IUPAC International Chemical Identifier (InChI) 73
4.4.4 Extensible Metadata Platform (XMP) 74
4.4.5 WebDAV for file handling 76
4.4.6 RDF Repository - Sesame 77
4.5 IMPLEMENTATION OF THE ARCHITECTURE 77
4.5.1 Sample Articles 77
4.5.2 Creation and Capture of XMP Metadata 78
4.6 PROCESS FLOW IN SEMANTICEYE 82
4.7 SYSTEM ARCHITECTURE 83
4.7.1 RDF Management (Figure 24, Orange) 84
4.7.2 XMP Service Suite (Figure 24, Blue) 84
4.7.3 XMP Uploader (Figure 24, Yellow) 85
4.7.4 Sesame Security 85
4.9 NAVIGATION OF JOURNALS 86
4.10 SEMANTICEYE RESULTS 89
4.10.1 Metadata Management 90
4.10.2 The SemanticEye Ontology 90
4.10.3 Validation and Trust in SemanticEye 91
4.10.3.1 Metadata Capture 92
4.10.3.2 Project Prospect 92
4.10.3.3 Ontology Add-in for Word 2007 93
4.10.3.4 Chemistry Add-in for Word 2007 93
4.10.4 Authorship and Digital Signatures 94
4.10.4.1 The validation of DOIs 95
4.10.4.2 An InChI agent 95
4.11 INTEGRATION OPPORTUNITIES OF SEMANTICEYE 96
4.11.1 The Open Archives Initiative 96

4.11.2 The Semantic Grid 96
4.11.3 The Semantic Wiki 97
4.12 ENTERING THE PLATEAU OF PRODUCTIVITY 98
4.13 INITIATIVES THAT WILL INFLUENCE SEMANTICEYE'S FUTURE 99
4.13.1 Open Access Publishing 99
4.13.2 Linked Data 100
4.13.3 oreChem 102
4.14 SUMMARY 103

5. CHEMICAL COLLABORATION NETWORKS 105

5.1 AN EARLY CHEMICAL COLLABORATORY ON THE WEB 105
5.2 INITIATION OF NEW COLLABORATIONS 106
5.3 EXPERT FINDER SYSTEMS 106
5.4 SEMANTIC ENABLEMENT OF SOCIAL NETWORKING 107
5.4.1 Inadequacies of Web 2.0 for collaborations 107
5.5 WEB 3.0 AND THE SEMANTIC WEB 108
5.6 FOAF (FRIENDS OF A FRIEND) 108
5.6.1 FOAF Aggregation 110
5.7 SPARQL SEARCH FOR SCIENTISTS WITH A COMMON INCHI 110
5.7.1 Identifying a scientist's collaborators in FOAF 110
5.7.2 Adding Scientist's Molecules of interest to the FOAF 111
5.7.3 Grouping authors by a molecule of interest 112
5.7.4 Aggregating FOAFs from Multiple Sources 113
5.8 THE CURRENT STATUS OF ONLINE RESOURCES 115
5.8.1 The Relevance of Existing Web 2.0 Technologies. 115
5.8.1.1 Blogs 115
5.8.1.2 Wikis 117
5.8.1.3 Podcasts 117
5.8.1.4 Social Networks 118
5.8.1.5 Mash-ups 119
5.8.1.6 Reference Management Software 119
5.8.1.7 Google Social Graph API 120
5.8.1.8 Facebook Open Graph API 121
5.9 POTENTIAL TECHNICAL SOLUTION 121
5.10 SUMMARY 122

6. CONCLUSION 124

6.1 THE 2022 PREDICTION 124
6.2 FUTURE SEMANTICEYE WORK 125

7. BIBLIOGRAPHY.....126

List of Figures

Figure 1: The CLIC homepage at the RSC website 33
Figure 2: A 3D molecule rendered in a separate browser window 33
Figure 3: Hyperwave HTML text insertion and metadata editing using a Web form 34
Figure 4: A CLIC article containing a PDB file 35
Figure 5: The bidirectional links of an HTML file with an embedded PDB file 36
Figure 6: A document template inserts navigation “buttons” as a page header 37
Figure 7: An XML example of a simple organisational chart 40
Figure 8: An example of an XML Schema that maps to a relational database schema 42
Figure 9: A 3D molecular Hyperglossary 52
Figure 10: Molecular data are archived with the articles 54
Figure 11: Molecular data are stored in digital archive external to the publisher 55
Figure 12: Molecular data are archived in a molecule database publisher-side 56
Figure 13: High level publishing architecture with a molecular database 57
Figure 14: The Gartner Hype Cycle 61
Figure 15: Semantic Web Stack 66
Figure 16: Overall dataflow of the RDF metadata lifecycle 71
Figure 17: InChI for Sucrose 73
Figure 19: The mouse-over feature for Acrobat. 76
Figure 18: Acrobat Professional XMP Tool 76
Figure 20: Example of XMP metadata created for a journal article 80
Figure 21: RDF Graph of an XMP example 81
Figure 22: Process flow to transfer metadata from the articles to the RDF repository 82
Figure 23: RDF Graph of the Chemical Journal Ontology 83
Figure 24: System Architecture of SemanticEye 83
Figure 25: WebDAV folder as it appears on the Windows XP client 85
Figure 26: Entry Point of the SemanticEye Ontology 87
Figure 27: The Title predicate is selected to reveal the 8 titles in the Ontology 87
Figure 28: All of the metadata are shown for a Document selected by title 88
Figure 29: The InChI is selected and articles which contain it are listed 88
Figure 30: The DOI is resolved by selecting "Get Article" 89
Figure 31: Data flow to create the Chemical Journal Ontology 91
Figure 32: Example of a Personal FOAF 108
Figure 33: Example of a group FOAF 109

Figure 34: SPARQL Query on SemanticEye Triples to identify collaborators 111
Figure 35: SPARQL Query to identify molecules of interest 111
Figure 36: Personal FOAF of SemanticEye output 112
Figure 37: SPARQL Query to group authors according to a molecule of interest 113
Figure 38: Group FOAF of SemanticEye output 113
Figure 39: Example FOAF Aggregation: SemanticEye ChemSem IC PWP 114

Preamble

From 1992 through 1997 I worked as a research assistant for Professor Henry Rzepa at Imperial College Department of Chemistry and in 1997 I registered for my PhD. I completed my research and this thesis subsequently whilst in full time employment. The research discussed draws from a wide and diverse body of 21 scholarly publications which I co-authored throughout this period and which are summarised here.

My research started with the development of EyeChem, a suite of software modules for molecular visualisation that will be discussed in Chapter 3 in this thesis. EyeChem was initially used to create visualisations of molecular orbital and electrostatic potential outputs from two computational chemistry programs, MOPAC and Gaussian. The very first scholarly publication that I co-authored in 1993 explored stereoselective epoxidation with the help of EyeChem visualisations of AM1, PM3 and *ab initio* 6-31G* calculations[1].

In the same period, the rapid advancement of Internet technologies and in particular the birth of the World Wide Web created significant research opportunities. My move into Internet based scientific research was marked by my development of another EyeChem application called eye2eye, an early exemplar of a chemical collaboratory. This work was discussed in my second publication[2] and will be summarised in Chapter 5 of this thesis. EyeChem's support for 3D animations of molecular vibrations and reaction pathways was explored within the "hyperactive molecules" exemplar whereby selecting a URL in an HTML page would launch EyeChem with an interactive 3D molecular animation downloaded. I co-authored 2 publications[3,4] on the subject.

A significant breakthrough came with the inception of the Virtual Reality Modelling Language, or VRML, which at the time was the Web's 3D hyperlinked graphics analogy to 2D HTML text. VRML 1.0 was based on SGI's Open Inventor format which fortuitously was also EyeChem's 3D format. I used this advantage to publish literally hundreds of the earliest VRML 1.0 files (relatively easy with EyeChem) for the benefit of a rapidly growing community that was determined to make the much hyped VRML a success. As with the eye2eye collaboratory, one of the goals of VRML was the sharing and editing of 3D "worlds" with remote collaborators. This was discussed in 2 subsequent publications[5,6] and an ACS talk[7]. I also prototyped a standalone molecular VRML authoring tool called Molecular Objects[8] which was based on the SGI Molecular Inventor toolkit.

My VRML work paved the way into chemical electronic publishing which, in 1995, was still nascent. As part of the CLIC project discussed in Chapter 2, I explored how VRML could add value to chemical journal articles and published two papers on the subject. One included several 3D molecular structures in VRML format[9] and the other included the 3D molecular "hyperglossary"[10], described in Chapter 3. The hyperglossary here is a VRML model that includes a scatterplot with hyperlinks to hundreds of VRML molecules with each one hyperlinked to an associated HTML file. This novel work was featured in a C&E News article[11] and incorporated into the magazine's front cover graphic.

The hyperglossary example underscored the enormous challenges in managing large volumes of Web files containing hyperlinks. To address this challenge within the CLIC project I investigated a sophisticated hypermedia system called Hyperwave and its capabilities in chemical publishing, VRML document management, HTML/VRML hyperlink maintenance and electronic conferencing[12,13,14,15]. The results are discussed in Chapter 2 of this thesis.

Towards the end of my Imperial College residency, I developed a script that auto-generated a table of contents (TOC) by the interrogation of Hyperwave's metadata database. All "chemical objects" within Hyperwave including HTML and supporting data such as molecular structures, images *and* URLs are assigned metadata and can be navigated consistently. This TOC script exemplified the importance of metadata management for the effective navigation of scientific publications and for the intermingling of journal articles with supporting data. It also set the stage for the later phase of my research.

At the start of 1998 I joined Molecular Simulations Inc. (MSI) where I worked as the Product Manager of WebLab. WebLab was the first Web based package supporting cheminformatics and bioinformatics and had much of the functionality found in modern electronic lab notebooks. During my three years at MSI I co-authored my final paper on VRML (which, incidentally, WebLab supported) for Chemistry in Britain[16], another widely read magazine where another one of our VRML scatterplots featured on the front cover. I also co-authored a talk on WebLab that was presented at an ACS conference[17]. With my MSI manager, I co-authored a paper on Internet related trends in cheminformatics[18].

In 2000 I left MSI to join GlaxoWellcome working briefly as a cheminformatician. Following the merger with SmithKline Beecham at the beginning of 2001 I moved into an Information Architect role where I developed my expertise in XML (I architected the first XML based production application in GSK) and scientific image management. In 2002 I had my first exposure to the Semantic Web when I was investigating the potential of eScience deliverables

for scientific imaging. In 2005 I joined the GlaxoSmithKline Clinical Imaging Centre as the Manager for Medical Imaging IT.

In order to concentrate on my career development, I took a six year hiatus from academic based research. My return to research in 2005, albeit part time, stemmed from my desire to implement the Semantic Web and in particular the Adobe Extensible Metadata Platform (XMP) metadata for scientific imaging. At the time, the Semantic Web was still very much within the realm of academia and I felt that in order to use the technology successfully within industry I would first need to subject myself to the research rigour worthy of a PhD.

Fortunately I was able to continue using chemical publishing as the basis of my research thereby maintaining the continuity with my earlier work. The SemanticEye exemplar I developed, repurposed the metadata management and TOC navigation discussed above within the Semantic Web. The work was described in an extensive publication[19] that was incorporated into Chapter 4 of this thesis. This publication was one of five articles selected from some thirty five journals to be featured in the November 22, 2006 edition of ACS News Service Weekly PressPac[20]. At around the same time it was also featured in a New Scientist Tech article[21]. I also gave a well received talk on it at an ACS conference[22]. The SemanticEye project established the importance of XMP in chemical publishing and the subsequent uptake of XMP by Elsevier, Nature and the Royal Society of Chemistry provides a clear testimonial.

The final stage of my research for this thesis returns to the chemical collaboratory metaphor by integrating, via SemanticEye, chemical publishing and the Friend of a Friend (FOAF) Semantic Web vocabulary for social networks. Several talks resulted, including one given at the ACS[23]. This work was also the basis of a peer reviewed book chapter[24] that was incorporated into Chapter 5 of this thesis.

“We shall not cease from exploration, and the end of all our exploring will be to arrive where we started and know the place for the first time”.

- T. S. Eliot

1. Introduction

1.1 The Nine Articles of Robert Boyle

This thesis begins with the incipient Philosophical Transactions journal published by the Royal Society in March 1665 and the very first journal devoted exclusively to scientific articles[i]. An inspection of its presentation confirms that not much has changed in 345 years other than the technological sophistication and efficiency of the journal production process. In terms of content, this single publication provides a microcosm of scientific diversity indicative of the landscape covered by the thousands of scientific journals today. Nine of its articles were written by Royal Society founding member Robert Boyle, who is widely regarded as the first modern chemist. These nine articles, the genesis of chemical scholarly publication, frame the problem space that the research in this thesis is attempting to address.

Locating Boyle's nine articles represents the first problem. It was not clear, by examining the table of contents, which articles actually pertain to him. A subsequent metadata search using Royal Society Publishing's search engine with the terms "year:1665" and "author:Boyle,R" generated three hits. Refining this search slightly to "year:1665" and "author:Boyle" generated the correct nine hits. "Boyle,R" was inappropriate to uniquely identify this famous author because three of the articles were attributed to Robert Boyle and other six to Mr. Boyle. Failing to impose a minimum level of structure in the journal resulted in at least one document management deficiency that has never been rectified to this day. **Chapter 2** will discuss research of the document management patterns required to ensure that chemical journal articles will always be easy to retrieve by the reader.

The lack of supporting data for any of Boyle's articles is not surprising. However little has changed in this regard in 345 years despite the fact that the enabling technologies for the preparation and archival of supporting data are now mature. This is the second problem that will be investigated in **Chapter 3** of this thesis. It is a complex problem due to a number of factors most of which are outside of the publisher's circle of

[i] Denis de Sallo's "Le Journal des Savants" came out two months earlier in January 1665 but contained a mix of scientific and non scientific content.

control. One of these factors is that of determining the appropriate file formats suitable for presentation, archival, retrieval and future proofing. Another is the establishment of a location where the supporting data should be kept. Yet another is the reluctance of authors to provide supporting data.

Automatically establishing relationships between articles, especially those articles that initially appear to be unrelated, reflects the third problem. If each of the nine articles of Boyle were to be identified by a single representative metadata term taken from their titles, a resulting set of terms would look like: *earthquake; barometer; country; foetal bird; baroscope; sea; animal blood transfusion; history of cold; monstrous calf*. Of the nine articles, the two with the terms *barometer* and *baroscope* can possibly be automatically inferred to be linked to Boyle's research. However, automatically establishing the link between the remaining seven articles with today's technologies or, to be more precise, the manner in which today's technologies are being implemented, would be impossible. Proper metadata management for journals, predicated by a unique way of identifying authors, is required and will be one of the subject matters on which **Chapter 4** of this thesis is based. The notion of trust is also explored within the context of Tim Berners-Lee's long term Web, or to be more precise, Semantic Web vision.

This leads on to the fourth problem which is that of collaboration. The fact that Boyle's nine articles and all the other articles in the journal are written by no more than one author signify a total lack of collaborative spirit, or trust during that period. As science became increasingly more expensive, an increase in scientific collaboration was inevitable and manuscript co-authorship became the norm. The scientific journal became not only an easy means of identifying existing collaborations but also a vehicle for identifying potential new collaborators. **Chapter 5** will present a forward looking perspective employing scientific publishing metadata and Web-based social networking to assist chemists in identifying scientific expertise, useful for locating potential collaborators.

1.2 Chemical Electronic Publishing (1976-1994)

The first prototype electronic journal was in 1976[25]. This information was initially in proprietary format and accessible using proprietary network protocols. Production costs along with the premium service required to access it made this information expensive. Costs gradually decreased and along with it, expectations rose. However, the uptake of

electronic chemical information only began to accelerate in the early 1990s once the Internet started to achieve a critical mass.

In 1993, adding structure to largely unstructured electronic documents emerged as a reaction to the limitations of e-mail. The approach taken was to separate e-mail messages into multiple parts which could be reconciled by intelligent software. The notion of Internet based document rendering with presentation did not yet exist. A major challenge was relating these objects to software that would render them appropriately.

With the widespread arrival of the World Wide Web in 1994 a medium for Internet based journal dissemination became available. Markup was now handled by Web browsers and chemical data was rendered by helper applications or plug ins. Rzepa *et al*[26] set out the possibilities here. Chemical publishing houses started exploring the Web in the hope of addressing the crisis state of scholarly chemical publishing, namely:

- Authors were under pressure to publish more while increasing the impact factor of their publications.
- The rising cost of journals, due largely to the increasing publishing overheads, was putting pressure on libraries to reduce journal subscriptions.
- Publishers were trying to preserve a customer base that was becoming increasingly accustomed to accessing chemical information on-line.
- Publishers wanted to uphold the journals' look and feel electronically even at the expense of providing an enriched user experience with more interactive documents.

1.3 A Unique Perspective (1995-2010)

This thesis will focus on the most recent fifteen years of chemical publishing between 1995 and 2010. Although this represents a relatively small proportion of the total 345 year period, it represents the period of transition for scientific publishing from print to electronic: in 1995 scientific publishers were just starting to investigate the Web and now (in 2010) virtually all have a strong Web presence. Because the research discussed in this thesis was mostly undertaken in two phases, between 1995 and 1998 and between 2004 and 2010, a unique perspective on chemical electronic publishing has been achieved.

An observation made during this fifteen year period, perhaps an obvious one, is that information technology evolves much more rapidly than the scientific publishing processes. Consequently, the technologies that publishers are currently using to achieve their Web presence were nascent during the first phase of this research and not ready to be investigated in depth. At the time, open hypermedia systems showed great potential and one such system, Hyperwave (formally Hyper-G), was investigated as part of the research reported in this thesis. The use of open hypermedia for electronic publishing was explored comprehensively by Hitchcock[27] and the suitability of Hyperwave in this regard was reported by Schmaranz[28] in 1996. Research using Hyperwave for chemical publishing, as discussed in this thesis, concurs with Schmaranz. However, Hyperwave failed to gain any traction in the publishing market, despite having moved from academic research into a commercial venture in 1997[ii].

An explanation was recently given by Vesse *et al*[29] as to why open hypermedia as a concept never significantly advanced despite its document and link integrity features. They argue that open hypermedia was providing a solution to a problem that became less significant as Web search engines, particularly Google, became more sophisticated. They also pointed out a distinction between the Internet and enterprise Intranets that was first reported by Ashman in 2000[30]. Unlike Internet Web documents, document and link integrity on corporate Intranets are important for good business practices. Thus, open hypermedia systems were better suited for Intranets, which justifies Hyperwave's current positioning towards the Intranet market. Microcosm[31], an alternative open hypermedia system developed at the University of Southampton, followed a similar route[iii]. Vesse *et al* go on to describe how the link integrity feature of open hypermedia systems may find modern applicability within the Linked Data metaphor on the Internet, thereby signifying that open hypermedia technologies are not necessarily redundant.

The incorporation of 3D molecular structures within the chemical journal was also investigated during the first phase of research and will be discussed in this thesis. The notion of a chemical journal publisher actively promoting the inclusion of supporting data

[ii] Hyperwave provides commercial enterprise content management solutions and is still trading today (www.hyperwave.com).

[iii] Microcosm was marketed commercially by Multicosm Ltd. (www.multicosm.com).

with its articles was first realised by Bachrach's Internet Journal of Chemistry (IJC)[32], now defunct. IJC was published during the 1998-2004 period, which neatly falls between the first and second phases of research for this thesis. IJC as a journal however, could not provide the catalyst needed for the established publishers to actively seek supporting data from authors and in a 2009 article[33] Bachrach states that this is still an unsolved problem.

A recent paper by Vision[34] highlights a number of cultural barriers associated with the provision of supporting scientific data. Taking these cultural barriers into account, he proposes a technical solution based on a dedicated and centrally managed institutional repository called Dryad[35]. The institutional repository approach has merit and staying power. However as is discussed in this thesis, chemical supporting data in general and molecular structures in particular have well established file formats that should be exploited, using specialised database technology possibly managed by publishers.

The second phase of research focuses on SemanticEye, an application of Semantic Web technology for chemical publishing which had hitherto not been investigated. The work draws from the content and metadata management experience gained during the first phase. When this work commenced in 2005 the Semantic Web was only starting to be investigated for chemistry, but as far back as 2001 Gkoutos *et al* explored the Semantic Web notion of trust within the context of the Chemical Markup Language (CML)[36] XML vocabulary. The use of the International Chemical Identifier (InChI) will be discussed in this thesis, prompted by the work of Coles *et al* in 2005[37] in the "InChIfication" of the Semantic Web. Taylor *et al* discuss uses of Resource Description Framework (RDF) vocabularies in Chemistry[38] and present chemical properties within an RDF ontology that incorporate the InChI. Willighagen has been maintaining an online tool to purpose InChIs within RDF[39].

An important enabling technology that was explored and presented in this thesis is the Adobe Extensible Metadata Platform (XMP), an RDF vocabulary for embedding and extracting metadata, which can include the InChI, within journal articles. At the time of this research, XMP was novel and had yet to be adopted for publishing. Fast forward to now. The uptake of XMP has been accurately foreseen: both Elsevier and Nature have

recently implemented XMP. The RSC are actively working with it and have released some example articles containing XMP in the near future[iv].

In 2001, Berners-Lee and Hendler outlined their vision of scientific publishing on the Semantic Web by way of a commentary in Nature[40]. Nine years later, their vision can still be considered futuristic. The inclusion of chemicals and reagents as metadata in published articles has not yet been realised. However, Frey's team have developed the Southampton Semantic ELN[41] for capturing such metadata directly from an electronic lab notebook (ELN) as RDF. Thus, although the ELN is out of scope for this thesis, it is recognised that realising the Berners-Lee and Hendler vision necessitates a mechanism for translating key ELN RDF terms into chemical journal XMP in the future. Although it is relatively early days, the Microsoft funded oreChem project[42], which is developing new models for the research and dissemination of chemical publications, is likely to play a role here.

The collaboration potential of SemanticEye was investigated whereby the creation of personal scientific profiles in the Friend-of-a-Friend (FOAF) RDF vocabulary[43] can be automated. Aleman-Meza *et al* discussed incorporating FOAF and related RDF technologies into a framework used for expert finding[44]. The goal of the SemanticEye-FOAF activity was not to provide an alternative framework, but to exemplify how SemanticEye could potentially plug in to such frameworks that understand the FOAF vocabulary. Li *et al* are investigating the use of FOAF within oreChem Chem$_x$Seer[45], a semantic digital library that extracts metadata from chemical publications. Again, it would be premature to determine how the oreChem activities and SemanticEye will ultimately align.

Bradley's Open Notebook Science[46] is a wiki where researchers share information about their ongoing projects in an open forum and encourage others to participate. This is a high profile example of a "chemical collaboratory" using Web 2.0 technology. In another example, Willighagen is interlinking chemistry related blogs via InChIs[47]. Although the SemanticEye-FOAF research is aimed at the next generation Web 3.0, the chemical collaboration potential of Web 2.0 technologies, as reported comprehensively by Williams[48], cannot be overlooked and therefore will be reviewed in this thesis.

[iv] The author would like to thank Colin Bachelor at the RSC for his input.

O'Reilly notes, however, that Web 2.0 technologies are not novel[49] and argues that similar services have been in use much earlier and Kolbitsch and Maurer argue that open hypermedia systems such as Hyperwave have been fulfilling this role[50].

2. Chemical Journal Web Document Management

The CLIC Consortium was a collaboration of the chemistry departments of three universities and one learned society, Imperial College, Leeds and Cambridge Universities, and the Royal Society of Chemistry. CLIC investigated new methods for chemical electronic publishing on the Web.

The research undertaken principally at Imperial College as part of CLIC will be reported in this chapter. This focused primarily on the use of Hyperwave, an open hypermedia solution, for achieving scalability when managing large volumes of structured journal documents. It will be demonstrated that this solution provided a good model for the future electronic publishing paradigm where journals, images, 3D graphics and collaboration are all linked.

Currently, virtually all scientific publishers use XML technologies for managing journals. This chapter will also examine the XML based workflow that scientific publishers currently follow.

In 2010, virtually all scientific publishers use Web document management solutions based on XML for managing journals. This chapter will also summarise the current publishing process around XML and provide an overview of the same.

2.1 The CLIC Project

When the Web first emerged, publishers initially perceived it as a means of supplementing paper journals, not as the basis for the publishing paradigm which it ultimately became. The goal of the JISC e-Lib funded CLIC Project[14] was to derive an understanding of the real potential of the Web for electronic publishing by modelling an electronic journal in a manner which exploited the virtues of the Web. CLIC used a selection of fifty RSC articles as its test bed. The articles were primarily from Chemical Communication but a sampling of articles from other journals was also included. One of the main deliverables for the CLIC consortium was structured and scalable chemical journal content management.

Because CLIC was an academic initiative, the total cost of ownership of a commercial implementation of the CLIC deliverables was not taken into account. It was recognised however that the long term commercially viability was important nonetheless.

The following sections summarise the main CLIC constituents which would reflect the requirements of any modern electronic journal.

2.1.1 Journal Presentation

Notwithstanding the need to uphold the journal's characteristic look and feel, restricting the electronic journal to a rigid style was not desirable. Ideally both the creator and the reader would have some control over the presentation. Moreover, any alteration of presentation should have no bearing on the underlying content that is being presented. This effectively allows content to be presented in multiple ways which is only possible when the presentation and content are separated.

2.1.1.1 SGML

The separation of presentation and content is the hallmark of the Standard Generalized Markup Language (SGML), an International ISO ratified language for describing electronic documents. It emerged in 1986 from generations of notes scribbled on manuscripts to instruct the typesetters. SGML provided device-independent and system-independent methods of representing text in electronic form. It was therefore appropriate for large publishing applications and was indeed the basis for HTML.

SGML introduced the notion of a separate document type definition (DTD) defining the constituent parts and their structures. The definition of a journal publication, for example, might be organised by title, author, abstract, etc. Anything lacking an abstract, according to this formal definition, would not be a formal publication, nor would an experimental section followed by an abstract. If a known DTD was specified within the document, a parser would then be used to process it and check that all the elements required for that document type were indeed present and correctly ordered. The DTD would either be embedded within the SGML document or referenced as a separate document.

SGML had severe limitations which spelled its demise. It was considered too general. It also contained too many extra smarts designed to minimize keystrokes in an era when bytes were expensive. As a result, SGML design and maintenance were extremely expensive. These limitations made SGML ill suited for the rapidly evolving Web which required more flexible standards.

2.1.2 HTML Creation

CLIC restricted its study of presentation format to the Hypertext Markup Language (HTML), the core language of the Web. PDF, which was created by Adobe in 1993, was considered out of scope because at the time, PDF could not handle embedded interactive chemical objects. Subsequent to CLIC in 2003, Adobe released the XML Data Package (XDP) as a companion to PDF. Since 2005, Adobe started supporting interactive 3D graphics in the U3D format embedded in PDF. As PDF has been a mainstay along with HTML for electronic articles, it was investigated as part of the second phase of research that will be described in Chapter 4.

HTML began in 1990 and went through a number of revisions for the first five years. A failed attempt at an HTML 3.0 specification made way for the release in 1997 of HTML 3.2, the first version of HTML that supported Cascading Style Sheets (CSS). HTML4, released the following year, was the last release before the W3C began work on an XML-based serialisation, called XHTML 1.0 which contained no additional new features other than an XML basis. Subsequently, the W3C focused on making XHTML more modular and began working on XHTML2. In parallel, a working party for an HTML Document Object Model (DOM) API was formed and lasted for around six years.

In 2003, after the publication of XForms for Web forms, a renewed interest in HTML occurred following the realisation that XML deployment was limited to new technologies like Really Simple Syndication (RSS). Nevertheless, the W3C chose to continue down the XHTML2 route. Consequently, Apple, Mozilla, and Opera joined forces to develop HTML5[51] under the auspices of a new body called WHATWG. Because the HTML5 specification was becoming more vendor-driven, implementation needs started taking on a higher priority. Thus, the scope of HTML5, while encompassing HTML4, XHTML1, DOM2 HTML and XForms, also included features that have traditionally been provided by plug-ins such as document editing, video play, and drag and drop.

In 2006, the W3C joined the HTML5 effort and are now publishing a parallel HTML5 specification that differs somewhat from the WHATWG version. Though not expected to be a W3C recommendation until well into the future (2022), parts of the HTML5 are now stable and will be incorporated into most browsers. HTML5 represents twenty years of progress but still permits the mixing of presentation, content and URLs, like HTML 1.0. In 1996, CLIC had explored surmounting this limitation via two different

avenues. The first, investigated at Leeds University, was to dynamically create HTML on demand from articles stored in SGML using a style sheet. The second, to be described later in this chapter, was to deploy the Hyperwave system which cleanly separates presentation, content and URLs.

2.2 Web Document Management Approach

2.2.1 Metadata Management

Metadata, or data about data, transforms raw data into knowledge. They provide representative characteristics about content entities in much the same way as a library card catalogue provides information about books within the library. The implementation of metadata can be more complicated than its definition would indicate. The distinctions between text based data objects and metadata are often ill defined and painstaking modelling is required to keep the metadata unambiguous.

Historically, metadata was centralised owing to the centralised nature of databases within mainframe computers. The current trend is moving towards more decentralised metadata models to coincide with the distributed information “islands” that are characteristic of the Internet. A publisher’s journal repository can be considered as one such “island”.

It is well known that Web search engines use HTML metadata to improve indexing. An HTML document holds its metadata within header fields. However, archiving metadata embedded within the content was not desirable for CLIC because HTML, images, and chemical supporting data could also be assigned metadata. Thus the management of metadata consistently for all the content could only be achieved with a more centralised approach.

Much has been written about metadata modelling and management and the reality is that, like data management, no magic bullet exists that will address all of the challenges. Ultimately, the selection of metadata, supporting technology, and standards should make the documents easier for both humans and software to locate and manage.

2.2.1.1 Dublin Core

Dublin Core[52] is an ISO standard for a minimum number of text metadata elements to describe and catalogue physical resources such as books, digital media, or text files.

"Dublin" refers to Dublin, Ohio, U.S., where the work originated. "Core" indicates that the metadata elements are an expandable "core" list. Dublin Core metadata are intended to be used cross-domain and have become standard in library and computer sciences. The implementation of Dublin Core makes use of XML and RDF, which will be discussed in Chapter 4.

The Dublin Core standard has two sets of elements, Simple and Qualified. Simple Dublin Core comprises fifteen elements (Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage and Rights). Qualified Dublin Core includes three additional elements (Audience, Provenance and RightsHolder) and qualifiers that improve the semantics of the elements in manners that may be useful in resource discovery. An early Dublin Core implementation for chemistry has been published[53].

2.2.2 Hyperwave

Hyperwave[54] is an open hypermedia system originally developed by a team of researchers at Graz University of Technology. Initially called Hyper-G, Hyperwave actually predates the Web as it is popularly known, and initially differed fundamentally from the Web ethos. It gradually implemented Web technologies while maintaining its attractive features. Hyperwave moved from an academic to a commercial venture[55] but was free to academics during the time of CLIC. It currently promotes its collaboration features as a competitive advantage.

Hyperwave's architecture, underpinned by a database, was highly advantageous compared to the unstructured way in which Web documents were then (and for many Web sites, still are) managed within folder trees. Hyperwave's database configuration consists of three servers which would not be unlike modern Web document management systems. The object server, initially a home grown object oriented database that was subsequently replaced by Oracle, stores all document metadata and URLs. The full text server was essentially a search engine that again was home grown and subsequently replaced by the Verity[v] product. Hyperwave's native full text engine that indexes HTML was used for the CLIC study. The document cache server stores and manages the actual documents.

[v] Verity was acquired by Autonomy in 2005.

Although, much of Hyperwave's functionality, such as user authentication and authorisation, and collaborative authoring, can now be found in commercial Enterprise Content Management packages, it offered additional distinctive and very useful features:

URL Management: URLs are managed as separate database objects with metadata. This provided the ability to manage URLs in a manner that was consistent with the way in which other documents were managed. Another useful feature was the ability to retrieve document metadata by appending a fixed string to the document's URL.

Multiple Presentation Options: Content creators need only focus on the text. HTML presentation is applied through the use of either CSS, document templates written in the Hyperwave's own scripting language, or server side JavaScript. A different interface to the same or different content can be provided for different user groups.

Multi Clusters: An HTML document in Hyperwave can consist of multiple HTML documents clustered together. The multi cluster document contains metadata separate to that of the individual documents that comprise the cluster.

Native SGML Support: Although never tested as part of CLIC, open hypermedia systems were designed to serve SGML documents. This would have been advantageous to publishers all of whom worked with SGML at the time.

Hyperwave was perhaps the most popular of several noteworthy open hypermedia systems, all of which failed to gain more than a niche following despite having features that were considered advanced for the time. Project Xanadu, the first hypertext project founded in 1960 by Ted Nelson, who coined the terms "hypertext" and "hypermedia", is still ongoing. Nelson's vision of hypermedia opposes that of Tim Berners-Lee: *"The Xanadu project did not 'fail to invent HTML'. HTML is precisely what we were trying to PREVENT-- ever-breaking links, links going outward only, quotes you can't follow to their origins, no version management, no rights management."* [vi] Like Hyperwave, the Xanadu hypertext model manages by-directional hyperlinks and the versioning of

[vi] Ted Nelson's Computer Paradigm, Expressed as One-Liners. http://xanadu.com.au/ted/TN/WRITINGS/TCOMPARADIGM/tedCompOneLiners.html (accessed 12 June, 2010).

content. The first release of an implementation was in 1998 and was considered incomplete.

2.2.2.1 URL Integrity

URL integrity in open hypermedia was actively researched in the late 1980s and early 1990s. In addition to Hyperwave, the Microcosm[31] project was considering this issue in depth. However, the scale of the Web became too vast for URL validation solutions. Indeed, the Web expanded so rapidly because content creators did not need to worry about URL maintenance since, in the Web metaphor, URLs were allowed to break.

As mentioned in Chapter 1, renewed interest in URL integrity was taking place in the early 2000s with the movement of the Web into corporate Intranets. Intranet content creators wanted assurances that URLs to and from their documents would continue to work throughout the document life cycles. It can also be concluded that publishers would have the same requirement for their Internet based journals.

This requirement led to new research in the area with some interesting findings. Replication and versioning technologies could permit readers to preserve Web content of their choosing using a browser plug-in[56]. The Robust Hyperlinks[57,58] method effectively automates the manual approach of using search engine technology to relocate missing Web documents. Ideally for publishers, URLs would never break in the first place!

2.2.2.2 Hyperwave URL Maintenance

Upon archival, Hyperwave strips URLs from the HTML and stores them as objects with metadata in its database. Within the HTML, the URLs are replaced with placeholders. URLs are mapped back on to the article placeholders when the article is accessed. Within the URL metadata, document identifiers specify the document within which the URL resides and the document(s) to which the URL is pointing. By this means, URLs become bi-directional and documents will not only know to where its URLs are pointing but also the documents that are pointing to it. The major advantage here is that if a document is removed all URLs that point to it become inactive and the URLs will no longer appear within the pages pointing to removed document. However, as the URLs still reside in the database, if the document is replaced at a later time, the URLs pointing to it will reappear as well.

Hyperwave's stateful server/server protocol guarantees URL maintenance across multiple Hyperwave instances. If a document is removed from one server, all links pointing to it from other Hyperwave servers are also removed. This feature is potentially very useful for an Intranet environment with multiple Hyperwave instances serving up large volumes of interlinked technical documents.

2.2.2.3 Offline Journal Production

Creating an archive of publications on optical media for distribution and offline viewing has always been tedious, requiring the organisation and validation of documents and their URLs. This was exemplified with the creation of the CD-ROM version of the ECTOC conference[59]. Certain functionality and URLs would be missing from the static offline version. Moreover, all text must strictly adhere to HTML rules for future proofing. Automating the creation of offline versions of journal content would require the Web content to be properly structured and validated. Hyperwave's CD Publisher tool was ideally suited for automated CD-ROM production.

2.3 CLIC Electronic Journal Production at the RSC

All of the work discussed herewith in this section was carried out by the RSC and represents the manner in which electronic chemical journals with enhancements would have been produced around 1996/7.

2.3.1 Manuscript Preparation

A set workflow was followed for creating the enhanced articles for CLIC. Articles were normally selected by the RSC for enhancement. Occasionally, articles were enhanced at the request of the author. The author would submit manuscripts as word-processed files. If the manuscript was not in Microsoft Word format, the RSC converted it into the same. Following peer review and editing, the RSC would convert the Word documents into structured HTML using a specialised program. In parallel, another program converted the Word documents into SGML for the typesetters. The structures, reactions and analytical data were represented as gif images.

When enhanced structures were provided, they were represented as PDB or MDL[vii] MOL files and rendered with the MDL Chemscape Chime plug in. Chime is based on the Rasmol viewer[60], is still available[61] and was confirmed to work in Microsoft Internet Explorer 8. It visualises 2D and 3D molecules and 3D proteins. In total, fifty enhanced articles were prepared most of which were from Chemical Communication. In addition to the enhanced articles, an HTML version of the Illustrated Contents List of Chemical Communication was also made available. The enhanced articles were placed on Web sites at the RSC, which used a standard Web server and Imperial College Department of Chemistry, which used Hyperwave.

2.3.2 CLIC article Navigation: Standard Web Server

At the RSC Web site readers would be directed to a separate CLIC location (Figure 1) where they would navigate to the enhanced articles using a hierarchical TOC that resembled a computer folder structure. The TOC was a separate HTML document in the left hand frame containing JavaScript. It was manually edited by the RSC whenever a new enhanced article was added. Figure 2 shows how the TOC would be used to navigate an enhanced article and to locate one of the two 3D molecules included with the article. The 3D molecules are rendered in the Chime plug in.

The CLIC articles were successfully rendered recently in Internet Explorer 8; yet they were prepared thirteen years ago and had not been touched since. This indicates the backward compatibility current Web browsers have with older versions of HTML and the staying power of JavaScript.

[vii] MDL is now Symyx (http://www.symyx.com/)

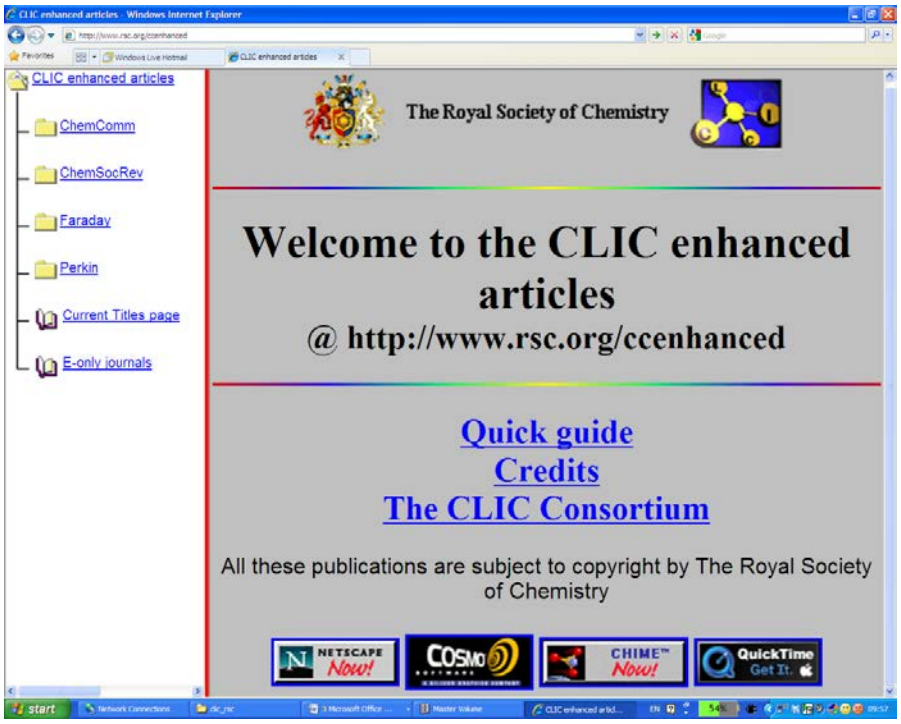


Figure 1: The CLIC homepage at the RSC website

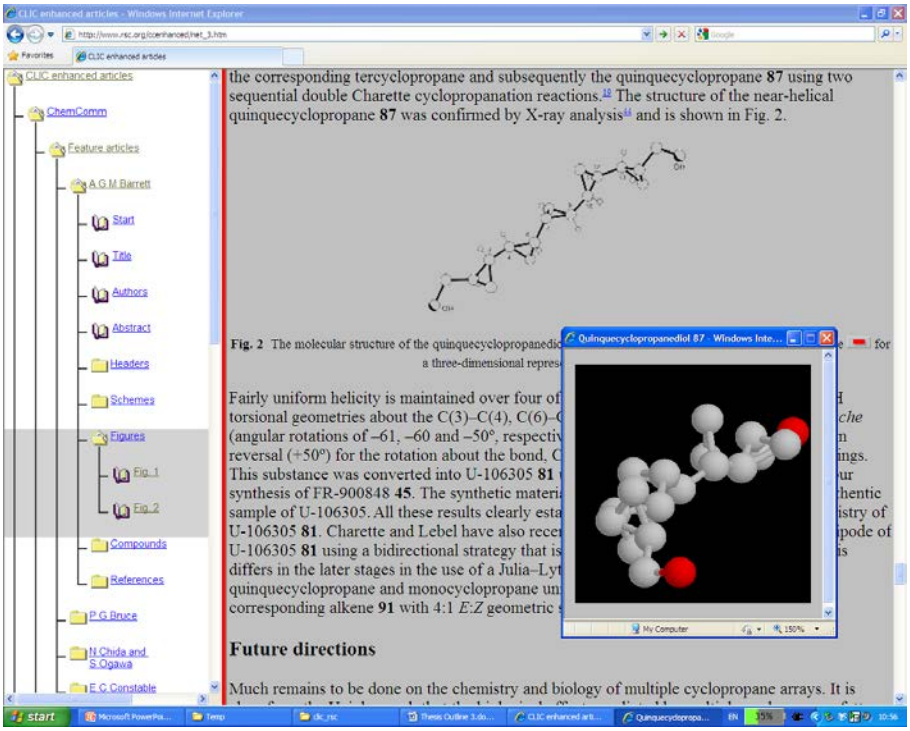


Figure 2: A 3D molecule rendered in a separate browser window

2.4 CLIC Electronic Journal Production with Hyperwave

This section describes the production of electronic chemical journals using Hyperwave and how it differs from the process described previously.

2.4.1 Manuscript Preparation

Once the HTML articles and supporting data files were prepared, as described previously, they were captured in Hyperwave. Documents could be captured individually using a Web browser (Figure 3) or in batch using a command line interface. Hyperwave's own client called Harmony would be used for more advance document managing and URL maintenance. Metadata would be assigned to the document upon capture and new metadata terms could be introduced at this point. When the HTML was inserted, all of the URLs were automatically stripped out of the HTML, stored in a metadata repository and replaced by placeholders in the text.

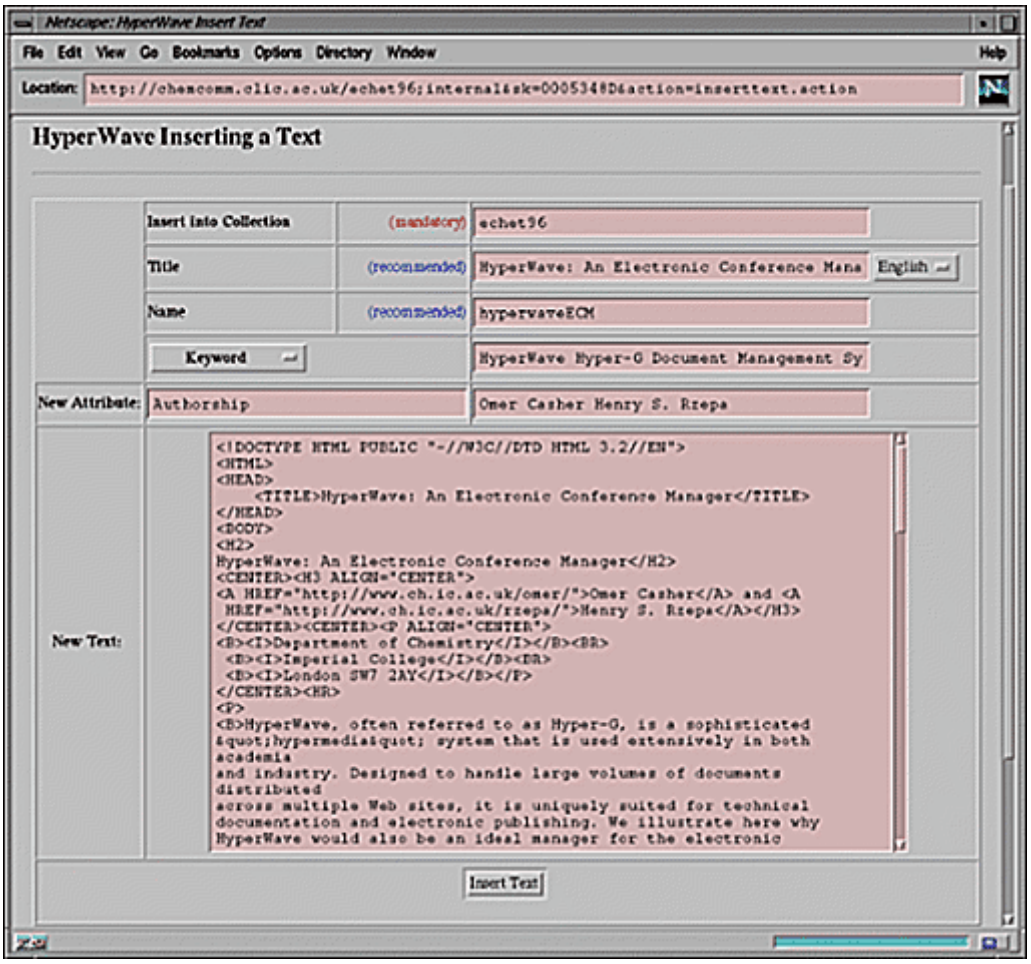


Figure 3: Hyperwave HTML text insertion and metadata editing using a Web form

Figure 4 displays one of the CLIC articles. It contains a supporting molecular structure represented as PDF in Chemscape Chime. For this example, the PDF molecule was stored in Hyperwave. A mechanism for storing supporting structures in a separate database will be described in Chapter 3. As with HTML, the structure had associated metadata, irrespective of where it was stored.

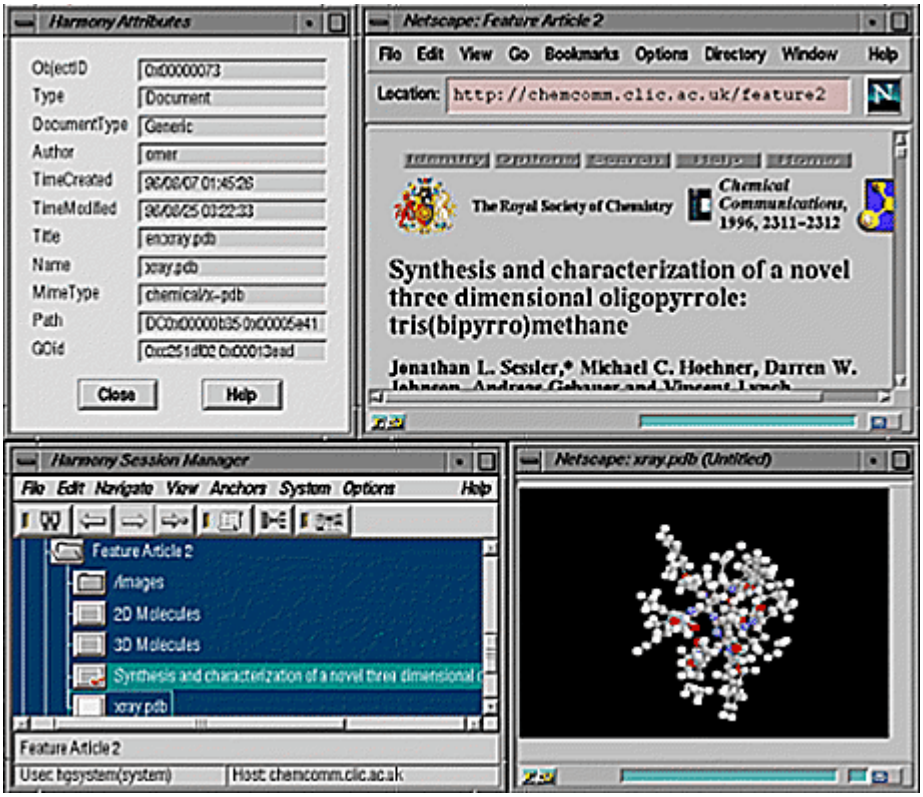


Figure 4: A CLIC article containing a PDB file

2.4.2 Link Maintenance for CLIC articles

Hyperwave's URL maintenance feature is illustrated in Figure 5. The HTML page contains an embedded PDF molecule, which was managed as a URL. The Harmony client displays the document's two URLs as well as all the URLs that point to it. The URL metadata include access rights and expiration dates. As manuscripts can be moved, removed, replaced or modified, such URL maintenance is invaluable. This URL maintenance feature should not be restricted to HTML and would also work for PostScript, PDF, images, MPEG movies, audio files and 3D molecular structures.

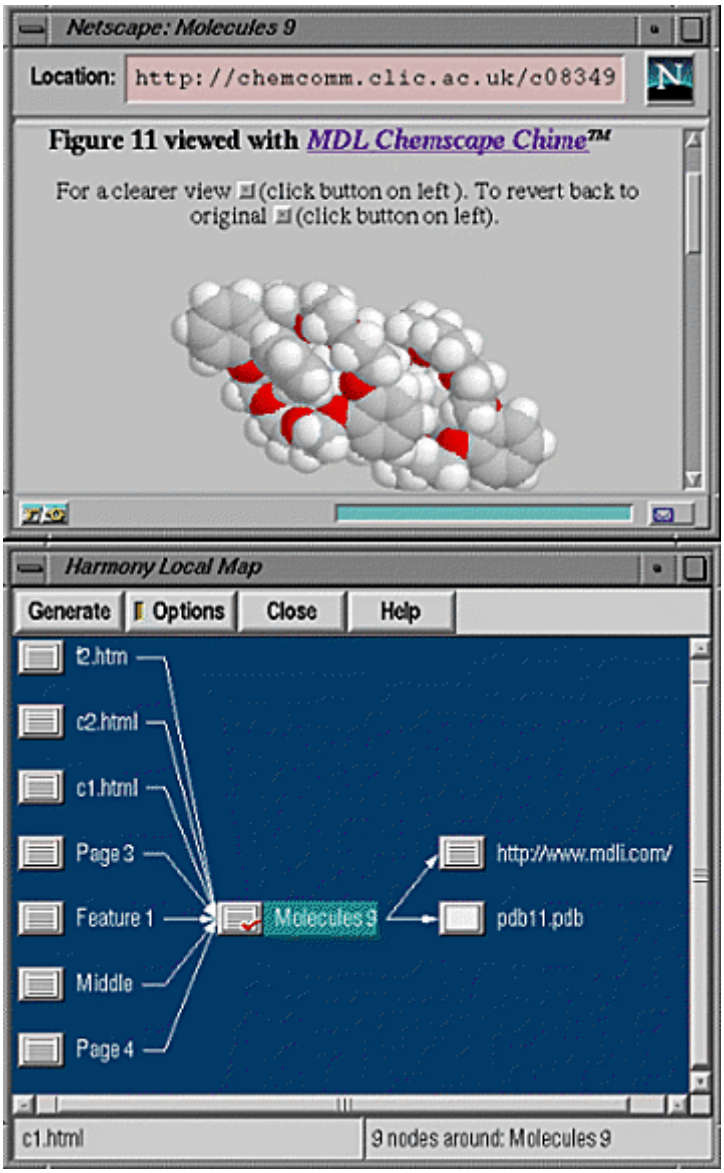


Figure 5: The bidirectional links of an HTML file with an embedded PDB file

2.4.3 CLIC article Navigation: Hyperwave

The presentation of the CLIC articles shown in Figure 1 and Figure 2 was recreated within Hyperwave but with one fundamental difference. The TOC described previously was part of the HTML content. Within Hyperwave it did not exist as HTML but was dynamically generated using a script that interrogated Hyperwave's metadata database for relationships. The process of dynamically generating TOCs from metadata interrogation was an important finding of this research. It is scalable and precludes the need for manual HTML editing. Furthermore, because it would only be generated on demand, it would never be out of sync with the content.

Hyperwave's document template feature, called PLACE, was investigated. PLACE provides dynamic Web page functionality in a language that is geared more towards content creators than programmers. Hence the language is at a higher level than that of Hypertext Preprocessor (PHP), Microsoft Active Server Pages (ASP), and Java Server Pages (JSP) but it lacks their flexibility. Nevertheless it proved to be suitable for the tasks required. Figure 6 shows how it was used to insert navigation buttons as a header to every page.

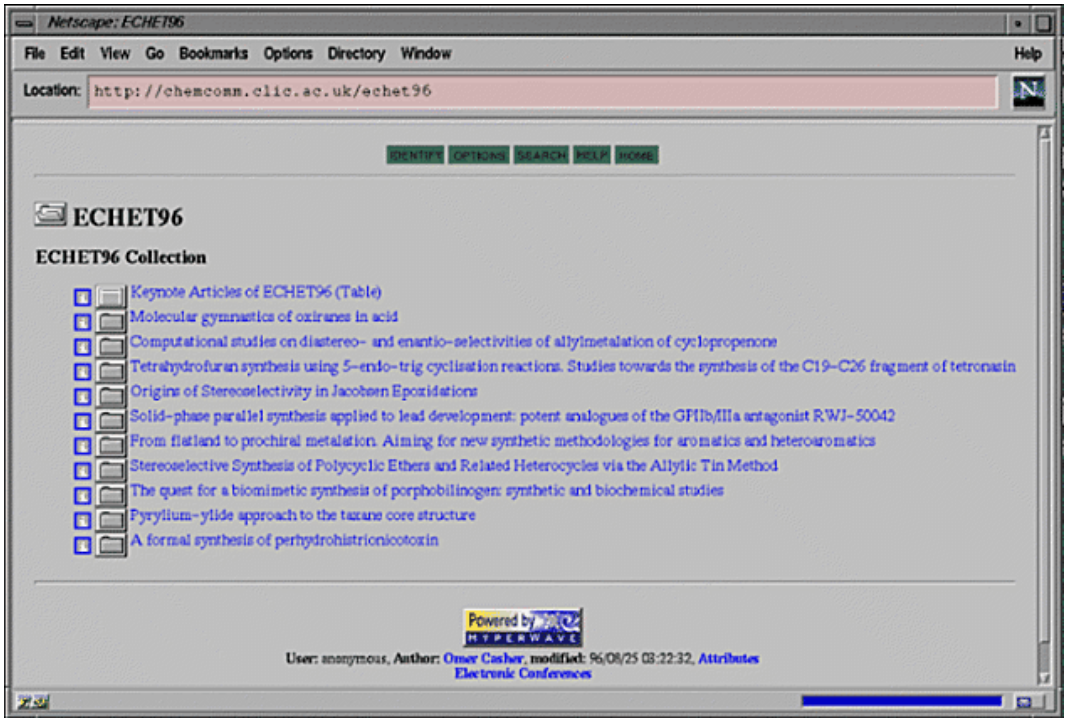


Figure 6: A document template inserts navigation "buttons" as a page header

A common problem encountered, even at present, when navigating to a Web page from the hit list of a search query is that the context of the Web page may be lost, i.e., the reader cannot be directed from the Web page to any other content on the same Web site because the page lacks the necessary links. The reader would therefore not know which site is serving that page. This problem was addressed for CLIC using the Hyperwave PLACE template that was written to insert, at the bottom of each HTML article page, a link which, when activated, navigated the reader to the exact location of the page in the TOC. The link was generated by metadata interrogation. A well structured publisher Web site would provide navigation capabilities at this level, or better.

Proper metadata management, as has been highlighted here, is a prerequisite for scalable scientific publishing. The approach described in this section was more recently

revisited within the realm of chemical publishing, RDF and the Semantic Web and will be discussed in Chapter 4.

2.5 Open Hypermedia and Collaboration

Chapter 5 will discuss where chemical publishing fits within a chemical collaboration network. In addition, Web 2.0 technologies, increasingly being used in a collaborative manner, will also be discussed. The Hyperwave model discussed in this chapter can be considered as an exemplar of a relatively early chemical collaboration network that is arguably more sophisticated than its modern counterparts. In particular, manuscript creation, publishing, and collaborative document creation were all integrated in a single environment. If the explosive popularity of Web 2.0 sites such as Wikipedia and Facebook provides any indication, open hypermedia may have been ahead of its time.

To explain this missed opportunity, certain aspects of Web 2.0 applications distinguish them from their hypermedia predecessors. The novelty is not *what* the modern services do but *how* they do it. All of them have one defining feature: The more people that use them, the better they get and the more respectable they become. Moreover, the relatively recent availability of the Internet to the masses resulting from the reduction in the cost of hardware and services has lead to the uptake of Web 2.0 by a critical mass of users which would never have been achievable in the 1990s.

A problem with open hypermedia systems was a relatively high barrier to entry; a certain amount of technical competence was required in order to derive benefit from them. In contrast, Web 2.0 users are drawn in to them by invitations from other users and find the new experience relatively intuitive, even if their technical skills are basic. Otherwise, the document management and versioning technologies that underpin Web 2.0 are actually narrow in focus and basic compared to open hypermedia technologies.

In summary, open hypermedia systems represent an important step in electronic publishing evolution as well as the development of this thesis. They provided some useful electronic publishing features, namely metadata management, that will continue to be explored throughout the remainder of this thesis. Certain other strong features, however, provided an undertow to open hypermedia's advancement, such as a relatively high barrier to entry and URL integrity that is better suited for corporate Intranets. For Internet based chemical Web publishing, alternative technologies proved to be more appropriate and will be discussed in the following sections.

2.6 Current Scientific Publishing Process

Most publishers have implemented eXtensible Markup Language (XML) repositories with Web content management capabilities, the most popular of which is the MarkLogic product. A brief summary of the publishing process, from manuscript submission through to publishing, is given here in order to illustrate where XML management fits within it[viii].

1. An author submits a manuscript, usually in Microsoft Word format.
2. A Document Object Identifier (DOI) is created for the manuscript and this DOI will never change.
3. The manuscript is sent by the editor to the referees for review.
4. If necessary, the manuscript is revised by the authors based on feedback from the referees.
5. If the revisions are satisfactory to the editor, the manuscript goes through the editorial cycle.
6. The manuscript is sent to offshore typesetters where it is marked up in XML against the publisher's XML Schema.
7. After proofreading, the XML manuscript is inserted into an XML repository containing the published journals.
8. Metadata are added to the article where appropriate.

Because of the core role of XML in journal production process it is summarised in the following sections. An emphasis will be on those XML standards that are used extensively within the MarkLogic product.

2.7 XML Standards

XML is a set of W3C standards or proposals. Since its creation in 1996 by John Bozak of Sun Microsystems, an international think tank of academic and industry visionaries formed working groups within the W3C to draft a large, dynamic body of documents that constitute the XML specification[62]. Some early XML based applications in the molecular sciences arena have been published. GXML[63] is a consolidation of diverse genomic data into an XML format suitable for querying. StarDOM[64] is a transcription from the STAR NMR format into an XML document object model, or DOM. In the chemistry arena, the

[viii] The author is indebted to the RSC for having shared their process for the benefit of this thesis.

ongoing Chemical Markup Language project, (CML)[65], is one of the first true XML testimonials.

Anybody is at liberty to extend the standards and build applications that understand the extensions. However, the widespread adoption of any modifications normally requires W3C approval. XML is a huge and growing standard and describing all of its features is outside of the scope of this thesis. However, a summary of its core features, most relevant to publishing, are summarised in the following sections.

2.7.1 XML Syntax

From the specification it can be seen that XML has only a few rules. One is that XML containers come in pairs. Container pairs can be nested inside one another. The nesting rule provides a simplified tree structure for every XML document. Each graphic and text component within the document represents a parent, child or sibling of some other component. Although trees cannot represent every type of information, they can represent most types understandable by software. Parsing trees programmatically is straightforward. An XML example is shown in Figure 7.

```
<Company>
   <Name>ABC Solutions</Name>
   <Employee>
      <Name>John Hander</Name>
      <Title>CEO</Title>
      <Employee>
         <Name>Dhir Pathak</Name>
         <Title>VP Finance</Title>
      </Employee>
      <Employee>
         <Name>Seth Jones</Name>
         <Title>VP Sales</Title>
      </Employee>
   </Employee>
</Company>
```

Figure 7: An XML example of a simple organisational chart

XML has adopted the Unicode standard, a character-encoding system that supports a combination of text in all the world's major languages. The primary advantage is that software that reads XML properly can deal with any combination of any of these character sets. Thus, XML enables the exchange of information not only between different computer systems but also across national and cultural boundaries.

The use of containers follows a rigorous set of rules enabling the data to be easily encoded in human readable form and unambiguously interpreted by software programs.

XML differs from HTML in a number of ways one of which is that the XML containers preserve the semantics of the subject matter. A fundamental difference is the way in which the presentation is handled. XML is derived from SGML which keeps presentation and content separate, thereby allowing the text to be stored in databases while the application that renders the text handles the presentation.

2.7.2 XML Validation

Because HTML browsers are somewhat forgiving, HTML documents are often not valid. XML rules are stricter; if XML does not follow the W3C specification, the parser will report it as being invalid. If an XML vocabulary will be supporting a focused subject area with multiple content creators, XML validation is usually necessary. This type of validation defines the elements that are permitted to be children of other elements along with the attributes that individual elements can have. XML documents without formal validation can be parsed and rendered. However, validation ensures that XML documents adhere to the content creator's guidelines and remain consistent. Two validation methods which have been defined by the W3C are described in the following sections.

2.7.2.1 Document Type Definition (DTD)

As with SGML the XML DTD specifies all the valid elements of a specific XML vocabulary along with the permitted ordering and attributes of XML elements. The DTD can be contained within the document itself or referenced externally. A DTD consists of a single root element that marks the beginning and end of a document. Within the root element, child elements are nested and these might contain their own child elements. An XML parser validates an XML document by determining whether its elements are structured to conform to the DTD.

Because XML is data centric as well as document centric it maps to database constructs. However, a database schema is more expressive than a DTD. DTDs have no notion of type information and constraints such as permissible value ranges. Database schemas also permit relationships and dependencies to be defined which is not possible with DTDs. The DTD may therefore be suitable for document centric XML but inappropriate for data centric XML.

2.7.2.2 XML Schema

Like a DTD, an XML schema describes the XML data. However an XML schema has richer data typing and can be mapped to a database schema as illustrated in Figure 8.

```
<schema targetNamespace=”myNamespace”>
 <element name=”table”>
  <element name=”rows” minOccurs=”0” maxOccurs=”*”>
   <element name=”tableColumn1”>
        <datatype source=”string”>
           <length value=”50”/>
        </datatype>
         </element>
         <element name=”tableColumn2”>
       ….
   </element>
</schema>
```

Figure 8: An example of an XML Schema that maps to a relational database schema

2.7.3 XML Namespace

XML namespaces provide a mechanism to uniquely qualify elements from different vocabularies. Consider, for example, the element <Name> which can have different meanings in different vocabularies. XML namespace will guarantee that this element is uniquely defined. The XML Namespace is essentially a Uniform Resource Identifier (URI), a generic term used to describe names and addresses that reference locations on the Web. Because URIs can be long, aliases are permitted. The namespace uses the xmlns attribute as follows:

```
<ElementName xmlns:prefixname=”AUniqueURI”>
```

Hence an “abcname” prefix can be defined as follows:

```
<abcname:Name xmlns:abcname=”http://www.abc.com/Name”>

<abcname:Name>John Hander</abcname:Name>
```

Not to be confused with the URL, when the URI within the xmlns definition is parsed, the parser will not attempt to fetch a document, which is not expected to exist[ix].

[ix] Although a namespace is used purely for name qualification, for programmatic convenience some XML developers choose to define namespaces to correspond to physical document locations. This approach, although legal, runs counter to the spirit of XML Namespaces.

The Namespace has the notion of element scope. By defining a namespace for a specific element, the element is in scope and the namespace is inherited by all the child elements. A child element can redefine a namespace prefix and any subsequent child element will inherit the new namespace.

2.7.4 Manipulating XML with DOM and SAX

Once an XML document has been parsed the data will most likely be manipulated. Two APIs are provided for this purpose. As discussed previously, an XML document is a container set where the containers are structured in a nested manner. An XML document is viewed as a tree structure with each container being a node in the tree. When an XML document is parsed, a Document Object Model (DOM), which is essentially a node tree representation of the XML, is stored in memory. The DOM API provides a complete set of operations to programmatically manipulate the DOM including navigating the tree nodes, creating and appending new nodes and removing nodes. Once the node tree has been modified it can be exported as a new XML document. Unlike the DOM, the Simple API for XML (SAX) does not build a node tree in memory. The advantage of SAX over DOM is that documents much larger than the available system memory can be parsed. The SAX approach is therefore well suited for very large XML files.

2.7.5 XPATH

The XML Path Language, or XPATH, is the W3C standard to traverse down containers in an XML document for subsets of interest. The syntax is simple; an XPATH is a node hierarchy to the element of interested. The first node is always the root node with the nodes separated by "/" and the name of the node element desired prefixed by "@". The result of an XPATH query can be a node hierarchy, a Boolean, a number or a string.

2.7.6 Extensible Stylesheet Language (XSL)

XML containers describe the information's meaning, independent of the display medium, which is advantageous for electronic publishing. However, containers provide no clues about how the information should appear electronically or in print. XSL style sheets are required to format the content for various rendering devices. The XML browser will read an XML document, fetch the required XSL style sheet, and use it to sort and format the information displayed on the screen. The rendered presentation is not necessarily XML. Although this process is transparent to the user, a style sheet can be tailored to the

individual user's needs. Style sheets can be used to render the information in any form such as Braille or speech.

The XML Stylesheet Language for Transformations (XSLT) is an XML construct that transforms XML documents into other XML documents. Its relationship to XSL is often ambiguous. XSL is a style sheet language for XML. It encompasses XSLT along with an XML vocabulary for specifying formatting.

XSLT elements belong to the XSL namespace:

<xsl:stylesheet version="1.0" xmlns:xsl=http://www.w3.org/1999/XSL/Transform>

More than one XSLT style sheet can be simultaneously applied to an XML document, the effect being identical to a single style sheet containing the constructs of each individual style sheet by the order in which each style sheet has been referenced. An XSLT processor normally uses XPATH and DOM to select the required information and perform the transformation. XSLT has become a mainstay in e-commerce, particularly in business-to-business (B2B) transactions where the two linked entities handle different vocabularies.

2.7.7 XML Linking Language (XLink)

XLinks are elements in XML documents that describe links between resources. The W3C standard defines a framework for unidirectional links, such as URLs, and bi-directional links as would be found in hypertext systems like Hyperwave. An example of a bi-directional link in action would be an electronic journal publication that knows which publications cite it. Hence XLinks can describe relationships among more than two resources. This is achievable because XLinks can contain metadata. The XLink elements within rendered XML documents will allow (a) multiple destinations to point from a single link, (b) direct insertion of the linked document within the page, (c) indirect links that point to database entries instead of other pages.

2.7.8 XML query language (XQuery)

XQuery is a query and programming language that queries XML data. It is normally instantiated within an XML document and uses XPath syntax to render specific parts of the document. It can also be used to construct new XML documents. The vocabulary is based on a tree-structured model of the XML content. XQuery 1.0 cannot update XML

documents or databases and lacks full text searching. These features will be part of a future release of the specification.

2.8 XML Databases

An XML database stores XML data for querying, exporting and serialising into the desired format. There are principally two categories of XML databases. The first is a native XML database which built around XML and may be based on the hierarchical database model although this not a prerequisite. The second is a traditional database, often relational, which internally maps XML to its own schema. A strong selling point for using an XML database is the improved efficiency attained by the ease of storage and conversion.

At a minimum the XML database model must support elements, attributes, XML Schema validation, XPath, and both DOM and SAX parsing. Like with Hyperwave, XML databases group documents into collections, often hierarchically structured like an operating system's directory structure. All XML databases support at a minimum XPath for querying against documents or collections and many support XQuery. Most support XSLT. Some XML databases support the XML:DB API (or XAPI) a vendor neutral interface for accessing the database from client applications.

2.8.1 MarkLogic

MarkLogic, founded in 2001 as a response to the growth of XML, is one of the leading XML database companies. It is widely used within the publishing industry by organisations such as Elsevier, Oxford University Press, the RSC, Nature Publishing, McGraw-Hill and the New England Journal of Medicine. It also has a following in other domains such as financial, defence, and government agencies.

MarkLogic claims that its native XML server can scale to over a hundred terabytes of content by using cloud computing[66] algorithms. It also supports XQuery and includes a full text search engine. Its rich functionality includes application services with high level APIs to enable the creation of flexible information centric applications including document management functionality. MarkMail is a useful utility that permits email to be managed and searched within the MarkLogic server.

As with Hyperwave, MarkLogic has a metadata repository where metadata are exposed using web services. Because of its native XML support the metadata can have

different XML structures. Applying XSLT translation to the XML enables the metadata to be accessed in a uniform manner.

Although MarkLogic was not implemented as part of this research, the potential of exposing its metadata to the Semantic Web via XSLT is potentially useful. Furthermore, it supports features such as XMP and WebDAV that are highly relevant to the latter phase of research discussed in Chapter 4.

2.9 Summary

In this chapter the CLIC project, one of the earliest exemplars of an electronic publishing model, was presented using a test bed of online RSC chemical journal articles. The article collection was marked up by the RSC in the Standard Generalised Markup Language (SGML) and rendered in HTML. Readers could then locate the articles with the help of a hierarchical navigation aid written in JavaScript.

The CLIC model became a template for electronic chemical journal implementations using scalable and structured content management packages. This chapter reported on one such implementation using Hyperwave, a sophisticated open hypermedia system which actually predated the Web. Hyperwave consists of a Web server and a database within a content management framework. This configuration addressed several of the shortcomings of the Web, namely scalability and bidirectional URL integrity. In keeping with the SGML (and subsequently XML) paradigm, presentation and content are kept separate. Hyperwave makes use of metadata management and document templates for structuring and presenting its content.

The Hyperwave implementation as part of CLIC consisted of inserting all of the CLIC articles and supporting data into the Hyperwave database. The JavaScript hierarchical navigation aid was replaced by a navigational aid that was dynamically generated by interrogation of the metadata in the database. One of the strongest features of this implementation is that documents could be reorganised without the need to update any URLs.

Although the Hyperwave approach addressed the Web's shortcomings it failed to make significant inroads within the Web. Nevertheless the Hyperwave chemical publishing exemplar provided an early look in to the framework that was to be adopted by scientific publishers, and a model that was the basis for Semantic Web research that will be discussed in Chapters 4 and 5.

The Extensible Markup Language (XML), which emerged late in the CLIC project cycle, is the established standard for Web document markup and arbitrary data exchange. Most scientific domains have associated XML vocabularies, the structure of which would be described in XML Schemas. XSLT is used to translate one XML vocabulary into another. Generic XML parsers would validate the hierarchical data structure and individual data field formats within an XML document against the XML Schema.

Having largely superseded SGML, XML has become the lingua franca for electronic publishing. There are literally hundreds of XML vocabularies in existence and the structured and scalable electronic chemical journal concept that Hyperwave underpinned can be almost completely realised using appropriate XML vocabularies and repositories such as MarkLogic.

3. Molecular Supporting Data Management

To a large extent, chemical electronic publishing is essentially an Internet based delivery of electronic articles that are no more than digital replicas of their paper counterparts. Most do not contain any supporting chemical data or semantics to enhance the reader's experience in a manner that would add scientific credibility, assist the peer review process and expedite follow on research.

One of the CLIC activities was to explore the use of a molecular database as a way of incorporating molecular structures within chemical journal articles in a manner that precludes the need for maintenance on the part of the author. A 3D graphics format was also investigated as an alternative means of presenting molecules based on the underlying data provide by the author.

3.1 The Elusive Enhanced Chemical Article

Prior to being published in 1996, the first CLIC article containing a supporting 3D structure vexed its author. He complained that the molecule, rendered in Chemscape Chime, was of poor resolution and did not project a desirable viewpoint. This was despite the fact that he did not play an active role in this enhancement other than having provided X-ray coordinates. Unbeknownst to him, the molecule embedded in the electronic article was an interactive 3D model. This story exemplifies the perception that readers had towards enhanced articles containing interactive data. Arguably this perception does not differ much in 2010 as enhanced chemical articles are still elusive.

The explanation for this impasse is the differing perceptions of enhanced data from the five principle stakeholders of a journal article.

3.1.1 Publisher

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 enhanced articles would improve the reader's experience, arguably they would not generate the necessary revenue to cover the ensuing support costs. Most

publishers would prefer to maintain their established revenue-generating infrastructures and leave the management of non revenue generating components to others. If the enhanced chemical electronic journal were to become viable, the publishers' core business model should not be disrupted. Nevertheless, although publishers do not proactively request supporting data from authors, they would not be resistant if they were provided, but would only be prepared to provide minimal support.

3.1.2 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[67].

Supporting data would greatly assist the peer review process as the data would ameliorate the manuscript.. The notion that supporting data can improve both the quality and efficiency of the peer review cycle should be a strong driver for their inclusion. In reality however, publishers may not even attempt to ensure that the presentation of the manuscript is suitable for peer review, let alone attempting to seek supporting data from the authors.

3.1.3 Archivist

Whereas the long term archival of printed journals has traditionally been a library service, for electronic journals, publishers assume the role of archivist, who would charge for access. For the archival of supporting data, the authors would assume the archivist role, as it is not a publisher activity. It has been predicted[68] that institutional repositories, which are central electronic archives set up in universities and centres of excellence, will play a part in the archival of supporting data. Authors would deposit the supporting data into 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[69] at MIT. They provide tools to capture, store, index, preserve and redistribute an organisation's electronic research material.

3.1.4 Reader

As with the reviewer, readers would be a prime beneficiary of supporting data, which would serve as a benchmark allowing readers to study, repurpose, and refine the analysis

methods. It is therefore perplexing that supporting data are still elusive despite the added value demonstrated by the CLIC project, the Hyperactive Molecules[3] work prior to CLIC, and the IJC[32] subsequent to it. Perhaps readers are not being more demanding here because for their own manuscripts they assume a different role, being that of the author.

3.1.5 Author

For the enhanced electronic chemical journal to achieve critical mass, authors would be tasked to provide, along with the electronic manuscript, data results. However, this is where the greatest resistance lies. Many chemists would be concerned that the sharing of supporting data to scientists outside of the research team risks misinterpretation and the reuse of data out of context. By far the greatest deterrent for authors however is amount of effort required in preparing supporting data[70]. “Climategate”, the University of East Anglia's (UEA) Climatic Research Unit (CRU) email controversy, was a high profile incident that manifested a "culture of withholding information"[71].

3.2 An Optimal Approach

Publishers may assume that mandating supporting data would incite scientists to publish elsewhere rather than to comply. In reality however the author has a proclivity for publishing in as respected a journal as possible. If a highly respected journal mandates the inclusion of results data, the burden of complying will be absorbed by the author(s), albeit begrudgingly. 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 to address. A more streamlined process for the preparation, storage and management of supporting data would also help.

The remainder of this chapter will discuss the technologies required for managing supporting 3D molecular structure data, with a goal of better integration, for the benefit of all five stakeholders. The actual preparation of supporting data by chemists is dependent on their workflows and is therefore outside the scope of this thesis. Also out of scope are other non 3D molecular chemical data types and formats, the creation and management of which may be governed by a totally different set of dynamics. However, the approach discussed here could very well be applicable in other domains.

The technical underpinnings for managing supporting 3D molecular structure data consist of a suitable file format and repository. The following sections discuss the options along with their pros and cons.

3.2.1 XML for Molecular Supporting Data

As XML is a mainstay in scientific publishing, supporting data should ideally be serialised in an XML vocabulary and provided with the correct version of the corresponding XML Schema if it exists. The archival of supporting data would be greatly facilitated if authors were to standardise on widely adopted XML Schemas such as that of CML[65], an open standard and one of the first domain specific XML vocabularies. CML actually comprises several published XML Schemas covering different chemical entities such as molecules, reactions and spectra. An alternative to CML for molecules is the MDL XDfile, a repurposing of the published and widely used SDfile, which contain molecular structure coordinates, connectivity information, and additional data fields. A number of chemistry software tools such as the popular Jmol[72] understand both CML and XDfile formats. Because both XML and CML started to gain traction subsequent to the CLIC project, supporting molecular structures for CLIC articles were represented in the well established SDfile format.

3.2.2 3D Graphics Formats for Presentation

As part of the CLIC project an alternative method of 3D molecular representation was investigated using the Virtual Reality Modelling Language (VRML) format and associated viewers. Figure 9 illustrates a "molecular hyperglossary"[10], essentially a 3D data mining model based on VRML and generated using a visualisation application called EyeChem. EyeChem[2] was developed at Imperial College from 1992-1998. Although it is no longer being actively developed, it is still being distributed and maintained by Nag as part of IRIS Explorer[73], a commercial visualisation package.

The hyperglossary, which can have thousands of files associated with it, starts off with a scatterplot showing oxygen contacts to the centroid of chlorobenzene. Only those contacts that penetrate the Connolly surface are highlighted. Selecting a given region of the scatterplot brings up a subset of the points to simplify selection. Each point is then hyperlinked to the corresponding structure and associated literature. This hyperglossary illustrates a sequence of hyperlinking ending with a publication; a 3D molecule is actually linked to the publication, not vice versa. The hyperglossary therefore becomes an alternative entry point to the journal. This notion of linking from the molecule to the publication is revisited later in this chapter.

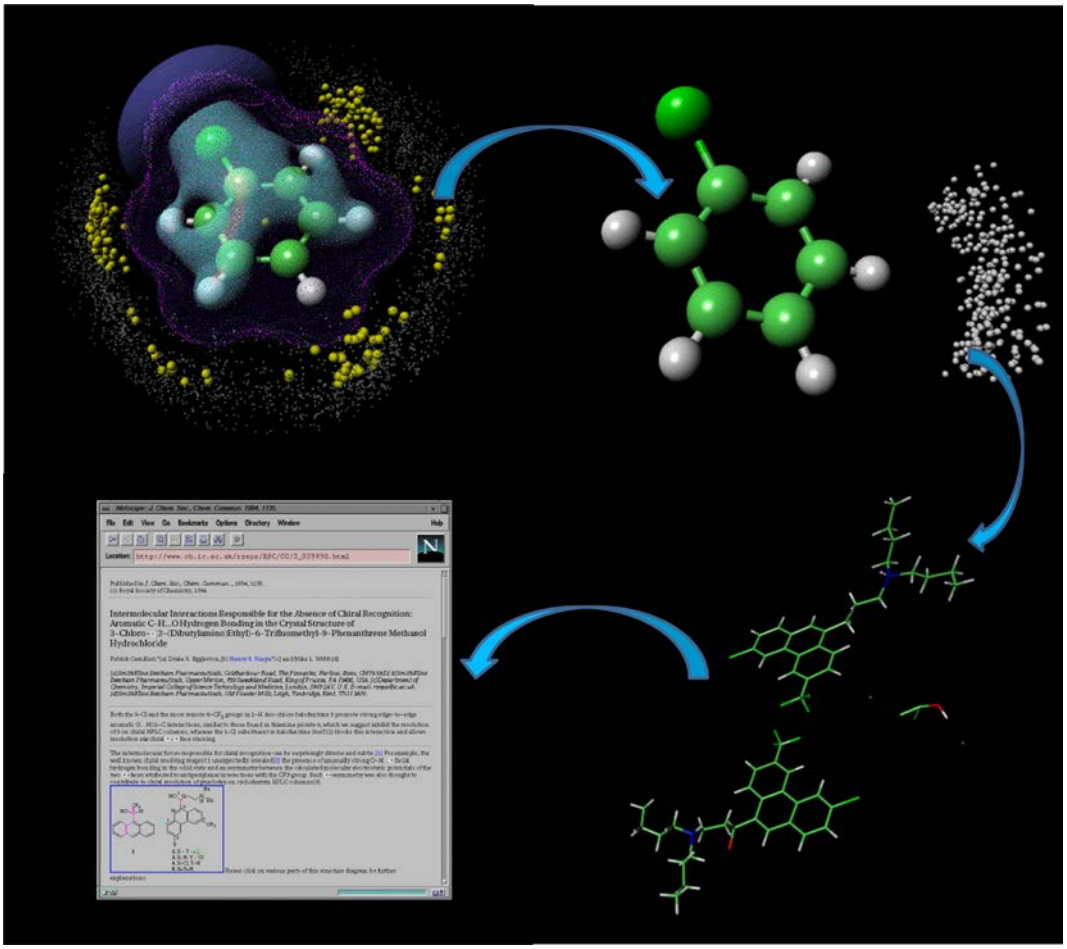


Figure 9: A 3D molecular Hyperglossary

VRML emerged in 1994 as a 3D extension to the two-dimensional ASCII character set describing HTML. VRML contained a set of three-dimensional primitives, useful for molecules such as spheres, cylinders, and nurbs (protein ribbons). It was designed to be used on the Web, hence its support for hyperlinked objects. SGI were actively involved in the VRML specification until its restructuring in 1998 following on from which VRML started losing momentum. It was ultimately replaced by X3D, essentially a repurposing of VRML97 into XML. X3D browsers are freely available and X3D export is supported in a number of applications. A VRML-to-X3D translator is also available.

X3D is only one of a number 3D graphics formats vying for dominance. Another is Universal 3D (U3D), a compressed file format standard for 3D defined by the 3D Industry Forum. The format is natively supported by the PDF format and 3D objects in U3D format can be embedded into PDF documents and interactively visualized by Acrobat Reader 7.0 and later. 3D Markup Language for Web (3DMLW) is another XML-based 3D file format. A 3DMLW plug-in developed by 3D Technologies R&D runs in

the popular web browsers. COLLADA is a COLLAborative Design Activity for 3D graphics. Originally created by Sony, the copyright is now shared between Sony and the Khronos Group, a member-funded industry consortium. It is popular among commercial games engines. WebGL is currently under development as the 3D graphics constituent of HTML5 and is also managed by the Khronos Group. The working party includes members from Apple, Google, Mozilla, and Opera. WebGL is based on OpenGL, a graphics library developed by SGI in 1992. It will run using JavaScript and not require a plug in. WebGL is being implemented for most major browsers. O3D was created by Google as another graphics language that runs within the Web browser. Google recently announced that O3D would be changing from a plug-in to a JavaScript library running on top of WebGL. Jmol is able to export to several of these formats including VRML, X3D and U3D.

3D graphics for chemistry appears to have had more championing and direction in the mid 1990s than in 2010. It is therefore not expected to be integrated with scientific publishing, in terms of 3D rendering or archival, any time soon despite its advantages and uses.

3.2.3 Molecular Repository Approach

In this section, three approaches for managing supporting molecular data are provided and the merits of each are discussed.

3.2.3.1 Archive Molecular Data with the Article

Supporting molecular data that are archived with the corresponding article would be managed in a manner identical to that of the associated assets embedded within, or linked from, the article. Publishers would therefore be utilising existing infrastructure to manage the supporting data although hardware infrastructure such as storage may need to be increased to deal with the additional data. In this approach, the supporting data live with the article (Figure 10) and would travel with it if the journals are migrated to a different repository. This approach mitigates the risk of supporting data being deleted, or the links to them breaking.

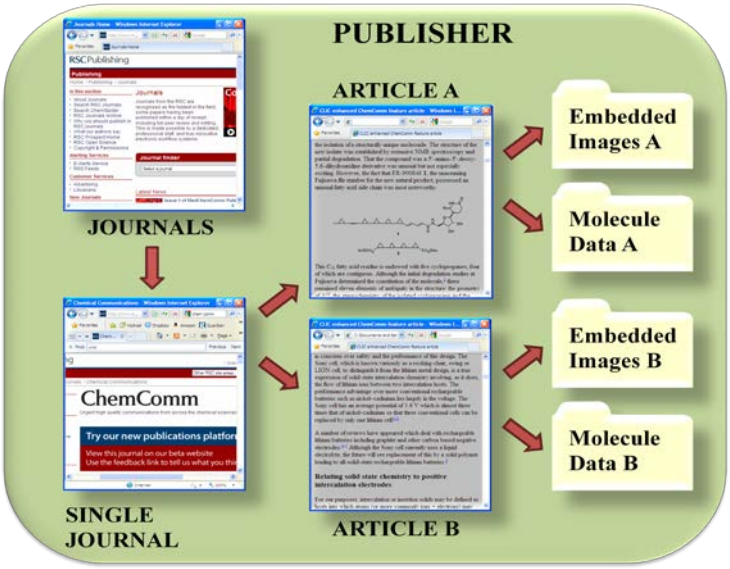


Figure 10: Molecular data are archived with the articles

The publisher would be taking on the role of the archivist who would need to consider the long term archival requirements of supporting data. The life cycle of the supporting data and the parent article differ. Whereas articles are expected to be available for posterity, supporting data do not have the same requirement. Normally, the desirability of supporting data is greatest at the period in which the article is first published, and diminishes over time.

The fact that most publishers use XML repositories would have a bearing on the choice of file formats. If an XML format is used, such as CML or XDFile, the XML features of the repository can potentially be leveraged in a useful manner. Some of the issues that would need to be resolved would include substructure or similarity searching and achieving bidirectional linking between the parent article and its supporting data

3.2.3.2 Archive Molecular Data in an External Archive

Unlike the management of supporting molecular data publisher-side as described previously, in the scenario here, the molecular data is being managed in an external digital archive outside out of the publisher's control (Figure 11). For example, SPECTRa, the Imperial College digital repository[74] has archived data from around 25 published articles thus far. The growing popularity of institutional repositories reflects the increased centralisation of data archival particularly for larger data sets. As a single institutional repository can service an entire academic department, or more, it can be used as an archive for supporting data. Because the lifecycle of content stored in an institutional

repository is down to content owners and not the publishers, it would be impossible to ensure that supporting data follow a consistent lifecycle. However the probability that they will be available increases closer to the time that they were archived which is normally when the research is most relevant.

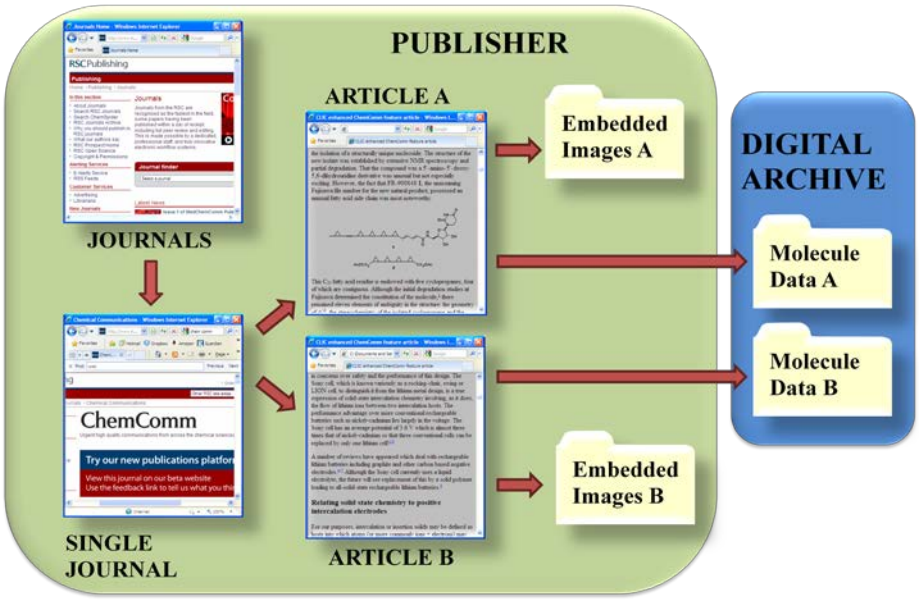


Figure 11: Molecular data are stored in digital archive external to the publisher

The publishers would have to create the URLs from the parent articles to the supporting data. It would be down to the custodians of the supporting data to assign the unique identifiers for the molecular data and to ensure that the links to them remain valid. Integration standards to institutional repositories would be helpful and the use of the Semantic Web for this purpose has been explored by Mason[75]. By also following a Semantic Web approach for chemical publishing, as will be discussed in Chapter 4, achieving bidirectional linking between the parent article and its supporting data would be trivial.

3.2.3.3 Archive Molecular Data in a Molecule Database

The archival of supporting molecular data within a dedicated database managed by publishers (Figure 12) represents the best approach from an integration, data and URL integrity and long term archival standpoint. The use of a molecular database has been explored in CLIC and will be discussed. However, only the larger publishers arguably have the economy of scale to manage molecular databases in house and the smaller

publishers may have to rely on external services. There would also be a worryingly strong incentive to sell access to the molecular data.

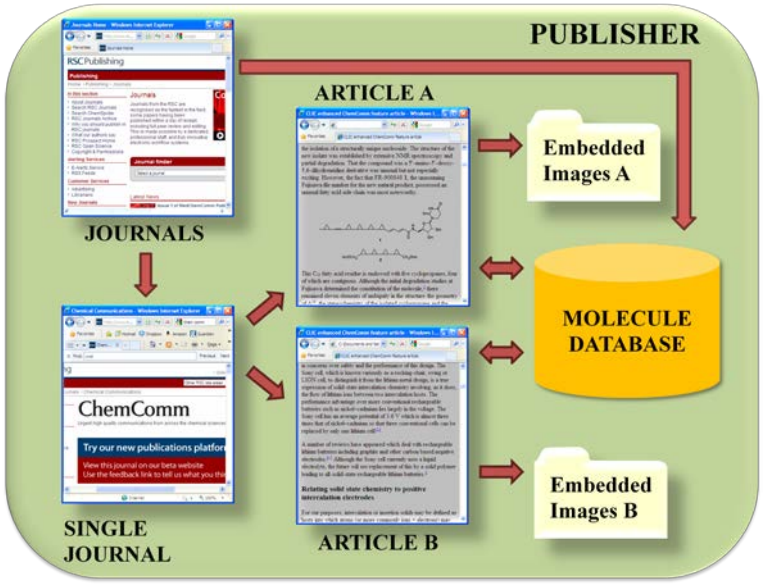


Figure 12: Molecular data are archived in a molecule database publisher-side

3.3 Management of Molecular Data in the CLIC Project

In the CLIC model discussed in Chapter 4, structures that were stored in Hyperwave could be searched against their metadata. The notion of publication metadata containing searchable chemical semantics was not explored during CLIC but will be discussed in Chapter 4. The aim instead was for 3D models to be retrieved by searching against structural features and, if available, chemical properties. Searching the structures in a chemically aware manner therefore required the implementation of a molecular database and appropriate integration with Hyperwave.

The MDL ISIS/Host database and the MDL Chemscape Server[x], a Web gateway to ISIS/Host, were installed. The Chemscape Web interface had similar functionality to that of the ISIS/Base client. Using server side and client side JavaScript, custom Web interfaces were able to be constructed to register and search for structures. Chemscape Chime Pro, an enhancement to Chime for communicating with the Chemscape server, would render the structures. Although the supporting data investigated in CLIC were

[x] Chemscape Server is no longer marketed. Symyx DiscoveryGate is now the Web interface to ISIS databases.

restricted to molecular structures, the Chemscape framework permitted other non molecular data sources to be plugged in.

Figure 13 illustrates the high level architecture of the configuration investigated as part of CLIC. The Hyperwave Web interface is used for the structure retrieval and metadata management features of the molecular database. The architecture would be similar if the Hyperwave components were replaced by those of an XML repository, such as MarkLogic, as discussed in Chapter 2.

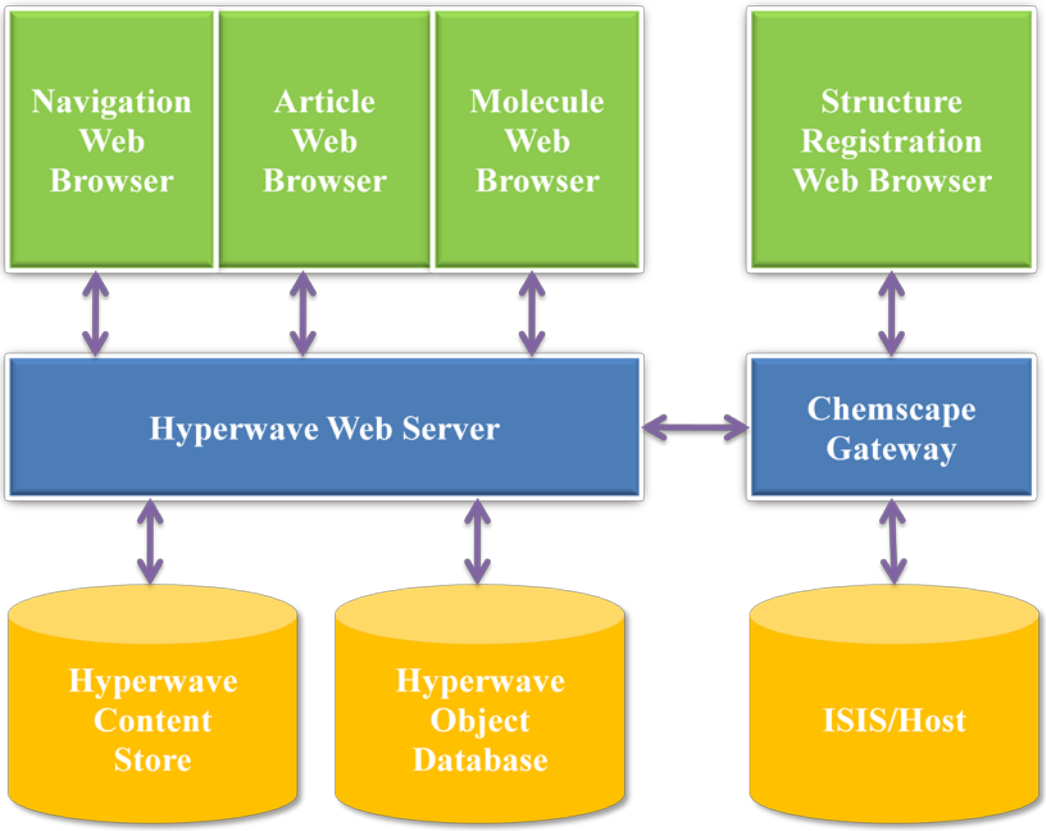


Figure 13: High level publishing architecture with a molecular database

The 3D models were either supplied by the authors or converted from their 2D representations. Most 2D models were supplied as ChemDraw files. Each structure was registered in ISIS/Host using the Chemscape Web client. One metadata term was a Chemscape URL containing the ISIS query required to retrieve a structure. Another was a unique identifier generated by Hyperwave. In Chapter 4 the use of the InChI as a unique identifier with chemical semantics will be discussed.

3.3.1 Management of URLs from Article to Structure

When a given molecular structure was registered in ISIS/Host using Chemscape, the Chemscape interface also issued a Hyperwave command that created the corresponding Hyperwave database object and appropriate metadata. When a structure was removed from ISIS/Host using Chemscape, the deletion of the URL object and metadata from Hyperwave were achieved as an administrator action: The URL object and metadata would be either logically or physically deleted from Hyperwave manually.

3.3.2 Structure Retrieval

All of the structures associated with a parent article were listed in the CLIC TOC that was implemented within Hyperwave, as described previously. Selecting a URL would fetch the structure and display it the Web browser. Substructure and similarity searches against the ISIS/Host database were performed within Chemscape. Molecules could therefore also be retrieved from the query hit list.

In this implementation the structures in the hit list did not contain associated URLs of the parent articles. Unlike VRML and X3D files, SDFiles and XDFiles do not support links. Consequently, the only mechanism by which hyperlinking could be achieved would be through a customisation of the Chemscape interface to include URLs to articles within hitlists. Building such an interface is time consuming, requires additional support, and would create a vendor lock in, which is undesirable when pursuing more open standards.

The alternative, analogous to the VRML approach, would be to directly embed, within the molecular model, URLs that a molecular viewer would fetch. Such a solution does not yet exist but it is not inconceivable. It has been investigated for Computational Chemistry using CML and XLink standards[76].

3.4 ChemSpider

ChemSpider[77] is a chemical database released in March 2007 and acquired by the RSC two years later. It aggregates and indexes chemical data from both open access and commercial databases and provides URLs to information of interest. The data consists of molecular structures and properties, curated literature with their associated URLs, chemical vendor catalogues, environmental data, toxicity data and analytical data. It contains millions of chemical structures including the Wikipedia chemical structure collection. Its entire content is searchable through a single common interface.

ChemSpider is proving to be a useful tool in automating chemical markup where chemical names are automatically identified within documents and web pages and converted to chemical structures as InChI strings. A variety of Web Services are provided for automating tasks such as structure conversion and data retrieval. The InChI conversion Web Service was integrated to an external service[47] which purposes the InChI into RDF. ChemSpider is therefore an ideal solution for managing supporting chemical data in a manner that the CLIC Chemscape implementation was aiming to achieve. It is also an appropriate fit within the Semantic Web metaphor that will be expanded in the following Chapter. ChemSpider merits future investigation for chemical publishing.

3.5 Summary

The rendering of supporting data files containing 3D structures was investigated as part of CLIC. Such files are usually rendered in freely available dedicated tools such Jmol. For CLIC, Chemscape Chime was used for embedding 3D rotatable molecules within articles. As an alternative to the SDfile and Chime, VRML, a generic 3D vocabulary for the Web, which emerged shortly after the Web's inception, was also investigated as a means of rendering 3D molecules within generic VRML viewers. VRML supports graphics primitives that are important to chemistry. It also supports URLs which could link to publications, a feature that is not widely available with traditional molecular file formats and browsers. VRML has been superseded by an XML version called Extensible 3D, or X3D.

Although CLIC focused on how supporting data could be archived within databases managed by publishers, such as the comprehensive ChemSpider, it is also conceivable that supporting data could be archived in resources outside of the publishers' control. The growing popularity of institutional repositories reflects a move to more increased centralisation of data archival. However, because the content stored in an institutional repository is managed by the authors and not publishers, there is no guarantee that the assets will always be available. Irrespective of where the supporting data are ultimately held, publishers need to proactively promote the adoption of generic global standards for supporting data, such as CML in order to ensure that supporting data remains future proof.

4. Chemical Journals and the Semantic Web

Chemical assets with semantically meaningful metadata would improve the discoverability of such assets, including journal articles, on the Internet. However, most chemical journals provide little in the way of useful metadata. The work carried out henceforth is a reaction to this and represents the second phase of research discussed in this thesis. When this work commenced in 2005, the Semantic Web was being adopted by the bio-ontology community but it was still largely unknown in chemistry.

An exemplar called SemanticEye was developed to demonstrate how journal articles available on the Web can be linked up using key metadata including InChI identifiers. The metadata model in this exemplar was not developed from first principles. Instead, it was inspired by the metadata management model for digital music, where music metadata both travels with the music file and are managed collaboratively in central repositories.

4.1 Drivers for Improved Chemical Electronic Publishing

4.1.1 The Hype Cycle

Since 2000, chemists read about 40% more journals than prior to the arrival of the Web, primarily because of the convenience provided by the Internet based delivery mechanism[78]. However, as was discussed in Chapter 3, the enhanced chemical electronic journal has yet to be realised, despite the added value. Most electronic media are still 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[79], the Web started to achieve saturation coverage around 2004 and from then on, further growth would not be attained by increased use but by applying new ideas.

Another perspective on Goodman's account can be derived from the application of Gartner's famous "hype cycle"[80] to chemical publishing. Hype cycles indicate when technologies move beyond the hype, offer practical benefits, and achieve widespread uptake. On the hype cycle curve (Figure 14) the Technology Trigger would have occurred around 1994 following the emergence of the Web. The Peak of Inflated Expectations

would have taken place around 1998 to coincide with the boom in popular Web technologies such as VRML. The Trough of Disillusionment would roughly correspond with the demise of the IJC which occurred in 2004, around the time of Goodman's Web saturation point. The year 2010 finds scientific publishing climbing up the Slope of Enlightenment where the widespread adoption of the XML repositories by publishers has been taking place. The Plateau of Productivity in roughly five years is likely to find a more "deconstructed journal" model[81], aligned to the positive aspects of the Web while addressing diverse stakeholder interests. This is the aim of the work described in this chapter.

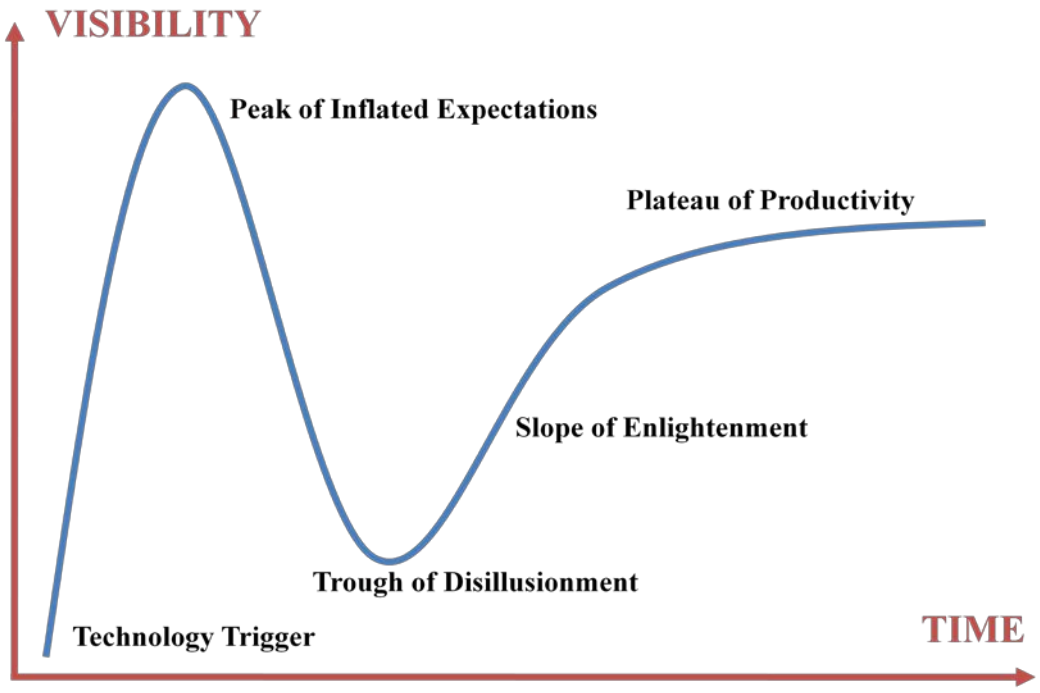


Figure 14: The Gartner Hype Cycle[xi]

4.1.2 The Need for Context and Community

King and Tenopir, who carried out a study on the 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[82]. Three trends here are noteworthy:

[xi] Reproduced from http://blogs.gartner.com/john_pescatore/files/2008/11/hypecycle.png (accessed 27 July, 2010).

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.

A follow up study has not been published for chemical journals but a more recent study was published for a broader spectrum of scientists[83]. In 2005, the percentage of readers who sourced articles by browsing had fallen by roughly 25%, having been offset by the increased use of citations and, more significantly, by the increase in recommendations from people (colleagues, collaborators, etc.). It is assumed that for chemists as a focus group the delta would be consistent. The drivers behind this delta set the stage for the research discussed in Chapter 5.

The fact that journal browsing consistently accounts for most 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 to shape a research community to which the reader would feel an association. However, the 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 on line searching. Hence, despite better communication, electronic publishing has yet to address the reader's need for attaining context and community amongst multiple electronic journal articles. This untapped opportunity is the focus of this study.

4.1.3 Defrosting the Chemical Journal

Herbert Samual's 1963 reference to the pre-digital library as a "cold storage"[84] of information, impersonal and inaccessible to both humans and computers, was picked by Hull *et al*[85] and reapplied to digital libraries in 2008. They make a case for the need for a thawing of the digital library and that second generation reference management applications such as Mendeley and Zotero are, they claim, the right place to start. In defrosting the chemical journal Hull *et al* are taking very much a reader perspective. From the five principle stakeholders the reader perspective is also taken in this thesis. The fundamental difference is that this thesis focuses on metadata management, and whilst

metadata management is intrinsic to second generation reference management, it does not stop there.

The “deep freeze” of the digital library in general and of electronic publishing in particular, is not caused by technology. Technology is a necessity but is not the driver. 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 standardisation that would enable the Web to be fully exploited. Thus, defrosting the existing situation must be carried out external to the journal. The following approach to achieving it is proposed here.

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

The more developed these relationships are, the more effective the Web will be as a medium for journal dissemination. The creation of the relationships can be carried out by one or more individuals, supported by an organisation 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.

4.2 Improving the Metadata

4.2.1 Fundamental Guiding Principles

Although there is no shortage of metadata models and management techniques in 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.

- Anticipate the reader’s working patterns and expectations. A solution which is cumbersome and confusing is potentially counterproductive.

- 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.
- Use open standards wherever possible.
- 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 modelled, the ensuing information and software architecture should ideally be a reuse of existing services. The metadata model must be simplified to minimise software complexity and maximise 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.
- 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.
- 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. Fortunately, one already exists.

4.2.2 Adapting a Successful Semantic Model

Two of the Web's greatest success stories, Google.com and Amazon.com use semantic models[86] though neither is based on the Semantic Web. Google presumably derives a semantic model of the entire Web and relies on hyperlinks between pages for relationships. This permits it to perform more intelligent queries than that of 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 personalised 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[87] and some interesting findings were uncovered.

- Digital music metadata is standardized and moves with the asset, while journal metadata is neither standardised nor embedded.
- Digital music metadata lookup services are collaborative and automate metadata retrieval for digital music. Journal metadata lookup services are not collaborative.
- 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 chapter, the term SemanticEye[19] is coined as a model for rectifying electronic journal shortcomings by adapting the digital music semantic model to chemical electronic publishing. However, unlike the 3 semantic models described above, the Semantic Web model is used as the starting point.

4.3 The Semantic Web

Tim Berners-Lee's vision of the Semantic Web is to "assist the evolution of human knowledge as a whole"[88]. 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 realising this vision, if it is indeed possible, requires some work. Nevertheless, now is a good time to adopt it as an enabler of electronic publishing for the following reasons:

- The Semantic Web principle has been subject to theoretical scrutiny for almost ten years.
- Adoption of its standards amongst the scientific community, particularly in bioinformatics, has been increasing.
- 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. 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 in 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.

4.3.1 The Semantic Web Architecture

The Semantic Architecture is not universally agreed and has been a source of constant debate among a community of developers, researchers and interested parties. The best known representation is Berners-Lee's Semantic Web stack[89] (Figure 15). 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 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.

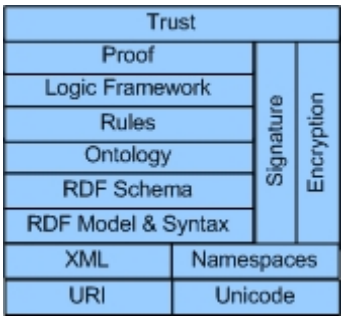


Figure 15: Semantic Web Stack

4.3.2 The Resource Description Framework

The Semantic Web metadata standard is the Resource Description Framework (RDF)[90], 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 graph 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 graph. 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 modelling, makes the content Semantic Web enabled nonetheless. This approach is counterproductive in the long run. The starting point for a Semantic Web solution should be a semantic model off of which the RDF vocabulary and supporting technologies would hang.

4.3.2.1 Web Ontology Language (OWL)

OWL, based on RDF, is a vocabulary used for defining ontologies. OWL adds constraints on the RDF to form a logical model suitable for inferencing. With OWL, the structure, meaning and relationship of any group of documents can be precisely defined. For example, two given ontologies, one containing the term “sugar” and the other containing the term “glucose”, can be semantically linked using the statement owl:IsEquivalentTo. Another OWL statement, owl:disjointClass can describe the fact that glucose and sucrose were disjoint although both are sugars.

OWL is better suited than RDF at the latter stages of a research activity when the knowledge model is well understood and OWL-based constraints can start to be applied.

In the earlier stages when the knowledge model is being built up, the unconstrained RDF, as is used in SemanticEye, is better suited.

4.3.2.2 Semantic Reasoning

The Semantic Web is aimed at furnishing the Web with an Inference Engine. RDF provides a dependable basis for reasoning about the meaning of an RDF expression. In particular, it supports rigorously defined patterns of entailment, the basis for defining reliable rules of inference in RDF data. RDF however specifies no mechanism for reasoning: The ability to reason, query or express logical relations, i.e. to help the Web "think" is down to the RDF creator. A semantic reasoner is a software application that infers logical consequences from a set of rules. The inference rules are commonly specified in OWL. Reasoning is normally carried out using logic algorithms or probability.

4.3.3 Semantic Searching

Semantic searching improves the user's search experience by understanding the contextual meaning of search terms to generate better targeted results. The user enters a phrase which provides context to an object of interest. There is no particular document that the user is trying to access. Instead the user is trying to locate a collection of documents that will answer the query. Semantics are used to rank the results. Unlike with popular search engines, the goal of semantic searching is to deliver the information that the user wants rather than having to sort through sometimes pages of keyword results.

The entire knowledge bank underpinning a semantic search service is a semantic network; each concept is represented by a node that is linked to other nodes by semantic relationships. In this manner, each concept is enriched with the characteristics and meaning of the nearby nodes. Depending on the search algorithm the semantic network can be derived from document metadata and/or computed using sophisticated knowledge management algorithms.

Because semantic searching represents a quantum leap over traditional searching, it is still not mainstream technology. A Semantic Search Survey[91] lists a plethora of solutions and tabulates a few of their technical features. Not surprisingly most of the solutions are within the realm of academia or start up companies. However the major search engines are starting to add, to their core products, semantic searching features.

Yahoo! SearchMonkey[92], released in 2008, is a true Semantic Web application that will index Web content, appropriately tagged with RDF metadata. This tagged content will become more semantically meaningful when included as part of a Yahoo! search result. In early 2009 Google announced the incorporation of semantic capabilities within its search engine[93]. The technology enhances, but does not replace its traditional keyword approach and unlike SearchMonkey it derives semantic meaning algorithmically from associations and concepts related to the search, not from direct RDF metadata interrogation within content. Microsoft Bing[94] takes yet another approach to semantic searching. Through its acquisition of Powerset, Bing enhances searching with Wikipedia semantic lookups. Bing also improves the user experience in four key vertical areas: purchase decisions, travel planning, health condition research, and locating a local business. Yahoo! and Microsoft recently entered into an agreement whereby Yahoo!'s search engine would be replaced by Bing with plans to incorporate SearchMonkey. It is still early days to determine if and how these technologies could be exploited for chemistry.

4.3.3.1 The SPARQL Query Standard

SPARQL, a recursive acronym which stands for SPARQL Protocol and RDF Query Language, is both a data access protocol and an RDF query language. Ratified in January 2008 and widely supported by RDF repositories, it is the *de facto* RDF query standard replacing a multitude of legacy query formats. SPARQL can be used to query diverse sources where the RDF data is either native or dynamically generated. The query results can either be hit lists or, more elaborately, RDF constructions. Hence one or more SPARQL queries across multiple RDF sources can generate useful aggregations as RDF.

4.3.4 RDF Triple Stores

A triple store is a database used for the persistence and accessing of RDF data. Recently there has been a major development initiative in query processing, access protocols and triple-store technologies. The number of triple stores being actively developed has mushroomed from initially two Open Source initiatives in the early 2000s, Jena and Sesame, to over ten. Jena TDB, Jena SDB, Virtuoso, BigData, Mulgara, Kowari, 3Store and RDF Gateway are also Open Source. BigOWLIM and AllegroGraph are commercial products and Garlik JXT and YARS2 are propriety and not distributed.

Triple stores can be divided into 3 categories based on their core architecture: in memory, native and non-memory non-native. In memory triple stores store the RDF graph in main memory. Although this provides the optimum performance, it is not a serious method for storing of large volumes of triples. They are however useful for performing certain operations like data caching and inferencing. Non native non memory triples stores, like Jena and Sesame, run on third party databases such as MySQL, PostsgreSQL and Oracle. Native triple stores, which provide databases built from the ground up, include Virtuoso, AllegroGraph, Mulgara, and Garlik JXT. Their popularity is increasing due to superior load times and ability to be optimized for RDF.

Native triplestores will ultimately have a performance advantage. A drawback however is that because of the manner in which triples are stored, querying a native triplestore efficiently is more challenging compared with querying relational databases.

4.4 SemanticEye Target Architecture

The 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 the approach taken in realising the target architecture.

4.4.1 Dataflow

Because SemanticEye is metadata driven, the Semantic Web model can be landscaped with the help of a dataflow diagram (Figure 16). The various functions of the dataflow can be summarised as follows:

- RDF metadata is captured by the reader using appropriate tools (A).
- RDF metadata is embedded in a manuscript (B) either by the reader, the publisher or both.
- The manuscript is submitted to the publisher (C).
- Following peer review (D) and possible revisions (B) the manuscript is published and made available on the internet (C).
- Journal articles are retrieved through searching or browsing using a Web browser (E).
- Articles can be saved either in a folder on the user's desktop (F) or in a WebDAV folder (G).

- Articles saved in a folder on the user's desktop (F) can be dragged and dropped into a WebDAV folder (G).
- Articles are uploaded to a folder (H) at the Metadata Agency.
- RDF metadata is extracted from the documents and stored in an RDF repository (I), thereby creating semantic associations between the contents of the repository.
- A representation of the RDF can be navigated by the reader in a Web browser (E). Journals and supporting data of interest can retrieved from the internet using appropriate resolving agencies.
- RDF in the repository (I) can be edited via the browser (E).

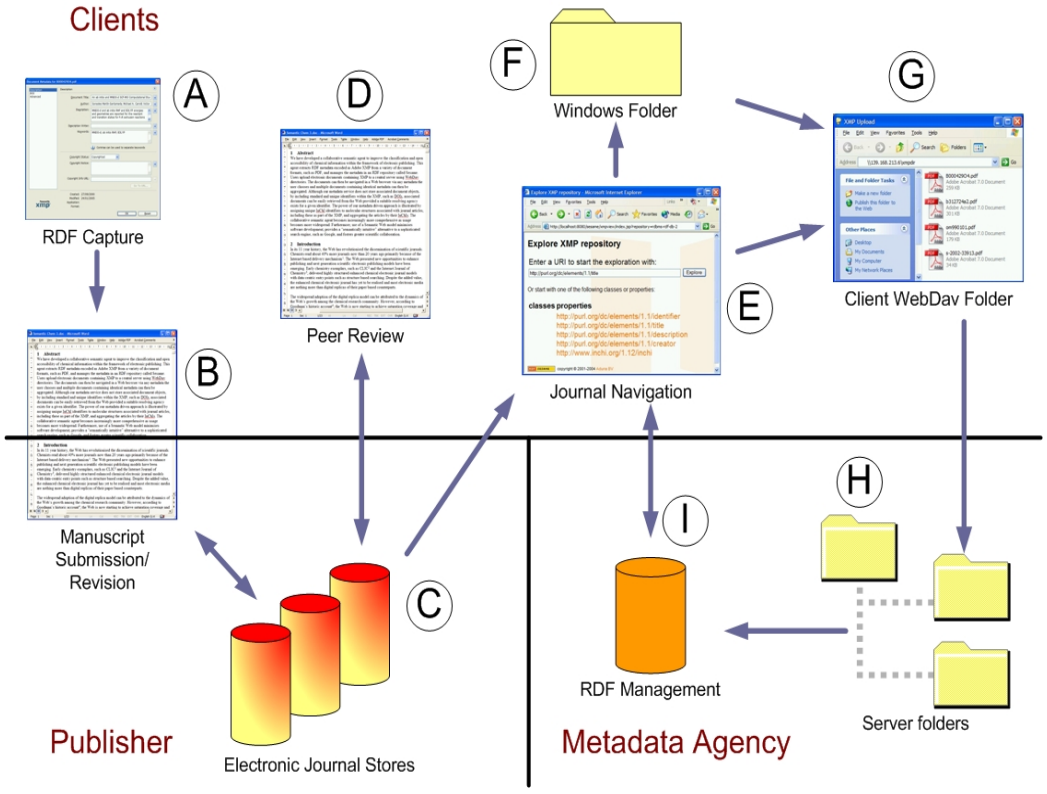


Figure 16: Overall dataflow of the RDF metadata lifecycle

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

4.4.2 The Document Object Identifier (DOI)

The DOI is an internationally recognised 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 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[95], considered the authoritative source of information on the DOI, provides the DOI namespace:

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

The use of a DOI resolving agency has three noteworthy advantages:

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.

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.

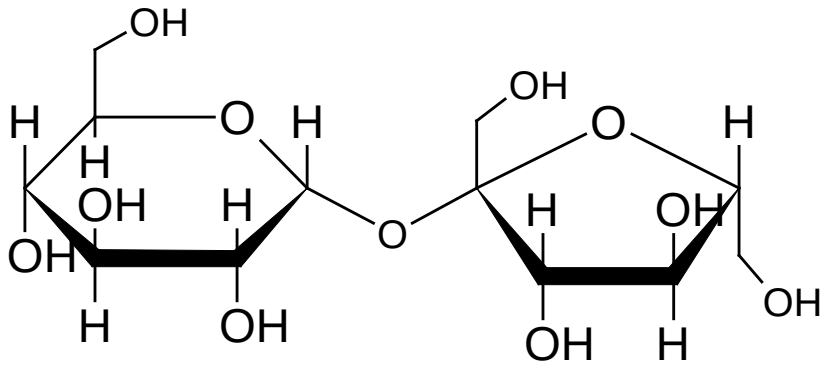
Discoverability: DOI registration identifies content that has been deemed to be of value and whose validity and currency are actively managed. Hence the use of DOIs for electronic publications potentially improves Google Search Rankings provided the DOIs resolve to the actual document and not to a DOI "landing page".

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.

4.4.3 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 serialising it existed. The InChI is the result of an IUPAC project[96] 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 in the future parse) it.

The InChI was developed to provide a standard human readable way to encode molecular information and to facilitate the Web based searching. The InChI expresses molecular information by taking a layered approach where each successive layer adds additional detail to the identifier. The layers are formula, connectivity (excluding formal bond orders), isotopes, stereochemistry and tautomers (on or off). Charges are not part of the InChI layers but are appended to the end of the InChI string. The specific layers are generated automatically from the structures and depend on the structural detail provided. The reconciliation of the InChI for searching has been discussed in detail[97]. An example of an InChI is shown in Figure 17.



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

InChI-key CZMRCDWAGMRECN-UGDNZRGBSA-N

Figure 17: InChI for Sucrose

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 non proprietary, 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. The InChI has been adopted by both public and commercial chemistry databases. Chemical structure drawing packages such as ACD/ChemSketch support it.

The InChI string has a major flaw. Its character set can break most search engines. To resolve this issue the InChIKey was introduced[98]. The condensed, 25-character InChIKey is a hashed version of the full InChI and is not human-readable.

SemanticEye makes no assumption and enforces no restriction on the level of detail that an InChI can have. If an InChI for a structure in Article A has more detail than an InChI for the same structure in Article B, Article B would currently not link to Article A. Because of the layered approach of the InChI, a facility (or Semantic Web agent) for linking up InChIs at lower layers of detail should in theory be straightforward to implement.

4.4.4 Extensible Metadata Platform (XMP)

XMP[99], an RDF standard promoted by Adobe and designed for the management and storage of documents, is a natural starting point for electronic publishing metadata. 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[100], 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 and provides an XMP-API for XMP handling within applications.

The uptake of XMP by scientific publishers was initially very limited due to the slow growth of RDF as a whole, as well as a number of caveats: XMP does not have an associated RDF Schema. Only a succinct specification, which includes the Dublin Core RDF Schema, was handled by Adobe software. Most XMP implementers would want to extend the vocabulary with their own RDF Schemas or external Schemas. A significant example is the Publishing Requirements for Industry Standard Metadata (PRISM) XMP[101] which was implemented to allow PRISM metadata to be embedded in multimedia objects accessible online.

XMP expects its RDF serialised in XML and does not support any of the other RDF representations. In contrast to the many possible ways of representing an RDF/XML

vocabulary (known as RDF polymorphism), XMP extensions require a documented[102] restrictive RDF serialisation to enable the correct handling by Adobe software. If XMP employed an appropriate ontology in OWL, such restrictions could be formally defined. Caveats notwithstanding, XMP is gaining traction amongst scientific publishers. Elsevier started embedding their articles with XMP

in 2007 and their use of XMP has been evolving based on experience[103]. Nature implemented a workflow solution to handle XMP capture whereby automated batch-processing extracts information from XML files and outputs standardized validated PDF files with the XMP embedded. All Nature articles published since December 2008 now contain XMP[104].

XMP is also supported by popular citation-management programs. Both Zotero[105] and Mendeley[106] read embedded XMP metadata from PDFs if available. However, neither of them will populate PDF documents with XMP metadata if the documents do not initially have them. CrossRef, the registrar of the Document Object Identifier (DOI), have release an experimental open source tool called PDFMark[107] that, given a DOI, will fetch the PRISM bibliographic metadata from CrossRef and insert it as XMP into a PDF.

XMP capture can be achieved manually by inputting metadata using a form tool (Figure 18) 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[108].

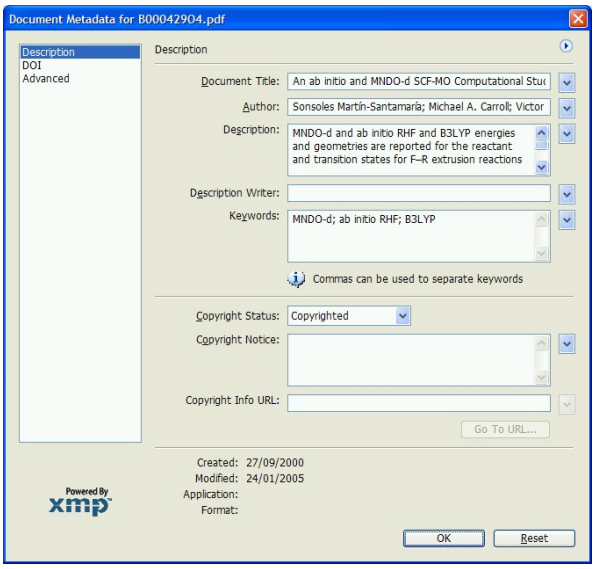


Figure 18: Acrobat Professional XMP Tool

Adobe's XMP support includes a useful mouse-over feature (Figure 19), whereby positioning the mouse cursor over a PDF document icon reveals a number of the XMP attributes.

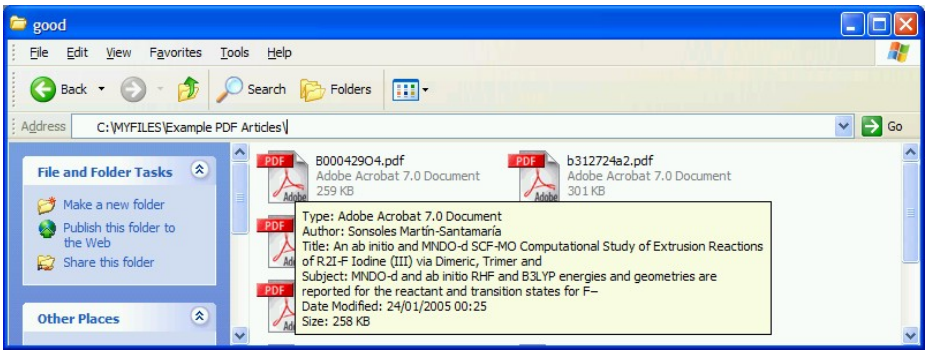


Figure 19: The mouse-over feature for Acrobat.

4.4.5 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. As its focus has been for general web-based file management, its versioning aspects have yet to be realised however. WebDAV is ideally suited for the required file upload capabilities. No bespoke software needs to be written and its support for drag and drop in particular is very useful.

4.4.6 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[109], 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 SPARQL, 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.

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.

4.5 Implementation of the Architecture

4.5.1 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 4 are covered in the standard XMP set[xii] whilst DOI and InChI required extensions to the XMP. Because no electronic articles containing XMP existed, the samples were enriched with XMP using the metadata tool provided in Adobe Acrobat Professional. Publisher-centric specific metadata such as the journal titles and the names of publishers are excluded from the XMP. Instead, the DOI is used to resolve to publishers' Web sites, effectively extending the Semantic Web model.

Of the 8 XMP enriched electronic articles which were prepared, two of them[110,111] 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 6 articles[112,113,114,115,116,117] . They were sourced using a SciFinder[118] substructure search of the molecule. ACD/ChemSketch was used to create its InChI representation. DOIs for the 8 articles were easily located on the Web.

4.5.2 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. An attempt was made to semi-automate this process with the help of the IsaViz RDF Editor. However, Acrobat Professional could not understand the RDF serialisation of the XMP that IsaViz exported. Consequently an Acrobat compliant version of XMP, an example of which is shown in Figure 20, had to be created by manual RDF editing. The XMP was then inserted into the article via the Acrobat Professional metadata tool (Figure 18). An RDF graph of the XMP used is shown in Figure 21.

[xii] The XMP "Description" property would be used in lieu of Abstract.

```
    <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. 20. Electrosynthesis of novel hypervalent iodobenzene
chlorofluoride derivatives and its application to 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 was successfully performed for
```

```
the first time and it was demonstrated that p-methoxyiodobenzene
chlorofluoride 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-
9a45efc4f438'
      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 20: Example of XMP metadata created for a journal article

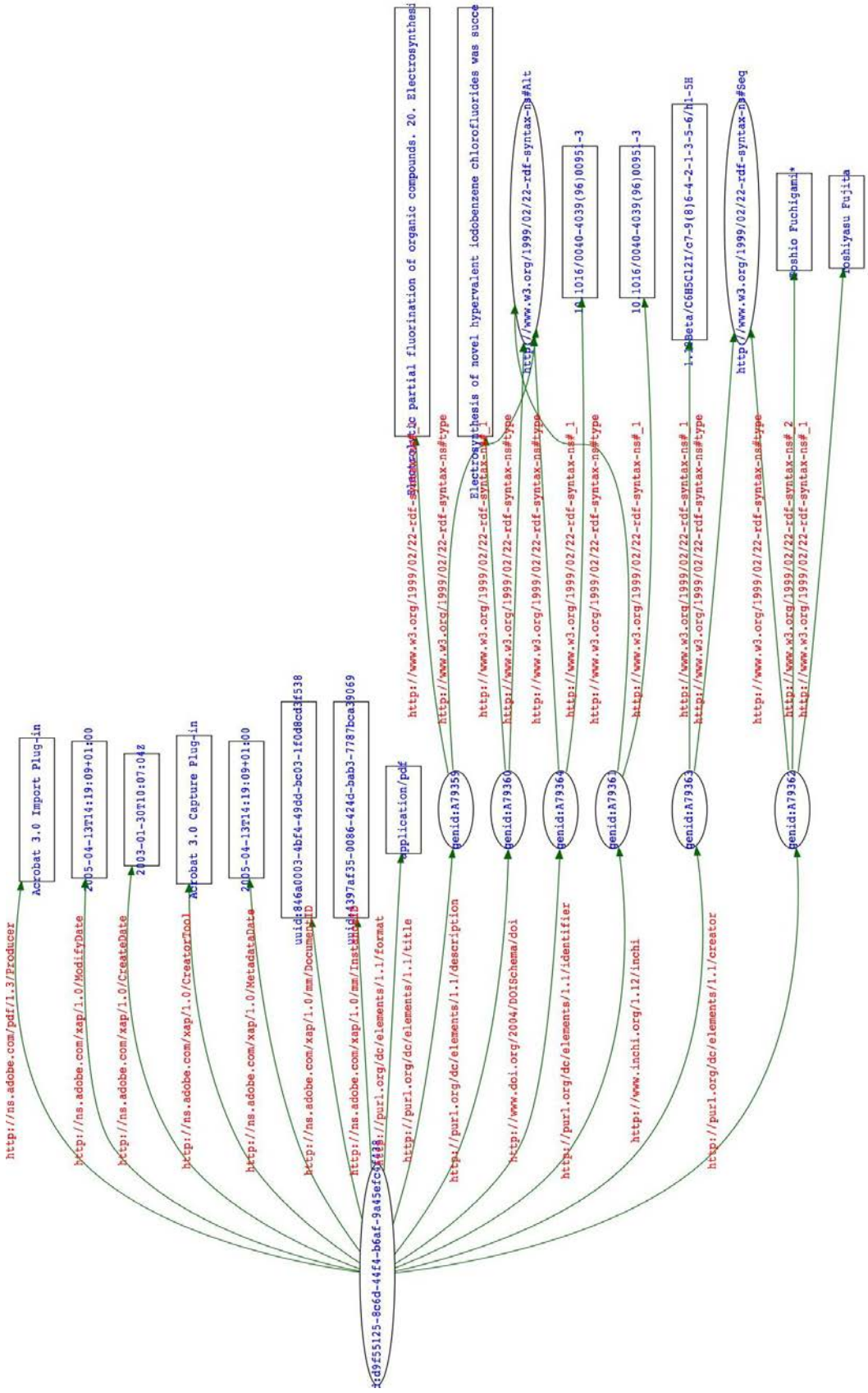


Figure 21: RDF Graph of an XMP example

4.6 Process Flow in SemanticEye

Figure 22 depicts the overall process flow for SemanticEye. This model expands on components F, G and H of the dataflow model (Figure 16). Swim lane notation is used to indicate the level in the 3 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.

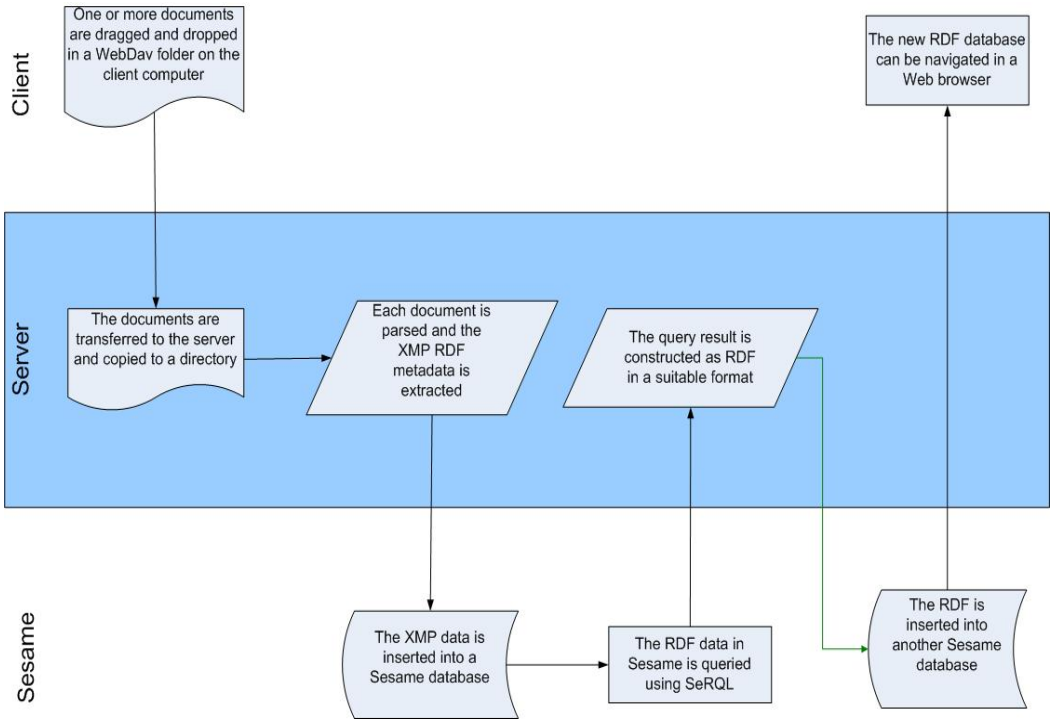


Figure 22: Process flow to transfer metadata from the articles to the RDF repository

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 21 versus Figure 23).

In keeping with the goal of the Semantic Web, it would be considered superfluous to include a complete list of citations within the SemanticEye ontology when many well established bibliographic databases fulfil this role. BIBO[119] is an example of a bibliographic ontology. In the Semantic Web metaphor, the SemanticEye ontology would simply link to BIDO via the DOI.

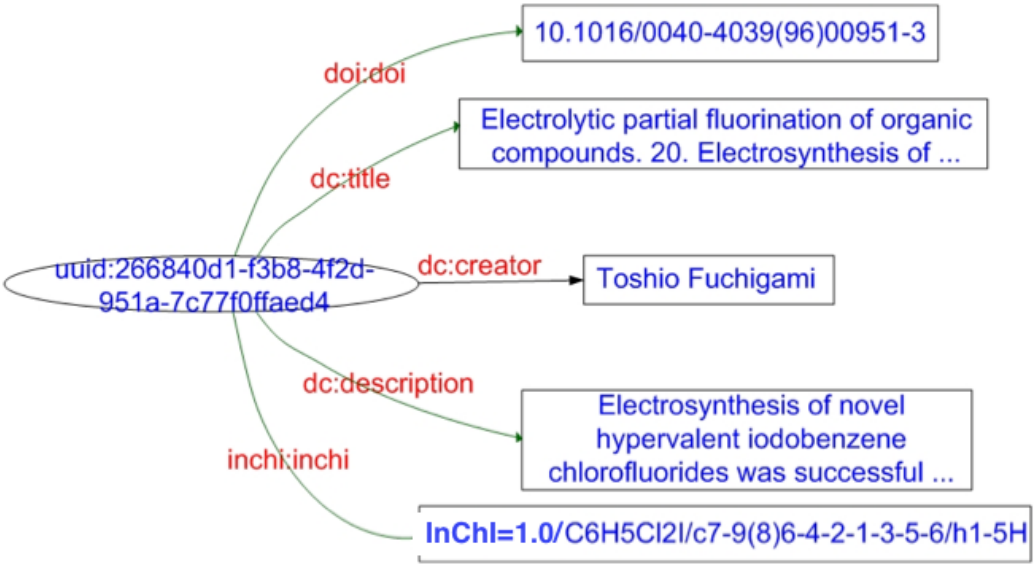


Figure 23: RDF Graph of the Chemical Journal Ontology

4.7 System Architecture

Figure 24 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 colour. All

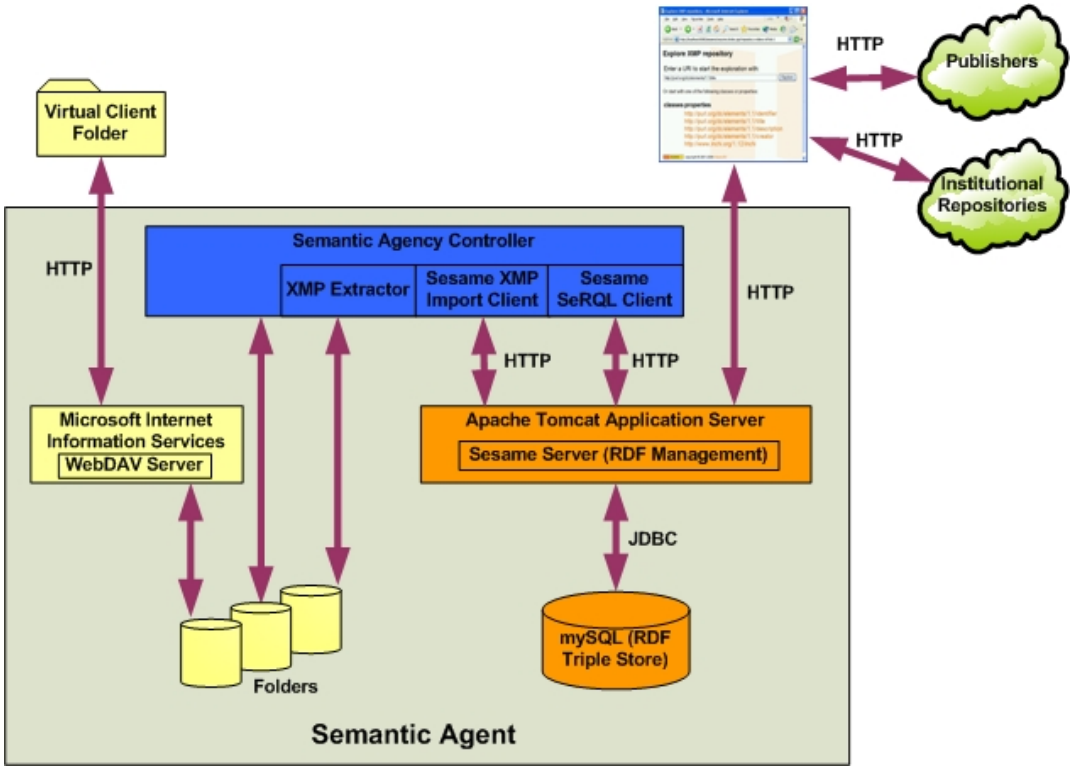


Figure 24: System Architecture of SemanticEye

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.

4.7.1 RDF Management (Figure 24, Orange)

Apache Tomcat is the preferred servlet container. It is Open Source and extensively tested with Sesame. Installation and servlet deployment was relatively straightforward. For RDF triple storage the Open Source relational database, MySQL, was installed 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 counter intuitive 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 manner that would not be confusing to the users and (importantly) without the need to customize the Sesame client.

4.7.2 XMP Service Suite (Figure 24, Blue)

The XMP Service Suite was a new development effort for this project. The Suite consists of 4 modules:

- **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 instantiate the XMP Extractor, Sesame XMP Import Client, the Sesame SPARQL Client and again the Sesame XMP Import Client in sequence.
- **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.

- **Sesame XMP Import Client:** This module, written in Java, will insert any RDF graph such as the XMP outputted by XMP Extractor into a specified RDF triple store database within Sesame.
- **Sesame SPARQL Client:** This module, written in Java, will run a specified SPARQL query on a Sesame database and output the result as RDF.

4.7.3 XMP Uploader (Figure 24, 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 25) via the "Add Network Place" Wizard.

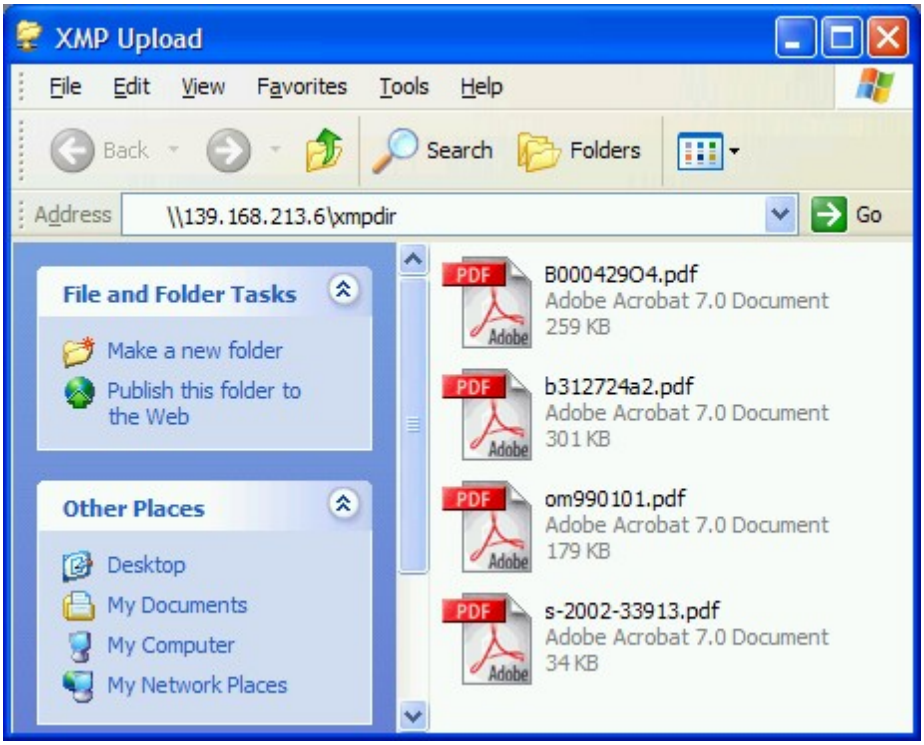


Figure 25: WebDAV folder as it appears on the Windows XP client

4.7.4 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.

4.9 Navigation of Journals

The figures in this section illustrate a typical workflow that a reader or semantic agent 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 through metadata that the RDF model (Figure 23) contains.

Figure 26 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 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 8 titles that currently exist in the SemanticEye ontology (Figure 27). 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 28). 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 29). 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 30).

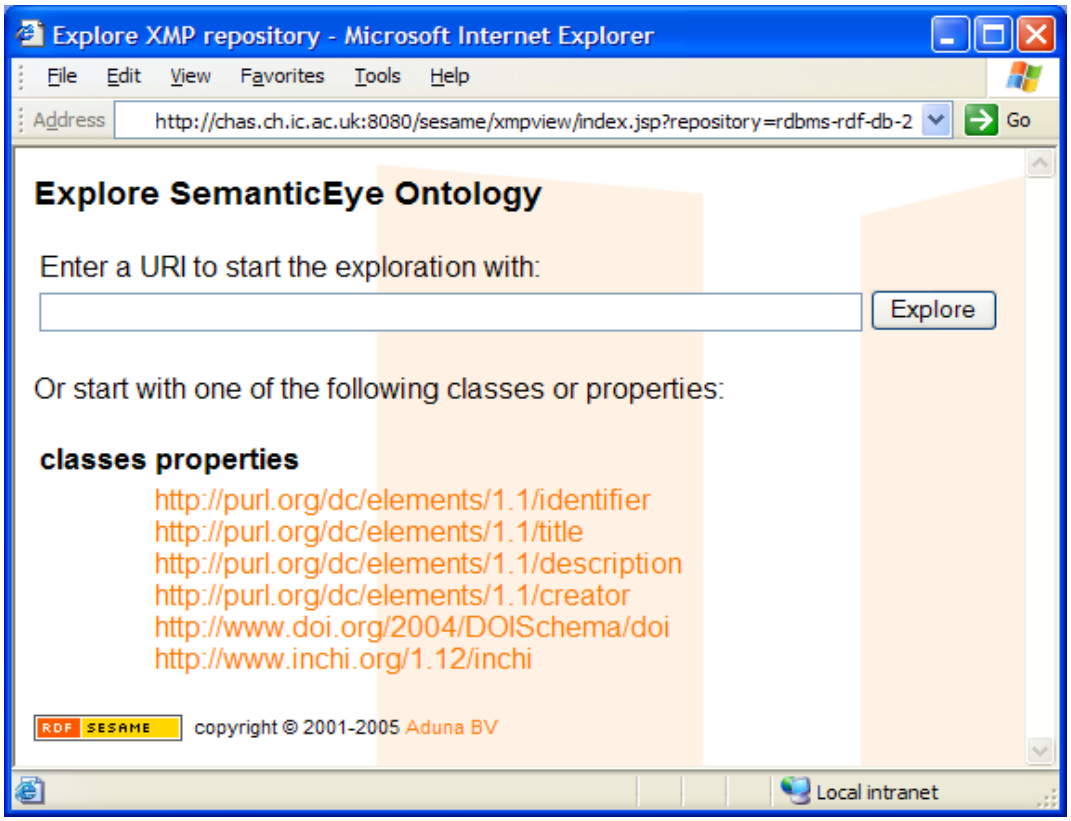


Figure 26: Entry Point of the SemanticEye Ontology

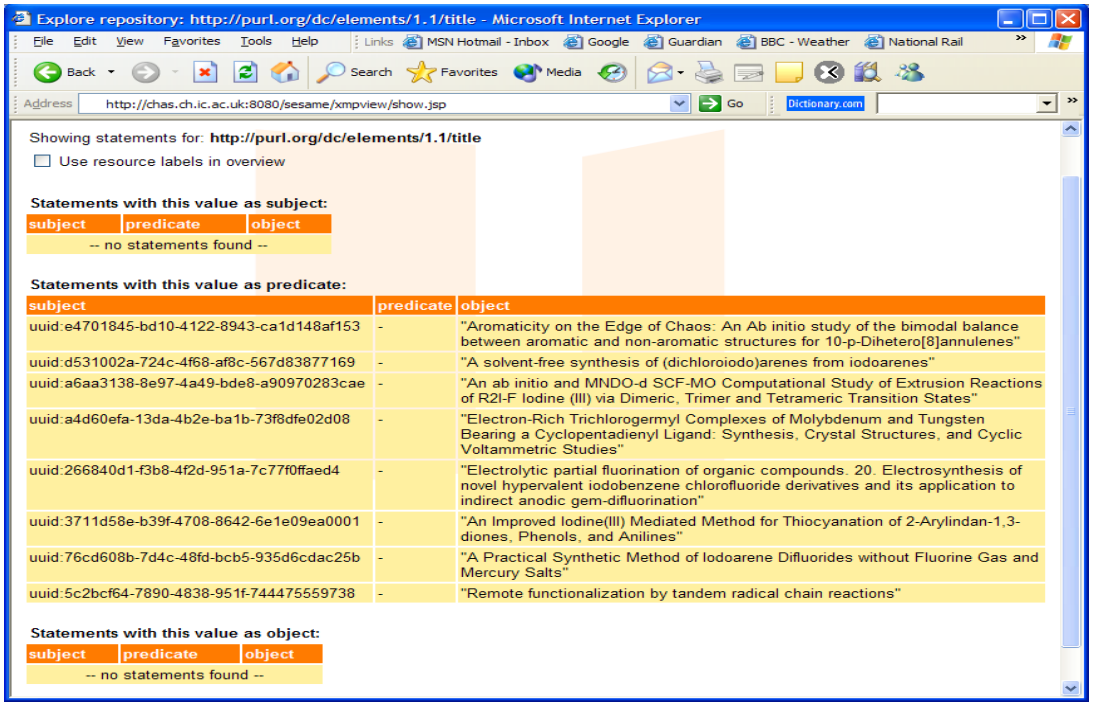


Figure 27: The Title predicate is selected to reveal the 8 titles in the Ontology

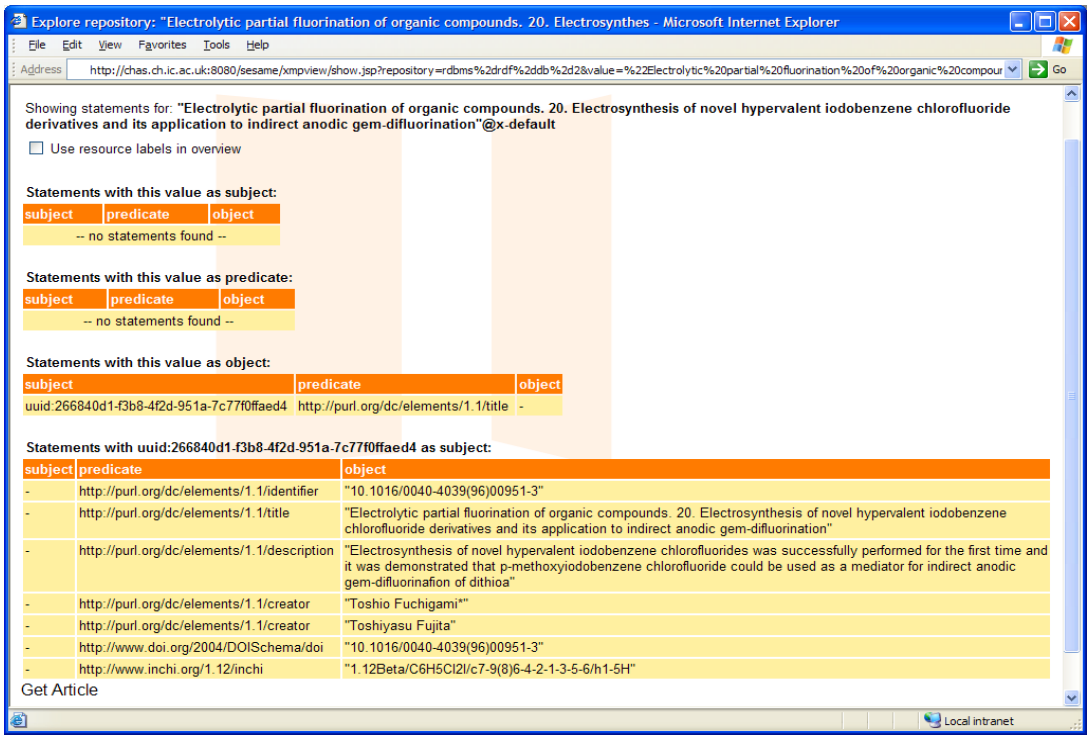


Figure 28: All of the metadata are shown for a Document selected by title

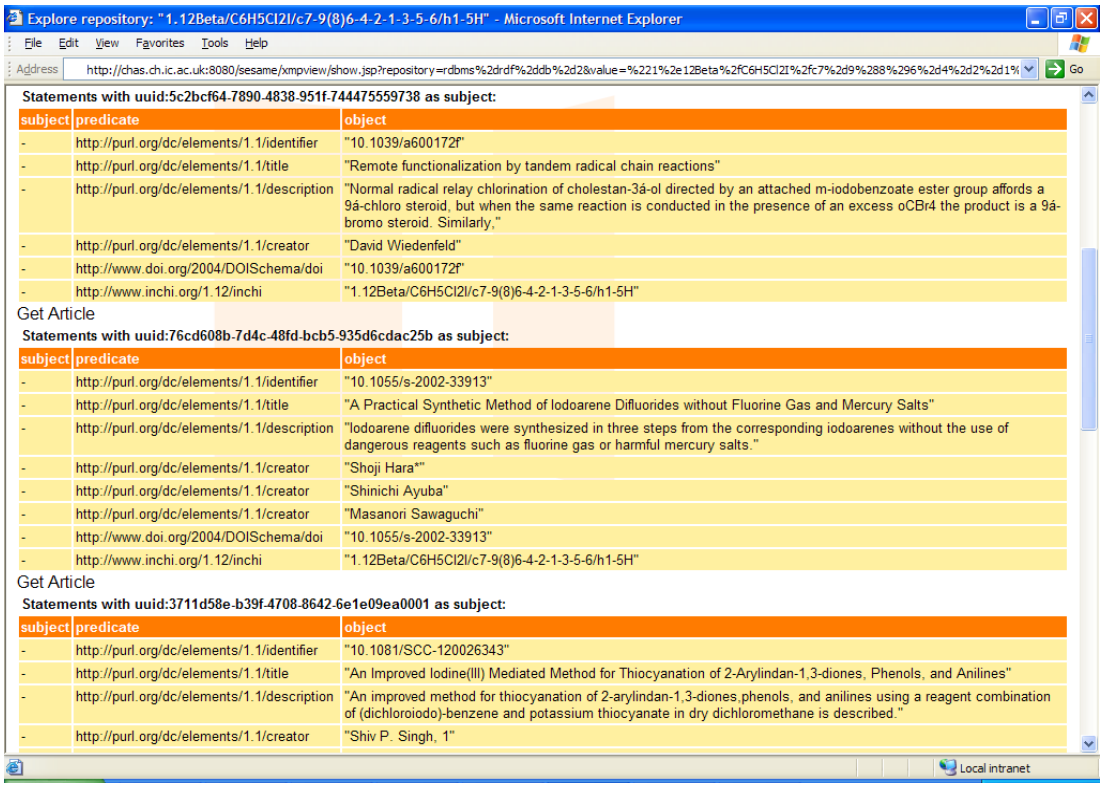


Figure 29: The InChI is selected and articles which contain it are listed



Figure 30: The DOI is resolved by selecting "Get Article"

4.10 SemanticEye Results

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:

- Adobe XMP RDF vocabulary
- Adobe XMP SDK toolkit
- InChI
- DOI
- WebDAV
- Sesame RDF repository
- SPARQL 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.

4.10.1 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 2 approaches here:

1. Embed the metadata within the content as is done with digital music files. The advantages here are the 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.

4.10.2 The SemanticEye Ontology

Because the centrally managed electronic journal metadata constitutes a molecule (InChI) centric knowledge bank, it is branded as the SemanticEye Ontology but not positioned at the Ontology level of the Berners-Lee architecture (Figure 15). In the pursuit of user uptake, it is expected to continually grow and to be continually optimised throughout its lifecycle. Fortunately, it is compact and therefore easier to maintain than a typical bio-ontology. More importantly, a small ontology is more 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 previously, the automated process of creating the SemanticEye Ontology follows these steps:

- Insert the XMP input into the XMP Triple Store.

- Run a SPARQL query to export appropriate RDF from the XMP triple store.
- 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 31) exemplifies the importance of a powerful RDF query language, such as SPARQL, to construct an optimised ontology.



Figure 31: Data flow to create the Chemical Journal Ontology

4.10.3 Validation and Trust in SemanticEye

As mentioned previously, most 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, in order to build up readership. It has been demonstrated that Semantic Web can provide the required technology framework. However, over the years publishers have mastered the editorial process of organising 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 15) means striving for the top level of the Berners-Lee architecture now.

1. Achieving trust requires, for the metadata and their relationships, a validation process which meets the following requirements:
2. Scientific publishers must mandate the inclusion of metadata within all manuscripts.
3. The capture of metadata must be automated and validated.
4. A means of uniquely identifying authors is needed.
5. The DOI must always resolve to its corresponding electronic publication.
6. 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.

4.10.3.1 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. Furthermore manual metadata entry for authors is considered disruptive and would be subjective and error prone without formal validation.

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, the manner in which Word 2003 handles metadata was investigated. It was found that the metadata does not map to Adobe XMP counterparts or maps incorrectly. Extending Adobe's XMP capture tool (Figure 18) has been ruled out as it would require the purchase one or more Adobe authoring applications. Commercial third party Microsoft Office metadata editors such as MetadataTouch[120] understand XMP. A freely available metadata editor with XMP support would be ideal. 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.

4.10.3.2 Project Prospect

Project Prospect[121] is an RSC project to enhance online journal articles. All of the RSC journals are in scope. Its goal is to make the science within RSC journal articles machine-readable through semantic enrichment. RSC editors will ultimately be annotating key data and concepts within articles, such as compounds, and linking them to Web resources such as database objects. This will transform the free text within an article to add new ways of identifying, retrieving and presenting the information within RSC publications.

One of the principal constituents of Project Prospect is the Experimental Data Checker[122] (OSCAR), a collaborative effort between University of Cambridge and the RSC. Oscar 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 to 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[123] journal where editorial control is minimal, the burden of tagging would most likely be borne by the manuscript author.

4.10.3.3 Ontology Add-in for Word 2007

The Ontology Add-in for Word 2007 is a joint effort between Microsoft and Creative Commons[124]. The functionality of this product is simple but powerful. Authors would be able to easily add scientific URLs as semantic annotations, to their manuscripts. The URLs are obtained from online ontologies and controlled vocabularies such as the National Center for Biomedical Ontology. URLs can also be extracted from Word documents that contain "tagged" terms. The widespread use of such a tool for URL creation would essentially realise the goal set out with the CLIC Hyperwave model thirteen years earlier.

4.10.3.4 Chemistry Add-in for Word 2007

Chem4Word is the product of a collaboration involving the University of Cambridge and Microsoft Research[125]. It is a Microsoft Word add-in which provides, at both the presentation and content levels, chemical semantics such as labels, formulae, and 2D structures. At the presentation level, Chem4Word creates "chemical zones," inlines of print-ready graphics of chemical information. At the content level, the chemical information is represented as CML and embedded within Microsoft DOCX documents. This enables chemical data to travel with manuscripts provided they are created in Word.

The chemical data can then be mined directly by applications which understand the CML vocabulary. DOCX, the Word 2007 constituent of the Microsoft's Open Office XML (OOXML) standard[126], is essentially a zip package containing the original text as XML markup along with images and other embedded objects stored as separate files. Chem4Word handles CML objects in this manner.

The Chem4Word project represents a good start in capturing chemical semantics within journal manuscripts, albeit in a Microsoft centric manner.

4.10.4 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 literature. An article describing its use in chemistry has been published[127]. 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. Somehow this change needs to be catalysed.

Digital signatures come in 3 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,[128] 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 include digital signatures either within the XMP or within separate Named Graphs[129]. At the same time publishing scientists need to be educated on the importance of the Digital Signature in order to help increase its uptake.

4.10.4.1 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 is not only vital for metadata management but 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.

4.10.4.2 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, such a facility will likely become available and linking SemanticEye to it should be straightforward.

4.11 Integration opportunities of SemanticEye

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. Three other Web-based frameworks were identified which potentially contain related information.

4.11.1 The Open Archives Initiative

The goal of the Open Archives Initiative (OAI)[130] is to deliver an interoperability framework for institutional repositories. Although closely related to the Open Access movement, the OAI technology and standards are much more broadly applicable than that of 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. For supporting data that is stored in institutional archives OAI-Identifiers for the data would need to be provided within the manuscript.

4.11.2 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 organisations to be drawn together into specialist “virtual organisations”. 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[131] attempts to address this aspect with the help of Semantic Web technologies. A related publication[132] presents an integration example where a DNA Grid and a Protein Grid are linked via Semantic Web repositories. SemanticEye would be in a position to exploit the Semantic Grid if the integration to the scientific resources which it exposes is straightforward.

4.11.3 The Semantic Wiki

"RDF and the Wiki-principle are a perfect match. The take-off of the semantic web is slowed down 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."[133]

Regular wiki pages have structured text, intended for humans to read and understand, and untyped URLs to other related pages. Semantic wikis enable humans to capture and store metadata about the pages along with typed hyperlinks, i.e., their relations to other pages. A set of questions is normally used to create the correct link type. Business rules can also be added to verify that the destination page is appropriate for that link type. For example a link of "scent" might only be appropriate when linking a sample with an odour type.

Semantic wikis use Semantic Web methods to address the search and information retrieval problems of Web 2.0 and its uncontrolled social tagging concept. Semantics may be either embedded within or separate from the wiki markup. For example, a Semantic wiki in medicinal chemistry could cover molecular targets within a particular disease area, such as oncology, and the compounds acting upon them. The page for one of these compounds would contain text information and semantic information, such as anti-carcinogenic properties, that a software agent would be able to disseminate. The wiki would then be able to automatically generate a list of related compounds, by listing all pages that are tagged with similar characteristics. Taking this simple notion to a more complex level, a Semantic Wiki could potentially support ontology reasoning and generate a complete ontology. Conversely, a Semantic wiki can inherit a pre-existing ontology.

Semantic wikis were first proposed in the early 2000s, and implementations emerged in the mid-2000s including the popular Semantic MediaWiki[134] (2005), Freebase[135] (2005) and OntoWiki[136] (2006). Most can export metadata as RDF. For querying, SPARQL may be directly supported. Otherwise, the exported RDF would need to be transferred to a triple store for querying. Semantic MediaWiki pages include formal notation within pages that are created by users. Other semantic wiki software derive formal notation from page titles or links. In either case, providing information through a formal notation enables agents to calculate new relations between pages. This feature would be useful for technical documentation where Semantic wikis are finding widespread use.

Project Halo[137], an early adopter of the Semantic wiki metaphor, develops "automated tutors" for students in physics, chemistry, and biology where sophisticated scientific questions posed in English are automatically answered. Another Semantic wiki exemplar, ChemSemWiki[138], illustrates how the Semantic wiki could underpin electronic publishing.

4.12 Entering the Plateau of Productivity

One of the mains aims of the research discussed in this chapter is to trigger the inclusion of XMP metadata within manuscripts by exemplifying the payback to the research community. However 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 the publishers.

Restricting SemanticEye's scope to electronic publishing restricts the focus on only a small portion of Web. However it is a high value portion with a well understood process and 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[139]. It is this confusion which explains why uptake of the Semantic Web has been slow and restricted to a few vertical domains such as bio-ontology where the payback is more apparent. Hence based on experience the following steps are proposed 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 users.

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. The solution was modelled by introducing two metadata objects, InChI and DOI, into existing Semantic Web technologies and the model was tested with a small sample set of articles. A scalable process has been established, but before moving on to the final scale up step, metadata validation needs to be implemented by the publishers. Fortunately, such a movement is underway with XMP capture being actively supported by both Elsevier (since 2007) and Nature Publishing (since 2008), and with the RSC piloting an XMP implementation.

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.

It was demonstrated how a strategic approach to metadata management for scientific electronic journal articles creates new research opportunities. The key requirement here is that scientific entities must be included in XMP metadata and be readily searchable using established search standards. SemanticEye exemplified this notion for chemical electronic publishing by the inclusion of molecular structures, uniquely defined as InChIs, within a Semantic Web metadata model. Returning to the hype cycle, once this notion achieves a critical mass of uptake by the publishers and readers, chemical publishing will have moved on to the plateau of productivity.

4.13 Initiatives that will Influence SemanticEye's Future

This section discusses three important activities which have been identified as having an influence on the future direction of SemanticEye and therefore cannot be ignored in the planning of future projects.

4.13.1 Open Access Publishing

Pressure from some quarters to make articles freely available has resulted in increasing attention being given to Open Access journals, 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[140], is slowly catching 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. For such a model to succeed, any cultural barriers will need to be well understood and addressed.

From a technical standpoint open access publishing should not have a significant impact on SemanticEye. From a cultural standpoint however, it may. The SemanticEye concept is reliant on the proactive involvement of scientific publishers for the inclusion of XMP metadata within manuscripts. As open access publishing is based on a business model which differs to that of traditional scientific publishing, a different approach for collecting XMP metadata is likely to be needed if open access publishers will not take responsibility for it.

4.13.2 Linked Data

Linked Data[141] uses the Web to create links between data from different sources. It would normally consist of heterogeneous databases within a single entity that historically have not easily interoperated, or of multiple databases maintained by multiple entities in different geographical locations. Linked Data is published on the Web with an explicitly defined meaning in order to be readable by agents and linked to and from other external data sets.

While documents on the Web are in HTML and connected by untyped hyperlinks, Linked Data rely on documents containing data in RDF format and uses the RDF to make typed statements that link arbitrary objects thereby creating what commonly referred to as a "Web of Data". Although it is still at an early stage, if the Web of Data ultimately achieves a critical mass of usage it could ultimately subsume the Semantic Web.

In 2006 Berners-Lee outlined a set of "Linked Data principles":

1. Use URIs as names for object.
2. Use the http:// scheme for URIs to enable manual lookup.
3. Useful RDF metadata should be provided with the URI to enable them to be located using SPARQL.
4. Links to other URIs should be provided with the URI to facilitate information discovery.

Hence URIs, HTTP, and RDF are the technologies which underpin Linked Data.

The Linking Open Data project[142], founded in January 2007 and supported by the W3C, identifies existing data sets that are available under open licenses, converts these into RDF according to the Berners-Lee principles, and publishes them on the Web. Participants in the early stages of the project were primarily researchers and developers in academia and startup companies. Since then the project has grown to include significant involvement from large organisations such as the BBC, Thomson Reuters, the Library of Congress, PubChem, PubMed and CAS. This growth is attributed to the open nature of the project.

A model of the linked resources within the Linked Data cloud has been published[143]. This model is akin to a hub and spoke model where certain resources have many links, the primary one being DBpedia. DBpedia is a project to extract structured information from Wikipedia. This structured information is then made available on the Web and is a principle entry point to the Linked Data Semantic Web.

On the Web, different information providers naturally publish data about the same entities and, as they may not know about each other, they introduce different URIs, or URI aliases, for the same entities. It is unrealistic for all information providers to agree on a single URI for a given entity. In order to still be able to track different resources referencing the same entity, it is common practice for information providers to set owl:sameAs links to URI aliases they know about.

Different communities have specific preferences for RDF vocabularies and Linked Data imposes no restriction on the choice of vocabularies. It is considered good practice, however, to reuse terms from well-known RDF vocabularies such as FOAF or Dublin Core wherever possible in order to make it easier for agents to process Linked Data. Only if these vocabularies do not provide the required terms should data providers define new, data source-specific terminology. New definitions should be made self-describing by the inclusion of an RDF Schema.

Linked Data can be crawled by following RDF links in a manner that is analogous to following URLs on the traditional Web. However, by working on the crawled data, search engines can provide query capabilities similar to those provided by conventional relational databases. Because the query results themselves are structured data, not just links to HTML pages, they can be processed by agents to create new Semantic Web

applications. The obvious advantage of an RDF link over a URL is that it describes the type of relationship between two data objects.

As mentioned previously much of the structured data on the Web from entities such as Amazon or Google are accessed through semantic layers that are basically Web 2.0 APIs. Compared to these APIs, Linked Data has the advantage of providing a single, standardized access mechanism instead of relying on diverse interfaces and result formats.

4.13.3 oreChem

The oreChem project[144] is a collaboration between chemistry researchers and information scientists to develop a novel infrastructure and services for information dissemination. A principal goal of the infrastructure is the sharing of data that are located in Web resources. It reuses open standards, primarily the Semantic Web and the Linked Data Effort[145], underpinned by the Open Archives Initiative-Object Reuse and Exchange (OAI-ORE)[146], thereby enabling the aggregation with non-chemistry resources. OAI-ORE has a data model and RDF syntaxes that describe the aggregations of Web resources and their relationships.

The initial work in the oreChem Project is re-purposing the core OAI-ORE data model for the chemistry domain. As with SemanticEye, this model has the chemical compound at its core and links to investigations and scientists. From this model, interfaces and APIs are designed to enable data exchange from resources. This infrastructure is being tested by adapting a number of existing chemistry resources following on from which new tools, such as chemical structure searching, will be developed. Future plans include the capture of chemistry metadata at the experimental source, and their incorporation into the infrastructure data model.

It is envisioned that an eScience offshoot called the eChemistry Web will ultimately emerge whereby molecules will link to publications that reference them, experiments that use them and the investigators that study them. It is not unlike the goals of the SemanticEye but has the advantage of a wide scale research and Microsoft sponsorship. OreChem has the notion of social networks. As will be shown in Chapter 5 SemanticEye also has this notion. However, unlike the SemanticEye project, oreChem's future goals include the development of novel analysis tools that will extract new knowledge from the eChemistry Web. The SemanticEye project precludes the

development of any new analysis tools but rather influences the adoption of the SemanticWeb by existing entities.

4.14 Summary

An exemplar was developed where chemical journal articles and supporting data from multiple publishers are linked up using a centralised metadata management approach following on from the Hyperwave experience. Given that around half of all chemical articles are sourced serendipitously, a primary deliverable here is a novel path to "knowledge discovery", achieved by the cross linking of related chemical articles via well structured terms.

A new application for this purpose was built called SemanticEye, the architecture of which adapts the successful semantic model used for digital music metadata management by popular applications such as iTunes and RealPlayer. Here, metadata for cross linking digital music are both embedded within the music files and stored in central repositories where the metadata can be updated collaboratively by the scientific community. SemanticEye repurposes this architecture for chemical publishing by pooling together a variety of standards, the core of which is the Semantic Web.

The Web and the Semantic Web differ in philosophy; the Web is designed for human navigation whereas the Semantic Web also permits the participation of software agents. The Semantic Web's metadata language is the well established Resource Description Framework, or RDF, a vocabulary consisting of triples where each triple specifies how a subject relates to an object via a predicate attribute. The object can, in turn, be the subject in another triple, thereby building up a semantic graph. SemanticEye's graph is a lightweight Ontology constructed collaboratively by editing metadata in an RDF repository called Sesame. The metadata include the attributes that are useful for locating an article. The three key attributes are the DOI, the InChI, and the author's name.

The DOI is a unique identifier pointing to any resource on the Internet. It has been a mainstay of scientific publishing for years, used for identifying scientific articles. DOIs, which permit resources to be cross-linked and discovered more easily, are maintained by a central resolving agency which routes a document's DOI to its actual location. The algorithmically generated and non human readable InChI is a unique label for molecules. The handling of the author's name, particularly by way of a unique identifier has

sociological implications which cannot be ignored. In order to ensure that SemanticEye only treats a person's identity within the context of scientific research, a study was carried out in Chapter 5.

Although SemanticEye's lightweight Ontology can be edited manually, building it up through automatic processes would be desirable and indeed achievable if each chemistry manuscript had associated metadata that was readily accessible. Embedding the metadata within the articles is analogous to digital music metadata and is robust because the metadata always travels with the assets. The technological enabler here is XMP, Adobe's RDF standard designed to be embedded in a variety of popular text documents and image formats. In this investigation, article metadata was encoded as XMP and embedded within PDF renditions. Authors would have to provide this metadata which could be obtained, for instance, from an electronic lab notebook.

As an analogy to digital music management, users would "drop" articles containing XMP into a WebDAV folder pointing to the Sesame server. The XMP metadata would then be extracted server-side from the articles, which are then deleted, and uploaded into Sesame where they are stored as part of the SemanticEye Ontology. This server-side parsing differs from that of digital music where the parsing normally occurs on the client. The Ontology can then be navigated via a Web browser for readers or via a Web Service for software agents. Documents can then be retrieved directly from the publisher by selecting the DOI. The power of this metadata driven approach is illustrated by including, within the Ontology, InChI identifiers for molecular structures and locating articles containing identical InChIs.

5. Chemical Collaboration Networks

The online searching of electronic publications and virtual networking in Web 2.0 social networks provide a novel approach for locating potential collaborations. Compared to conferences, such virtual activity is global and can represent a significant time savings in identifying centres of activity and expertise. By combining publication metadata with metadata from Web 2.0 resources, a scientist's Web profile would be broadened to include metadata about research activities, useful for locating (and being located by) new collaborators. Enabling the latter activity is one of the goals of Web 3.0, which represents the next evolutionary step in the Web cycle and is predicated on the adoption of a Semantic Web approach to content management.

A proof of concept is presented in this chapter featuring two Semantic Web tools, SemanticEye and FOAF, the "friend-of-a-friend" RDF vocabulary for social networking. A dynamic approach to generating FOAF profiles will be demonstrated whereby SemanticEye is used to output a FOAF serialisation of its ontology by querying it with SPARQL. FOAF information from other RDF sources can similarly be queried and then aggregated with the SemanticEye FOAF thereby building up a social network. Although the aim of this work is to provide insight into how chemical publishing would ultimately fit within the Web 3.0 metaphor, potential uses for FOAF vis-à-vis Web 2.0 technologies will also be investigated.

5.1 An Early Chemical Collaboratory on the Web

The Internet has transformed scientific research from the individual, who carried out research within the four walls of a laboratory, into a collective where much research is undertaken in a very distributed fashion within a collaboratory. A collaboratory is where participants use the Internet to access shared scientific data and communicate with colleagues who are normally situated remotely.

Eye2eye was an early exemplar of an Internet based chemical collaborator[5,6] and an early implementation of a Web helper application. It was basically an integration of EyeChem, and a Mosaic API called the Common Client Interface (CCI). Multiple sites which were running both Mosaic and eye2eye were able to see the same molecular visualisations and each other's changes.

As whiteboarding became the dominant Internet technology to share applications and data interactively, eye2eye never evolved much further. Nevertheless, the importance of the Internet in general and the Web in particular for a chemical collaboration network was demonstrated 1995 and revisited recently.

5.2 Initiation of New Collaborations

Currently nearly all scientific collaborations still begin face-to-face[147] with the initial contact made primarily by networking at seminars, conferences and meetings. However seminars provide only a partial coverage of all potential collaborators. Additional means would be required help to broaden the pool and the Internet is an obvious place to start.

There are instances of collaborative works being published where the participants have never met, a study on the topological properties of Möbius Annulenes[148] being a noteworthy example. This illustrates that seminars are not only means of initiating collaborations and that alternative ways using the Internet can be exploited provided that potential collaborators are locatable on the Web. A full elaboration of this has been published[24]. Irrespective of how the collaboration started, once a project is underway, remotely situated researchers normally maintain contact virtually via the Internet.

5.3 Expert Finder Systems

Finding the relevant experts in industry as well as academia is an important challenge. Locating scientific experts requires defining areas of expertise at the outset. The scientist's background must be then matched accordingly, to better enable the accurate weighting of the expertise. These activities require data sources from which expertise knowledge can be derived. Whilst publication data are obviously a suitable source, many other disparate RDF data sources also exist[149] such as curriculum vitae (CV)[150]. The particular appeal of publication data, however, is that they disclose with whom researchers collaborate.

Expert Finder systems identify candidate experts and rank them based on their estimated expertise on a given topic. Online profiles, biographies, email, and social networks are target data sources. Like a CV one's expertise changes over time and accurately assessing it is challenging. Well established expert finder systems exploit either structured data from data silos such as email programs, messages exchange forums, ontologies, and scientific publication repositories. Various disciplines have used numerous techniques for expert discovery that have been quite successful in their

respective domains. However, most domain specific techniques lack semantics, thereby limiting the outreach and the level of granularity.

5.4 Semantic enablement of Social Networking

5.4.1 Inadequacies of Web 2.0 for collaborations

The Web improves the channelling of scientific ideas whereby a good idea, irrespective of the status of its source, can easily be brought to the attention of others on the Web and a community can be formed. Such communities, which enable like-minded people to meet, often make use of Web 2.0 applications which facilitate interactive information sharing and collaboration on the Internet. In contrast to traditional websites where users are passive information consumers, a Web 2.0 site allows its users to contribute to the website's content. Web 2.0 sites currently serve as an Internet mechanism for establishing contacts between potential collaborators, and for individuals to try to establish semantic connections between ideas.

Despite the opportunities, the concept still has major shortcomings which could deter scientists. In particular, many such sites tend to be information silos. Users entering them must manually maintain a separate personal profile for each site, and most are designed to be accessed by humans *via* Web browsers. Absent from many of these sites are semantic data, metadata and interfaces which would enable the participation of software agents that could perform useful functions such as connecting users who have common interests, but who are not necessarily members of the same social network. Indeed such a notion might run counter to prevailing business models of Web 2.0 social networks whereby users along with the content they create are essentially ring fenced. One interesting exception is Plaxo Pulse[151], which implements Google's OpenSocial API[152]. Indeed, in a landscape where social networking would start to open up, established search engines such as Google and Yahoo could ultimately help with locating potential collaborators.

In the case of finding collaborators, no single social network provides a broad enough pool to be a serious alternative to locating collaborators at e.g. highly focused scientific conferences. As will be discussed later, what has become known as Web 3.0 technologies may have the potential for rectifying these shortcomings, thereby creating new opportunities for chemical discovery and innovation.

5.5 Web 3.0 and the Semantic Web

Although not strictly the Semantic Web, semantic layers in popular Web sites are being employed to introduce users to people or products of interest. A semantic layer in Amazon.com's database provides targeted marketing by tracking its customers' shopping habits and introducing them to potential products of interest. Likewise, LinkedIn, a social network geared towards business people, tracks its customers' profile and relationships in a semantic layer and introduces them to possible contacts. The next generation, colloquially known as Web 3.0 adopts the Semantic Web approach to content discovery.

The following sections present an exemplar illustrating how a Web 3.0 Chemical Collaboration Network would be built up from using SemanticEye and perhaps the most well established Semantic Web RDF vocabulary.

5.6 FOAF (Friends of a Friend)

FOAF is essentially an RDF vocabulary of terms for constructing machine readable profiles used in home pages and social networks. It allows the semantic expression of "friends", publications, public collaborations and other activities. Its architecture is decentralised: FOAFs are expected to be maintained within home pages, not in centralised databases. An example of a personal FOAF is given in Figure 32. This example makes use of what is considered a unique personal identifier in FOAF, a textual representation of applying the SHA1 mathematical functional to a 'mailto:' URI for an Internet mailbox.

```
<foaf:Person>
 <foaf:name>Omer Casher</foaf:name>
 <foaf:mbox_sha1sum>0cd8fa86a146f025e635ad43f64537ecdc7fd408
 </foaf:mbox_sha1sum>
 <foaf:workplaceHomepage rdf:resource="www.gsk.com"/>
 <foaf:homepage rdf:resource="http://www.omercasher.com/"/>
 <foaf:knows> <foaf:Person>
 <foaf:name>Henry Rzepa</foaf:name>
 <foaf:mbox_sha1sum>0d26f3beb220c20f1fced27930005e5f0d0d23c2
 </foaf:mbox_sha1sum>
 <foaf:homepage rdf:resource="http://www.ch.ic.ac.uk/rzepa/"/>
 <foaf:workplaceHomepage rdf:resource="http://www.ic.ac.uk/"/>
 </foaf:Person> </foaf:knows>
</foaf:Person>
```

Figure 32: Example of a Personal FOAF

The FOAF RDF vocabulary is expressive enough to allow for the creation of rich scientific profiles which, within a social network setting, can lead to "research serendipity" where potential collaborators are located, perhaps by chance. However, the uptake of FOAF has been slow in Chemistry presumably because maintaining personal FOAFs is time consuming and the value to scientists has not been effectively demonstrated. As with SemanticEye, the widespread adoption among chemists will only be achieved if it is driven by influential organisations, namely the publishers.

Nevertheless some grass roots adoption has been taking place, one example being the Blue Obelisk organisation[153] which maintains a group FOAF of its members (Figure 33). Better FOAF software tools, particularly those which understand chemical semantics, are needed.

```
<foaf:Group>
 <foaf:name>Blue Obelisk</foaf:name>
 <foaf:member>
 <foaf:Person>
 <foaf:name>Henry Rzepa</foaf:name>
 <foaf:homepage rdf:resource="http://www.ch.ic.ac.uk/rzepa/"/>
 <foaf:workplaceHomepage rdf:resource="http://www.ic.ac.uk/"/>
 </foaf:Person>
 </foaf:member>
<foaf:member>
 <foaf:Person>
 <foaf:name>Egon Willighagen</foaf:name>
 <foaf:homepage rdf:resource="http://chem-bla-ics.blogspot.com/"/>
 </foaf:Person>
 </foaf:member>
….
</foaf:Group>
```

Figure 33: Example of a group FOAF

A number of FOAF extensions would be of useful for helping to find experts. One such extension, FOAFRealm[154] is a user profile management system with access control and social networking features. FOAFRealm enables the sharing of personal taxonomies across a social network. A user's collection is assigned an expertise value based on the quality of the profile information and users can see each other's expertise value on given topics.

5.6.1 FOAF Aggregation

FOAF aggregation is essentially a process which combines multiple FOAFs to create broadened profiles, without the need for manual editing. FOAF aggregation is attractive for communities where devolved control is required. However, using an individual's personal information outside of the individual's control has sociological issues. FOAF aggregations therefore need to be handled appropriately. Because the goal of this investigation is primarily to add value to scientific collaborations, FOAF aggregations are restricted to the machine assisted establishment of scientific collaborations within social networks. The following sections describe how this is achieved.

5.7 SPARQL search for Scientists with a Common InChI

Using a suitable SPARQL query, SemanticEye would be able to output all of an author's collaborators as FOAF "friends" which can then be easily added to the author's personal FOAFs. Alternatively, SemanticEye can output, as a FOAF group, all of the scientists associated with a common InChI identifier. No manual editing of these outputs is needed.

FOAF aggregations could feed into any social network which understands FOAF. However, most mainstream social networks do not support FOAF, the one noteworthy exception being LiveJournal[155].

Automatically created SemanticEye FOAFs enable all scientists who actively publish to have up-to-date scientific profiles, the values of which can potentially be improved by aggregation with FOAF outputs from additional sources. A preliminary investigation of the automatic creation of FOAF has been published[156] whereby FOAFs are constructed from RDF inputs and purposed into HTML using PHP.

The following sections illustrate the creation of such an aggregation with the help of SPARQL.

5.7.1 Identifying a scientist's collaborators in FOAF

SemanticEye is queried for all co-authors associated with any articles having "rzepa" as one of its authors (Figure 34). The SPARQL CONSTRUCT feature is used to output the RDF triples serialised as FOAF.

```
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

CONSTRUCT {
<http://www.ch.ic.ac.uk/people/rzepa> foaf:knows _:v2 .
_:v2 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> foaf:Person .
_:v2 foaf:name ?creator .
}
WHERE
{
 ?document1 dc:creator ?author .
 FILTER regex(?author, "rzepa", "i") .
 ?document2 dc:creator ?creator .
 FILTER (?document2 = ?document1 && !regex(?creator, "rzepa","i"))
}
```

Figure 34: SPARQL Query on SemanticEye Triples to identify collaborators

5.7.2 Adding Scientist's Molecules of interest to the FOAF

SemanticEye is queried for the InChI metadata associated with any articles having Henry Rzepa as one of its authors (Figure 35) and a FOAF is outputted as RDF triples. The InChI namespace used here is "http://www.inchi.org/1.12/" and it is given the alias "inchi". The output is then aggregated with the FOAF containing scientific collaborators, as presented in the previous section, by a simple concatenation of the 2 RDF triple outputs. The FOAF aggregation is illustrated as XML in Figure 36.

```
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX inchi: <http://www.inchi.org/1.12/>
CONSTRUCT {
 <http://rdf.openmolecules.net/>   <http://www.inchi.org/1.12/inchi>
 ?inchi .
}
WHERE
{
 ?document1 inchi:inchi ?inchi .
 ?document1 dc:creator ?author .
 FILTER regex(?author, "rzepa", "i") .
}
```

Figure 35: SPARQL Query to identify molecules of interest

```
<foaf:Person>
 <foaf:name>Henry S. Rzepa</foaf:name>
 <foaf:knows><foaf:Person>
 <foaf:name>Dave Widdowson</foaf:name>
 </foaf:Person></foaf:knows>
 <foaf:interest>
 <rdf:Description rdf:about=" http://rdf.openmolecules.net"
rdfs:label="Molecule">
 <inchi:inchi>INChI=1/C6H12O2/c1-2-4-8-6-5-7-3-1/h1-
6H2</inchi:inchi>
 <inchi:inchi>INChI=1/C21H17ClN4O3/c1-12-23-20-21(29-
12)26(14(3)28)24-19(15-7-5-4-6-8-15)17-11-16(22)9-10-
18(17)25(20)13(2)27/h1-3H3,4-11H/b24-19-</inchi:inchi>
 </rdf:Description></foaf:interest>
 <foaf:publications>
 <foaf:Document rdf:about="http://dx.doi.org/10.1039/b000429o"
rdfs:label="Journal”>
 <dc:title>An ab initio and MNDO-d SCF-MO Computational Study of
Extrusion Reactions of R2I-F Iodine (III) via Dimeric, Trimer and
Tetrameric Transition States</dc:title>
 </foaf:Document></foaf:publications>
 </foaf:Person>
```

Figure 36: Personal FOAF of SemanticEye output

5.7.3 Grouping authors by a molecule of interest

SemanticEye is queried for all journal authors associated with an InChI molecule of interest (Figure 37) and outputs a FOAF Group of the authors (Figure 38).

```
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX inchi: <http://www.inchi.org/1.12/>

CONSTRUCT { <http://www.inchi.org/1.12/inchimolecule> rdf:type
foaf:Group .
<http://www.inchi.org/1.12/inchimolecule> foaf:member _:v1 .
_:v1 rdf:type foaf:Person .
_:v1 foaf:name ?creator .

WHERE
{
 { ?document inchi:inchi ?inchi .
```

```
 ?document dc:creator ?creator .
 FILTER regex(?inchi, "INChI=1/C6H5Cl2I/c7-9", "i") . }
}
```

Figure 37: SPARQL Query to group authors according to a molecule of interest

```
<foaf:Group>
 <foaf:name>INChI=1/C6H5Cl2I/c7-9(8)6-4-2-1-3-5-6/h1-5H
 </foaf:name>
 <foaf:member>
 <foaf:Person>
 <foaf:name>David Wiedenfeld</foaf:name>
 </foaf:Person>
 </foaf:member>
 <foaf:member>
 <foaf:Person>
 <foaf:name>Rashmi Pundeer</foaf:name>
 </foaf:Person>
 </foaf:member>
....
</foaf:Group>
```

Figure 38: Group FOAF of SemanticEye output

5.7.4 Aggregating FOAFs from Multiple Sources

This section illustrates how an individual's scientific profile can be built up automatically by aggregating SemanticEye FOAF outputs with FOAF outputs from other resources, in this case ChemSem and the Imperial College Professional Web Pages.

ChemSem[157] is a seminar notification system, which includes the CML XML vocabulary, and was built using open source software. It outputs an RDF site summary (RSS 1.0) feed which can potentially intermingle with other RSS feeds based on common metadata, and which provides information about individuals giving seminars at the institute. Another source of personal information comes from the Professional Web Page (PWP) database, with the content of this database such as publication lists populated automatically using technology provided by Symplectic[158]. Using an XML schema definition (XSD) and an XML stylesheet transform (XSLT) the database can output a personalised FOAF for each academic.

The FOAF aggregation from the three resources for Henry S. Rzepa is illustrated in Figure 39.

```
<foaf:Person>
 <foaf:name>Henry S. Rzepa</foaf:name>
 <foaf:title>Professor</foaf:title>
 <foaf:knows>
 <foaf:Person>
 <foaf:name>Dave Widdowson</foaf:name>
 </foaf:Person>
 </foaf:knows>
 <foaf:interest>
 <rdf:Description rdf:about="http://rdf.openmolecules.net/"
     rdfs:label="Molecule">
 <inchi:inchi>INChI=1/C6H12O2/c1-2-4-8-6-5-7-3-1/h1-6H2</inchi:inchi>
 <inchi:inchi>INChI=1/C21H17ClN4O3/c1-12-23-20-21(29-12)26(14(3)28)24-
     19(15-7-5-4-6-8-15)17-11-16(22)9-10-18(17)25(20)13(2)27/h1-3H3,4-
     11H/b24-19-</inchi:inchi>
 <inchi:inchi>INChI=1/C18H28/c1-2-4-6-8-10-12-14-16-18-17-15-13-11-9-7-5-
     3-1/h1-2,7,9-10,12,17-18H,3-6,8,11,13-16H2/b2-1-,9-7-,12-10-,18-17-
     </inchi:inchi>
 </rdf:Description>
 </foaf:interest>
 <foaf:publications>
 <foaf:Document rdf:about=http://dx.doi.org/10.1039/b000429o
     rdfs:label="Journal">
 <dc:title>An ab initio and MNDO-d SCF-MO Computational Study of
     Extrusion Reactions of R2I-F Iodine (III) via Dimeric, Trimer and
     Tetrameric Transition States</dc:title>
 </foaf:Document>
 </foaf:publications>
 <foaf:publications>
 <foaf:Document
     rdf:about="http://www.ch.ic.ac.uk/pss/index.php?page=view&amp;action
     =view&amp;ID=633" rdfs:label="Seminar">
 <dc:title>Organic Computational chemistry, lecture 1</dc:title>
 <dc:description>An introduction to molecular modelling, scope of
     lecture, basic definitions and scales of molecular
     models</dc:description>
 </foaf:Document>
      </foaf:publications>

     ...
```

Figure 39: Example FOAF Aggregation: SemanticEye ChemSem IC PWP

5.8 The Current Status of Online Resources

In the previous sections, the automated creation of a scientific profile was demonstrated. Such a profile is invaluable for improving a scientist's coveted visibility via a "Web presence" and, in turn, their chances of establishing new collaborations. A good Web presence can be established, at a minimum, by a well-populated and up to date list of publications together with a home page describing current research and areas of expertise. Aggregated FOAFs fit well here and can be incorporated now with minimal effort.

The Web however provides additional channels for improving visibility that are particularly advantageous for up-and-coming scientists without a well-known research portfolio and publication list. A good idea, irrespective of the status of its source can easily be brought to the attention of others on the Web and a community can be formed. Such communities, which enable like-minded people to meet, are often underpinned by Web 2.0 technologies.

5.8.1 The Relevance of Existing Web 2.0 Technologies.

There is a prevailing notion that the Semantic Web and Web 2.0 are competing technologies. It can be argued however that the two are actually complementary and should really be exploiting each other's strengths. Indeed some Web 2.0 technologies are beginning to link in to the Semantic Web by various means. For example, the Semantically Interlinked Online Communities (SIOC) project[159] provides a framework for the connection and interchange of information from Web 2.0 portals. Such communities comprise users, their postings, and the discussion forums to which they subscribe. The SIOC RDF schema describes online community constructs. By aggregating all the SIOC "topics" that are associated with a particular user's posts across multiple sites, the user's topics of interest and related expertise can be established. The following sections discuss Web 2.0 technologies and cite several relevant examples.

5.8.1.1 Blogs

Blogs (from: "Web log") are user-created Web sites, usually maintained by an individual, with regular entries of commentary, descriptions of events, or other material such as graphics or video. Entries are commonly displayed in reverse-chronological order. Many blogs provide commentary or news on a particular subject, whilst others function as more personal online diaries. A typical blog combines text, images, and links to related

resources. The open interactivity of comments is an important feature. Something of a new medium, the purpose of a blog is still being actively explored by its proponents.

Several blogs are noted here to illustrate the diversity of approaches. The totally synthetic blog[160] came to substantial prominence by highlighting a report in a noteworthy journal[161] describing the use of sodium hydride as an oxidant. Within days, a community had coalesced around discussion of the scientific issues raised in the article, and the eventual outcome was its formal withdrawal. The blog posting highlighted the problems faced with the modern refereeing processes, and is notable for the speed of events catalysed by its posting. Another blog[162] has come to the fore for its crusading approach to many important problems and issues facing scientific publishing. Significantly, it reaches many communities that conventional chemistry journal publishing does not. Other blogs address a more conventional chemistry audience. One with a wide readership[163] highlights, often with critical comment, recent interesting publications in computational chemistry; many journals also now offer a similar service for their own readership. "The only thing worse than being talked about is not being talked about".[164] It is no longer considered sufficient to simply publish a new scientific result; it must be exposed as widely as possible.

"Blogging" is thus seen as a relatively effortless way to connect with like-minded people, to improve visibility and become part of a community. Blogs are also a particularly useful way to share interesting or perhaps even constructively critical findings that would otherwise not be published. They are also very effectively and rapidly indexed by search engines such as Google.

Many exemplars of Semantic blogging have been published with one of the earliest coming out of HP Labs[165]. A relatively recent Semantic blog was created[166]. This blog makes use of Zemanta[167] a tool for creating in-text Web hyperlinks that integrates to important Linked Data resources like DBpedia[168]. It also includes an automatic tagging service[169] to detect and highlight common terms, expressing them as RDF metadata declarations within the blog itself. The detected terms themselves however contain no overtly chemical meanings, and it would be desirable to link such a service to chemical and molecular ontologies.

5.8.1.2 Wikis

A Wiki is a collaborative website where content is created and maintained by multiple authors using a simple online editor. This concept is sometimes described as write-many-read many, and is thus distinct from a blog which tends to be associated with an individual. The best known example is Wikipedia, within which the chemical content is now impressively large. Access to a Wiki can be either open or restricted. An underlying document management system supports version control and a layer of metadata including comments on the reasoning behind edits. Because a Wiki is multiple-authored, attribution in a scientific sense is difficult. It can be the case that the metadata available for any individual contributor does not allow any clear identification of their real identity or indeed credentials relevant to their contribution. This is a real issue for Wikipedia itself, and the extent to which the content contained there can be trusted. Many variations in the concept have evolved to address these various issues, including a commercial product known as Google Wave[170].

5.8.1.3 Podcasts

A podcast – the term being a merging of "iPod" and "broadcast" – is an audio or video recording which is uploaded to a "podcast feed" along with an RSS file containing metadata about the content and the author. Podcasts are created from Open Source or commercially available software. Participating Web pages would normally point to the feed location. Users subscribe to the podcast and are notified on their MP3 players whenever new content is available. In addition to iTunes, podcast search engines include Podscope and Odeo.

Podcast software normally support the two established versions of the RSS standard: RSS 2.0 (where RSS stands for "Really Simple Syndication") is an XML vocabulary where core metadata elements are defined within an XML schema. RSS 1.0 (where RSS stands for "RDF Site Summary") is based on RDF and can be processed by any RDF processor. The Dublin Core namespace is normally included. Additional XML namespaces are allowed by both versions but consideration needs to be made about how they would by handled by downstream applications. For example, if for scientific podcasts InChIs were included in the RSS, RSS 1.0 would be more appropriate as the podcasts could then be located on the Semantic Web.

A number of podcasts services for the chemistry community are provided by publishers. These include Nature's Chempod[171] and the RSC's Chemistry World Podcast[172]. They are also used to some extent by individuals to disseminate lectures and accompanying notes. The merits of podcasts for chemistry lectures (along with the deficiencies of PDF) have been also been discussed[173]. This article also discusses the value of incorporating CML into podcast RSS 2.0 metadata.

A Semantic Web RDF vocabulary for podcasts, called Podcast Pinpointer has been published[174]. This vocabulary uses as its basis RSS 1.0 along with Dublin Core and FOAF. Another RDF vocabulary which uses a modified version of Podcast has also been published[175]. The synergy between these approaches and those of SemanticEye merit further investigation.

The management of metadata for MP3 media assets is well established and can be readily applied to Podcasts. It is therefore a matter of course that a Semantic Web model for podcasts has been investigated[176].

5.8.1.4 Social Networks

Social networks consist of a community of individuals with a common interest or purpose. The most popular social networking sites, including MySpace and Facebook, have millions of users. Normally a user would log on to a social network site and interact with its user interface to access a variety of services including email, instant messaging, picture and file sharing, blogging, and voice over IP. An alternative entry point supported by some social networks such as Twitter is text messages from mobile phones. A principal feature of social networks is the ability to rapidly create user groups against any topic of interest such as a geographic location or a scientific domain.

There are literally hundreds of social networks on the Internet and joining one normally requires registration and the creation of a public profile. Other more specialised social networks are open only to members of a particular group such as a place of work or a profession. A number of social networks, such as BiomedExperts[177], target a scientific membership. Profiles generally are site specific and need to be recreated for each site. At a given social network site, users would normally connect by searching the sites' metadata for names or other common features. Connections can be established by contacting "friends" directly or by joining groups. Most sites carry out local metadata interrogation in order to introduce users with common features to each other.

Because a FOAF is essentially social network metadata, the support of FOAF by established social networks has been widely discussed and software tools exist which convert FOAF to and from formats understood by established social networks such as Facebook. Plaxo has been providing direct FOAF support for some time[178].

5.8.1.5 Mash-ups

A mashup combines data or functionality from two or more Web resources into a single service or presentation and normally targets the general public or the enterprise. Creating a mashup is usually straightforward involving open APIs and data sources used in manners that were not originally intended. Google Maps is an example of a resource with an open API that is frequently employed within mashups pertaining to locating specific entities. For example, a virtual venue for an electronic conference could be established as a mash up of a social network and a podcast feed. The notion that exposing content as RDF will automate the mash-up process has been investigated[179].

5.8.1.6 Reference Management Software

Reference management software enables scholars to record and track bibliographic citations. Once a citation has been recorded, it can be utilised in generating bibliographies for scholarly articles, books, etc. These packages normally consist of a database in which full bibliographic references can be entered. A selected list of articles can be formatted in the different formats required by publishers. Most packages provide Microsoft Word plug-ins to enable a reference list in the appropriate format to be produced automatically as an article is written. Other features include lookups of popular online libraries such as Google Scholar and PubMed. Bibliographic databases contain all article citations published in a particular discipline or group of disciplines, a prime example being the ISI Web of Knowledge[180]. Reference management tools are not bibliographic databases, but have the ability to import publication lists from the same.

Several of the more popular Reference Management packages are particularly noteworthy. Mendeley, which emerged in 2008, is a commercial attempt at providing reference management within a Web 2.0 social research environment. It essentially adapts the model of the popular social music service Last.fm where users share their listening habits and access the listening patterns of millions of other users to discover new artists or songs. Mendeley users are able to share and discover research patterns by

publishing bibliographic lists. As with all Web 2.0 social networks, the Mendeley user base is built up through user invitations.

Users drag and drop research articles into a client and the articles are automatically transferred to mendeley.com. Mendeley's algorithms then attempt to extract citation metadata from external resources, namely CrossRef, ArXiv, PubMed and Google Scholar, and from XMP metadata in the articles themselves if they exist. The articles may not yet have been published, in which case metadata may have to be added manually. The algorithms (which are not openly published) establish a scientist's expertise in a field by metadata interrogation as well by the number of times that articles are recommended.

Zotero[181] is another well established Reference Management package available as Open Source. It provides many of the metadata management features of Mendeley. It also supports XMP extraction. However, Zotero differs in concept; it does not provide Web 2.0 collaboration capabilities as it is purely client based, running as a Firefox plug in, not a Web 2.0 application.

CiteULike[182], developed at the University of Manchester in 2004, was one of the earliest Web-based social bookmarking tools geared towards scientists. It is similar in concept to del.icio.us where users tag a document URI with metadata and share the tags online. The software is proprietary but the metadata it collects is openly available. Additional features include normalising tags for those articles that have been added by multiple users and determining the number of readers of a given article. Articles are normally accessed many more times than they are actually cited. CiteULike also incorporates metadata from EndNote and BibTeX.

5.8.1.7 Google Social Graph API

The Google Social Graph API[183] exposes an individual's public Internet connections. It is expected that software agents would query this information to offer users streamlined "add friends" functionality. The API interrogates public web pages from Google indexes, such as MySpace profile pages, and publicly declared relationships, such as hyperlinks between friends' LiveJournal pages. The indexing of open standards, such as FOAF, is supported. Moreover, Google have also been trying to avoid any sociological issues associated with creating a mashup of an individual's Web data outside of the individual's immediate control.

The Social Graph API would appear to be adding desirable open social networking context to Goggle indexes.

5.8.1.8 Facebook Open Graph API

Trying to change the Web from a document centric view to a user centric view is part of Facebook's business model and underpins its Open Graph API. Open Graph will enable any Web document to include Facebook Widgets, particularly Facebook's "like" feature which will allow documents to link to favourite documents, essentially creating a Semantic Web with Facebook at its core. Although this is a step forward for both Facebook and the Semantic Web, the privacy implications of this type of feature, or indeed for the Semantic Web in general, cannot be ignored. On a positive note Open Graph will help to derive a better understanding of the privacy pitfalls in the Semantic Web.

5.9 Potential Technical Solution

By querying the SemanticEye ontology for InChIs using the SPARQL standard, communities of chemists based around common structures of interest can be automatically established. Representing such communities as FOAFs creates the possibility for expression within Internet social networks thereby making them a viable alternative to electronic conferences for locating potential scientific collaborators.

There are, however two caveats to realising this goal. The first is the fact that the uptake of the Semantic Web amongst the mainstream scientific publishers is still limited although growing. The second is that support for FOAF amongst the mainstream social networks is also limited. For SemanticEye there is a potential technical solution to these caveats that merits further investigation. This would entail the following approach:

- Extract of all of the structures from within journal articles, using software tools such as Project Prospect if necessary, and convert the structures into InChIs.
- Populate the SemanticEye ontology with the InChIs and corresponding metadata.
- Provide FOAF user groups for given structures to the Google Social Graph API by querying SemanticEye with SPARQL.

Populate social networks which support the Google Social Graph API, such as Plaxo Pulse.

5.10 Summary

Many, if not most, new collaborations between scientific researchers who have hitherto never met are still initiated at conferences, seminars and meetings during networking sessions. In the age of tightening travel budgets however, alternative means of finding collaborators represents a new opportunity. The Internet is the obvious place to start given that its value for expertise location is well documented.

Scientific electronic publishing is effectively an Internet collaboration network since most journal articles have more than one author. This network would broaden significantly if scientific article metadata from multiple publishers were joined up, ideally on the Semantic Web. Some of the popular Web 2.0 social networks, such as LinkedIn, also provide some collaboration network capabilities because they can be used to locate domain exerts by the interrogation of members within specialist groups. Social networks aimed at scientists such as BiomedExperts improve on this capability.

The repurposing of publishing metadata within the social network metaphor was investigated in order to model a chemical collaboration network. The FOAF RDF standard, which describes people and their network of “friends” and activities, was utilised. FOAFs have traditionally been authored manually and maintained within personal home pages or Web sites and not in more central RDF databases. By aggregating multiple FOAFs automatically created from SemanticEye and other resources, a social network is built up. Because the FOAF language is RDF, FOAF aggregation is readily achieved by concatenating two or more FOAF RDF graphs containing at least one common value.

Within Web 2.0 social networks, each user manually creates and maintains a personal profile. However, in contrast to the openness of FOAFs, Web 2.0 social networks provide no facility for importing FOAFs or profiles created on other social networks. This is a major drawback and consequently the investigation described in this chapter would better fit the Web 3.0 Open Social Network where FOAF would likely be a standard. Nevertheless, the ability to semantically link Web 2.0 sites has been widely investigated.

From the SemanticEye Ontology, FOAF RDF graphs can be constructed containing relationships between scientists that are established by common articles (DOIs) and common molecules (InChIs). This is achieved using SPARQL, the RDF

query standard. The values of a SPARQL query can either be represented as a simple list or as an RDF graph. By applying an appropriate SPARQL query on the SemanticEye Ontology, a FOAF RDF graph can be automatically created. This mechanism would enable all scientists who actively publish to have up-to-date scientific profiles without any manual intervention.

A SPARQL query against a single person in SemanticEye produces a FOAF scientific profile containing collaborators and molecules of interest. A different SPARQL query against a single InChI molecule in SemanticEye produces a group FOAF listing all the scientists that have an interest in that particular molecule. Since a molecule can be associated with one or more articles, this would be a useful means for a chemist to locate potential collaborators who are interested in a common molecule. The broadening of the FOAF scientific profile is achieved by aggregating it to FOAF outputs from other scientific sources. This chapter presented an experimental FOAF aggregation from 3 sources: SemanticEye, ChemSem, and IC PWP.

The intersection between scientific electronic publishing and social networking can be considered an early Web 3.0 exemplar where information resources are integrated using an established Semantic Web standard, FOAF. Although Web 3.0 is still conceptual, its potential as a medium for disseminating scientific publishing in a novel manner that leads to scientific collaborations has been demonstrated. This model exemplifies the critical role that the Semantic Web plays in completing the evolution of scientific electronic publishing.

6. Conclusion

Sandweiss predicts that *"…the future of scientific publishing will involve major innovations in electronic aids to read the literature and in new journals with new goals that wrestle with the flood of information"*[184].

This paints a desirable picture that can slowly start to be realised as scientific publishing progresses along the "Plateau of Productivity" on the hype curve. Scientific publishers however must first approach the Semantic Web strategically and foster compliance from authors. In addition, Semantic Web frameworks such as Linked Data must have achieved a critical mass of subscribers. Once these fundamentals are firmly in place, the evolution of scientific publishing will have completed, and Sandweiss' prediction is likely to be fully realised. The question is: When?

6.1 The 2022 Prediction

Tim Berners-Lee made his first World Wide Web proposal in 1989[185] and its public release is widely accepted to have occurred in 1993 with the release of the Mosaic[186] browser. After seventeen years, scientific publishers may view the evolution of journals on the Web as complete or nearing completion. From the perspective of the reader however, there is a long way to go. It is predicted in this thesis that the chemical publishing evolution will actually complete some time from now, around 2022. Therefore, the evolution of electronic publishing is only three fifths of the way there.

The 2022 prediction is based on two projections. The first projection is the planned ratification of HTML5 by the W3C in 2022. Undoubtedly, HTML5 will alter the presentation of the chemical journal in interesting and novel ways, one of which will be a presentation of text and supporting data integrated seamlessly. HTML5 will also be the interface to the future collaboratory. Science will routinely be discussed in Web 3.0 social networks and publications will be created collaboratively in Semantic Wikis. A more speculative idea is that Semantic Wikis will be able to import semantic data directly from second generation electronic lab notebooks. A de facto 3D graphics format based on WebGL will also be mature and used by molecular viewers.

The second projection casts scientific publishing into the "disruptive technology" metaphor: As one technology achieves a critical mass of users, another disruptive technology is just around the corner waiting to dislodge it. This assertion has governed

most every industry and has been rigorously analysed in Clayton Christensen's business bestseller "The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail."[187]. This book uses a number of high profile examples, primarily from the computer industry, to underline a common pattern. Large organisations often fail to adopt nascent disruptive technologies because these do not fit within customer requirements, business models, or corporate cultures. Consequently, the early adopters of disruptive technologies tend to be younger leaner outfits looking to create new markets for their innovations. As the uptake of disruptive technologies grow and start to develop mass market appeal, these early adopters grow as well and start becoming a threat to the larger established companies who then have to adapt in order to survive. Sometimes it is too late and the established players diminish in stature or even fail.

Christensen's model may be applicable to the scientific publishing industry and the impact that the Web has had on it vis-à-vis open access journals. However, the publishing industry is relatively conservative and change happens at a much slower rate compared to that of the industries analysed in his book. Nevertheless, a recent investigation[188] applied the mathematical model derived by Christensen to the growth rate of open access journals and made a provocative conclusion. Open access journals will account for 50% of all academic journals between 2018 and 2019 and 90% by 2022. These percentages are debatable, as is the form that the open access journal will ultimately assume in 2022. However, it is not unrealistic to predict that traditional and open access publishing will be in a steady state by 2022 to coincide with the evolution of scientific publishing from paper to paperless reaching completion.

6.2 Future SemanticEye Work

SemanticEye was originally conceived and developed as a reaction to the "Trough of Disillusionment" in the hype cycle, and scientific publishing will always be the principle driver for its continued existence. As part of future research, SemanticEye will be applied to molecular supporting data as discussed in Chapter 3, and the problem space will be broadened to incorporate scientific image data in order to fully align with professional activities and interests.

7. Bibliography

[1] Casher, O.; O'Hagan, D.; Rosenkranz, C. A.; Rzepa, H. S.; Zaidi N. A. Electronic Effects In π-Facially Stereoselective Epoxidation Of Phenyltrifluoromethylpropenol. *J. Chem. Soc., Chem. Commun.* **1993,** 1337-1340.

[2] Casher, O.; Rzepa, H. S.; Green, S. M. EyeChem 1.0: A Modular Chemistry Toolkit for Collaborative Molecular Visualisation. *J. Mol. Graphics* **1994,** *12*, 226-230.

[3] Casher, O.; Chandramohan, G.; Hargreaves, M.; Leach, C.; Murray-Rust, P.; Sayle, R.; Rzepa, H. S.; Whitaker, B. J.; Hyperactive Molecules and the World-Wide-Web Information System. *J. Chem. Soc., Perkin Trans 2.* **1995,** 7-11.

[4] Rzepa, H. S.; Leach, C.; Casher, O. Solvation Difference Maps as Probes of Intramolecular Hydrogen Bonding: An Application of Hyperactive Molecules. *Proceedings of the First Electronic Computational Chemistry Conference (ECTOC)* [CD-ROM], **1995**, Paper 2.

[5] Casher, O.; Rzepa, H. S. A Chemical Collaboratory using Explorer EyeChem and the Common Client Interface. **1995,** *Computer Graphics*, 29, 52-54.

[6] Chemical Collaboratories using World-Wide Web Servers and EyeChem Based Viewers, Casher, O.; Rzepa, H. S. *J. Mol. Graphics* **1995,** *13* (5) 268-270.

[7] Rzepa, H. S.; Casher, O. *ACS Division of Computational Chemistry (COMP)*, Proceedings of 210th ACS National Meeting, Chicago, IL, **1995**.

[8] Casher, O.; Rzepa, H. S. The Molecular Object Toolkit: A New Generation of VRML Visualisation tools for use in Electronic Journals. 14th Eurographics Conference, Imperial College, London, UK, **1996**.

[9] Casher, O.; Rzepa, H. S.; Suñer, G. A. Studying Perturbation Theory with Explorer EyeChem and VRML. *Electronic Conference on Trends in Organic Chemistry (ECTOC)* [CDROM], **1996**, Paper 70.

[10] Casher, O.; Leach, C.; Page, C. S.; Rzepa. H. S. Advanced VRML based chemistry applications: A 3D molecular hyperglossary. *J. Mol. Struct. (THEOCHEM)* **1996,** *368*, 49-55.

[11] Krieger, J. H. Doing Chemistry in a Virtual World. *Chem. Eng. News*, **1996**, *74* (50), 35–41. DOI: 10.1021/cen-v074n050.p035.

[12] Rzepa, H. S.; Casher, O.; Whitaker, B. J. A Paradigm Shift in Chemistry Electronic Publishing. Proceedings of the International Chemical Information Conference, Nimes, Fr, **1996**, 141-148.

[13] Casher, O.; Rzepa, H. S. HyperWave: An Electronic Conference Manager. *Proceedings of Electronic Conference on Heterocyclic Chemistry (ECHET)*, **1996**.

[14] 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. Information Networking* **1996**, *1*, 61-70.

[15] Rzepa, H. S.; Whitaker, B. J.; Goodman, J. M.; Casher, O.; Riddick, D.; Griffiths, J.; James, D.; Hildyard, C. *ACS Division of Computational Chemistry (COMP)*, Proceedings of 214th ACS National Meeting, Las Vegas, NV, **1997**, COMP-045.

[16] Casher, O.; Page, C.; Rzepa, H. S. Virtual Reality Modelling Language (VRML) in Chemistry, *Chemistry in Britain*, **1998**, 34, 26-30.

[17] Güner, O. F.; Casher, O.; Shah, A. V.; Hempill, C. *ACS Division of Chemical Information (CINF)*, Proceedings of 218th ACS National Meeting, New Orleans, LA, **1999**.

[18] Güner, O. F.; Casher, O. Role of the Internet in chemoinformatics: Recent developments. *Current opinion in drug discovery & development*, **1999**, 2, 204-208.

[19] Casher, O.; Rzepa, H. S. SemanticEye: A Semantic Web Application to Rationalize and Enhance Chemical Electronic Publishing. *J. Chem. Inf. Model.* **2006**, *46*, 2396 -2411. DOI: 10.1021/ci060139e.

[20] PressPac: Nov. 22, 2006. http://portal.acs.org/portal/PublicWebSite/pressroom/presspacs/CTP_004285 (accessed 9 November, 2010).

[21] 'iTunes approach' could aid scientific discovery – tech – 23 November 2006 – New Scientist. http://www.newscientist.com/article/dn10651-itunes-approach-could-aid-scientific-discovery.html (accessed 9 November, 2010).

[22] Casher, O.; Rzepa, H. S. SemanticEye. *ACS Division of Chemical Information (CINF)*, Proceedings of 232nd ACS National Meeting, San Francisco, CA, **2006**.

[23] Casher, O.; Rzepa, H. S. Using SemanticEye and FOAF to add value to the scientific collaboration process. *ACS Division of Chemical Information (CINF)*, Proceedings of 235th ACS National Meeting, New Orleans, LA, **2008**.

[24] Casher, O.; Rzepa, H. S. Using Semantically-enabled Components for Social-Web based Scientific Collaborations. *Enhancing Learning with Online Resources, Social Networking and Digital Libraries* [Book] ACS. **2010**.

[25] Brief history of scholarly journals. http://panizzi.shef.ac.uk/elecdiss/edl0001/ch0200.html (accessed 10 June, 2010).

[26] Rzepa, H. S.; Whitaker, B. J.; Winter, M. J. Chemical Applications of the World-Wide-Web. *J. Chem. Soc., Chem. Commun.* **1994,** 1907-1910.

[27] Hitchcock, S. Perspectives in electronic publishing: Experiments with a new electronic journal model. Multimedia Research Group, Department of Electronics and Computer Science, University of Southampton [PhD Thesis] **2002**.

[28] Schmaranz, K. Professional Electronic Publishing in Hyper-G: The Next Generation Publishing Solution on the Web. *J.UCS.* **1996,** *2* (9), 650-658.

[29] Vesse, R.; Hall, W.; Carr, L. Preserving Linked Data on the Semantic Web by the application of Link Integrity techniques from Hypermedia. *WWW2010* Raleigh, NC, **2010**.

[30] Ashman, H. Electronic document addressing: dealing with change. *ACM Comput. Surv.* **2000,** *32* (3) 201-212.

[31] Fountain, A. M.; Hall, W.; Heath, I.; Davis, H. C. Microcosm: an open model for hypermedia with dynamic linking. Hypertext: concepts, systems and applications. [Book] Cambridge University Press. 298-311. New York, NY, **1992**.

[32] 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 28 Jan, 2006).

[33] Bachrach, S.M. Chemistry publication – making the revolution. *J. Cheminform.* **2009,** *1* (2).

[34] Vision, T.J. Open Data and the Social Contract of Scientific Publishing. *BioScience* **2010,** *60* (5).

[35] Dryad. http://datadryad.org/ (accessed 6 July, 2010).

[36] Gkoutos, G.; 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. Comp. Sci.* **2001,** *41* (5), 1124–1130. DOI: 10.1021/ci000406v.

[37] Coles, S. J.; Day, N. E.; Murray-Rust, P.; Rzepa, H. S.; Zhang, Y. Enhancement of the Chemical Semantic Web through INChIfication. *Org. Biomol. Chem.* **2005,** *3*, 1832-1834. DOI: 10.1039/b502828k

[38] Taylor, K. R.; Gledhill, R. J.; Essex, J. W.; Frey, J. G.; Harris, S. W.; De Roure, D. C. Bringing Chemical Data onto the Semantic Web. *J. Chem. Inf. Model.* **2006,** *46*, 939-952. DOI: 10.1021/ci050378m.

[39] chem-bla-ics: RDF-ing Molecular Space. http://www.chem-bla-ics.blogspot.com/2007/07/rdf-ing-molecular-space.html (accessed 1 June, 2010).

[40] Berners-Lee, T.; Hendler, J. Publishing on the Semantic Web. *Nature* **2001,** *410*, 1023-1024.

[41] Frey, J., et al. CombeChem: A Case Study in Provenance and Annotation using the Semantic Web. *International Provenance and Annotation Workshop (IPAW'06)* Springer. Chicago, Il, **2006**.

[42] Lagoze, C. The oreChem Project: Integrating Chemistry Scholarship with the Semantic Web. *WebSci'09* Athens, Greece, **2009**.

[43] FOAF Vocabulary Specification. http://xmlns.com/foaf/spec/ (accessed 1 June, 2010).

[44] Aleman-Meza, B.; Bojars, U.; Boley, H.; Breslin, J. G.; Mochol, M.; Nixon, L. J.; Polleres, A.; Zhdanova, A. V. Combining RDF vocabularies for expert finding. *Proceedings of the 4th European Semantic Web Conference (ESWC2007)* Springer. 235-250. Innsbruck, Austria, **2007**.

[45] Li, N.; Zhu, L.; Mitra, P.; Mueller, K; Poweleit, E.; Giles, C. L. oreChem ChemXSeer: a semantic digital library for chemistry. *Proceedings of the 10th annual joint conference on Digital libraries* ACM. 245-254. Gold Coast, Queensland, Australia, **2010**.

[46] Bradley, J.C. Open Notebook Science Using Blogs and Wikis. **2007,** http://www.precedings.nature.com/documents/39/version/1 (accessed 1 June, 2010) DOI:10.1038/npre.2007.39.1.

[47] Chemical blogspace - Chemical Compounds. http://cb.openmolecules.net/inchis.php (accessed 4 June, 2010).

[48] Williams, A.J. Internet-based tools for communication and collaboration in chemistry. *Drug Discov. Today* **2008,** *13*, 502–506.

[49] O'Reilly, T. What Is Web 2.0. Design Patterns and Business Models for the Next Generation of Software. **2005,** http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html (accessed 3 April, 2010).

[50] Kolbitsch, J.; Maurer, H.; The growing Importance of e-communities on the web. *Web Information Systems and Technologies.* Springer **2007,** *1* (1), 19-37.

[51] HTML5. http://dev.w3.org/html5/spec/ (accessed 1 July, 2010).

[52] Dublin Core Metadata Initiative. http://dublincore.org/ (accessed 2 June, 2010).

[53] Gkoutos, G. V.; Rzepa, H. S. Use of Metadata in Chemical Content Relevancy Ranking. *Electronic Conference on Synthesis in Organic Chemistry (ECSOC)* [CDROM] **1999**.

[54] Maurer, H. HyperWave: The Next-Generation Web Solution. [Book] Addison-Wesley, Harlow, UK, **1996**.

[55] Hyperwave. http://www.hyperwave.com/ (accessed 1 June, 2010).

[56] Veiga, L.; Ferreira, P. Repweb: replicated web with referential integrity. *SAC '03: Proceedings of the 2003 ACM symposium on Applied computing* ACM. 1206-1211, New York, NY, **2003**.

[57] Phelps, T. A.; Wilensky, R. Robust hyperlinks: Cheap, everywhere, now. *Digital Documents: Systems and Principles*. [Book] Springer, 514-549, **2004**.

[58] Harrison, T. L.; Nelson, M. L. Just-in-time recovery of missing web pages. *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia* ACM. 145-156. New York, NY, **2006**.

[59] Rzepa, H. S.; Goodman, J. M.; Leach, C. (Editors). *Proceedings of the First Electronic Conference on Trends in Organic Chemistry (ECTOC-1)* (ISBN 0 85404 899 5) [CDROM **]** RSC. **1996**.

[60] RasMol and OpenRasMol. http://rasmol.org/ (accessed 4 June, 2010).

[61] Symyx: Downloads. http://www.symyx.com/downloads/index.jsp (accessed 5 July, 2010).

[62] Extensible Markup Language (XML). http://www.w3c.org/XML/ (accessed 4 June, 2010).

[63] Stokes, A.; Matsuda, H.; Hashimoto, A. Making high-level queries on diverse genome data: A structured genome document database system based on GXML and GQL. *Genome Inf. Ser*. **1999,** *10*, 176-185.

[64] Linge, J. P.; Nilges, M.; Ehrlich, L. StarDOM: From STAR Format to XML. J. *Biomol NMR*. **1999,** *15* (2), 169-172.

[65] Murray-Rust, P.; Rzepa, H. S. Chemical markup, XML, and the worldwide web. 1. Basic Principles. *J. Chem. Inf. Comput. Sci.* **1999,** *39* (6), 928- 942.

[66] Armbrust M.; Fox, A.; Griffith, R.; Joseph, A.; Katz, R.; Konwinski, A.; Lee, G.; Patterson, D.; Rabkin, A.; Stoica, I.; Zaharia, M. Above the Clouds: A Berkeley View of Cloud computing. [Technical Report] No. UCB/EECS-2009-28. **2009,** University of California at Berkley, CA.

[67] Weller, A. C. Can editorial peer review survive in a digital environment? *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 16 Jan, 2006).

[68] 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.

[69] DSpace Federation. http://www.dspace.org/ (accessed 5 Feb, 2006).

[70] Campbell E. G.; Clarridge B. R.; Gokhale M.; Birenbaum, L.; Hilgartner, S.; Holtzman, N. A.; Blumenthal D. Data withholding in academic genetics: Evidence from a national survey. *JAMA.* **2002,** *287*, 473–480.

[71] Randerson, J. Climate researchers 'secrecy' criticised – but MPs say science remains intact. The Guardian (London) [Newspaper article] **2010,** 31 March. http://www.guardian.co.uk/environment/2010/mar/31/climate-mails-inquiry-jones-cleared (accessed 27 July, 2010).

[72] Jmol: an open-source Java viewer for chemical structures in 3D. http://jmol.sourceforge.net/ (accessed 8 July, 2010).

[73] Numerical Algorithms Group – IRIS Explorer. http://www.nag.co.uk/Welcome_IEC.asp (accessed 10 July, 2010).

[74] Downing, J.; Murray-Rust, P.; Tonge, A. P.; Morgan, P.; Rzepa, H. S.; Cotterill, F.; Day, N.; Harvey, M. J. SPECTRa: The Deposition and Validation of Primary Chemistry Research Data in Digital Repositories. *J. Chem. Inf. Model.* **2008,** *48* (8), 1571-1581.

[75] Mason, H. J. Integrating Institutional Repositories into the Semantic Web. Faculty of Engineering, Science, and Mathematics, School of Electronics and Computer Science, University of Southampton [PhD Thesis] **2008**.

[76] Couch, P.A.; Sherwood, P.; Sufi, S.; Todorov, I. T.; Allan, R. J.; Knowles, P. J.; Bruin, R. P.; Dove, M. T.; Murray-Rust, P. M. Towards Data Integration for Computational Chemistry. *Proceedings of the UK e-Science All Hands Meeting* (ISBN 1-904425-53-4) **2005**.

[77] ChemSpider - Database of Chemical Structures and Property Predictions. http://www.chemspider.com/ (accessed 18 December, 2009).

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

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

[80] Linden, A.; J. Fenn, J. Understanding Gartner's Hype Cycles. Gartner **2003**.

[81] Smith, J. W. T. The Deconstructed Journal - a New Model for Academic Publishing. *Learned Publishing*. **1999,** *12*, 79-91.

[82] King, D. W.; Tenopir, C. 25 Year trends in information seeking and reading patterns of chemists. *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 16 Jan, 2006).

[83] Tenopir, C.; King, D.W. Electronic journals and changes in scholarly article seeking and reading patterns. *D-Lib Magazine* **2008,** *14*. http://www.dlib.org/dlib/november08/tenopir/11tenopir.html (accessed 18 July, 2010).

[84] Ratcliffe, S. The Concise Oxford Dictionary of Quotations. [1963] Oxford University Press, USA **2006**.

[85] Hull, D.; Pettifer, S. R.; Kell, D. B. Defrosting the Digital Library: Bibliographic Tools for the Next Generation Web. *PLoS Comp. Biol.* **2008,** *4* (10).

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

[87] Howison, J.; Goodrum, A. Why can't I manage academic papers like MP3s? The evolution and intent of Metadata standards. *Proceedings of Colleges, Code and Copyright, ACRL Publications in Librarianship no 57*, College Park, MD, **2004**; ACRL: Chicago, 2005.

[88] Berners-Lee, T.; Hendler, J.; Lassila, O. The Semantic Web. *Sci. Am.* **2001,** *5*, 34-43.

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

[90] 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 27 Feb, 2006).

[91] Semantic Search Survey. http://swuiwiki.webscience.org/index.php/Semantic_Search_Survey (accessed 9 May, 2010).

[92] SearchMonkey. http://developer.yahoo.com/searchmonkey/ (accessed 9 May, 2010).

[93] Google search gets semantic. http://www.google.com/hostednews/afp/article/ALeqM5iYRn_3Frd4j_CCTp0e7t4eXLWAag (accessed 9 May, 2010).

[94] Bing. http://www.bing.com/ (accessed 9 May, 2010).

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

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

[97] Coles, S. J.; Day, N. E.; Murray-Rust, P.; Rzepa, H. S.; Zhang, Y. Enhancement of the Chemical Semantic Web through INChIfication. Org. Biomol. Chem. 2005, 3, 1832-1834.

[98] International Union of Pure and Applied Chemistry. http://www.iupac.org/inchi/release102.html (accessed 20 July, 2010).

[99] Adobe XMP: Adding Intelligence to Media. http://www.adobe.com/products/xmp/ (accessed 2 May, 2010).

[100] Scalable Vector Graphics (SVG). http://www.w3.org/Graphics/SVG/ (accessed 18 December, 2009).

[101] PRISM Specificatoins. http://www.prismstandard.org/specifications. (accessed 19 Feb, 2006).

[102] XMP Specification Part 1. http://www.adobe.com/content/dam/Adobe/en/devnet/xmp/pdfs/XMPSpecificationPart1.pdf (accessed 1 October, 2010).

[103] CrossTech: Metadata in PDF: 2. Use Cases. http://www.crossref.org/CrossTech/2007/08/metadata_in_pdf_2_use_cases.html (accessed 10 June, 2010).

[104] Nascent: XMP Labelling for Nature. http://blogs.nature.com/wp/nascent/2008/12/xmp_labelling_for_nature.html (accessed 10 June, 2010).

[105] Zotero. http://www.zotero.org/ (accessed 10 June, 2010).

[106] Mendeley v0.9.6.1 released | Mendeley Blog. http://www.mendeley.com/blog/progress-update/mendeley-v0-9-6-1-released/ (accessed 10 June, 2010).

[107] CrossRef Labs: PDFMark. http://labs.crossref.org/pdfmark/pdfmark.html (accessed 1 July, 2010).

[108] 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.

[109] Broekstra, J.; Kampman, A.; van Harmelen, F. Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. *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,** 54-68.

[110] 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.* **2000**, *2,* 2158–2161. DOI:10.1039/b000429o.

[111] 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[8annulenes. *Phys. Chem. Chem. Phys*. **2004,** *6,* 310-313. DOI:10.1039/b312724a.

[112] Wiedenfeld, D. Remote functionalization by tandem radical chain reactions. *J. Chem. Soc., Perkin Trans. 1*. **1997,** *3,* 339-347. DOI:10.1039/a600172f.

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

[114] 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.

[115] 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. *Synthetic Comm*. **2003,** *33,* 4037-4042. DOI:10.1081/SCC-120026343.

[116] 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+.

[117] 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.

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

[119] The Bibliographic Ontology. http://bibliontology.com/ (accessed 11 June, 2010).

[120] MetadataTouch- Powerful Multi-Format Metadata Editor. http://www.digitalconfidence.com/MetadataTouch.html (accessed 29 July, 2010).

[121] Batchelor C.; Corbett, P.; Semantic enrichment of journal articles using chemical named entity recognition. *Proceedings of the 45th Annual Meeting of the ACL Demo and Poster Sessions* **2007,** 45-48. http://www.aclweb.org/anthology-new/P/P07/P07-2012.pdf (accessed 14 June, 2010).

[122] 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.

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

[124] Microsoft Releases Open Tools to Enhance Scientific Research Efforts Building on Science Commons Ontologies. http://www.microsoft.com/presspass/press/2009/mar09/03-11MSCreativeCommonsPR.mspx (accessed 18 July, 2010).

[125] Chem4Word – Microsoft Research. http://research.microsoft.com/en-us/projects/chem4word/ (accessed 4 Apr, 2010).

[126] Introducing the Office (2007) Open XML File Formats. http://msdn.microsoft.com/en-us/library/aa338205.aspx (accessed 21 April, 2010).

[127] 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.

[128] IETF/W3C XML-DSig Working Group. http://www.w3.org/Signature/ (accessed 20 July, 2010).

[129] Carroll, J. J.; Bizer, C.; Hayes, P.; Stickler, P. Named Graphs, Provenance and Trust. *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, NY **2005**, 613-622. DOI:10.1145/1060745.1060835.

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

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

[132] Brown, M. C. What is the semantic grid? IBM developerWorks. **2005,** http://www-128.ibm.com/developerworks/grid/library/gr-semgrid/ (accessed 14 Mar, 2006).

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

[134] Semantic MediaWiki. http://semantic-mediawiki.org/ (accessed 16 May, 2010).

[135] Freebase – A wealth of free data. http://www.freebase.com/ (accessed 16 May, 2010).

[136] AKSW : Projects / Onto Wiki. http://ontowiki.net/ (accessed 16 May, 2010).

[137] PROJECT HALO. http://www.projecthalo.com/ (accessed 16 May, 2010).

[138] Why Wiki - ChemSemWiki. http://www.ch.ic.ac.uk/wiki2/ (accessed 16 May, 2010).

[139] Marshall, C. C.; Shipman, F. M. Which Semantic Web? *Conference on Hypertext and Hypermedia*, Proceedings of the 14th ACM Conference on Hypertext and Hypermedia, Nottingham, UK, 2003; Ashman, H., Brailsford, T., Carr, L., Hardman, L., Eds.; ACM Press: New York, **2003,** 57-66. DOI: 10.1145/900051.900063.

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

[141] Bizer, C.; Heath, T.; Berners-Lee, T. Linked Data - The Story So Far. *Int'l J. on Semantic Web and Information Systems (IJSWIS)*, **2009,** *5* (3).

[142] Linking Open Data. http://esw.w3.org/SweoIG/TaskForces/CommunityProjects/LinkingOpenData (accessed 6 April, 2010).

[143] About the Linking Open Data dataset cloud. http://richard.cyganiak.de/2007/10/lod/ (accessed 6 April, 2010).

[144] Lagoze, C. The oreChem Project:Integrating Chemistry Scholarship with the Semantic Web. http://journal.webscience.org/112/2/websci09_submission_10.pdf (accessed 5 April, 2010).

[145] Bizer, C.; Cyganiak, R.; Heath, T. Linked Data - Connect Distributed Data across the Web. http://linkeddata.org/ (accessed 5 April, 2010).

[146] Lagoze, C.; Van de Sompel, H.; Johnston, P.; Nelson, M.; Sanderson, R.; Warner, S. ORE Specification and User Guide Table of Contents. Open Archives Initiative **2008,** http://www.openarchives.org/ore/1.0/toc (accessed 5 April, 2010).

[147] Laudel, G., Collaboration, creativity and rewards: why and how scientists collaborate. *International Journal of Technology Management* **2001**, *22*, 762–781.

[148] Rappaport, S. M.; Rzepa, H S. Intrinsically Chiral Aromaticity. Rules Incorporating Linking Number, Twist, and Writhe for Higher-Twist Möbius Annulenes. *J. Am. Chem. Soc.* **2008,** *130*, 7613-7619. DOI: 10.1021/ja710438j.

[149] Aleman-Meza, B.; Bojars, U.; Boley, H.; Breslin, J. G.; Mochol, M.; Nixon, L. J. B.; Polleres, A.; Zhdanova, A. V. Combining RDF Vocabularies for Expert Finding. *Proc. ESWC 2007*, Innsbruck, Austria, **2007,** 235-250.

[150] Bojārs, U.; Breslin J.G. ResumeRDF: Expressing skill information on the Semantic Web. *1st International ExpertFinder Workshop.* Berlin, Germany **2007**.

[151] Plaxo. http://www.plaxo.com/ (accessed 22 December, 2009).

[152] OpenSocial - Google Code. http://code.google.com/apis/opensocial/ (accessed 22 December, 2009).

[153] BlueObelisk. http://www.blueobelisk.org/ (accessed 7 July, 2010).

[154] Kruk, S.R.; Decker, S. Semantic Social Collaborative Filtering with FOAFRealm. *Intl. Semantic Web Conference (ISWC2005)*, Galway, Ireland, **2005**.

[155] LiveJournal FOAF. http://community.livejournal.com/ljfoaf/ (accessed 18 December, 2009).

[156] Grimnes, G.; Schwarz, S.; Sauermann, L. RDFHomepage or "Finally, a Use For Your FOAF File". *Proceedings of Semantic Web Scripting Workshop at ESWC06* Budva, Montenegro, **2006**.

[157] Rzepa, H.S.; Wheat, A.; Williamson, M.J. ChemSem: An Extensible and Scalable RSS-Based Seminar Alerting System for Scientific Collaboration. *J. Chem. Inf. Model.* **2006,** *46*, 985-990. DOI: 10.1021/ci0504115.

[158] Symplectic. http://www.symplectic.co.uk/ (accessed 15 June, 2010).

[159] Semantically Interlinked Online Communities. http://sioc-project.org/ (accessed 26 April, 2010).

[160] Docherty, P. http://totallysynthetic.com/blog/?p=1903 (accessed 15 June, 2010).

[161] Wang, X.; Zhang, B.; and Wang, D. Z.; *J. Am. Chem. Soc*. **2009**, DOI: 10.1021/ja904224y and 10.1021/ja910615z. (withdrawn 23 December, 2009).

[162] Murray-Rust, P. http://wwmm.ch.cam.ac.uk/blogs/murrayrust/ (accessed 15 June, 2010).

[163] Bachrach, S. http://comporgchem.com/blog/ (accessed 15 June, 2010).

[164] Wilde, O. http://www.quotationspage.com/quotes/Oscar_Wilde/31 (accessed 15 June, 2010).

[165] Cayzer, S. Semantic blogging and decentralized knowledge management. *Communications of the ACM* **2004,** *47* (12), 47–52.

[166] Semantic Blogs – Henry Rzepa. http://www.ch.ic.ac.uk/rzepa/blog/?p=1516 (accessed 15 June, 2010).

[167] Solc, T. Automatic generation of in-text hyperlinks in web publishing. http://www.tablix.org/~avian/blog/papers/generation_of_intext_hyperlinks.pdf (accessed 15 June, 2010).

[168] Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. DBpedia: a nucleus for a web of open data. *6th international semantic web conference and 2nd asian semantic web conference (ISWC/ASWC2007).* **2007,** Aberer, K.; Choi, K-S.; Noy, NF.; Allemang, D.; Lee K-I.; Nixon, L.J.B. *et al.* editors.

[169] Grossman, D. http://www.dangrossman.info/wp-calais-archive-tagger/ (accessed 16 May, 2010).

[170] Google Wave - Communicate and collaborate in real time. http://wave.google.com/ (accessed 14 June, 2010).

[171] Podcasts : chemistry@nature.com. http://www.nature.com/chemistry/podcast.html (accessed 14 December, 2009).

[172] Chemistry World Podcast. http://www.rsc.org/chemistryworld/podcast/ (accessed 18 December, 2009).

[173] Rzepa, H.S.; Cass, M.E. Progress towards a Holistic Web: Integrating OpenSource programs, Semantic data, Wikis and Podcasts. Spring 2006 Confchem. 2006. http://www.stolaf.edu/depts/chemistry/bh/confchem/confchem_s2006.htm (accessed 18 October, 2010.

[174] Hogan, A.; Harth, A.; Breslin, J.G. Podcast pinpointer: a multimedia semantic web application. 2nd European Workshop on Integration of Knowledge, Semantics and Digital Media Technology. **2005**.

[175] Shukla, N.K.; Agrawal, A. Semantic Mediacasting and Collaborative Feed Sharing. TENCON 2006. 2006 IEEE Region 10 Conference. **2006**.

[176] Hogan, A.; Harth, A.; Breslin, J.G. Podcast pinpointer: a multimedia semantic web application. *2nd European Workshop on Integration of Knowledge, Semantics and Digital Media Technology*. EWIMT. **2005,** 61-66.

[177] BiomedExperts: Scientific Social Networking. http://www.biomedexperts.com/ (accessed 22 December, 2009).

[178] Smarr, J. Technical and privacy challenges for integrating FOAF into existing applications. *The 1st Workshop on Friend of a Friend, Social Networking and the Semantic Web*. **2004**.

[179]Ankolekar, A.D.; Vrandecic, M.; Krotzsch, D. ThanhTran, The two cultures: mashing up Web 2.0 and the Semantic Web. *Proceedings of the 16th International Conference on the World Wide Web (WWW2007).* **2007.**

[180] Web of Knowledge - Science - Thomson Reuters. http://wokinfo.com/ (accessed 14 June, 2010).

[181] Zotero. http://www.zotero.org/ (accessed 14 June, 2010).

[182] CiteULike: Everyone's library. http://www.citeulike.org/ (accessed 1 October, 2010).

[183] About the Social Graph - Social Graph API - Google Code. http://code.google.com/apis/socialgraph/docs/ (accessed 22 December, 2009).

[184] Sandweiss, J. The Future of Scientific Publishing. *Phys. Rev. Lett.* **2009,** *102* (19) 190001.

[185] The Original Proposal of the WWW. http://www.w3.org/History/1989/proposal.html (accessed 10 June, 2010).

[186] NCSA Mosaic -- September 10, 1993 Demo. http://totic.org/nscp/demodoc/demo.html (accessed 10 June, 2010).

[187] Christensen, C. M. The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail [Book]. Harvard Business School Press, Boston, MA, **1997**.

[188] Lewis, D.W. How to Think about the Pace of Substitution of Open Access Academic Journals for Traditional Subscription Journals. **2010,**

https://scholarworks.iupui.edu/bitstream/handle/1805/2030/How+to+Think+about+the+Pace+of+Substitution+of+Open+Access+Journals+for+Subscription+Journals.pdf?sequence=1 (accessed 10 June, 2010).