Institution: University of Cambridge

Unit of Assessment: 8 - Chemistry

Title of case study: Chem4Word

1. Summary of the impact (indicative maximum 100 words)

From 2008-2010, Peter Murray-Rust developed a Chemistry Add-in for Microsoft Office Word, which enables users to insert and modify searchable, semantically rich chemical information within a Microsoft word document and for the data to be stored and manipulated in a semantically rich manner. The Add-in has been downloaded over 400,000 times. It was one of the first projects from Microsoft Research for which a public release under an open source license was obtained. This project demonstrated to a wide audience new semantic approaches to computing in chemistry. Chem4Word has impacted on education, publishing and science in industry and academia.

2. Underpinning research (indicative maximum 500 words)

Peter Murray-Rust joined Cambridge as a lecturer in the Department of Chemistry in 2000, and then Reader in (2004) until retirement (2008), remaining research-active in the Department thereafter. Together with Henry Rzepa (Imperial College London, since 1995) they established the first fully operational system for managing complex chemical content entirely in Extensible Markup Language (XML), which defines a set of rules for encoding documents in a format that is both human- and machine-readable. Chemical Markup Language (CML), which adapts XML to chemistry, is the first and the most ambitious application of semantic computing in chemistry. CML is the pioneering XML-based language for chemistry, providing a uniform, extensible system for representing, storing, and transmitting chemical information. CML supports 4 main methods of creating semantic chemistry: 1) human authoring (as in conventional articles, reports, laboratory notebooks, etc.), 2) conversion of chemical data from legacy formats, 3) creation of semantic chemistry through computer program output and 4) machine extraction of chemistry from unstructured and semi-structured material (e.g. electronic chemistry publications). CML also enables a range of concepts to be modelled, including molecules, reactions, and chemical metadata. CML has become an important component of many chemical information systems, including toolkits, structure editors, and other software.[1,2&3]

Based on the CML framework and Murray-Rust's vision of creating a simple way for chemists worldwide to insert “semantically intelligent” chemical information into documents using existing desktop applications such as Microsoft Office Word, the Chemistry Add-In for Word (Chem4Word) Project was initiated in 2008 in a collaboration between the Department of Chemistry and Microsoft Research. The new Microsoft Office Word format (.DOCX) would allow for chemical entities to be authored, manipulated, and stored as CML files within a DOCX package in a user-friendly manner, and the resulting files could easily be queried and mined for chemical data whether that was expressed in the document as a name, formula, image, or bold number reference.

Other chemical drawing tools are available, such as ChemDraw, which is used by many chemists to create publication-ready, scientifically meaningful drawings. A limitation of software like ChemDraw is that individuals, groups and organisations now have a large number of Word Documents containing embedded ChemDraw objects that cannot easily be searched in a chemically meaningful way. Additionally, only humans can create and interpret the chemical structures created by ChemDraw. Other limitations of some of the currently available software include limitations on the types of chemical structures that can be created, e.g. some are limited to only generating organic molecules. The Chem4Word project set out to simplify the process of inserting and modifying chemical information from within Microsoft Office Word, and also to have the chemical information stored and manipulated in a semantically rich manner. Working closely with Microsoft engineers, Murray-Rust and colleagues provided the underlying research for the design, semantic descriptions and ontology, software algorithms, and chemical knowledge for the program. Microsoft funded a postdoctoral scientist, Dr. Joe Townsend, to design and implement the software.

Murray-Rust and his group, together with Microsoft Research and the Microsoft Office Word team, jointly defined and developed the features that led to beta (March 2010) and version 1.0 (February 2011) releases of the Chemistry Add-in for Microsoft Office Word. This Add-in makes it possible not only to author chemical content in Word, but also to include the meta-data behind the structures, which means that the chemical information can be represented in a variety of ways: 2D chemical structures, names, chemical formulae and importantly, a variety of semantically encoded data. In other words, instead of a static picture one can search and select with a few clicks what information should appear in the embedded fields.[4]

Chemical Mark-Up Language
In addition to Chem4Word, the following projects (in whole or part involving the Department of Chemistry Unilever Centre for Molecular Informatics) have included development of CML specifications, software or applications: Molecular Standards for the Grid (DTI/EPSRC) 2002-2005, eMinerals (NERC), Materials Grid (DTI/Unilever/Accelrys), Sciborg (EPSRC), eCrystals, (JISC/Soton), SPECTRa (JISC), SPECTRa-T (JISC), Crystal editor (IUCr), Crystal repository (Dept of Chemistry), Polymer Informatics (Unilever), MDL2CML (MDL), Openbabel (Merck).[5]

3. References to the research (indicative maximum of six references)

1. Chemical markup, XML, and the world wide web. 6. CMLReact, an XML vocabulary for chemical reactions. Holliday GL, Murray-Rust P, Rzepa HS. J Chem Inf Model. 2006 Jan-Feb; 46(1): 145-57. (*****)
2. CML: Evolution and design. Peter Murray-Rust and Henry Rzepa. J Cheminform. 2011 Oct 14; 3(1): 44. doi: 10.1186/1758-2946-3-44. (*****)
3. The semantics of Chemical Markup Language (CML): dictionaries and conventions. Peter Murray-Rust et al. J Cheminform. 2011 Oct 14; 3:43. doi: 10.1186/1758-2946-3-43.
4. Chemistry Add-in for Word. Microsoft Research. http://research.microsoft.com/en-us/projects/chem4word/
5. http://www-pmr.ch.cam.ac.uk/wiki/Funded_CML_Projects.

(*****) References that best indicate the quality of the research.

Grant Information:

- Grant No: MAZA042, RG50497; PI: Peter Murray-Rust; Grant title: Chem4Word; Sponsor: Microsoft; Period of Grant: 1-10-2007 to 30-9-2009; Value of Grant: £106,404.80.
- Grant No: MAZA056, RG59518; PI: Peter Murray-Rust; Grant title: Chem4Word; Sponsor: Microsoft; Period of Grant: 31-3-2010 to 31-3-2013; Value of Grant: £115,445.

4. Details of the impact (indicative maximum 750 words)

Microsoft Research has supported, built, and collaborated on a large number of Office extensions in a number of scientific domains, but the Chemistry Add-in for Word remains the most highly downloaded project developed to date, that number exceeding 400,000 as of July 2012. (See Corroboration Letter LC1).

Chem4Word enables both humans and machines to understand the underlying semantics of the documented chemistry and expose semantically rich chemical information across the global chemistry and chemical information community.

The Chemistry Add-in for Word has been widely recognised by educators, publishers and software developers as making it easier for students, educators, and chemists to insert and modify semantically searchable chemical information, such as labels, formulae and 2D depictions, from within Microsoft Office Word. Designed for and tested on both Word 2007 and Word 2010, it makes chemistry documents open, readable and easily accessible, not just to other chemists, but also to other (robotic) technologies using the widely adopted extensible markup language (XML). The Chemistry Add-in supports both publishing and data-mining scenarios.[1]

Impact on Publishing

Chem4Word has been widely welcomed by publishers. The following testimonials serve as an example of this: "The future of research will be powered not only by ever more rapid dissemination of ever large quantities of data, but also by software tools that 'understand' something about science. These tools will behave intelligently with respect to the information they process, and will free their human users to spend more time doing the things that humans do best: generating ideas, designing experiments and making discoveries. Chem4Word is one of the best examples so far of this important new development at the interface between science and technology." Managing Director, Digital Science, Macmillan Publishers.[2]

"The IUCr is delighted to see the release of v.1 of Chemistry Add-in for Word under an open-source development license - this has great potential for authors to enrich the semantic content of their articles, and for publishers to leverage this semantic content in creating ever more useful and powerful active publications." Research and Development Officer, International Union of Crystallography.[2]

Impact on Education

Chem4Word is promoted by Microsoft as a tool for teachers and students as a Chemistry teaching and learning aid.[3 & 4]

Impact on Open Source Publishing

Much of the work of science depends on having appropriate tools available to analyse experimental data and to interact with theoretical models.

The Chem4Word research collaboration was unique at the time, with software development happening both within Microsoft and the University of Cambridge, and it provided a model for future collaborations between Microsoft Research and academic institutions worldwide. Additionally, the Chem4Word project was one of the first Microsoft Research projects for which it sought public release under an open source licence from the project's inception. Releasing software in this way was extremely rare for Microsoft at the time, but was recognised as an important component in building long-term collaborative development communities. The Chem4Word project was instrumental in formulating new policy at Microsoft Research to define and grow a broader open-source strategy, and contributed to the establishment of the Outercurve Foundation, whose mission is to enable the exchange of code and understanding among software companies and open source communities. As a direct result the source code for the Chem4Word Add-in was released under an open-source licence and the intellectual property was assigned to the Outercurve Foundation to facilitate broader community involvement and governance, leading the way for future releases of Open software via Microsoft and other previously closed vendors.

"Releasing software under open source licenses was extremely rare for Microsoft at the time, but was recognized as an important component in building long-term collaborative development communities. The Chem4Word project helped Microsoft Research to define and grow a broader open-source strategy, which has led to a large number of open source releases, and ultimately to the establishment of the Outercurve Foundation." Quote from Director Scholarly Communication Microsoft Research (see Corroboration Letter, LC1).

Impact on Chemical Informatics

In 2012, Peter Murray-Rust was a joint recipient (with Henry Rzepa) of the Herman Skolnik Award presented by the American Chemical Society Division of Chemical Information, an award that recognises outstanding contributions to and achievements in the theory and practice of chemical information science and related disciplines. According to the ACS "*Their work has had a huge impact in the fields of chemical document analysis, chemistry on the Internet, and in the orchestration of a viable strategy for making electronic chemistry information as widely accessible and usable as possible in our information age.*"[5]

6. **Sources to corroborate the impact** (indicative maximum of 10 references)

Letter of corroboration available for audit
LC1 Director Scholarly Communication, Microsoft Research, Redmond, WA 98052

References in the public domain

1. http://chem4word.codeplex.com/
2. What people are saying about the Chemistry Add-in for Word http://research.microsoft.com/en-us/projects/chem4word/quotes.aspx
3. http://www.decd.sa.gov.au/it/files/links/Bringing_a_1_to_1_Progra_1.pdf
4. http://cie.acm.org/articles/microsoft-research-connections-collaborating-reinvent-education/
5. http://bulletin.acscinf.org/node/245

Users/Beneficiaries who can be contacted to corroborate claims
Manager Informatics, Royal Society of Chemistry Publishing (verify benefit to publishers)
Managing Director, Digital Science, Macmillan Publishers (see Reference 3)
Research and Development Manager, International Union of Crystallography (see Reference 3)