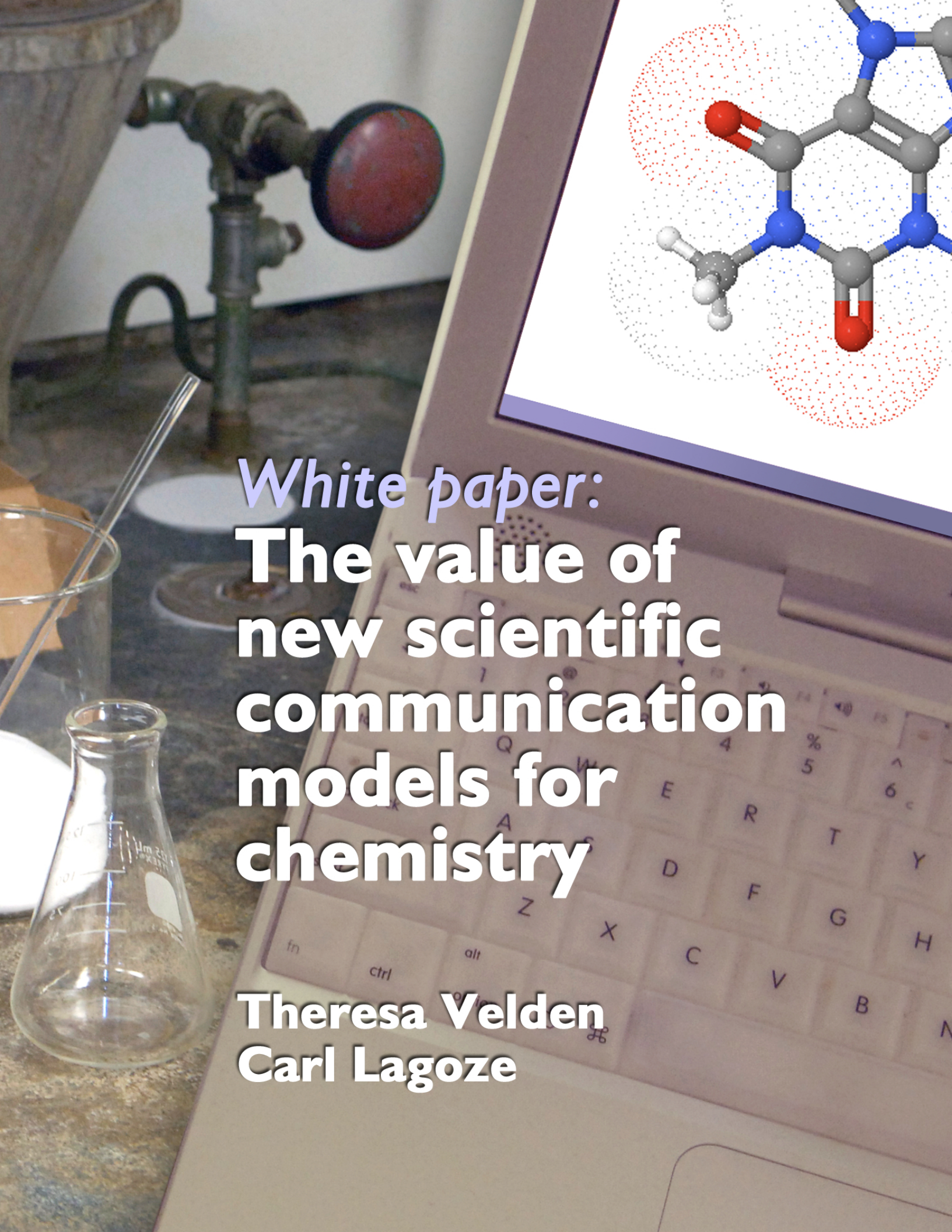

White paper:

The value of new scientific communication models for chemistry

Theresa Velden
Carl Lagoze

White Paper:
The Value of New Scientific Communication Models for Chemistry

Objective:

- To serve as input for an international workshop of academic and industrial chemists, information scientists, publishers, scientific society representatives, and information service providers to validate and refine the analysis presented here, and to develop recommendations for action and further research

Contributions:

- Assesses status of new models of scientific communication in chemistry, concluding that take-up is limited and critical mass to stimulate widespread usage is missing

- Forwards hypotheses to explain this finding and points to open questions

- Offers background information to provide a common starting point to workshop participants with a variety of professional backgrounds by:
    1. Pointing to connections to related activities and discourses in publishing reform movement and eScience

    2. Describing the scientific communication system in chemistry and projects to develop new models in chemistry

    3. Proposing relevant field characteristics of chemistry

Authorship of this Document:

The authors of this white paper are Theresa Velden and Carl Lagoze who gratefully acknowledge contributions from Roald Hoffmann, Leah Solla, Hilary Spencer, Jeremy Frey, John Wilbanks, Clifford Lynch, and Jean-Claude Bradley. The responsibility for the content as presented remains that of the authors alone. This white paper is informed by a workshop that was held on October 23-24, 2008, which was attended by the people listed below. This white paper does not represent views or opinions of all workshop participants or of the organizations with which they are affiliated.

Workshop Participants (Washington DC, 23-24 October 2008)
"New Models for scholarly Communication in Chemistry"

Colin Batchelor - Royal Society of Chemistry

Christine Borgman - University of California, Los Angeles

Jean-Claude Bradley - Drexel University

Jeremy Frey - University of Southampton

Roald Hoffmann - Cornell University

Jane Hunter - University of Queensland

Peter Jerram - Public Library of Science

Susan King - American Chemical Society (*Attended only on October 23*)

Carl Lagoze (Principal Investigator)- Cornell University

Clifford Lynch - Coalition for Networked Information

Peter Murray Rust - University of Cambridge

Heinz Saller - InfoChem GmbH

Leah Solla - Cornell University

Hillary Spencer - Nature Publishing Group

Theresa Velden - Cornell University

John Wilbanks - Science Commons

Executive Summary

This paper is intended as a starting point for discussion on the possible future of scientific communication in chemistry, the value of new models of scientific communication enabled by web based technologies, and the necessary future steps to achieve the benefits of those new models. It is informed by a NSF sponsored workshop that was held on October 23-24, 2008 in Washington D.C. It provides an overview on the scientific communication system in chemistry and describes efforts to enhance scientific communication by introducing new web-based models of scientific communication. It observes that such innovations are still embryonic and have not yet found broad adoption and acceptance by the chemical community. The paper proceeds to analyze the reasons for this by identifying specific characteristics of the chemistry domain that relate to its research practices and socio-economic organization. It hypothesizes how these may influence communication practices, and produce resistance to changes of the current system similar to those that have been successfully deployed in other sciences and which have been proposed by pioneers within chemistry.

The fact that the perspective presented in this paper is not unanimously shared across the board of stakeholders within chemistry was evident from the comments of some participants of the October 2008 workshop to the draft of this paper. Change in established systems is difficult and inevitably disrupts practices that are considered essential by established stakeholders. The revised version of the paper that you are now reading acknowledges this and highlights issues of disagreement among the stakeholders represented at the workshop. Further, the analysis in this paper is incomplete with regard to the many different research fields within chemistry. Additional work, deepening, and validating the analysis presented in this paper is needed. Hence, we see this document as only a first step and propose it as the basis of a second, broader workshop. This workshop would include a broad range of chemists, both from academia and industry, and other stakeholders in the scientific communication system in chemistry, as well as researchers who study transformation processes in the sciences. The aim of such a workshop would be to critically discuss and further develop the analysis presented here, and to design concrete recommendations on

- How to assess the value of new scientific communication models in chemistry?
- How to catalyze desirable changes?
- What aspects require further exploration and research?

We suggest this document and the proposed second workshop have broader value. We believe that the domain of chemistry with its cautious approach to new communication models constitutes a valuable case study for transformation processes in scientific communication in the Digital Age. Efforts to innovate scientific communication will benefit from an increased understanding of discipline and research field specific factors, which can be acquired through the discussions and analyses that this paper aims to initiate.

Acknowledgements

Funding for the 'New Models for Scholarly Communication in Chemistry' workshop held in Washington DC on October 23-24, 2008 was provided by the National Science Foundation through grant IIS-738543 SGER: Advancing the State of eChemistry. Additional follow-on support for research and report writing was provided by Microsoft. The content of this white paper does not necessarily represent the views or opinions of the funders, contributors, or workshop participants or of the organizations with which they are affiliated. Thanks to Carol Minton-Morris for the cover art.

Request for Feedback

The authors welcome feedback on this document. E-mail addresses are:

- Theresa Velden – tav6@cornell.edu
- Carl Lagoze – clagoze@gmail.com

1 Introduction

This white paper builds on discussions about new scientific communication models in chemistry at a two-day NSF-funded workshop in October 2008 in Washington D.C. It initiates assessment of the status of scientific communication in chemistry, and reflects on its evolution in the context of new web-based information and communication technologies (ICTs).

The intention of this document is to provide a starting point for discussion. The analysis presented here is preliminary only, and is incomplete in its coverage of the many different areas of research in chemistry. Also, the initial draft of this white paper that was circulated among the participants of the Washington workshop did not find unanimous agreement. It is our impression that within the chemistry community the topic is controversial and perceived as disruptive to the status quo. Hence the authors of this document have decided that rather than homogenize the original draft to a consensual document, to instead highlight where substantial disagreement exists. We expect this to be informative for further analysis and dialogue between stakeholders. We propose a follow-up workshop with a broad range of participants and stakeholders of scientific communication in chemistry. The aim of this workshop would be to re-examine, improve, and extend this analysis, to define open research questions, and to explore opportunities for joint action to evolve the communication of chemical information along with new web-based information and communication technologies.

The World Wide Web, the Semantic Web, and the social networking tools of Web 2.0[1], together with the digitization of content and the increasing processing power of computers, are claimed to revolutionize the ways researchers communicate with one another, disseminate results, collaborate and share data and knowledge (Arms 2007). Indeed, in a number of scientific fields new models of scientific communication have emerged that are based on these technologies. Many of these have inspired wide community participation. Examples are: the public dissemination of manuscripts[2] of research articles through the e-print server arXiv[3] in various fields of physics, mathematics, and quantitative biology (Gunnarsdóttir 2005), the sharing of research data through public databases such as GenBank[4] in genomics and the integration of this data

[1] Web 2.0 refers to recent web developments that are more interactive than early web-based services, and enable web users to participate in the creation of web content (by sharing videos or images, by providing commentary, editing shared resources like wikipeadia etc.) See http://en.wikipedia.org/wiki/Web_2.0 for more information.
[2] Authors post on arXiv their manuscripts as unrefereed preprint before or in parallel to submitting it to a journal, and often also the revised, refereed 'postprint'.
[3] http:/arxiv.org
[4] http://www.ncbi.nlm.nih.gov/Genbank/

with the increasing open access journal literature in biomedicine (Benson 2007) through PubMed Central[5], and community-based open peer review and discussion in interactive online journals[6] in geosciences (Pöschl 2008).

The success of new web-based models in neighbouring disciplines and at the periphery of chemistry (such as drug design) stands in contrast to the lack of comparable success stories in chemistry. Why do similar initiatives in chemistry fail to gain critical mass and widespread usage?

So far, a number of technical developments have been undertaken to pave the way for the exchange of chemical information on the web. Examples of these initiatives include a standardized mark-up language (CML), and a computable identifier for organic molecules, the IUPAC International Chemical Identifier (InChI), open source tools for the manipulation and management of chemical information (Blue Obelisk, Guha 2006), and the use of free, hosted Web 2.0 services to support Open Notebook Science (Bradley 2008). The pioneers of these developments promote a vision where the entire life cycle of research - from planning of research projects and experiments, through conducting experiments in the lab, to analysis of results, and finally to dissemination and publication - is supported by capture, storage, and interlinking of the underlying (raw and derived) research data (Frey 2009). The benefits expected are: increased transparency of the research process, improved verifiability of research results and their reproducibility, increased efficiency in local management of data in research teams, increased efficiency of global research through the ability to re-use data, opportunities for new forms of research by processing and data-mining large aggregated data sets, and facilitation of distributed and open research on under-researched areas (such as tropical disease) that lack commercial incentives, but would benefit from 'crowdsourcing'[7].

At this point in time, we find projects that demonstrate technical and conceptual feasibility. So far, however, these initiatives fail to take on critical mass to become an integral part of the scientific communication system in chemistry. Outside of specific subfields like cheminformatics or crystallography, few chemists seem to perceive these developments as opportunities to enhance scientific communication practices (Todd 2007).

Hence, while projects and ideas proliferate, is chemistry as a discipline ready to take them on? What is needed to make them successful? Will we eventually see substantial changes to the current models of scientific communication in chemistry, if only delayed in comparison to other disciplines? Or will the current

[5] http://www.pubmedcentral.nih.gov/
[6] http://www.atmospheric-chemistry-and-physics.net
[7] i.e. the mobilization of a large number of independently contributing volunteers. See for further explanation of the term http://en.wikipedia.org/wiki/Crowdsourcing. For an example of its use in astronomy to classify galaxies in a large survey of the sky see 'Galaxy Zoo' https://www.galaxyzoo.org/.

scientific communication system in chemistry remain essentially unaltered? If so, is this to be taken as an indication of an extreme mismatch between the opportunities offered by new ICT's and the research practices and values of the chemistry community? Or are there particular barriers that could be sensibly addressed? Or does the existing system in chemistry already provides the best possible match to research needs, meaning that no fundamental changes are expected from new technological capabilities apart from incremental improvement? Also, is it misleading to suggest that there is a deficiency of the current scientific communication system in chemistry, and to imply that other disciplines are more advanced in their adoption of web-based communication models? For instance, one could argue that there is no significant difference in the way chemistry has taken up web-based ICTs compared to other disciplines. Innovations like arXiv or PubMed Central have been adopted in specific fields within physics and the life sciences respectively, but not uniformly across the entire disciplines. So how would one sensibly compare developments in different disciplines and learn from such comparisons?

As a starting point for such a discussion we need to understand the particular ways in which scientific communication supports knowledge production in chemistry and how this can be meaningfully compared to other disciplines.

The authors of this document believe that the field of chemistry is an instructive case study of factors influencing the adoption of new models for scientific communication at large. So far, talk about the revolutionary impact of the web and other ICTs on scientific communication seems to ignore the rather large differences in take-up among various disciplines. An analysis of the reasons why that is so, and a deeper understanding of factors that shape the scientific communication system in chemistry, will not only benefit initiatives to introduce new scientific communication models in the domain of chemistry, but should be informative also for activities in other scientific domains that expose characteristics of 'long tail science'[8].

This white paper presents results of an initial workshop on this issue. We propose a broader workshop that aims to engage various groups of professionals (chemists, publishers, information service providers, representatives of scientific societies, and information scientists) to validate and deepen the analysis presented in this paper, and to develop recommendations on:

- How to assess the value of new scientific communication models in chemistry?
- How to catalyze desirable change?
- What aspects require further exploration and research?

This document develops a preliminary analysis of the value of new scientific communication models in chemistry. It is intended as an input for the workshop

[8] This term refers to a field of research dominated by large numbers of small collaboration units. It has been coined by Jim Downing in contrast to large-scale collaborative 'big science' such as experimental high-energy physics (Murray-Rust 2008).

and hence aims to address and inform people with a range of different backgrounds. The paper outline is as follows: Section 2 provides some background on terminology used in this paper and describes connections to and insights from existing work on publishing reform and e-science, or cyberinfrastructure initiatives. Section 3 provides an overview of chemical information services and their evolution with the web. Section 4 presents a tentative analysis of the particular characteristics that distinguish chemistry from other scientific disciplines, and their implications for scientific communication. Section 5 summarizes our conclusions, highlights points of dissent, and closes with defining the aims for the envisioned second workshop.

2 Background: Terminology & Related Discourses

In the first part of this section we define key terms used in this paper. Then, to put the discussion of new models for scientific communication in chemistry into the context of related movements, we briefly review publishing reform and e-science initiatives. We will highlight insights on the dynamics of transformation that may be useful in the analysis of the value of new scientific communication models in chemistry.

2.1 Terminology

When we use the term 'scientific communication' we refer to variety of practices of information exchange between scientists directed at the generation of scientific knowledge. These practices include informal communication between researchers in all kinds of settings; in seminars, at workshops or conferences, in writing (e.g. emails), by telephone, in corridors, via videoconferencing, or face-to-face. They also include formalized vehicles such as the publication of articles in peer-reviewed journals. The distinction between informal and formal communication stems from library science and rests on whether the communication becomes part of the enduring, and archived record of science (i.e. will be archived and can be cited). The boundary is somewhat fuzzy, as it does not map 1:1 onto the status of an item as having been peer-reviewed or having been made public. Informal communications can be private, but need not be – e.g. a talk given at a conference, although public, would be considered part of informal communication. And monographs may not have been peer-reviewed, but are still considered formal communication because they are included in the record of science.

When we speak of the 'scientific communication system' we refer to a combination of conventionalised practices, technical infrastructures (such as printing houses, libraries), communication vehicles (such as journals, books, conferences, etc.), institutions and people. The term 'system' suggests a coherent assembly of parts to support the flow of scientific communication in its entire lifecycle, from inception of ideas, where scientific information may be one of the inputs, through the informal discussion of work in progress, to formal publication in journals, and indexing and archiving for future retrieval.

Scientific domains differ in how their scientific communication system is realized[9] and this realization may evolve as technologies and disciplines evolve. Certain core functions of the system though are not expected to change regardless of the actual form scientific communication systems may take in the future (Roosendaal and Geurts, 1998). These functions are registration (to establish priority), certification (to validate results), awareness/distribution (to learn about

[9] e.g. whether preprint dissemination is a valued part, or whether conference papers rival journal articles in reputation as they do in computer science.

and gain access to new results), and archiving (to preserve the scientific record). Sometimes, a fifth function, recognition or reward is added, pointing to the increasing emphasis put on the reputation of publishing outlets and citation – based measures of impact for career evaluations and funding decisions (de Sompel et al., 2004). This latter function may be seen as a secondary (derived) function of the scientific communication system, whereas the aforementioned primary functions directly concern the production, dissemination, and longevity of reliable scientific information.

The constituent parts of scientific communications systems, such as scientific journals or preprint servers, have been described as 'communication regimes' (Hilgartner 1995), or 'socio-technical interaction networks' (Kling 2003) to emphasize their historical contingency and the fact that they are social as much as technical arrangements, where the social and the technical configurations mutually shape one another.

Further, in this paper, we repeatedly use the term 'new models of scientific communication' to refer to new forms of scientific communication. These models exemplify ways in which web-based technologies can be used to support new practices of scientific communication[10]. When reading the term 'model' here, consider a) that it refers to a complex socio-technical arrangement and not just a technical system, and b) that the question of transferability of such a model from one scientific field to another is problematic. The terms 'communication regime' or 'socio-technical interaction network' are more descriptive, but the simpler term 'model' is more appropriate for the mixed audience of this paper.

2.2 Publishing Reform

The term 'publishing reform' encompasses a variety of ideas and initiatives to improve the scientific communication system. Advocates conceive of the advent of the web as a fundamental transition from the 'Gutenberg Era' where printing dominated the communication of scientific information to a new 'Digital Era' (Harnad 1991, Giles 1996, Borgman 2000). The web, and with it electronic publishing, is seen as a 'disruptive technology' (Christensen 1997) that undermines established business models of the scientific publishing industry, as it offers unanticipated, ease of the distribution of content along with new, unprecedented capabilities (Odlyzko 2002). The network character of the web enables multiple actors to contribute services to the communication system. For example, core functions such as registration, awareness, certification, and dissemination that were formerly bundled together in the centralized production system of the paper-based journal, may now be disaggregated and performed by an interplay of services, such as preprint servers (registration, awareness), so called overlay journals providing peer-review services (certification), and search

[10] Our use of the term 'model' does not imply that such a communication form is only some idealized, and abstracted idea that will never work in production - on the contrary, we are talking about actual forms of communication

engines providing access to online versions (dissemination) (de Sompel et al. 2004). Further, some envision that scientific publishing should not only make publications, data, annotations accessible to humans, but it should also be 'machine-readable' making it possible for computers to combine information from distributed sources and compute new results.

In this subsection we will briefly review major strands of publishing reform, and provide some observations on the dynamics of the transformation of scientific communication systems. We only touch on but not analyse in depth the financial and economic issues of such reforms.

2.2.1 Open Access

Open access is concerned with free-to-the-reader web-based access to the results of publicly funded research in order to optimize their impact and ease of use and re-use. It addresses both price barriers for access to research results (such as journal subscription prices), as well as permission barriers, that is restrictions on the re-use of scientific material even for legitimate scientific purposes (Suber 2007). A primary concern is access to the peer-reviewed scientific journal literature. However, research data and cultural heritage are considered within the wider scope of open access.

Open Access Definition

"By 'open access' to this literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited."

Budapest Open Access Initiative
February 14, 2002

http://www.soros.org/openaccess/read.shtml

An important driver of open access has been economic: the volume of scientific articles published as well as journal subscription prices have risen dramatically over the last two decades, while academic library budgets have almost stagnated (Cope & Kalantzis 2009, Edlin & Rubinfeld 2004). Scientific journal publishing has seen growing market concentration, with a large portion of scientific journal content now being published by a few commercial publishers: Elsevier, Springer, Taylor & Francis, and Wiley-Blackwell. Society publishers have been more moderate in their pricing policies than commercial publishers, who ask higher prices and have higher profit margins. This divergence is growing further, with some publishers being more aggressive about increasing

prices than others (McCabe et al. 2006). These trends have produced the so-called 'serials crisis' where libraries are forced to cancel journal subscriptions, to streamline their collections by buying into comprehensive electronic bundle deals of large publishing houses, and to cut their monograph acquisition budget to be able to afford access to the journal literature for their patrons. Limitations on access to the journal literature have particularly impacted smaller, less well-funded institutions and interdisciplinary researchers who require specific literatures from a broad range of fields.

The idea of 'open access', free-to-the-reader, seems attractive as a means of addressing the limitations in access to the journal literature due to subscription ('toll') barriers. In the long run it implies switching journal publishing from a subscription based business model to one that offers readers access to published articles for free with production costs covered by other means e.g. by an article processing fee from the author or a research funding institution. In this manner, journal publishing costs are packaged with the much larger research costs (staff, equipment) and paid at the source, and thereby maximise the impact of the published output by making access free to the reader. A number of issues are associated with a switch to a pay for publication model that play out differently in different disciplines. For example in humanities research equipment costs are usually low. Therefore the addition of even a modest amount of publishing costs to research funding would be perceived as quite high and daunting. Hence, a truly viable and provable sustainable economic model for open-access publishing across the board has yet to emerge.

A second important driver of open access is the anticipated value of an integrated network of scientific information that would become possible once published research information is openly available on the web. Web technologies offer new capabilities to link and integrate research results, to search and mine information, and to re-use data. To take full advantage of these opportunities, content needs to be widely accessible across databases and publisher platforms. In particular, data-driven sciences as well as interdisciplinary research would benefit from seamless access and powerful tools to exploit scientific information.

At this point in time several 'open access' declarations[11] have found support by hundreds of universities, research institutes and funding organizations. They highlight two alternative routes that authors can take to provide open access to their research articles. The so-called 'golden road' of open access publishing:

[11] Budapest Open Access Initiative 2002, online at: http://www.soros.org/openaccess/read.shtml
Bethesda Statement on Open Access Publishing 2003, online at: http://www.earlham.edu/~peters/fos/bethesda.htm
Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities 2003, online at: http://oa.mpg.de/openaccess-berlin/berlindeclaration.html

the article is published in journal that grants open access to it[12]. Or the 'green road' of self-archiving: authors deposit the final version of their peer-reviewed manuscript in an institutional repository[13], or in a subject-based repository such as the preprint server arXiv, or PubMed Central[14]. arXiv is notable because it has a versioning system to accept revised, final refereed versions of preprints. This deposit is not intended to replace publishing in a peer-reviewed venue. Rather, it is intended to return some level of control over the dissemination of publications back to their originators, authors and research funders. Such repositories function as a back-up that ensures access to research output independent of an interested reader's institutional access to journals, the fate of publishers' online archives, or eventual disputes about access conditions. In the bio-medical sciences funding organizations such as Wellcome Trust (U.K.) and NIH (U.S.A) require their grantees to deposit their articles in PubMed Central. At this point in time, approximately 50% of publishers and about 63% of journals allow authors to post the final peer-reviewed version of their article on their personal web page or deposit it to the open access repository of their home institution[15]. Services such as Google Scholar provide indexing of author homepages and institutional repositories in addition to journal content.

2.2.2 Preprint Servers

The high-energy physicist Paul Ginsparg set up the physics preprint server arXiv on the Internet in 1991, a few years before the advent of the web. By now, several fields have adopted web-based mechanisms for the dissemination of preprints, working papers, or final peer-reviewed articles – that is, research reports that have undergone minimal or extensive quality control. Most such services have some form of screening or accreditation mechanism such that the material posted can be expected to be at least of refereeable quality (Ginsparg 2004). Major services are RePeC for economists, SSRN for social scientists, arXiv serving now a number of quantitative sciences, and the recently introduced Nature Precedings for the life sciences. In computer science a slightly different model has been adopted. Computer scientists often post their manuscripts on their home pages, where a specialized search engine, CiteSeer, indexes them and makes them available for searching and download.

---

[12] There are genuine open access journals that provide open access to all its articles, as well as 'hybrid' journals that require subscriptions for accessing regular articles and only provide open access to selected articles for which their authors have paid an additional publication fee. A current list of open access journals is provided by the Directory of Open Access Journals (DOAJ) at http://www.doaj.org

[13] The term 'institutional repository' refers to a web based document server set up and maintained by an institution such as a University or a research organization to collect and publicly disseminate digital artefacts of its intellectual output (papers, proceedings, learning material etc.).

[14] Possible subject to an embargo period of six or 12 months – depending on the journal's policy.

[15] See statistics from the SHERPA/ROMEO project at http://romeo.eprints.org/stats.php

Advocates and users claim that the advantages of these services are manifold. First, they provide immediacy of access to the latest research results to anyone in the scientific community who might be interested in them. In addition, they serve as a comprehensive archive of the literature of a field that normally is dispersed over a large variety of sources such as journals or conference proceedings. Finally, to the extent that they include content outside the limitations of peer review, they provide a long tail of research results with items at the end of the tail that may reveal value at some later time. The majority of content within preprint servers eventually appears also in the peer-reviewed journal literature. Because of that preprint servers frequently serve a complementary function, and do not substitute for peer-reviewed publication venues.

2.2.3 Open peer review

Peer review is highly valued by the scientific community, though not considered to be without flaws (Ware 2008). Open peer review generally aims at using the immediacy of the online medium to involve a broader range of experts in the evaluation of the merits and shortcomings of a research article. Open peer evaluation may take two forms: the use of online tools to facilitate an open version of traditional peer review that happens before publication of an article; or enabling all readers to submit post publication commentary on traditionally peer reviewed and published articles. In addition, open peer review aims to increase the transparency of the peer review process itself by either disclosing the identity of the peer reviewers to the authors, or by publishing the reviewers' comments along with the article (or both) (Falagas 2007).

The disclosure of the reviewers' identities at least to the authors, if not publicly, is thought to avoid unfair or careless treatment of an author by the reviewer and to foster a constructive dialogue between the two. The British Medical Journal, an open access journal, adopted this practice already in the 90's, and further complemented it with the option for post publication peer commentary. The journals of the European Geosciences Union publish the peer review comments, as mentioned in the introduction of this paper. If a submitted manuscript meets minimal quality criteria that make it worthy of evaluation and discussion, the reviewers' comments[16] are published along with the manuscript in an online discussion forum. In this manner the scientific community is invited to get involved in the debate. After a period of a few weeks, the authors are asked to revise their manuscript based on the discussion feedback. The authors may also defend their position. In the end, the finalized article version is published in the journal. Expected benefits are: increased quality of submitted articles (to avoid embarrassment by criticism in the discussion forum due to carelessness in the manuscript preparation), elimination of unjustified suppression of publications due to reviewer bias, early availability of new results, encouragement of scientific discourse that evaluates and contextualizes new research results, and – because

[16] Reviewers may request to remain anonymous – hence their comments are public but not necessarily their identity.

review comments and comments are permanently archived and given a citable ID - rewarding of reviewers for effort put into the preparation of careful reviews and criticism (Pöschl 2008).

2.2.4 Data Publishing

Data publishing aims at making research data publicly available in a re-usable format in order to a) support legitimate academic reuse and exploitation of data and avoid unnecessary duplication of work, and b) enable validation of reported research results through investigation of the underlying data.

A critical issue for data publishing is the variance in the motivation of researchers to share data at some point. The factors involved include the stage of research, the type of data, the investment made in its generation, the effort needed to make it available in a useful form, and the value of the data for the community as a shared resource versus the competitive advantage of keeping it private for further exploitation. Because of these factors, different trade-offs exist for making ones data publicly available (Hilgartner 1997, Birnholtz 2003). Data publishing efforts focus in particular on increasing the incentive for making data publicly available, and rewarding the effort needed to make them useful for third parties e.g. by providing sufficient descriptive and contextual information (so-called metadata). One strategy is to raise the status of publicly releasing data sets to the status of formal scientific publications by making them part of the scientific record, i.e. by ensuring some form of quality control, long-lasting access, and a standard identifier to enable formal citation of data sets (Klump et al. 2006).

One may broadly distinguish two forms of data publishing based on their different scope: one aims at making available the specific primary data that underlie a scientific article's claims. The other regards data sets by themselves as worthy of publication; hence no original research claim is necessarily linked with the data set. Two major incentive schemes for data publishing are at work in areas where data publishing has become the norm: either journals require publication of primary data (e.g. in crystallography), or (as in genomics), funding agencies request deposition of all data created with help of the research funds into a public database – independent of or prior to an article publication (Swan 2008).

2.2.5 Science Blogs

Science blogs, that is web blogs[17] on scientific topics, belong to the realm of informal communication. Usually authorship is controlled, one or several authors publish on their web blog, at more or less regular intervals, entries of a few lines to entire essays. Typically the writing style is conversational, and humorous content gets mixed with posts of a more serious tone. Science blogs come in different flavors. Some are dedicated to science communication. They educate and communicate excitement about research to a lay audience, or aim to balance

[17] Definition: http://en.wikipedia.org/wiki/Weblog

what is perceived as one-sided reporting or political bias e.g. on topics such as evolution or climate change. Some aim at an academic discussion and provide e.g. commentary on published scientific literature or initial thoughts on unpolished scientific ideas. And a good many are more like personal diaries, providing emotional release and social exchange about the trials of day-to-day research. Many represent a mix of all three flavors. At this point in time most science bloggers are assumed to be less than 30 years old. They are journalists, teachers, graduate students or young researchers (Bonetta 2007). Hardly any established scientists maintain a web blog – after all blogging regularly is very time consuming (Wilkins 2008).

Science blogs are a very recent genre. In most disciplines they do not constitute widely recognized academic resources. Nevertheless they seem to serve some small, but potentially global communities very well as a means of exchange of ideas and the development of arguments. Web 2.0 tools such as web blogs, wikis[18], and social networking sites facilitate participation in informal global scientific communities. The question is whether these will remain fringe phenomena or become part of the mainstream organization and communication in science.

Science Web

The vision to create a global 'knowledge space' out of the integration of the various layers of distributed scientific information discussed above, requires technologies and agreements that support interoperability and integration. It recognizes that the simple use of the web as a network of hyperlinks between documents is not sufficient. A more expressive and flexible alternative is to view the web as a network of "linked data" (Bizer et al. 2007)– a platform for a rich application layer. Ingredients for realizing such a vision are Semantic Web technologies and policy agreements that facilitate the exchange and integration of content on the web.

Semantic web technologies enable the sharing and re-use of data across

The promise of semantics

"If we can get the world promised by semantic computing, chemistry and other scientific disciplines will be transformed, much like the web has transformed culture and commerce. We'll be able to get precise answers to complicated questions, we won't have to maintain dozens of tabs in a browser, and using Excel to integrate data will be as quaint as using a sliderule. Since knowledge represented in this way tends to look an awful lot like a network, we will find out if Metcalfe's Law applies to knowledge the way it's applied to computers and to documents - will there be exponential increases in the value of what we know once we hook it up the right way?"

John Wilbanks, Vice President Science Commons

[18] Definition: http://en.wikipedia.org/wiki/Wiki

applications[19] that support the integration of papers and databases, the linking of data across the web, and the ability to query thousands of databases from a single endpoint. To realize this vision, a set of standards for describing facts, relationships, ideas, and data need to be defined[20]. There are two primary schools of thought on how to achieve this. One is deeply rooted in the traditions of the artificial intelligence field, concerned with building machine-readable representations of knowledge structures, relationships between concepts, definitions of entities, and so forth (broadly speaking, "ontologies"). The second is more deeply rooted in Web 2.0 principles, leveraging the emergence of structures as they are defined on the fly by users - tag clouds, "folksonomy," shared bookmarks, and so on.

The majority of the semantic work to date in the sciences has revolved around the formalization of ontologies. These would establish unique names for entities and relationships that would allow the interconnection of disparate information that concerns the same entity (e.g. molecule or gene). Machine-interpretable languages that describe ontologies and their instantiations would allow computers to map and aggregate relationships (e.g. between a compound and a protein, or a molecular structure and physical properties) from articles, databases, or anywhere else that the network reaches.

Apart from agreements and developments on technical standards, the vision of an integrated knowledge space requires 'social engineering' to ease the sharing of scientific content. To facilitate the declaration of use rights that conform to academic values (such as proper attribution) and encourage the re-use and integration of scientific data, Creative Commons has launched the Science Commons[21] Initiative. It is developing best practices for licensing access and reuse of scientific literature, databases, and materials (samples). To inform its recommendations it conducts proof-of-concept projects in the life sciences that demonstrate the value of Semantic Web technologies for building integrated open access knowledge spaces that serve research in pharma companies, university, industry, and government alike.

2.2.6 Dynamics of Change

Notably, scientific fields differ in the kind of models of scientific communication that materialize and thrive (Cronin 2003, Fry & Talja 2007). The three successful models respectively in physics, biomedicine and the geosciences, mentioned in the introduction, are targeted at different aspects of the scientific communication process. ArXiv is about fast, and unhindered access to the research literature – for everyone, independent of status or personal connections. The notion of democratization drove the initial project of setting up a web-based arXiv service (Ginsparg 1996). GenBank provides public access to

[19] W3C Semantic web Activity http://www.w3.org/2001/sw/
[20] Such as RDF, OWL, and the Object Reuse and Exchange (ORE) Protocol of the Open Archives Initiative,
[21] http://sciencecommons.org/

nucleotide sequences to facilitate the validation of research results and to enable further research, thereby producing a valuable resource for a large-scale data driven science. The interactive journal model of the European Geosciences Union aims to overcome shortcomings of the traditional peer review process. Its originators regarded traditional closed and anonymous peer review as insufficient to ensure quality of published articles and rigour of debate (Pöschl 2004). This suggests that scientific communities differ in their perception of where their existing system fails them and may be improved through new models.

Scientific communication is tightly interlinked with disciplinary practices and cultures. As much as new scientific communication models extend capabilities and support new practices, they also represent continuity by building on pre-existing practices and values. Take the example of high-energy physics where a 'preprint culture' existed long before the arXiv preprint server was created (Till 2001). Preprints were widely disseminated by their authors or their institutes. This practice seems to have gone beyond the occasional exchange of preprints between individual researchers that is practised as well in other fields (including chemistry), as those preprints were systematically collected and indexed in libraries of major research centres in high-energy physics. Thus the creation of a website to facilitate the dissemination of preprints was seen as a natural evolution of the field's communication practices.

As a result, when new models are transferred from one field to another, they may need to be significantly altered before they can become successful. For example, the arXiv preprint model, successful since the early 90's in high-energy physics, failed entirely when attempted in the early 2000's in chemistry by a subsidiary of the publisher Elsevier[22]. Also, the initial proposal for PubMed Central was modelled after arXiv, but the concept had to be significantly altered before it would take-off - no longer as a preprint server but as a repository of published, peer-reviewed journal articles in biomedicine (Kling 2004). Again, this indicates that conditions for the introduction of new models differ across fields and disciplines. At the same time we do see diffusion of new models to other disciplines[23] – so there seem to be sufficient commonalities in some cases to allow diffusion of new models across disciplinary boundaries.

Distinctions such as pure vs. applied sciences seem too coarse to explain the ways fields shape their scientific communication system, and the appropriate level of granularity for analyzing such differences (discipline, sub-discipline, research specialty) is neither obvious nor necessarily uniform across science.

The classification of research fields by a combination of social and intellectual properties developed by Whitley (2000) seems to have some purchase for explaining differences in scientific communication practices. It distinguishes fields

[22] Chemistry Preprint Server (CPS) http://www.sciencedirect.com/preprintarchive
[23] This is exemplified by the growth of arXiv to serve also communities beyond high-energy physics, both within physics as well as in other disciplines such as mathematics, biology, and economics.

of high or low task uncertainty (i.e. agreement on what constitute valid research problems and appropriate methods), and high or low mutual dependency between researchers in order to contribute and achieve recognition (e.g. dependencies resulting from the need to form large collaborations). These field characteristics influence what researchers perceive as effective and legitimate communication practices, and hence influence adoption of new practices such as the ease of data sharing or the acceptance of preprint literature in a field. According to this theoretical framework, data sharing is facilitated in fields with high task certainty and high mutual dependency. In contrast, it is extremely complex if task uncertainty is very high (Birnholtz & Bietz 2003). Further, if in a field both functional and strategic interdependence are high, such as in high-energy physics, then this creates high interdependency between scientists and high concentration of research efforts and goals. This seems to correlate with high interpersonal recognition and high levels of trust and accountability – factors that have been suggested as conducive for preprint use (Frey & Talja 2007).

Another observation concerns how radical a transformation is brought about by new scientific communication models. In spite of the talk about 'a revolution in scientific communication' (Harnad 1991), so far new models of scientific communication as discussed here tend to complement rather than replace existing models. GenBank complements and is integrated with the scientific journal regime (Hilgartner 1995) in that journals require deposit of the primary data an article is based on in the database. Similarly, the use of the preprint server arXiv (Kling 2004) for dissemination does not replace journal publication. Even now, when in some fields in physics the entire community shares its research articles on arXiv, authors still submit those same articles to journals for publication. The reason being, presumably, that in most disciplines[24] the publication of journal articles has become a key metric for career advancement.

Nowadays the reward system in science is increasingly coupled with the formal communication system by relying on journal impact factors and citation measures like the Hirsch index to assess scientific quality and researchers' performance[25]. This development in combination with rigid academic stratification between institutions and within institutions presents a strong reason for scientists to be risk-averse in experimenting with new models of scientific communication (Armbruster 2007). Hence the way the reward system in science is set up presents an inhibitor to any research-driven change in the scientific communication system that focuses on its communicative function rather than its role as a proxy for the assessment of research performance.

---

[24] Highly collaborative research communities such as high-energy physics present an exception. Here the epistemic subject is no longer the individual author but rather the entire collaboration as testified by the hundreds of authors listed on articles. In this area other modes of evaluation for career advancement have evolved (Knorr-Cetina 1999).
[25] A trend of quantization is probably not as strong in the USA as it is in the U.K. as part of the RAE, and in continental Europe, where according to anecdotal evidence the evaluation of a list of candidates for a leading research position may start and end with the comparison of the gradient of change of the candidates' h-indices.

Another important factor would seem to be the role scientific societies take on in supporting new models. Given their mission to support scientific communication and the dissemination of knowledge, they would be expected to take a leading role in exploring the potential of the web for enhancing scientific communication. As publishers of major journals, though, they are affected as much as commercial publishers by threats to their proven business models[26], and their reactions range from constructive cooperation[27] to what has been perceived by many as strong opposition to open access[28]. Hence the dynamics of change in a field may depend critically on the investment its scientific societies have made into the existing system, the positions they take, and the control they can exercise.

In the next chapter we will review the status of new models of scientific communication in chemistry and discuss how these tensions play out. Before that, we briefly turn to a closely related development: e- science or cyberinfrastructure.

2.3 E-Science & Cyberinfrastructure

Recent initiatives to develop new capabilities for research through the integration of advanced computing, information, and communication technologies are mostly known as 'e-science' in the U.K. (Hey & Trefethen 2002), and as 'cyberinfrastructure' in the U.S.A. (Atkins et al. 2003). By providing access to distributed resources such as high-performance computers, large-scale data storage, applications, data, and last but not least people, cyberinfrastructure is proposed to enable new forms of science and to support multidisciplinary research into otherwise intractable scientific problems. Many of these so-called 'grand challenge' problems (e.g., climate change) extend across multiple disciplinary boundaries.

Major e-science and cyberinfrastructure funding initiatives started up in the U.K., the European Union, and the U.S.A. in the early 2000's. The most likely candidates to benefit from e-science approaches are data-driven sciences, such as those targeted by the U.K.'s early pilot projects in particle physics[29], astronomy[30], bioinformatics[31] or chemical combinatorics[32]. More recently, the scope has been

[26] Although not-for-profit organizations, most scientific societies use revenue made with journal publishing or other information services to support other activities in accordance with their mission.
[27] E.g. American Physical Society, and Institute of Physics, UK, who both host a mirror of the arXiv server.
[28] E.g. American Chemical Society (ACS). See (Michaelson 2008, Biello 2007) for details of its opposition to open access, PubChem and the NIH mandate on open access.
[29] http://www.gridpp.ac.uk/
[30] http://www.astrogrid.org/
[31] http://www.mygrid.org.uk/
[32] http://www.combechem.org/

broadened to include the social sciences (Berman & Brady 2005) and humanities (ACLS 2006).

The definitions of what e-science or cyberinfrastructure is about tend to shift (Freeman 2007). Sometimes they emphasize communicational aspects, such as support of distributed collaborations and 'virtual organizations', more often they highlight powerful tools such as high-performance computing and large-scale data storage (Schröder & Fry 2007). In either case e-science and cyberinfrastucture are closely related to scientific communication through the question of how the scientific communication system will be able to support these new forms of research and the anticipated increase in digital content. In order to address, this challenge the term 'cyber-scholarship' has been devised (Arms 2007).

Essentially, there are two reasons to include e-science and cyberinfrastructure in the background section of this white paper on new scientific communication models in chemistry. First, e-science could become an important driver for the creation of new scientific communication models to support 'cyber scholarship'. Second, there is an observable parallelism in attitudes within publishing reform initiatives and in e-science programs. In both, when change is framed only in the context of technology, then strategy and expectations are guided by problem statements and solutions that are defined only in terms of the limitations or possibilities of that technology (Kling et al. 2003, Wouters et al. 2008). In consequence, the social and cultural dimensions of realizing e-science vision are systematically underestimated.

The lessons learned from the first wave of e-science and cyberinfrastructure projects demonstrate the need to think of the 'human infrastructure' (Lee 2006, Berman 2001) as major challenge in e-science:

1. The first lesson concerns the interaction of people working together on realizing cyberinfrastructure and e-science capabilities: The interdisciplinary collaboration between cyberinfrastructure technologists (computer scientists, software engineers, etc.) and domain scientists (biologists, chemists, physicists, etc.) suffers if a technology-centric view prevails. This is captured in the following recommendation from a 'lessons learnt' report to NSF:

> *"Careful technologists will take the time needed to understand fully how users currently work, and why, rather than simply assuming that the innovations they propose are an inevitable improvement… Users understand what they need and, moreover, why they do things the way they do, which is not always apparent to others outside the community. Technologists need to understand this point, as well to understand that most researchers are not technologically naïve. In other words, cyberenvironment development is a two-way street—users need to be able to describe to technologists how they work, and technologists need*

to be able to explain to users how a community cyberenvironment can enable them to do even more." (Spencer 2006 et al.).

2. A second lesson concerns the awareness of the type of problems that need to be solved: 'soft issues' suffer from a lack of attention and turn into major stumbling blocks for the success of e-science projects. Examples are security and authentication, anonymization of social science micro-data, usability, uptake and use, contractual arrangements for inter-institutional and cross-sector collaboration, and trust between stakeholders (Schroeder & Fry 2007).

3. A final lesson learned concerns the interdependencies between technical and social solutions: layer-models of technical systems[33] that are very popular in software engineering suggest a clean separation of technical and social sphere. In reality, the boundary between technical and social is flexible and in constant negotiation (Edwards et al. 2007). Hence, instead of thinking of a one-directional impact of cyberinfrastructure on science and research practices, e-science technologies need to be understood as co-constructed with the research environment in which they are developed and deployed, e.g. "What the technology is, who the scientists involved are, and what they aspire to achieve are co-produced" (Hine 2008).

2.4 Conclusions

Observations both from publishing reform initiatives as well as from e-science projects confirm that the availability of new technologies is not enough to generate fundamental and widespread change. Instead we may expect a field or discipline-specific co-construction of what research and scientific communication in a discipline is about, and what technologies and services serve it best (Hine 2008). To find acceptance, new models for scientific communication need to fit well with research practices and community values. Hence to assess the value of a new model a simple needs-analysis that focuses on functional capabilities is not enough. Instead a deeper understanding is needed of the socio-technical system of scientific communication in the scientific domain in question.

[33] For e-science they typically depict a grid layer at the bottom, then a layer of middle-ware, and a top layer of user specific applications. See Kling et al (2003) for a critique of these models in the context of scientific communication models.

3 The Scientific Communication System in Chemistry

The first part of this section provides an overview of established elements in the scientific communication system in chemistry. The second part of this section reviews the status of recent web-based innovations to scientific communication in chemistry.

3.1 Established System

We will focus our discussion on scientific journals, which are central for the record of science in chemistry, and databases, since comprehensive access to the chemical literature and information on chemical substances is crucial in most areas of chemical research.

In addition to the primary and secondary literature types of journals and databases, we briefly mention here further important elements: tertiary literature such as monographs, encyclopedia that also represent important knowledge resources, and workshops and conferences that are part of informal scientific communication. Smaller meetings are valued for intensive discussions, and the many Gordon conferences dedicated to chemical topics indicate their success within the chemistry community as a model of informal communication. Larger society meetings are important to gain visibility, stake out ones territory, and engage in social networking. Scientific results get communicated at conferences in talks, abstracts, and posters, but the enduring scientific record gets established through publication in peer-reviewed journals.

3.1.1 Chemistry Journals

The cumulated number of chemistry articles roughly doubles every 14 years[34] (Behrens 2006), and CAS currently processes each year about 700,000 abstracts in chemistry and related fields (Hoffmann 2007). The two largest non-profit publishers in chemistry are the scientific societies the American Chemical Society (ACS) and the Royal Society for Chemistry (RSC) who together publish about 35,000 articles per year (Garson 2004). Although commercial publishers publish 94% of the chemical literature, the most prestigious general chemistry journals are

[34] There exists disagreement in the literature whether the growth of scientific literature since World War II is best described by a an exponential model which has started levelling off in the 80's (Laviere 2008, Price 1963) or whether it is better described by a quadratic growth model (Behrens 2006) with a constant acceleration and whether the supposed decrease in growth rate is just an artefact of the use of the exponential model. Note that underlying this global trends are local, field specific trends, where literatures of specific subfields follow a life-cycle evolution of accelerated growth, stabilization and decline, and may show over a time period doubling times of 3 years and less (Braun et al. 1977).

society journals, the *Journal of the American Chemical Society* (JACS), and *Angewandte Chemie* of the German Chemical Society (GDCh), published in cooperation with the commercial publisher Wiley VCH. JACS publishes twice as many articles than the latter, and is the most cited chemistry journal in terms of total citations, whereas the Angewandte has the highest impact factor for a general chemistry journal publishing original research articles. This spring, Nature Publishing Group launched *Nature Chemistry* - clearly positioned as a competitor to the other two top chemistry journals[35].

To an author who seeks an appropriate forum for publishing his or her research, the journal system in chemistry exposes a 3-layered structure: of highest public-relations value is 'getting a publication into' *Science* or *Nature*; one step down on the ladder of public attention, but still very prestigious within chemistry, are publications in *JACS* or *Angewandte*. In those two journals about 4,500 articles get published annually. Especially for younger chemists aiming to build an academic career, publishing in these general chemistry journals with high visibility is very attractive. But also senior scientists may seek to assert their reputation by having a large number of publications in either venue[36]. The large bulk of published chemical information appears in hundreds of chemistry journals that are specific to a subdiscipline such as organic chemistry or physical chemistry, or are dedicated to a smaller, specialised research field. Whereas the prestige journals only publish very short, condensed reports, the longer articles that are published in specialist journals provide comprehensive scientific evidence for new discoveries. The content of these journals represents a crucial accumulated archive of chemical knowledge of enduring value to current research in chemistry. Within each field a further stratification of journals exists. Submission decisions depend on a number of factors like the type of results one plans to report[37], the balance between communicating widely, and targeting ones message to a specific community of specialists, as well as social obligations[38].

Journal use habits seem to be a little more agnostic to journal reputation. Although the top journals are more regularly browsed for 'important' science, an important entry point for literature researchers are literature databases, such as SciFinder or Web of Science that provide equal access to a wide range of chemistry journals (Schummer 1999). In addition, recommendations of specific

---

[35] See blog post of the chief editor of Nature Chemistry at http://blogs.nature.com/thescepticalchymist/2009/02/chemistry_countdown_complete.html (23 February 2009)

[36] A tendency favored by the journals: e.g. Angewandte publishes on its website a list of authors ranked by their total number of publications in the journal http://www3.interscience.wiley.com/journal/26737/home/2002_mostfreqauth_62-08.pdf

[37] Some journals are particularly suited to short communications with the main aim to add data sets with minimal theoretical analysis and interpretation to the scientific record.

[38] e.g. to contribute to a special issue honouring a senior researcher, or to submit to the journal edited by a colleague.

articles (or monographs) by co-workers who are aware of your immediate research interests are highly valued[39].

Co-authorship has become very common for original research papers in chemistry, with only 1% of papers being single author papers (Cronin 2004). On average papers are published by 3-4 authors, a number that may further vary between sub-disciplines as collaboration patterns differ[40]. The smallest unit of publishable research, that is the typical article length, is between 2-8 pages.

ACS has been one of the first chemistry publishers to experiment with electronic versions of research articles already in pre-web times in the 80's (Garson 2004). Nowadays all major chemistry journals are available online. They mostly reproduce the paper-based article in electronic form by offering pdf files for printing and sometimes an HTML version of the article. The online versions of the journals play to the strong visual orientation of chemists for processing information (Hoffmann & Laszlo 1989) by offering graphical abstracts in the table of content listings (see Figure 1).

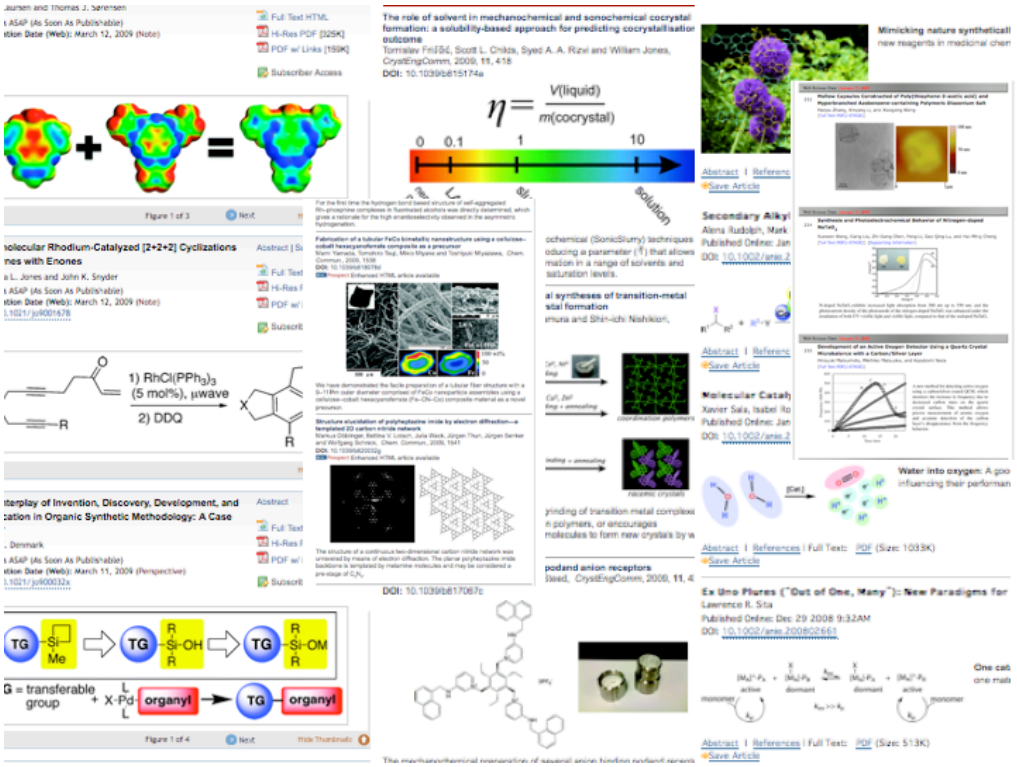

Figure 1: Visual abstracts in the online versions of chemistry journals (screenshots from (1) The Journal of Organic Chemistry/ACS, (2) ChemComm/RSC, (3) CrystEngComm/RSC, (4) Chemistry Letters of Chemical Society of Japan, (5) Angewandte Chemie/GDCh)

[39] [unpublished] Observation in ongoing field study of communication cultures in chemistry by Velden & Lagoze.

[40] This is suggested e.g. by the differences in international collaboration found between analytical chemistry and organic chemistry by (Glänzel & de Lange 1997).

Electronic publishing should increase the speed of the publication process (Garson 2004), not just for online-only journals but also for print journals[41]. So far, the specific advantages of online-only journals have found little uptake among chemists. RSC launched the online-only journal ChemPhysChem in 1998 in order to take advantage of the web for speedy publication and to include colour images, animated graphics and movies in their article. At the end of 2003 the journal was discontinued. RSC explained that the submitted articles contained fewer features for electronic enhancement than they had expected, and that print journal production had sped up to such an extent that "It has become clear that there is now no real advantage to an e-only journal in the physical chemistry field."[42]

The competitive push to further increase the immediacy of publication seems to come from authors' experiences with the practices of society publishers in the life sciences many of whom have adopted programs to release manuscripts upon acceptance. ACS has responded to author demand by introducing the web-release of peer-reviewed accepted manuscripts in a pilot project for selected biochemistry and pharmacological journals.

Certification of articles through peer review organized by journals is clearly highly valued in the chemistry community. Peer review is seen as preventing factual mistakes (if not always), and enhancing the editorial quality of articles. It has been suggested that chemists perceive the web medium with the suspicion that it may endanger the integrity of the scientific record: new web–based publishing models might undermine rigorous peer review, facilitate the manipulation or misuse of the electronic article copy, and fail to ensure perpetual access (Searing & Estabrook 2001).

3.1.2 Chemistry Databases

Since the 60's, high-quality, comprehensive databases have been serving the chemistry community both in academia and in industry. These databases can be broadly classed into three categories of information: literature and patent databases, structure and reaction databases, and factual databases (chemo-

[41] The bulk of the increase in the speed of publication is probably attributable to the use of electronic submission systems. Kling and Swygart-Hobaugh (2002) found that the Journal of Biological Chemistry and Journal of the American Chemical Society decreased their publication delays respectively by 70 and 47 days on average, between 1980 and 2000 through the use of Internet based tools. However, Kling and Swygart-Hobaugh also found that publication delays at several journals in other disciplines such as economics increased during this time.

[42] Editorial, Phys. Chem. Chem. Phys., 2003, 5, 24-v-xvii, DOI: 10.1039/b309812p, Accessed online on 12 March 2009 at http://www.rsc.org/delivery/_ArticleLinking/DisplayHTMLArticleforfree.cfm?JournalCode=CP&Year=2003&ManuscriptID=b309812p&Iss=24

physical properties incl. spectra). A number of databases provide access to several of these information types. Table 1 lists some of the widely used databases in chemistry. For a more detailed overview on chemical databases see (Engel & Zass 2007).

Type	Name	Main Content[43]
Literature & patents	Chemical Abstracts CA (CAS)	> 28 million publication records from chemistry, biochemistry, chemical engineering from 9,500 journals plus patents, proceedings, theses etc.
	BIOSIS (Thomson)	15 million publication records from biosciences, biomedicine from 5,000 journals and US patents
	MedLine/PubMed (US National Library of Medicine)	14 million publication records from 4,300 journals
	Science Citation Index SCI (Thomson)	> 30 million publication records from science, technology, medicine from 3,700 journals (5,800 in expanded version)
	Derwent World Patent Index WPI (Thomson Derwent)	14.5 million patent families
	INPADOC (European Patent Office)	34 million patent families
Structure & reactions	CAS Registry (CAS)	28 million organic and inorganic compounds plus ~ 57 million peptides, proteins, nucleic acids
	Beilstein (Elsevier)	9 million organic compounds
	Gmelin (GDCh/Elsevier)	2 million inorganic and metal-organic compounds
	CASREACT (CAS)	> 10 million reactions
	SPRESI (InfoChem)	4.5 million organic compounds, 4.5 million organic reactions
	Inorganic Crystal Structure Database ICSD (NIST, FIZ Karlsruhe)	~ 100,000 inorganic crystal structures (as of 3/2009)
	Cambridge Structural Database CSD (CCDC)	~ 400,000 crystal structures of small organic molecules and metal-organic compounds (as of 3/2009)
	Protein Databank	~ 50,000 crystal structures of proteins, nucleic acids, and complex assemblies (as of 3/2009)
Factual	Beilstein (Elsevier)	400 different data fields for spectra, thermo-chemical data etc.
	Gmelin (GDCh/Elsevier)	200 different data fields for spectra, thermo-chemical data etc.
	NIST Web Book	Physical and spectroscopic data for ~70,000 species (as of 3/2009)
	SpecInfo (Wiley VCH)	150,000 organic compounds, NMR, IR and mass spectra

Table 1: Examples of major widely used databases of chemical information

[43] as given in (Engel&Zass 2007), unless otherwise noted.

Access to databases is vital in all areas of chemical research. Common scenarios of database use include:

- Planning the synthesis of a chemical substance: The novelty of the intended product is ensured by searches in literature, patent, and structure databases, and the planning of synthesis routes is supported by lookup in reaction databases.
- Characterizing a chemical substance: To relate the chemical composition of reaction products to their molecular structure one obtains results from many different types of measurement such as NMR, IR, mass spectroscopy, x-ray diffraction. Besides these routine methods applied in molecular chemistry, many more methods are available requiring more specialized equipment and user training. The observed properties are then checked by the synthetic chemist against the properties of known substances in the respective factual databases.
- Understanding structure function relationships: For curiosity-driven research into the relationships between molecular or supramolecular structure and functionality, extensive information on functional properties (optical, thermal, solution, electrical, magnetic, biological activity) as well as electronic structure information is needed.
- Design of functional material: To design novel materials with specific functionality from molecules, several material properties of these molecules need to be known. Often some properties are not captured in databases, but can be calculated from structural and related property information using algorithms based on quantum mechanics or mean field approximation.

CAS, as well as Beilstein and Gmelin, are widely used by most practicing chemists, and in particular synthetic chemists. They have a century long history, and have been built into valuable resources deeply rooted in chemists' research practices. In particular access to the CAS structure and patent databases is critical to double check the potential novelty of a substance in the context of patenting which makes them into indispensable resources that are licensed (from CAS and Elsevier respectively) especially by industry at considerable costs.

Content creation

The number of known chemical substances reported in the literature has grown exponentially since the second half of the 19th century; the doubling time for organic substances is about 13 years, that of inorganic substances almost 24 years. Today, on average almost 2 new substances are reported in every chemistry paper (Schummer 1997a). Most data on chemical substances in the databases are extracted from the published journal or patent literature. Depending on the level of indexing and validation this is very labor-intensive expert work.

Crystallographic databases deviate somewhat from this post-publication extraction model. In crystallography, experimental data is highly standardized. An estimated 95% of crystallographic data is in a particular standard format, the Crystallographic Information File (CIF) (Swan 2008). Crystallographic databases

support 'private communications', that is, deposit of a data set by its creator independent of a journal publication accompanying it. The percentage of content acquired by direct deposit e.g. in the CDS database seems to be rather low though (Allen 2004). Crystallographic databases further have agreements with many journals to act as repository for primary data accompanying articles and hence authors are requested to deposit their crystallographic data directly in the database (Allen 2004, Swan 2008) and to note the database acquisition number in the article. Still the majority of content also in crystallographic databases is harvested from journals post-publication and consequently the content of the databases mostly corresponds to what has been published in the literature. It is estimated that only 20% of the crystallographic data produced in research laboratories is publicly released in databases (Coles 2005).

Access

Good access tools are important to find relevant information in databases. Even for the same databases, vendors offer different interfaces such as STN Messenger, SciFinder, and SciFinder Scholar that differ in the sophistication of query mechanisms and the extent of access that they offer to the information contained in the database or an assembly of databases. Initially, in the early 70's, database retrieval was text-based. Today, visual representations of structures or substructures can be drawn and submitted as searches.

Of particular value are visualization tools for inspecting results, e.g. the three-dimensional representation of a molecule, because a molecule's spatial arrangement is indicative of its physical properties - "for chemists […] three-dimensional representation of shape may be a matter of life or death" (Hoffmann 2007).

Integration

Since a chemical substance is often described by different names, a key role for integration of information from different databases is the CAS registry number, a sequential number that uniquely identifies each new chemical substance as it is indexed by the CAS Registry database[44]. Using this number a chemical substance and information pertaining to it can be looked up and connected across diverse databases. As the owner of the registry database CAS holds a monopoly on assigning and granting use of CAS registry numbers, and permits third party services to use up to 10,000 such numbers without licence or paying a fee. Thereby it can control and restrict the efforts of any third party service (and potential competitor) that would like to make use this 'gold standard' for the identification of chemical substances exceeding the 10,000-limit.

Given this situation, an alternative, non-proprietary identifier has been proposed to enable unrestricted global chemical information integration. It is called InChI and has been introduced by the International Union of Pure and Applied

[44] The CAS Registry was set up in the 60's when electronic data processing was introduced by CAS to cope with the growing volume of primary literature, and has become available as a structure searchable database in 1980 (Engel & Zass 2007).

Chemistry (IUPAC). It differs from CAS registry numbers in that it is derived from the structure of a chemical compound. It covers organic molecules, as well as inorganic, organometallic and coordination compounds[45]. Since CAS numbers are so widely used, a major challenge for adoption of the InChI is to gain visibility and wide acceptance. InChIs have found support by NIH, NIST, Thomson (SCI), Chemspider, Nature, RSC, and Elsevier who have started using InChIs in their databases or publications. Further, RSC and Chemspider have announced in December 2008 that they will cooperate in providing an InChI key resolver service[46]. Unfortunately the world of chemistry is complex and InChIs (since derived from structural formula) work only well in those cases where there is a 1:1 correspondence between molecule structure and chemical compound, and to know when that is the case is not trivial (Murray-Rust 2009). So the question of an open, universal chemical identifier authority remains unsolved.

3.2 Recent Web-Based Innovations and Experimentations

In this section we provide a brief overview of the status of recent efforts to use the web to enhance the communication and management of chemical information. Generally it can be said that these are pioneering efforts that explore possibilities and provide proof of concept. So far they have not found a wide uptake such that they would change or extend scientific communication practices in chemistry in a fundamental way. A major hurdle, according to many of the pioneers, is the proprietary regimes in which most chemical data reside that restrict access to and the integration of chemical information.

3.2.1 Semantic Chemistry Web

The vision for a Semantic Web in chemistry is driven by interests in large-scale data mining in cheminformatics to support drug discovery (Neumann 2005, Tetko 2005), as well as the perceived benefit of integrating work of small scale labs by better managing their diverse data. The semantics aid the discovery and reliable re-use of data, and reduce ambiguity for later automatic processing of data (Frey 2009).

To realize a Semantic Web of chemistry information one has to overcome at least three challenges: First an engineering problem - how to design the structures and systems that allow semantic markup of chemistry content on the web. Good progress has been made in developing a markup language for chemical information. The Chemical Markup Language (CML) developed over the last decade by Murray-Rust, Rzepa, and colleagues covers a wide variety of chemical information, molecular structures, material structures, spectroscopic, analytical, crystallographic, and computational data. It is complemented by

[45] InChI project homepage at http://www.iupac.org/web/ins/2008-033-1-800

[46] RSC Press release at http://www.rsc.org/Publishing/News/RSCandChemSpiderdevelopInChIResolver.asp

further specialized mark-up languages for analytics, thermochemical and thermophysical property data, mathematics or reactions[47].

Less far developed is the next layer consisting of RDF triple statements connecting marked-up entities and adding semantics to their relationships. Both RDF (a generic machine-readable framework for encoding semantic information) and the development of chemical ontologies are still in their infancy (Adams 2009). Furthest developed is ChEBI (Degtyarenko et al. 2008), a free dictionary whose name indicates its orientation towards the life sciences: "Chemical Entities of Biological Interest"[48]. Nevertheless some exemplary projects demonstrate the feasibility and value of modelling aspects of chemistry to enhance the communication and reuse of chemical information (see green box below). ChemAxiom[49], which aims to become a broad and deep chemistry ontology is currently under development.

Second, there is the relative novelty and immaturity of the field and, in particular, its supporting technologies. The construction of useful semantic technical standards and systems is non-trivial technically, requiring much diligence in the choice of unique technical names for entities, ensuring that relationships are logically consistent, and that the knowledge represented is accurate. There is no equivalent of WYSIWYG, or a web browser, for semantic knowledge, and querying semantic information can be a high barrier for the naive user. Advocates of Semantic Web solutions in chemistry concede that they are not yet competitive to well-established and more scalable solutions to current problems, but argue that the versatility of Semantic Web enabled resources and data and their ability to integrate chemistry data with that of related disciplines (e.g., life sciences) will have the edge for providing cost-efficient solutions to new problems, as soon as more general-purpose tools are widely available (Frey 2009). Recently, Microsoft Research has initiated research in this area, aiming to encourage the development of

CAS in the 'New Information Order'

"Reflecting on our own experience, my observation would be that compared to twenty or more years ago, the most advanced technologies for publishing are much more ubiquitous and available. Technology tends not to distinguish information providers today; rather, the combination of technology/content and its application or market specialization does."

Robert J. Massie, President CAS
(Feb 25, 2008, Miles Conrad Lecture)

[47] See Adams (2009) for an overview and pointers into the literature of mark-up languages.
[48] Online information at http://www.ebi.ac.uk/chebi/
[49] Development home page at http://bitbucket.org/na303/, motivation explained at http://wwmm.ch.cam.ac.uk/blogs/adams/?p=195

open standards that provide support for chemical information and scientific communication processes and leverage them in Microsoft products. The oreChem project[50], funded by Microsoft Research, is an international, multi-institutional effort to develop and deploy the infrastructure and tools for Semantic Chemisty.

The third challenge is chemistry-specific, non-technical, and represents a critical barrier to overcome if semantic technologies are to bring benefits to chemistry. That challenge is the access problem. To make a semantic chemistry web valuable for chemists, a huge effort has to be spent to retrofit the existing knowledge into new formats. This is a significant challenge - chemistry is an old, complex field, with knowledge in multiple languages, document formats, document types, spanning a long time of relevance. And since the technologies are still evolving, a variety of approaches to that retrofit have to be tried before the one that serves chemists best can be identified - the ontology approach, the tagging approach, or a mix of the two.

Unfortunately, the vast majority of accumulated chemistry knowledge is locked up behind pay firewalls - the field "has ceded the dissemination of data and knowledge almost entirely to commercial entities in the form of publishing businesses" (Adams 2009) in a way that many other scientific disciplines have not. To the extent that publishing companies hold valuable content, they are key players in any retrofit of existing knowledge into new formats. Unfortunately some of those entities perceive control over content as their key competitive advantage, as the strategic assessment from the President of CAS, Robert Massie, of the role of CAS in the 'New Information Order'[51], indicates (see insert). This places the publishing industry in sole control of any retrofit with the potential to disable the ability of entrepreneurs and academics to experiment on the retrofit to semantic technologies.

This distinguishes chemistry from the bio-medical field, where the quantity of available data and the scale of public funding is much greater. Biology has evolved a significant set of norms that almost mandate data sharing at many levels. For example, the Bermuda Principles require the rapid release of genomic data from large scale sequencing projects. The organizations involved (the Wellcome Trust and the NHGRI) sought the involvement of key journals to mandate data deposit as a part of publications, and implemented funding requirements of openness for published scholarly articles[52]. Consequently web services and Semantic Web technologies flourish in this domain.

---

[50] http://research.microsoft.com/en-us/projects/orechem/

[51] This was the title of the conference where Massie was invited to deliver the Miles Conrad Lecture.

[52] The National Human Genome Research Institute's Rapid Data Release Policy: http://www.genome.gov/10506376

3.2.2 Semantic Publishing

A showcase for the semantic enrichment of journal publications is the Prospect Project[53] by the RSC. This project provides access to enhanced articles (Batchelor 2007) in which a reader can select different highlighting options that display in the text in different colours entities including gold book terms (IUPAC Gold Book), chemical terms (ChEBI), biology terms (gene, sequence, and cell ontology), and compounds. These highlighted terms are then linked to further information. For example, clicking on a highlighted biology term provides a definition of the term, a link to the ontology, and further links to related articles on the topic. Clicking on a highlighted compound provides additional information such as the compound name, any synonyms, the InChI full identifier and the InChI key, SMILES strings for the compound, a downloadable CML file, a 2-D graphic of the compound, a list of other RSC enhanced articles that contain that compound, and links to find the identified compound on PubChem, or within SureChem's patent database.

In spite of a considerable amount of automation through language processing tools, producing such enhanced articles requires qualified curatorial staff with significant domain knowledge in chemistry. Hence the question remains how such efforts will scale across the entire journal literature. The problem would be eased if the authors, who have the best domain knowledge with regard to their own publication, would contribute to the mark-up of chemical information, and would be enabled to do so painlessly with easy-to-use tools.

The Chem4Word project is a collaboration of scientists with Microsoft working to develop applications to support authoring documents in chemistry[54]. As a first step in this direction, Microsoft is releasing plug-ins for Word 2007[55] that support reading and writing of XML documents that follow the standard of the National Library of Medicine that is used by publishers and PubMed. These plug-ins will also support the ontology-based semantic mark-up of named entities (Shotton 2009).

3.2.3 Electronic Lab Notebooks and Open Notebook Science

Electronic lab books help to capture data at the source, when they are created in the laboratory. The Semantic Electronic Lab Notebook (ELN, see insert below), pursues a Semantic Web approach in the capture of data and facilitates internal management and eventual publication of the data created (Hughes et al. 2004). The emphasis in this approach is on capturing the data during the research process in digital format with as little effort for the researcher as possible (e.g. by having instruments providing the data in standard formats that can be read into the ELN). This way a complete as possible provenance trail of data

[53] http://www.rsc.org/Publishing/Journals/ProjectProspect/index.asp
[54] Project info at http://research.microsoft.com/en-us/projects/chem4word/
[55] Download from http://tinyurl.com/5szjly, http://tinyurl.com/abc4c7

could be established that then later would facilitate re-use of the data as well as public dissemination if so intended.

The vision of Open Notebook Science (Bardley et al 2009) is to make the entire primary record of a research project available almost in real time, as it is created. Its advocates expect an increase in efficiency of the scientific process through greater transparency: through provision of all contextual details of measurements, through sharing of information on failures and negative results that may avoid duplication of vain efforts, through triggering unanticipated collaboration between partners whose research interests align or complement one another, and through development of a collaborative community of scientists that share data and may join forces in analysing results and deriving conclusions. To get started on this vision of open notebook science even without advanced software to support chemical information intelligently, the UsefulChem project[56] at Drexel University started in 2005 pragmatically by combining existing Web 2.0 tools such as free, hosted blog and wiki services to manually capture and publish data of their anti-malarial compound synthesis and testing studies on the web. In this way they have found collaborators across the world to work with in their research efforts. In another project, they integrated web services within Google Spreadsheets to calculate solubilities directly from NMR spectra[57]. The Wikipedia page on "Open Notebook Science" [58] keeps track of the groups using open notebook science.

3.2.4 Data Publishing

The idea of data publishing is to make experimental data publicly available in a way that they a) can be referred to and cited by a stable identifier, and b) are provided in a re-usable format[59]. This would increase the value of experimental data by allowing the scientific community e.g. to double-check a result that has been reported in the literature, to conduct alternative analyses of the data possibly not anticipated by the creators of the data, to recalculate quantities from the data when new calculation methods become available, or to aggregate data for data mining.

Several variants of making data available on the web for re-use have been tried. One approach has the goal of making data published in the literature more easily accessible, e.g. either by encouraging authors to include data in standardized formats in the supplemental sections of articles, or by using machine learning-based automated methods to extract underlying data from tables and figures in the text (see Liu 2007). Another, proactive approach (in

---

[56] Home page of the UsefulChem project at: http://usefulchem.wikispaces.com/
[57] For details see: http://usefulchem.blogspot.com/2009/03/semi-automated-measurement-of.html
http://usefulchem.blogspot.com/2009/05/streamlining-automated-solubility.html
[58] Info at http://en.wikipedia.org/wiki/Open_Notebook_Science
[59] This implies e.g. for spectra to not just represent the graph as an image in bitmap format but to provide the actual underlying data points.

contrast to post-hoc extraction) conceives of data sets themselves as publishable, independent of an accompanying article. The latter approach would potentially increase the amount of data available, since the need to create an article before a data set can be publicly released means that only a small fraction of the data created in laboratories is ever made publicly available.

Post-hoc extraction of data from published literature

One way of making the data that is dispersed in the journal literature more visible is by aggregating this data through web-based harvesting and extraction of data sets from journal websites. This works well in a domain like crystallography where with CIF a global standard for the formatting and description of crystallographic data (Hall 1991) exists and is widely used (Murray-Rust et al. 2004). In 2006 the IUCr was awarded the ALPSP Award for Publishing Innovation for their efforts in creating CIF and checkCIF, an online service operated by the IUCr that checks the consistency and integrity of crystal structure determinations reported in CIF format. Through the agreement on a standard data format and automated checking of consistency of such files a certain level of quality control becomes possible that is not always achieved through human peer review and manual abstraction of data from published articles.

Crystal Eye is a web-based service that demonstrates the benefit of such a standard format for data publishing[60]. This fully automated service harvests CIF files from journal web pages, converts the information to CML, and provides web pages for easy browsing of the crystallographic data, including 2-dimensional and 3-dimensional rendering of the structures using the open source rendering software CDK and Jmol. It also provides bond length histograms derived from the accumulated data that link back to the underlying data sets, and publishes RSS feeds to provide a news feed on newly published data.

Nevertheless, this model of post-publication data extraction to support data re-use has some problems, even for crystallographic data. Not all journals publishing articles that contain crystal structures expose CIF files of these structures, and of those that do, not all allow a web-based robot to automatically harvest the CIF files from their journal websites (IUCr and RSC do allow automated harvesting). Some publishers such as ACS claim copyright on supplemental data files published along with the articles[61]. Whereas the original data submitted by the authors as such are facts that do not underlie copyright, the specific representation of such data e.g. in a CIF file formatted by a publisher is subject

[60] Home page of CrystalEye at http://wwmm.ch.cam.ac.uk/crystaleye/

[61] A typical statement in the online version of an article in JACS reads: "Electronic Supporting Information files are available without a subscription to ACS Web Editions. All files are copyrighted by the American Chemical Society. Files may be downloaded for personal use; users are not permitted to reproduce, republish, redistribute, or resell any Supporting Information, either in whole or in part, in either machine-readable form or any other form."

Showcases of Semantic Chemistry Web

Semantic Electronic Lab Notebook – *Capturing chemical data when created*

The Southampton Semantic ELN was built as part of the CombeChem e-Science project (www.combechem.org, Taylor 2006) to demonstrate the use of a lightweight semantic model in the planning and execution of a synthetic organic chemistry project. The unique feature of this ELN software (http://smarttea.org) was the use of the language of the Semantic Web to record information on both materials and processes and the links between them. In the project, advantage was taken of the advance planning and recording of experiments, which is necessary in order to comply with the UK regulations on the control of substances hazardous to health (COSHH). This plan was used to produce a digital framework for the experiments that acts both as a guide to the sequence of experimental steps in the laboratory and a framework on which to hang the record of the observations. The ELN exists 'in the cloud' and is accessible via a web interface (e.g. for planning and recall) and via a tablet interface for ready access and recording in the laboratory during the experiments. The ELN has been very well received by those using it in the laboratory, and has led to a significant increase in the quality of the notebooks, and in the parallel or subsequently used paper notebooks. It has led to much clearer and more thoughtful approach to planning and performing experiments." (Frey 2009). A more general approach using the Blog metaphor has also been deployed to facilitate and study collaborative experimentation and discussion enabling the linking of laboratory experimental data, information and discussion.

OSCAR 3 & Project Prospect– *Extracting chemical information from articles*
http://oscar3-chem.sourceforge.net/
http://www.rsc.org/Publishing/Journals/ProjectProspect/

OSCAR3 is an open extensible system for the automated annotation of chemistry in scientific articles, which can process thousands of articles per hour. It attempts to identify chemical names including some enzymes and reaction names, ontology terms, and chemical data. Where possible, the chemical names that are detected are annotated with structures, either via lookup or name-to-structure parsing ("OPSIN"), and with identifiers from the chemical ontology ChEBI. RSC has built on OSCAR in their reward winning Project prospect that publishes selected, semantically enhanced articles from all RSC journals (Corbett 2006).

AnnoCryst – *Collaborative annotation of crystal structures*

"The aim is to enable geographically distributed teams of crystallographers and chemists to collaboratively discuss, compare, assess, and make comments on macromolecular crystal structures either before or after they have been published in public online databases. The AnnoCryst system enables annotations to be attached to 3D crystallographic models retrieved from either private local repositories (e.g., Fedora) or public online databases (e.g., Protein Data Bank or Inorganic Crystal Structure Database) via a Web browser. The system uses the Jmol plugin for viewing and manipulating the 3D crystal structures but extends Jmol by providing an additional interface through which annotations can be created, attached, stored, searched, browsed, and retrieved. The annotations are stored on a standardized Web annotation server (Annotea), which has been extended to support 3D macromolecular structures. Finally, the system is embedded within a security framework that is capable of authenticating users and restricting access only to trusted colleagues." (Hunter et al 2007)

to copyright, and a publisher can restrict re-use of such data files (Swan 2008). To ensure a seamless flow of factual information, publishers would need to adopt more liberal policies, and allow authors to regulate access to supplemental data published along with their article through permissive licence for data re-use as promoted by the Science Commons Project. Such liberal data re-use policy by publishers is supported e.g. by the Association of Learned and Professional Society Publishers (ALPSP) in their declaration on Open Data[62]:

In summary, the core requirements for this approach to data publishing are:

- Declaration of re-use rights for data sets in the published literature
- Cooperation of publishers to allow automated harvesting
- Agreement on standardized open formats for data and routine use of such formats by authors and publishers

Direct deposit of data in a repository

An alternative to the post-hoc approach of extracting data from the published literature is to make data sets publicly available through direct deposit in a web-based repository by their creators. In this manner, data sets are published by themselves, independent of or in parallel of a journal publication, thereby increasing the amount of available data significantly.

For example, in crystallography the x-ray diffraction image, that is the raw data from which the 3-dimensional structure that is encoded in a CIF file is derived, is not routinely published in the journal literature. Also those data files are so big that central storage is not necessarily a scalable solution. Therefore the Australian TARDIS project[63] has created a federated service, where the information on the available diffraction images is provided on a central server, but the actual data files reside on local repositories run by universities and other research institutions (Androulakis et al. 2008). This approach is still in an early development stage and not yet deployed for full production use.

Data repositories that focus on crystal structures are eCrystals in the U.K. and ReciprocalNet[64]in the U.S.A., and the Crystallography Open Database[65] (COD). eCrystals serves as a web-based archive of crystal structures generated by the Southampton Chemical Crystallography Group and the EPSRC U.K. National Crystallography Service. Each data set submitted to the archive gets a globally unique Digital Object Identifier (DOI) and a web page is generated that displays the data, including a 3-dimensional rendering using Jmol. The archive contains only a couple of hundred structures, making it sub-critical in size at this point. ReciprocalNet is a distributed database network providing access to crystal structure data from crystallographic service facilities of about a dozen universities mostly in the U.S.A. To allow parallel publication in a journal, sometimes a delay

[62] http://www.alpsp.org/ForceDownload.asp?id=129
[63] TARDIS project info at http://tardis.edu.au/
[64] Home page of ReciprocalNet at http://www.reciprocalnet.org
[65] Home page of Crystallography Open Database at http://www.crystallography.net/

of a year or more can occur between creation of the data and its public release in the database. It is part of the National Science Digital Library (NSDL) and emphasizes applications in the education domain, and it hosts a prize-winning collection of crystal structures of 'common molecules'. The Crystallographic Open Database is more rudimentary in how it represents the data - it simply offers CIF files for download. Its originators understand their service as a light, open-access version of the proprietary databases CSD, ICSD, Crysmet and ICDD[66]. It has currently about 75,000 entries, and is complemented by databases for predicted crystallographic structures (PCOD) and predicted powder diffraction patterns (P2D2) with more than 100,000 entries each.

Data that is part of a journal publication will have undergone a certain level of validation as part of the journal peer review, whereas for independently published data sets different mechanisms for quality assurance may need to be found. In crystallography the checkCIF service may minimize this problem, but for other type of areas this may well be a concern that needs to be addressed in other ways. It has been suggested that 'open' databases may benefit by allowing users to report and flag errors or inconsistencies in the data for review or even participate in real-time curation of the data (Williams2008a).

In summary, the core requirements for this approach to data publishing are:

- A sufficient level of curation to ensure data quality and consistency
- Sustainability of the data repositories
- Motivating scientists to deposit their data.

3.2.5 Finding Chemistry Data on the Web

As the amount of open web-based chemical data grows, search engines and web crawlers that aggregate chemical information and provide services on top of these different resources become more relevant. An example for a chemistry search engine under development is the NSF funded project ChemXSeer[67], a chemistry search engine modelled after the CiteSeer search engine for computer science literature, but with added intelligence to deal with chemical information. It identifies chemical entities (names, formula) in full text documents, and extracts data from tables in documents, using novel ranking functions to display the search results. Another example is ChemSpider, a free online search engine for chemical structures. The central aim for this search engine is to answer the question "is there specific information about my chemical" and to provide access to it where it is open access, or link to the commercial provider where access is restricted. ChemXSeer seems to be technically more ambitious, trying to extract data from the legacy literature (i.e. full text documents in pdf format), ChemSpider in its core is a more lightweight aggregator expecting formatted data – though it is configured to integrate further services and hence may well

[66] The International Center for Diffraction Data provides a database on powder diffraction data http://www.icdd.com

[67] Project homepage at http://chemxseer.ist.psu.edu/

be extended in the future, in particular since it has recently been acquired by RSC.

Ultimately, both services will rely on the quality of data out there on the web and in open databases. The builders of ChemSpider are well aware of the problem of data that is not carefully manually curated. Aggregating data from different sources is challenged by conflicting or ambiguous identifier assignments and the fact that structures are commonly incorrect. The ChemSpider vision is to use robots to enable data curation but to rely ultimately on humans by building a 'structure centred community' around its search engine and to use a crowdsourcing approach to correct errors and to increase the quality of the data[68].

3.2.6 Preprint Servers

The first attempt to establish a web-based preprint server in chemistry failed. In August 2000, ChemWeb.com, a subsidiary of the commercial publisher Elsevier, launched a preprint server for chemistry[69] (CPS) that was modelled after the physics ArXiv (Brown 2003). Submissions were screened for general appropriateness, but not peer-reviewed. In its 4 years of existence, CPS attracted only about 900 submissions by authors – regarded insufficient to justify further development. CPS was discontinued in May 2004.

Several reasons for failure have been suggested. First of all, the ACS journal editors usually do not accept a manuscript that has been already publicly disseminated as a preprint. This almost certainly provided a strong disincentive to chemists to post their work on a preprint server. Indeed a study on the "Use and Non-Use of E-Print Archives for the Dissemination of Scientific Information" (Lawal 2002) found that at least 30% of chemists said they would consider using a preprint server if it was not against journal policies. On the other hand, these figures imply that a majority of chemists would not use preprint servers for other reasons, and we speculate that the general attitude among many chemists is captured well by the following view of preprint servers expressed by the ACS journal editors in 2004[70]:

---

[68] Presentation slides at http://www.slideshare.net/AntonyWilliams/crowdsourcing-collaborations-and-text-mining-in-a-world-of-open-chemistry-presentation

[69] Homepage of CPS at http://www.sciencedirect.com/preprintarchive?url=/CPS

[70] Accessed online on 20 Feb 2009 at http://pubs3.acs.org/instruct/preprints.html. Note that in the meantime this page has been taken down and the URL now points to journal specific statements that were released in 2008. Although they may vary between journals they generally clarify that: ACS journal editors consider for publication only original work that has not been previously published and is not under consideration for publication elsewhere; content that has been made publicly available, either in print or electronic format, and that contains a significant amount of new information, if made part of a submitted manuscript, may jeopardize the originality of the submission and may preclude consideration for publication.

> *"The disadvantages of preprint servers include: the potential for flooding the literature with trivial and repetitious publications, thus making extraction of reliable and valuable information more difficult; absence of peer review; possible premature disclosure with inadequate experimental details or supporting data; premature claims of priority; potential lack of proper references and credit to prior work; abuse of multiple revisions or updates; possible lack of duration and long term archiving."*

A skeptical view on the good of preprint servers was also expressed by one of the participants attending the Washington workshop. The view expressed was specifically concerned about: 1) the quality of presentation, especially graphics, 2) The lack of any control on the ethical granting of credit and precedence - both of these issues were felt to be dealt with very well in the editorial and the peer review process; and more generally a concern of 3) giving in to a culture of priority, and 4) allowing a narrowing of perspective to happen as preprint servers might encourage a specialized approach to literature (reading only what is of interest to me) rather than a broadening one, thereby generating a system to enlarge and connect the scientific outlook of a young scientist.

We may conclude that no 'preprint culture' exists in chemistry that assigns value to publicly disseminating their not yet peer-reviewed manuscripts. On the contrary, to disclose information before priority has been established through a formal journal publication seems to be perceived as too risky. In contrast, in some areas of physics the exchange of preprints between scientists and between their institutions has a long tradition, pre-dating the web by several decades. Physicists that use ArXiv trust that their colleagues respect the time stamp that an item receives on submission to the server as establishing priority. Also they seem to value having all the ongoing work in their field freely accessible from a single web server with little concern for the lack of peer review. Eventually, most of the preprint manuscripts submitted to arXiv are published in peer-reviewed journals and a system of checks and balances prevents the server from being flooded with submissions that are either irrelevant or of poor quality. One interpretation of these differences between disciplines in their attitudes towards preprint servers suggests that they are connected to differences in the social and intellectual organization of scientific fields. It is assumed that physics is characterized by a higher degree of intellectual coherence, and by stronger interdependencies between researchers. Because scientific communities in physics subfields are more tightly interconnected, issues of trust play out differently and reduce the risk associated with the posting and reading of preprints (Fry & Talja 2007).

In June 2007, Nature Publishing Group launched *Nature Precedings*[71], a preprint server targeted at researchers in the life sciences. *Nature Precedings* is intended to enable researchers in the life sciences "to openly share preliminary findings, solicit community feedback, and claim priority over discoveries by posting

[71] Homepage of Nature Precedings at http://precedings.nature.com/

preprint manuscripts, white papers, technical reports, posters, and presentations." A DOI is assigned to each submitted item such that it can be cited unambiguously by this identifier, and the system includes a mechanism for version control. It remains to be seen how usage and acceptance of *Nature Precedings* will evolve. Current submission statistics show about a hundred submissions in the chemistry subject category. Most of these are cross-listed with other subject areas, in particular biotechnology, bioinformatics and pharmacology. Only 20% of the submissions fall squarely into the core chemistry domain (i.e. are not cross listed). Take-up of this service by chemists is so far relatively low, which is not surprising, since the service is targeted at the life sciences.

3.2.7 Open Access to Journal Literature

The 'green road' to open access, i.e. the so called self-archiving by authors of their peer-reviewed articles in either institutional repositories or subject specific repositories (such as PubMed Central for the life sciences) has not yet found significant uptake in chemistry. We have already discussed the lack of use of a subject specific preprint server in chemistry, and there is no chemistry specific server to collect peer-reviewed 'post-prints' either (such as NIH has set up for bio-medicine with Pubmed Central). An increasing number of universities and other research institutions such as the Max Planck Society or CNRS operate institutional repositories to capture their research output such as the full text of published articles, but deposit rates are relatively low and use is scattered across all disciplines (Jones 2008, Zuber 2008). Only a small percentage of an institution's research output tends to be captured in its institutional repository unless a deposit mandate requires researchers to deposit their articles in the institutional repository (Sale 2006). The argument has been made that in order to find better acceptance institutional repositories would need to be more closely geared towards a specific discipline's communication culture and thereby tie in with researchers information seeking needs (Kingsley 2008). Evidence of the differences between individual discipline's use of institutional repositories is scarce, but a case study of seven representative institutional repositories suggests that chemists are not less likely to contribute than e.g. physicists (Xia & Sun 2007).

Journal publishers have increasingly adapted their editorial policies to tolerate various forms of self-archiving of final peer-reviewed version of articles by researchers (on author homepages, in institutional repositories or subject specific archives). Those policies vary in some details, e.g. sometimes an embargo period is imposed such that the article cannot be released on a repository until 6, or sometimes12 months after publication, or some publishers allow the publisher's formatted pdf to be used, while other publishers allow only the deposit of the author-revised accepted manuscript. Especially in the life sciences there has been significant pressure from funding agencies mandating some form of open access for publications reporting results of research that they fund (Wellcome Trust in 2003, NIH in 2008). In chemistry, comparable pressure from organizations

funding non-corporate research has been lacking. However, some of the large publishers in chemistry today have a somewhat self-archiving friendly policy: RSC allows authors to post the final publisher pdf of their article on their personal website, and to deposit the author's version after 12 month embargo on institutional or subject specific repository, Elsevier and Springer allow self-archiving of the peer-reviewed version (though not publisher pdf). Notable exceptions are ACS, which at present does not allow any form of self-archiving[72], and Wiley VCH, which for Angewandte Chemie only allow self-archiving of the unrefereed version, i.e. the preprint. Both publishers make concessions only in the case of the existence of funder mandates (such as NIH) in which case they allow authors to comply with these mandates.

The 'golden road' to open access, that is journals operating under an open access principle by providing open access to the articles that they publish is making some modest inroads into the chemistry domain. Most open access journals are newly founded journals, and hence have to build up reputation over years. Examples of new open access chemistry journals that are also indexed by Thomson Reuters (ISI) are *Beilstein Journal of Organic Chemistry*[73] founded in 2005 (about 40 articles in 2008, no article publication fee as journal is funded institutionally by Beilstein Institute), *Molecules*[74] founded in 1996 (about 240 articles in 2008, article processing fee of 780 US$), and *Chemistry Central*[75] founded in 2007 (25 articles in 2008, article processing fee of 1,250 US$). The IUCr's *Acta Crystallographica Section E: Structure Reports Online* is an example of an established journal that has switched (in 2008, article processing fee 150 US$) from a subscription journal to an open access journal. It publishes about 5,000 articles each year and hence contributes a much more substantial number of articles to the open access literature in chemistry than the newly founded journals. Open access journals usually have a policy of waiving publication fees for under-funded authors e.g. from developing countries.

Instead of switching the business model for their entire journal operation from a subscription funded model to an author or sponsor-funded model, a number of established journals in chemistry have taken an intermediate step: they give authors the option of paying a fee for providing open access to their article on the journal web site. Within this hybrid model publishers charge authors between several hundred to about 3,000 US$ to grant open access to their articles as they are published on the journal's website. Major publishers that offer this option for article fees between 1,000 and 3,000 US$ are ACS ('ACS Author Choice'), Elsevier ('Sponsorship Option'), Springer ('Open Choice'), Wiley-Blackwell ('OnlineOpen'), and RSC ('RSC Open Science')[76]. Uptake among ACS authors

[72] but for NIH funded authors – after 12 months embargo their articles are deposited on Pubmed Central to comply with the NIH mandate.
[73] Journal homepage at http://www.beilstein-journals.org/bjoc/
[74] Journal homepage at http://www.mdpi.com/
[75] Journal homepage at http://journal.chemistrycentral.com/home/
[76] See the Sherpa/Romeo database on journal open access policies for further information http://www.sherpa.ac.uk/romeo/PaidOA.html

The Scientific Analysis of Scientific Journal Literature

It is difficult to quantify the proportion of the open access journal literature in the sciences (and its eventual growth), and to make meaningful comparisons between research fields. To obtain at least some ballpark figures we turn to Ulrich's periodical directory that captures about 25,000 current peer-reviewed journals in any field of knowledge. The database has its own proprietary subject classification with about 100 top- level subject headings (e.g. CHEMISTRY or PHYSICS). We select for our comparative analysis of a few broad scientific fields journals that are peer-reviewed and indexed by some abstracting & indexing service – to ensure some basic level of scientific reputation of the journals that we include*. We end up with the following figures (as of 5 April 2009):

	# journals (refereed, indexed & abstracted)	**subset that is open access**	
CHEMISTRY	*678*	*48*	*(**7.1** %)*
PHYSICS	*662*	*49*	*(**7.4** %)*
ASTRONOMY	84	*7*	(**8.3** %)
BIOLOGY	2,372	252	(**11.3** %)
MEDICAL SCIENCES	4,130	471	(**11.4** %)

These are high-level numbers. We do not know to how many articles the journal numbers correspond to, as journal size may vary with subject area and with publication mode (subscription vs. open access). With the journal as unit of analysis we further fail to include open access articles published in hybrid journals under an article-specific open access scheme.

To refine this and similar analyses we face two obstacles: one concerns getting comprehensive access to the primary data on scientific literature that is held in proprietary databases (Ulrich's, SCI, CAS, SCOPUS…), which must be integrated for such analyses. Only two or three research groups in information science worldwide pride themselves in having comprehensive access to SCI data, since costs for acquiring the data are substantial. The other issue concerns the level at which such comparisons between subject fields are meaningful. It has been shown that subject classifications developed by indexing services are optimized for retrieval but do not correspond well to the journal groupings that emerge from the literature itself e.g. through citation links between journals (Rafols 2009). These patterns of self-organization in the scientific literature may provide more meaningful aggregates for comparisons of publication practices between research areas. To what extend sub-groupings need to be considered and distinguished in order to capture scientific communication cultures is an open research question.

* There are some serious issues with the bias (language, geography) of indexing services like the Science Citation Index, but Ulrich's considers a long list of abstracting and indexing services such that we assume a reasonable balance.

seems to be minimal – between August 2006, when the option was introduced, and October 2008 less than 300 articles have been published as open access articles in ACS journals under this scheme. Some research institutions have done package deals with publishers (e.g. University of California and Max Planck Society with Springer) that provide them with an institution-wide subscription to the publisher's journals plus waiving of the open access fee to provide open access to all articles submitted to these journals by authors form the respective institution.

A preliminary quantitative analysis of the golden road to open access (see insert above) suggests that this strategy is most prevalent in biology and the medical sciences, where more than 11% of the refereed journals are open access journals. The proportion of open access journals seems to be lowest in Chemistry (7.1%), but interestingly the number for physics is similarly low (7.4%). These numbers that count open access at journal-level possibly overestimate the fraction of articles published in open access mode, since open access journals are often new journals and hence likely smaller in size than established subscription journals. On the other hand, these numbers do not include additional open access articles that are published by hybrid journals.

3.2.8 Use of Web 2.0 Tools in Chemistry

The Chemical Blogosphere

There is scattered use of web blogs in chemistry. There are individuals that maintain blogs dedicated to some aspect of chemistry, with the subject matter including personal experiences in the lab and beyond, conference reports, informal discussion of papers, synthesis protocols, ideas, science politics and science publishing. As a vehicle for informal scientific communication they are far from being mainstream and widely used at this point. For an entry point into the emerging chemistry blogosphere see the index of chemical webblogs on 'Chemical Blogspace' http://cb.openmolecules.net/blogs.php or individual blogs such as:

- *Molecule of the Day* http://scienceblogs.com/moleculeoftheday/,
- *Carbon based Curiosities* http://www.coronene.com/blog/ or
- *Totally Synthetic* http://totallysynthetic.com/blog/.
- *YoungFemaleScientist* http://youngfemalescientist.blogspot.com

Most blogs have a so-called 'blogroll' section with links to further blogs recommended by the blogger.

Then there seem to be cases where a group of scientists uses blogs and shared wikis to coordinate their research activities, and to reach out to a virtual community with a core of members that regularly meet face-to-face. An example is the 'Blue Obelisk' group of chemists, programmers and

informaticians[77]. Note that the 'Blue Obelisk Award' that is given to people who significantly promote Blue Obelisk activities is not a virtual prize but an artefact that has to be handed over in a face-to-face meeting ("there has to be physical meeting - they are not delivered by post"[78]), reemphasizing the role of face-to-face meeting for this community.

To what extent these communications represent scientific communication aimed at the dissemination and validation of results or just 'scientists communicating', and hence part of the social and political glue in the production of scientific knowledge, is open to debate. The boundary is, and always has been fuzzy, independent of the medium used. Still, the question arises how the new global, web-mediated forum changes the impact of those kinds of informal communications that hitherto were limited to face-to-face encounters, e.g. in the corridors of research labs or at coffee breaks of scientific meetings, but now reach across larger social and geographic distances (Todd 2007).

Publishers seeking interaction with their audiences

Publishers also experiment with Web 2.0 tools. They launch web blogs to add some level of interactivity and informal communication around their journals, and to act possibly as sounding boards to pick up moods and trends from the web-articulate subsection of the scientific community. Examples are:

- Nature: *The Sceptical Chymist* http://blogs.nature.com/thescepticalchymist/
- RSC: *Chemistry World Blog* http://prospect.rsc.org/blogs/cw/

In particular ACS is experimenting with many different Web 2.0 methods to package and deliver journal content, and to try to build virtual communities around it. The general chemistry journal *JACS* has a web-based sandbox called *JACSβ* where it offers demos of new features (such as audio readings of communication type articles) and invites feedback from users before considering them for inclusion in the journal website proper. It is a testing ground for features to include in the *JACS* journal and the ACS publishing platform more generally. With a portal on nano science called *Nanotation*[79] ACS is trying to build a virtual community around the topic. Its mission statement reads[80]:

> - *"Promote nanoscience and nanotechnology*
> - *Save students and researchers time by providing a portal to content in the field that interests them*
> - *Free, easy access, current, high-impact, forward-looking, broad scope*

[77] Homepage http://blueobelisk.sourceforge.net/wiki/Main_Page
[78] Peter Murray-Rust, 30th May 2007 http://wwmm.ch.cam.ac.uk/blogs/murrayrust/?p=343
[79] Nanotation home page at http://community.acs.org/nanotation/
[80] *Susan King, Workshop Washington DC, October 2008*

- *Forum for scientists (especially early in their careers) to promote their work and interact with others in the field"*

Many of these uses of Web 2.0 tools seem to be to reach out not only to researchers, but also to non-experts such as students or teachers, and are aimed at facilitating re-use of material for educational purposes and at facilitating communication across disciplinary boundaries.

A Quick Win: Increasing the Value of Supplementary Data

Jeremy Frey

(a**) Make the data files that produce any graphs and tables in your publications available** as ascii data. This is not hard to do; the files exist as they have been used to make the graphs (export from the graph program if necessary, or have the excel file). E.g. make the **digital spectra files** available and useable. Most spectra are now exportable in standard formats together with a moderate amount of metadata

(b) The **journals need to ensure that these supplementary files are indexed and given a DOI** (a persistent, citable identifier); this is already done in the eCrystals approach x-ray crystal structure data (ecrystasls.soton.ac.uk).

(c) **The format in which this data is kept on the journal website should be such that it can be easily reused as input for further calculations**. Should the journals largely hold files in the less optimal format of pdf, authors ideally need to ensure that the layout of the ascii files that will be converted to pdf are simple enough for the tools available to convert the pdf back to useful formats. Better still, encourage that the supplementary data is held in formats that are understood in the community.

(d) **In those fields where repositories exist, an ideal business solution and academic solution would be to have journals agree to allow the data to be posted in those data repositories**, **as well as in the paper** and to link from the journal articles to the data. This is already done with crystal structures data files where the structure files are held in the Cambridge Structural data base (CSD) and linked via a code for validated structures that allows one to find the data in the CSD run by (Cambridge Crystallographic Data Centre) CCDC

(e) A quick gain will be **better import/upload facilities for spectra** based around a basic repository for the spectra. Trials of this exist but the problem is the commercial sustainability. An important issue is to provide a DOI or a similar identifier for each spectrum. (an extension of the problem raised in (b)) This is probably be done at an institution level and the spectra exposed to the outside and linked to from journal articles rather than attempt a single repository to which the spectra are handed over. When appropriate, the INChI provides a very good identifier for the molecule (but a URI for a specific example of the spectrum is still needed). We could imagine a YouTube or Flicker for Spectra.

How such a Flickr for spectra could be used e.g. for educational purposes is demonstrated by the *Spectral Game* implemented by Bradley et al (2009) on top of the Open Data service of ChemSpider.

(e) Given that we now need to package together more material - the journal paper, the supplementary materials, any material in other repositories, etc - the role of the overlay journal becomes more interesting**. The overlay journal can provide the necessary link between these parts and then can also describe views and comments on any of the parts (validation & discussion).** Such collections are beginning to exist. Some societies provide an overlay for example of all articles of interest to a sub-community whichever of the society's journals they appear in. Extending this to link to material elsewhere is possible but may run into issues of ownership.

4 Chemistry Distinguished

In this section we examine the distinctive characteristics of chemistry and hypothesize how they may shape its communication practices and have impact on the manner in which scientific communication in chemistry will further evolve. This is a preliminary list only and is proposed as a seed for further discussion. The discussion should not only address validity of the assertions here, but also should attempt to sharpen the specific context in which the stated characteristics apply. Certain characteristics will be relevant in other scientific disciplines, but we posit that the interplay of these factors distinguishes chemistry from other disciplines and constitutes the unique context of scientific communication in chemistry. We expect that the question of the manner in which these and other factors impact changes in the scientific communication system in chemistry should be one of the central issues of the proposed follow-up workshop.

In the organization of this section we make a broad distinction between characteristics inherent to chemistry research, and characteristics of the larger socio-economic organization of research in chemistry. The former are closely linked with the material research culture and the epistemic culture of chemistry. The latter relate to aspects of the social, political and economic organization of the discipline and its communication system. Presumably neither type of characteristic is easy to change, nor are these characteristics entirely independent in their historical evolution. Still, the research-inherent factors change only with the type of research being done and the mode of research. They are important to understand in order to appreciate the specificity of scientific communication practices in chemistry. This is especially for non-chemists. In contrast, the socio-economic organization seems more contingent and amenable to change without fundamentally challenging research aims and practices in chemistry.

The question of the stability of these factors and the forces influencing them (e.g. increasing funding for interdisciplinary and collaborative research) would be an important question to consider at the proposed workshop.

4.1 Research Practices

4.1.1 Focus on Creation

Chemistry has been characterized as distinct from other sciences by being first of all the "Art, Craft, and Business of substances and their transformation" and only secondly a "Science" (Hoffmann 2007). This not only refers to the huge economic significance of the chemical industry that is coupled tightly to the chemical profession (see section 4.2), but also to the ways knowledge is produced in chemical research. If, for the exercise of distinguishing chemistry from other sciences, we focus on the chemical core of chemistry (Schummer 1998), we find

a predominance of synthesis. Chemists create the empirical objects whose material properties they then investigate. Producing new chemical substances is a central research activity in chemistry (Schummer 1997a). "Making something" is a leitmotiv of chemical research, and it has been argued that this orientation overshadows the discovery mode that is typical for physics and biology. Synthesis does not fit comfortably into a Popperian model of science progressing by refuting hypotheses – what hypothesis is disproven by the synthesis of a molecule? Synthesis brings chemistry close to engineering and to high art.

For the preparative chemist succeeding in a synthesis requires intuition and tacit knowledge, making it often seem to be more of an art than a science. It foregrounds an affirmative mode of representing research results rather than one that is critical, and where failure is taken as refutation of a hypothesis (Hoffmann 2007). What is, is - however limited the detailed understanding is of how one got there.

Implications for scientific communication: A large part of the chemical literature reflects this 'stamp collecting' mode of chemical research. Since the Second World War the number of new chemical substances reported per chemically relevant article has doubled to reach a mean of about 1.7 by the mid 90's (Schummer 1997a). The emphasis in synthetic papers where these new substances get reported is on the synthesis itself, and the application of the new substance in generating other new substances (Lipkus et al. 2008) – not on the analysis or theory development (Hoffmann 2007, Schummer 1997). With a lack of emphasis on reproducibility in order to promote theory development, research papers can afford to be incomplete in the description of the synthetic protocol. Only a tiny portion of synthetic papers (published in Organic Synthesis http://www.orgsyn.org/) undergo rigorous peer-review that aims at reproducing the results to confirm a reliability and hence reusability of synthesis protocols. So far, chemists can be very productive researchers measured by their publication output without providing a comprehensive and reusable documentation of the data underlying their research. Hence within this model there is little push for using web technologies to publish data more widely in a reusable manner.

4.1.2 Long Tail Science

Chemistry can be characterized as a long tail science (Murray-Rust 2008). It is a field of research dominated by large numbers of small research producing units (in contrast to the large-scale collaborations of high-energy physics). The typical chemistry research group is lead by a principal investigator (PI) and composed of 5-15 graduate students and postdocs, collocated at a single site. Autonomy of these groups is high, since conducting successful research can be done with minimal reliance on other research groups. A group can fall back on a local service infrastructure to provide routine measurements, and there is occasional collaboration with other groups to exchange samples or provide a measurement

service that is not available at the home institution. When collaborations between groups occur they are characterized by a limited reliance on one another (Walsh & Bayma 1996).

Implications for scientific communication: the predominantly non-collaborative mode of research in chemistry reduces the incentive to make use of new technologies to facilitate data sharing and research collaboration. This contrasts with the need to share seamlessly information within the large collaborations of high-energy physics that gave rise to the invention of the World Wide Web at CERN. Furthermore, the success model of an autonomous research group makes secrecy rather than openness an effective communication and research strategy. For instance, it makes it desirable to keep a doctoral dissertation coming out of a research group hidden in the maize of a physical library instead of posting it online - to ensure that the group can mine the knowledge included in the dissertation and optimally exploit it through publications.

4.1.3 Longevity of Scientific Literature and Data

Another hallmark of chemistry as a science is the enormous knowledge base of scientific data that has been accumulated over more than a century. The millions of data sets reported in the literature and captured in databases that record chemical structures and the physical and chemical properties of chemical substances retain immediate value for the majority of current research in chemistry. This knowledge rarely gets outdated, and comprehensive access to this knowledge base is vital for conducting chemical research. Further, the accuracy of data and unambiguous identification of substances is critical not least for the safety of chemists working with them in the research lab.

Implications for scientific communication: For any transformation of the scientific communication system including the legacy data is not just a nice-to-have feature, but essential for the use of a new system to chemists. So are good mechanisms for quality control and identification of the data to make it safe to use.

4.1.4 Non-Digital Practices

Preparative chemistry as 'art and craft' emphasizes manual practices. Even though instrumentation in most labs is increasingly computerized, it is only partially integrated into the workflow, and a large part of the everyday work consists of manual manipulation of experimental set-ups and substances. Hence, although computers have become important in chemistry research, e.g. for information retrieval or to read out measurement data, the computer is not the central tool for generating chemical knowledge (if we exclude subfields such as quantum chemistry or cheminformatics). Tacit knowledge and intuition play a

great role, as does scribbling on paper. Usually lab space and computer spaces (for writing or doing information research) are well separated.

Implications for scientific communication: It is a challenge to integrate digital data capture into the workflow of chemists in the lab. The experience of gaps between the physical medium and digital medium may well be more pronounced among lab chemists than other scientists whose work is more integrated with computers.

4.1.5 Computerized Chemistry

While the majority of synthetic academic chemists pursue synthesis as a manual craft, there are certain areas in chemistry (medical chemistry, pharmaceutics, computational chemistry) that heavily rely on automation and computing for their research goals. In particular, industry has invested in combinatorial chemistry and high-throughput screening in order to increase chances for discovery of new substances for drug design.

Implication for scientific communication: Chemists working in these areas are required to develop advanced IT skills and data sharing practices. Further, activities in these fields do increase the need for an integrated electronic information infrastructure to support automated data capture and exchange of data in standardized formats (Farrusseng 2008). An open question is, to what extend do these trends influence chemical research in other chemical subfields, and whether there will be a disconnect between infrastructures build in industry and those in the academic domain.

4.1.6 Diversity of Research Cultures in Chemistry

So far, we have used a broad brush to bring out the distinct features of chemistry in comparison to other sciences. In reality, and as indicated in the two previous sections, research cultures within chemistry are all but unified. There is a plurality of historical research traditions, methods and goals that research fields within chemistry adhere to, as well as a variety of interdisciplinary projects (Schummer 1998). A conventional classification would distinguish subfields such as inorganic, organic, physical, analytical, polymer chemistry, biochemistry, physical and theoretical chemistry, and chemical physics. In recent times these subfields have been added to – materials chemistry, green chemistry, environmental chemistry, and chemical biology.

But this is not the only way to distinguish research subcultures. More natural might be one of the two following broader distinctions:

- A major division is between those people who make molecules and get their structure and properties, in contrast to those who study their properties. In the first group are synthetic organic and inorganic chemists, and polymer chemists; in the latter are physical, theoretical, and analytical chemists. The former – synthetic organic and inorganic – probably publish a little more, and the unit of publishable research is smaller.
- An alternative division is synthesis (I made it!), analysis (what do I have?), mechanism (how did it happen?) and theory (why, oh theorist, why?). The synthesis and analysis people share one subculture, the mechanism and theoretical people another.

Implication for scientific communication: One may speculate that different subfields are served by the existing system in different ways, and to a different degree of satisfaction. Consequently the perceived need for improvement, and readiness to innovate will differ across subfields in chemistry. E.g. for organic and synthetic chemists SciFinder/CAS and Beilstein are central resources, organized very effectively around chemical structures, the 'lingua franca' of the organic chemist. For theorists and physical chemists factual databases are of greater importance than to the former group, and holes in their coverage, and the ability to combine data are perceived as failures of the existing system. Further, the configurability of new web-based technologies makes it possible to support a greater plurality of communication practices in the future. Hence our analysis needs to take the diversity of research cultures in chemistry into account. It is an open research question how many different types of communication cultures in chemistry need to be distinguished for such an analysis.

4.2 Socio-Economic Organization

4.2.1 Proprietary Nature of Chemical Information

Chemistry is distinguished from most other disciplines in that the chemical information that is produced in everyday academic research is of considerable relevance for a huge, profitable chemical industry that is the prime user of that information. One important effect of this is the need to strategically plan IP protection for new substances developed in industry labs or in collaboration with academic groups. Hence, information resources, in particular the accumulated chemical information in databases, represent an extremely valuable commercial resource. Further, confidentiality with respect to the use of this resource, to avoid tipping of competitors about the research direction a company is taking, is important.

Many fields of academic chemistry and industry are closely interlinked. Academics train PhD students and 70% of those students find their work in the chemical industry. Industry funds some fellowships for training of students, as well as research projects. Chemistry professors quite often act as consultants in industry. Some chemists move between employment in academia and in industry, and back. There is a strong group identity of chemists across the academic-corporate divide (Laszlo 2006). When academic chemists work on industry-funded research projects, they have to accept a certain trade-off with regard to their ability to gain scientific credit by publishing results, versus the need to keep results secret to protect the industrial partner's interest in exploiting the results for commercial gain. This obligation to secrecy may be temporary in those cases where a patent needs to be filed first before results can be published. In other cases though, the obligation to secrecy may extend indefinitely when patenting is avoided because a patent violation would be impossible to prove, and filing a patent would only tip off the competitor.

"Socialized Science"

"… Their [open-access advocates'] unspoken crusade is to socialize all aspects of science, putting the federal government in charge of funding science, communicating science, and maintaining the archive of scientific knowledge. If that sounds like a good idea to you, then NIH's open-access policy should suit you just fine."

Rudy M. Baum, Editor-in-Chief
Chemical & Engineering News
Editorial "Socialized Science",
20 Sep. 2004, C&EN 82(38), p. 7

Implications for scientific communication: For a number of reasons, chemists are more secretive about details of their research in formal and informal communication. The proprietary nature of chemical information and the commercialization of chemical information seem to be widely accepted among academic chemists. This contrasts with other sciences with less market penetration, where the ideal of open sciences and a more socialised approach to scientific information prevails (Walsh & Bayma 1996). This might explain a cultural inertia that mutes the rallying cry of the open access and open science movements in the chemistry community. Certainly this is a sentiment played on by the editorial on *Socialized Science* that appeared in *Chemical & Engineering News* as part of ACS's lobbying against an open access mandate for NIH funded research (see insert).

4.2.2 Industry - Academia Balance in Chemistry

Some academic chemists feel that the relationship between industry and academia is unbalanced, that industry is in some sense 'feeding-off' academia. They point out that academia produces two vital inputs for the chemical industry, trained PhDs and published scientific results, without proper compensation. Industrial researchers read the scientific literature but they publish themselves only sparsely, because their careers do not depend on it, and because they want to keep their research strategies and goals secret from competitors. So far the consumption of literature by industry is taxed through journal subscription fees.

Implication for scientific communication: The fact that industry's consumption of chemical information is disproportional to its production of scientific publications makes switching business models to an open access publishing model problematic. A producer-pays model would imply 'free-riding' by those who consume but do not equally contribute. Depending on what the proportion of publishers' journal subscription income from industry is (said to be 25% for ACS), a producer-pays model could raise a substantial financing problem for the public sector research funding (and subsequent publication of that research). It also explains why some academic chemists are particularly skeptical of current proposals for open access business models, as they feel that industry would profit inappropriately.

4.2.3 ACS's Global Responsibility

The world's largest scientific society, ACS[81] is a non-profit organization[82] but nevertheless behaves very much like a commercial entity with regard to the information services it develops and offers. This behavior may derive from the dominant role of members and customers from the commercial sector in the society, who tend to perceive of chemical information mainly as an economic asset, rather than as a common good. There is evidence, however, that this business orientation of ACS causes frictions with its academic membership[83], in

[81] with about 155,000 members, including 19,000 international members.

[82] "…publicly supported, federal income tax exempt organization pursuant to Sections 501 (c) (3) and 509 of the Internal Revenue Code of 1986, as amended." from 'About ACS' http://portal.acs.org/portal/Navigate?nodeid=225

[83] The level of salaries and bonuses that ACS officials receive has raised the eyebrows of some academic chemists, and spurred some accomplished members of ACS to speak out publicly. ACS defended its position by referring to the large membership, the $420-million annual revenue, and the $1-billion in assets and the need to offer salaries competitive in comparison to employers with a similar-size operation (Jacobson 2004). Hence, if one acknowledges ACS Publishing and CAS as being massive businesses, then the remuneration would seem to be in line with that size of business.

particular in the context of the Society's responsibility for the global scientific communication system in chemistry.

With regard to its journal publishing operation ACS , like many other society publishers offers good value for money[84], while gaining revenue to fund other parts of the society's operation. With CAS it runs a second economically successful information service operation, that has been reported to have generated $250-million revenue for ACS in 2007 (Trager 2009). Since academic institutions seem to get substantial discounts for access to the CAS database via STN, SciFinder or SciFinder Scholar, a large part of this revenue is presumably generated from corporate customers in the chemical and pharmaceutical industry.

With over more than one hundred years of existence, ACS has built a collection of high-quality and well-reputed chemistry journals and the CAS databases into an invaluable resource of accumulated, quality-controlled chemical information. It is one of the dominant chemistry publishers, and in the case of the CAS Registry it controls a resource of critical global relevance to almost anyone doing research in chemistry.

Through the web, the opportunity now exists to network chemical knowledge and to build collaborative, shared services. The evolution of the open science web produces considerable tension with the established privately-funded proprietary information system in chemistry. In reaction to this evolution, it appears that ACS and CAS are attempting to maintain their economic assets and sustain the proprietary regimes that have worked so well in the past. Examples of this behavior include the vigorous opposition displayed by ACS to the scope of NIH's PubChem database (Kaiser 2005). This raised considerable unrest among academics and in the library community[85]. Indeed, the nature of a business model that balances investment in quality-ensured services and harnesses the power of open, integrated and shared resources is unclear at this point and is the subject of active investigation.

There is some evidence of a more open attitude by CAS to open solutions. The standard practice of CAS is to require a licence fee from third party services that want to make use of CAS identifiers to link information on more than 10,000 chemical substances. This way it can exert control over any third party services that it may perceive as evolving into a potential competitor. Recently CAS made

---

[84] See http://www.journalprices.com/ and listing of ACS journals in 'good value' category.

[85] documented in an exchange of letters between the president of ACS and the president of NIH, available from a website on the topic of "The American Chemical Society and NIH's PubChem" by the University of California Office of Scholarly Communication at http://osc.universityofcalifornia.edu/news/acs_pubchem.html

a laudable, if guarded, concession by agreeing to cooperate with Wikipedia Chemistry to provide accurate assignment of CAS identifiers - to current substances "that are of widespread general public interest." (i.e. either elements or substances cited at least 1,000 times in the literature). A free web-based service called *Common Chemistry* that allows users to look up names or identifiers of 7,800 common chemical substances was launched in December 2008[86].

4.3 Non-Chemistry Specific Factors

We have thus far focused on factors that distinguish chemistry from neighbor disciplines such as physics, life sciences, or computer science. The question arises of how generic, non-chemistry specific factors play out. Do discipline specific factors dominate or do commonalities with other scientific disciplines outweigh discipline-specific features in shaping the future evolution of scientific communication in chemistry? Or is there a domain specific interplay of these factors?

Implications for scientific communication: As the de-facto global registration authority for chemical substances, CAS identifiers are a vital component of any future web-based information system based on shared chemical resources. Whereas a closed subscription–based system may have been a sensible financing model in times predating the web, CAS's proprietary policy towards large-scale use of the identifier system undermines widespread experimentation and innovation by third parties that rely on the accurate integration of chemical information. Such integration is necessary for a web of shared chemical resources. Hence, CAS's proprietary policy presents a major stumbling block for new models gaining critical relevance to chemists, and achieving wide uptake.

One example of a trend that is independent of chemistry is the manner in which major publishers approach scientific publishing as a commercial enterprise and position themselves strategically to gain profits. In their strategies they exploit the increasingly close link between publishing and the evaluation of research performance. As long as policy-makers, politicians, and science administrators buy into this paradigm, it represents a strong force shaping the scientific publishing market and its products. Issuing more and more specialized journals is one of the strategies to increase market share that several larger publishers pursue. These new journals then play a role in the system by which scientists are evaluated, tenured, and promoted. An example of how publisher strategies and evaluation practices interact is the new series of Nature journals (such as *Nature Materials* or *Nature Chemistry*). *Nature* is a distinguished brand, and because publications in those high prestige journals have considerable impact on the evaluation of an individual's publication record, this new set of high prestige

[86] Service online at http://www.commonchemistry.org/

journals enters the market with support of the scientists, who welcome the new vehicle for enhancing their C.V. However, these costly new *Nature* journals further squeeze precarious library budgets quite considerably, at the expense of more specialised journals. Another example is the publisher Elsevier. Elsevier has invested in establishing a new bibliographic database (SCOPUS) and is now offering new scientometric evaluation tools (SciVal[87]) on top of *SCOPUS*, thereby entering into competition with Thomson ISI that hitherto domineered the market for bibliometric evaluation services.

These trends are troubling because they emphasize scientific publication as a proxy for actual research performance. This emphasis on the evaluative aspect of scientific publications marginalizes the primary function of scientific publications – to communicate and provide evidence for new scientific results. How do these trends affect chemists' expectations in new models of scientific communication, and their willingness to experiment with new models?

We suggest that to understand the forces at play and the dynamics of the evolution of scientific communication in chemistry, such generic, non-chemistry specific factors need to be taken into account.

[87] Product information at http://scival.com/

5 Conclusions and Aims of a Future Workshop

We have argued in this paper that some of the new web-based models of scientific communication that have made an impact in other disciplines do not seem to have equal value or impact in chemistry. Nevertheless, pioneering chemists and chemistry publishers are actively exploring the value of web-based information and communication services for chemists, and are developing components for what could be called a semantic chemistry web. So far though these efforts have found neither recognition nor widespread use by the average chemist. Tangible hurdles exist like the lack of an open global, curated, and authoritative identifier system. The question we raise is whether there are further incompatibilities between what could be envisaged as an open, integrated web of chemistry information, and the research and communication culture within chemistry.

In the previous section of this paper we have put forward a number of observations on specific characteristics that distinguish the chemistry domain from other scientific disciplines, and that may shape the evolution of scientific communication in chemistry. This list is preliminary and needs further validation, including the sharpening of the contextual circumstances in which these observations hold. We suggest that further research into communication practices in different research fields within chemistry is needed in order to identify areas of perceived failure of the existing system and to assess the potential value of new models. In particular, further research is needed to determine the number of communication cultures coexisting within chemistry that need to be distinguished before any meaningful comparison can be made with other scientific fields.

5.1 Points of Dissent

As already mentioned in the introduction of this paper, the argument presented here does not find unanimous support among all participants of the workshop held in October 2008. We have benefited from the feedback and comments we received to the paper draft, and we have worked those comments into the revised version of the document. A few critical points though remain that we distil here to highlight areas of disagreement:

- ACS position on open access: whereas we interpret in this paper ACS actions and published statements by its staff as opposed to open access and sometimes actively working against it, ACS representatives claim that the Society has not taken a position on open access.

- The role of the CAS identifier for building an open chemistry web: we claim in this document that the proprietary nature of this identifier and the licensing policy of CAS inhibit innovation by third parties. ACS

representatives claim that the ownership of the CAS registry by ACS and the licensing policy and practices for the CAS identifier system are not monopolistic in character, and do not put severe restrictions on third party services who want to make use of it for the identification of chemical substances.

- One of the respondents criticised that many of the points we list in section 4 do not appropriately characterize the research and communication culture in chemistry. The respondent claimed:

    - Chemists are *neither* secretive *nor* non-collaborating
    - Chemists *have* fully embraced the web as a communication vehicle and productivity-enhancing tool. Hence, it is *inaccurate* for us and others to claim that chemists fear that the web medium may endanger the scientific record because of a belief that new web–based publishing models might undermine rigorous peer review, facilitate the manipulation or misuse of the electronic article copy, and fail to ensure perpetual access
    - A large part of the chemical literature *does not* reflect a 'stamp collecting' mode of chemical research, and the view that "The emphasis in synthetic papers where these new substances get reported is on the synthesis itself, and the application of the new substance in generating other new substances (Lipkus et al. 2008)" *represents a narrow view* on the field of chemistry
    - The implication that "there is less incentive in chemistry to publish data widely in a reusable manner than other fields of science" *lacks evidence*.

In particular the last series of points are part of the larger rejection of the hypothesis that the domain of chemistry differs in any substantial way from other disciplines' take up of new web based technologies to support scientific communication.

We defer further discussion to the readership of this paper, and only highlight these points here as points for discussion at the follow-up workshop proposed in the next section.

5.2 Aims of Future Workshop

We intend this document to be a starting point for discussion. As acknowledged before, the analysis presented here is initial and incomplete in its coverage of the many different areas of research in chemistry. We suggest that it would benefit from review and discussion by a broad range of professionals with an interest in the matter (academic and industrial chemists, publishers, information service providers, representatives of scientific societies, information and social scientists, funding agencies). We suggest bringing those various stakeholders of scientific communication in chemistry together at a second, international workshop with the following aims:

1) Enhance and extend the analysis presented here and scrutinize the hypotheses about chemistry characteristics and global factors that influence the scientific communication system in chemistry;

2) Gain clarity on
- The value of various new communication models for chemistry;
- The obstacles preventing their realization;

3) Include a wide range of stakeholders in the discussion of how to assess the value of new models for chemistry and of opportunities for joint action;

4) Develop an international research program that
- Addresses open research questions on factors shaping communication cultures in chemistry, determines how to conduct appropriate comparisons between fields and disciplines, and develops an understanding of the dynamics of change processes of the scientific communication system in chemistry and other fields;
- Provides the framework for large-scale deployment of technology with which new models of scholarly communication in chemistry can be tested and evaluated.

We believe that the topic deserves attention to ensure that opportunities provided by new technological capabilities are not missed, but also to ensure that particular, discipline specific contexts within which the scientific communication system evolves are investigated to guide initiatives, highlight non-technical challenges, and avoid failures and waste of resources.

References

ACLS (2006),'Our Cultural Commonwealth The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences'.

Adams, N. (2009) Semantic Chemistry. In Semantic Universe. Online at http://www.semanticuniverse.com/articles-semantic-chemistry.html

Allen, F. (2004), 'High-throughput crystallography: the challenge of publishing, storing and using the results', Crystallography Reviews 10(1), 3-15.

Androulakis S, Schmidberger J, Bate MA, Degori R, Beitz A, Keong C, Cameron B, McGowan S, Porter CJ, Harrison A, Hunter J, Martin JL, Kobe B, Dobson RC, Parker MW, Whisstock JC, Gray J, Treloar A, Groenewegen D, Dickson N, Buckle AM. (2008) Federated repositories of X-ray diffraction images. Acta Crystallogr D Biol Crystallogr. Jul;64(P7):810-4.

Armbruster, C. (2007), 'Moving out of Oldenbourg's long shadow: what is the future for society publishing?' Learned Publishing 20(4), 259-266.

Arms, W. & Larson, R. (2007),'The Future of Scholarly Communication: Building the Infrastructure for Cyberscholarship', Technical report, Washington, DC: National Science Foundation and the Joint Information Systems Committee.

Atkins, D. E.; Droegemeier, K. K.; Feldman, S. I.; Garcia-Molina, H.; Klein, M. L.; Messerschmitt, D. G.; Messina, P.; Ostriker, J. P. & Wright, M. H. (2003),'Revolutionizing Science and Engineering Through Cyberinfrastructure', Technical report, NSF Blue Ribbon Panel.

Batchelor, C.R., and Corbett, P.T. (2007) Semantic enrichment of journal articles using chemical named entity recognition. Proceedings of the ACL 2007 Demo and Poster Sessions, pages 45-48, Prague, June 2007.

Behrens, H. & Lankenau, I. (2006), 'Wissenschaftswachstum in wichtigen naturwissenschaftlichen Disziplinen vom 17. bis zum 21. Jahrhundert', Berichte zur Wissenschaftsgeschichte 29(2).

Benson, D.; Karsch-Mizrachi, I.; Lipman, D.; Ostell, J. & Wheeler, D. (2007), 'GenBank.', Nucleic Acids Research 35, D21.

Berman, F. & Brady, H. (2005). Final report: NSF SBE-CISE workshop on cyberinfrastructure and the social sciences. Retrieved online on 8 April 2009 at http://www.sdsc.edu/about/director/pubs/SBE/reports/SBE-CISE-FINAL.pdf

Berman, F. (2001), 'The Human Side of Cyberinfrastructure', EnVision 17(2), 1.

Biello, D. (2007) Open Access to Science Under Attack. Scientific American, 26 January 2007. Retrieved March 8, 2009 from http://www.sciam.com/article.cfm?articleid=60AADF2C-E7F2-99DF-383C632C90DD1AA5

Birnholtz, J. P. & Bietz, M. J. (2003),Data at work: supporting sharing in science and engineering, in 'GROUP '03: Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work', ACM Press, New York, NY, USA, pp. 339-348.

Bizer, C., Cyganiak, R. and Heath, T. (2007) How to Publish Linked Data on the Web. Free University of Berlin, 2007. Online tutorial available from http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/.

Bonetta, L. (2007), 'Scientists enter the blogosphere', Cell 129(3), 443-445.

Borgman, C. L. (2000), 'Digital libraries and the continuum of scholarly communication', Journal of Documentation 56 (4), 430.

Bradley, J.-C. (2008) CDI-Type I: Chemistry Crowdsourcing using Open Notebook Science. Nature Precedings : doi:10.1038/npre.2008.1505.1 : Posted 9 Jan 2008.

Bradley, J.-C.; Lancashire, R.J.; Lang, A.S.I.D.; & Williams A.J. (2009) 'The Spectral Game: Leveraging Open Data and Crowdsourcing for Education' Journal of Cheminformatics 1(9) doi:10.1186/1758-2946-1-9.

Jean-Claude Bradley, Rajarshi Guha, Andrew Lang, Pierre Lindenbaum, Cameron Neylon, Antony Williams, and Egon Willighagen (2009) Beautifying Data in the Real World. In: Beautiful Data - The Stories Behind Elegant Data Solutions, Toby Segaran, Jeff Hammerbacher (Eds), O'Reilly ISBN 10: 0-596-15711-8 | ISBN 13: 9780596157111.

Braun, T.; Lyon, W. & Bujdoso, E. (1977), 'Literature growth and decay: an activation analysis resume', Anal. Chem 49(8), 682A-688A.

Brown, C. (2003), 'The role of electronic preprints in chemical communication: Analysis of citation, usage, and acceptance in the journal literature', Journal of the American Society for Information Science, 362-371.

Christensen, C. (1997), The innovator's dilemma: when new technologies cause great firms to fail, Harvard Business School Press.

Coles, S.; Frey, J.; Hursthouse, M.; Light, M.; Carr, L.; DeRoure, D.; Gutteridge, C.; Mills, H.; Meacham, K.; Surridge, M.; Lyon, L.; Heery, R.; Duke, M. & Day, M. (2005),'The 'end to end' crystallographic experiment in an e-Science environment: From conception to publication.'

Cope, W, and Kalantzis, M (2009). "Signs of epistemic disruption: Transformations in the knowledge system of the academic journal" First Monday [Online],14(4), 17 March 2009.

Corbett, P., and Murray-Rust, P. (2006) High-Throughput Identification of Chemistry in Life Science Texts. CompLife 2006, LNBI 4216, pp. 107-118, 2006.

Cronin, S. D. L. B. K. (2004), 'Visible, less visible, and invisible work: Patterns of collaboration in 20th century chemistry', Journal of the American Society for Information Science and Technology 55(2), 160-168.

Cronin, B. (2003), 'Scholarly communication and epistemic cultures', New Review of Academic Librarianship 9(1), 1-24.

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M. and Ashburner, M. (2008) ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res. 36, D344-D350.

De Roure, D et al. (2008) The design and realisation of the Virtual Research Environment for social sharing of workflows, Future Generation Computer Systems, In Press, Corrected Proof, Available online 5 July 2008, ISSN 0167-739X, DOI: 10.1016/j.future.2008.06.010. http://www.sciencedirect.com/science/article/B6V06-4SX9FTN-4/2/e44404603ec05e03f8add717d5069d25

de Solla Price, D. J. (1963), Little Science, Big Science, Columbia University Press.

Edlin, A. S. and Daniel L. Rubinfeld. (2004). "Exclusion or efficient pricing? The 'big deal' bundling of academic journals," University of California, Berkeley, at http://repositories.cdlib.org/blewp/art167/, accessed 8 April 2009.

Edwards, P.; Jackson, S.; Bowker, G. & Knobel, C. (2007),'Understanding Infrastructure: Dynamics, Tensions, and Design', Technical report, Report of a Workshop on History and Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures.

Engel, T. & Zass, E. (2007), 'Chemical Information Systems and Databases', Comprehensive Medicinal Chemistry II 3(13), 265-291.

Falagas, M. (2007), 'Peer review in open access scientific journals', Open Medicine 1(1).

Farrusseng, D. (2008) High-throughput heterogeneous catalysis, Surface Science Reports, 63(11), 487-513, DOI: 10.1016/j.surfrep.2008.09.001.

Freeman, P. (2007), 'Is'designing'cyberinfrastructure - or, even,'defining'it - possible?", First Monday [Online]12(6), 4 June 2007.

Frey, J.G., et al (2008), 'The Laboratory Blog-Book: How a laboratory blog notebook has developed to support, and in turn has been influenced by, experimental laboratory practice', 4th International Conference on e-Social Science. http://www.ncess.ac.uk/events/conference/programme/workshop1/?ref=/programme/fri/4cfrey.htm

Frey, J.G. (2009), The value of the Semantic Web in the laboratory, Drug Discovery Today, Vol. 14, No. 11-12. June 2009, pp. 552-561.

Fry, J. & Talja, S. (2007), 'The intellectual and social organization of academic fields and the shaping of digital resources', Journal of Information Science 33(2), 115.

Garson, L. (2004), 'Communicating original research in chemistry and related sciences', Accounts of Chemical Research 37(3), 141-148.

Giles, J. (2007) 'PR's 'pit bull' takes on open access'. Nature 445, 347; 2007. Retrieved online on 8 April 2009 from http://www.nature.com/nature/journal/v445/n7126/full/445347a.html

Giles, M. (1996), 'From Gutenberg to Gigabytes: Scholarly communication in the age of

cyberspace', Journal of politics, 613-626.

Ginsparg, P. (1996) 'Winners and losers in the global research village. In Elliot, R. & Shaw, D. (Eds.) Proceedings of the ICSU Press-UNESCO Conference on Electronic Publishing in Science'. Retrieved online on 25 February 2009 at http://www.library.uiuc.edu/icsu/ginsparg.htm (copy at http://arXiv.org/blurb/pg96unesco.html)

Glänzel, W. & De Lange, C. (1997), 'Modelling and measuring multilateral co-authorship in international scientific collaboration. Part II. A comparative study on the extent and change of international scientific collaboration links', Scientometrics 40(3), 605-626.

Gordon & Gordon (2004) 'Cyber-Enabled Chemistry Workshop Report', CHE NSF.

Guha, R.; Howard, M.; Hutchison, G.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J. & Willighagen, E. (2006), 'The Blue Obelisk-Interoperability in Chemical Informatics', Journal of Chemical Information and Modeling 46, 991-998.

Gunnarsdottir, K. (2005), 'Scientific Journal Publications: On the Role of Electronic Preprint Exchange in the Distribution of Scientific Literature', Social Studies of Science 35(4), 549-579.

Hall, S. R., Allen, F. H. and Brown, I. D. (1991). "The Crystallographic Information File (CIF): A New Standard Archive File for Crystallography", Acta Cryst., A47, 655-685

Harnad, S. (1991), 'Post-Gutenberg Galaxy: The Fourth Revolution in the Means of Production of Knowledge', Public-Access Computer Systems Review 2(1), 39-53.

Hey, T. & Trefethen, A. E. (2005), 'Cyberinfrastructure for e-Science', Science 308(5723), 817-821.

Hilgartner, S. (1997),Access to Data and Intellectual Property: Scientific Exchange in Genome Research, in 'Intellectual Property Rights and Research Tools in Molecular Biology, Summary of a Workshop Held at the National Academy of Sciences, February 15-16, 1996.

Hilgartner, S. (1995), 'Biomolecular Databases: New Communication Regimes for Biology?', Science Communication 17(2), 240-263.

Hine, C. (2008), 'Systematics as Cyberscience: Computers, Change, and Continuity in Science', MIT.

Hoffmann, R. (2007), 'What might philosophy of science look like if chemists built it?', Synthese 155(3), 321-336.

Hoffmann, R. & Laszlo, P. (1989), 'Representation in chemistry', Diogenes 37(147), 23.

Hughes G. et al (2004) 'The Semantic Smart Laboratory: A system for supporting the chemical eScientist'. Org. Biomol. Chem., 2 (22), 3284-3293 DOI: 10.1039/b410075a

Hunter, J.; Henderson, M. & Khan, I. (2007), 'Collaborative Annotation of 3D Crystallographic Models', Journal of Chemical Information and Modeling 47(6), 2475.

Jacobson (2004) 'Chemical Society Draws Fire for Leader's High Pay'. The Chronicle of Higher Eduacation _ Research and Publishing 51, (12), p. A19 Retrieved online on 8 April 2009 from http://chronicle.com/weekly/v51/i02/02a01902.htm

Jones, C.; Darby, R.; Gilbert, L. & Lambert, S. (2008), 'Report of the Subject and Institutional Repositories Interactions Study', CMJ 7, 9.

Kaiser, J. 'Scientific Databases: NIH, Chemical Society Look for Common Ground', Science Magazine, 2 September 2005, DOI: 10.1126/science.309.5740.1473a

Kingsley, D. (2008), 'Repositories, research and reporting: the conflict between institutional and disciplinary needs', in 'VALA2008 - Libraries, Technology and the Future'.

Kling, R.; Spector, L. B. & Fortuna, J. (2004), 'The real stakes of virtual publishing: The transformation of E-Biomed into PubMed central', Journal of the American Society for Information Science and Technology 55(2), 127-148.

Kling, R.; McKim, G. & King, A. (2003), 'A Bit More of It: Scholarly Communication Forums as Socio-Technical Interaction Networks', Journal of the American Society of Information Science and Technology 54, 47-67.

Kling, R. & Swygard-Hobaugh, A. J. (2002). The internet and the velocity of scholarly journal publishing (No. WP- 02-12) http://rkcsi.indiana.edu/archive/CSI/WP/WP02-12B.html

Klump, J.; Bertelmann, R.; Brase, J.; Diepenbroek, M.; Grobe, H.; H□ck, H.; Lautenschlager, M.; Schindler, U.; Sens, I. & W□chter, J. (2006), 'Data publication in the open access initiative', Data Science Journal 5(0), 79-83.

Knorr Cetina, K. (1999), 'Epistemic Cultures - How the Sciences Make Knowledge', Havard University Press.

Lariviere, V.; Archambault, E. & Gingras, Y. (2008), 'Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900-2004)', Journal of the American Society for Information Science and technology 59(2).

Laszlo, P. (2006), 'On the Self-Image of Chemists, 1950-2000', HYLE-International Journal for Philosophy of Chemistry 12(1), 99-130.

Lawal, I. (2002), 'Scholarly communication: the use and non-use of e-print archives for the dissemination of scientific information', Issues in Science and Technology Librarianship 36, 1-16.

Lee, C.; Dourish, P. & Mark, G. (2006), 'The human infrastructure of cyberinfrastructure', Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work, 483-492.

Lipkus, A.; Yuan, Q.; Lucas, K.; Funk, S.; Bartelt Iii, W.; Schenck, R. & Trippe, A. (2008), 'Structural Diversity of Organic Chemistry. A Scaffold Analysis of the CAS Registry', J. Org. Chem 73(12), 4443-4451.

Liu, Y.; Bai, K.; Mitra, P. & Giles, C.L. (2007), 'TableSeer: automatic table metadata

extraction and searching in digital libraries', in *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, (Vancouver, BC, Canada, 2007), ACM.

McCabe, M.J., Nevo, A., and Rubinfeld, D.L. (2006), "The Pricing of Academic Journals" (November 17, 2006). Berkeley Program in Law & Economics, Working Paper Series. Paper 199. Retrieved online on 8 April 2009 from http://repositories.cdlib.org/blewp/art199

Michaelson, B. (2008) 'Viewpoints: The American Chemical Society and Open Access', Issues in Science and Technology Librarianship. Winter 2008. Accessed online on 8 March 2009 at: http://www.istl.org/08-winter/viewpoint.html

Murray-Rust, P. (2009) 'Identifiers: why we need them and when CAS and InChI don't mix', petermr's blog - A Scientist and the Web. Unilever Centre for Molecular Informatics. Accessed online on 8 April 2009 at http://wwmm.ch.cam.ac.uk/blogs/murrayrust/?p=1541

Murray-Rust, P. (2008) 'Big Science and Long-tail Science.' petermr's blog - A Scientist and the Web. Unilever Centre for Molecular Informatics. Accessed online on 8 April 2009 at http://wwmm.ch.cam.ac.uk/blogs/murrayrust/?p=938

Murray-Rust, P. (2008a), 'Chemistry for everyone', Nature 451(7179), 648.

Murray-Rust, P.; Rzepa, H.; Tyrrell, S. & Zhang, Y. (2004), 'Representation and use of chemistry in the global electronic age', Organic & Biomolecular Chemistry 2(22), 3192-3203.

Neumann, E. (2005) 'Finding the critical path: applying the semantic web to drug discovery and development'. Drug Discovery p.25.

Odlyzko, A. (2002), 'The rapid evolution of scholarly communication', Learned Publishing 15(1), 7-19.

Pöschl, U. & Koop, T. (2008), 'Interactive open access publishing and collaborative peer review for improved scientific communication and quality assurance '. Information Services & Use 28, 105-107.

Pöschl, U. (2004), 'Interactive journal concept for improved scientific publishing and quality assurance', Learned Publishing 17, 105-113.

Rafols, I. & Leydesdorff, L. (2009). 'Content-based and Algorithmic Classifications of Journals: Perspectives on the Dynamics of Scientific Communication and Indexer Effects.' Journal of the American Society for Information Science and Technology, 60(9), 1823-1835.

Roosendaal, H. E. & Geurts, P. A. T. M. (1998), 'Forces and functions in scientific communication: an analysis of their interplay'. In 'CRISP 97 Cooperative Research Information Systems in Physics'.

Sale, A. (2006) 'The acquisition of open access research articles'. First Monday [Online], 11 (9), October 2006, URL: http://firstmonday.org/issues/issue11_10/sale/index.html

Schröder, R. & Fry, J. (2007), 'Social Science Approaches to e-Science: Framing an Agenda', Journal of Computer-Mediated Communication 12, 563-582.

Schummer, J. (1999), 'Coping with the Growth of Chemical Knowledge: Challenges for Chemistry Documentation, Education, and Working Chemists', Educación quimica 10(2), 92-101.

Schummer, J. (1998), 'The Chemical Core of Chemistry I: A Conceptual Approach', Hyle: International Journal for Philosophy of Chemistry 4, 129-162.

Schummer, J. (1997), 'Scientometric studies on chemistry II: Aims and methods of producing new chemical substances', Scientometrics 39(1), 125-140.

Schummer, J. (1997a), 'Scientometric studies on chemistry I: The exponential growth of chemical substances, 1800-1995'. Scientometrics 39(1), 107-123.

Searing, S. & Estabrook, L. (2001), 'The Future of Scientific Publishing on the Web: Insights from Focus Groups of Chemists'. Portal: Libraries and the Academy 1(1), 77-96.

Spencer Jr, B.; Butler, R.; Ricker, K.; Marcusiu, D.; Finholt, T.; Foster, I. & Kesselman, C. (2006),'Cyberenvironment project management: lessons learned'. Technical report, NSF Grant CMS-0117853.

Shotton, D. (2009), 'Semantic publishing: the coming revolution in scientific journal publishing'. Learned Publishing 22(2), 85-94.

Suber, P. (2007) 'Open Access Overview'. last revised June 19, 2007, Online at http://www.earlham.edu/~peters/fos/overview.htm

Swan, A. (2008) To Share or not to Share: Publication and Quality Assurance of Research Data Outputs. Report commissioned by the Research Information Network (RIN). Annex: detailed findings for the eight research areas (June 2008).

Taylor, K.R., Essex, J.W., Frey, J.G., Mills, H.R., Hughes, G. and Zaluska, E.J. (2006) The semantic grid and chemistry: experiences with CombeChem. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4, (2), 84-101. (doi:10.1016/j.websem.2006.03.003)

Tetko, I.V. (2005) 'Computing chemistry on the web'. Drug Discovery Today, Volume 10, Issue 22, 15 November 2005, Pages 1497-1500, DOI: 10.1016/S1359-6446(05)03584-1.

Till, J. (2001), 'Predecessors of preprint servers'. Learned Publishing 14(1), 7-13.

Todd, M. (2007) 'Open Access and Open Source in Chemistry'. Chemistry Central Journal 2007, 1:3. (doi:10.1186/1752-153X-1-3)

Trager, R. (2009) 'Web chemistry progresses InChI by InChI.' Chemistry World News Item 06 January 2009. Retrieved online on 8 April 2009 from http://www.rsc.org/chemistryworld/News/2009/January/06010901.asp

Van de Sompel, H.; Payette, S.; Erickson, J.; Lagoze, C. & Warner, S. (2004), 'Rethinking Scholarly Communication -- Building the System that Scholars Deserve'. D-Lib Magazine

10(9).

Walsh, J. P. & Bayma, T. (1996), 'Computer Networks and Scientific Work'. Social Studies of Science 26(3), 661-703.

Ware, M. (2008),'Peer review: benefits, perceptions and alternatives'. Technical report, Publishing Research Consortium.

Whitley, R. (2000), 'The intellectual and social organization of the sciences'. Clarendon Press.

Wilkins, J. (2008), 'The roles, reasons and restrictions of science blogs'. Trends in Ecology & Evolution 23(8), 411-413.

Williams, A. (2008a), 'A perspective of publicly accessible/open-access chemistry databases'. Drug Discovery Today 13(11-12), 495-501.

Wouters, P.; Vann, K.; Scharnhorst, A.; Ratto, M.; Hellsten, I.; Fry, J. & Beaulieu, A. (2008),'Messy Shapes of Knowledge-STS Explores Informatization, New Media, and Academic Work'. The Virtual Knowledge Studio in: The Handbook of Science and Technology Studies'. MIT Press, pp. 319-352.

Xia, J. & Sun, L. (2007), 'Assessment of self-archiving in institutional repositories: depositorship and full-text availability'. Serials Review 33(1), 14-21.

Zuber, P. (2008), 'A Study of Institutional Repository Holdings by Academic Discipline'. D-Lib Magazine 14(11/12), 1082-9873.