What shall we do? Challenges and opportunities of the coming changes in science publishing

Vitek Tracz, presented by Matthew Cockerill
BioMed Central, Science Navigation Group, London, UK

Report from the editors

First of all let me say what this talk is not going to be about, I'm not trying to persuade anybody that Open Access is a good thing or that it is necessary or even that it is possible. What I'm going to ask you to do, for the purposes of this talk, is to simply assume that it is the future and that it *is* going to happen. That may be easier for some of you than for others but let's see where it takes us as a thought experiment.

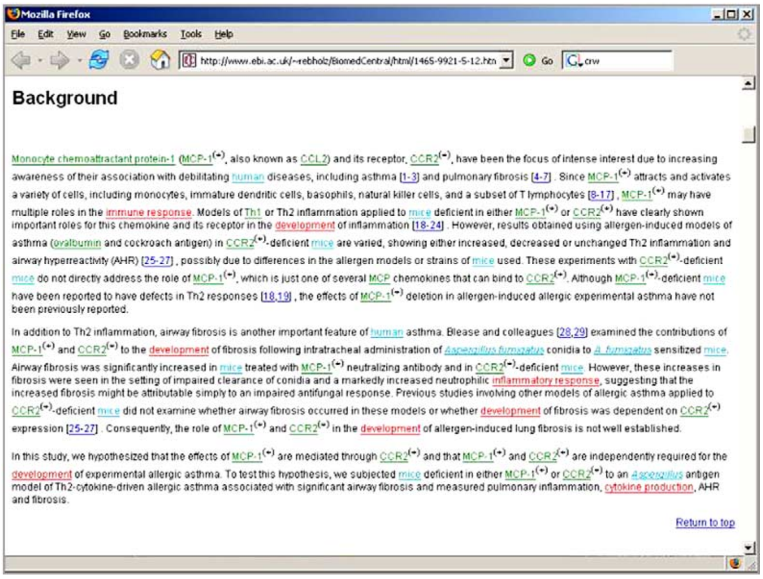
What does this mean for science publishing? One of the things it means is that, in a way, it's the end of the easy life. Selling academic research is a pretty easy thing to do because the academic community needs that research. Essentially it sells itself. But in the short term, in a world where research flows freely, publishers will need to demonstrate much more clearly that they are adding value in the services they offer if they are to survive. Looking to the future, it may still be tempting for publishers to defend the attractive and profitable model of selling access to research, but that carries the very big danger of preventing them from moving forward, and may lead them to fall behind in terms of developing the services that will add value in the long term. So what do publishers need to do if Open Access *is* going to happen? They must re-invent the role of the publisher. And I think it's interesting to look at the analogy of the internet as a general principle. What is special about the internet is not just the electronic transfer of data. We have known for many, many years how to send data from place to place electronically. But the birth of the internet provided a set of neutral standards which allowed data to flow in packets completely freely, whatever kind of information that data carried. And several things have resulted from that. One of them is the growth in the business of sending packets around, providing hardware to manage those packets of data. Companies like Cisco who make routers have done very well, but their business has become a kind of commodity business where it's all about offering the best priced performance. On the other hand, those open standards have also created entire new markets offering the opportunity to add value in lots of different ways by building on the foundation of the free-flowing data. So you have companies like Ebay and Skype; entirely different services taking as their starting point the fact that data can flow freely whatever it carries, and using it for trading things or for having fun conversations. And what we at Science Navigation Group see in the future for publishing is a need to look at some of those higher level services where publishers can offer ongoing value to the scientific community and the research communities.

So in my talk I want to cover a few ideas for the types of services that publishers may be able to offer those communities in the future. They fall broadly into three categories. They are linked by the fact that the future of publishing is all about recognizing that it's not paper. Electronic publishing makes lots of new things possible and simply providing articles in electronic form, as if they were electronic paper, isn't the long term future. It's certainly progress, it's wonderful but it's not where the action will be in the future. One area where the action is starting to happen has to do with adding meaningful structure – and 'semantics' is a very prevalent buzz word right now – to the articles, and not just to the articles but also to the raw data. Making a set of the data isn't just strings of zeros and ones, and the articles aren't just strings of words, but they actually have a structure which represents the knowledge that they contain.

Another active area is the development of tools to actually mine that knowledge and to model the current state of scientific knowledge; to draw inferences, to figure out new hypotheses, to identify inconsistencies and see if those inconsistencies are informative. Lastly, an area that has become quite prominent in the last couple of years, and which seems to have a lot of potential, is figuring out how to take advantage of the collective knowledge of the community in ways that go beyond the traditional model of publishing a scientific article. So there are already various ways in which the web is being used to take collective knowledge and gather it together in non-traditional ways.

First of all, in terms of adding structure to scientific research articles, lots of publishers have experimented over time with trying to collect more structure in scientific articles at the time of writing. It's a big challenge. This is an example of a collaboration between BioMed Central and Wolfram Research, the people who make Mathematics. And it's a structure offering tool: the author has created an entire scientific manuscript where all of the structure to produce a full excel is now retained in the article. And it includes the mathematics as live equations; equations that can be copied and pasted and solved. You can enter all the details of the authors and the affiliations in a nice structured way. Obviously the challenge with all these things is what will motivate authors to use them? It may be nice to collect more structured information so you can do more things with the articles when they are published electronically, but why would an author do this? It seems pretty clear that the answer has to involve having an infrastructure in place so that authors get immediate benefit from having created an article that has structure, and also providing tools that mean that when an author is creating structured articles using these tools it's actually easier, quicker and involves fewer errors than using traditional unstructured tools. And this sort of thing is starting to happen. Just an example of how you make fewer errors when using smart tools: I think most people are familiar with typing to their email programme, not having to remember every single email address but typing the first name of somebody and then getting a list of all the different people who they know with that name. And that works, it means you don't make mistakes with your email addresses and it's quicker. The same thing can potentially apply dynamically to genomes and chemical structures. So you get two benefits; the author makes fewer mistakes and wastes less time and you are also able to add structure right at the time of creating the manuscript, which can then flow through an entirely electronic process all the way to the final electronic version of the article and on to the reader's desk where the reader can then work with that structured content, be it a chemical structure or anything else.

That leads into this buzzword of semantic enrichment, which is the broader concept of making sure that all of the entities, of the facts expressed in scientific articles are not just expressed in a way to make them readable by humans but are expressed in a way to make them readable by all kinds of computer programs that may be sifting through publications and the web. I have already said that in an ideal way, in an ideal world, you would capture this from the author, who is in the best position to express what


Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.ebi.ac.uk/~rebholz/BiomedCentral/html/1465-9921-5-12.htm
Go
crw

Background

Monocyte chemoattractant protein-1 (MCP-1$^{(-)}$, also known as CCL2) and its receptor, CCR2$^{(-)}$, have been the focus of intense interest due to increasing awareness of their association with debilitating human diseases, including asthma [1-3] and pulmonary fibrosis [4-7] . Since MCP-1$^{(-)}$ attracts and activates a variety of cells, including monocytes, immature dendritic cells, basophils, natural killer cells, and a subset of T lymphocytes [8-17] , MCP-1$^{(-)}$ may have multiple roles in the immune response. Models of Th1 or Th2 inflammation applied to mice deficient in either MCP-1$^{(-)}$ or CCR2$^{(-)}$ have clearly shown important roles for this chemokine and its receptor in the development of inflammation [18-24] . However, results obtained using allergen-induced models of asthma (ovalbumin and cockroach antigen) in CCR2$^{(-)}$-deficient mice are varied, showing either increased, decreased or unchanged Th2 inflammation and airway hyperreactivity (AHR) [25-27] , possibly due to differences in the allergen models or strains of mice used. These experiments with CCR2$^{(-)}$-deficient mice do not directly address the role of MCP-1$^{(-)}$, which is just one of several MCP chemokines that can bind to CCR2$^{(-)}$. Although MCP-1$^{(-)}$-deficient mice have been reported to have defects in Th2 responses [18,19] , the effects of MCP-1$^{(-)}$ deletion in allergen-induced allergic experimental asthma have not been previously reported.

In addition to Th2 inflammation, airway fibrosis is another important feature of human asthma. Blease and colleagues [28,29] examined the contributions of MCP-1$^{(-)}$ and CCR2$^{(-)}$ to the development of fibrosis following intratracheal administration of *Aspergillus fumigatus* conidia to *A. fumigatus* sensitized mice. Airway fibrosis was significantly increased in mice treated with MCP-1$^{(-)}$ neutralizing antibody and in CCR2$^{(-)}$-deficient mice. However, these increases in fibrosis were seen in the setting of impaired clearance of conidia and a markedly increased neutrophilic inflammatory response, suggesting that the increased fibrosis might be attributable simply to an impaired antifungal response. Previous studies involving other models of allergic asthma applied to CCR2$^{(-)}$-deficient mice did not examine whether airway fibrosis occurred in these models or whether development of fibrosis was dependent on CCR2$^{(-)}$ expression [25-27] . Consequently, the role of MCP-1$^{(-)}$ and CCR2$^{(-)}$ in the development of allergen-induced lung fibrosis is not well established.

In this study, we hypothesized that the effects of MCP-1$^{(-)}$ are mediated through CCR2$^{(-)}$ and that MCP-1$^{(-)}$ and CCR2$^{(-)}$ are independently required for the development of experimental allergic asthma. To test this hypothesis, we subjected mice deficient in either MCP-1$^{(-)}$ or CCR2$^{(-)}$ to an *Aspergillus* antigen model of Th2-cytokine-driven allergic asthma associated with significant airway fibrosis and measured pulmonary inflammation, cytokine production, AHR and fibrosis.

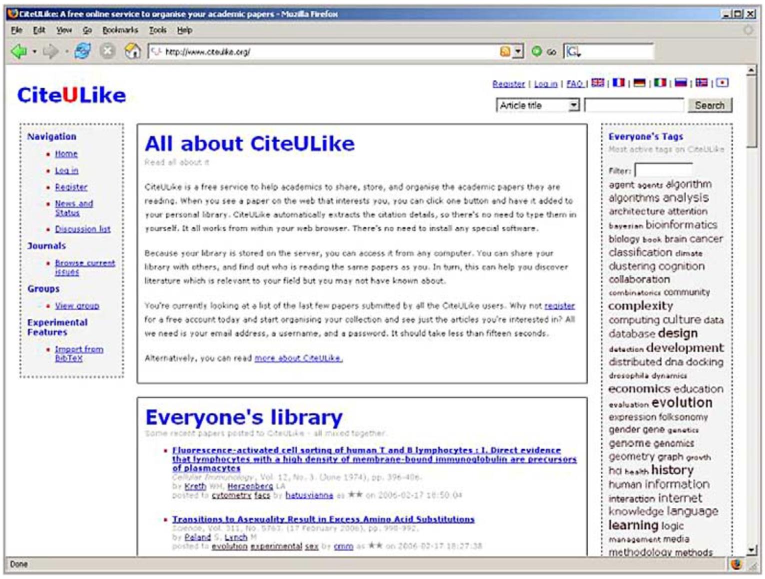
Return to top

they mean, but the fallback is to be able to use mining tools which attempt to automatically assign the correct concepts to the correct terms.

Very importantly, it's not just about taking this technology and applying it to traditional scientific articles, it's also about looking at the types of data that are being created all the time, like the results of micro rates experiments or clinical trial results and seeing that this data has a lot more potential to be used. Not all that is being published right now. If we can make it available in a structured way so it can be sifted, you will be able to find just the data you are looking for – and the matter data which tells you how that experiment can be done. Research funding can be used a lot more efficiently if the results of the experiments being done can be reused. So it's quite exciting that finally, after many of years of promises with not very much delivered, the appropriate standards and technologies are starting to emerge. For example: in the U.S. (a collaboration between Stanford and UCSF or Berkeley I believe) there is a national centre for biomedical ontology which is taking on the responsibility for blessing certain authoritative ontologies, which are sort of controlled vocabularies for representing biological or medical concepts. That kind of standardisation, combined with the latest generation of text-mining tools, puts us in a position to have a lot of articles and data which are not just words, but actually represent the concepts in a way that the new generations of tools can mine.

So this is an example, just to show you, of the abstract from a BioMed Central article that has been run through one of the European bio-informatic institutes' latest concept matches. The different colours represent different standard ontologies which have been used to identify different species names, different drugs, different diseases, and as more and more of this standardisation happens, there's much more potential to add value to content in that way. For example, the Open Access to content, such as that published by BioMed Central, is really helping to facilitate this research work. BioMed Central Corp has more than 15,000 articles available and has greatly helped a lot of the researchers who are working on this kind of full text mining.

The last point I promised to touch on is the idea of how a community's collective knowledge can potentially be gathered together using modern technology in a somewhat innovative way. I think the best known example right now would be Wikipedia, which takes the whole encyclopedia concept, which has traditionally been very top down and all based on assigning the relevant experts, and turns it upside


CiteULike: A free online service to organise your academic papers - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.citeulike.org/
Go
CiteULike
Register | Log in | FAQ |
Article title
Search
Navigation
Home
Log in
Register
News and Status
Discussion list
Journals
Browse current issues
Groups
View group
Experimental Features
Import from BibTeX
Everyone's Tags
Most active tags on CiteULike
Filter:
agent agents algorithm algorithms analysis architecture attention bayesian bioinformatics biology book brain cancer classification climate clustering cognition collaboration combinatorics community complexity computing culture data database design detection development distributed dna docking drosophila dynamics economics education evaluation evolution expression folksonomy gender gene genetics genome genomics geometry graph growth hci health history human information interaction internet knowledge language learning logic management media methodology methods
Done

All about CiteULike

Read all about it

CiteULike is a free service to help academics to share, store, and organise the academic papers they are reading. When you see a paper on the web that interests you, you can click one button and have it added to your personal library. CiteULike automatically extracts the citation details, so there's no need to type them in yourself. It all works from within your web browser. There's no need to install any special software.

Because your library is stored on the server, you can access it from any computer. You can share your library with others, and find out who is reading the same papers as you. In turn, this can help you discover literature which is relevant to your field but you may not have known about.

You're currently looking at a list of the last few papers submitted by all the CiteULike users. Why not register for a free account today and start organising your collection and see just the articles you're interested in? All we need is your email address, a username, and a password. It should take less than fifteen seconds.

Alternatively, you can read more about CiteULike.

Everyone's library

Some recent papers posted to CiteULike - all mixed together.

- Fluorescence-activated cell sorting of human T and B lymphocytes : I. Direct evidence that lymphocytes with a high density of membrane-bound immunoglobulin are precursors of plasmacytes
Cellular Immunology, Vol. 12, No. 3. (June 1974), pp. 396-406.
by Kreth WH, Herzenberg LA
posted to cytometry facs by hatusvianna as ★★ on 2006-02-17 18:50:04
- Transitions to Asexuality Result in Excess Amino Acid Substitutions
Science, Vol. 311, No. 5763. (17 February 2006), pp. 990-992.
by Paland S, Lynch M
posted to evolution experimental sex by cmm as ★★ on 2006-02-17 18:27:38

down, and says: what can you achieve by having something of a free for all, but with a filtration system, that allows the best things to survive? And there is also an interesting debate right now between the encyclopedian puritanical nature as to the role of American use of e-moles.

There is also a lot of work going on with the idea of applying Wiki-technology to biomedical publishing which I'm not going to discuss. But I'll just mention a couple of other examples of how community based approaches could be of interest.

This is a small scout website, produced by a student I believe, using some simple technology and the available data on the internet. It brings together information from various different sources and it combines it to identify which recently published articles have been cited by what had been referred to. (Not exactly cited in a traditional publishing sense, which would take maybe six months to a year for the whole process to happen.) Maybe it appeared in a news story in the Guardian, or has been mentioned by another scientist in their blog. And it takes all this information from many different sources and says: look – here's what is interesting at the moment on the web in any particular field – or search for the keyword. Here is an interesting thing. And this isn't based on any organisation doing a lot of expensive work, hiring people to choose what's most interesting. It's based on sifting in the same way that Google sifts the web. This is taking a very specific area and using the available tools via access feeds and saying what's hot in science right now.

Another example which are, in a way, even more interesting in their potential, are tools simply managing bibliographies, managing reference information. It's a central part of most scientists' life. It's both for the purpose of keeping track of the research that's important to them and also in terms of preparing manuscripts.

This is increasingly moving online and so people are managing their bibliographical information via website. This is a free website for doing that. But once people start to manage that bibliographic information online it opens up many other possibilities. For example, on a system like this, BibTeX, you share your bibliographic information with others. If you choose to do that it means the information about what you find interesting is contributed to the community's knowledge of what everybody else finds to be interesting. And you can identify other individuals whose interests seem to overlap with yours or whose opinions you trust, and you can automatically be alerted to articles which they find to be of interest.

So all of these different ways have taken the collective knowledge and the collective activities going on in science and combined them with technology to add value. And I think that publishers will need to embrace a lot more of these different possibilities for adding value if they are to move forward. I don't think that by any means all their activities will be replaced by technology or tools. They will still play an important role. We believe there is also a very, very important role for publishers in editorial curation. A lot of these additional layers of services will need curation to ensure high standards. For example: Science Navigation Group has a civil faculty 1000 which involves 1500 different scientists all evaluating the literature. And that goes through a quite sophisticated editorial process of quality control. But there is going to be a transition from focussing on the low level raw research, which is almost equivalent to the packets of data going across the internet, to focussing on the higher layers of information and how publishers can generally add value.