The new scholarly universe: are we there yet?

Based on a paper presented at the UKSG One-Day Conference, London: 'A Problem Shared? Understanding shared services and the drive for efficiency in scholarly communications', on 16 November 2011

Despite long-standing threats of disruption to scholarly publishing, the community has remained remarkably unchanged over the last two decades.

However, underlying this apparent inertia, there is an alternative approach that may emerge over the next few years as known problems move into the foreground, and a number of key social and technical issues are resolved. Furthermore, the industry of science and scientific communication are not isolated from the current economic epoch: collaboration and cost-effectiveness are strong drivers for change.

The technological components of this new scholarly universe – natural language processing, semantic technologies such as taxonomies, ontologies and linking – will come together with collaborative innovations inspired by Web 2.0 social media.

The final components in recreating a new publishing model are identity and reward, and these are being resolved by the forthcoming Open Research and Contributor ID (ORCID) project and the altmetrics movement respectively.

Background

For 20 years, the publishing world has been living under the threat of disintegration in the face of new technology. This threat has clearly failed to materialize, and yet that which has changed within scholarly publishing has been structurally trivial. Exchangeable metadata have displaced printed books of abstracts, open access publishing has subtly altered the onus on 'who pays', and occasionally, articles are enriched with data or a video. The game of publishing still consists of passing a chunk of writing around the world, even if it is now expressed in pixels rather than pigment.

The last two decades have witnessed a continuing evolution, not a revolution.

As a consequence, our industry has remained remarkably unchanged: scholarly publishing has all the appearance of a monolith, a construction that is impervious to change, expandable and resilient.

If we are to extrapolate from the recent past, beyond the present day, we might conclude that there is no obvious threat to the profession.

Publishing is a very comfortable world for all the people and institutions that inhabit it. It is a known world where the mechanisms are understood by all stakeholders; and it is enriched with peripheral industries of performance measures, marketing, fulfilment and syndication.

Emerging threats to the status quo

However, there are a number of emerging issues that threaten to radically change traditional publishing models, and these may combine with current

MICHAEL TAYLOR
Research Specialist
Elsevier Labs

"The last two decades have witnessed a continuing evolution, not a revolution. As a consequence, our industry has remained remarkably unchanged: scholarly publishing has all the appearance of a monolith …"

technological and social trends to finally realize the disruption that has been promised for so long. These can be summed up as:

- the number of papers being published is increasing at an ever growing rate

- the growth of publishing activity in Asia is a challenge to our ability to organize data

- the amount of data being produced is overwhelming

- there is increasing pressure from funding organizations to reduce costs

- newer generations of researchers expect to be able to collaborate.

Scholarly publishing continues to grow – even though the economic upheavals have taken their toll. This is especially true for Asiatic output, where the growth rate continues to accelerate.[1] There are over one million papers published every year – and that total is growing at a greater rate than ever before. New disciplines are being created – sometimes transient and reactive – and the result is that we have an industry so fluid that the total number of new journals published every year can only be estimated with a margin of error of 2,000. And for an industry largely focused on Latin character metadata, the disproportionate growing in Asian publishing adds a new and challenging layer to the management of the infrastructure.

New advances in scientific technique and computer science have meant that the amount of data being produced (particularly in medicine, physics and genetic research) is so massive that it is a serious challenge to the IT infrastructure of research organizations.[2]

This growth – in all senses – means that it is getting harder to manage the constant stream of information, of data, of critical knowledge – and to correctly judge the relevance and accuracy of all of these findings.

Partially as a consequence of the increasing complexity and volume of publishing, there is increasing criticism that traditional publishing models are also failing to maintain their core values: replicability and validity. People accustomed to open social networks expect to be able to make their own judgements, and are increasingly critical of delivery mechanisms that obfuscate methodologies and data.

To understand the potential fracture points of the industry, we need to understand its primary functions – both explicit and implicit. The motivation to publish was originally to disseminate knowledge, and then to register the provenance of discovery. But as the enterprise expanded beyond a relatively small group of people, it evolved, and became more industrialized. Disciplines began to coalesce, and as quality became an increasing concern, peer review evolved to ensure a certain level of scholarly work. Journals and editorial boards brought focus and topicality, enabling scholars to focus on relevant articles.[3]

> "To understand the potential fracture points of the industry, we need to understand its primary functions – both explicit and implicit."

It is logical to maintain allegiance to the status quo for several reasons, not least of which is the existing infrastructure. All the stakeholders have a lot of investment in the way we communicate today – for many of us, it is what we do – but perhaps the most important reason for its enduring success has been that publishing has become the key method of judging success. Researchers get credited – through promotion, invitations to conferences, grants and tenures – on the basis of what they feed into this system.

These threats and weaknesses have been getting more serious over the last two decades, but it is only when coupled with the desire and the ability to change that the threat becomes near. The ubiquity of computing power and social networking fulfil these last two requirements.

New developments

When we consider the future of scholarly communication, we can categorize new developments by using three facets: we can make the constituent elements of the published

article usable in their own right; we can change how articles can relate to the scholarly universe; and we can see how the contributors to the article can collaborate and compete with each other.

These first two facets are intimately connected – if an article is to be richly linked to the rest of the scholarly universe, then its constituent elements will be exposed and should be reusable.

Scholarly articles tend to fall into a sequence of known sub-elements (abstract, methodology, data, etc.). Some parts have been extracted and used away from the main article for many years – title, abstract and references, whilst some are just finding their niche.

Computerized experimental pathways

Computerized experimental pathways are among the first to become publishable in their own right. Many disciplines have their own platforms that enable the computerization or computer-management of the experimental process, such as Taverna[4]. Researchers are able to construct complex analyses that run over data stored either privately, or on open data repositories, thus enabling the researcher to replicate and modify the experimentation process. These pathways are stored in a reusable format and may be shared with other researchers, via services such as MyExperiment[5] that lean heavily on the social experience of networking sites.

Citation count

Although citation count provides the raw ingredient upon which scholarly impact is calculated, these figures are only a raw count and do not take into account the meaning of the citations. There is significant work to classify citations into several categories: for example, whether a reference supports another paper, or contradicts it, or develops from it[6]. Citation analysis offers the real possibility of qualifying the current use of citation count, which underpins current performance metrics. This would enable us to contextualize the number of citations, so judgements could be made about whether a paper was truly influential, or whether its apparently high impact was a reflection of the errors within.

Additional tools and platforms

Citation counts are only one of the ways in which an article is connected to the rest of the world. We can also expose the subject and claims made within an article to the outside world, by using standardized terms to define their contents and inter-relations.

There are standardized, formal ways of identifying and defining inter-relationships of many of the elements of scientific research – taxonomies, ontologies and databases of entities. These elements – that form the backbone of the Semantic Web – can be used to define many basic concepts, for example, chemical compounds and their synonyms, genes and gene effects, medical conditions, definitions of people and places and years.

This enables a contributor to refer to *Gene A* and *Condition B* and *Drug C* and be absolutely certain about what they mean by using those definitions. To get deeper into the nuance of the meaning, there are teams working in the area of scientific discourse, and their work goes further, lets us say: "*Gene A* has *Effect X* on *Drug C* when *Person type Y* has *Condition B*".

What these tools give research is a formal, computerized way of exposing what is being said inside an article, and how it relates to the rest of the scholarly universe – quite independently of language, or localized expression. It gives us the possibility of developing tools and platforms that can look and retrieve things that are absolutely relevant, without any doubt or coincidence of meaning.

> "What these tools give research is a formal, computerized way of exposing what is being said inside an article, and how it relates to the rest of the scholarly universe …"

Connectivity

Interchangeable experimental methodologies, citation analysis and the work in describing the arguments and facts within an article are all ways

of connecting the article to the rest of the universe whilst optimizing the usefulness and reusability of the article's constituent elements. The third facet – how to connect the contributors that create these works – may seem superficially unrelated.

Over the last few years, there have been several platforms that have been described as the 'Facebook for scientists'. Some of them have been apps on Facebook itself – but very few of these specialized platforms seem to survive very long. In part, this is because researchers use social networking tools in the same manner as everyone else, and adapt their behaviour to fit – for example, using Twitter to highlight published work whilst socializing on Facebook.

The social platforms for researchers that have succeeded have been well-differentiated. Tools such as Citeulike, Mendelay and Zotero are becoming established as mostly social reference management systems, ways of spreading knowledge and sharing it with colleagues or with the general public. Some go a step further and take on a notebook function too, enabling people to share experimental details within a team. MyExperiment allows for exchange of experimental workflow.

So rather than talking loosely about the 'Facebook for scientists', the platforms we see developing are those that enable the social exchange of both tiny fragments of data about reading recommendations, criticism, claims and suggestions – as well as executable experimental objects.

Identity and reward

All these elements of an alternative scholarly publishing model exist now: some are mature technologies or well-discussed standards with several generations of revision behind them, and some of them are only being explored as research papers and downloadable developmental code. But in all instances, they are currently not plugged into any measurement of human achievement, or automatically referenced to our scholarly records.

What we seem to be missing from the emerging publishing model is a clear system to define merit and provide rewards. Many innovations enable distribution and collaboration, but if people are to collaborate – especially if they are to spend additional time making their research communications more transparent and usable – how does the system ensure proper identification and how does it use that identification to reward those efforts?

> "What we seem to be missing from the emerging publishing model is a clear system to define merit and provide rewards."

There may be ways of identifying a particular gene or a particular drug. We can use a DOI to identify a paper. But up until now, there has been no way to guarantee that you can identify an individual. Of course, you could use a Scopus ID, or an institutions' web page – but these things are not necessarily open to all or persist beyond our working relationship. There is no open and universal system, no system that guarantees that a scholar can be identified, and identified uniquely and accurately.

Fortunately, later in 2012, the Open Researcher and Contributor ID organization (ORCID)[7] will be issuing unique identifiers for researchers and research contributors. You will be able to self-claim your ORCID ID and establish your scholarly record in a public arena, with records being pulled in from publishers, institutions and professional organizations – and corrections you make on orcid.org will be fed back to its partner organizations.

Your ORCID ID will go beyond the traditional publishing world. If you make any research claims, say, about gene/condition interaction, or you publish a large database, or make a comment or clarification on a colleague's paper – you will be able to freely publish this wider activity on a public and authoritative platform.

This free and authoritative platform will be open to all interested parties and the database of wider contributions could be used to address the issue of the failings of the current impact factor calculations. ORCID will enable emergent researcher activity to be combined with traditional citation figures. Calculating impact metrics is a science, whereby hypotheses and

predictions can be made about academic importance. This area is under active development by the altmetrics movement[8].

So, this work goes some way to completing an alternative scholarly universe: ORCID plus a new analysis of metrics will make sense of an unrelated confusion of initiatives and offer a structure for recognition and reward for the researchers who invest in their own work. Perhaps it will enable a renewed focus on scholarly activity, and enable less time to be taken on the curation of career statistics.

Conclusion

The relationship between scholarly research trends and the economic climate has been fully demonstrated by Science-Metrix[9]: The organizations that promote scholarly work are as affected by the economic climate as anyone else. Consequently, they have a keen interest in reducing cost and in advancing innovation and are favouring more open and complete forms of publishing in our new scholarly universe. New kinds of collaboration are becoming not only possible and desirable, but mandatory too.

The time is right for change: for things that might have once seemed strangely esoteric, diverse and unconnected to come to the forefront. Technologies such as natural language processing, ontologies and taxonomies are uniting with social media and open standards and the unification is driven by social and economic change.

We need all of these to become connected in our new scholarly universe. And this is where the future for publishing lies – in curating and connecting the content and contributors of the scholarly communication.

> "New kinds of collaboration are becoming not only possible and desirable, but mandatory too. The time is right for change: for things that might have once seemed strangely esoteric, diverse and unconnected to come to the forefront."

References and notes

1. Science-Metrix, *30 Years in Science: Secular Movements in Knowledge Creation,* 2010, Canada
http://www.science-metrix.com/30years-Paper.pdf (accessed 6 January 2012).

2. Big guns turn sights on cancer-causing genes, *The Register*:
http://www.theregister.co.uk/2011/11/17/storage_research/ (accessed 6 January 2012).

3. Clarke, M, *Why Hasn't Scientific Publishing Been Disrupted Already?,*
http://scholarlykitchen.sspnet.org/2010/01/04/why-hasnt-scientific-publishing-been-disrupted-already/ (accessed 10 January, 2012)

4. Taverna:
http://www.omii.ac.uk/wiki/Taverna (accessed 11 January 2012).

5. myExperiment:
http://www.myexperiment.org (accessed 11 January 2012).

6 Two very influential researchers in the area of citation are Agnes Sandor of Xerox and David Shotton of Oxford University.
Agnes Sandor:
http://www.xrce.xerox.com/Research-Development/Document-Content-Laboratory/Parsing-Semantics/People/Agnes-Sandor/(language)/eng-GB
(accessed 6 January 2012).
David Shotton:
http://www.zoo.ox.ac.uk/staff/academics/shotton_dm.htm (accessed 6 January 2012).

7. ORCID:
http://orcid.org/ (accessed 6 January 2012).

8. Altmetrics: a manifesto:
http://altmetrics.org (accessed 6 January 2012).

9. Science-Metrix, ref. 1.

Further reading

For more information on the future of scientific publishing, there is a whitepaper on http://www.force11.org/white_paper

Michael Taylor, Research Specialist, Elsevier Labs
Elsevier Limited, UCD/Labs 2/East, The Boulevard, Langford Lane, Kidlington, Oxfordshire OX5 2GB UK
Tel: +44 7775004449 │ E-mail: mi.taylor@elsevier.com │ Twitter: #herrison