Key Issue

The technologies that oil the supply chain

Based on a presentation given at the UKSG one-day conference 'Usercentric: new strategies for scholarly communication', London, November 2010

ROSS MACINTYRE
Senior Manager
Mimas, The University of Manchester

For the Key Issue this time *Serials* asked Ross MacIntyre to present an overview of some of the standards that underpin the information industry. What he delivers is an all-you-can-eat banquet, to whet your appetite for the intricacies of these key technologies.

As most of the standards used are discrete technologies, which tend to be talked about in disjointed ways, I've arranged them on a series of platters. I'll start off with an appetiser, move on to a main course, we'll then have the cheese section and finish with dessert and digestif.

Appetiser

Twenty years ago, the web was born[1]. Let's start by talking about stuff that is based on the rather unassuming uniform resource locator (URL) format[2], for example:
http://hostname[:port]/path[?searchwords] .

Let's recall some of the URL-based linking mechanisms. We started off in the good old days with explicit file names. For a journal article, it might include the author name(s), for example:
http://www.ariadne.ac.uk/issue44/petrie-weisen/intro.html .

This took (and still takes) you directly to the article. However, any subsequent amendment may have resulted in losing the link ('http 404 Not Found'), and the convention was not predictable.

An improvement came when the URL included a property of the thing that was being pointed to. An ISSN might be appropriate when creating a journal-level link, for example:
http://www.jstor.ac.uk/journals/10624783.html.

Things could improve further if the URL contained either a recognized identifier for the object or a standard set of attributes of the item.

The identifier route was successfully taken forward as the digital object identifier (DOI)[3], which is literally a persistent identifier for a digital object; assigned to that object at or around the time of its creation. Simplistically, you could think of it as like a car chassis number, i.e. as long as that object exists, the number is associated with it. Importantly, the chassis number does not change just because the owner does. CrossRef[4] was the application that made the DOI 'come alive' and had a wonderful clarity of purpose, essentially 'get me to the article'. The CrossRef database contains DOIs and metadata, including the online location(s) of the object. You just match the metadata and you'll get the DOI, which can be embedded within article references, even for material that is not published by you.

This was a publisher-initiated initiative, but it has had wide acceptance, so much so that, for example HEFCE (Higher Education Funding Council for England) accepted DOIs as identifiers of published material in the RAE (Research Assessment Exercise) submissions and it will be intrinsic to its successor, the REF (Research Excellence Framework).

The OpenURL is slightly different. There are effectively two parts: a 'Base URL', which refers to the location of OpenURL resolver software deployed by an organization, such as an academic institution's library, and a 'Context Object', which describes the item of interest using an agreed syntax. (If you look at the URL format given at the start of this article, the Base URL is to the left of the '?' and the Context Object to the right.)

The idea behind this, and linking it to a particular institution, is so that they are best placed to identify further material that concerns the thing that's in your Context Object, the item of interest. And you will be able to get through by virtue of being a member of that institution. (The best example of being frustrated at asking things, and being turned down, is Monty Python's Cheese Shop sketch.[5])

The OpenURL was taken forward by NISO (Z39.88-2004) and recently, a NISO workgroup, IOTA[6], has been formed to look at the assessment of the quality of the contents of the Context Object.

Main course

Increasingly, libraries found that less and less of their collection was physically in the library. It was on the publisher's site and often elsewhere too. When users used these various sites, the usage data was captured there. If the library wanted to collect the usage data, they had to download them separately from each publisher's site, quite a tedious process: logging into admin interfaces (each with unique different usernames, structures, etc.), then load the downloaded data into some kind of consolidation tool to calculate the usage.

However, the real 'pain points' were around the transfer of the data. There were different formats for the data, different terminology used, different counting rules. A cross-sector body, the Publisher and Library Solutions Group (PALS) established COUNTER[7] in 2002, which defined a Code of Practice covering the terminology and the report formats. The reports had to be made available in XML, much better for machine processing.

Retrieving that data was also time consuming. The idea behind SUSHI[8] was to come up with a way of automating the transfer of those (COUNTER-compliant) usage stats to the library. SUSHI (NISO standard Z39.93-2007) is a transfer protocol, meaning that both publisher and library need to be configured to run SUSHI. This appears to have been an inhibitor to change, as currently very few libraries have configured SUSHI, even if it is supported within their systems, so, consequently, publishers are not supplying stats via SUSHI in the vast majority of cases.

Once the library has got the stats, they may want to do some kind of cost/use calculation, so they need price information. Where from? Well, it's from the publisher, or their subscription agent(s), or consortium administrator. There is then some kind of calculation done, so that the cost for use is determined.

Now, it would be sensible if standards applied, so that the price and associated information was presented in a standard format. A trade standards body called EDItEUR[9] is behind ONIX, a family of XML-based standards relevant to books and serials. ONIX-SPS defines a family of messages used for transmitting information about serials products and subscriptions, including the price information. Using ONIX-PL covers 'Publications Licenses', i.e. all the stuff to do with licensing. So there is a way of that data being formatted in a standard, predictable fashion, which will be especially useful if the library has invested in an electronic resource management (ERM) system.

If the library does have an ERM, sitting behind it there will be a 'knowledge base' (KB) containing things like lists of the journal titles available from various publishers and aggregates. Where does that information come from? Well, once again, it's coming from 'outside' and there are potential 'pain points' on these data flows, eased if they are standardized transmissions. Non-standard reports and inaccurate lists cause problems in the KB. And you're using the knowledge base not just for 'A to Z' lists of resources, but also to drive your link resolver.

That's where KBART (Knowledge Bases and Related Tools)[10], a UKSG/NISO initiative comes in, standardizing on the data provided to and by a knowledge base, and advocating best practice. It is trying to improve the data that's there to support OpenURL linking, improving the reliability and quality.

So we've got standards applying across multiple information flows, which if exploited fully, offer a nicely oiled chain – technology working with and for the community. However, if/when things change in the environment, then that change has to be fully reflected so that the technology continues to work. Reflecting the change typically requires various 'things' to be (re-)configured.

Consider journal titles changing publisher. In 2007, EBSCO logged 2,667 unique titles that moved from one publisher to another, which required EBSCO to make 20,000-25,000 changes to their title file. And the libraries themselves get notifications that a title has moved from publisher A to publisher B, who may or may not be somebody

they actually deal with. Additionally, there may be some conditions associated with that transfer.

This is where Transfer[11] comes in. It is a publisher-championed initiative, taken forward by UKSG. The goal of Transfer is to enable collaborative working to ensure no one loses access, or, if there is some kind of disruption, that it is kept to a minimum. It places explicit obligations on the publisher who is transferring the title and the publisher who is receiving it. These include: online access, subscription lists, URL of the journal, DOI update, etc. It has actually been quite successful, but it does need to be pushed further and all members of the community should look into Transfer and support it.

Cheese

The next platter consists of two COUNTER-related initiatives, but they came from different sectors.

Firstly, there was general interest in the possible establishment of a 'Journal Usage Factor', as an additional measure to Thomson Reuters'[12] 'Journal Impact Factor', i.e. a factor based on downloads rather than citations. You might think of a download as a unit of consumption, while a citation is a recommendation. Professor Carol Tenopir refers to 'reach' and 'prestige'. UKSG initiated the Usage Factor Project[13] to explore what might be possible/feasible. The following calculation was proposed:

$$\text{Usage Factor} = \frac{\text{Total usage over period 'x' of articles published during period 'y'}}{\text{Total articles published during period 'y'}}$$

(N.B. 'x' is the usage period, 'y' is the publication period

During Phase 2 of the project, which was funded by a number of industry organizations, a large trial was undertaken using a COUNTER-like format, but at article level. Consequently, it included (NISO) article version[14] and a DOI to identify the article. Real journal usage data, from seven leading publishers, was analyzed by John Cox Associates and Frontline GMS. Technical obstacles were minimized by using a format that the industry was already familiar with. A Progress Report summarizing Phase 1 and 2 will be published in the first quarter of 2011.

Secondly, there has been an increasing interest in article-level usage, though from the repository community. Articles may now appear in a number of locations. They could be on the publisher's site, but could also be in local institutional or subject repositories. If it is in a subject repository, say PubMed Central, the content is available in the US, UK and Canada. So if you want to get some kind of global overview of use, you need to be able to combine the usage data.

A collaborative project, PIRUS(1&2)[15], was funded by JISC to take this forward. As COUNTER only applies at journal level, an Article-level Report (AR1) was defined. Code has been developed (for ePrints, DSpace and Fedora) to allow the pull or push of the usage data between repositories and a prototype 'central clearing house'. The data is in XML and uses the Context Object to identify the article. Repositories don't always have the DOI, but it is included where available. The use of a Context Object-based mechanism had already been used by MESUR[16] (MEtrics from Scholarly Usage of Resources) in the US. It has made its way into PLoS[17] (Public Library of Science), who make various metrics available, and Open-AIRE[18] (Open Access Infrastructure for Research in Europe).

Dessert

Lastly, an initiative called ORCID[19] (Open Researcher & Contributor ID), whose membership draws in representatives from all areas of the community.

There have been a number of separate initiatives to provide researchers and other entities with an identifier to associate with their research outputs. If you possess multiple identifiers, you would naturally like to be able to re-use and link the associated profiles. All research stakeholders will benefit. Importantly, it is intended to work with others, not compete with them. It is also important that the system 'scales'. In a recent talk by Dr Salvatore Mele of CERN, he gave an example of 'hyperauthorship', where a 4.5-page article included an additional 9.5 pages of authors. Anyone processing all those names and affiliations would appreciate doing so in a consistent way.

Digestif

A final observation is that the truly successful and useful technologies have come about when the community has come together, irrespective of the initiator. It is important to try and avoid proprietary solutions, which are inevitably psychotic. UKSG is here to help avoid psychoses and to encourage the community to accept a shared view of reality.

Bon appétit!

References

1. http://info.cern.ch/ (accessed 21 January 2011).
2. http://www.w3.org/Addressing/URL/url-spec.txt (accessed 21 January 2011).
3. http://www.doi.org/ (accessed 21 January 2011).
4. http://www.crossref.org/ (accessed 21 January 2011).
5. http://www.youtube.com/watch?v=B3KBuQHHKx0 (accessed 21 January 2011).
6. http://www.niso.org/workrooms/openurlquality (accessed 21 January 2011).
7. http://www.projectcounter.org/ (accessed 21 January 2011).
8. http://www.niso.org/workrooms/sushi/ (accessed 21 January 2011).
9. http://www.editeur.org/ (accessed 21 January 2011).
10. http://www.uksg.org/projects (accessed 21 January 2011).
11. http://www.uksg.org/transfer (accessed 21 January 2011).
12. http://thomsonreuters.com/products_services/science/academic/impact_factor/(accessed 21 January 2011).
13. http://www.uksg.org/usagefactors (accessed 21 January 2011).
14. http://www.niso.org/workrooms/jav (accessed 21 January 2011).
15. http://www.cranfieldlibrary.cranfield.ac.uk/pirus2/tiki-index.php (accessed 21 January 2011).
16. http://www.mesur.org/MESUR.html (accessed 21 January 2011).
17. http://www.plos.org/ (accessed 21 January 2011).
18. http://www.openaire.eu/ (accessed 21 January 2011).
19. http://www.orcid.org/ (accessed 21 January 2011).

Acknowledgment

The description of the library perspective draws heavily on a talk given by Oliver Pesch (EBSCO) at the NFAIS meeting on 10 November 2010, who kindly agreed to its inclusion.