OAI-PMH: Interview with Tony Hammond

Martin Fenner, Gobbledygook

May 25, 2009

Most of us find, store and sometimes read scientific papers electronically. Although abstracts and full-text papers are usually available as web pages in HTML format, PDF is clearly the preferred format for storing and printing papers.

But publishing scientific papers in electronic form obviously requires more than providing the content in HTML or PDF format. We want to find the papers we are interested in on the journal homepage or in a digital library (e.g. PubMed), and for this we need metadata about the paper. The metadata could simply be an digital object identifier (DOI), but the metadata could also contain important information required to find a paper in a search strategy (e.g. authors, title, publication date or keywords).

As Duncan Hull et al. noted in a PLoS Computational Biology paper last year (Defrosting the digital library: bibliographic tools for the next generation web), metadata are often disconnected from the data, and there are no universally agreed standards to represent these metadata.

But why should we as scientists care about the technologies used to publish and distribute a paper? We shouldn't forget that these technologies could allow new and innovative ways to find and read scientific papers. One simple example: storing the metadata in the PDF file (using XMP) could make it much easier to import a large collection of PDF files into a reference manager.

One such initiative to provide the metadata of a scientific paper is OAI-PMH. I asked Tony Hammond from Nature.com a few questions about this newly supported protocol, as well as some more general questions about metadata provided by the Nature Publishing Group journals.

1. Can you describe what OAI-PMH is and does?

Well, we'll first need to unpack that double acronym. OAI-PMH is the Protocol for Metadata Harvesting which comes from the Open Archives Initiative. It's traditionally been known simply as OAI, although these days it's more properly referred to in full as OAI-PMH in view of the arrival of a new sibling protocol: OAI-ORE, which stands for Object Reuse and Exchange. The two protocols are

complementary: OAI-PMH deals with metadata harvesting, while OAI-ORE is concerned with content aggregation and compound digital objects.

In brief, OAI-PMH provides an interoperability framework for networked repositories to exchange metadata records on their holdings. Such metadata typically includes bibliographic-type descriptions of repository items, such as title, authors and other identifying information. At a technical level OAI-PMH provides a very simple Web-based API for querying a repository by item, by set of items or by date range with data records being returned in an XML format which can be validated by a W3C XML Schema. A basic metadata format using Dublin Core is available from all OAI-PMH implementations, while other richer community-specific metadata formats are encouraged for a fuller semantic exchange.

2. What is PRISM Aggregator Message (PAM)?

I posted an entry PRISM Aggregator Message on CrossTech recently which delved into the PRISM Aggregator Message (PAM) format and provided some specific details and also reasons why we chose to use it. Basically our choice of PAM stems from our earlier work with RSS and PRISM.

Our first foray into semantic descriptions was with RSS feeds. While still at Elsevier I had collaborated with Timo Hannay and Ben Lund at Nature Publishing Group and wrote up a piece for XML.com entitled Why Choose RSS 1.0? which described how journal RSS feeds could be enhanced with PRISM metadata. Specifically that piece reviewed the different RSS strains and came out strongly in favour of RSS 1.0 which being an RDF application was truly extensible and could accommodate new vocabularies. The identification of PRISM as a useful vocabulary was largely based on noting the adjacency of two RDF use cases in an early edition of the W3C RDF Primer: RSS and PRISM. I had been reading up on RSS and RDF and I suddenly thought "Bingo!", we can add in PRISM terms to the RSS feeds since PRISM is a very handy supplement to Dublin Core and provides support for enumerated fields such as volume, issue, and page number. Another advantage for publishers was that PRISM is a simple term set and did not require any specialist library knowledge. A later paper by the same authors The Role of RSS in Science Publishing published with D-Lib Magazine made a more in-depth review of RSS in scholarly publishing and concluded again that RSS 1.0 with its transparent support for new vocabularies such as PRISM was the right way to go.

We wanted to provide this standard article description that we were shipping with RSS also with OAI-PMH records. Unfortunately we could not simply move over the RDF properties into an OAI-PMH delivered RDF/XML payload. The problem with OAI-PMH is that it requires a record to be constrained by a W3C XML Schema. This is where PRISM Aggregator Message (PAM) comes in.

PAM is an application of PRISM and defines a W3C XML Schema for XML content aggregators to use. The message is simply structured as a sequence of

one or more articles each with a body section for XHTML content and a head section with PRISM metadata. The body sections are optional so PAM can be used exclusively as a simple metadata packaging format as we have chosen to implement for OAI-PMH.

3. How does PAM relate to the NLM DTD?

Well, both are schemas for content. This was my blind spot with regard to PAM for many years. Our focus was on metadata exchange. PRISM was a vocabulary for metadata, or rather a set of closely related vocabularies. No problem. And yet there was this thing called PAM which was busying itself with content. And we had no interest in that as we had our own Nature-specific DTD. But the PRISM folks were very enthusiastic – so clearly there was a need.

It only dawned on me relatively slowly that the PAM DTD (also available in W3C XML Schema form since PRISM 2.0) was the correspondent of the NLM DTD in the scholarly world. It provides an interchange format for content elements. So the real impetus behind the development of the PRISM metadata vocabularies was as a direct support to XML content exchange although content is not actually required to be present and the metadata alone can be used in different applications.

4. What are typical use cases for OAI-PMH?

The main use of OAI-PMH is to sync, or perhaps a better term would be align, the holdings descriptions of digital repository collections, where the word "repository" should be understood in its broadest context. The protocol exposes a machine interface for the robot harvesting of records from a data provider which a service provider will build upon to provide value add services. But this machine interface can sometimes also be accessed through a user interface (such as that provided by the Nature OAI-PMH service) which adds a browser onto the repository records.

For additional uses of OAI-PMH a good start point is the paper by Herbert Van de Sompel, Jeff Young and Thom Hickey in D-Lib Magazine on Using the OAI-PMH ... Differently. This paper notes that the metadata formats used by OAI-PMH are any that can validated by a W3C XML Schema and that therefore OAI-PMH is nothing less than "a medium for incremental, date-sensitive exchange of any form of semi-structured data". OAI-PMH clearly has legs.

5. Why should researchers care about OAI-PMH?

I'm not sure that researchers should want to have any specific knowledge of OAI-PMH. Bear in mind that at heart this is an infrastructural technology which exists down in the data pipes and service conduits and is analogous to the fibre or radio channels that deliver broadband services to users.

Applications that make use of the syndicated metadata records that OAI-PMH provides for, however, are another matter. Those are very definitely things that researchers will care much about. It is difficult to distil those consumer applications into any specific categories as the services platform that OAI-PMH supports is very broad, ranging from general to domain-specific descriptions to full text records, and beyond.

6. What are the different data formats that descriptions of Nature.com articles are provided? How do they differ?

At Nature Publishing Group we are working towards a common delivery architecture for our metadata. We are defining a core set of properties and making descriptions built from that property set available across multiple channels. Currently we are focussing on the basic bibliographic record which supports reference linking so that links can be made back to the platform. But once the channels are established they can be readily amplified to carry additional properties using the various channels' native packaging mechanisms. Metadata channels include both standalone services (e.g. RSS, OAI-PMH, etc.) as well as content objects themselves (HTML, PDF, etc.).

On the services side we began almost six years ago with RSS feeds, which being RSS 1.0 are full RDF/XML documents. The open data model implicit in RDF means that we can readily add new properties into our feeds. A case in point is the InChI identifier for chemical substances which we are currently working towards adding into our RSS feeds for Nature Chemistry, similarly to what the Royal Society of Chemistry has done earlier with their Project Prospect.

We have just released our OAI-PMH service which provides the same basic property set as RSS but in PAM format – see the post A Catalog for Nature.com. As alluded to above OAI-PMH comes from an earlier generation of protocols that put more rather emphasis on validation (or packaging the elements) than on data modelling (or relating the elements). That is, while the records served by the OAI-PMH server are validatable using W3C XML Schema the properties they contain are not directly reusable within an open, cross-application context such as is provided for by RDF. We do though have a simple stylesheet that can generate RDF/XML from these OAI-PMH records.

On the content side we added in the same properties to our HTML using META tags a year ago now – see the post Nature.com adds metadata. These properties are easily extractable by tools as simple key/value pairs. While this metadata is not directly representable as RDF it can be readily generated. We are anyway moving to make this property set more accessible by adding in RDFa which has now gone mainstream following Google's recent announcement of rich snippets and their support for RDFa.

We also began at the end of that year a programme to embed XMP packets into all of our newly published PDF files, again using the same basic property set – see the post XMP Labelling for Nature. The XMP packets are essentially simple

RDF/XML documents.

You can begin to see where all this is going. We are aiming to make all our descriptions conformant to a common data model, i.e. to RDF. That way, regardless of the distribution channel used the data delivered down that pipe can be merged into the common semantic graph.

7. What tools can researchers use to retrieve these descriptions?

Many of these channels are amenable to repurposing. The metadata they carry can be consumed within application-specific contexts, or it can be extracted from the channel medium for use in a wider generic context. Consider, for example, an RSS feed which can be used directly by a desktop or Web-based RSS reader. But it can also be mined for its metadata content, trivially in this case since the medium is already RDF/XML. Or consider again the metadata within an XMP packet in a PDF document which can be read by a viewer application (e.g. Adobe Acrobat) and presented to the user in a "Document Properties" display. But it can also be extracted simply by locating the XMP packet and reading the single XML child element which is itself a full RDF/XML document.

So one could say there are two classes of tools, those that operate at an application or specific layer and those that operate at a more generic layer, albeit with some preprocessing steps to unpack the metadata.

I should really expand here on OAI-PMH specifically since this is new for us. The primary means of interacting with an OAI-PMH server is via its service endpoint. Obviously to manage pagination seamlessly (the resumption token provides a cursor into the result record set) a library or tool is of enormous assistance. The OAI website provides a reasonably full listing of PMH tools available.

Our OAI-PMH server implements Jeff Young's OAICat which is a Java servlet webapp providing a repository framework which also comes with a forms interface for testing. This interface is especially useful for occasional use, e.g. a single "GetRecord":http://www.nature.com/oai/html/getRecord.html or call to "ListRecords":http://www.nature.com/oai/html/listRecords.html (or "ListIdentifiers":http://www.nature.com/oai/html/listIdentifiers.html), although repetitive calls to ListRecords (or ListIdentifiers) would quickly become tedious.

For harvesting we have used for testing purposes the ruby-oai gem for Ruby by Ed Summers and Will Groppe which includes a library and simple client which can also be run interactively as a shell. Note that this gem is not listed on the OAI-PMH tools page.

We have also made use of the open-source Java client OAIHarvester2, again by Jeff Young of OCLC. We used this for test harvesting of our full record collection as it was a robust implementation and we had earlier run into some problems with the Ruby app as it is not as finished as it might be, although it remains very configurable and easy to use. Our intention is to proceed with the Ruby

app for incremental harvesting. We're aiming to become consumers of our own services for quality control purposes.

8. What are your responsibilities at Nature.com?

I work within the Platform Technologies group on infrastructural projects supporting discovery and access across the nature.com platform, especially those that are built upon open technologies. My job handle is Application Architect although we're working on deconstructing that. I don't have line responsibilities but do supervise the development of our new interfaces.

We are corralling these various standards – some come from the wider Web world, some from the digital library community, some come from industry, and some are closer to home – under the general moniker of Public Interfaces. We also have a new documentation centre for this which is located on our Librarian Gateway.

I do maintain an active presence with various external bodies, not unsurprisingly given that the focus of my work is on defining and building interfaces. I have been from the beginning very involved in CrossRef and the development of DOI and related technologies. Most recently I have worked with a CrossRef WG to draw up a best practices document for scholarly publishers and RSS, and we are now starting work on a companion document for embedding metadata. I am also on the PRISM WG and a regular contributor.

Other activities I'm involved with include being a member of the SRU Editorial Board and of the eJournal Joint Technical Panel on Legal Deposit here in the UK. I have previously been a member of the OpenURL Standard Committe that developed ANSI/NISO Z39.88, and worked as a member of the OAI-ORE TC and the JISC PALS Metadata and Interoperability WG.

9. What did you do before starting to work for Nature.com?

Before working with Nature I was with Elsevier (2001-2004) as part of their Advanced Technology Group, and prior to that I worked with Academic Press (1997-2001) as Head of the Online Resource Activity. With Academic Press I was part of Electronic Publishing team that managed IDEAL – one of the first successful large journals platforms that applied consortial site licenses.

My previous experiences included a long-ish stint (1986-1995) with NATO SACLANTCEN in La Spezia, Italy, as Scientific Editor and subsequently as Head of the Information Branch. The facility was named SACLANT ASW Research Centre when I joined and had been transmuted into NATO Undersea Research Centre by the time I left (the anti-submarine warfare aspect deftly tidied away). Same mission though, to work on basic research (from oceanography to operational research) that would aid the NATO nations in their submarine detection programmes. (It now seems to be commuted to NURC alone.)

Prior to that I worked (1982-1985) as Assistant to the Editors for the North-Holland Publishing Company journal Nuclear Physics A based in Copenhagen, Denmark, and previous to that as a Research Associate in the Space Physics Laboratory at the University of Kent at Canterbury, UK where I collaborated in building a micrometeoroid sensor experiment which was deployed by the Space Shuttle.

10. Do you want to talk about future plans for metadata at Nature.com?

Sure. I guess I should separate out metadata management per se from metadata delivery and discovery. As a company we have an active programme underway to review wholesale our various ontologies and vocabularies in order to coordinate and streamline them. We also have ongoing initiatives to add in text mining to our production workflow and to address the entity extraction problem. Having better and richer vocabularies and new terms is one step. The next step is how to communicate that value in an open and structured manner to consumer applications.

My particular focus is on delivery channels. As I mentioned earlier we are working towards providing a notion of public interfaces: a set of open interfaces for delivering standard object descriptions. Complementing the OAI-PMH service which provides a catalog for nature.com we are also currently working on an SRU (Search and Retrieve by URL) service to support structured searching. And that would also be accessible through simple OpenSearch conventions.

We will be extending our page markup to include RDFa which will not only provide metadata in RDF format but will also localize those descriptions to the content fragment so that cut-and-paste operations will scoop up any descriptive markup along with the actual content.

We are now close to completing our support across our full title range for XMP in PDFs. A related development will be to embed XMP into our images (JPEGs and GIFs) so that all of our main resources then become self-describing. It is a wonder that we have gotten thus far in online publishing sending out content entities which are not unambiguously labelled.

And beyond this all lie the promises and challenges of the Semantic Web.

Figure 1: