Insights into Nature’s Data Publishing Portal

Posted March 30, 2016 by Andreas Blumauer

In recent years, Nature (http://www.nature.com/) has adopted linked data technologies on a broader scale. Andreas Blumauer was intrigued to discover more about the strategy and technologies behind. He had the opportunity to talk with Michele Pasin and Tony Hammond who are the architects of Nature’s data publishing portal (http://www.nature.com/ontologies/).

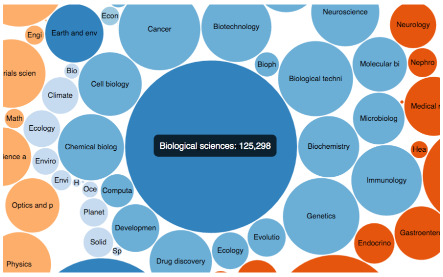
Semantic Puzzle: Nature’s data publishing portal (http://www.nature.com/ontologies/) is one of the most renowned ones in the linked data community. Could you talk a bit about its history? Why was this project initiated and who have been the brains behind it since then?

Michele Pasin: We have been involved with semantic technologies at Macmillan since 2010. At the time it was primarily my colleague Tony Hammond who saw the potential of these technologies for metadata management and data sharing. Tony set up the data.nature.com (http://data.nature.com) portal in April 2012 (and expanded in July 2012), in the context of a broader company initiative aimed at moving towards a ‘digital first’ publication workflow.

The data.nature.com (http://data.nature.com) platform was essentially a public RDF (http://en.wikipedia.org/wiki/Resource_Description_Framework) output of some of the metadata embedded in our XML articles archive. This included a SPARQL (http://en.wikipedia.org/wiki/SPARQL) endpoint for data about articles published by NPG from 1845 through to the present day. Additionally the datasets include NPG product and subject ontologies. These datasets are available under a Creative Commons Zero waiver.

The data.nature.com (http://data.nature.com) platform was only for external use though, so it was essentially detached from the products end users would see on nature.com. Still, it allowed us to mature a better understanding of how to make use of these tools within our existing technology stack. It is important to remember that in the years the company has been investing a considerable amount of resources on an XML-centered architecture, so finding a solution that could leverage the legacy infrastructure with these new technologies has always been a fundamental requirement for us.

More recently, in 2013 we started working on a new hybrid linked data platform, this time with a much stronger focus on supporting our internal applications. That’s pretty much around the time I joined the company. In essence, we made the point that in order to achieve stronger interoperability levels within our systems we had to create an architecture where RDF is core to the publishing workflow as much as XML is. (By the (https://semantic-web.com/wp-content/uploads/2016/03/Bildschirmfoto-2016-03-29-um-16.44.39.png)way if you are interested in the details of this, we presented a paper about this at ISWC 2014 (http://www.nature.com/ontologies/about/publications/).) As part of this phase, we also built a more sophisticated set of ontologies used for encoding the semantics of our data, together with improved versions of the datasets previously released.



The nature.com ontologies portal (http://www.nature.com/ontologies/) came out in early 2015 as the result of this second phase of work. On the portal one can find extensive documentation about all of our models, as well as periodical downloads in various RDF formats. The idea is to make it easier for people – both within the enterprise and externally – to access, understand and reuse our linked data.

At the same time, since user engagement level on data.nature.com was not as good as expected, we decided to terminate that service. In the future, we plan to keep releasing periodic snapshots of the datasets and the ontologies we are using, but not a public endpoint in the immediate future.

Semantic Puzzle: As one of your visions (http://www.nature.com/ontologies/about/vision/) you’re stating that your “primary reason for adopting linked data technologies is quite simply better metadata management”. How did you deal with metadata before you started with this transition? What has changed since then, also from a business point of view?

Michele Pasin: Our pre–linked data approach to dealing with metadata and enterprise taxonomies is probably not unheard of, especially within similar sized companies: a vast array of custom-made solutions, varying from simple word documents sitting in someone’s computer, to Excel spreadsheets or, in the best of cases, database tables in one of our production systems. Of course, there were also a number of ad-hoc applications/scripts responsible for the reading/updating of these metadata sources, as often they would be critical to one or more system in the publishing workflow (e.g. think of the journal’s master list, or the list of approved article-types).

To provide a superior content discovery experience for our audience and to facilitate emergent behaviours and interactions

To evolve a data model which is highly responsive to new and legacy business needs and can drive an in-flight publishing operation

It is worth stressing that the lack of a unified technical infrastructure aspect was a key problem, of course, but not the only one. In fact I would argue that addressing the lack of a centralized *data governance* approach was even more crucial. For example, most often you would not know who/which department was in charge of a particular controlled vocabulary or metadata specification. In some cases, no single source of truth was actually available, because different people/groups were in charge of specific aspects of a single specification (due to their differing interests).

Hence you need a certain amount of management buy-in to implement such a wide-ranging approach to metadata; moving to a single platform and technical solution based on linked data was fundamental, but an equally fundamental organizational change was also needed. Even more so, if one considers that this is not a time-boxed project but rather an ongoing process, an approach which pays off only as much as you can guarantee that as new products and services get launched, they all subscribe to the same metadata management 'philosophy'.

Semantic Puzzle: One of the promises of Linked Data is that by "using a common data model and a common naming architecture, users can begin to realize the benefits and efficiencies of web scaling." Could you describe a bit more in detail into which eco-system your content workflows and publishing processes are embedded (internally and externally) and why the use of standards is important for this?

Tony Hammond: We operate with an XML-based workflow for documents where we receive XML from our suppliers and store that within an XML database (MarkLogic (http://www.marklogic.com)). Increasingly we are beginning to move towards a dynamic publishing solution from that database. We are also using the database to provide a full-text search across all our content. In the past we had various workflows and a small number of different DTDs to reconcile, although we are currently converging on a single DTD. To facilitate search across this mixed XML content we abstracted certain key metadata elements into a common header. This was managed organically and was somewhat unpredictable both in terms of content model and naming.

By moving to a linked data solution for managing our metadata which is based on a single, core ontology we bypass our normalized metadata header and start to build on a new simpler data model (triples) with a common naming architecture. In effect, we have moved from a nominally normalized metadata to a super-normalized metadata which uses web standards for data (URI, RDF, OWL).

Semantic Puzzle: Your contents are also multimedia (image, video, …). How do you embed this non-textual contents into your linked data ecosystem? Which gateways, tools and connectors are used to bridge your linked data environment with multimedia?

Tony Hammond: Some years ago we embarked on a new initiative internally to streamline our production workflows. Our brief was to support a distributed content warehouse where digital assets would be stored in various locations. The idea was to abstract out our storage concerns and to maintain pointers to the various storage subsystems along with other physical characteristics required for accessing that storage.

In practice our main content was housed as XML documents within a MarkLogic XML database and associated media assets (e.g. images) were primarily stored on the filesystem with some secondary asset types (e.g. videos) being sourced from cloud services.

To relate a physical asset (e.g. an XML document, or a JEPG file) to the underlying concept (e.g. an article, or an image) we made use of XMP packets (a technology developed by Adobe Systems and standardized through ISO) which as simple RDF/XML descriptions allowed us to capture metadata about physical characteristics and to relate those properties to our data model. An XMP packet is a description of one physical resource and could be simply linked to the related conceptual resource.

We started this project with an RDF triplestore for maintaining and querying our metadata, but over time we moved towards a hybrid technology where our semantic descriptions were buried within XML documents as RDF/XML descriptions and could be queried within an XML context using XQuery to deliver a highly performant JSON API. These semantic descriptions enclosed minimal XMP documents which described the storage entities.

Semantic Puzzle: Nature links its datasets to external ones, e.g. to DBpedia or MeSH (http://www.nature.com/ontologies/datasets/linksets/subjects/). Who exactly is benefiting from this and how?

Michele Pasin: I would say that there are at least two reasons why we did this. First, we wanted to maximize the potential reuse of our datasets and models within the semantic web. Building *owl:sameAs* relationships to other vocabularies, or marking up our ontology classes and properties with subclass/subproperty relationships pointing to external vocabularies is a way to be good 'linked data citizens'. Moreover, this is a deliberate attempt to counterbalance one of our key design principles: minimal commitment to external vocabularies. This approach to data modeling means that we tend to create our own models and define them within our own namespaces, rather than building production-level software against third party ontologies. It is worth pointing out that this is not because we think our ontologies are better – but because we want our data architecture to reflect as closely as possible the *ontological commitment* of a publishing enterprise with decades of established business practices, naming conventions etc. In other words, we aimed at creating a very *cohesive* and *robust* domain model, one which is resilient to external changes but that also supports semantic interoperability by providing a number of links and mappings to other semantic web standards.

> *Pointing to external vocabularies is a way to be good 'linked data citizens'*

The second reason for creating these links is to enable more innovative discovery services. For example, a nature.com subject page about *photosyntesis* could surface encyclopedic materials automatically retrieved from DBpedia; or it could provide links to highly cited articles retrieved from PubMed using MeSH mappings. This just scratches the surface of what one could do. The real difficulty is, how to do it in such a way that the overall user experience improves, rather than adding up to the information overload the majority of internet users already have to deal with. So at the moment, while the data people (us) are focusing on building a rich network of entities for our knowledge graph, the UX and front end teams are exploring design and interaction models that truly take advantage of these functionalities. Hopefully we see these activities continue to converge!

Semantic Puzzle: How do you deal with data quality management in general, and how can linked data technologies help to improve it?

Tony Hammond: We can distinguish between two main types of data: documents and ontologies. (And by ontologies we also comprehend thesauri and taxonomies.) Our documents are created by our suppliers using XML and are amenable to some data validations. We use automated DTD validation in our new workflow and by hand DTD validation in the older workflows. We also use Schematron rulesets to validate certain data points but these address only certain elements. We have a couple hundred Schematron rules which implement various business rules and are also synchronized with our ontologies.

(http://poolparty.biz)Our ontologies, on the other hand, are by their nature more curated datasets. These are mastered as RDF Turtle files and stored within GitHub. These are currently maintained by hand, although we are beginning now to transition some of our taxonomies to the PoolParty taxonomy manager (https://www.poolparty.biz/). We have a build process for deploying the ontologies to our XML database where they are combined with our XML documents. During this build process we both validate the RDF as well as running SPIN rules over the datasets which can validate data elements as well as expanding the dataset with new triples from rules-based inferencing.

Semantic Puzzle: For a publisher like Nature it is somehow “natural” that Linked Data is used. How could other industries make use of these principles for information management?

Tony Hammond: The main reason for using linked data is not to do with publishing the data (and indeed many other data models are generally used for data publishing), but with the desire to join one dataset with other datasets – or rather, the data within a dataset to the data within other datasets. It is for this reason that we make use of URIs as common (global) names for data points. Linking data is not just a goal in publishing data but applies equally when consuming data from various sources and integrating over those data sources within an internal environment. Indeed, arguably, the biggest use case for linked data is within private enterprises rather than surfaced on the open web. Once that point is appreciated there is no restriction on any industry in being more disposed to using linked data than any other, and it is used as a means to maximize the data surface that a company operates over.

> *The biggest use case for linked data is within private enterprises rather than surfaced on the open web*

Semantic Puzzle: Where are the limits of Linked Data from your perspective, and do you believe they will ever be exceeded?

Tony Hammond: The limits to using linked data are more to do with top-down vs bottom-up approaches in dealing with data, i.e. linked data vs big data, or data curation vs data crunching. Linked data makes use of global names (URIs), schemas, ontologies. It is highly structured, organized data.

Now, whether it is feasible to bring this level of organization to data at large or whether data crunching will provide the appropriate insights over the data is an open question. Our expectation is that we will still need to use ontologies – and hence linked data – as an organizing principle, or reference, to guide us in processing large datasets and for sharing those data organizations. The question may be how much human curation is required in assembling these ontologies.

Michele Pasin: On a more practical level, I’d say that the biggest problem with linked data is still its rather limited adoption on a large scale. I’m referring in particular to the data publishing and reuse aspect. On this front, we really struggled to get the levels of uptake the business was expecting from us. Consider this: we have been publishing metadata for our entire archive since 2012 (approx. 1.2m documents, resulting in almost half a billion triples). However very few people made use of these data, either in the form of bulk downloads or via the SPARQL API we once hosted (and that was then retired due to low usage). This is in stark contrast with other – arguably less flexible – services we make available, e.g. the OpenSearch APIs, or a JSON REST service, which often see significant traffic.

Last year we gave a paper at the Linked Science (http://linkedscience.org/events/lisc2015/)workshop (affiliated with ISWC 2015) with the specific intent to address the problem within that community. What seemed to emerge is that possibly this has to do with the same reason why this technology has been so useful to us. RDF is an extremely flexible and powerful model, however, when it comes to data consumption and access, the average user cares more about simplicity than flexibility. Also, outside linked data circles we all know that the standard tech for APIs is JSON and REST, rather than RDF and SPARQL.

> *Lowering the bar to the adoption of semantic tech*

The good news though is that we are seeing more initiatives aimed at bridging these two worlds. One that we are keeping an eye on, for example, is JSON-LD. The way this format hides various RDF complexities behind a familiar JSON structure makes it an ideal candidate for a linked data publishing product with a much wider user base. Which is exactly what we are looking for: lowering the bar to the adoption of semantic tech.

About Michele Pasin

Michele Pasin (https://www.linkedin.com/in/michele-pasin-0a02231) is an information architect and product manager with a focus on enterprise metadata management and semantic technologies.

Michele currently works for Springer Nature (http://www.springernature.com/gp), a publishing company resulting from the May 2015 merger of Springer Science+Business Media and Holtzbrinck Publishing Group’s Nature Publishing Group, Palgrave Macmillan, and Macmillan Education.

He has recently taken up the role of product manager for the knowledge graph project, an initiative whose goal is to bring together various preexisting linked data repositories, plus a number of other structured and unstructured data sources, into a unified, highly integrated knowledge discovery platform. Before that, he worked on projects like nature.com's subject pages (http://www.nature.com/subjects) (a dynamic section of the website that allow users to navigate content by topic) and the nature.com ontologies portal (http://www.nature.com/ontologies/) (a public repository of linked open data).

He holds a PhD in semantic web technologies from the Knowledge Media Institute (http://kmi.open.ac.uk/) (The Open University, UK) and advanced degrees in logic and philosophy of language from the University of Venice (http://venus.unive.it/philo/) (Italy). Previously, he was a research associate at King's College Department of Digital Humanities (http://www.kcl.ac.uk/artshums/depts/ddh/index.aspx) (London), where he developed on a number of cultural informatics projects such as the People of Medieval Scotland (http://www.poms.ac.uk/about/) and the Art of Making in Antiquity (http://www.artofmaking.ac.uk/). Online Portfolio: http://www.michelepasin.org/projects/ (http://www.michelepasin.org/projects/).

Michele Pasin will give a keynote at this year's SEMANTiCS conference (http://www.semantics.cc). SEMANTiCS Leipzig 2016 (http://www.semantics.cc)

About Tony Hammond

Tony Hammond (https://www.linkedin.com/in/tonyhammond) is a data architect with a primary focus in the general area of machine-readable description technologies. He has been actively involved in developing industry standards for network identifiers and metadata frameworks. He has had experience working on both sides of the scientific publishing information chain, from international research centres to leading publishing houses. His background is in physics with astrophysics.

Tony currently works for Springer Nature, a publishing company resulting from the May 2015 merger of Springer Science+Business Media and Holtzbrinck Publishing Group's Nature Publishing Group, Palgrave Macmillan, and Macmillan Education.

CATEGORIES

Knowledge Modelling (https://semantic-web.com/category/knowledge-modelling/), Linked Data (https://semantic-web.com/category/linked-data/), Semantic Applications (https://semantic-web.com/category/semantic-applications/)

TAGS:

enterprise linked data (https://semantic-web.com/tag/enterprise-linked-data/), Linked Data (https://semantic-web.com/tag/linked-data/), Linked Data & Open Data (https://semantic-web.com/tag/linked-data-open-data/), Quality assurance (https://semantic-web.com/tag/quality-assurance/), Resource Description Framework (https://semantic-web.com/tag/resource-description-framework/), Semantic Web (https://semantic-web.com/tag/semantic-web/), taxonomy management (https://semantic-web.com/tag/taxonomy-management/), XML (https://semantic-web.com/tag/xml/)

Our Core Product & Services

- PoolParty Semantic Suite (/poolparty-semantic-suite/)
- Semantic Web Starter Kit (/semantic-web-starter-kit/)
- Semantic Data Management (/semantic-data-management/)
- PoolParty Academy (/poolparty-academy/)

Our Research Focus

- R&I at SWC (/innovation/)
- Research Projects (/research-projects/)
- Trending Topics (/blog/)

Security

SECURITY CiS CERTIFICATION INFORMATION ISO 27001:2013 No.00403/0

SWC has been successfully certified according to ISO 27001:2013