Linked data experience at Macmillan: Building discovery services for scientific and scholarly content on top of a semantic data model

Tony Hammond and Michele Pasin
Macmillan Science and Education, The Macmillan Campus, London, N1 9XW, UK
{tony.hammond, michele.pasin}@macmillan.com

Background

Macmillan Science and Education is a publisher of high impact scientific and scholarly information and publishes journals, books, databases and services across the sciences and humanities. Publications include the multidisciplinary journal *Nature*, the popular magazine *Scientific American*, domain specific titles and society owned journals under the *Nature Publishing Group* and *Palgrave Macmillan Journals* imprints, as well as ebooks on the *Palgrave Connect* portal.

We have recently implemented a linked data architecture at the core of our publishing workflow with an archive of over 1m articles and a publication rate in the 100s of articles per day. We build on a common metadata model defined by an OWL 2 ontology. To meet acceptable page response times we have evolved a hybrid storage/query platform. User-facing queries are resolved against RDF/XML document includes using XQuery with execution speeds of 10s-100s of milliseconds depending on complexity, whereas data enrichment and integration is managed at the ETL layer using both SPARQL query/update together with SPIN and Schematron rules and bespoke code (Java/Scala).

Recently released discovery products on the nature.com[1] platform are based squarely on this linked data foundation and include subject pages[2] as a new navigational paradigm, and also bidirectional links between articles and related articles.

In general, we have found that by building on top of a rich and consistent data model we can provide new navigational pathways for users to discover and explore different facets of our content. Using such a simple entity–relationship model coupled with global addressing allows us to be truly web scalable.

Infrastructure

Our data model is realized using a linked data technology stack which provides a number of significant benefits over traditional approaches to managing data. The use of RDF encourages the use of a *standard naming convention*, and makes this generally accessible by enforcing a global naming policy. It provides a *higher-level semantic focus* for operations which means that we are less susceptible to

[1] http://www.nature.com/
[2] http://www.nature.com/subjects

syntax anomalies. Also, by building on an open data model it allows for *flexible schema management* consistent with an agile approach to software development. And finally it facilitates a simple means of maintaining *dataset descriptions* by allowing us to partition the data space using the named graphs mechanism.

To realize this common data model our Science and Scholarly division has developed the Content Hub as part of the ongoing New Publishing Platform programme. (This extends our earlier linked data work from 2012 with the public-facing query service at data.nature.com[3]. We have since retired this service but will continue to make data snapshots available at the same address.) All our publishable content is aggregated within the Hub which thus presents as a simple logical repository. In practice, the data is distributed across multiple physical repositories. The ontology organizes the conceptual data model as well as managing the physical distribution of content within the Hub using sidecar XMP packets for asset descriptions.

Our two core capabilities in managing the Hub are *content management* and *content discovery*. Structured content is maintained in XML format and is held within a MarkLogic repository. MarkLogic also provides us with a powerful text search facility. By contrast, discovery metadata is modelled in RDF (constrained as an OWL 2 ontology). The discovery metadata is further enhanced by using object-oriented RDF rule sets: knowledgebase contracts for data sharing, and SPIN rules for data generation and data validation and also Schematron rules for RDF/XML validations.

Challenges

Initially we had intended to query the graphs directly in a triplestore and had developed a linked data API for this purpose. In practice, we found that our implementation was not fit for purpose and failed in two critical dimensions: performance and robustness. Typical result sets were being delivered in seconds or tens of seconds, whereas we were tasked to achieve ~20 ms, some 2–3 orders of magnitude faster. Additionally for reliability we required a clustered solution, but the triplestore we had implemented was unclustered and non-transactional.

Since our immediate concern was to support our online products we decided on a tailored API that directly reflected the data model. Our main principles were that the API should be chunky not chatty, i.e. we aim to provide all required data in a single call; that data should be represented as it is consumed, rather than how it is stored; that it support common use cases in simple, obvious ways; that it ensure a guaranteed, consistent speed of response for more complex queries; and that it build on a foundation of standard, pragmatic REST using collections and items.

This led to our developing a hybrid system for storage and query of the data model. The data is throughout modelled in RDF but is replicated and distributed between RDF and XML data stores. We have added semantic sections as

[3] http://data.nature.com/

RDF/XML includes within our XML documents as well as creating standalone RDF/XML documents for our core and domain ontologies. Retrievals based on document ID are now realized with XQuery, and augmented by in-memory key/value lookups, yielding acceptable API response times. SPARQL queries are currently restricted to build time data assembly.

Future aims are threefold: 1) to grow the data model with additional things and relations as new product requirements arise; 2) to open up the user query palette to more fully exploit the graph structure while maintaining an acceptable API responsiveness; and 3) to create an extended mindshare and understanding throughout the company in the value of building and maintaining the discovery graph as a real enterprise asset.