LINKED DATA STANDARDS AND INFRASTRUCTURE FOR SCIENTIFIC PUBLISHING

Elsevier Position Paper for the W3C Workshop on Linked Enterprise Data Patterns

Bradley P. Allen
Elsevier Labs
b.allen@elsevier.com

MOTIVATIONS FOR LINKED DATA IN SCIENTIFIC PUBLISHING

The nature of scientific publishing is changing. The knowledge originally contained in scholarly books and journals is becoming liberated from formats based on the constraints of the printed page, and is being repackaged for a wide variety of online distribution channels, made accessible through APIs used by a ecosystem of content suppliers and partners to craft information solutions and delivery channels, and increasingly discoverable through social networks, search engines and researcher portals and Web sites.

As a consequence, scientific publishers such as Elsevier are changing the way we think about and work with scholarly content. We are beginning to address the fact that digital scholarly content is constantly changing and growing in type, structure and quantity. Specifically, we must:

- Build content acquisition, production and management systems that support with equal capability and flexibility a broad range of content types, treating as first-class objects video, audio, images, datasets, metadata and knowledge organization systems in addition to journals and books, as well an equally broad range of delivery formats across desktop, notebook, tablet and mobile computing devices.
- Make it easy for our authors, editors, suppliers and customers to discover and access, across all our content assets, information in fragments smaller than the traditional units of publication (e.g., down to the level of specific words, phrases, images, and table cells in articles or book chapters, or key frames and segments in videos) and make it easy to for them to aggregate and compose these fragments into new content, and to integrate content across application silos and organizations.
- Incorporate support for emerging knowledge organization systems (i.e. industry-standard taxonomies, ontologies, metadata vocabularies and named entity registries) to better organize our content, make it more discoverable, and make it easier to mine and visualize the underlying structure and patterns buried in the content for the purpose of furthering scientific knowledge and understanding.

Linked data provides a way to leverage the power of Web architectural standards and formats to achieve these ends. Linked data for these purposes can be created in a variety of ways. It may be created automatically and in an unattended fashion by having applications crawl content repositories or the linked data Web, or be produced in a semi-automated fashion embedded into traditional content production processes. In either case, publishers such as Elsevier are beginning to use linked data in the context of mature content production and management systems and will be adapting and leveraging existing content format and metadata standards to allow linked data to be produced and delivered within these systems.

ELSEVIER'S APPROACH TO USING LINKED DATA IN THE SCIENTIFIC PUBLISHING PROCESS

To aid in these efforts, we have identified a set of tasks across the different use cases for linked data within the scientific publishing process. Specifically, they are:

- Create linked data during the authoring and editorial process (e.g., by supporting offline annotation by authors and editors)
- Maintain linked data through production (I.e. make linked data compatible with our production process)
- Link to Elsevier content from Web (to allow Web content to link and drive traffic to Elsevier content, facts, concepts and relationships)
- Link from Elsevier content to Web (to allow Elsevier content and linked data to leverage and make reference to Web-based knowledge organization systems and related Web content)
- Append enhancements to published content (to allow new enhancements to be added at any stage of the content life stage after content has been published)
- Create content bundles using enhancements (to use enhancements to craft specialty- or topic-specific views of Elsevier content)
- Support import/export of enhancements (to allow integration with third-party sources of enhancements and knowledge organization systems)

To support the production and management of linked data about scientific publications in support of these use cases, Elsevier has created a number of specifications for content "satellites," which are XML documents that incorporate RDF named graphs providing specific types of linked data for scientific publishing (e.g. for expressing basic scholarly asset metadata, subject taxonomies or the results of subject tagging a specific journal article or book section based on a given taxonomy.) Content satellites capture the notion of the primacy of the journal article or book section in the production and management of scientific publications, and can be easily accommodated by existing XML-centric content production and quality control systems for validating submitted content.

We are further extending our existing taxonomy management tools and workflows to accommodate linked data standards and best practices, and by leveraging existing RDF vocabularies such as SKOS and SKOS-XL. New taxonomy and ontology development is beginning to be performed using these standards, and third-party suppliers of content enhancements such as subject tagging of journal articles are beginning to deliver content adhering to these standards. The standards we have developed for content satellites address the need in the scientific publishing process to address issues of provenance and quality.

To store, manage and provide access to and discovery of content satellites and the linked data they contain, we created a Linked Data Repository to provide services and APIs for applications built by Elsevier and third parties to store and retrieve scholarly linked data as either named graphs or sets of RDF statements. The Linked Data Repository complements existing content repositories that are focused on high-throughput, low-latency delivery of content through Web services and applications, or in the archival management of traditional book and journal content. Semantic enhancement, access and discovery will be supported through these services, by storing named graphs containing linked data for any and all content. The Linked Data Repository supports the expression of provenance metadata for linked data, and provides an API that allows retrieval of URIs for all linked data registered for a given document, author or other identifier. It is optimized for high-volume read-write of RDF documents, provides REST APIs for ease of integration, and supports existing access and entitlement mechanisms to support third-party use.

Search engines that index our content (such as Elsevier's Science Direct) are currently the principal mechanism for content access and discovery. Existing faceted search infrastructure that is in place today is being adapted to use linked data in the indexing key core content (articles, book chapters, author and affiliation records and taxonomy terms) allowing users to then find and follow linked data relationships between them and other content types. However, approaches and user interfaces are being developed to make it easy to leverage linked data for faceted

or ontology-driven browsing and other forms of enhanced content delivery, such as linked data mashups and visualizations, microsites or topic pages for concepts, authors, and organizations, and learning objects created from syllabi generated for use with books or journal articles.

CHALLENGES IN THE ADOPTION OF LINKED DATA AND HOW THE WORKSHOP CAN HELP

The following is a list highlighting three areas of issues that we have grappled with during the above-described work in adding linked data to our publication processes.

- Ease of technology adoption by IT organizations supporting scientific publishers
    - Best practices for URL and namespace governance
    - Approaches to the globalization/localization of knowledge organization systems and linked data
    - RDF serializations and deployment design patterns that ease the cognitive burden on application developers (e.g., clarifying practical and easy-to-communicate approaches to cool URLs and issues arising from HTTP-Range-14, and teasing apart competing standards for HTML markup with linked data such as RDFa and HTML5 microdata)
- Standards for vocabularies, taxonomies and named entity registries relevant to scholarly content
    - Standards for annotation of scholarly content (e.g. CITO, SWAN, SIOC)
    - Standards for named entities crucial to scholarly publishing (e.g. author identifiers and profiles in existing systems like Elsevier's Scopus and projected systems such as ORCID)
- Requirements for RDF stores and tools that address scientific publishing requirements
    - Standards supporting best practices associated with scholarly publication workflows (e.g. RDF named graphs, provenance and access and entitlement)
    - Validators for linked data for quality control in publishing high-value and/or mission-critical scholarly content (e.g. medical guidelines and research studies)
    - Support for free text search within RDF stores
    - Clarifying and choosing between competing approaches to specifying RDF queries (e.g. through RDF APIs vs. SPARQL)

Some of these challenges are specific to the use of linked data in scholarly publishing; others are of more general interest. Our motivation in submitting this position paper to the W3C Workshop on Enterprise Linked Data Patterns is to be able to discuss these challenges with colleagues in other organizations that are beginning to apply linked data principles, towards the end reducing risks and costs associated with adopting linked data technology in a commercial setting. Our sense is that there are a lot of existing systems built by this community that, discussed in the framework of this Workshop, can provide design patterns and best practices that address this goal.