Against the Grain

Volume 25 | Issue 1 Article 9

February 2013

From Discovery to Delivery: Publishing Opportunities on the Semantic Web

Daniel Mayer
TEMIS, daniel.mayer@temis.com

Recommended Citation
Mayer, Daniel (2013) "From Discovery to Delivery: Publishing Opportunities on the Semantic Web," *Against the Grain*: Vol. 25: Iss. 1, Article 9.
DOI: https://doi.org/10.7771/2380-176X.6410

From Discovery to Delivery: Publishing Opportunities on the Semantic Web

by **Daniel Mayer** (VP of Corporate Marketing, TEMIS) <daniel.mayer@temis.com>

In the seminal 2001 article he published in *Scientific American*, **Tim Berners-Lee** described "A new type of Web Content that will be meaningful for computers" and that "will unleash a revolution of new possibilities." Current trends in Online Publishing may well prove he was right, thanks to a new breed of technologies that exploit natural language processing to make text meaningful for computers.

Signs show that these semantic technologies are now being applied by significant portions[1] of the Publishing ecosystem to add strategic value to their business, driven in particular by three priorities:

- To make their existing products as compelling as possible
- To package their content innovatively
- To derive productivity gains in editorial workflows

Semantic Content Enrichment Defined

The key service offered by these semantic technologies to Publishers is the automated enrichment of their content with domain-specific metadata. Some of the typical forms this takes are as follows.[2]

- **The extraction of a *"document fingerprint"*** of most relevant terms used in a document. These are typically identified by comparison to a lexical model that reflects the distribution of vocabulary that is deemed "normal" in comparable documents. Document fingerprints can be used to classify documents according to predefined categories or to identify documents that are topically similar to an original.
- **The recognition of certain objects of interest — also called *entities* — mentioned**. This involves disambiguation between homonyms (such as between the fruit "orange" and the company "orange") and normalization across synonyms (the same concept may have several different names). As we will see below, extracted entities help boost the efficiency of search processes.
- **The recognition of *relationships*** that interlink entities (for example a merger between two companies). Among other methods, the analysis of syntax helps recognize the role played by each entity in the relationship, thus enabling a precise understanding of the event or situation described (for example helping understand which company is the acquirer and which is the target).

Makes Content More Compelling

The first area of application for semantic content enrichment lies in making existing products more compelling for customers by helping them access information efficiently and more effectively, in particular through:

A boost to the quality of the search experience

- More effective navigation in content
- Increased insight thanks to context and perspective

As we mentioned briefly above, entity extraction recognizes concepts as they appear in text, all the while overcoming homonymy and synonymy. This enables features such as ***"search by concept"*** or ***"by taxonomical category"*** (find all documents mentioning an automobile manufacturer) and helps retrieve a ***more complete and precise set of results*** relevant to the original query.

Entity extraction furthermore enables the deployment of ***"navigational facets,"*** an interactive mechanism that provides deep insight into search results along key entity types of interest (presenting, for example, the names of people most commonly or most uncommonly associated with a given topic or event), enabling the effective narrowing down of the search effort to the most relevant documents, but also acting as a navigational cue, for example by suggesting directions of research that may not have been foreseen at query time and would not have been suggested by a simple ranked list of results.

Extracted entities can be further leveraged to ***recommend*** to the reader further information that can help ***place the document in proper context***:

- First, by ***inserting structured knowledge*** that pre-exists regarding some of the objects mentioned: for example, the relevant encyclopedic articles concerning certain chemicals or illnesses,[3] or the biographic profiles of the authors of the original article, or a list of most frequent writers on these topics.
- Second, by ***recommending links to related content***: for example, other scientific articles that are closely similar in the topics they cover, patents mentioning the same chemical or chemicals belonging to the same class, articles discussing the same class of chemical compounds applied to other illnesses, or clinical study reports mentioning the same patient category in the context of the same illness.

The relevance of such recommendations can be enhanced by adapting these recommendation strategies to the individual profile of the reader, leading to ***personalization***.

Lastly, the extracted metadata may be ***analyzed*** and ***visualized*** through dedicated statistical tools, semantic relationship graphing, or clustering widgets that offer perspective in to the matter at hand. In a biomedical setting this might help researchers:

- understand which side-effects are the most commonly associated with chemical compounds similar to the one mentioned in an article, or
- cross-analyze patient populations and side-effects for the specific chemical at hand based on available literature citing it, or
- graph the relationships between related chemical compounds, the illnesses they treat, and the known side effects produced.

The above features provide a boost to the efficiency of end-user search processes by reducing their need for foresight, helping them navigate corpora, proactively delivering additional highly relevant content recommendations to the end-user in the form of inserted knowledge and links to related content, and helping them gain perspective by analyzing it.

By promoting an ***engaging*** and ***immersive experience***, semantic content enrichment can be expected to promote differentiation as well as enhance the stickiness of Information Portals and ultimately increase their usage.

Helps Package Content in Innovative Formats

Perhaps the most exciting area for semantic content enrichment is in developing new information products and new ***types*** of information products.

As we pointed out earlier, it is possible to leverage semantics to automate the extraction of relationships (facts) from content, or conveniently perform large-scale analyses on such corpora. This streamlines the assembly of commercial ***Knowledge Bases*** that can be inserted in the growing Linked Open Data space and ***Analytical Reports*** on virtually any subject covered in text.

In a similar spirit, semantics also enable the automated aggregation of content related to any given topic, however focused, across silos or repositories. This is powering the emergence of ***Topic Pages*** and ***Microsites***[4] that can collate not only the relevant journal articles, reports, and books on a given topic but also the most recent related media coverage, real-time data,

continued on page 22

From Discovery to Delivery ...
from page 20

profiles of key individuals involved, links to relevant discussion forums and job postings, as well as multimedia offerings related to the topic.

A point worth highlighting is that semantics provide the means to perform such topical aggregation at any scale, including the finer ones where the topic at hand is so focused that a dedicated editorial investment would have been difficult to justify given the small scale of the corresponding audience.[5] In turn, this means that ***the number of such Topic Pages that can be produced is virtually unlimited, enabling the cost-effective publication and maintenance of as many pages as needed at little incremental cost***. In its most forward-thinking incarnation this idea points to the possibility of letting the audience prompt the packaging of such pages dynamically through query mechanisms, even going as far as bespoke "audience-of-one" information products.

It is interesting to note as well that any type of content can be mashed up, including non-text, for example images, videos, and sound, provided that it carries appropriate metadata that enables it to be linked. An emerging commercial application of this is ***Contextual Advertising***: the linking of highly relevant advertising to the original content. As an example of this tactic, an end-user reading an article about education might find ads for books or seminars on the topic, or for local private schools. This tactic enables the targeted placement of ads in front of highly focused communities of interest that can be identified by leveraging the type of content being accessed.[6]

Lastly, the combined advent of Web Services and associated APIs and their recent adoption as a growing pathway for accessing content are enabling the ***proactive delivery*** of content within a large — potentially universal — set of end-user workflow applications. This will make it possible for end-users to get answers to their information needs without having to explicitly ask for them, simply as a function of the predictive quality of link-building strategies. The shapes that this will take have yet to be seen, but it is foreseeable that at least certain end-users will value the opportunity of finding highly relevant content within their ongoing workflow rather than having to disrupt it to turn to a dedicated, expert interface for information research. The commercial payoff is that these end-users will view the suppliers of such workflow integration as unique among their peers. From the Publisher perspective, such integrations may provide not only a new delivery mechanism for their information products, but also an entirely new set of contexts for selling them. We posit that such proactive ***Delivery*** may also represent a significant evolution to the ***Discovery*** paradigm of Information Access traditionally embodied by Search and Exploration activities.

Whether through ***Knowledge Bases***, ***Topic Pages***, ***Contextual Advertising***, or ***Workflow Integration***, the deep metadata provided by semantic content enrichment opens the door to new growth opportunities for Publishers, thanks to innovative content packages and formats that are highly relevant for their audience.

Productivity Gains

Lastly, semantic content enrichment finds an immediate area of application in the context of editorial processes, where metadata is originally added to content. Here, automation provides productivity gains by helping overcome the limits associated to traditional manual metadata contribution. The benefits are an ability to process more content at greater speed and to provide metadata that is more consistent and describes contents in greater depth.

We showed earlier in this article how semantic content enrichment could provide end-users efficient navigation and deep insight into large quantities of content. The same benefits can be used by Product Development teams to gain insight into the unstructured informational assets that they may have at hand. A typical case concerns the monetization of archives whose content may be initially unclear. In such a context, semantic content enrichment provides Product Development teams with deep insight into these archives and helps assess which subsets may be marketable. When applied to existing archives — a category of readily available yet underutilized informational assets — semantic content enrichment therefore brings the potential of ***increased Return on Assets***.

The metadata produced through semantic content enrichment may also find further applications in boosting the efficiency of a wide range of internal processes where its ability to reveal the nature of informational assets can be used for example to appropriately match content to reviewers (supporting efficient peer reviewer identification) or to recommend subscriptions to customers by analyzing usage patterns at comparable institutions.

As a consequence, when leveraged by Editorial and Product Development teams, semantic content enrichment can become ***a core productivity tool*** that provides flexibility and practicality to key processes, lowering time-to-market and increasing overall return on assets.

Conclusion

Early market signals show that semantic content enrichment is bringing disruptive enhancements to existing Information Access and Content Management paradigms supporting the Publishing business.

By increasing the relevance of search results and enabling powerful navigation in content, by boosting insight with related content recommendations and knowledge linking, and by enabling metadata-driven analytics, it provides end-users the benefit of finding relevant content in closer grasp, if not embedded directly within their workflow. It also provides Publishers a powerful tool to optimize key Editorial and Product Development tasks, to boost audience engagement with and usage of more differentiated online information products and it is thus a formidable enabler of growth through new formats, products and channels, and an invitation to push the traditional envelope of their business and stake out new territory on the Semantic Web.

Author's Bio

Daniel Mayer is VP of Corporate Marketing at **TEMIS**, a leading provider of semantic content enrichment solutions for the Enterprise. He is responsible for promoting, as well as shaping, the flagship Luxid product range and its roadmap with a particular focus on STM Publishing and Enterprise Information Management. Prior to joining **TEMIS**, **Daniel** served for the past 12 years in a variety of marketing and product strategy functions throughout the IT value chain. **Daniel** holds both a Masters degree in Business from **HEC**, France, and a Masters in Computer Science from **ENST**, France.

About TEMIS

TEMIS helps organizations to structure, manage, and leverage their unstructured information assets. Its flagship platform, Luxid,

continued on page 24

against the grain

people profile

VP Marketing, TEMIS
207 rue de Bercy, 75012 Paris
Phone: +33 1 80981145 • Fax: +33 1 80981101
<daniel.mayer@temis.com> • *www.temis.com*

Daniel Mayer

BORN AND LIVED: Born in Paris, France. Lived between Paris and the U.S. East Coast (NYC).

PROFESSIONAL CAREER AND ACTIVITIES: 15 years in Product Management and Marketing across the IT value chain.

FAMILY: Married with three children.

FAVORITE BOOKS: *Crossing the Chasm*, **Geoffrey Moore**.

PHILOSOPHY: Growth starts at the edge of your comfort zone.

From Discovery to Delivery ...
from page 22

identifies and extracts targeted information to semantically enrich content with domain-specific metadata. Luxid enables professional publishers to efficiently package and deliver relevant information to their audience, and helps enterprises to intelligently archive, manage, analyze, discover, and share increasing volumes of information. Founded in 2000, **TEMIS** operates in the United States, Canada, UK, France, and Germany, and is represented worldwide through its network of certified partners. *http://www.temis.com*

Endnotes

1. A 2011 study among scientific journals publishers by Publishing Research Consortium revealed that 46% of respondents were currently applying these technologies to their content.
2. The reader should take care that many approaches exist, and some offer only a subset of the capabilities listed here.
3. Such structured knowledge could be found in proprietary assets such as a knowledge base, or through openly available linked data repositories (The following link can be used as a starting point for more information on this topic: *http://en.wikipedia.org/wiki/Linked_Data*).
4. In recent years the term "mash up" (both a verb and a noun) has been used in a slightly larger sense to designate such content aggregation tactic by the technical community. Topic Pages and Microsites can be considered examples of mashups.
5. This participates in the beneficial effects of the Long Tail as popularized by **Chris Anderson** in his seminal *Wired* article (*http://www.wired.com/wired/archive/12.10/tail.html*) and subsequent book.
6. Or through profiling, for example, by statistically analyzing individual end-user's content of interest.

against the grain **people profile**

Chris Banks

University Librarian and Director, Library, Special Collections & Museums, University of Aberdeen
The Sir Duncan Rice Library, University of Aberdeen
Bedford Road, Aberdeen, AB24 3AA
Phone: +44 (0) 1224 273384 • <c.banks@abdn.ac.uk>
www.abdn.ac.uk/library

BORN AND LIVED: Born in London. Lived in London, Belfast, London, Aberdeen.

EARLY LIFE: Shaped (positively) by living in Belfast and benefitting from the very supportive schooling there. My career owes everything to that.

PROFESSIONAL CAREER AND ACTIVITIES: Two music degrees. Worked at **English National Opera** and the **British Library** (20+ years) before moving to Aberdeen.

FAMILY: One very supportive husband who doesn't mind that I live and work over 550 miles from home. And one cat who likes a choice of laps (and therefore welcomes when there is more than one in London).

IN MY SPARE TIME: I enjoy singing.

FAVORITE BOOKS: Well-written ones

PET PEEVES: Procrastination; inefficiency; intolerance.

PHILOSOPHY: Life is special and to be treasured. Each day could be my last, and I want to make a positive difference.

MOST MEMORABLE CAREER ACHIEVEMENT: Working with a wonderful team to achieve a wonderful new library in Aberdeen.

GOAL I HOPE TO ACHIEVE FIVE YEARS FROM NOW: That we will still be full of positive ideas as to how to make a positive difference to those studying, researching, and visiting our new library (and that we'll have won a few prizes on the way).

HOW/WHERE DO I SEE THE INDUSTRY IN FIVE YEARS: The relationship between publishers, libraries, and LMS suppliers will fundamentally change over the next five years. Our goal is to continue to ensure that the academic endeavour reaches the widest audience and makes the biggest and most relevant impact. Our role as curator/custodian will evolve to fully embrace the born digital. We will share more services and, if sense can prevail, will not be penalized for doing so. Libraries will continue to be responsive to these new environments but will remain relevant.

Spaces and Clouds: The Library as Destination and Launch Pad

by **Chris Banks** (University Librarian and Director, Library Special Collections and Museums, University of Aberdeen) <c.banks@abdn.ac.uk>

Abstract

This article considers both the physical and online spaces that together comprise the university library and study environment for many of today's students. It looks at some of the evidence which can be used to inform decision-making in terms of space optimization, eliminating barriers to online access, maximizing collection development budgets both in terms of targeting acquisitions and ensuring that collections are discoverable, and using process improvement techniques in order to maximise staff effectiveness.

University Investments

In the last four years the **University of Aberdeen** has invested over £57m in a new University Library, a Special Collections Centre, a Conservation Centre, and a new museum (King's Museum). Further investment has seen the introduction of a single resource discovery layer which searches all locally-held and subscribed resources in all formats. Finally, there has also been substantial evidence-based investment in online resources, including journal backfiles and eBooks. Over 80% of the current collection development budget is spent on electronic resources.

Evidence-based Investment in Online Resources and Tools

The evidence base for the targeted acquisition of journal backfiles included the publisher's own record of click-through attempts from discovered titles to full text. Using this evidence, together with an unfunded priority list of backfiles prepared by academics resulted in significant additional use of the newly-subscribed content. Furthermore, the addition of bibliographic records for the eBook content to the library's own catalogue resulted in

continued on page 26