

Peer Review:

Scientific Publishing: Disruption and Semantic Build-Up



Frank Hellwig

Frank Hellwig is studying Publishing at Oxford Brookes University (MA) and Leipzig University of Applied Sciences with a focus on scientific publishing. He has worked in digital publishing, for a journal subscription agency, in book retail, event management and development aid in Germany, UK, France and Bangladesh.

E-mail: hellwig.frank@gmx.net

Abstract

A new technology paired with a viable business model will have disruptive impact on incumbent companies in a specific market, if they do not re-evaluate and update their business models accordingly. As the Internet matures, Semantic Web technologies enable applications for meaning-based and dynamic filtering and processing of information, which has a disruptive impact on scientific publishing.

This article calls for publishers to adopt semantic technologies and emphasises the “need to include a semantic strategy in their business models” (Hawkins 2009). With a focus on journals as the ‘cash cow’ of scientific publishers, it assembles debates about disruption and general tendencies in scientific publishing. An introduction to Semantic Web, Text Mining and Semantic Publishing is given as well as various examples of product developments, company partnerships and acquisitions related to semantic technologies. Finally, different ways of acquiring semantic annotation data and financial aspects of semantic enhancements are discussed.

Keywords: scientific publishing, science, semantic web, linked data, natural language processing, text mining, business, disruption

Disruptive Innovation

According to Clayton Christensen (2008), disruptive innovation makes a product simpler and more affordable. Disruptive innovation comprises an enabling technology and a business model that can deliver this solution more cost effectively. Such innovations have disruptive effects on established companies, as managers tend to compare investing in a new business model (full cost) with leveraging what is already in place (marginal cost). This causes them to think that business model innovation

DOI: 10.1163/095796509X12777334632744

is not attractive. New entrants in contrast, without a comparison, create what needs to be created (Christensen 2008).


Concerning the implications of the advanced process of scientific literature becoming available in digital formats, Bruck (2008) mentions P2P networks, Open Access / open archive publishing and inter-community trading as “challenges” for publishers.

Cope and Kalantzis (2009) in *Signs of epistemic disruption: transformation in the knowledge system of the academic journal* describe “disruptions of scholarly work”. For example, they suggest that pre-publications erode the significance of publications. In some areas, conference proceedings, for their immediacy, and reports become more important than journal articles, and authors and institutions insist that articles be published in their own institutional repositories or on personal websites – “legally or illegally, with or without reference to the publishing agreement they have signed” (ibid). Cope and Kalantzis identify as further drivers of disruption that knowledge these days is produced by a whole host of organisations, and more knowledge is produced within the networked interstices of the Social Web where amateurs mingle with professionals.

David Bousfield (2009), Vice President and Lead Analyst of research and advisory firm Outsell, mentions the first of four “disruptive forces” for the STM market as Open Access. He notes: “Springer’s purchase of BioMed Central and the launch of Nature’s *Communications* [journal] both represent significant landmarks in the adoption of this disruptive business model by for-profit publishers.” The Open Access “business model that has infected mainstream STM publishing is working its way through legal, tax, and regulatory content, and also permeates the co-creation of news and market research” (Stratigos et al. 2009). Also, funding institutions increasingly demand an Open Access approach (Lunn 2010).

By number of articles Open Access penetration is estimated to be 9.8 per cent, and the current Open Access market accounts for 3.3 per cent of the total journal publishing market, growing at 11.3 per cent per year (Pollock 2009). “It [Open Access] represents a less benign model to publishers, as it breaks the monopoly of ownership of must-have units of content and will bring price elasticity, and

therefore reduced revenues, to the primary journal publishing market” (ibid). Although significantly mitigated by increasing R&D spending in the event of widespread take-up of Open Access, market value will shrink by an estimated 57 per cent (ibid).

 **Scientific publishing is shifting from selling static pieces of content toward access-centred models for dynamic content in multiple formats coupled with value-added services.**

Michael Nielsen (2009) in his article *Is scientific publishing about to be disrupted* claims “that scientific publishing is in the early days of a major disruption” and that “those publishers that don’t become technology driven will die off”.

Michael Clarke (2010), asking why scientific publishing has not been disrupted already, examines the potential for disruption by listing five functions of journals. He asserts that beyond dissemination and registration, for which journals are no longer needed, there are three additional functions that journals serve which have developed over time: validation, filtration and designation.

Regarding validation Clarke writes: “To date, no one has succeeded in developing a literature peer-review system independent of journal publication”. Concerning filtration, various new tools, instead of replacing journals, “rely on the filtration provided by journals” (ibid). Clarke continues:

While there is the possibility that recent semantic technologies will be able to provide increasingly sophisticated filtering capabilities, these technologies are largely predicated on journal publishers providing semantic context to the content they publish. In other words, as more sophisticated filtering systems are developed – they tend to augment, not disrupt, the existing journal publication system.

As funding and career advancement decisions are based on scientists' publication record (designation), Clarke, at best, sees a shift away from journal toward article-based metrics. But even then, "change would likely be incremental rather than disruptive" and such a transition "would likely be measured not in years but in decades" (ibid).

Finally, he argues that the main reason why the latter three functions were not easily replaced is that they are not technology-driven but "cultural functions" and therefore not vulnerable to disruption.

Also concerning publishers' value proposition filtration, Hagenhoff (2006) acknowledges that progress in metadata, harvesting and Semantic Web technologies enable increasingly reliable selection and aggregation of research papers, and therefore the commercial publishers are no longer needed. Accordingly, Berners-Lee and Hendler (n.d.) explicitly describe the disruptive impact of semantic technologies on scientific publishing.

Bernard Lunn (2010) assumes scientific publishing is not yet disrupted and lists the key elements of the STM process according to the journal functions outlined by Clarke. First copy costs are approximately 80 per cent, 96 per cent of articles are available electronically, and "Companies like Wiley, McGraw-Hill, Elsevier, Wolters Kluwer and Springer ... [are] in good financial health" (ibid). Therefore, the Internet's superiority regarding distribution is not the disruptive force in STM publishing. Also, innovative technologies for registration, validation (peer review) and filtration are not main drivers for disruption in Lunn's view. In contrast to Clarke, Lunn believes the element

designation ... where the researcher gets credit ... may be the main impediment to disruptive change. ... Journals have a power law distribution, like a network effect. The best journals attract the best articles, which have the biggest impact on academic reputation and so on. ... But we see the same power law distribution in social networks. ... As these peer networks do not require the intermediation of a journal brand, they are fundamentally disruptive (ibid).

Furthermore, Lunn writes, in some disciplines, recognition with an Open Access system may start to have a serious impact on academic reputation (designation) (ibid).

Although there are good reasons to deploy semantic technologies predicated on journal publishers' content as Clarke suggests, there is no reason why there should not be a more efficient filtration system not predicated on journals and metadata provided by publishers, by acquiring metadata from other sources and in other ways through authors' and readers' 'contribution'. With Open Access paving the way, it is not evident that a new system would even need publishers' acquiescence for leveraging their assets. Semantic filtration technologies not only "augment" the journal publishing system, but will have disruptive effects on publishers as well.

According to Julia Lane's (2010) call for attention to flaws of existing metrics and her suggestions for "measuring all activities that make up academic productivity" to "make science metrics more scientific" Clarke's affirmation of funding criteria is not entirely convincing. Also, his notion that filtration and designation are not technology-driven but "cultural" functions and therefore are not prone to disruption is unconvincing as a culturally-driven function may also be technology-driven.



Publishers are becoming information solutions providers and scientific support service providers.

Finally, because scientific search tools are improving, more scientists will publish Open Access articles (Lunn 2010). It can be argued that publishers' paid content business models are kept alive because distribution of scientific literature is controlled by them due to distribution being bundled with the publishers' other value propositions of filtration and designation, which are not yet undermined by new technologies. Hence, Open Access beyond author-fee models could take off just as Open Access content becomes available in formats allowing filtration and designation to be done by machines and new hybrid solutions based on semantic technologies. Thereby, semantic technologies could unfold a disruptive potential regarding publishers'

paid content business models by enforcing adoption of Open Access.

Summarising, it can be said that a significant decline in revenues in journal publishing in the coming years is realistic. Thereby, advanced filtration technologies, social networking services and Open Access models can be identified as drivers for disruption. Accordingly, Lunn's (2010) conclusion "that we are on the cusp of disruptive change and that it will be brought on by the implementation of social networking and semantic technology" might be a good starting point for further assessments.

Where the Value Goes

Phillips (2009) in *The Future of Journal Publishing* quotes Brunelle's (2006) report emphasising a "basic shift in business models that is mandated by a move from a journal economy of scarcity (print world) to a journal economy of plenty (online world)" with completely new players flooding the market with free content. Semantic technologies allow grey literature to become more visible while the increasing amount of freely available data will trigger higher demand for science services (Hagenhoff 2006). At the same time the peer review article becomes less relevant.

Grant allocating bodies and researchers themselves rely on the primary research output data rather than text as the main means for evaluation ... The research article often is the gateway into a world of simulations, data analysis, modelling etc. ... It has 'links' to similar databases, to bibliographic databases, has links to images, maps and structures ... but it becomes less essential as standalone entity (Brown and Bouldstone 2008).

Reasons why journals will be less satisfying in the future are the increasing speed of research, the static character of the journal as well as the fact that it is a single mode of communication and relatively isolated (Morris 2009). In the domain of "economics, top authors are moving away from top journals altogether" (Ellison 2007). The concept of an "article within an issue within a journal becomes redundant. Instead users will 'subscribe' to those items that are specifically relevant to their needs irrespective of source" (Brown and Bouldstone 2008). Likewise the proportion of pure

data and the importance of data publishers such as WesternGeco have risen significantly during the last few years (ibid).

Clarke (2010), although opposing the notion of a disruption process in scientific publishing, acknowledges that "new technologies are opening the door for entirely new products and services built on top of – and adjacent to – the existing scientific publishing system". Beside mobile technologies and Open Data standards he lists semantic technologies as particularly promising. Citing King and Tenopir (2000) he explains:

the cost of journals is small relative to the cost, as measured in the time of researchers, of reading and otherwise searching for information ... Which is to say that the value to an institution of workflow applications powered by semantic and mobile technologies and interoperable linked data sets may exceed that of scientific journals. If such applications can save researchers ... significant amounts of time, their institutions will be willing to pay for that time savings and its concatenate increase in productivity.

Clarke also confirms that there will be "a downward pressure on journal pricing" and he refers to Nielsen, emphasizing "that acquiring expertise in information technology (and especially semantic technology) – as opposed to production technology – is of critical importance to scientific publishers". Also Bousfield (2009) emphasises a trend that publishers

move up the value chain. ... Elsevier, Thomson Reuters, and Wolters Kluwer Health are rapidly diversifying their STM publishing divisions in order to add more value to their offerings ... They are all moving away from traditional article publishing into areas that require enterprise scale content aggregation and analysis.

Products based on data mining flourish with Open Access licences, and repository services such as "abstracting and indexing, semantic search and discovery tools, and new ways of presenting the scholarly article ... all can add value and enable publishers to charge for their services" (Pollock 2009). Scientific publishing is shifting from selling static pieces of content toward access-centred models for dynamic content in multiple formats coupled with value-added services. It also has been

indicated that the boundaries between resources themselves and discovery services are increasingly permeable (Brown and Boulderstone 2008). We

will witness a strengthening in secondary information systems over primary publications, with a significant growth in A&I [abstracting and indexing] platforms over the next 3-5 years. Services such as ScienceDirect (Elsevier) and Web of Science (Thomson) are paving the way (ibid).

In this context the “extent of value add provided will determine price and market acceptability ... [and] larger publishers in particular are looking at redefining their business and moving from a content-focus to a service orientation” (ibid). Publishers are becoming information solutions providers and scientific support service providers.

Whether through a disruptive change or an incremental process, Open Access is on the rise, and value is migrating away from journals and content altogether toward technology (Nielsen 2009), workflow applications (Clarke 2010) and services related to the scientific process (Brown and Boulderstone 2008). David Shotton predicts for the coming decade a decrease in the value of raw text while the value of semantic services that help readers to find actionable data, interpret information and to extract knowledge will increase (Shotton 2009).

Generally there are four main areas of growth for scientific information markets: emerging markets (in particular BRIC countries), mobile content and services, the non-library and the extensive-knowledge worker sectors (those people outside the research institutional environment) (Brown and Boulderstone 2008) and finally peer networks and Social Semantic Web services. All four of these areas of growth comprise tendencies of a ‘journal economy of plenty’, decline of importance of the journal article and value shifting toward workflow applications and value-added services related to scientific content – with semantic technologies playing a pivotal role.

Semantic Web

Data can describe any thing (entity) and relationships between things, and both (things and their relationships) are also described in scientific literature.

According to the World Wide Web Consortium (W3C) the main goal of the Semantic Web, or Linked Data, is to extend the principles of the Web from documents to data. Thereby, data itself should be related to one another in a way that it can be shared, reused, integrated and processed across disparate applications and to reveal possible new relationships. “That allows a person, or a machine, to start off in one database, and then move through an unending set of databases which are connected not by wires but by being about the same thing” (W3C 2010).

There is an ongoing debate whether semantic mark-up or machine learning (machines learning to ‘translate’ and ‘read’ human language) will eventually get us to the holy grail of a Semantic Web (O’Reilly 2009). Velden and Lagoze (2008) outline two primary schools of thought about how to achieve standards for semantic mark-up:

One is deeply rooted in the traditions of the artificial intelligence field, concerned with building machine-readable representations of knowledge structures ... (broadly speaking, “ontologies”). The second is more deeply rooted in Web 2.0 principles, leveraging the emergence of structures as they are defined on the fly by users - tag clouds, “folksonomy”, shared bookmarks, and so on. The majority of the semantic work to date in the sciences has revolved around the formalization of ontologies (ibid).

The focus in this article therefore will be on semantic mark-up and ontologies although in later sections examples for social networks in this context will be given and the relevance of social networks for acquiring semantic metadata will be discussed.

In order to achieve the goal of reusing and integrating data across disparate applications, relationships among data between any two resources must be described in common formats using a common ‘language’ for recording how the data relates to real world objects (W3C 2009). The Resource Description Framework (RDF) is a basic data model providing formats and, together with RDF Schema and specific ontologies, a syntax and in this sense ‘language’ for data of a specific domain. In RDF relationships are described in ‘facts’ and

each fact is decomposed into three parts, known as a 'triple': the subject (what is it about?); the predicate (what type of fact or assertion are we making?); and the object (what are we asserting?). In RDF anything that is not a simple data point (e.g. a weight) or label (e.g. a name) must have a unique Web identifier: a URL. This allows machines to follow those links to accumulate new information, types of assertions, etc. (Dodds 2009).

RDF can be described in XML or incorporated in XHTML using RDFa or Microformats. The Public Library of Science (PLOS) for example has RDFa deployed on all of its articles.

Beside RDF, ontologies are essential for cross-platform data integration and processing. An ontology contains definitions of concepts and relationships between concepts in a specific knowledge domain, and includes a taxonomy which represents sub- and super-class relationships between concepts. These concepts describe entities (any thing or object). Standards for semantic markup achieved through formalisation of ontologies "would establish unique names for entities and relationships that would allow the interconnection of disparate information that concerns the same entity" (Velden and Lagoze 2008). Data can be cross-referenced into countless other databases enabling data reuse, and metadata is amenable to automated processing (Shannon 2006). Hence, through the use of RDF and ontologies, scientific literature can be coherently annotated with meaningful and dynamic metadata to enable computers to aggregate and integrate this data, and process assertions made in the literature.

Before further describing how semantic metadata is used in Semantic Publishing it is useful to understand the ramifications of Text Mining.

Text Mining – Components and Layers of Metadata

Text Mining is based on Natural Language Processing (NLP) and Statistical Analysis (SA) (Bousfield 2010), and it is roughly equivalent to text analysis. The UK's National Centre for Text Mining (NaCTeM), a major Text Mining research group and service provider (funded by Nature Publishing Group), describes Text Mining as the proc-

ess of discovering and extracting knowledge from unstructured data. This incorporates information retrieval to gather text, information extraction to identify entities and relationships, annotation and adding of semantic metadata, and finally, Semantic Search and Text Mining to find associations among the pieces of information from many different sources. Information extraction can also implicate recognition of events, opinions, attitudes, certainty and contradictions (Ananiadou 2010). Furthermore, Text Mining can include the use, improvement and production of ontologies. Bousfield (2010) states that for Text Mining a corpus of terms is required which "could be a simple dictionary of terms or a highly structured taxonomy/ontology", and that only in the most sophisticated NLP applications are RDF triples used.

Text Mining refers to an array of different technologies and applications that can be employed in a consecutive way or separately, depending on the system where they are integrated. In accordance with these separate applications, Text Mining generates metadata in different layers: about syntax, semantic entities, relationships, events and meta-knowledge about events (Ananiadou 2010). Semantic Search as one of the final steps of Text Mining can leverage metadata of all layers. To find relevant results it goes beyond general search techniques and considers additional sources, which are related to a user's search terms by using these semantic metadata. Therefore, the result documents do not need to include the query terms, and terms can be disambiguated more easily.

Semantic Publishing

Definition

In his paper *Semantic publishing: the coming revolution in scientific journal publishing* Shotton (2009) defines Semantic Publishing as:

anything that enhances the meaning of a published journal article, facilitates its automated discovery, enables its linking to semantically related articles, provides access to data within the article in actionable form, or facilitates integration of data between papers. Among other things, it involves enriching the article with appropriate metadata that are amenable

to automated processing and analysis, allowing enhanced verifiability of published information and providing the capacity for automated discovery and summarization. These semantic enhancements increase the intrinsic value of journal articles, by increasing the ease by which information, understanding and knowledge can be extracted. They also enable the development of secondary services that can integrate information between such enhanced articles (ibid).

Similar to NaCTeM's description of layers of metadata, De Waard (2010) distinguishes features of Semantic Publishing in entity-based features ('enhanced entities'), relationship-based features (based on triples/facts), and features based on knowledge representation models more complex than relationships.

Enhanced entities

Enhanced entity features require an ontology and entities to be identified and described in RDF. They facilitate disambiguation, improved retrieval precision, and the possibility to search entities through specifying the entity's semantic type. Furthermore it allows users to access additional information about an entity by linking to external sources such as primary research data or visual content.

Features based on RDF Triples

Triple-based features enable more advanced filtering and aggregation based on the detection of associations among concepts or entities, and between a query term and concepts. RDF triples allow fact-based querying as realised with MEDIE, a search engine developed by NaCTeM. MEDIE's search interface facilitates queries as a combination of subject, verb and object. For example, a search for 'p53 activates' brings up documents detailing interactions (relationships) where p53 activates other entities. Results can be shown as sentences or in a tabular form, and they can be sorted by section of the article where they appear. With a built-in ontology, queries can be expanded to find similar relationships matching similar verbs such as 'induce' instead of 'activate'. With triple-based features Liked Data can be 'consumed' and processed to infer new knowledge in real-time.

Beyond RDF Triples

De Waard (2010) calls for going beyond triple-based features to understand the structure of text and narratives as a whole and "the way knowledge is transmitted through discourse". She calls attention to the fact that "without a hypothesis and an interpretation, a data set by itself is not going to transmit any knowledge; ... scientific papers are stories that persuade with data". Together with Groza et al. (2009) she gives a comparative overview of existing "Discourse Representation Models ... with the goal of sketching a unified model" (ibid). Furthermore, the Semantic Web Health Care and Life Sciences (HCLS) Interest Group of the W3C is in the process of developing a "set of elements that can describe the key rhetorical features of a scientific paper" (De Waard 2010).

Also going beyond metadata carried by RDF triples are Topic Maps that can be used to describe relationships between not just two entities, but any number of nodes.

Semantic Publishing – Examples

There has been significant engagement by publishers and other companies to employ semantic technologies for enhanced content and platforms, new science support services and publishing workflow tools. The following examples are roughly grouped together based on the purpose of the application, specific features such as Semantic Search, or companies involved.

Journal Articles

The Semantic Biochemical Journal offers entity enhancement within PDF documents. Reflect, developed by the European Molecular Biology Laboratory (EMBL), is a service that identifies entities of any given web page, highlights them and displays a text box with additional content and links when hovering over the term. Elsevier's Cell Press partners with EMBL to use the tool. The *ChemSpider Journal of Chemistry* is an experiment conducted by the Royal Society of Chemistry (RSC) to demonstrate semantic enhancement of journal articles. Another article by Reis et al. (2008) was hand-crafted by David Shotton (2009) and his team to show the potential of Semantic Publishing.

Enhanced Platforms

Coraal is an effort to index all of Elsevier's ScienceDirect with a set of triples. De Waard (2010) describes how the initiative fell short because users did not understand how to use the triple-based interface. "As interfaces grow more sophisticated and disappear, triples will probably be used by computers but not be exposed to users" (De Waard 2010). The British Library announced the launch of the new UK PubMed Central Open Beta website which is enhanced using large scale Text Mining technology created by NaCTeM. The Royal Society of Chemistry's (RSC) journal, *Molecular Biosystems*, uses a new standard HTML mark-up known as Project Prospect that provides mark-up of chemical entities based on a compendium and various ontologies. Clicking on a chemical name links to its structural formula, a list of synonyms, a specific identifier, or patents involving use of that chemical (Velden and Lagoze 2008).

TEMIS, the leading provider of Text Analytics solutions, has entered into a software licensing and service agreement with Springer Science+Business Media providing them with software for their Semantic Linking project on SpringerLink. This project involves enrichment of journals with hyperlinks into major reference works. TEMIS is also partnering with Thomson Reuters and other publishers. Another initiative by Springer is its SpringerLink Beta platform for "experimenting with statistical text approaches designed to identify semantically related articles" (Bousfield 2010). Aside from TEMIS, Springer also cooperates with the National Center for Biotechnology Information (NCBI), the organisation behind popular sources such as PubMed, PubMed Central and PubChem. TEMIS also partners with Nature Publishing Group (NPG) to identify chemical entities in journal articles and establish links to additional sources of molecular information stored by PubChem and ChemSpider (ibid).

A database for chemical entities was recently launched by the European Bioinformatics Institute (EBI). The database, called ChEBI (Chemical Entities of Biological Interest), includes entity relationships via cross-links to UniProt, the world's most comprehensive repository of protein annotation data. One reason for Macmillan (NPG) to acquire

SureChem, a chemical patent search platform, is its 'semantic expertise'. Wiley offers semantic features with its Blackwell Reference Online platform, and the UK's Joint Information Systems Committee (JISC) is involved in a whole range of Text Mining projects including a joint effort with RSC.

Semantic Search

Publishers can develop their own Semantic Search applications or cooperate with search engine providers by purchasing a technology license, accessing their facilities through an Application Programming Interface (API), or by integrating the solution provider's applications in their own web properties. May (2009) recommends that publishers cede lead generation to the general search providers altogether and then focus on Semantic Search to "add an element of discovery" to their own sites and also to "use external content to supplement".

May also introduces NetBase, a research, Semantic Search, indexing, Text Mining and NLP solution designed to help answer R&D questions.

NetBase is targeting large publishers and companies that want to create their own vertical search tools. NetBase enables an organization to apply a topic or industry-specific lens across its internal content as well as external sources – including the open web – and extract meaning and insight from an aggregated view (ibid).

NetBase parses a sentence into not just entities, but into causal relationships between entities, and discerns entities independent of taxonomy (ibid). NetBase is licensing its technology to drive Elsevier's illumin8 product, an R&D research support tool sold via a subscription model (Pollock Outsell 2008). The key differentiator of NetBase is that by not relying on taxonomies "it scales across subject areas with no need for investment in domain expertise". Pollock (2008b) wonders whether Elsevier's taxonomy-agnostic approach could better compete with platforms like Wolters Kluwer's OvidSP, and "realise operational efficiencies by retiring its multiple vertical search platforms".

Another Semantic Search service is NextBio, offering ontology-based search, collaboration and data sharing tools for the life sciences. The solution integrates "corporate and public data from

next-gen sequencing and microarray technologies". It detects data correlations and enables researchers to intelligently mine this data in real-time. Therefore new knowledge may be discovered and research costs can be decreased by reducing the number of redundant experiments. Elsevier is one of NextBio's partners and uses the technology to enhance its ScienceDirect platform. Another Reed Elsevier company, LexisNexis, has integrated Semantic Search technology for its full range of intellectual property research products through an alliance with Pure Discovery.

Wolters Kluwer included Semantic Search capabilities in some of its online services last year, and Collexis, a major solution provider, cooperates with several publishers to integrate Semantic Search capabilities in their properties. Collexis for example has agreements with CABI, a provider of information and databases in the life sciences, the American Institute of Physics (AIP), a leading society publisher, and Thomson Reuters for its data mining solution Thomson Collexis Dashboard. Another solution provider, Transinsight, offers GoPubMed, a Semantic Search tool built on top of PubMed using the Gene Ontology (GO) and the Medical Subject Headings (MeSH) vocabulary. PubMed is accessing the MEDLINE bibliographic database and other citation databases, abstracts and full text articles on life sciences, biomedicine and biology. The US National Library of Medicine provides PubMed and MEDLINE. Also Medline.Cognition by Cognition Technologies, and NaCTeM's MEDIE are tools for searching MEDLINE based on semantic metadata.

The adoption of Freemium models as well as cross-publisher Text Mining services are strongly related to the publisher's Open Access strategy.

DeepDyve, an online rental service for scientific literature, makes content accessible that often requires specialist knowledge about a domain's vocabulary and is not indexed by traditional search engines. DeepDyve bridges these barriers with semantic technologies and provides an advanced yet easy interface that allows queries of up to 25,000 characters. Hakia Semantic Search relies on Yahoo's search index (May 2009), and develops vertical search sites for specific topic areas such as medicine by giving priority in relevance to preselected credible sources such as PubMed (May 2008). Hakia also has a researcher community based on semantic technologies on its roadmap (ibid). A service already combining social networking and Semantic Search is ResearchGate. More Semantic Search tools include Lexxe, Yebol, Exalead, Microsoft's Powerset, SenseBot, Ask, Evri, which recently merged with Twine, and Sig.ma, a semantic information mashup browser.

Furthermore, relevant for Text Mining generally are the solution providers Linguamatics and Semantic, and the Fraunhofer Institute for Applied Information Technology (FIT) (Bousfield 2010). Linguamatics markets solutions for healthcare and life sciences that can support 'fact-based' queries (ibid). Semantic is already partnering with Wiley-Blackwell, OUP, Brill, Palgrave Macmillan and CABI, and just recently completed a consultancy project for NPG examining the use of taxonomies. Thomson Reuters acquired Discovery Logic to enhance its research analytics, decision support and workflow solutions with Text Mining technology, and Modus Operandi, TopQuadrant and Equentia are also offering semantic technology solutions for publishers.

Social Networking Services

Collexis and Elsevier have just announced an agreement with the University of North Carolina to link granting data and other forms of institutional information with bibliometric information from Scopus to create a social networking tool across the combined 15,000-strong faculties (Bousfield 2010). BiomedExperts uses Collexis' technology for mining the PubMed database for extracting relationships authors have with publishers and other authors, and pre-populates profiles with these

connections. The “networks can be extended by searching for individuals associated with articles that have similar ‘semantic’ profiles” (Strohlein 2010). Similarly, the AIP’s UniPHY networking platform for physical scientists is pre-populated with author profiles.

The Knowledge Media Institute provides a tool for “online collaborative ‘sense-making’ that lets users create, trace, refute, or agree with one another’s claims” (De Waard 2010). Finally, Mendeley is a ‘social’ reference management tool offering research trend analysis as well as recommendations of related material and like-minded academics based on semantic analysis and collaborative filtering. It just opened up an API for selected partners, and is projected by Techcrunch’s O’Hear (2009) to overtake Thomson Reuters’ Web of Science this year.

Publishing Workflow Tools

Beyond enhancement of the publishing platform, search and social networks semantic technologies can also be used for publishing workflow tools to increase editorial performance and decrease costs. Collexis Reviewer Finder is using Semantic Search for finding reviewers by going beyond measuring metrics of scholarly communication and analysing the funding market. The tool has been deployed by several large publishers (Pollock 2008a), and recently has also been licensed to the American Association for Cancer Research (AACR). Although there are all sorts of workflow tools benefiting from semantic technologies, the focus will be on tools for acquiring annotation data.

Dimensions of Semantic Annotation

De Waard (2010) describes three dimensions of semantic annotation: granularity, means, and moment in the workflow process when they are acquired. In regard to granularity, annotations can be gathered to describe entities, triples, claims, documents and whole collections. Concerning means, annotation data can be acquired, automated, semi-automated and manually. Moreover, semantic metadata can be attached by authors; by editors and during production; by readers and users; and during curation and Text Mining processes.

All three dimensions of annotations are essential to assess commercial viability of enhancing

content with annotations, and publishers have to balance out quality of and added value deriving from annotations on the one hand, and costs for acquiring annotation data on the other hand.

Automated Annotations

One way for publishers to acquire annotations is to integrate an automated tagging tool in their workflow. Thomson Reuters’ OpenCalais for example breaks down unstructured content into named entities and facts and returns RDF-data. The tool is backed by powerful Text Mining and machine learning techniques. The data derived through OpenCalais including Linked Data connections are also supported by OpenPublish, a suite for the social publishing platform Drupal, as well as Oracle, a major database used by publishers such as Pearson. Nstein Technologies, part of Open Text and provider of online publishing solutions, offers a Text Mining Engine (TME) for identifying key concepts and context, and encasing content in a layer of rich, semantic metadata. Nstein produces either content in pure RDF format or generates web pages in XHTML + RDFa, enabling content mash-ups, content repackaging and high page-ranking.

Tools like OpenCalais and Nstein’s TME are meant to transform the industry by automation of the annotation process. Nstein, on its website, writes:

ProQuest, one of the world’s largest information aggregators, uses Nstein to automatically tag over 75,000 articles from 2100 journals. A typical worker could handle 10 journals a day; the Nstein solution automates the work of 210 people, resulting in major cost savings.

Annotation by Authors and Editors

However, manual annotation is described as the ‘gold-standard’ of semantic metadata. The Royal Society of Chemistry has taken the lead in pioneering aspects of Semantic Publishing as part of its routine production schedule. Its semantic mark-up is undertaken by skilled domain-specialist editors supported by Text-Mining software. The additional information “adds tremendous value to an article” (Kidd 2007).

Easy-to-use tools are essential for author-created annotations. Chem4Word is a project in collaboration with Microsoft developing applications for

creating semantically rich chemistry information within Word documents (Velden and Lagoze 2008). The Structured Digital Abstracts (SDA) used for the European Biochemical Society's FEBS Letters contains a feature for author-curated data on relationships. The FEBS Journal published by Wiley-Blackwell has a similar semi-automatic approach. One kind of semi-automated annotation is suggesting automatically generated metadata to authors who then 'verify' them in a simple selection procedure.

De Waard (2010) mentions the Okkam project, a Microsoft Word plug-in "to aid author-curated entity (and hopefully soon triple) annotation". An example of a project engaged in identifying argumentation structure in articles beyond entity relationships is SALT (Semantically Annotated LaTeX), a LaTeX-based authoring tool where authors can create rhetorical relationships between sentences (De Waard 2010).

Shotton et al. (2009) tested various automation procedures for annotation and acknowledge that human intervention can be minimised, and automated text processing will help enhancements become affordable and routine. Also, Ruiz-Casado et al. (2006) argue that data about semantic entities and relationships can be extracted with minimal editorial revision.

Collaborative Annotation

Social networking tools are increasingly used within the scientific communities for reference management and bookmarking. StemBook, an online community for stem cell research, uses the Science Collaboration Framework (SCF) (Das et al. 2009) enabling researchers to provide shared semantic context for content using established vocabularies and Text Mining. SCF supports RDF and "captures the semantics of the relationships among the resources and structures discourse around the resources" (ibid). Within ChemSpider, acquired by the RSC last year, the Project Prospect enhanced HTML is supported. With RDF technology layered onto the system, this allows access to entities from over 200 data sources. It also allows community contribution for "cleaning up and improving the quality of the data" (Williams 2009). SWAN (Semantic Web Applications in Neuromedicine) is a project by the Alzheimer Research Forum and oth-

ers to develop a semantically structured framework (Passant et al. 2009) for the Alzheimer disease research communities by providing Social Web applications that allow researchers to author, curate and connect data. It also provides a modularized ontology to foster reusability and integration with other ontologies. A Social Web service can also be used for developing ontologies. The National Center for Biomedical Ontology (NCBO) pursues a Web 2.0 approach for its BioPortal ontology library which supports the researcher community to bring structure and order to an ever increasing amount and complexity of biomedical ontologies (Noy et al. 2009).

To 'allocate' the annotation workload to communities of researchers (rather than editorial staff) it is necessary to bridge the gap between the ambiguity of social tagging and the consistency of controlled ontologies. This would support the editorial process and decrease costs. To achieve this Li et al. (2009) propose models to organise social annotations from a semantic perspective by clustering users, Web pages and annotations according to semantic properties, and by recommending Web pages to users. Also, Guan et al. (2010) with experiments on data sets crawled from Delicious and CiteULike demonstrated the feasibility of document recommendation based on social data. A project called Entity Describer lets users of NPG's social bookmarking tool Connotea tag Web resources with terms from structured knowledge representations (Good 2007). Other semantic collaborative annotation tools include AnnoCryst, Bibsonomy, Fuzzy and ZigTag.



Paid content subscriptions are becoming subscriptions for knowledge discovery services with semantic technologies increasingly determining the subscription price.

Costs and Revenues for Semantic Enhancement

By encouraging author contribution and collaborative annotation as well as using other kinds of data from social networks such as a person's connections, costs for the acquisition of semantic annotations can be reduced. Furthermore, Paschke et al. (2006) examine "Semantic Web Technologies for Content Reutilization Strategies" concluding that efficient content reutilisation leads to decreasing "production and transaction costs by reducing search and coordination costs of editors" (Schulze 2005).

Semantic technologies can be also used to increase revenue by improving the placement of advertising (Bousfield 2010), as shown by Knewco. The Knewco platform provides relevant content and advertisements for readers of life science and healthcare articles based on "concepts on the page, not just keywords" (Brown and Boulderstone 2008), and information not just about the selected paper, but the entire set of established and potential knowledge about the subject related to the paper.

Other revenue streams related to semantic technologies can be established using Freemium models. A user for example could get certain facts extracted from a paper for free while the actual evidence, the complete paper, is accessible only for premium subscribers – or it could be the other way around. The adoption of Freemium models as well as cross-publisher Text Mining services are strongly related to the publisher's Open Access strategy. Some argue that 'restrictive' Pay-Wall models would cripple Text Mining services (Brown and Boulderstone 2008, Velden and Lagoze 2008). On the other hand, services in the sense of Linked Data, not Open Data, could be implemented by opening up the publishers' assets just to partners involved in a specific service.

The gain in popularity of open licensing models and increasing interconnection of data gives reason to consider the Market Place model introduced by Forrester analyst Rotman Epps (2009). With this model publishers would make their content accessible for third-party developers, allowing them to mash-up information via an API and to build their own Semantic Web applications and businesses

based on the publishers' data. Content dissemination and the publisher's revenues attached to the content and its reutilisation could be increased driven by the third-parties' own business rewards.

Finally, as Brown and Boulderstone (2008) emphasise, value-added services will be the distinctive criteria for competing knowledge platforms which today generate the main part of publishers' revenue through subscriptions. The services attached to and integrated in the platform are merging with the resources themselves. Paid content subscriptions are becoming subscriptions for knowledge discovery services with semantic technologies increasingly determining the subscription price.

Conclusion

Semantic technologies change the scientific process in fundamental ways. They enforce publication of raw data, and with dynamic and meaning-based metadata they enable processing of assertions and more complex argumentative structures automatically, in real-time, across very large corpora of data, and by referring to entities rather than words. This undercuts the practices of 'manually' dealing with 'static' documents representing these entities using 'incompatible' domain-specific 'languages'. By sharing and interconnecting research data, redundant work can be minimised, and many products used today for filtering, 'translating' research results and conducting research will become obsolete.

The Internet, as a technology for distributing documents, did not disrupt scientific publishing. In contrast, Semantic Web technologies beside social networking services and Open Access models can be identified as a driver for disruption regarding scientific publishers' value propositions of information filtration and researcher designation as well as their paid content business models. A significant revenue decline in the primary journal publishing market can be expected.

At the same time value migrates away from the journal article toward the content discovery platform, workflow applications for scientists and science support services, with semantic technologies increasingly determining their value.

Semantic technologies also represent major opportunities for scientific publishers to leverage additional and nascent business opportunities.

Through the adoption of semantic technologies, publishers can add value to their content and platforms, develop new content products and 'secondary' value-added services, reuse content to decrease costs, and develop publishing workflow tools to improve editorial performance. The most cost-effective way to acquire semantic metadata must be found and opportunities assessed for leveraging these data to increase revenue using different business models such as advertisement and Freemium models.

The estimated total STM Text Mining sales value of \$40 million (Bousfield 2010) compared to the \$24 billion total value of the STM segment (Bousfield 2009) is significant, and new services and companies based on semantic technologies are taking off. This trend will continue toward mass adoption, and scientists and funding organisations will allocate their resources following more effective solutions, which will be based on semantic technologies. Publishers should anticipate the changes and (further) shoulder the long-term investment to be able to offer these advanced solutions first hand. They "need to include a semantic strategy in their business models if they plan on succeeding" (Hawkins 2009).

Recommendations

Existing semantic technologies and related companies and products should be analysed in a more comprehensive way. It is also recommended to analyse in detail publishers' different revenue streams to solve the contradiction between predictions of disruption in scientific publishing and other predictions of total value of the STM information market rising until 2012 (Bousfield 2009). Furthermore, an analysis of long-term costs and revenues, including business models, for semantically enhanced products should be conducted. Challenges for implementation of semantic technologies should be rationalised, and the range and nature of 'pre-competitive' engagement in standards and cross-publisher services examined. Additionally the wider effects of Semantic Publishing on innovation and generativity in science are relevant topics worth exploring.

Finally, it is recommended to analyse diversification strategies of current Semantic Publishing players to assess and anticipate scientific information market development related to semantic technologies. □

References

- Ananiadou, S., 2010. *Text Mining: unlocking the literature in STM publishing*. Available at: http://www.alpsp.org/ngen_public/article.asp?aid=185590 [Accessed June 9, 2010].
- Berners-Lee, T. and Hendler, J., n.d. *Scientific publishing on the 'semantic web'*. Available at: <http://www.nature.com/nature/debates/e-access/Articles/bernerslee.htm> [Accessed May 24, 2010].
- Bousfield, D., 2010. *From "Text Mine!" to Text Mining: STM Text Analytics Come of Age*. Available at: <https://clients.outsellinc.com/insights/index.php?p=11189> [Accessed May 24, 2010].
- Bousfield, D., 2009. *Scientific, Technical and Medical Information: 2009 Market Forecast and Trends Report*. Outsell, Inc.
- Brown, D. and Boulderstone, R., 2008. *The Impact of Electronic Publishing: The Future for Publishers and Librarians*. München: K.G. Saur.
- Bruck, P., 2008. *Multimedia and E-Content Trends: Implications for Academia*. 1st ed., Wiesbaden: Vieweg Teubner.
- Brunelle, B., 2006. *Publishers Speak Up On Open Access: Big Promise, Small Uptake*. Outsell, Inc.
- Christensen, C.M., 2008. *Reinventing Your Business Model*. Harvard Business IdeaCast 122. Available at: <http://itunes.apple.com/podcast/harvard-business-ideacast/id152022135> [Accessed February 23, 2010].
- Clarke, M., 2010. *Why Hasn't Scientific Publishing Been Disrupted Already?*. The Scholarly Kitchen. Available at: <http://scholarlykitchen.sspnet.org/2010/01/04/why-hasnt-scientific-publishing-been-disrupted-already/> [Accessed April 13, 2010].
- Cope, B. and Kalantzis, M., 2009. *Signs of epistemic disruption: transformation in the knowledge system of the academic journal*. In: *The Future of the Academic Journal*. Oxford: Chandos.
- Das, S. et al., 2009. *Building biomedical web communities using a semantically aware content management system*. *Briefings in Bioinformatics*, 10(2), 129-138.
- De Waard, A., 2010. *From Proteins to Fairytales: Directions in Semantic Publishing*. *IEEE Intelligent Systems*, vol. 25, no. 2, pp. 83-88, Mar./Apr. 2010, doi:10.1109/MIS.2010.49.
- Dodds, L., 2009. *The Web's rich tapestry*. *Learned Publishing*, 22(4), 275-280.
- Ellison, G., 2007. *Is Peer Review in Decline?* NBER Working Paper. Available at: <http://www.nber.org/papers/w13272.pdf>.
- Good, B., 2007. *Bridging the gap between social tagging and semantic annotation: E.D. the Entity Describer*. *Nature Precedings*: hdl:10101/npre.2007.945.1: Posted 7 Sep 2007.
- Groza, T. et al., 2009. *A Short Survey of Discourse Representation Models*. *Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009)*, Washington DC, USA. Available at: <http://ceur-ws.org/Vol-523/Groza.pdf>.
- Guan, Z. et al., 2010. *Document Recommendation in Social Tagging Services*. *Proceedings of the 19th international conference on World Wide Web, Raleigh, USA, 2010*. Available at: <http://portal.acm.org/citation.cfm?id=1772731>.
- Hagenhoff, S., 2006. *Internetökonomie der Medienbranche*. Göttingen: Universitätsverlag Göttingen.
- Hawkins, D.T., 2009. *Publishing 2.0: SSP Part 2*. *Information Today*, 26(8).
- Kidd, R., 2007. *Semantic enrichment boosts information retrieval*. Available at: http://www.researchinformation.info/features/feature.php?feature_id=127FirefoxHTML\Shell\Open\Command [Accessed May 24, 2010].
- Lane, J., 2010. *Let's make science metrics more scientific*. *Nature*, 464(7288), 488-489.
- Li, H. et al., 2009. *Exploring Social Annotations with the Application to Web Page Recommendation*. *Journal of Computer Science and Technology*, 24(6), 1028-1034.
- Lunn, B., 2010. *Semantic Wave Hits STM Publishing, Part 1: Current Cash Cows*. *Semantic Web*. Available at: http://www.semanticweb.com/features/semantic_wave_hits_stm_publishing_part_1_current_cash_cows_154355.asp [Accessed April 13, 2010].
- May, N., 2008. *Hakia: Upstart Semantic Search Player You Need to Watch*. Outsell, Inc.
- May, N., 2009. *NetBase: Enabling the Next Stage of Information Consumption – Today*. Outsell, Inc.
- Morris, S., 2009. *'The tiger in the corner': will journals matter to tomorrow's scholars*. In: *The future of the academic journal*. Oxford: Chandos.
- Nielsen, M., 2009. *Is scientific publishing about to be disrupted?* Michael Nielsen. Available at: <http://michaelnielsen.org/blog/is-scientific-publishing-about-to-be-disrupted/> [Accessed April 13, 2010].
- Noy, N.F. et al., 2009. *Harnessing the Power of the Community in a Library of Biomedical Ontologies*. *Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009)*, Washington DC, USA. Available at: <http://sunsite.informatik.rwth-aachen.de/>

- Publications/CEUR-WS/Vol-523/Noy.pdf.
- O'Hear, S., 2009. *Mendeley, the Last.fm-of-research, could be world's largest online research paper database by early 2010*. Available at: <http://eu.techcrunch.com/2009/11/18/mendeley-the-last-fm-of-research-could-be-world%E2%80%99s-largest-online-research-paper-database-by-early-2010/> [Accessed May 24, 2010].
- O'Reilly, T., 2009. *Google's Rich Snippets and the Semantic Web*. O'Reilly Radar. Available at: <http://radar.oreilly.com/2009/05/google-rich-snippets-semantic-web.html> [Accessed May 24, 2010].
- Paschke, A. et al., 2006. *Semantic Web Technologies for Content Reutilization Strategies in Publishing Companies*. International Conference on Web Information Systems and Technologies (WEBIST'06), Setubal, Portugal, 2006.
- Phillips, A., 2009. *Business models in journal publishing*. In: *The Future of the Academic Journal*. Oxford: Chandos.
- Passant, A. et al., 2009. *SWAN/SIOC: Aligning Scientific Discourse Representation and Social Semantics*. The 1st International Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009) at the 8th International Semantic Web Conference (ISWC 2009), Washington, DC, USA, 26 October 2009. Available at: http://www.johnbreslin.org/files/publications/20091026_swad2009.pdf.
- Pollock, D., 2008a. *Collexis' Reviewer Finder Takes Nothing for Granted*. Outsell, Inc. Available at: <https://clients.outsellinc.com/insights/index.php?p=10696> [Accessed May 24, 2010].
- Pollock, D., 2008b. *Elsevier's Illumin8 Shines a Light on Semantic Search Applications*. Outsell, Inc. Available at: <https://clients.outsellinc.com/insights/index.php?p=10702> [Accessed May 24, 2010].
- Pollock, D., 2009. *An Open Access Primer – Market Size and Trends*, Outsell, Inc.
- Reis, R.B. et al., 2008. *Impact of Environment and Social Gradient on Leptospira Infection in Urban Slums*. R. E. Gurtler, ed. PLoS Neglected Tropical Diseases, 2(4), e228.
- Rotman Epps, S., 2009. *Eight Models For Monetizing Digital Content*, Forrester Research.
- Ruiz-Casado, M. et al., 2006. *From Wikipedia to Semantic Relationships: a Semi-automated Annotation Approach?*. 1st Workshop on Semantic Wikis: From Wiki to Semantics, at the 3rd European Semantic Web Conference (ESWC 2006). Budva. Available at: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.74.3115>.
- Schulze, B., 2005. *Mehrfachnutzung von Medieninhalten : Entwicklung, Anwendung und Bewertung eines Managementkonzepts für die Medienindustrie*. 1st ed., Lohmar, Köln: Eul.
- Shannon, V., 2006. *A 'more revolutionary' Web*. The New York Times. Available at: http://www.nytimes.com/2006/05/23/technology/23iht-web.html?_r=2 [Accessed April 13, 2010].
- Shotton, D., 2009. *Semantic publishing: the coming revolution in scientific journal publishing*. Learned Publishing, 22(2), 85-94.
- Stratigos, A.C., Strohlein, M. and Watson Healy, L., 2009. *Information Industry Outlook 2010: A New Dawn, New Day, New Decade*. Outsell, Inc.
- Strohlein, M., 2010. *2010 – The Year of Reckoning: Five Crucial Technologies for Information Publishing*. Outsell, Inc.
- Tenopir, C. and King, D.W., 2000. *Towards electronic journals: realities for scientists, librarians, and publishers*. Washington DC: Special Libraries Association.
- Velden, T. and Lagoze, C., 2008. *The Value of New Scientific Communication Models for Chemistry*. White Paper, Workshop for New Models for scholarly Communication in Chemistry, Washington DC, 23-24 October 2008.
- W3C, 2010. *W3C Semantic Web Activity*. Available at: <http://www.w3.org/2001/sw/> [Accessed April 13, 2010].
- W3C, 2009. *W3C Semantic Web FAQ*. Available at: <http://www.w3.org/2001/sw/SW-FAQ#whatarebuildingblocks> [Accessed April 13, 2010].
- Williams, A., 2009. *Community curation helps chemical information*. Research Information. Available at: http://www.researchinformation.info/features/feature.php?feature_id=230FirefoxHTML\Shell\Open\Command [Accessed May 24, 2010].