



Text in a Data-centric World

PAUL GINSPARG
Cornell University

FIRST MET JIM GRAY WHEN HE WAS THE MODERATOR of the database subject area of arXiv, part of the expansion into computer science that arXiv initiated in 1998. Soon afterward, he was instrumental in facilitating the full-text harvest of arXiv by large-scale search engines, beginning with Google and followed by Microsoft and Yahoo!—previous robotic crawls of arXiv being overly restricted in the 1990s due to their flooding of the servers with requests. Jim understood the increasing role of text as a form of data, and the need for text to be ingestible and treatable like any other computable object. In 2005, he was involved in both arXiv and PubMed Central and expressed to me his mystification that while the two repositories served similar roles, they seemed to operate in parallel universes, not connecting in any substantive way. His vision was of a world of scholarly resources—text, databases, and any other associated materials—that were seamlessly navigable and interoperable.

Many of the key open questions regarding the technological transformation of scholarly infrastructure were raised well over a decade ago, including the long-term financial model for implementing quality control, the architecture of the article of the future, and how all of the pieces will merge into an interoperable whole. While answers have remained elusive, there is reason to expect significant near-term progress on at least the latter two



questions. In [1], I described how the range of possibilities for large and comprehensive full-text aggregations were just starting to be probed and offered the PubMed Central database as an exemplar of a forward-looking approach. Its full-text XML documents are parsed to permit multiple “related material views” for a given article, with links to genomic, nucleotide, inheritance, gene expression, protein, chemical, taxonomic, and other related databases. This methodology is now beginning to spread, along with more general forms of semantic enhancement: facilitating automated discovery and reasoning, providing links to related documents and data, providing access to actionable data within articles, and permitting integration of data between articles.

A recent example of semantic enhancement by a publisher is the Royal Society of Chemistry’s journal *Molecular BioSystems*.¹ Its enhanced HTML highlights terms in the text that are listed in chemical terminology databases and links them to the external database entries. Similarly, it highlights and links terms from gene, sequence, and cell ontologies. This textual markup is implemented by editors with subject-matter expertise, assisted by automated text-mining tools. An example of a fully automated tool for annotation of scientific terms is EMBL Germany’s Reflect,² which operates as an external service on any Web page or as a browser plug-in. It tags gene, protein, and small molecule names, and the tagged items are linked to the relevant sequence, structure, or interaction databases.

In a further thought experiment, Shotton et al. [2] marked up an article by hand using off-the-shelf technologies to demonstrate a variety of possible semantic enhancements—essentially a minimal set that would likely become commonplace in the near future. In addition to semantic markup of textual terms and live linkages of DOIs and other URLs where feasible, they implemented a reorderable reference list, a document summary including document statistics, a tag cloud of technical terms, tag trees of marked-up named entities grouped by semantic type, citation analysis (within each article), a “Citations in Context” tooltip indicating the type of citation (background, intellectual precedent, refutation, and so on), downloadable spreadsheets for tables and figures, interactive figures, and data fusion with results from other research articles and with contextual online maps. (See Figure 1.) They emphasize the future importance of domain-specific structured digital abstracts—namely, machine-readable metadata that summarize key data and conclusions of articles, including a list of named entities in the article with precise database iden-

¹ www.rsc.org/Publishing/Journals/mb

² <http://reflect.ws>, winner of the recent Elsevier Grand Challenge (www.elseviergrandchallenge.com).

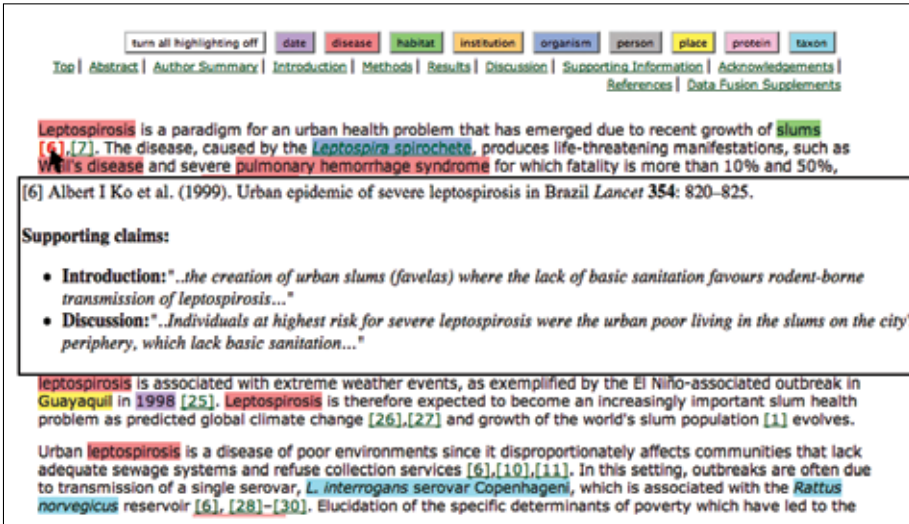


FIGURE 1.

A screenshot of “Exemplar Semantic Enhancements” from <http://imageweb.zoo.ox.ac.uk/pub/2008/plospaper/latest>, as described in [2]. Different semantic classes of terms are linked and can be optionally highlighted using the buttons in the top row. Hovering the mouse pointer over an in-text reference citation displays a box containing key supporting statements or figures from the cited document.

tifiers, a list of the main results described via controlled vocabulary, and a description, using standard evidence codes, of the methodology employed. The use of controlled vocabularies in this structured summary will enable not only new metrics for article relatedness but also new forms of automated reasoning.

Currently, recognition of named entities (e.g., gene names) in unstructured text is relatively straightforward, but reliable extraction of relationships expressed in conventional text is significantly more difficult. The next generation of automated knowledge extraction and processing tools, operating on structured abstracts and semantically enhanced text, will bring us that much closer to direct searching and browsing of “knowledge”—i.e., via synthesized concepts and their relationships. Further enhancements will include citation network analysis, automated image analysis, more generalized data mashups, and prekeyed or configurable algorithms that provide new types of semantic lenses through which to view the text, data, and images. All of these features can also be federated into hub environments where



users can annotate articles and related information, discover hidden associations, and share new results.

In the near term, semantic text enhancement will be performed by a combination of semi-supervised tools used by authors,³ tools used by editors, and automated tools applied to both new and archival publications. Many legacy authors will be unwilling to spend time enhancing their documents, especially if much additional effort is required. Certainly many publishers will provide the markup as a value-added component of the publication process—i.e., as part of their financial model. The beneficial effects of this enhancement, visible to all readers, will create pressure in the open sector for equally powerful tools, perhaps after only a small time lag as each new feature is developed. It is more natural to incorporate the semantics from the outset rather than trying to layer it on afterwards—and in either case, PDF will not provide a convenient transport format. With the correct document format, tools, and incentives, authors may ultimately provide much of the structural and semantic metadata during the course of article writing, with marginal additional effort.

In the longer term, there remains the question of where the semantic markup should be hosted, just as with other data published to the Web: Should publishers host datasets relevant to their own publications, or should there be independent SourceForge-like data repositories? And how should the markup be stored: as triplestores internal to the document or as external attachments specifying relationships and dependencies? As knowledge progresses, there will be new linkages, new things to annotate, and existing annotations that may lead to changed resources or data. Should it be possible to peel these back and view the document in the context of any previous time frame?

To avoid excessive one-off customization, the interactions between documents and data and the fusion of different data sources will require a generic, interoperable semantic layer over the databases. Such structures will also make the data more accessible to generic search engines, via keyword searches and natural-language queries. Having the data accessible in this way should encourage more database maintainers to provide local semantic interfaces, thereby increasing integration into the global data network and amplifying the community benefits of open access to text and data. Tim Berners-Lee⁴ has actively promoted the notion of linked data

³ For example, Pablo Fernicola's "Article Authoring Add-in for Microsoft Office Word 2007," www.microsoft.com/downloads/details.aspx?familyid=09c55527-0759-4d6d-ae02-51e90131997e.

⁴ www.w3.org/DesignIssues/LinkedData.html



for all such purposes, not just by academics or for large and commonly used databases. Every user makes a small contribution to the overall structure by linking an object to a URI, which can be dereferenced to find links to more useful data. Such an articulated semantic structure facilitates simpler algorithms acting on World Wide Web text and data and is more feasible in the near term than building a layer of complex artificial intelligence to interpret free-form human ideas using some probabilistic approach.

New forms of interaction with the data layer are also embedded in discussions of Wolfram|Alpha,⁵ a new resource (made publicly available only after this writing) that uses substantial personnel resources to curate many thousands of data feeds into a format suitable for manipulation by a Mathematica algorithmic and visualization engine. Supplemented by a front end that interprets semi-natural-language queries, this system and its likely competition will dramatically raise user expectations for new forms of synthesized information that is available directly via generic search engines. These applications will develop that much more quickly over data repositories whose semantic layer is curated locally rather than requiring centralized curation.

Much of the recent progress in integrating data with text via semantic enhancement, as described above, has been with application to the life sciences literature. In principle, text mining and natural-language processing tools that recognize relevant entities and automatically link to domain-specific ontologies have natural analogs in all fields—for example, astronomical objects and experiments in astronomy; mathematical terms and theorems in mathematics; physical objects, terminology, and experiments in physics; and chemical structures and experiments in chemistry. While data-intensive science is certainly the norm in astrophysics, the pieces of the data network for astrophysics do not currently mesh nearly as well as in the life sciences. Most paradoxically, although the physics community was ahead in many of these digital developments going back to the early 1990s (including the development of the World Wide Web itself at CERN, a high-energy physics lab) and in providing open access to its literature, there is currently no coordinated effort to develop semantic structures for most areas of physics. One obstacle is that in many distributed fields of physics, such as condensed-matter physics, there are no dominant laboratories with prominent associated libraries to establish and maintain global resources.

⁵ www.wolframalpha.com, based on a private demonstration on April 23, 2009, and a public presentation on April 28, 2009, <http://cyber.law.harvard.edu/events/2009/04/wolfram>.



In the biological and life sciences, it's also possible that text will decrease in value over the next decade compared with the semantic services that direct researchers to actionable data, help interpret information, and extract knowledge [3]. In most scientific fields, however, the result of research is more than an impartial set of database entries. The scientific article will retain its essential role of using carefully selected data to persuade readers of the truth of its author's hypotheses. Database entries will serve a parallel role of providing access to complete and impartial datasets, both for further exploration and for automated data mining. There are also important differences among areas of science in the role played by data. As one prominent physicist-turned-biologist commented to me recently, "There are no fundamental organizing principles in biology"⁶—suggesting that some fields may always be intrinsically more data driven than theory driven. Science plays different roles within our popular and political culture and hence benefits from differing levels of support. In genomics, for example, we saw the early development of GenBank, its adoption as a government-run resource, and the consequent growth of related databases within the National Library of Medicine, all heavily used.

It has also been suggested that massive data mining, and its attendant ability to tease out and predict trends, could ultimately replace more traditional components of the scientific method [4]. This viewpoint, however, confuses the goals of fundamental theory and phenomenological modeling. Science aims to produce far more than a simple mechanical prediction of correlations; instead, its goal is to employ those regularities extracted from data to construct a unified means of understanding them *a priori*. Predictivity of a theory is thus primarily crucial as a validator of its conceptual content, although it can, of course, have great practical utility as well.

So we should neither overestimate the role of data nor underestimate that of text, and all scientists should track the semantic enhancement of text and related data-driven developments in the biological and life sciences with great interest—and perhaps with envy. Before too long, some archetypal problem might emerge in the physical sciences⁷ that formerly required many weeks of complex query traversals of databases, manually maintained browser tabs, impromptu data analysis scripts, and all the rest of the things we do on a daily basis. For example, a future researcher with seamless semantic access to a federation of databases—including band structure properties and calculations, nuclear magnetic resonance (NMR)

⁶ Wally Gilbert, dinner on April 27, 2009. His comment may have been intended in a more limited context than implied here.

⁷ As emphasized to me by John Wilbanks in a discussion on May 1, 2009.



and X-ray scattering measurements, and mechanical and other properties—might instantly find a small modification to a recently fabricated material to make it the most efficient photovoltaic ever conceived. Possibilities for such progress in finding new sources of energy or forestalling long-term climate change may already be going unnoticed in today’s unintegrated text/database world. If classes of such problems emerge and an immediate solution can be found via automated tools acting directly on a semantic layer that provides the communication channels between open text and databases, then other research communities will be bootstrapped into the future, benefiting from the new possibilities for community-driven scientific knowledge curation and creation embodied in the Fourth Paradigm.

REFERENCES

- [1] P. Ginsparg, “Next-Generation Implications of Open Access,” www.ctwatch.org/quarterly/articles/2007/08/next-generation-implications-of-open-access, accessed Aug. 2007.
- [2] D. Shotton, K. Portwin, G. Klyne, and A. Miles, “Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article,” *PLoS Comput. Biol.*, vol. 5, no. 4, p. e1000361, 2009, doi: 10.1371/journal.pcbi.1000361.
- [3] P. Bourne, “Will a Biological Database Be Different from a Biological Journal?” *PLoS Comput. Biol.*, vol. 1, no. 3, p. e34, 2005, doi: 10.1371/journal.pcbi.0010034. This article was intentionally provocative.
- [4] C. Anderson, “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete,” *Wired*, June 2008, www.wired.com/science/discoveries/magazine/16-07/pb_theory. This article was also intentionally provocative.