Standard Domain Ontologies: The Rate Limiting Step for the Next Big Change in Scientific Communication

Division of Chemical Information
American Chemical Society
Chicago, April 27th, 2007

Allen Renear
Electronic Publishing Research Group
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

Last change 3/30/07. Comments to renear@uiuc.edu

Abstract

The long awaited emergence of high-function scientific publishing may, finally, be near. There will soon be the tools, structured data, and communication infrastructure that will allow researchers to use new and innovative strategies for taking advantage of computationally available representations of scientific information. As this happens, the use of traditional publishing artifacts like journals, abstracts and articles will be increasingly away from simply finding and reading and towards more direct and efficient computer-supported exploitation. Several important social and technological trends are converging to make this possible. We focus here on the role of standard domain ontologies and their potential for interaction with changing user behavior in search environments.

In a nutshell…

What?

The first major internet-driven change in scientific publishing is over.

It ended when network access to PDF files became routine.

The next, which builds on the last, may be near.

It is the old dream of radical new functionality for scientific communication.

Why … or rather, why *now* ?

Permanent driving forces

increasing quantity and complexity of relevant information, increasing competition

Changing background expectations

advanced interactivity is now routine, and expected

Detectable direction of motion

a trajectory away from "finding" and then "reading" and towards others ways of assessing and exploiting intellectual content.

Recent enabling conditions [*]

technological advances and social changes aligned with this trajectory and that sustain and shape it.

New occasions

a re-emergence of evangelism — this time from domain practitioners.

Enabling conditions…

Two enabling conditions are particularly important.

Evolving changes in user behavior in comprehensive indexing and navigation environments:

Users rapidly navigate more and more articles, spending less and less time with each and attempting to assess and exploit content without reading the article.

The explosion of computationally accessible domain models in the form of XML markup languages, metadata systems, conceptual models, and ontologies

- This I think is the rate-limiting step for the next revolution in scientific publishing

What are we talking about?

The grand old dream of radical new functionality as envisioned by Paul Otlet, Vannevar Bush, Douglas Engelbart, and Ted Nelson:

— advanced navigation and viewing optimized for browsing and analysis,

— computationally available data accessible with discipline-specific tools,

— typed hypertext linking with links as first class objects,

— data-driven intereactive diagrams and graphics

— computable equations,

— thoroughgoing interoperability with other tools

… and so on, and on

The new scientific journal seemed imminent in the mid-80s,

…so we were astonished that the 80s dragged on without the revolution

… finally it seem to be starting, in 1992,
with the *Online Journal of Current Clinical Trials*

… only … only …

Only…

Only it didn't

But something else happened, a different revolution.

> Not without value, but not what we were looking forward too, at all.

Now I'm back, returned to tell you…

Ok … and it didn't happen in 1992

… and it didn't happen in 2002

But it *could* happen around **2012**

… no, *really*, it could! (it will)

Why now? (again)

- Permanent driving forces
- Changing background expectations
- Detectable direction of motion
- New occasions
- Enabling conditions

Permanent driving forces

- growing quantities relevant information
- increasing complexity of relevant information
- an increasingly competitive research environment

These are powerful and permanent

… and, at some point…

differences make a difference

"Nowadays … sets of relevant papers [are] identified that surpass human capability for reading, interpretation, and synthesis."

– Barend Mons "Which gene did you mean?"
BMC Bioinformatics 6:142 2005

Detectable direction of motion

There is trajectory away from "finding" and "reading" and towards new ways of assessing and exploiting, or mobilizing, intellectual content.

The goal of researchers searching and navigating the literature is *not* to find something to read…

… It is to *avoid reading*

Avoiding reading…

- Indexing and citation analysis help us decide whether or not articles are relevant…
  … *without reading them.*

- Abstracts and literature reviews help us take advantage of articles…
  … *without reading them.*

- The articles we do read, in their analyses and summaries help us take advantage of other articles…
  … *without reading them.*

- Text mining and data mining for "undiscovered public knowledge" help us take advantage of articles…
  … *without reading them.*

- Colleagues, and, best of all, graduate students, help us take advantage of articles…
  … *without reading them.*

Changing background expectations

Advanced functionality is now routine, and expected.

Consider the variety and level of functionality on shopping, news, travel, and stock trading sites

New occasions

New champions …

The current transition to e-journals seems to be welcomed by many — but not us …

The *datument* is a hypermedia document accessible to robots and humans … for transmitting "complete" information including content and behaviour.

… the machine is … semantically aware of the document content [through] domain-specific XML components ...

… [machine] understandability may require ontological (meaning) or semantic (behaviour) support for components. Neither are yet fully formalised but within domains it is often possible to find that certain concepts are sufficiently agreed that programs from different authors will behave in acceptable manners on the same documents.

We argue that a cultural change in our approach to information is needed.

P. Murray-Rust and H. S. Rzepa, "The Next Big Thing: From Hypermedia to Datuments," *Journal of Digital Information*, 5:1 2004

New champions cont'd.

Imagine what could be achieved if articles, rather than consisting entirely of free-form natural languages, contained explicit assertions about biological knowledge in unambiguous machine readable form … some progress is being made…

… for example … you should be able to cut and paste the equation below into any MathML aware application…

Mathew Cockerill, Editorial, *BMC Bioinfomatics*, 6:140 2005.

Enabling conditions

There are specific technological advances and social changes aligned with this trajectory and that sustain and shape it.

- tremendous improvements in functionality, interoperability, and efficiency of basic communication and networking technology: networking, hardware and software, underlying protocols, etc.

- Industrial and social infrastructures that support systemic social/technological changes

- ⁕ New trends user behavior in indexing and navigation environments will create demand for new ways to engage with scientific literature.

- ⁕ Explosion of domain specific XML schemas, models, and ontologies.

Changing user behavior in SSEs

SSE's = Scholarly Search Environments.

- … comprehensive indexing, search, and navigation environments such as Google Scholar, Thomson ISI's SCOPUS, Citseer, and related environments that support navigation are already extremely important to many scholars

The *SSE trance*.

In SSEs researchers engage with the literature as if playing a video game

SSE users…

- rapidly, almost subconsciously develop queries likely to find known items, or retrieve subject or topic result sets, etc.
- track references backward and citations forward,
- dodge publisher sites, commercial integrator sites, and appropriate copies to hunt for open-access copies
- make rapid relevance judgments: assessments of impact, and quality,

How strange!

- this is almost sub-cognitive, kinaesthetic, even trance-like,
- users often unable to easily articulate what they were doing or why
- sessions are routinely described as successful — even though no article to was ever *read*.

And inside …?

Documents are skimmed rapidly, making use of key components

> …engineers describe a common pattern for utilizing document components by zooming in on and filtering information in their initial reading of an article. They tend to first read the abstract, then skim section headings. Next they look at lists, summary statements, definitions, and illustrations.
>
> … they disaggregate and reaggregate article components for use in their own work … perhaps by using a marker to highlight … perhaps by creating a mental register
>
> B. Schatz et al. "Federated Search of Scientific Literature" *IEEE Computer*, 1999.

And inside… ?

The goal is not to find an article to read.

It is to find, assess, and exploit relevant information, often in the form of equations, data, and other technical expressions.

Informant:

…I used the sections of the papers for the equations. I even wouldn't read all the other parts of the article.

l look for specific surface tensions, experimental measurements.

I recently looked for the efficiency of an electric motor … I had to just search the entire database for the term 'electric motor'; you can spend hours looking this way.

I sometimes need to look specifically at other methods and theories.

A. Bishop. "Document Structure and Digital Libraries: how Researchers Mobilize Information in Journal Articles". *Information Processing and Management*, 1999.

What's going on here?

Again, users routinely describe sessions as successful even though no article to "read" was located and read.

Because the goal is not to find an article to read,

The goal is to avoid reading articles.

Relevant empirical research: reading time

Amount of reading

- Until the mid-late-1990s the number of articles read by researchers appeared steady.
- Since then the number of articles "read" has been climbing, and apparently rather steeply (perhaps c. 30%),
- However time spent reading is constant
  - and so reading time per article is dropping, *fast*.
- Time searching and browsing appears to have *doubled* from 1984 to 2000.
- Time reading and browsing on the screen is steadily going up.
- etc…

C. Tenopir, "How Electronic Journals are Changing Scholarly Reading Patterns". CONCERT 2006.
See also papers by Tenopir and King, and others 2003-2006;

Relevant empirical research: Searching behavior

...Now we see what the migration from traditional to electronic sources has meant…

We are all *bouncers* and *flickers*, and the success of Google is a testament to that, with its marvellous ability to enhance and amplify this flicking and bouncing

… This analysis of the searching behaviour of digital consumers tells us … more than that, it also shows us how people develop knowledge.

A slightly irritated father watching his young daughter using the remote to flick from one television channel to another … asks why she cannot make up her mind and she answers: she is not attempting to make up her mind *but is watching all the channels*.

She, like our bouncers, *is gathering information horizontally*, not vertically.

D. Nicholas, P. Huntington, P. Williams, Tom Dobrowolski, "Re-appraising information seeking behaviour in a digital environment: Bouncers, checkers, returnees and the like". *Journal of Documentation* 60:1 2004 [adapted by ahr].

cf. additional papers by Nicholas, Huntingon, Jamali, Hamid, Monopoli, and Watkinson and from the Ciber Virtual Scholar research programme

What we still need to know

When *leading users* engage with the literature in circumstances that are *optimal* and *exemplary*

what *exactly* are they doing …

or *trying to do …*

or *would try to do* … if they could?

… *what are they thinking*?

The explosion of domain models in the sciences

[The second enabling condition]

Standards for interoperability

- Global standards for *serialization interoperability*
  (e.g., XML)
  [Adoption: nearly total]

- Global standards for *syntactic interoperability*
  (i.e. RDF(S), OWL)
  [Adoption: rapid growth underway]

- Global standards for (general) *semantic interoperability*
  (e.g, Cyc, SUMO, BFO, Dolce, etc.)
  [Adoption: slight]

- Domain standards for domain specific *semantic interoperability*
  (XML schemas, conceptual models, and domain ontologies)
  [Adoption: varies widely across fields]

Global standards for serialization interoperability

XML

- a metalanguage for document and data markup languages.
- serializes a particular very well-understood data structure
  - a directed acyclic graph
  - with labeled nodes,
  - and attribute/value pairs on nodes,
  - and data content in the leaf nodes.

- Easy to use and read
- Widely used for data modeling as well as serialization

Within 5 years (since 1998) [or as SGML, 20 years]

- Complete domination of serialization and interchange on the web
- Complete domination of content modeling for documents

Global standards for syntactic(*) interoperability

RDF:

a standard for

dyadic predicative assertion

"Herman Melville is the author of Moby Dick"
"Melatonin modulates glutamate toxicity"

Has a web-oriented XML serialization

RDFS

a standard for

introducing domain vocabularies

defining basic semantic relationships
(sub/super classes, sub/super relation types, domain/range restriction

Has a web-oriented XML serialization

OWL:

an RDF(S) standard standard for predicate logic

allows choice in the space of the expressiveness vs. efficiency tradeoff.

OWL Full: equivalent to first order logic
expressive, but allows intractable and undecidable queries.

OWL DL: equivalent to "description logic"
less expressive, but decidable and generally fairly efficient query processing.

OWL Lite: further restrictions
less expressive still, but very efficient query processing.

Standards for global semantic interoperability

- So-called “upper ontologies”, general descriptions of the world
  - Physical things, abstract things, collections, artifacts, places, times, persons, etc.
- Examples: Cyc, SUMO, BFO, Dolce, etc.
- Of theoretical interest, but only isolated practical applications so far.
- Originally thought to be essential wide spread use of ontologies — but turned out not to be.

Standards for domain semantic interoperability

- XML schemas (e.g. MathML, MGED, CML, etc.)
- Conceptual models (expressed in EER, UML, etc),
- Domain ontologies (expressed in RDFS, OWL)
- As well as controlled vocabularies, thesauri, etc.

Scientific ontologies

- Within the scientific disciplines, as in industry, there has been an explosion of standardized models and ontologies
  - In biomedical sciences there is an enormous investment, with some really astonishing successes (Gene Ontology)

- These are typically expressed in XML markup languages

- And sometimes, but not always, implemented in RDF and OWL.

- They are typically not intended to support publishing.
  - These are systems for making scientific information interoperable and computationally available.

Evolving and exploiting these ontologies will be the foundation for the next revolution in scientific communication.

A speed bump in the road ahead:

- Many conceptual models are articulated in XML schemas.
- But XML schemas are sub-optimal for this task — XML is really a *data structure serialization* language, not a *modeling* language.
  - and therefore human intervention is still involved in interoperability and going to scale
- Remedying this problem will require layering a *formal semantics* on top of existing XML schemas.
- This can be down with RDF and OWL and will make XML schema-based domain models genuine ontologies with the highest possible level of interoperability and functionality.

What we still need to know

1) When leading users engage with the literature in circumstances that are optimal and exemplary

what exactly are they doing …or trying to do … or would try to do … if they could? … *what are they thinking*?

2) How can we develop fully formalized XML-based domain models in a way that scales, but still delivers functionality?

XML is a optimal for serializing a data structure
… not, necessarily, formalizing a model

XML-based information modeling is today where database modeling was in 1970 (Codd)

Concluding…

- Users want, need, will welcome, the tools that will support their increasingly fast-paced, indirect, and horizontal use of the literature.
- And it is now practical to provide them.
- With the convergence of driving forces the use of scientific articles will become even more innovative and indirect, including not only new integrative browsing, linking, analysis, and filtering tools but also new text mining and literature-based discovery applications.
- Users will soon be working with a number of articles at once, from many different publishers, and only in an extended sense would we characterize what is going on as “reading”.
- There will be new infrastructures and services to support these practices and the changes entailed will alter the strategic dynamics of STM publishing, as well as the professional lives of researchers..

How do we find out what we still need to know?

Projects on *user behavior* in the use of scientific literature being designed by

Carole L. Palmer, Allen H. Renear
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

Projects on *semantics for XML domain models* being designed by

Dave Dubin, Allen H. Renear
Electronic Publishing Research Group
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

Questions?