New Developments in Electronic Publishing
AMS/SMM Special Session, Houston, May 2004
ECM4 Satellite Conference, Stockholm, June 2004
p. 5-7.

The Meaning of Scientific Documents

Romeo Anghelache

Max-Planck-Institut für Gravitationsphysik
Albert-Einstein-Institut, Berlin, Germany
http://www.aei.mpg.de *http://romeo.roua.org*

Abstract. We enumerate here a couple of pressing issues, related to semantic authoring and preservation, in the context of digital creation, administration and usage of scientific documents; accordingly, some present and future solutions to these are sketched out.

1. Current status

Currently, the authors of scientific documents use a TEX editor or an alternative, proprietary software solution.

Meanwhile, some related standards have emerged, DocBook for structuring generic documents (books, articles), MathML and/or OpenMath for semantically clean authoring of mathematical expressions, OmDoc for rigorously maximizing a document's use by computing machines, MARCXML for representing and communicating bibliographical records and related metadata, then Unicode for unambiguous specification of international or domain specific symbols, and so on.

All of the above standards, except the latter, are using the XML specifications. XML itself being a step made from the SGML standard towards the generic user needs for reasons of simplicity and effectiveness.

And still, the scientist is in the same situation as 5–10 years ago, while authoring his articles or books: no clue as to how these standards can be of help to him, no effective, open source or otherwise, tools, to help him make use of them effortlessly. Why?

2. Stating the issues

Part of the answer is that neither the publishers, nor the librarians helped the author become aware of or concerned with the fate of their own written works. This awareness was not an urgent matter in the paper publishing era (the article will last as long as the paper sitting on a shelf), but in the digital document era it becomes a real issue: it is easy and cheap to create multiple versions or multiple copies of a digital document, so how can the author make sure that these versions are not being corrupt in the process or their rendering is not broken at a later time (when the reader accesses it), or that they are stored in a place where an indexing spider can find it?

The answer to this question is of a much higher priority than, say, digital access rights, unless one chooses to protect a corrupted representation of one's work.

The answer is bound to rely on the open standards noted in Section 1.

In comparison to these, proprietary formats and proprietary document authoring solutions do not guarantee an appropriate rendering (or meaning) in the future (be it near or far), unless they commit to a standard semantic vocabulary (or a set of them) which should be used by the author while editing his document.

Defining vocabularies with a meaning (that is, with a formally defined way to use them) is an exciting research topic today (the steps and standards needed to create ontologies in the digital era, are detailed by others [1]), but one cannot reasonably expect an author to suddenly jump from writing plain text or mathematical expressions directly to using ontology defined concepts, simply because the authoring process becomes tedious and would resemble more to computer programming; practically the author is still helpless in ensuring that his work will be reachable and useable after a period of time.

The ontologies are more helpful in extracting and managing the knowledge created by the authors and machines. We are, though, concerned here mainly with the knowledge creation process of which the human is the author.

The need for an effective authoring solution, positioned between being useful directly to the machines and being plain simple to humans to type, is becoming obvious [2]. A bias towards protecting the time of the human authors will be present at sketching a solution in the following sections.

3. What do I mean? To whom?

These are common questions in the author's mind: the meaning of his work is its capability of being used for a purpose (whether intended or not).

A handwritten article will have a meaning to an appropriately educated human; a computer typed text will have a meaning to some rendering, printing or indexing software (this is the lowest level semantic layer in a digital document) and a different meaning to the final human reader (presumably the highest level of semantics humans really care about); again, a scan of an old article will have a meaning for the graphical rendering software, another meaning for the character recognition software and a different meaning for the final human reader.

We note, even if it sounds to some as a trivial statement, that an article is, in all cases, meant to be found, read and used by a human being: it is, in short, a message.

The machines can help in the messaging process: index an article, act in a certain way while a specific expression is found (flag a misspelling, validate an expression or start an external process), advertise the presence of the article to the interested audience, check its consistency according to the available semantic rules, render it on different media, append a reader's comments to a section of it, store it in the appropriate digital library slot and relate its presence to the other neighbouring articles, keep a version history of it, assemble it with other documents according to an editor's, or library user's, request.

These functionalities depend on the availability of the semantic layers in the digital document. Some of these layers can be hinted by the author: the computer cannot even infer where a paragraph starts unless the author types some specific keys, it also cannot relate accurately concepts (the consequence of this is the inability of getting effectively useful search results) without the author's hints to a vocabulary of concepts.

4. Bounds for a semantic solution

The cardinality of this set of hints should stay minimal while maximizing the functional space to which the document can be made part of. The fuzzy constraint to this problem is the author's patience: he has always the alternative of creating a semantically flat document at the cost of his editors' time and his audience's time and size (a cost which is almost invisible at the time of authoring).

One can name the above requirement: user-friendliness of the authoring package.

But also, the author of a scientific article wants to communicate something and to preserve that message for future readers.

This requirement means: the authored document has to have a well defined structure. Well defined, in turn, means that the document should satisfy the following conditions, at the end of the authoring process:

1. be created in an open format which is platform neutral (XML),
2. contain enough information to locate it (administrative metadata: author, date etc.),
3. contain enough hints for a librarian to store, preserve and manage it (document structure definition for a validating procedure),
4. contain enough hints for a publisher to render it or relate it to other documents, (TEX like suggestions about how some symbol should look like, neighbouring conceptual domains),
5. contain enough hints for the reader to locate and use it (using consistently semantic vocabularies defined in open standards, e.g. MathML-content; and using keywords as often as, and wherever, necessary).

References

[1] Davies, J.; Fensel, D.; van Harmelen, F. (eds.): *Towards the semantic web*. Jon Wiley and Sons, 2003.

[2] Handschuh, S.; Staab, Steffen; Volz, Rafael: *On deep annotation*. http://www.aifb.uni-karlsruhe.de/WBS/sha/publication.html, 2003

Received August 5, 2004