XML & SGML for STM publishers

Toulouse November 6, 2002

Hubert PEDURAND, Marie-Elise FREON &/or Peter ROGERS

Jouve S.A.

11, bd de Sébastopol Paris 01

hpedurand@jouve.fr

mefreon@jouve.fr &/or progers@jouve.fr

Résumé – Pour publier encore plus vite et au moindre coût, les Editeurs en Sciences Techniques et Médecine ainsi que les bibliothèques ont été contraints de recourir au savoir faire de prestataires de services. Le présent communiqué explique 3 processus de publication (pour du contenu déjà publié et du nouveau contenu) utilisé aujourd'hui par plusieurs éditeurs afin d'avoir leurs contenus sur le web le plus rapidement possible et à bas prix. De la même manière que les grandes bases de données sont maintenant largement accessibles au domaine universitaire, il est dorénavant vitale pour tous les éditeurs quels que soient leur taille, de publier aussi sur le web avec des données hautement structurées. De plus, ce contenu hautement structuré est prêt pour de la diffusion multimédia (e-books, papier, bases de données, cédérom, etc.). Le communiqué présentera aussi le pour et le contre de chacun des processus, les différentes possibilités en sortie et leurs usages.

Abstract - To drive prices down and to publish faster, STM publishers and libraries have been forced to call on suppliers to help them out. This paper is designed to show 3 different publishing models (for already published content and new unpublished content) used by different publishers today to get their content onto the web in the shortest time possible and at low cost. As large databases on the web are now becoming structured and widely accessible to the wider Interlectual society, it is now vital that all publishers whatever the size, publish on the web with a high level of structured data. Moreover this highly structured content is ready for multi-media distribution (e-book devices, paper, database, CD-ROM etc.) The paper will also present the advantages and disadvantages of each model for the publisher and the supplier, the possible electronic outputs supplied and their uses.

1. Technological developments in STM publishing

1.1 Introduction

After years of trial programs with select publications, leading STM publishers are in the midst of rolling out online versions of all of their journals and preparing the way for books. Initially positioned as complements to their print counterparts, online journals are the foundation for a new generation of STM digital libraries, and a potential massive new source of revenue for their publishers. Behind the scenes we found a very different picture at many STM publishers. We found strong evidence of the use of SGML and XML, both for bibliographic headers and for full text. We also found tremendous acceptance of PDF, again, as a format for both archiving and delivering print journals.

1.2 SGML, HTML, XML, etc.? why?

1.2.1 SGML versus HTML ?

It is not easy to summarize SGML in a few words. The most famous and most used SGML application is HTML. It goes against the SGML philosophy that aims at separating the document content tags from the formatting tags. HTML includes a mix of formatting tags and content tags, i.e. structural tags (but in a limited number).

The danger in mixing formatting tags and structural tags becomes obvious when in need of changing the appearance of the text quickly. Let's say that a bibliography of 500 titles is tagged in HTML, and foreign words, monograph titles and periodical titles are in italics. You are required to change the style sheet according to the APA standards that specify that titles should be underlined. How can you make the changes without affecting foreign words? Get some coffee ready… You can not use the "Find & Replace" feature as there is no difference between a <i> tag and another <i> tag. A visual separation must be made between titles in italics and foreign words in italics. If the tags were <periodical title>, <foreign word> and <monograph title>, only one change would have been required in the external file that includes the DTD style sheet. SGML allows semantic searching.

HTML tags can not describe every class of e-documents available on the Web. With HTML, everything must fit in the same mold. In summary, there are three main benefits to the HTML language :

a) ***SIMPLICITY :*** HTML is easy to learn and understand.

b) ***LINKS :*** Hyperlinks can be created very easily, and there is only one way to link objects over the Internet: the <A ...> tag. The link source and target will need to be tagged.

c) ***PORTABILITY :*** Given the small amount of tags, it is very easy to integrate the HTML DTD into Netscape and Explorer browsers. No need to worry whether or not the other software will be able to read our HTML tags.

However, HTML shows the following weaknesses:

a) ***INTELLIGENCE :*** SGML allows us to transmit our "intelligence" of the text through specific tags. Since there is no concept of specific content in HTML, search engines experience significant problems and the recall rates are unacceptable with HTML.

b) ***ADAPTATION*** : It was not possible to display HTML tables before version 3 of the HTML language. The release of version 3 took quite some time due to the long debates of the HTML committee within the W3C. Version 3 was slowed down by "religion wars" within the W3C, where those in favor of structured documents were facing those in favor of simplicity. To display SGML tables, a solution developed externally (by military people) was included in a new DTD: that's all.

c) ***MAINTENANCE*** : HTML links are often broken. HTML was designed as if Web objects could not move. Problem is they can not stop moving! This would be like if we were placing our documents in a library by saying "This book will be in the third floor, second row, fourth shelf and in the fourteenth position from the right." What happens when the twelfth book is picked up? Error 404, document not found! A broaken link means potential lost revenue to the publisher.

This is just as primitive as that. It is therefore necessary to improve the HTML link system. Mixing formatting tags and structural tags is another maintenance problem that makes it hard and tedious to reuse the text.

SGML is the solution to the main HTML weaknesses. But this solution is to the detriment of the main HTML benefits. SGML is not easy, texts need to be validated, and hyperlinks are more complete but they use several methods of increased complexity. In addition, SGML is not as portable as HTML over the Internet. Additional software needs to be installed to view a SGML document and the DTD has to be sent with the document

Beyond the wonders that SGML can make, it has not reached critical mass among the millions of Internet users, twelve years after its recognition as an ISO standard. At the end of 1996, a workgroup of the W3C consortium was created to address HTML issues. The goal was to find a solution midway between the HTML simplicity and the SGML complexity, meaning some kind of a "light" SGML. The names that were suggested for this new language reveal the work spirit of these developers: MGML (Minimal Generalized Markup Language), SLIM (Structured Language for Internet Markup), MAGMA (Minimal Architecture for Generalized Markup Applications). In the end, the committee voted in favor of XML (eXtensible Markup Language) which is easier to market as an acronym.

XML 1.0 became an official W3C recommendation on February 12, 1998. XML combines the SGML strengths and the HTML simplicity. The SGML standard consists of 300 pages whereas the XML standard only has 32 pages.

1.3 Why is XML important?

W3C is a consortium of the main industries (130) involved in Web development. For a standard to become an official W3C recommendation, it has to go through every review step showing that it is a stable standard that developers can use to build. XML did not reach this status until February 1998. This explains why XML is not widely used on the Web. Among the main writers of XML, there was a Microsoft representative, a Netscape representative and a Text Encoding Initiative representative. XML is included in the development projects of the industry’s leading companies.

1.3.1 XML consists of 3 main parts:

* DTD. We reviewed the significance of the DTD concept in our previous note. A DTD can be used in XML, but it is not compulsory. When using DTD, the document will be called “valid”, i.e. it will use and comply with this DTD. When not using DTD, the XML document will need to be “well shaped” without any tag inconsistency. For instance, every attribute will need to be between quotation marks (HTML is more tolerant regarding quotation marks); items can not be empty; each X item will have to end with a </X> closing tag; items that are empty in HTML, such as an horizontal line, will need to be written as <HR/> or <HR></HR> in XML. In addition, the document will need to specify that it does not use any DTD with the beginning tag: <?XML version="1.0" standalone ="yes"?>. But don’t worry those who have designed their SGML DTD already can still use it (with a few modifications) for XML publishing, and yes, for all main publishers have a DTD for their online publications .

* XSL means Extensible Style Language. This language is used to define style sheets that are related to XML documents. The XSL file will define that an XML item needs to be displayed with a specific font, a specific color, etc. With XSL, the creator of the document can make such decisions for a better control of the document look. Reference can also be made to an existing public XSL file. XSL is based on 2 style sheet standards: Cascading Style Sheet, that is starting to be used with HTML files, and DSSSL (Document Style Semantics and Specification Language), that is a more complex style sheet standard. Once again, developers have chosen a midway solution as XSL uses both CSS and DSSSL. XSL is still being developed. Actually, XSL is only a W3C “Note”. This means that XSL still has to go through three additional steps, i.e. “Draft”, then “Suggested Recommendation”, and then “Official Recommendation”. (Note that the style sheet template process, that XSL uses, is quite advanced as the second version of CSS, i.e. CSS2, has become an official W3C recommendation, and DSSSL is an ISO standard).

* XLL means Extensible Link Language. It is the hyperlink description language in XML. XLL is the second part of the XML standard. XLL is at the stage of a W3C work document (July 1997) known as Extensible Markup Language (XML): Part 2, Linking. Once approved, XLL will address several issues related to broken hyperlinks that can be currently found on the Internet. XLL is based on standard ISO 10744, (Information Technologies: Hypermedia/event structuring language), better known as HyTime. XLL also includes functionalities available in the DTD of the TEI. Among other things, XLL will allow bi-directional links, links to Internet targets not previously tagged, links that can be managed in a file outside of the document instance, and also link attributes that can define the type of link (link to a definition, external link, etc.). In the end, such improvements should get rid of the “404 not found” message for good.

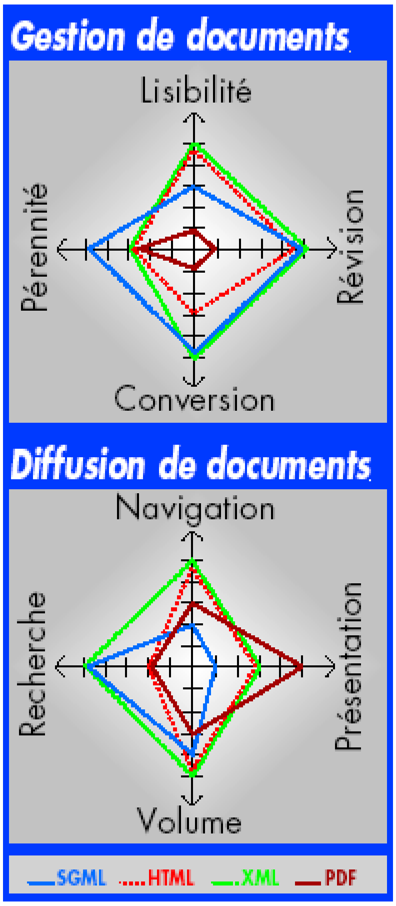
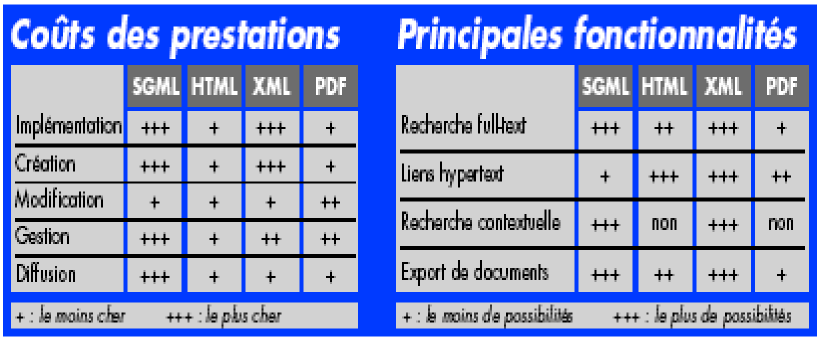


FIG .1 : FLUX COMPARISON (© JOUVE – 2000)



Coûts des prestations

	SGML	HTML	XML	PDF
Implémentation	+++	+	+++	+
Création	+++	+	+++	+
Modification	+	+	+	++
Gestion	+++	+	++	++
Diffusion	+++	+	+	+

+ : le moins cher   +++ : le plus cher

Principales fonctionnalités

	SGML	HTML	XML	PDF
Recherche full-text	+++	++	+++	+
Liens hypertext	+	+++	+++	++
Recherche contextuelle	+++	non	+++	non
Export de documents	+++	++	+++	+

+ : le moins de possibilités   +++ : le plus de possibilités

FIG .2 : COSTS & FUNCTIONALITIES (© JOUVE – 2000)

1.4 The STM market

	STM Publishers	*Number of Journal titles (Y2K)*
1	Elsevier (Netherlands)	*1500*
2	Kluwer Academic (Netherlands)	*507*
3	Blackwell Publishing (USA)	*308*
4	Springer (Germany)	*284*
5	John Wiley & Sons (USA)	*271*

TAB. 1 : TOP 5 STM PUBLISHERS RATED BY NUMBER OF JOURNALS (JOUVE – 2000)

Around 9000 STM Journals are available (ISI database). According to Bennett Zucker, vice president of Really Strategies, "Publishers in the STM market typically need to produce highly structured, complex packages in both traditional and new media. The STM publisher's early-adopter experience offers excellent guidelines for those in other segments."

The STM sector has become very concentrated: today, there are only 5 major players in the private STM publishing market. All the big private publishing houses are multinational, because investments in the sector are very large and extend over several years. An STM publisher has to wait 4 to 5 years for a journal to become profitable and books too, demand considerable investment. That’s why publishers are creating more and more international subsidiaries, in order to publish their books and journals in different languages and get a return on their investment more quickly.

2. PRISM

2.1 General purposes

The Publishing Requirements for Industry Standard Metadata (PRISM) specification defines an XML metadata vocabulary for syndicating, aggregating, post-processing and multi-purposing magazine, news, catalog, book, and mainstream journal content. PRISM provides a framework for the interchange and preservation of content and metadata, a collection of elements to describe that content, and a set of controlled vocabularies listing the values for those elements

The working group focused on metadata for:

- General-purpose description of resources as a whole
- Specification of a resource’s relationships to other resources
- Definition of intellectual property rights and permissions
- Expressing inline metadata (that is, markup within the resource itself).

Like the ICE protocol [ICE], PRISM is designed be straightforward to use over the Internet, support a wide variety of applications, not constrain data formats of the resources being described, conform to a specific XML syntax, and be constrained to practical and implementable mechanisms.

Relationship to Other Specifications:

2.2 XML

PRISM metadata documents are an application of XML [W3C-XML].Basic concepts in PRISM are represented using the element/attribute markup model of XML. The PRISM specification makes use of additional XML concepts, such as namespaces [W3C-XML-NS].

2.3 Resource Description Framework (RDF)

The Resource Description Framework [W3C-RDF] defines a model and XML syntax to represent and transport metadata. PRISM uses a simplified profile of RDF for its metadata framework. Thus, PRISM compliant applications will generate metadata that can be processed by RDF processing applications. However, the converse is not necessarily true. The behavior of applications processing input that does not conform to this specification is not defined.

2.4 Dublin Core (DC)

- The Dublin Core Metadata Initiative [DCMI] established a set of metadata to describe electronic resources in a manner similar to a library card catalog.
- The Dublin Core includes 15 general elements designed to characterize resources.
- PRISM uses the Dublin Core and its relation types as the foundation for its metadata.
- PRISM also recommends practices for using the Dublin Core vocabulary.

2.5 NewsML

- NewsML [IPTC-NEWSML] is a standard from the International Press Telecommunications Council (IPTC) aimed at the transmission of news stories and the automation of newswire services.
- PRISM focuses on describing content and how it may be reused. While there is some overlap between the two standards, PRISM and NewsML are largely complementary.
- PRISM's controlled vocabularies have been specified in such a way that they can be used in NewsML. The PRISM working group and the IPTC are working together to investigate a common format and metadata vocabulary to satisfy the needs of the members of both organizations.

2.6 Information and Content Exchange (ICE)

- The Information and Content Exchange protocol manages and automates syndication relationships, data transfer, and results analysis.
- PRISM complements ICE by providing an industry-standard vocabulary to automate content reuse and syndication processes. To quote from the ICE specification [ICE]:

  Reusing and redistributing information and content from one Web site to another is an ad hoc and expensive process.

The expense derives from two different types of problem:

- Before successfully sharing and reusing information, both ends need a common vocabulary.
- Before successfully transferring any data and managing the relationship, both ends need a common protocol and management model.

Successful content syndication requires solving both halves of this puzzle.

2.7 eXtensible Rights Markup Language (XrML)

XrML [XRML] is a specification developed by ContentGuard, Inc.It specifies the behavior of trusted digital rights management systems and repositories.Unlike XrML, PRISM assumes that the sender and receiver of a PRISM communication already have a business arrangement that is specified in a contract. PRISM's focus is on lowering the costs of complying with that agreement. Thus, it provides a standard means of expressing common terms and conditions. XrML takes on a much harder problem,controlling the behavior of end-user applications and devices such as printers and tape drives to prevent unauthorized reuse of the content.PRISM specifies as little as possible about the internal behavior of systems.Thus, PRISM's treatment of derivative use rights is complimentary to, but separate from, the rights and uses that are specified in XrML.

3. STIX

3.1 General purposes

After years of planning, a group of scientific publishers today formally announced the Scientific and Technical Information Exchange (STIX) font creation project and the launch of the STIX web site at http://www.stixfonts.org. The STIX publishers aim to develop a comprehensive set of fonts for mathematics and other special characters used in Scientific, Technical, and Medical publishing. The web site provides information for potential users within the scientific and publishing communities, and a special area for software developers who may want to incorporate support for the STIX Fonts into their products.

Six publishers -- the American Chemical Society (ACS), the American Institute of Physics (AIP), the American Mathematical Society (AMS), the American Physical Society (APS), Elsevier Science, and the Institute of Electrical and Electronic Engineers (IEEE) -- came together to design, fund and manage the STIX project. They have awarded the font development contract to a respected font development company, which has begun the process of designing and delivering nearly eight thousand characters/glyphs. The design submissions of the various character sets are currently being evaluated by a Technical Review Committee consisting of representatives of the six participating publishers.

There is currently a clear need for a new font set for mathematical and other scientific symbols, especially in the area of on-screen display in electronic publishing. Today, scientists must assemble scientific symbols and special characters from a variety of fonts, many of which may vary in character style, positioning, or size. The resulting documents typically have an unsatisfactory, jumbled appearance. Even more importantly, when posted to a web site, these documents may not be properly rendered unless the viewers of the document have all of the same specialized fonts available on the computer workstations they are using. This new set of fonts, known as the STIX Fonts, will solve both of these problems, serving the scientific and engineering community in the process from manuscript creation all the way through to final publication, both in electronic and print formats. It will unify support for all special symbols and alphabets into a single, comprehensive font set.

3.2 Why is it important that the STIX Fonts be compatible with Unicode™?

"Unicode™ is the universal character encoding scheme for written characters and text. It defines a consistent way of encoding multilingual text that enables the exchange of text data internationally and creates the foundation for global software." (*The Unicode Standard, Version 3.0,* The Unicode Consortium, Addison-Wesley: Reading, Massachusetts, 2000, p. 1.) It is also the native language of XML, which is very important for the use of the STIX Fonts in rendering scholarly scientific communications online. Web browsers support Unicode™ character representation, and therefore should be able to directly reference every STIX Fonts glyph, eventually.

3.3 Other technologies required to accurately represent scholarly scientific communications on the Web?

The other primary requirement is the need to be able to accurately render mathematical formulas, especially built up mathematical expressions that may involve complex fractions (such expressions are often called display equations within the publishing industry). One solution to such rendering is MathML, an XML application created by the World Wide Web Consortium. Web browsers must support MathML for this requirement to be fully solved. This support may be via a MathML plugin (like the Adobe Acrobat Reader plug-in for viewing PDF files) or via adding support for MathML to the browser, itself. The STI Pub companies expect to see MathML support from the major browsers (either natively, or via plug-ins) within the next 12 – 18 months.

Most publishers have provided documents in PDF form, which exactly replicates the printed page, but is difficult to enhance with internal and external linking. Some publishers have also provided documents in HTML, with all special characters, in-line formulas, and display equations rendered as graphic images. These HTML documents permit rich linking, although the quality of presentation suffers because the images are not re-sizable as the point size of the text is enlarged or reduced.

3.4 Will the STIX Fonts work on all computers and with all software applications?

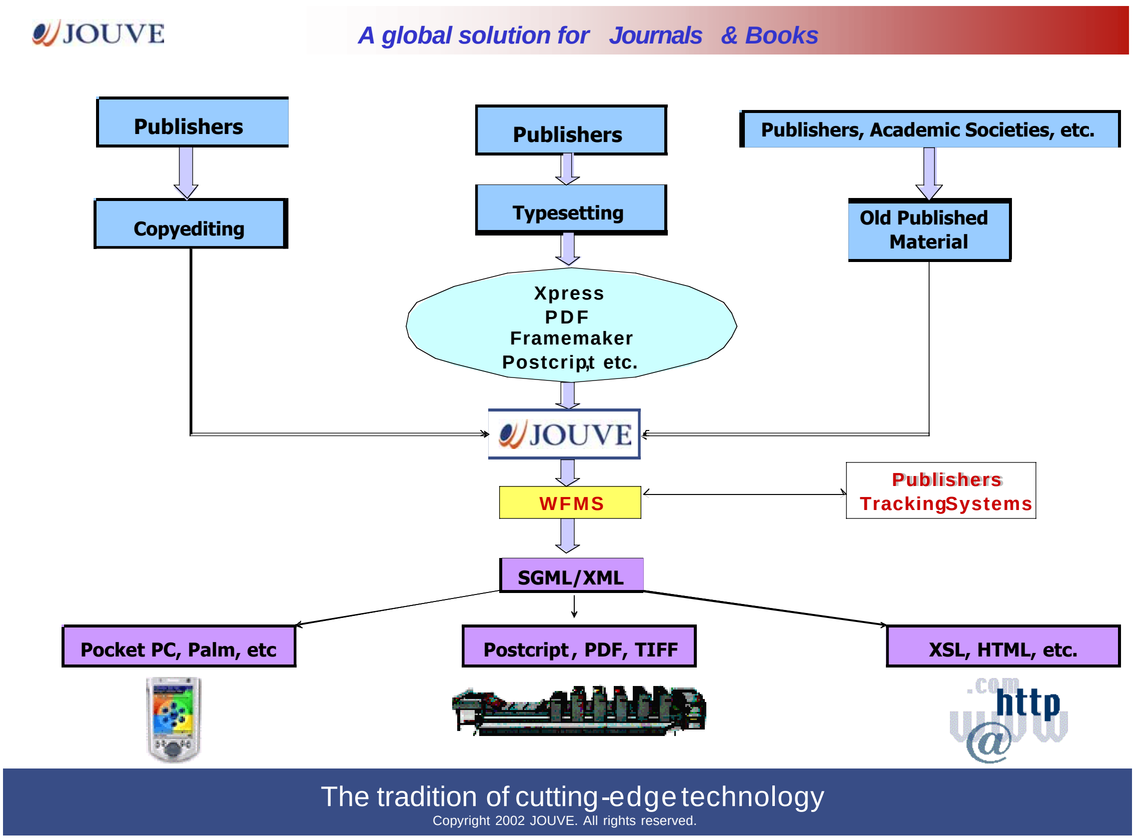
The STIX fonts will be made available, under royalty-free license, to anyone, including publishers, software developers, scientists, students and the general public. Target for completion of the project is the Fall of 2003. By making the fonts freely available, the STIX project hopes to encourage the development of applications that make use of these fonts. In particular the STIX project will create a TeX implementation that TeX users can install and configure with minimal effort. TeX is a computer language designed for typesetting, with particular application to mathematics and other technical material.

The goal is to make the applicability of the STIX Fonts as wide as possible. At a minimum, the STIX Fonts will work with computers running current versions of Windows, Macintosh, and most UNIX (including Linux) operating systems. They will work with most TeX applications, and we are working with a range of software developers to ensure that many applications involved with the scholarly scientific communication process will fully support the fonts. If an application has support for Type 1 (i.e., PostScript) or OpenType fonts and support for Unicode character representations, it can use most of the STIX characters/glyphs. If, in addition, it supports building or rendering complex mathematical expressions using the STIX Fonts, it can be said to fully support the STIX Fonts. As software developers make announcements regarding STIX Fonts support, we will report them on this web site. The STIX Fonts have been designed to provide the same functionality as the basic font used for viewing Web pages. Accordingly, the STIX Fonts should be set as the "web page font" on your browser. The browser will then know to map each Unicode character value in the document to the appropriate glyph in the STIX Fonts.

The STIX mission will be fully realized when:

- Fully hinted PostScript Type 1 and OpenType font sets have been created.
- All characters/glyphs have been incorporated into Unicode representation or comparable representation and browsers include program logic to fully utilize the STIX font set in the electronic representation of scholarly scientific documents.

4. One unique structure for a multi-diffusion



JOUVE

A global solution for Journals & Books

Publishers

Copyediting

Publishers

Typesetting

Xpress
PDF
Framemaker
Postcript etc.

Publishers, Academic Societies, etc.

Old Published Material

JOUVE

WFMS

Publishers TrackingSystems

SGML/XML

Pocket PC, Palm, etc

Postcript, PDF, TIFF

XSL, HTML, etc.

The tradition of cutting-edge technology

FIG. 3 : THREE DIFFERENTS PUBLISHING MODELS

❖ Front Matter

- ➢ The program recognizes elements by keywords, databases (city, country), syntax, links between elements.

❖ Body

- ➢ Program structures section and section title (numbered or not)
- ➢ Author highlighting is maintained
- ➢ Program doubts are signaled with tags <process instructions>

❖ References

- ➢ Comparison of each reference with several bibliographic databases (~ 10 000 000 references)
- ➢ 80 % structured correctly

    - Uncorrectly structured references are rekeying
- Tables and figures
    - Structuring of cross-ref, tables and figures
    - Structuring completion
    - Missing element, errors, tables presentation
- Quality control
    - Internal QCTools (complementary controls)
    - Clients QCTools
    - Quality control sheet (sent with the proof)

5. Conclusion

For the least advanced STM publishers, digital technology represents at least 10 to 20% of their activity; this figure rises to more than 50% for those publishers which are on the cutting edge of technology. Scientists were involved in launching the Web and began using Internet early on to exchange their work. This had an immediate impact on STM journals. All the big publishing houses now offer a dual approach, combining a paper version with on-line subscription. With the Internet, they have improved publication turnaround time and interactivity with the scientific community, and updating in particular has become much quicker. It remains that a structured language is the best way to keep control of the data

In term of permanence “If you are not XML you’re dead!”. At the moment, the debate is focused on the development of common standards: the publisher must be capable of delivering information on all media, without being dependent on the platform used by the reader. Just as for the cinema or video, the pressure created by consumer demands for simple tools will bring about the advent of single standards.

Références

- Online ressource for markup language technologies
http://www.oasis-open.org/cover/

- The xml industry portal
http://www.xml.org/xml/news_market.shtml

- Xml from the inside out
http://www.xml.com/

- W3C architecture domain
http://www.w3.org/XML/

- XML tutorial web site
http://www.w3schools.com/xml/default.asp

- Goldfarb's XML Handbook
www.xmlhandbook.com

For more information visit the STIX Fonts web site at http://www.stixfonts.org.

STIX Fonts -- Related Web Sites

- American Mathematical Society STIX project page http://www.ams.org/STIX

- Unicode™Standard web site

http://www.unicode.org/

- World Wide Web Consortium MathML Standard

http://www.w3c.org/Math

- Adobe Solutions Network OpenType Developer Program

http://partners.adobe.com/asn/developer/type/opentype.html

- Microsoft Typography web site

http://www.microsoft.com/typography/default.asp

- Adobe OpenType web site

http://www.adobe.com/type/opentype/main.html

Others:

http://www.elsevier.com

http://www.wolterskluwer.com

http://www.blackwell-science.com/

http://www.springer.de

http://www.wiley.com

http://www.jouve.fr

http://www.jouve.com

SUMMARY SESSION 3