Content Management and XML: the needed interaction

Ricardo Eíto Brun

Citación recomendada: Ricardo Eíto Brun. *Content Management and XML: the needed interaction* [en linea]. "Hipertext.net", num. 3, 2005. <http://www.hipertext.net> [Consulted: 12 feb. 2007]. .

1. A Brief Introduction to XML language

XML (eXtensible Markup Language) language began to be developed in September 1996 under the auspices of the W3C with a clear proposal: design a tags language optimised for Internet. XML should continue with HTML simplicity but with the expressive capacity of its previous version, SGML.

The edition of XML documents follows the following objectives:

- Distinguish the content and structure of the documents from their presentation in paper or on screen.
- Make its structure and informative contents explicit.
- Create documents that can be easily exchanged and processed in heterogeneous computing systems.

To reach these objectives, XML proposes a format where the tags in the text of the documents are intercalated, in order to distinguish the various structural parts or elements of the same. The main characteristics are the following:

- The possibility of descriptive marking. In XML the tags. have the function of differentiating the informative contents of the documents against their use in HTML, where the tags are used to indicate how the content should be visualised. XML does not specify a valid group of tags, but provides rules that allow us create new vocabulary or group of tags applicable for the coding of various types of documents.
- The distinction between document structure and presentation. In XML a clear difference is established between the document structure and presentation. The XML document tags do not indicate how the document should be presented. To indicate how a document should be presented on screen or on paper, a separate style sheet should be created, and then associate it with the document.

2. Continuous evolution of the language

XML has given rise to a large number of initiatives related with the exchange and coding of content and metadata and has become a widely accepted option to facilitate its management and recovery. The list of initiatives is extensive, and even though many of them have not achieved the same level of acceptance, we have numerous examples of the successful application of this language in the academic, business and institutional fields.

During 2004 we have seen the publication of new proposals and initiatives related to the use of the language. Specifically, the definite publication of the first version of the UBL vocabulary (Universal Business Language) for the exchange of commercial documents should be

highlighted, as well as the selection of the Danish government of this vocabulary for their electronic administration projects; the development of the DITA vocabulary (Darwin Information Typing Architecture) by OASIS for the codification and re-use of technical documentation; or the launching of new versions of computer programmes such as DB XML of Sleepycat Software, Astoria 4.3 or the purchase of Advent Publishing by Arbortext.

The working lines that receive the greatest attention during 2004 were the following:

A growing interest in the specifications oriented to the integration of computer applications by means of data exchange through the so-called "web sites. This interest has been materialised in the publication of new specifications facilitating the orchestration and coordination of interfaces between applications (BPEL4WS, BPML, etc.) or in the exchange of XML data in binary format.

More support by the international normalisation organisms to the specifications destined to achieve standardisation of the electronic business between companies. A proof of this is the publication, as ISO standard, of the ebXM [1] specifications, and the recent update of the UDDI standard UDDI (Universal Description and Discovery). These two standards have again awoken the interest which slowed down in previous years due to the decrease of investment in information technology and distrust in the markets related to Internet.

2004 concluded with the holding of the tenth anniversary of the W3C Consortium (World Wide Web Consortium) with a symposium held on the 1 st of December in Boston. The presence of sponsors in the event of competing technology companies should be highlighted, proving the widely accepted agreement regarding the standards published by W3C. Amongst the communications presented in this event, the XML impact and its function in the creation, management and distribution of information and content should be highlighted.

These contributions are only a sample of the protagonism reached by XML language since its creation 8 years ago, and of the recognition of its importance in the creation, management processes, and, above all, in the exchange of information and know-how.

But when we refer to *content management* we see that this term is not always linked to the use of XML language. What relationship can we establish between this practice and the format? What do we refer to when we talk about content management?

3. Definition of content management

In the drafting of a definition for content management we find the same difficulties when talking about document management or know-how management. This difficulty is mainly due to the fact that this term is being used in various contexts.

Currently, the term content management is put on the same level as the group of functions that are an integral part of this type of programme. Content management can also be compared with a wider planning focused on the global management of information resources of an institution, or company, by means of web technology (Internet e Intranet). This focus gives the facilitating role to technology, and the greatest weight is on the aspects related to the identification of internal and external information resources, their valuation, management and efficient treatment.

A third definition of content management comes from the area of management of electronic publications and documents. The use of this term refers to the application of a series of techniques and tools for the coding, storage and distribution of publications in digital format. It is in this scope where the use of the markup languages -initially SGML and then XML - has been constant due to their open character, independent from suppliers and specific hardware/software platforms.

3.1. Origin of document management

Document management arose as the reply to the difficulties experienced in the management and maintenance of web sites. The first web sites characterised by static content and made up of HTML pages and image files gave way to more sophisticated solutions where new technologies made the dynamic generation of content possible. The CGI programmes were initially used for this, and then the dynamic pages written in server script languages such as ASP (Active Server Pages), JSP (Java Server Pages) or the popular PHP, which allowed the extraction of information of relational databases.

The architecture of web sites became more and more complex with the adoption of these technologies. At the same time, a reply had to be given to the need to keeping the publications alive and facilitate access to information updated in real time by multidisciplinary teams made up by authors, editors, graphic designers and technical personnel. This problem encouraged some computer system suppliers to design applications facilitating the management of all the components or resources part of a web site. Amongst these components, the content itself was included (this being able to be material in various formats: text, video, audio, etc.) as well as dynamic pages, style sheets or script files implementing the visual and dynamic aspects of the site.

Content management has had considerable commercial success. Numerous organisations have adopted a computer application of this type. At all times the suppliers have given specific replies to the problems present in the daily practice of the organisations. To identify these problems, we should ask ourselves questions such as: are we sure that the users really access the latest content of texts that have been updated? Do we depend on the availability of a person - usually the administrator of the web site - to put at the users' disposal certain information? Does the publishing of new information following the guidelines and style notes adopted by an organisation require too much work and time? To what extent is it possible to have a follow-up of the status where the validation and approval of a certain article or text is possible? Can a user publish and edit content with ease and independence, without having to know HTML language details, style sheets, or the physical distribution of the web site in folders, subfolders, etc?

These difficulties could be summarised in the following points:

- Difficulties related to the need to have technical know-how to be able to publish information and contribute to the site content. This is translated into a dependence on the technical personnel to keep the site updated.

- Difficulty in manually maintaining the integrity of the information and ensure that only the approved content are available for the users, in accordance with an established access and security policy.

- Incapacity to programme the publication of content at certain times and during specific periods of time.

- Difficulty to coordinate the work of an interdisciplinary team made up by heterogeneous profiles (authors, editors, designers, programmers, etc.).

Not having an application and work procedures facilitating the management of a web site undoubtedly leads to negative consequences, amongst which the following should be highlighted:

- The web site administrator and technical personnel end up being overloaded with the increase in the number of requests and tasks related to the publication of content and site administration.

- The contents are not available when these are expected to be, delays occur affecting the validity and integrity of the information.

- The update of browsing tools is very difficult as pages are added and withdrawn from the site, as this is a process requiring the manual update of multiple pages.

- It is very difficult, if not impossible, to re-use contents (for example, publish the same contents for an external web site users, and for those of an Intranet with different staff would necessarily involve a duplication of tasks and contents.

The management systems of commercial contents - and the *open source* alternatives- attempt to provide a solution to these problems, and have defined the content of management discipline as we know it today. In this way, we can confirm that an organisation manages its contents correctly if it solves the problems previously mentioned in an orderly manner, and if it implements the tools and procedures necessary to solve to the previous problems.

3.2. Functional description of a content management system

The application of content management will integrate the necessary tools for the staff in charge of maintaining and managing a web site (usually the marketing, communication departments, etc) to be able to easily update the content of the site without needing all the HTML codification details or the physical location of the pages in the web server. The functions that these applications usually include are the following:

- Maintenance of the physical and logical structure of the site.

- Creation of new contents and edition of the existing contents by means of templates, usually by means of a web browser.

- Automatic maintenance of the site browsing and hyperlinks between pages.

- Approval, revision and validation of the content until these are made public on the web site.

- Content validity periods.

- Changes and revision control.

- Sharing of the content between various pages (connected pages).

3.2.1. Maintenance of the logical structure of the site

Two structures can be differentiated on a web site. On one hand, the physical structure of the HTML, PHP, etc., pages making it up. This structure is decided upon by the way in which the pages are physically stored in the web server (folders, subfolders, tables in databases, files, etc.). On the other hand, we have a logical structure that establishes the way in which the pages will be shown to the web site visitors (grouped in sections, subsections, etc.). A correspondence between the physical and logical structure of the site is usual, although this correspondence does not have to be one hundred per cent of the cases.

In a content management system, the logical structure of the site is managed by means of a hierarchy of "groups' or "sections'. In each section the pages including relevant information for this section are grouped. The content management application will be in charge of maintaining the physical structure in a transparent way for the users and administrators, for them "not to worry' about the details regarding the physical storage of the contents.

3.2.2. Content edition by means of templates

In order to avoid the complexity of HTML language to the users, the creation of new pages and

modification of exiting pages is done by means of the use of templates. The templates are HTML forms where the authors can type text, add images, hyperlinks, and any other element features to the web pages.

In a template there are editable and non-editable sections. The authors can only modify the content of editable sections, with which the integrity of the fixed elements of the page is guaranteed (heads, menus, images, etc.)

When a user creates content by means of a template, what it is being done is adding a text to a content repository, where the text is saved independently from the fixed parts of the text.

A content management tool should allow us the creation of various templates for the various types of pages that are going to be used at our site. Each template can have as many fields as necessary. The edition can be carried out from a standard browser to which visual utilities are added allowing the user to mark and format texts for them to be shown in bold, cursive, by means of lists, etc.

The concept of template allows guaranteeing the consistency in content presentation.

3.2.3. Automatic maintenance of the navigation structure

One of the most frequent problems in the management of a web site is the maintenance of the hyperlinks that determine the browsing of the site. Although this is not a problem in the web sites with few pages, the problem is bigger in sites with a high number of pages and that are frequently updated.

In a content management application the hyperlinks are generated by a series of queries in the page repository managed by the content manager, so the options available in the menus, indexes, etc., are always dynamically generated and reflect the real status of the site at each moment.

3.2.4. Approval, revision and validation of content

Another function that the content management applications usually offer are the content approval and validation cycles. The pages can have an associated approval cycle, in such a way that whenever an author creates a new page or modifies and existing one the approval of the changes by an authorised user will be requested.

In the event of pages and content subject to approval cycles, the changes are not visible on the web site until the editor approves them.

3.2.5. Page validity periods

Another problem that should be faced by those in charge of maintaining the content of a web site arises from the fact that certain content have a validity period of time, after which the withdrawal of these from the site is necessary.

Without a content management application, the maintenance of these sites is extremely problematic. The problem is bigger with the web sites with financial information, data on the prices of products, etc., in which the guarantee of the information being available at the correct moment - not before and not after- is a critical requirement.

With a content management application we will be able to indicate, for each individual page, group of pages or even for a section what its validity period is going to be, in such a way that these contents are only available at the web site during the indicated validity period.

In this section we have to take into account that the visibility should not only be applied to the page itself, but also to all the references and hyperlinks directing to the former.

3.2.6. Change and revision control

The change control allows maintaining the history of the contents the page. Thus, when a page is modified, instead of overwriting the contents and then losing it, the system creates a new version and will keep the previous state of the page and information.

3.2.7. Connected pages: sharing content between pages

On certain occasions, the need of sharing the same content in various pages of the same web site or of different web sites (for example, between an internet site and an intranet site, or between the users registered and those not registered in a public site). Without a content management system, the need to share content between various pages would oblige us to duplicate the same content in various files. This involves the maintenance problem and requires greater control, than if that content is to be modified, the changes have to be made on more than one page. It is also necessary to know exactly how many pages share the content, for the update of the site/s to be possible and consistent.

With a content management application we can share the content between pages easily, applying various templates.

3.3. Technical description of a content management system

The functions described in the previous sections require a technical infrastructure and the use of various types of technology. Content management uses the following types of technology: dynamic pages (PHP, JSP, ASP, etc.); databases and repositories; work flow and messaging flow management; and integration between applications.

3.3.1. Dynamic pages

The dynamic pages -also called "server scripts'- make up the basis of the current web sites. These are files, similar to HTML pages, in between - together with labels and tags of HTML language- orders or commands written in a programming language. When the web server receives a request for a dynamic page, instead of directly sending it to the client browser from which the request is coming - as it happens in the case of HTML pages- the orders or commands are interpreted or executed to then send the web browser the result obtained from the execution, which will be HTML code.

3.3.2. Databases and repositories

A content management application needs storing a great deal of information. We have to add to the texts and contents themselves, the metadata and properties of each one of its pages or sections. Within these, we include the administration metadata (relative to page validity, its creator, approval status, etc.) and those necessary to facilitate its later recovery (key words, classification codes, etc.). Apart from this the content management application should be acquainted with the structure of the sites managed, the sections within each site, the pages included in each section, the approval cycles associated to each page, the possible status, and the information regarding which users are authorised and responsible for validating each type of contents.

Therefore, an information content repository is required, with which the various system users interact (authors, editors, programmers, designers and consumers and users of the information). This repository can take various shapes: a relational database, a file system, a XML data repository or a combination of these.

3.3.3. Metadata

By metadata we understand the information associated to the various objects forming part of the system content manager system repository, facilitating administration management and recovery. In the previous section we have mentioned various examples of metadata applicable to a page, section to any other type of resource level.

Amongst the functions that the metadata have, the possibility for personalising the site is highlighted: the comparison of the metadata associated with a page with the preferences established for the users registered make the personalisation and selective distribution of the information possible, adapting the content shown on the site or that are distributed by means of notifications by e-mail to the specific needs and interests of each user.

3.3.4. Workflow management

The workflow management allows organising sequences of activities surrounding one or more pages. This element has a functional part (being one of the features desirable in any content management application), but also a technical component.

The use of a content management application of the workflow automation does not require the complexity required by other environments. Nevertheless, it is a key part to ensure that all the information has been duly revised and validated before being made public.

The subsystem of workflow management can include some of the most advanced characteristics:

- Definition of various revision levels for each type of content.
- Definition of the various revision levels for each type of content.
- Integration with an electronic mail system.
- Definition of the task execution order to be completed for a specific content.
- Definition of the periods available to complete each one of the revisions.
- Application of the various levels in task assignment, distinguishing between sequential flows (where all the users should complete task after task of those assigned) and parallel (where a task is simultaneously assigned to various users, in such a way that this is completed when one of them carries it out or part of them carry it out).
- Information on the approval status of each element managed by the system.

3.3.5. Content and application integration

As the Web grows it is very complex to maintain and internally generate contents attractive to the users. This has made the organisations search for external information suppliers that can provide contents in the shape of articles, summaries, references to external sites, etc., in a format allowing its integration in the web site. This activity has generated a business: the syndication or addition of contents.

The addition processes have an important added value for the companies using these services, as there is professional filtering of the growing data and content volume and distribution of only those contents relevant to the customer company.

The syndication requires the use of two components: a) formats normalised for the transfer of

information via the net and b) a mechanisms allowing the customer company to download the contents added that are of its interest in an unattended way, at a pre-established periods of time.

The first component closely related with the use of XML language and various vocabularies that have been designed for this purpose, RSS (Resource Site Summary) being the most popular with its various versions and ATOM.

The second component is related to the web services technology, to which we will refer later on, and that allow us to access functions available in remote web servers by means of a URL (Uniform Resource Locator) which will return a XML document with the data resulting from the execution of said function on remote equipment.

4. The function of XML in content management

Once we have defined the functions characterising a content management application, two questions should be answered: what function does XML language have in this type of system? And what are the advantages we can have with its application?

4.1. XML as basis for content storage

The content managed and published by means of a web site can be stored in various ways, although to guarantee its later re-use and recovery we should consider the advantages offered by XML as storage language compared to other alternatives such as HTML.

The importance over the past few years of the so-called native XML databases should be mentioned. With this term we refer to the databases that store and manage a collection of XML documents without carrying out any type of previous transformation. In this model, the XML model is the main information storage unit.

Amongst the main exponents of the native XML databases we can highlight commercial systems such as Tamino, from the German company Software AG, Textil, or the open source system DBXML, that can be obtained free of charge.

To facilitate the edition of content in XML, the main suppliers of edition tools have published utilities that allow interaction with these and the data repository of the content management application. Examples of this integration can be found in the proposals from the Altova, Blast Radius, XYEnterprise or Stylus companies.

4.1.1. XML as a model for metadata representation

Some of the metadata systems that have been published over the past few years have opted for XML language as the main mechanisms for the representation and coding of the same.

If we opt to use XML for the coding of metadata we should consider the need for the availability of an indexing and recovery system allowing the discrimination of documents from the content of specific elements or attributes. That is to say, the search system should not only allow the search of a full text (that is, be able to recover the document if this has a combination of specific words), but also if said words appear within a specific element or in any of the descending elements.

4.1.2. XML as exchange and content integration means

XML is not a format to code texts and documents, but of a family of specifications establishing the way in which said texts can be processed and presented. Specifications such as XSLT, DOM or Xpath make the processing of XML documents possible based on various vocabularies by

means of various programming languages (Visual Basic, Java, etc.), using as common, standard and clearly documented model.

The possibility of obtaining XML documents by means of the net and process them with ease for any purpose (for example, to integrate them in a repository or database, or to visualise them as part of our web site) offer extreme flexibility and open the doors to any type of integration.

4.1.3. Conclusions

In the previous sections the main characteristics of content management have been defined, and we have seen how this term is closely linked to a type of computer applications that have arisen over the past few years as a solution to the problems derived from web site management.

The main areas of application of the document management application have been described, and we could find various scenarios of their use; depending on the scenario, the importance given to one or the other will be greater or smaller.

In this way, if we think of applications oriented to electronic commerce between companies, where the exchange of structured commercial documents is key, the greatest importance will be in the use of XML as exchange format. Nevertheless, if we think of applications oriented to information publishing, the application of the language as information storage means and as format for the creation and edition of documents will be more relevant.

In any event, the power of the language is evident in the various scenarios, as proven the adoption of the same carried out by the computer system suppliers who, in the long turn, are those who have provided some meaning to this discipline.

5. Bibliography

Boiko, Bob (2001) The Content Management Bible. New York: Wiley. ISBN 076454862X
Bussler, Christoph (2003). B2B Integration: concepts and architecture. Berlin [etc.]: Springer. ISBN 3-540-43487-9.
Donovan, Truly (1997) Industrial-Strength SGML: an Introduction to Enterprise Publishing. New Jersey: Prentice Hall PTR. ISBN 0-13-216243-1.
English, Bill. (2003) Microsoft Content Management Server 2002: A Complete Guide. Boston [etc.]: Addison-Wesley. ISBN 0321194446.
Hammersley, Ben (2003). Content Syndication with RSS. Beijing [etc.]: O'Reilly. ISBN 0596003838.
Marcos, Mari Carmen; Baiget, Tomàs (2003). "Integración y personalización de contenidos: Factiva". En: *El Profesional de la Información* , vol. 12, no. 1, p. 26-34.
Meloni, Julie (2004). Plone Content Management Essentials. Indiana: SAMS. ISBN 1-8000-382-3419.
Nakano, Russell. (2002) Web Content Management: a Collaborative Approach. Boston [etc.]: Addison-Wesley. ISBN 0-201-65782-1.
Rockley, Ann. (2002) Managing Enterprise Content: A Unified Content Strategy. New York: New Ridder Press. ISBN 0735713065.
XML Data Management (2003). Chaudhri, Akmal B.; Rashid, Awais1; Zicari, Roberto (eds.) Boston [etc.] : Addison-Wesley. ISBN 0-201-84452-4.

6. Notas

[1] Corresponds to the family of standards ISO 15000 Electronic Business eXtensible Markup Language, that attempt to promote the advantages of the Internet electronic commerce

between companies of any sector and size. These standards establish a framework for process definition, commercial documents, company register, etc. In the case of UDDI, a web company and organisations directory model based on the exchange of XML information is established. [volver]