2006

Automated content-management systems

Rowe, Neil C.

Monterey, California. Naval Postgraduate School

Encyclopedia of Digital Government, ed. A.-V. Anttiroiko & M. Malkia, Hershey, PA, USA: The Idea Group, 2006
http://hdl.handle.net/10945/35996

Automated content-management systems

Neil C. Rowe
Cebrowski Institute
U.S. Naval Postgraduate School
Monterey, CA 93943 USA

INTRODUCTION

The World Wide Web quickly evolved as a valuable resource for organizations to provide information and services to users. Much initial development of Web pages was done haphazardly. This resulted in many information gaps and inconsistencies between pages. Departments with more available time created more and better-designed Web pages even when they were no more important. Personnel who created Web pages would move to other jobs and their pages would become obsolete, but no one would bother to fix them. Two copies of the same information on the Web would become inconsistent when only one was updated, leaving the public wondering which was correct. Solutions were needed. We survey here the principal solution methods that have been developed.

This is a chapter in the *Encyclopedia of Digital Government*, ed. A.-V. Anttiroiko & M. Malkia, Hershey, PA, USA: The Idea Group, 2006.

BACKGROUND

"Content management" has recently become a popular term encompassing ways to manage Web pages, online databases, and print documents more consistently (Boiko, 2002; Hackos, 2002). "Content" means an organization's information assets. Since Web pages have become the primary means for organizations to publish information today, the primary focus of content management is on Web pages (Goodwin & Vidgen, 2002; Proctor et al, 2003). Content management is "Web page bureaucracy", imposing a set of policies and rules for creating pages, implementing them, updating them, and reusing their content for new purposes. Bureaucracy is not necessarily bad, since no one wants an organization (especially a government one) that is inconsistent or incompetent. Governments are required by law to provide certain services, and a bureaucracy of Web pages can assure that Web services are delivered properly and fairly. So although content management is not unique to digital government, it is an especially important and essential technology for digital government. But content management, like any bureaucratic innovation, does stifle some creativity, impose additional restrictions, and add time to create and use pages.

A variety of commercial products are available for content management, ranging from standalone applications for Web-page authoring to comprehensive systems that control every aspect of an organization's Web pages. The term ?content-management software? can refer to any of these. Costs range from free (for open-source software) to millions of dollars, and systems are rarely compatible with one another. So an organization must do a careful study before embarking on content management. Useful case studies of development of systems are available (Lerner, 2000; Weitzman et al, 2002; Kunkelmann & Brunelli, 2002; Dudek, 2003).

TASKS OF A CONTENT-MANAGEMENT SYSTEM

Typically content management is divided into collection, management, and publication (Boiko, 2002):

- Collection facilities obtain information ("content") with such things as Web authoring tools, specialized word-processing software, media managers and editors, and format conversion.
- Management facilities control the approval mechanisms and information flow of content. Most systems store content for pages in a database or "repository" along with metadata describing the form of the content. Management facilities ensure that checking and approval is done by specified people before content is made public, and they can also test content errors and track different versions of content.
- Publication facilities convert content into polished public presentations in Web pages, print documents, or various forms of media. They provide templates for selecting information and providing a consistent appearance. Publication management also includes efficient management of Web sites.

We now discuss in more detail the tasks of a content-management system (see Figure 1). Collection facilities comprise authoring, conversion, and editing; management facilities comprise workflow control and the content repository; and publication facilities comprise publication management, publication templates, electronic publications, and print publications.

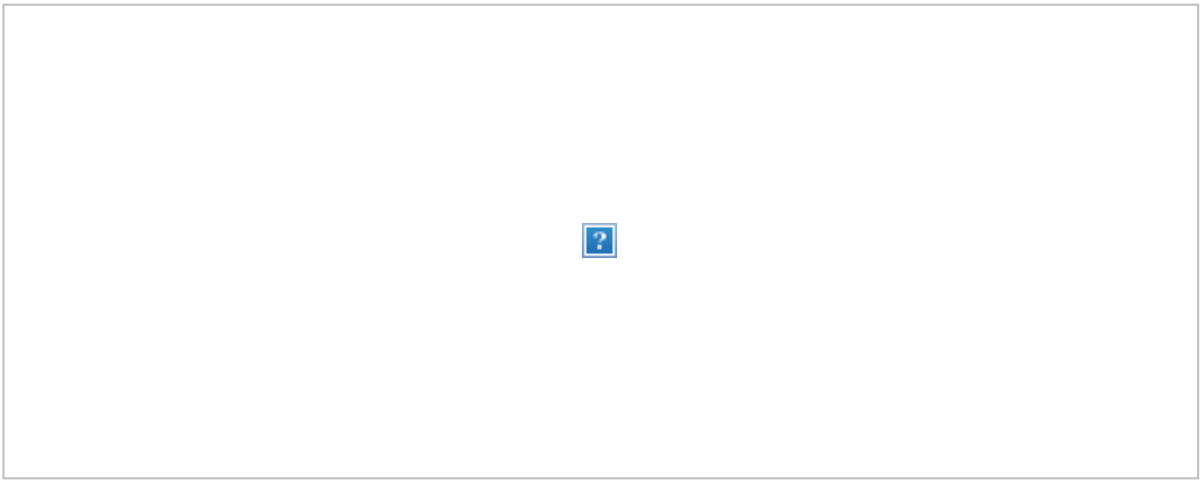

Figure 1: Outline of the content management process.

Authoring

General-purpose text editors can create content for content-management systems, but editors specifically designed for Web authoring like Microsoft Front Page and Macromedia Dreamweaver are often better, and the more-structured editors accompanying comprehensive content-management systems are even better. An organization can mandate starting templates for its Web pages with such tools, into which the content must be fitted and apportioned. A template can specify the types of information allowed and/or required on a page, general information about the page, and its layout.

Templates require "metadata" information about each chunk of content to manage it properly. This can include (among other things):

- author;
- who needs to approve it;
- the software that created it;

- when it was created;
- when it was last revised;
- when it becomes effective;
- when it becomes obsolete;
- when update-reminder messages should be sent;
- to whom update-reminder messages should be sent;
- fonts needed to display it (if they matter);
- how the text should be aligned and justified (if it matters);
- links needed to other documents;
- keywords that help describe it; and
- classification of the content type.

The authoring tool must obtain this information, but it should not need to ask the author for most of it if the tool is designed properly; otherwise, metadata requirements can quickly develop into a serious point of contention between authors and their organization. Author name, software, and special formatting information can be obtained from defaults set when a user first uses the authoring tool. Creation and revision dates can be obtained from the operating system. Effective and obsoleting dates can default to specified durations or times after the revision date (so for instance, class schedules at a university become effective at the beginning of each quarter and obsolete at the end). Formatting can be specific to the type of content selected by the user before starting. Keywords and content classifications can be obtained from authors via menus, but it can still be burden for them, as it has required many hours by librarians for print publications over the years. It helps to have different keyword menus for different types of content, and to use defaults for types where possible. For instance, all content from the purchasing department can have keywords "purchasing", "acquisition", and "contracts". Keywords can also be guessed from page titles and abstracts, and classification can be estimated by text-analysis methods (Varlamis et al, 2004), but this is less accurate.

Conversion

Much important content of organizations comes from sources other than Web pages. So a content-management system needs tools to convert a variety of documents to the format of the system. This includes such things as converting image files from GIF format to JPEG format, and documents from Word format to PDF format. Audio and video often require conversion since several incompatible formats are currently competing with one another. Conversion also includes formatted editing such as stripping blank lines or rearranging the columns of a table from a text-formatted database. When reusing information from other sources, copyright and usage restrictions may apply, so rights management software (Fetscherin & Schmid, 2003) may be necessary to track this, but this is not common with government content.

Electronic publications can also automatically acquire content from across the Internet. This can be done by specialized programs called "aggregators" and "bots" (Heaton, 2002) but they require programming. For example, an organization's Web page can be programmed to automatically show the latest weather report, news headlines, and boss's pronouncement as acquired from other pages.

XML (Extensible Markup Language) is essential today for organizing chunks of content, and most content-management software uses it. It is a generalization of the Web language HTML that allows for structuring and labeling of arbitrary data. So acquisition of content usually entails a conversion into XML (Surjanto, Ritter, & Loeser, 2000).

Editing

After content has been created or acquired, an organization normally sends it to someone for checking of style, appearance, coverage, and consistency with organization policy. Style is traditionally checked by human "copy editors", and involves examination of spelling, grammar, usage, rhetoric, and consistency. Content may also need to be edited to conform to organization policy in matters such as length, technical detail, use of color images, and accessibility to disabled users (W3C, 1999). Editing may also address metadata, since while a good authoring tool can fill in some metadata, authors may not be consistent in such metadata as keywords.

Editing electronic content may also involve segmentation and/or aggregation of content into user-friendly pieces, an issue more important than with traditional print documents. The editor may get many small documents (especially if they are generated automatically) and need to combine them into a larger one, as in collecting information for a phone directory. But also an editor may

have a document that is too large and need to partition it into subdocuments. Partitioning is essential in formatting for small handheld devices like personal digital assistants (PDAs), but can be valuable with any Web pages since users get lost easily in large documents when they lack the tactile feel of turning pages.

Workflow Control

Most automated content management manages the chain of approval necessary to publish. If done right, automation of approval management can significantly increase the productivity of a government organization since often much time is wasted in obtaining approvals. It can also eliminate many tedious manual transfers of files and possible errors in doing so. One needs to define, for each category of document or content, the sequence of people who process it (with substitute people when they are absent), what kind of processing they provide, and what time constraints must be fulfilled ("workflow control"). For instance, instructions for completing a new government form could be drawn up by a technical writer, sent to their supervisor for approval, sent to a copy editor for style and language improvements, sent to a Webmaster who controls the Web site, posted on the test site for debugging, and then moved ("migrated") after a time to the public Web server. As examples of time constraints, a department could specify posting of the current version of its forms catalog every three months, and a review of press releases by on-staff lawyers could be bypassed if they do not complete it within two days.

Workflow control also must include procedures for handling mistakes in publications since they can be costly. So it must be possible to return a Web site or electronic publication to a previous version, what is called "versioning control" (Nguyen, Munson, & Thao, 2004). Mistakes can be reduced by posting content to a staging or test server first (not necessarily on a separate machine from the public server); this is important for testing when content is dynamic or requires special software.

The Content Repository

Most content-management systems use a database, the "repository", to hold the pieces of content after authoring, conversion, editing, and approvals. This can range from a basic file system to a full database system. In many cases this should be a single centralized database for simplicity, but there are advantages in flexibility and robustness to distributing the information over several sites (Luo, Yang, & Tseng, 2002; Cranor et al, 2003); peer-to-peer methods can even be used to distribute content (Hausheer & Stiller, 2003). Methods for object-oriented databases can be helpful because content usually represents a variety of object types. Media files will need to be stored separately with pointers to them since their sizes can vary so much. The database will also need to support versioning and archiving.

Another benefit of automated content management is the ability to systematically check for inconsistencies like links that become invalid or pages that are moved. So deletion or moving of content in the repository should trigger changes to all content pointing to it. Broken external links can be found automatically by systematic periodic checking.

Publication Management

Finally, content management creates publications automatically or semi-automatically from the content in the repository including its metadata. Templates can be defined to assure consistent presentation style, including how the information is partitioned into pages and cross-linked. Metadata labeling of content permits creation of a diverse set of documents from the same information, including both print and electronic publications. It can also permit personalization of publications to the user using "cookies" and other forms of user tracking. For instance, the system can remember the type of display device the user had last time, the language they preferred, and the topics in which they were interested (Huang & Tilley, 2001).

The structure of publications is critical to usability so content management needs to address this, using feedback from potential users. This includes the partitions into subdocuments, titles and headings, links between documents (hyperlinks in electronic documents), navigation aids to help users locate themselves in a set of documents, and use of dynamic content. Electronic documents also greatly benefit from good indexes to the content, and keyword-based access using the indexes; although search engines such as Google index sites, an index specific to a site helps users find things faster.

Content management also includes management of the hardware and software that supplies content to users. Caching of frequently-requested information helps response time (Candan et al, 2001). Distributing popular content redundantly across different sites can

reduce overloading problems and increase reliability. Management can include automated copyright management. It also can include methods for presenting multimedia data such as choosing the size of an adequate window for playing video over a limited-bandwidth Web connection.

ACCESS CONTROL FOR SECRECY

All governments have secrets they must protect. Examples include secrets to preserve the privacy of citizens (like tax information), secrets to preserve fairness (like early economic data), secrets to facilitate negotiation (like diplomatic secrets), and secrets to protect public safety (like military secrets). Secrecy requires more powerful content control than that of a content-management system, to control who has access to the content and how it can be used. Secret information must be segregated on separate computers and networks where it cannot be accessed or transferred without tight restrictions. It entails many other security measures, like passwords, encryption, access-control lists, monitoring of systems by intrusion-detection software, monitoring for inadvertent electronic emanations, and control of the physical devices on which secrets reside. Declassification or revealing of secrets is needed on occasion but must be subject to stringent safeguards.

Military and diplomatic organizations have elaborate systems for handling such "classified" information (Landwehr, Heitmeyer, & McLean, 1984). The U.S. military is representative, with four basic levels of Unclassified, Confidential, Secret, and Top Secret. In addition, the levels can have "compartments" which designate a subtopic to which the classification applies, such as a compartment for nuclear secrets. An individual must have a "clearance" for both the level and compartment to access the information. All information must be used in appropriate designated facilities and have its classification printed on it.

FUTURE TRENDS

Content management is becoming increasingly common to all Web sites of organizations. It will soon be unthinkable for organizations to develop Web sites without it, and governments especially need content management to provide reliable and consistent services to their citizens. The current large number of products and vendors will probably decrease as a few products become popular. Few major innovations in the facilities of content-management systems are likely with the exception of multimedia support, but the systems will become increasingly flexible and increasingly well human-engineered as vendors gain experience with users.

CONCLUSION

Everyone complains about bureaucracy in government, but some bureaucracy is necessary to provide systematic, consistent, and complete service to citizens. Content-management systems extend bureaucracy to an organization's publications by deriving them systematically from stored electronic data that has been carefully checked through a series of approval steps and then presented in a standardized way. As such, they can increase the speed and reliability of government procedures and reduce workloads by permitting reuse of content for many purposes. But like all bureaucracy, content management risks getting out of control and impeding operations if not managed properly since its complexity can encourage growth of a technocratic elite unresponsive to citizens. To avoid this, it is essential to keep content-management systems as simple as possible and to manage them responsibly for the benefit of citizens.

REFERENCES

Boiko, B. (2002). *Content management bible*. New York: Hungry Minds.

Candan, K., Li, W.-S., Luo, O., Hsiung, W.-P., & Agrawal, D. (2001, May). Enabling dynamic content caching for database-driven web sites. *Proceedings of ACM SIGMOD International Conference on Management of Data (SIGMOD Record, 30* (2)), Santa Barbara, CA, 232-243.

Cranor, C., Ethington, R., Sehgal, A., Shur, D., Sreenan, C., & Van der Merwe, E. (2003, June). Content management: Design and implementation of a distributed content management system. *Proceedings of 13th International Workshop on Network and Operating Systems Support for Digital Audio and Video*, Monterey, CA, 4-11.

Dudek, D., & Wieczorek, H. (2003, September). A simple web content management tool as the solution to a web site redesign. *Proceedings of 31st ACM SIGUCCS Conference on User Services*, San Antonio, TX, 79-81.

Fetscherin, M., & Schmid, M. (2003, September). Comparing the usage of digital rights management systems in the music, film, and print industry. *Proceedings of Fifth International Conference on Electronic Commerce*, Pittsburgh, PA, 316-325.

Goodwin, S., & Vidgen, R. (2002, April). Content, content, everywhere ... time to stop and think? *Computing and Control Engineering Journal, 13* (2), 66-70.

Hackos, J. (2002). *Content management for dynamic Web delivery*. New York: Wiley, 2002.

Hausheer, D., & Stiller, B. (2003, September). Design of a distributed P2P-based content management middleware. *Proceedings of 29th Euromicro Conference*, 173-180.

Heaton, J. (2002). *Programming spiders, bots, and aggregators in Java*. San Francisco, CA: Cybex.

Huang, S., & Tilley, S. (2001). Issues of content and structure for a multilingual web site. *Proceedings of 19th International Conference on Computer Documentation*, Santa Fe, NM, 103-110.

Kunkelmann, T., & Brunelli, R. (2002, September). Advanced indexing and retrieval in modern content-management systems. *Proceedings of 28th Euromicro Conference*, 130-137.

Landwehr, C., Heitmeyer, C., & McLean, J. (1984, August). A security model for military message systems. *ACM Transactions on Computer Systems, 2* (3), 198-222.

Lerner, R. (2000, September). At the forge: content management. *Linux Journal, 77es*, 14.

Luo, M.-Y., Yang, C.-S., & Tseng, C.-W. (2002, March). Web and e-business application: Content management on server farm with layer-7 routing. *Proceedings of ACM Symposium on Applied Computing*, Madrid, Spain, 1134-1139.

Nguyen, T., Munson, E., & Thao, C. (2004, May). Versioning and fragmentation: Fine-grained, structured configuration management for web projects. *Proceedings of 13th Conference on World Wide Web*, New York, 433-442.

Proctor, R., Kim-Phuong, L., Najjar, L., Vaughan, M., & Salvendy, G. (2003, December). Virtual extension: content preparation and management for e-commerce Web sites. *Communications of the ACM, 46*(12), 289-299.

Surjanto, B., Ritter, N., & Loeser, H. (2000, June). XML content management based on object-relational database technology. *Proceedings of International Conference on Web Information Systems Engineering*, 1: 70-79.

W3C, (1999). Web content accessibility guide. Retrieved from July 15, 2004, from http://www.w3.org/TR/WAI-WEBCONTENT/.

Varlamis, I., Vazirgiannis, M., & Halkidi, M. (2004, June). THESUS, a closer view on Web content management enhanced with link semantics. *IEEE Transactions on Knowledge and Data Engineering, 16* (6), 685-700.

Weitzman, L., Dean. S., Meliksetian, D., Gupta, K., Zhou, N., & Wu, J. (2002, April). Transforming the content management process at IBM.com. *Proceedings of Conference on Human Factors in Computing Systems*, Case Studies of the CHI 2002 / AIGA Experience Design Forum, Minneapolis, MN, 1-15.

Terms

aggregator: Software for automatically creating documents by collecting it from a set of designated Internet sites.

classified information: Secrets (like military capabilities) that a government protects by special access controls (like identification cards and passwords).

content: Any information formatted primarily for display to humans, as opposed to internal data of a computer.

content management: Policy and software for managing the electronic publications of an organization.

metadata: Any data describing other data, such as size and type information for a media file; essential to management of databases and other information systems.

migration: Copying content from one site to another, as from a test server to a public server when it has been approved for release.

repository: In content management, the database holding the tagged content from which publications are constructed.

versioning: Software to keep track of versions of publications, different either in date or audience, so that when changes are made to parts of the content those changes will appear simultaneously in all the versions with those parts.

workflow control: Management of who sees and approves content, and in what order, before it can be published.

XML: Extensible Markup Language, a general language for structuring information on the Internet for use with the HTTP protocol, an extension of HTML; currently the most important language for flexibly sharing information between computer systems.