Automated Content-Management Systems

Neil C. Rowe
U.S. Naval Postgraduate School, 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.

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, Kim-Phuong, Najjar, Vaughan, & Salvendy, 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 (Dudek & Wieczorek, 2003; Kunkelmann & Brunelli, 2002; Lerner, 2000; Weitzman et al., 2002).

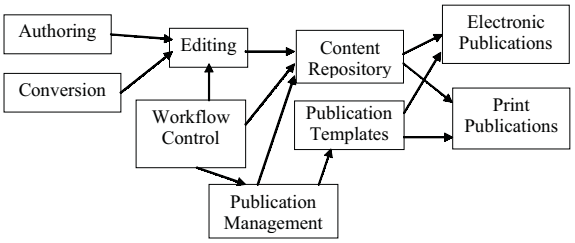
TASKS OF A CONTENT-MANAGEMENT SYSTEM

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

Figure 1. Outline of the content management process



repository; and publication facilities comprise publication management, publication templates, electronic publications, and print publications.

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
- 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, Vazirgiannis, & Halkidi, 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 usu-