Automating the internal workflow of the Miguel de Cervantes Digital Library

Alejandro Bia
E-mail: abia@dlsi.ua.es

Javier Pozo
E-mail: jpu@alu.ua.es

Biblioteca Virtual Miguel de Cervantes Saavedra
Universidad de Alicante, Alicante E-03080, Spain
http://cervantesvirtual.com/
Tel: 34-96-5903400 #2518

ABSTRACT

We describe the digital-book-production flow of the Miguel de Cervantes Virtual Library, from book acquisition up to Internet publishing, highlighting the main requirements and design considerations of the workflow system.

KEYWORDS: digital libraries, DL system architecture, DL development

INTRODUCTION

Although our library covers many different areas, from a "library of voices" up to academic thesis, the vast majority of our present digital books are public domain classic works of hispanic writing from the 12^{th} century up to these days, including narrative, theater, poetry, history and other subjects. Many professionals and technicians take part in the development of our digital books: librarians, scanner operators, correctors, markup specialists and computer technicians. We will describe the production process of digital books and the information system that supports it, that we call "the workflow system".

THE WORKFLOW SYSTEM

The production process begins with a bibliographic search to find interesting available books to digitalize. After the selection of new literary works is complete the librarians elaborate the orders to be sent to various sources (conventional libraries, bookstores, publishers, private collectors in the case of rare books, etc.). Some books are bought, others borrowed. The information related to the orders is stored in the system (e.g. order number, issuing date, reception date, return date when the book is borrowed) along with technical bibliographic information associated to each book, that compose the bibliographic card.

The source physical books and the produced digital books do not always relate in a one-to-one basis. In some cases, a physical book will give birth to many digital books as is the case of collections or "complete works" that may be split into several digital books [1]. Titles may differ also since some works are known by many titles according to different editions. Upon reception, the librarians mark the received source titles as available, and entry records are created for each digital book to be produced from them. A unique code that will identify the d-book permanently is assigned at this point where the production process begins. This code is used within the workflow system, and also in the names of all the files related to the book (during production and also for publication).

At every stage of the production process start and end date-time information is recorded along with the operator identification for follow up and production control purposes.

At any time during the production process, the librarians can access the records of the d-books under development to modify bibliographic-catalog information. Cataloging refers to data like the subject of the work, its universal decimal classification, authors and collaborators (with birth and death dates, if available) and a series of author-title search keys that will simplify the location and retrieval of d-books to the users of the DL.

The first production stage is scanning and optical character recognition (OCR). The resulting output is a set of files of two classes: scanned images and after OCR, text documents. The former are stored in backup media for future projects. The latter, after quality control, pass to the correction stage. At quality control, if too many errors are detected, measures are taken to adjust the scanning-OCR process to improve the resulting output, since a high rate of mistakes rise the time-cost of the rest of the process. As our library handles books of many centuries of hispanic writing (including Spanish, Cata-

[1] In a DL there is no reason to group different literary works as it is done on a printed book, since the criteria used for traditional books do not apply to their digital siblings. However, there are exceptions. Literary experts may decide to group poems from different collections into a single digital book.

lan and Galician languages), peculiar problems arise which are not trivial, like the spell-checking of ancient writings. Projects to tackle this time-dynamics of language are being considered by our research team.

The next stage in production is correction, where specialists in literature, history and linguistics correct not only OCR errors but also mistakes of the original publication, sometimes making side-by-side comparisons between different editions.

Although some works of long extension, may be fragmented to facilitate the correction task, it is desirable that a single person corrects each book, to take advantage of the specialized knowledge she/he acquires during the correction process of its contents and its peculiarities. In case of splitting, this has to be reflected in the workflow system by adding an extension (A, B, C ...) to the identifier code of the split parts).

The following stage is markup. If the work was fragmented for correction, the markup operator joins the fragments. It is also advisable that the markup be accomplished by the same person that performed the correction, due to her/his familiarity with the contents, but sometimes a different highly skilled markup specialist may be preferred. So the system was designed to allow flexibility of task assignment. Markup consists of applying marks with display format indications, but most specially marks that indicate how the digital book must be built. These marks are concerned with index inclusion, chapter fragmentation, hypertext linking and graphics insertion, among others.

After markup, documents go through a final revision before being passed to subsequent stages. Documents that pass this check are considered final products, and are preserved and stored accordingly. From this point on, processing is automatic, and although we now generate a few publishing formats, we expect to generate many more based on these revised corrected marked-up files.

Finally, the d-book generation stage is a fully automated process where the marked-up document files, and a set of templates are processed together by a parser program that produces the HTML d-book files ready for immediate Internet publishing. The use of templates allows us to generate different sets of books with the same look and functionality within a given set, and a different look and functionality for other sets. We use this approach to maintain two different portals of the same DL: the original one, Miguel de Cervantes [2] with books in Spanish language, and the Joan Lluis Vives [1] with books in Catalan language. [2]

Once the generated d-book has been checked and properly catalogued, it can be published in the Internet pages of the library for public access.

[2] Our library specializes in books written in any of the Iberian family of languages. A new portal for Galician books is about to be opened to public access.

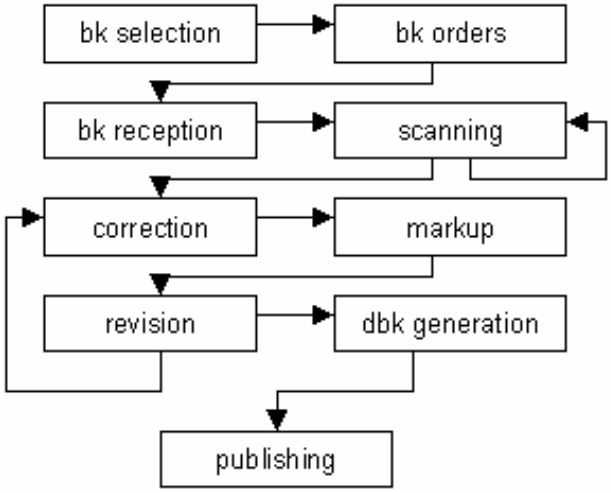

Figure 1: Workflow diagram

MANAGEMENT INFORMATION SUBSYSTEM

A fundamental pending aspect of the workflow system is that it must provide the tools for production an quality control. It should be able be to generate production reports by date range, indicating the work accomplished by each person at every stage, as well as global productivity and quality statistics.

CONCLUSION

The workflow system of the Biblioteca Virtual Miguel de Cervantes Saavedra is being developed according to an incremental development schedule. The first stage is operative since May 1999. Apart from the infrastructure to support the workflow requirements explained above, it includes automatic document backup facilities, essential in a DL that handles thousands if documents.

When completed, it will combine the power of a database system with the functionality of a document management system (like those used for Software Configuration Management in the software industry). It is expected to automatically indicate when a piece of work under development takes too long at a given stage, as well as to prevent almost any possibility of file loss, misplace or misnaming, common errors when handling high amounts of documents.

REFERENCES

1. Biblioteca virtual joan lluis vives. http://lluisvives.com/, 1999.

2. Biblioteca virtual miguel de cervantes saavedra. http://cervantesvirtual.com, 1999.