elpub2002 Proceedings — J. A. Carvalho et. al. (Eds.) — VWF Berlin, 2002

Digital Binding of Multiple Manifestations and Collections of Literary Works

Hugo Amorim and José Borbinha

Instituto Superior Técnico, Technical University of Lisbon, Lisbon, Portugal
National Library of Portugal, Lisbon, Portugal
{hugo.amorim, jose.borbinha}@bn.pt

Abstract. This paper addresses the problem of managing, handling and archiving multiple digital manifestations of literary works. The problem was motivated by a real requirement at the National Library of Portugal. It is proposed an approach for the problem based on the concept of digital binding of multiple manifestations of digital works, taking into account their technical and intellectual contents as well as their metadata and possible contexts of usage. It is thereafter described a framework where one can manage and relate several manifestations, bound in a new object containing technical and bibliographic information about the literary work whose manifestations were included. The model is implemented as a repository managed by an XML framework, comprising the data and a metadata model, as also the related editing, management and contents transforming tools.

1 Introduction

The problem of dealing with a nation's cultural heritage in a digital world has been hardly studied and discussed, leading to the definition of best practices and standard procedures whose aim is to maintain and preserve these cultural assets.

Specific techniques for traditional archiving and cataloguing were already defined and are widely used by the entities responsible for preserving literary heritages. However, these techniques are usually vague for a particular subset, which is getting bigger nowadays, of a country's cultural heritage: the digital and multiple manifestations of literary works.

Thinking of these limitations and having to deal with this problem at the National Library of Portugal, we developed a framework to help professionals who manage these genres of literary works and also to allow anyone to easily explore repositories of digitised works.

The main problem to be addressed is the problem of the digital binding of all the contents that compose one or more manifestations of a specific literary work. A requirement is to store with the multimedia contents (possibly a combination of images, texts and sound) also the metadata for content identification (bibliographic metadata) as well as the structural metadata of the contents.

2 The Problem

The National Library of Portugal is a legal deposit library (in Portugal the deposit of printed works is enforced by law). Has a result of that context, the National Library owns a large and valuable collection of printed literary works, as also of manuscripts related with these works or their authors.

Among all these publications, one can find some extremely rare and valuable works, such as manuscripts from the Middle Age, first or unique editions of reference works and others, some hand glossed by the authors. To preserve these special contents, some being deteriorated, the access has to be sometimes restricted for the general public.

To overcome these limitations, it has been carried out an initiative for digitisation, basically aiming to produce simple images of those contents. The original idea was to put these images available to the general public through a web browser, at least on a restricted intranet but, when possible, on the Internet. But, producing the images (and we are talking about millions of images) is not enough; it is necessary to provide feasible ways to manage and store them, as also to explore those valuable contents.

Another initiative currently on-going at the library is the digitisation and production of digital talking books. Talking books are simply human speech recordings of people reading printed books, without any kind of dramatization, but just narrating. Those special "readings" are being converted to digital audio, to provide more easy access and also for better long-term preservation. The main purpose of this activity is to provide the visual-impaired and print-disabled community with an enhanced product, more understandable and easy to use. Once again we have the same problem: how to archive, manage and explore this kind of literary work's manifestations.

Finally, the National Library has been also promoting among the creators and publishers the digital deposit, not only for the digital published and digital-born works but also for versions of printed works (especially for new or recent editions).

Considering all these sources of digital contents, which will make our National Digital Library, there can be identified in our collection multiple manifestations of the same work such as for example a digitised copy of a first edition, plus a textual transcript of it (in ASCII, HTML, etc.), a speech recording of someone reading that same edition, or even also a digital-born version of a recent edition. In some cases, we can even find digitisations and digital text versions of foreign language translations of those works.

In the next sections we present one part of our proposal to the problem of store, manage and access to those classes of contents (a system were one can manage and relate all these manifestations and respective multimedia contents). The aim is to define a new concept of digital binding for all of those manifestations, covering not only the ones we might have now but also those that we might acquire in the future.

3 The framework

The main requirement for this work was to create a framework where the contents could be structured in a way that, when searching and find a particular work, one could learn about and access all the manifestations and versions of that same work existing at the National Digital Library.

In this model there is a central entity – the digital work – that acts as an information aggregator, virtually binding different manifestations (as also their possible different versions). This aggregated information comprises information about the work (author, title, subject, etc.) and about the manifestations themselves (edition, genre, etc.). This later information can be further divided into information about the manifestation itself – bibliographic information – and information about the manifestation contents and their structure – structural and descriptive information. Actually, all this is metadata about a work and the contents that compose their manifestations.

Additionally, an effort has been made, when defining the metadata containers, in order to allow the establishment of links between the several manifestations at the time of binding them into a new digital work. These links must allow the cross-manifestation reading and exploring of a literary digital work.

As mentioned before, another requirement for the framework was the ability to incorporate any kind of digital or digitised manifestation in the collection, despite their source or format. To achieve this it is necessary to build a set of tools capable of generating a digital work, given the contents and its metadata.

Although with these tools we can easily create a virtually bound digital work, and despite the quality and completeness of its contents, it consists in a set of related multimedia files and remains almost a machine-readable work. It is necessary to develop a set of tools or filters that, given a digital work, can bind its contents according to a specific platform, defining views and ways to explore, and making them understandable to humans.

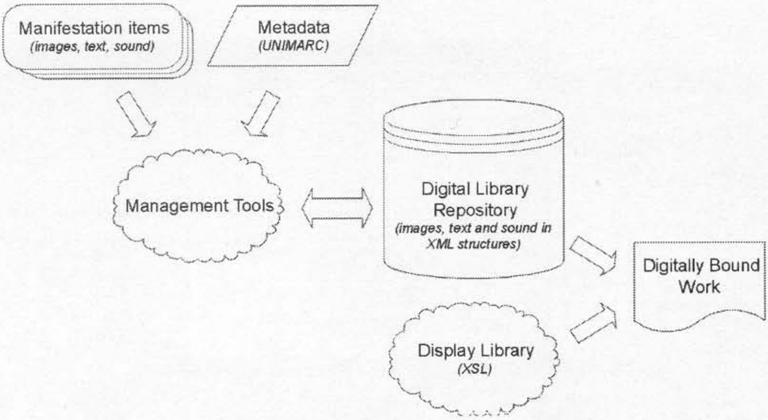


Fig. 1. *The framework and its data flow*

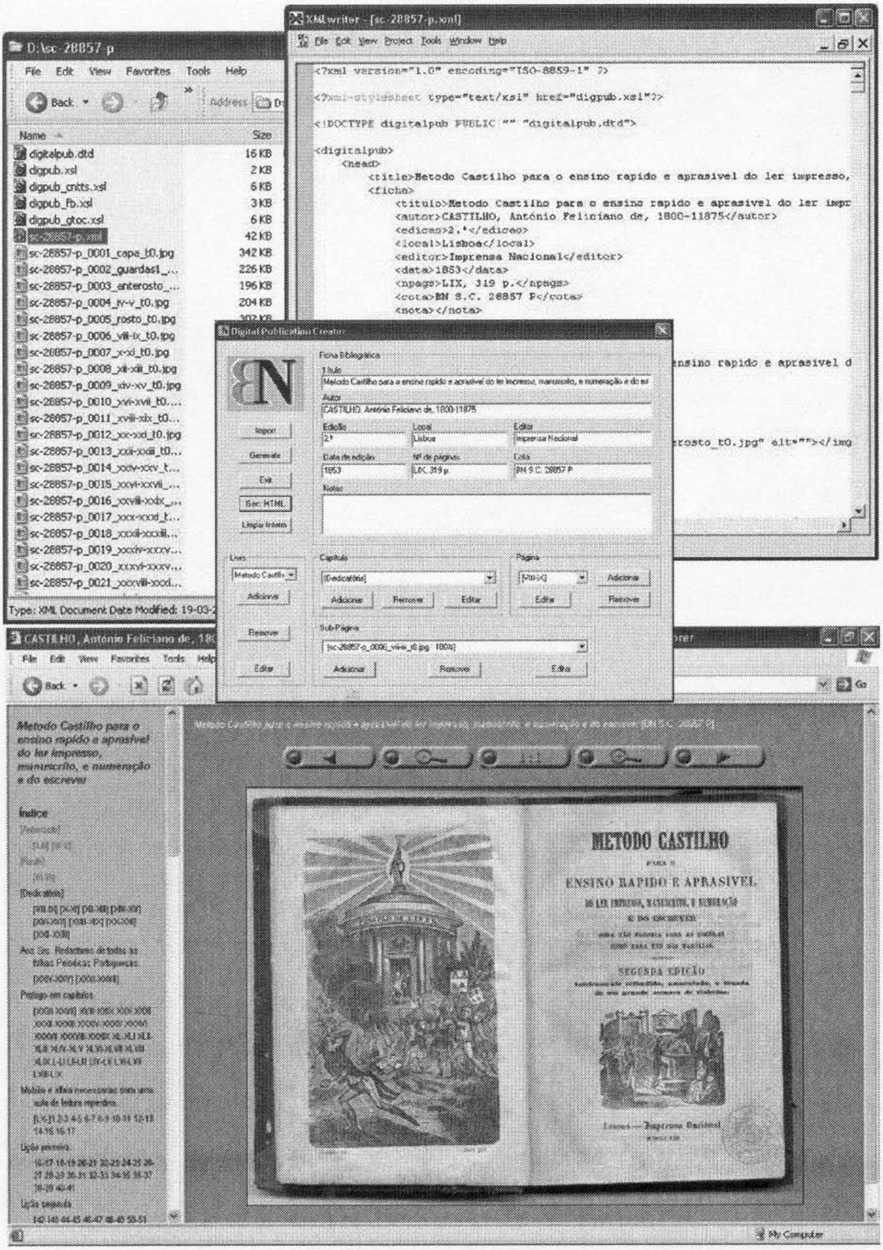
These are the three main components of our proposed framework. To achieve these functionalities it was therefore necessary to go through three steps:

1. specify and build an XML structural metadata model to describe and structure the contents that composes the manifestations and its bibliographic metadata (originally in UNIMARC);
2. create the tools to import and manage the digital or digitised manifestations and to make its virtual binding, creating and managing digital works (basically, XML data structures);
3. create a library of style sheets capable of processing the data in the digital works and display it given a specific access platform, allowing the definition of pre-specified views or paths to explore the work and its bibliographic metadata.

To get a general idea of what we are talking about, in Figure 2, we show a sample of a digital work bound with this framework. In the upper left corner is possible to see a list of the files that compose the work's manifestations; on the opposite right corner an extract of the XML file produced by the management tool in the centre. At the bottom is presented a snapshot of the visualization, using one of our display libraries, in a web browser.

3.1 The Structural Metadata Model

The structural metadata model is what actually defines the digital work. It defines both the logical structure of the work and the containers for the descriptive metadata about the work and respective manifestations. This way, the model specifies a well-defined and open structure for the central entity in a digital work, which is the


XMLwriter - [sc-28857-p.xml]
File Edit View Project Tools Window Help
D:\sc-28857-p
File Edit View Favorites Tools Help
Back
Address
Digital Publication Creator
Import
Generate
Exit
Ger. HTML
Limpar Interm
Adicionar
Remover
Editar
CASTILHO, António Feliciano de, 18
Metodo Castilho para o ensino rapido e aprasivel do ler impresso, manuscrito, e numeração e do escrever
Indice
[Dedicatória]
Prologo em capitulos
Lição primeira
Lição segunda
METODO CASTILHO
PARA O
ENSINO RAPIDO E APRASIVEL
SEGUNDA EDIÇÃO
Lisboa — Imprensa Nacional
My Computer

```
<?xml version="1.0" encoding="ISO-8859-1" ?>

<?xml-stylesheet type="text/xsl" href="digpub.xsl"?>

<!DOCTYPE digitalpub PUBLIC "" "digitalpub.dtd">

<digitalpub>
    <head>
        <title>Metodo Castilho para o ensino rapido e aprasivel do ler impresso,
        <ficha>
            <titulo>Metodo Castilho para o ensino rapido e aprasivel do ler impr
            <autor>CASTILHO, António Feliciano de, 1800-11875</autor>
            <edicao>2.ª</edicao>
            <local>Lisboa</local>
            <editor>Imprensa Nacional</editor>
            <data>1853</data>
            <npags>LIX, 319 p.</npags>
            <cota>BN S.C. 28857 P</cota>
            <nota></nota>
```

Name	Size
digitalpub.dtd	16 KB
digpub.xsl	2 KB
digpub_cntts.xsl	6 KB
digpub_fb.xsl	3 KB
digpub_gtoc.xsl	6 KB
sc-28857-p.xml	42 KB
sc-28857-p_0001_capa_t0.jpg	342 KB
sc-28857-p_0002_guardas1_...	226 KB
sc-28857-p_0003_anterosto_...	196 KB
sc-28857-p_0004_iv-v_t0.jpg	204 KB
sc-28857-p_0005_rosto_t0.jpg	302 KB
sc-28857-p_0006_viii-ix_t0.jpg	
sc-28857-p_0007_x-xi_t0.jpg	
sc-28857-p_0008_xii-xiii_t0.jpg	
sc-28857-p_0009_xiv-xv_t0.jpg	
sc-28857-p_0010_xvi-xvii_t0....	
sc-28857-p_0011_xviii-xix_t0...	
sc-28857-p_0012_xx-xxi_t0.jpg	
sc-28857-p_0013_xxii-xxiii_t0...	
sc-28857-p_0014_xxiv-xxv_t...	
sc-28857-p_0015_xxvi-xxvii_...	
sc-28857-p_0016_xxviii-xxix_...	
sc-28857-p_0017_xxx-xxxi_t...	
sc-28857-p_0018_xxxii-xxxiii...	
sc-28857-p_0019_xxxiv-xxxv...	
sc-28857-p_0020_xxxvi-xxxv...	
sc-28857-p_0021_xxxviii-xxxi...	

Type: XML Document Date Modified: 19-03-2

Ficha Bibliográfica

Título: Metodo Castilho para o ensino rapido e aprasivel do ler impresso, manuscrito, e numeração e do es

Autor: CASTILHO, António Feliciano de, 1800-11875

Edição: 2.ª

Local: Lisboa

Editor: Imprensa Nacional

Data de edição: 1853

Nº de páginas: LIX, 319 p.

Cota: BN S.C. 28857 P

Notas:

Livro: Metodo Castilh

Capítulo: [Dedicatória]

Página: [VIII-IX]

Sub-Página: [sc-28857-p_0006_viii-ix_t0.jpg 100%]

Fig. 2. *An example of a digital work bound in this framework*

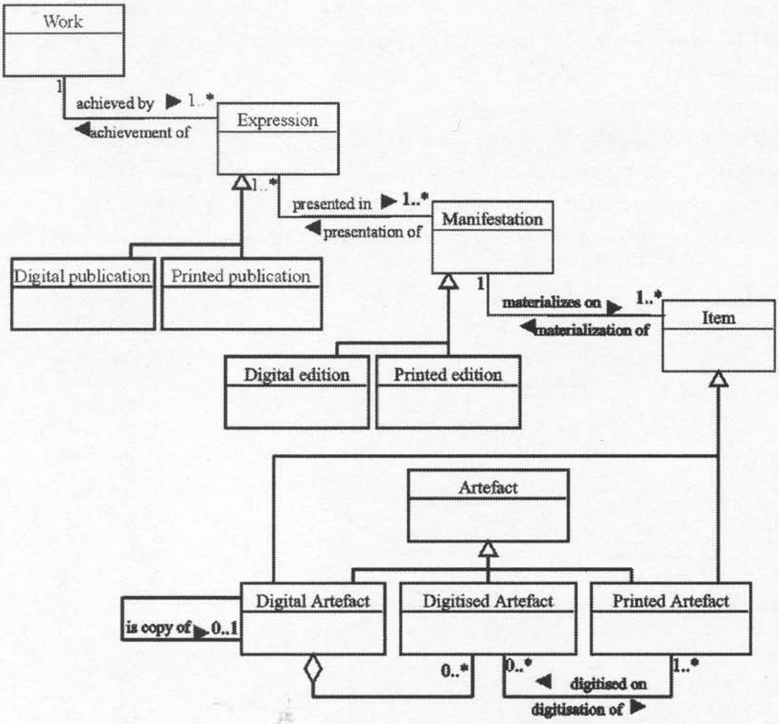


Fig. 3. *IFLA's Functional Requirements for Bibliographic Records (a proposal for the new paradigm)*

root of each work in our XML managed repository of digital multi-manifestation collections.

The model, which is still being further improved, intends to be a hierarchical model, following the IFLA's model for functional requirements for bibliographic records [1]. Basically, the main difference is related with the levels we intended to implement, since we bypass the expression layer because we only relate manifestations from the same expression of a work (see Figure 3). However, we are studying the possibility to expand this model to accommodate at least metadata about other expressions of the same work, at least their bibliographic records.

To coherently and systematically define this model and with the purpose to make this structural metadata model machine-understandable, providing the ability to easily know if a specific digital work is compliant with the defined model, a DTD[1] has been specified.

[1] Document Type Definition

The digital work is then composed by an XML[2] document, compliant with the defined DTD that holds the metadata model. As depicted in Figure 4, two main sections compose this document:

- a section of bibliographic metadata (title, author, edition, editor, subject, content type. . .) and metadata for cataloguing and classification;
- a section of structural metadata containing the description of the multiple manifestations of that work and establishing eventual relations between them.

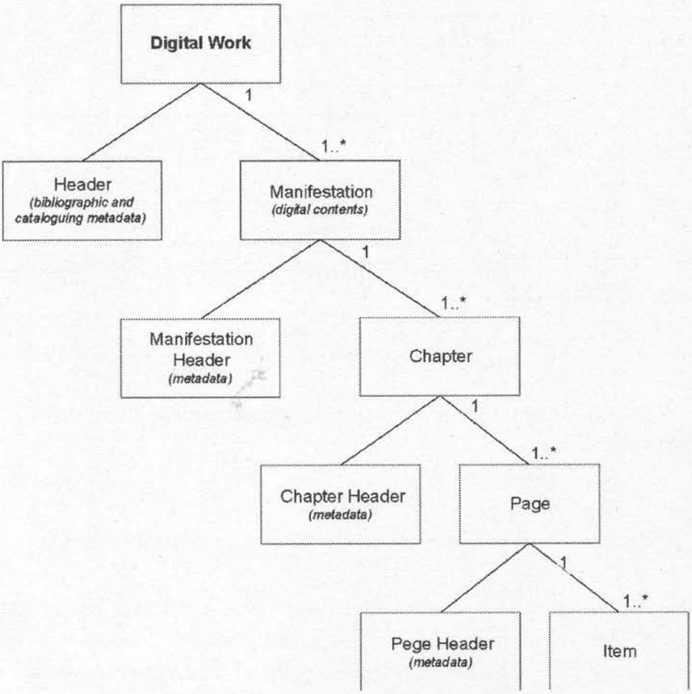


Fig. 4. *The Structural Metadata Model – main entities and relationships*

Each manifestation is further divided into chapters, pages and items. Similarly to the header section of the digital work, each manifestation section can have also a metadata subsection for specific information (such as type of manifestation, date of production, edition. . .). The same applies for chapters and pages.

[2] Extensible Markup Language [2]

This is the model used in the actual prototype. Based on the experience already acquired, we are currently analysing a generalization of this schema, in order to make it more flexible to accommodate less traditional works. Instead of the three level hierarchy, classifying the work's sections as chapters, pages and items, we are attempting to create an n level hierarchy composed by a recursive set of composing elements. This is what we call of "section" in Figure 5.

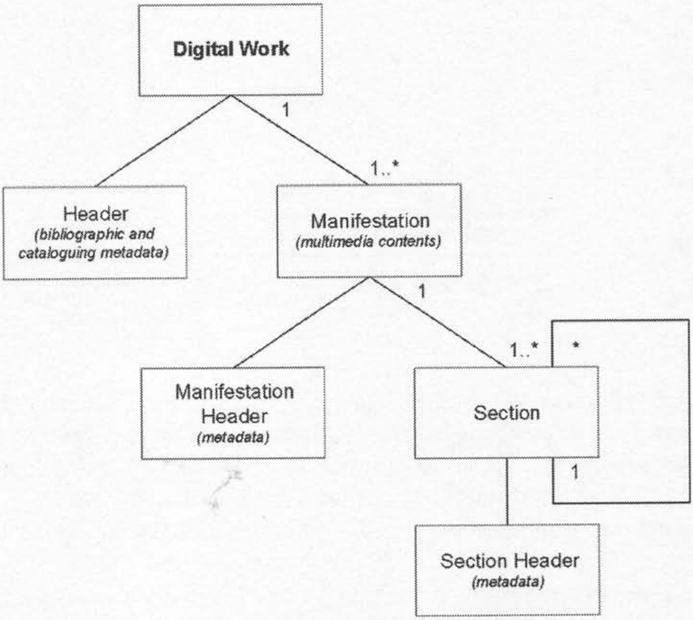


Fig. 5. *The new structural metadata schema under development.*

This later approach has some important advantages, such as the possibility to digitally bind works not necessarily composed by chapters and pages, as well works that are composed by nested chapters, or even unique non conventional works, resulting of the physical binding of heterogeneous contents (a usual practice in the early days of the printing industry, where individuals used to by first the printed works in simple sheets, not bound, which were latter on bound with other works, making unique pieces). Though, the embedded information of the first approach about the type of section is lost, but that can be overcome if this type or class is specified and assigned to the section as an attribute.

Independently of the model for a digital work, it is required to define also a higher level of aggregation to represent the concept of collection. A collection is usually a set of literary works, either related or independent. In our framework this can be easily represented by a new entity called precisely "Collection", which might comprise several digital works. However we decided to define it also as a recursive entity, allowing not only the aggregation of digital works but also the aggregation of other collections (see Figure 6).

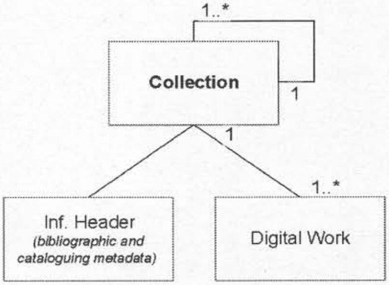


Fig. 6. *The higher level – a Collection*

The concept of collection is important not only for the management of the digital repository of literary works but also for the final user. This way, we can provide the contents to the readers dynamically grouped by usual or real collections, as we can find in the printed works domain. It can also be possible for the user to demand a custom collection, grouping his/her favourite works for future exploration or reading.

The representation of this new concept in the framework is achieved by a new XML document that has its own DTD, defining a model similar to the one depicted in Figure 6. This file contains the descriptive and bibliographic metadata about the collection and also structural metadata specifying the components of the collection, either digital works or other collections. Some specific tools that are discussed in the next sections aid the production of all these files.

3.2 Managing Tools

To create a new digital work it is necessary to bind the contents that compose its digital manifestation, or its multiple manifestations, into a new XML document, according to the correct DTD. This is a particularly boring task if performed by manually editing the XML document (a digitised version of a 400 pages work is made of 400 image files...). The same happens when we try to change or merge previously bound works.

To make this task more efficient and less boring we developed a set of tools – which we denominate "managing tools" – to generate the digital work and collection XML documents.

This way, it is easy to generate and edit the digital binding, without any need to know XML or having to know the metadata model, as this is embedded in the application that is responsible for the compliance with it.

To create the digital works, using these tools, one just has to fill the metadata fields, through a form, insert the multimedia contents by dragging and dropping them on the appropriate locations, and click on the *generate* button – the application does all the rest: the generation of the remaining structural metadata and the creation of the XML document depicting the digital work. Further editing of these digital works is also easily made by the application, which allows the editing of previously bound work.

New tools are currently being developed to allow the merging of two or more digital works and the creation of collections. These tools generate a new XML document, containing references to the XML documents that aggregate the information for each digital work as well as its metadata (bibliographic and structural) about the collection itself and about the several digital works. The application automatically downloads the bibliographic records from the appropriate bibliographic metadata server running on the National Library (the bibliographic record is retrieved in UNIMARC from the central bibliographic database, in a specific XML encoding).

The same tools allow us to extract the digital works and collections from the repository, independent and unrelated manifestations and bind them into new custom digital works collection. This way it is possible to export from the repository a specific part or manifestation of a collection or digital work, for instance, into a CD-ROM answering to a readers demand.

All these tools guarantee that the XML documents produced are valid and compliant with the defined DTDs. However, using these tools, we can only obtain well-defined and structured files, referencing, identifying and classifying the multimedia contents that compose the digital works or the collections. We still cannot see what those files reference. Although the data layer is ready and fully functional we need a presentation layer. The next section presents one of the solutions that might act as a presentation layer for this framework.

3.3 Display Libraries

Ranging from applications interpreting the XML documents and displaying the multimedia contents to an open source display level composed by some kind of rules interpreted by a public domain application, several systems could be developed to the presentation layer of this framework.

The former systems might have the advantage of being robust and having high performances presenting the contents to the users. Although the fact of being bound to a specific platform and working as a black box that no one can understand or change are major drawbacks.

Considering this, we adopted a solution similar to the later system referenced – an open solution that works on every platform and that anyone can adapt or rebuild.

To achieve this, we created a set of Display Libraries – collections of filters and rules embedded in style sheets that, benefiting from the well-defined structure of the XML document, know how to extract the relevant information and how to present it to the reader.

With this collection we can effectively bind a digital work to almost any display device, format, protocol or platform we want. The data structure is well defined – it is only necessary to define what is expected at the output.

With this mechanism to extract and display data from the digital work and since we can express details till very refined level, it is easy to combine several manifestations of one work in the same display. Moreover, it is possible to give the user the ability to choose which manifestations he wants to explore and thereafter display them synchronously on-screen.

At the moment, we are developing a set of XSL[3] style sheets to display image-based manifestations of works. These style sheets extract the metadata from a digital work, displaying it on several static XHTML documents, allowing the user to explore this type of manifestations almost like exploring a real (printed) work.

4 Conclusions and Future work

Although there is too much work to do we already have some results. We finished a prototype that allows us to build digital works of image-based manifestations only. This prototype is already being used at the National Library and has been generally accepted by the people who used to do the managing of the digitised image based works, saving time and achieving better results in the process of storing these contents.

A few hundreds of digital works are already bound with the managing tool developed for this framework, and almost 90% of them were built from the beginning with zero problems. The tool aids the user[4] by asking for general work identification and descriptive metadata and allows him to easily rebuild the structure of the work from the snapshots of books that compose the printed work's manifestation.

These digital works are now stored in the collections repository. The XML documents composing the digital works solved yet another problem: finding and

[3] Extensible Stylesheet Language [3]

[4] the digital work creator, in this case

accessing a specific part of a digital work as they work as an index of the multimedia contents that compose the manifestations.

Our Display Library collection has already a few components – sets of XSL style sheets that, given an XML document from a specific digital work, generate a set of static XHTML[5] pages depicting the content of the selected digital work. The user can then explore the work in a way very similar to the exploitation of a normal printed manifestation.

These collections of style sheets allowed us to automatically generate thousands of XHTML pages on-line at the National Digital Library site that, otherwise, would have to be done manually, in a very hard and time consuming way. The successive refinements done to the style sheets, in order to present contents in different ways, gave us the opportunity to enrich our Display Library with several versions of similar libraries.

Other Display Libraries are currently being developed, such as a library to generate web pages on the fly, as we explore the contents of a collection in a web browser. In other words, the XHTML pages are generated dynamically as the user explores the work rather than previously. The main advantage is the instantaneous visibility of any change in the style sheets.

The next step is to continue to upgrade the framework, supporting more manifestations and adding more functionalities (both improving the managing tools for other type of manifestations and creating XSL filters to digitally bind the contents to other languages and platforms). By now the priority is to finish the specification of the DTD, supporting the new concept of "Collection".

References

1. IFLA. Functional Requirements for Bibliographic Records. Final Report. `http://www.ifla.org/VII/s13/frbr/frbr.pdf`
2. W3C. Extensible Markup Language (XML). `http://www.w3.org/XML/`
3. Extensible Stylesheet Language (XSL). `http://www.w3.org/Style/XSL/`
4. XHTML^{TM} 1.0: The Extensible HyperText Markup Language `http://www.w3.org/TR/xhtml1/`

[5] Extensible HyperText Markup Language [4]