Information Processing and Management 48 (2012) 931–955

Contents lists available at SciVerse ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Modeling, encoding and querying multi-structured documents

Pierre-Édouard Portier, Noureddine Chatti, Sylvie Calabretto, Elöd Egyed-Zsigmond *, Jean-Marie Pinon

Université de Lyon, LIRIS UMR 5205 – INSA LYON, 7, avenue Jean Capelle, 69621 Villeurbanne Cedex, France

ARTICLE INFO

Article history:
Received 24 July 2008
Received in revised form 30 August 2011
Accepted 28 November 2011
Available online 31 March 2012

Keywords:
Multi-structured document
XML
MultiX
Multi-structured document querying
XQuery

ABSTRACT

The issue of multi-structured documents became prominent with the emergence of the digital Humanities field of practices. Many distinct structures may be defined simultaneously on the same original content for matching different documentary tasks. For example, a document may have both a structure for the logical organization of content (logical structure), and a structure expressing a set of content formatting rules (physical structure). In this paper, we present MSDM, a generic model for multi-structured documents, in which several important features are established. We also address the problem of efficiently encoding multi-structured documents by introducing MultiX, a new XML formalism based on the MSDM model. Finally, we propose a library of Xquery functions for querying MultiX documents. We will illustrate all the contributions with a use case based on a fragment of an old manuscript.

1. Introduction

1.1. Document structuring

Document structuring is used in many applications such as document exchange, integration and information retrieval. Several types of structures (physical, logical, semantic, ...) (Nanard & Nanard, 1995; Poullet, Pinon, & Calabretto, 1997) have been defined for several specific uses. Moreover, a document can actually be a vehicle for various media types that can themselves introduce other structural layers (such as the temporal dimension of an audio track).

A single document can be used in many contexts. Thus, its content might be presented through many structures. In this case, the structures are said concurrent or parallel, since they share the same content. Humanities provide numerous instances of such structures. For example, the study of medieval manuscripts often implies the creation of concurrent hierarchical structures. First, we can consider a ubiquitous and trivial case of overlapping: the physical book-structure of a manuscript (a sequence of pages, columns, lines, etc.) and its syntactical structure (a sequence of sentences, words, etc.). Less trivial would be a structure of the sequences of damaged characters. Fig. 1 is an extract of such a medieval manuscript fragment with its transcription. It should be noted that damaged characters are overlapping with words and words are overlapping with lines. The emphasis on multi-structured documents comes with the possibility of formally encoding documentary structures with digital representations. The fact that we find numerous examples of multi-structured documents in the Text Encoding Initiative (TEI) guidelines (TEI Consortium, 2011) should prove it. Among those examples, we can mention in verse drama, the structure of acts, scenes and speeches that often conflicts with the metrical structure. Indeed, poems often

* Corresponding author. Tel.: +33 4 72 43 62 97; fax: +33 4 72 43 87 19.
E-mail address: elod.egyed-zsigmond@insa-lyon.fr (E. Egyed-Zsigmond).

0306-4573/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ipm.2011.11.004

Fig. 1. A fragment of an old manuscript and its transcription.

provide multiple concurrent structures. Poem 1 is an example of two stanzas from a poem by Lewis Carroll[1] with verses, speeches, and syntactic elements producing overlapping concurrent structures.

1.2. Construction of multi-structured documents

1.2.1. Introduction

The work presented here will deal with the central issue of encoding multi-structured documents. However, in order to make clear the advantages of our model over existing solutions, we now introduce the important but little-studied problem of the construction of multi-structured documents. Indeed, in a large majority of real life situations, documents are not *a priori* given but have to be constructed. Moreover this construction is in most cases the work of a team. Thus, it is to be expected that the partition of the annotations in a number of different structures comes from this collaborative work. In other words, the construction of multi-structured documents is a dynamic process. How do structures emerge? How to check on their coherence? ... In order to tackle these important issues, we need an adequate encoding for multi-structured documents.

In a previous work (Portier & Calabretto, 2009, 2010) we explicitly studied the construction of multi-structured documents. Although the formal representation we then used was based on RDF rather than XML, this previous work offers a well-adapted and non-trivial applicative context for the XML based model we now propose. Moreover it gives us the opportunity to motivate the key choices behind our proposition.

1.2.2. Context

We studied how multi-structured documents are constructed in a multi-users context composed of philologists. Our work is based on experience gained working with Humanities researchers building digital editions of large archives of (mainly handwritten) manuscripts from various epochs. Digital editing covers the whole editorial, scientific and critical process that leads to the publication of an electronic resource. In the case of manuscripts, editing mainly consists in the transcription and critical analysis of digital facsimiles, that is to say, the creation of a textual document associated with the images of a handwritten manuscript. We discovered that multi-structured documents construction was at the heart of their work. Indeed, they need to let coexist a multiplicity of structures in order to access a document according to many interpretations. Thus, we proposed a methodology promoting the emergence of multiple structures in a multi-users context.

1.2.3. Scenario

This study gave us a lot of scenarios similar to the following: a philologist finds a consistent subset about medicinal plants in a stack of pages of consequent size. He creates a new collection from this subset. He creates an association named "mainSubject" between this collection and the topic "Medicine". He starts to transcribe the collection and annotates some intervals of the text with terms such as "quotation" and "prescription". Later, he may discover that this collection is in fact a preparatory work for another piece of the archive. He then creates an association named "preparationFor" between the two collections, etc.

How is it that, for example, a user chooses to place the term "citationTitle" within a structure named "Citations" while he affects the term "line" to the "Physical" structure? This kind of question brought us to define a methodology for the construction of multi-structured documents.

1.2.4. Methodology

First of all, we only consider content addressable by concrete intervals: characters intervals in a text, time intervals in an audio or video document. In order to offer some unity to the various structures emerging from the work of a group of users, we introduce an *a priori* rule: a structure must form a hierarchy. In other words, the annotated intervals of a structure should never overlap. By dynamically checking the validity of this rule we managed to ease the collaborative construction of multi-structured documents. We should now briefly illustrate this idea with the previous example of an old manuscript (see Fig. 1).

We assume that the researchers made use of annotation terms such as: "line", "w" (for word) and "dmg" (for damaged characters). The transcription process continues until a word is overlapping with two lines (see strong & dashed lines of Fig. 2).

[1] http://www.gutenberg.org/files/13/13-h/13-h.htm.