# Electronic Publishing of ADB Editions

José Carlos Ramalho
jcr@di.uminho.pt
University of Minho

José Luís Santos
jlsantos@aeiou.pt
Critical Software

**Abstract**

Arquivo Distrital de Braga (ADB), the second largest Portuguese Archive, owns thousands of important master pieces, published according to merely typographic techniques. For the great majority, their existence in electronic support is still a mirage.

However, some years ago, ADB and the Department of Computer Science at the University of Minho have created a synergy aiming at modernizing, through computing processes, the unique cultural value the Archive owns.

The goal of this document is to provide a technological overview on the work developed at ADB along the years in the arena of electronic publishing. It includes the presentation of a practical case.

## 1 Introduction

The Arquivo Distrital de Braga[1] (ADB) and the Department of Computer Science at the University of Minho[2] have been uniting efforts in order to *recover old documents in paper format* through computing processes. Therefore, and under the scope of some University of Minho courses, one has been electronically publishing documents of extreme historic and cultural value.

*Electronic Publishing* is the process of producing structured digital formats from paper or binary (or simply non-structured) files. Desired digital formats are typically those suitable for on-line viewing or paper printing. This process can be divided in three fundamental phases:

**Document Analysis** Specification of a grammar, expressed in a DTD, declaring the hierarchy of structural elements and their properties.

**'Up-translation'** Transformation of data represented in an arbitrary format into valid XML instances. Valid XML instances are correct XML documents respecting the grammar defined in the previous phase. In order to achieve this one applies techniques of (semi-)automatic processing[3].

**'Down-translation'** Translation of valid XML instances, created in the previous phase, into desired formats. In the concrete case of the document presented in 2, one wishes to generate the following formats:

**LaTeX** For post-generation of PostScript (PS) or Portable Document Format (PDF) instances suitable for high-quality paper printing;

**HTML** For web publishing.

[1] `http://www.adb.pt`
[2] `http://www.di.uminho.pt`
[3] In [Cha] one can find some approaches and tools used within this phase.

## 2 Case-Study

This section introduces the reader with the electronic publishing process of *Bullarium Bracarense* ([VA86]), an ADB book of great historic relevance.

*Bullarium Bracarense* aggregates the so-called bulls or pontifical letters that were officially and solemnly emitted. These bulls aim at promulgate doctrinal decisions, saints canonization, measures of ecclesiastical discipline, jubilees or even administrative or pastoral measures, such as creation, division, union or extinction of cathedrals.

Besides the summaries of bulls, *Bullarium Bracarense* includes a synchronic table made up of archbishops and popes, a list of bibliographic references, a list of abbreviations, and anthroponomical and toponymical indexes.

In 2.1, 2.2 and 2.2, one describes, respectively, the three phases mentioned in 1 that make up the electronic publishing process, for the particular case described herein.

The material here presented is based on the report [San01].

The terms 'bull', 'summary of bull', 'pontifical diploma' and 'summary of pontifical diploma' are used interchangeably to refer to *summary of bull*.

### 2.1 Grammar Specification

*Bullarium Bracarense* is a structurally heterogeneous document. Thus, both the definition of its grammar and its processing need special attention. The definition of elements and attributes, compiled in a DTD, personifies the flexibility and mutability of the information they contain.

Given the cardinality of elements constituting the DTD, one only presents those which are, from the practical point of view, the most important ones[4].

What follows is the enumeration of the highest-level elements[5] within the DTD hierarchy. These elements actually represent the major parts the book is divided in.

1. Introduction
2. Synchronic Table
3. Bibliography
4. Summaries
5. Appendix

```
<!ELEMENT bulario (intro, synctable, bibsiglas, summaries, appendix)>
```

The 'Summaries' part will be the one emphasized, as it effectively contains the set of pontifical diplomas summaries, i.e. bulls. Although the book contains a chapter for indexes, the DTD does not consider them, since they are automatically generated from the chapter 'Summaries'.

Pontifical diplomas are grouped in centuries, which means that each century contains a chronologically sorted list of bulls[6].

```
<!ELEMENT summaries (century)+>
<!ELEMENT century (bula+)>
```

Each bull is composed by a header, a body and a block of bibliographic information.

```
<!ELEMENT bula (header, content, addon)>
```

[4] The complete DTD can be consulted at `http://www.di.uminho.pt/~jcr/PROJS/bb/BBOnline/bb.dtd`.

[5] Just below the root element.

[6] In fact, these lists are also *numerically sorted*, as each bull is assigned a unique numerical identifier.

The header contains the sender and recipient of the bull:

```
<!ELEMENT header (who, whom*)>
```

The body of a bull can contain information regarding people and/or places[7]

```
<!ELEMENT content (#PCDATA | anthropos | topos)*>
```

Bibliographic information optionally contains a sentence in Latin, indications of whether the bull resides, previous publications, references in other documents and a remarks field.

```
<!ELEMENT addon (quoted?, library*, publ?, ref?, note?)>
```

The attributes of the element bula are:

**id ID #REQUIRED** Unique identifier ('B' concatenated with an order number).

**role (BULA | BREVE) "BULA"** Classification of bull as of type 'bula' or 'breve'.

**syear CDATA #REQUIRED** Year of issue.

**eyear CDATA #IMPLIED** Final year of issue (only exists if there is an interval of time [syear; eyear]).

**yabout (yes | no) "no"** Boolean representing the accuracy regarding the value of the year of issue.

```
("no", "yes") = (inaccurate, accurate)
```

**ysure (yes | no) "yes"** Boolean representing the certainty regarding the value of the year of issue.

```
("no", "yes") = (uncertain, certain)
```

**month CDATA #IMPLIED** Month of issue.

**day CDATA #IMPLIED** Day of issue.

**dsure (yes | no) "yes"** Boolean representing the certainty regarding the value of the day of issue.

```
("no", "yes") = (uncertain, certain)
```

**where CDATA #IMPLIED** Local of issue.

**wsure (yes | no) "yes"** Boolean representing the certainty regarding the local of issue:

```
("no", "yes") = (uncertain, certain)
```

**sure (yes | no) "yes"** Boolean representing the certainty regarding the veracity of all attributes.

```
("no", "yes") = (uncertain, certain)
```

NB: The definition of the complete grammar has not preceded, *in its entirety*, the 'up-translation' phase: the initial DTD has been adjusted along with the edition and validation processes, inherent to that phase.

[7] Elements representing these entities are explained in 2.4.

## 2.2 From paper to a valid XML instance

Once defined the grammar governing the document, one proceeds with the 'up-translation' phase. This phase comprises digitalizing a *Bullarium Bracarense* exemplar and obtaining a valid XML instance. It includes optical recognition, edition and validation.

### 2.2.1 Digitalization and Optical Recognition

This process followed an usual procedure, where book pages are manually entered in a *scanner*. The result of this process was a series of RTF files that were immediately converted into ASCII and grouped in a single file. This file was the basis for creating the final valid XML instance.

### 2.2.2 Processing and Structuring

This phase consisted of correcting errors resulting from the OCR (Optical Character Recognition) process as well as structuring the document according to the rules inherent to the DTD, through the introduction of appropriate tags.

Substitutions via `Vim` and `Sed`, and application of scripts written in `Awk` and `Perl` were the means used for correcting OCR errors. These errors followed an identified set of patterns. *Parsers* for verifying the validness of the document complemented this process.

## 2.3 Writing of Translators

Once obtained the well-formed and valid XML instance, one proceeds with writing translators responsible for generating LaTeX and HTML formats.

The following programs were written:

**`bb2tex`** Translator of valid XML instances into LaTeX.

**`bb2html`** Translator of valid XML instances into HTML. It generates the following files:

- **`bb.html`** *Frame set* composed of two frames representing the table of contents and the contents.
- **`bbtoc.html`** *Frame* containing the table of contents. Each element of this table is an *hyperlink* for elements within the contents;
- **`bbcts.html`** *Frame* containing the contents.
- **`bbcts.css`** *Cascade Style Sheet* for `bbcts.html`;
- **`bbtoc.css`** *Cascade Style Sheet* for `bbtoc.html`;
- **credits.html** Acknowledgments.

These translators were written in `Perl`, using the `XML::DT` module. They automatically generate the indexes (subsection 2.4). A series of style-related variables can be easily configured.

Additionally, an XSL *style sheet* (`bb.xsl`) was written. This *style sheet* translates `bb.xml` into HTML. To use it, one only needs to point the URL of any XML-XSL-aware browser to file `bb.xml`.

Analogous to `bb2html`, `bb.xsl` also generates a *frame set* with two *frames*, one containing the table of contents, the other the contents. Making use of a set of `javascript` functions (residing in `bb.js`), these *frames* are not generated in the form of files, but in-memory. The table of contents contains, besides what the original does, direct *hyperlinks* to every single bull. This is achieved through bull order numbers.

## 2.4 Indexes

As already stated, *Bullarium Bracarense* contains two indexes: *anthroponomical* e *toponymical*. Elements representing people and places are, respectively, anthropos and topos:

- anthropos

```
<!ELEMENT anthropos (#PCDATA)>
<!ATTLIST anthropos
fullname CDATA #IMPLIED
role CDATA #IMPLIED>
```

**fullname** Person's full name. For the purpose of generating the *anthroponomical index*, the value of this attribute has priority over the contents of the element. In other words, this is the value that will appear as an item of that index.

**role** Designation of person's position (e.g., degree).

For instance, the index entry

```
DIAS, Gonçalo - cónego de Viseu, docs. 351
```

corresponds to element

```
<anthropos role="cónego de Viseu">
Gonçalo Dias
</anthropos>
```

that appears on bull 351.

The same index entry can be associated to more than one bull references.

- topos

```
<!ELEMENT topos (#PCDATA)>
<!ATTLIST topos
contents CDATA #IMPLIED
name CDATA #IMPLIED
geoinfo CDATA #IMPLIED
role CDATA #IMPLIED>
```

**contents** Similar to attribute `fullname` of anthropos, this attribute prevails over the element contents, *solely* for the purpose of generating the *toponymical index*, meaning that its value will be the one appearing as an item of that index.

**geoinfo** Geographical information.

**name** Complementary geographical information.

**role** Expression that designates contextual information over the local.

For instance, the entry

```
PALÊNCIA - [cid. Espanha]: doc. 145;
bispo, docs. 68, 70, 73, 328;
concílio provincial, doc.14;
deão, doc. 68
```

was generated from the following elements:

```
<topos geoinfo="cid. Espanha">
Palência
</topos> <!-- doc. 145 -->

<topos geoinfo="cid. Espanha" role="bispo">
Palência
</topos> <!-- doc. 68 -->
<topos geoinfo="cid. Espanha" role="bispo">
Palência
</topos> <!-- doc. 70 -->
<topos geoinfo="cid. Espanha" role="bispo">
Palência
</topos> <!-- doc. 73 -->
<topos geoinfo="cid. Espanha" role="bispo">
Palência
</topos> <!-- doc. 328 -->

<topos geoinfo="cid. Espanha" role="concílio provincial">
Palência
</topos> <!-- doc. 14 -->

<topos geoinfo="cid. Espanha" role="deão">
Palência
</topos> <!-- doc. 68 -->
```

Along the document processing, two structures were filled with anthroponomical and toponymical information. A glimpse of one of those structures is:

```
%topos = (
  'Monção' => { 'convento'  => [1, 2, 3, 4, 5],
                'mosteiro de S. Francisco' => [6, 7, 8, 9, 10],
              },
  'Braga'  => { 'Bom Jesus'  => [21, 52, 63, 74, 85],
                    ...      => [...],
              },
    ...
);
```

Keys of this associative *array* are the names of cities, villages, etc.; its co-domain is another finite function that relates specific places of those cities to a set of bulls referenced in the context of those places.

In the end, and as a *post-operation*, the contents of these trees are dumped onto either LaTeX and/or HTML elements.

## 3 Conclusion

The power of structured documents resides not only in the ability to expeditiously generate several document formats but also in the possibility to perform contextual searches over them, similarly to what happens with databases.

Web publishing of ADB books completely eliminates any temporal or spatial obstacles, since its availability allows for consulting at any time and at *one-click* distance.

EAD (Encoded Archival Description) is a standard for representing inventory-like data, such as registers, indexes and other documents created by archives, libraries, museums or repositories of originals. Adapting the different DTDs of the already electronically published ADB books to EAD would be a stimulating challenge and a way of unifying all those grammars into a syntactically common and widely accepted scheme.

## References


[Cha] François Chahuneau. Current approaches to SGML up-translation. Technical report, A.I.S. S.A., 35, rue du Pont, 92200, Neuilly, France. http://www.oasis-open.org/cover/fcha.html.

[San01] José Luís Santos. Publicação Electrónica de Edições Esgotadas do Arquivo Distrital de Braga. Technical report, Universidade do Minho, 2001.

[VA86] Maria da Assunção Jácome Vasconcelos e António de Sousa Araújo. *Bulário Bracarense — Conjunto de Sumários de Diplomas Pontifícios dos Séculos XI a XIX Existentes no* ADB. Edições do Arquivo Distrital de Braga/Universidade do Minho. Arquivo Distrital de Braga/Universidade do Minho, 1986.