Journal of the Text Encoding Initiative

Issue 2 | February 2012
Selected Papers from the 2010 TEI Conference

A TEI P5 Manuscript Description Adaptation for Cataloguing Digitized Arabic Manuscripts

Mohammed Ourabah Soualah and Mohamed Hassoun

OpenEdition Journals

Electronic version
URL: http://journals.openedition.org/jtei/398
DOI: 10.4000/jtei.398
ISSN: 2162-5603

Publisher
TEI Consortium

Electronic reference
Mohammed Ourabah Soualah and Mohamed Hassoun, « A TEI P5 Manuscript Description Adaptation for Cataloguing Digitized Arabic Manuscripts », *Journal of the Text Encoding Initiative* [Online], Issue 2 | February 2012, Online since 03 February 2012, connection on 20 April 2019. URL : http://journals.openedition.org/jtei/398 ; DOI : 10.4000/jtei.398

A TEI P5 Manuscript Description Adaptation for Cataloguing Digitized Arabic Manuscripts

Mohammed Ourabah Soualah and Mohamed Hassoun

1. Introduction

1 Digitization presents an effective means to reveal manuscripts and other treasures long hidden in our libraries. Yet, simply providing images of the manuscript is not sufficient as the images do not convey many important aspects of the manuscript including

- codicological description (codex, binding, etc.)
- paleographic description (handwriting, etc.)
- manuscript transmission history

2 Although no single standard to catalog these types of features has emerged, TEI P5 Manuscript Description module (henceforth "TEI-ms") has come nearest to providing an adequate response to manuscript cataloguing problems as a whole. It offers a concise set of descriptive elements for encoding metadata about manuscripts. Nevertheless, TEI-ms is not as comprehensive as it might be as it is not particularly suited to describe Arabic manuscripts—a deficiency that this article addresses. We are thus advocating an expansion of the TEI-ms module to via a limited number of elements and attributes that would allow for a more accurate and unambiguous description of Arabic manuscripts.

2. Current Methods for Providing Online Access to Digitized Arabic Manuscripts

3 While it would be ideal for libraries to scan manuscripts in their entirety, providing users with access to high-resolution surrogates of the original, this alone is of limited use to

scholars because the text is not searchable. It is still not practicable for the majority of institutions to provide fully searchable versions of these manuscripts as the recognition of manuscript handwriting is in an embryonic stage. Until the time when providing full text versions of the manuscripts is feasible, manuscript cataloguing proves to be a realistic solution for providing access, and it has proved extremely successful in several Arabic manuscript digitalization projects.

4 The experience of the National Library of France (BnF) provides an interesting case in point. Indeed, the BnF is one of the pioneers in Arabic manuscripts digitization online publication. Their metadata is based on EAD (Encoded Archives Description) (Bosc 2006), but since EAD lacks elements needed for Arabic manuscripts, such as for recording the incipit (Guesdon 2008), it is insufficient.

5 There are other projects that also deserve to be mentioned. For example, the Bibliotheca Alexandrina has digitized and published online six manuscript catalogs describing a total of 4071 manuscripts (Ziedan 2010), on a site developed by a company called System Online. The bibliographic records are brief, containing only four parts: author information, the implicit content (Awâlu-hu), the explicit content (Âkhiru-hu), and the general condition of the manuscript (al-nuskha). This is insufficiently descriptive for many scholarly uses. Other Arabic manuscript cataloguing projects, such as a recent project of the National Library of Tunisia (BNT), are based on the UNIMARC format (Ghédira Dakhli 2010). At the BNT, catalog records were limited to two levels of description for expediency. Moreover, the fixed size of some MARC fields makes this format unsuitable. Indeed, the Wellcome Library in London realised the limits of MARC: after using MARC21, it decided to create TEI-MASTER metadata for the "Haddād Manuscripts" collection online project[1].

6 The weaknesses in description mentioned above are absent in TEI-ms. Indeed, TEI-ms was designed with medieval manuscript description in mind and contains an exhaustive set of descriptive elements for encoding metadata about manuscripts. The root element, `<msDesc>`, contains seven child elements: `<msIdentifier>`, `<head>`, `<msContents>`, `<physDesc>`, `<history>`, `<additional>` and `<msPart>`. Nevertheless, as we outline below, an expansion of the current element set is necessary to effectively describe Arabic manuscripts.

3. Adapting TEI-ms for Description of Arabic Manuscripts

7 The first aspect of the solution is a means to indicate transliteration (representing writing of one script in another script). Indeed, the Arabic language is written in a syllabic script of 28 letters whose vowels, long or short, are placed at the top or the lower part of the consonants, which complicates their representation in transliterated form (Hassoun 1987). For example, the transliteration of the Arabic name "إبن بطوطة" into Latin characters is "Ibnu baŢŢuŢa," but more characters are required in the Latin script than in the Arabic script.

8 The second aspect of the solution involves adding elements to TEI-ms required to describe ancient Arabic manuscripts while still respecting the general structure of TEI-ms.

3.1. Transliteration

9 Arabic script is used to write several languages: not only Arabic but also Turkish, Persian, Berber, and others. Transliteration is helpful when one does not know a language script but understands the language. Since vowel markers are rarely used in written Arabic, transliteration customarily involves insertion of the missing vowels. By providing transliterated catalogue entries much wider access to the records is possible. We thus propose to transcribe authority file elements such as name, subject, and title into the Latin characters, greatly facilitating access by these specific entries.

10 TEI uses the `@xml:lang` attribute to define the language of the content and descendants of a particular element, and this attribute may also contain a script subtag. For example, `xml:lang="ara-Latn"` defines Arabic language written in the Latin script. This notation is simple and very useful, but it presents a semantic weakness, because its form does not contain any information about the transliteration system. This solution can be implemented as follows:

11 The use of `<langUsage>` which is an element of `<profileDesc>` allows defining the language used in the TEI. We can use this element specificities to define the romanized form of a term as shown by the following example:

```
<langUsage>
        <language ident="ara">Arabic language</language>
        <language ident="ara-Latn">Arabic language with latin script</
language>
</language>
…
<msContents>
        <title xml:lang="ara"> الحضارة العربية القديمة </title>
        <title xml:lang="ara-Latn"> al-ḥaḍārah al-`rabīah al-qadīmah </
ref>
        <respStmt>
                <name xml:lang="ara"> عبد الرؤوف النمامشي </
name>
                <name xml:lang="ara-Latn"> `abd ar-ru'ūf an-namāmshī </
name>
        </respStmt>
</msContents>
```

12 To overcome this weakness, we propose to implement transliteration by using a specific model based on TEI components. It uses the `<ref>` element with the `@xml:lang`, `@target`, and `@type` attributes:

- `<ref>`: contains the transliterated form
- `@xml:lang`: on the target element defines the language of the element content; on the `<ref>` element, defines the transliteration system used
- `@target`: refers to a location where the term to transliterate is mentioned

- `@type`: for an instance of transliteration, takes the *"transliteration"* value

13 This can be seen in the following example:

```
<msContents>
        <title xml:id="Tr01" xml:lang="ara">
الحضارة العربية القديمة
        </title>
        <ref target="#Tr01" type="transliteration"  xml:lang="Latn">
                al-ḥaḍārah al-`rabīah al-qadīmah
        </ref>
        <respStmt>
                <name xml:id="Tr02" xml:lang="ara">
عبد الرؤوف النمامشي
                </name>
                <ref target="Tr02" type="transliteration" xml:lang="Latn">
                        `abd ar-ru'ūf an-namāmshī
                </ref>
        </respStmt>
</msContents>
```

14 The entire transliteration is carried out using a `<ref>` element, which can point to any TEI element.

3.1.1. Discussion

15 The first solution provides a very simple and effective automated processing of the text encoding form. Nevertheless, it remains incomplete, because of the absence of the main described concept which is the "transliteration". Indeed, this aspect is inferred by the structure of the text, but it is not implicitly quoted.

16 In addition, the expression "`<language ident="ara-Latn">` Arabic language with Latin script`</language>`", defined in the `<teiHeader>` part, considers the "ara-Latn" value as an identified particular language aspect—which is not correct in our case. That is why we propose another way which provides more and significant semantics to the transliteration.

17 Our proposed solution appears to be complex and may consume much time rather than the first one, described above, but it provides an implicit explanation of the described concept.

18 These examples show the richness and the power of the TEI to describe the same concept in two different ways. Our objective is not to show the relevance of a solution compared to another, but our principal concern is to highlight the concept of transliteration, which is still remains a concept not yet taken into account implicitly by TEI-Guidelines.

3.2. Adaptation for Different Manuscript Cataloguing Modes

19 Ancient Arabic manuscripts can be made up of several volumes. In this case, we speak about a multi-volume manuscript. Each volume represents a single component written by a particular author and covering a different subject than the others.

20 However, an important question has to be asked: How will this category of manuscript be catalogued? Two methods are proposed:

- *specimen cataloguing* (Gavin 2003): One and only one bibliographic record is associated with the entire manuscript. This cataloguing mode is appropriate to the TEI-ms structure, which uses the `<msPart>` element to describe each distinct volume.
- *volume cataloguing* (Prince Al-Saud Foundation 1990): A bibliographic record is associated with each volume of the manuscript. There are as many bibliographic records as volumes. This cataloguing mode unfortunately requires information to be repeated for each volume for a single manuscript, which is contrary to cataloguing principles.

21 Since we desire to describe each volume of a manuscript without repeating information, we propose that information common to all volumes (generally speaking, the manuscript description) be shared among all volumes and not repeated.

22 **Example:** If the author is the same for all volumes in a manuscript, then author information will be described only in the main bibliographic record. The others simply to refer to it (see fig.1).

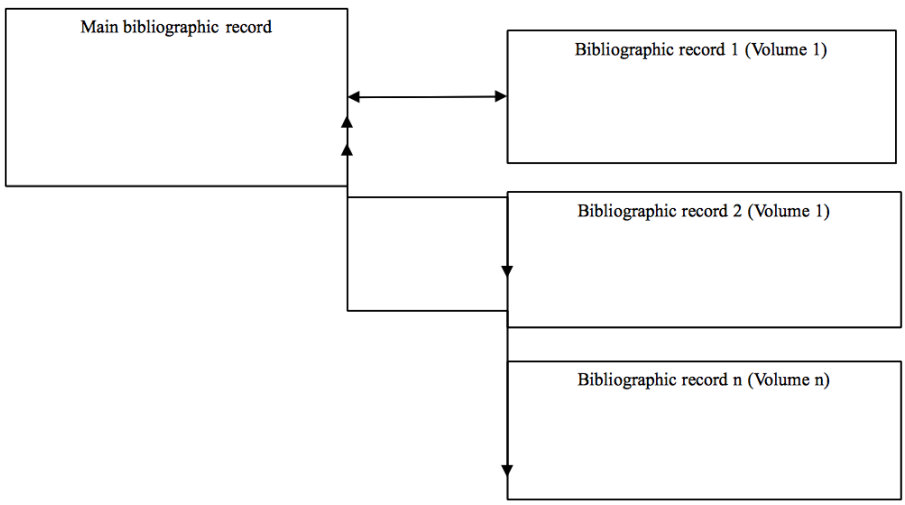

Figure 1: Multi-volume manuscript bibliographic record structure.

23 The `<msPart>` element is useless for the volume cataloguing mode. That's why we propose to instead use the `<idno>` element with the `@type` attribute, which takes one of two values:

- *main*: representing the main record
- *vol*: representing any manuscript volume

24 Example: We consider any multi-volume manuscript identified by the number 34.

1. The main bibliographic record will be defined by: `<idno n="34" type="main"/>`
2. The first volume will be defined by: `<idno n="34" type="vol">1</idno>`

3.3. Author

25 The TEI defines <persName> element to encode a personal name. It may contain several other elements for the components of a name: <surname>, <forename>, <roleName>, <addName>, <nameLink> and <genName>. Here is an example from the TEI Guidelines:

```
<person xml:id"HOC001">
        <persName>
                <surname>Hoccleve</surename>
                <forename>Thomas</forename>
        </persName>
        <birth notBefore="1368"/>
        <occupation>poet</occupation>
        <!-- other personal data -->
</person>
```

26 This example shows that the author's name is formed by a surname and a forename. Although western medieval manuscripts are suitable for this kind of description, this is not true for classical Arabic names.

3.3.1. Classical Arabic Name Structure

27 Classical Arabic names are generally, composed of six different elements: the patronymic, the ism (إسم), the laqab (لقب), the kunyah (كنية), the khitab (خطاب) and the nisbah (نسبة). Each is defined below:

- **Patronymic**: introduced by "Ibn" (إبن - son of);
- **Ism**: a given name;
- **Laqab**: a word or an expression applied to an eminent person to evoke a real or an assigned quality (Khairy 2006). Sometimes the laqab is used as a forename and at other times it's used as a "celebrity name";
- **Kunyah**: a mark of distinction applied to prominent figures to honor them. For example, "Abū-Yūsuf" (أبويوسف) is often used for someone who is called Ya'qūb (يعقوب) (Khairy 2006);
- **Khitab**: an honorific name, which is usually ended by the suffix al-Dīn (الدين) (Khairy 2006);
- **Nisbah**: an adjective formed by using the suffix ī in order to indicate the person origin, his birth place, or his residence. It represents the relationship name, which can be a genealogical, political or ideological affiliation of a person. Example: the mathematician "Al-khawarizmii" refers to "Khuwarism", name of a place located in Central Asia.

3.3.2. TEI-ms Adaptation for Describing Arabic Names

28 Our goal is use TEI elements to encode the components of classical Arabic names when available and create new elements only when necessary. Consequently, we propose adaptations as follows:

- **<surname>**: used to define the *laqab*. The use of this element to define the *laqab* is justified by the similarity that exists in using the *laqab* in the Arabic old names and <surname> in western naming model. Moreover, the @type attribute could take different values depending on the signification that represents the *laqab* in the name. @type takes the following values:
  - **"celebrity"**: defines the celebrity name;
  - **"laqab"**: used when the name is given relative to the person's characteristics;
- **<forename>**: used to define the ism'
- **"ism"**: used to define the person name which is given at birth.

29 The other components of the old Arabic names (*patronymic name*, *kunyah*, *khitab*, and *nisbah*) can be defined by using the <addName> element, as in the OCIMCO[2] project, with a @type attribute, which may take *kunyah*, *khitab*, or *nisbah* as its value.

30 The *nisbah* may a geographical qualifier, a political qualifier, or any other word about the ideological affiliation of person. To take into account this case, we propose to use the @subtype TEI attribute within the <addName> element. @subtype attribute defines the real context of the nisbah. It may have several values such as "political", "ideological", etc.

31 **Example:**

```
<addName type="nisbah" subtype="place">
        Al-khawarizmii
</addName>
```

32 Note that each name element defined above would also have a transliteration.

33 **Example:**

```
<author>
        <persName>
                <forename xml:id="Tr03" type="ism" xml:lang="ara">
 الحسين
                </forename>
                <ref target="#Tr03" type="transliteration"
xml:lang="Latn">
                        al-Husayn
                </ref>
                <addName xml:id="Tr04" type="kunyah" xml:lang="ara">
إبن سينا
                </addName>
                <ref target="#Tr04" type="transliteration"
xml:lang="Latn">
                        Ibnu Sīna
                </ref>
        </persName>
</author>
```

3.4. The Copyist

34 The Arabic manuscripts codicology attaches great importance to the copyist. He is considered alongside the author and is often one of his disciples. The disciple transcribes the teacher's knowledge through various processes, such as:

- Attribution of transcription license (iğāza);
- Listening to the teacher's speech (sama'a);
- Ideas confrontation (muqabalāt);
- Reading (qāri).

35 Hence, the copyist is a very valuable intellectual source. TEI-ms does not contain a way to record the copyist. But, it defines the `<docAuthor>` (document author) element, which contains the author of the document.

36 This element can also be used for the role of copyist, with the `@type` attribute distinguishing the copyist roles: "*listner*" (sama'at), "*license*" (iğaza), "*reader*" (qarī) or "*comparison*" (muqābalāt).

3.5. Logical Structure of the Manuscript

37 Arabic manuscripts deal with different topics: theology, astronomy, grammar, etc. The manuscript structure differs from one subject to another. Arabic manuscripts are structured into *kytab* (book), *juz* (part), *bab* (door) and *fasl* (scene). This ranking has no equivalent in the western documents structure.

38 Note that this classification is not widespread. We find some manuscripts with an undefined classifying system, while others present only one or two logical structure

criteria (for example, some present only the *bab* structure while others present the *bab* and *fasl* structure simultaneously).

39 In order to describe the Arabic manuscript logical structure, the <rubric> element, which is a <msItem> sub-element, is the best candidate. Moreover, in its current form, the <rubric> has two useful attributes: @type and @n, which will indicate respectively the logical structure (*kytab*, *juz*, *fasl*, or *bab*) of the manuscript and the associated number of the relevant heading in the body of the manuscript.

40 **Example:**

```
<msContents>
        <msItem n="1">
                <locus>fol.23r</locus>
                <rubric type="kytab" n="1">علم الفلك</rubric>
                <locus>fol.54r</locus>
                <rubric type="kytab" n="2"> الطب</rubric>
        </msItem>
</msContents>
```

41 Thus, the <rubric> sub-elements combination may give rise to the automatic generation of a table of contents even without encoding divs in the body of the manuscript.

3.6. The Manuscript History

42 The <acquisition> element is a sub-element of <history>. It contains useful information about the reason that the manuscript is held within the conserving institution. For example, the manuscript may have been acquired by purchase, gift, loan, *waqf*. The purchase and donation procedure permit the institution to become the rightful owner of the manuscript. The loan procedure is usually done for a digitization or a special study on the manuscript. In this case, it would be interesting to identify the owner.

43 Sometimes the presence of the manuscript in the institution is the result of a *waqf*. This term is a synonym of *habs* in Arabic: it means a legacy property of a Muslim or Christian institution (Troupeau 2002). The usufruct may remain with the manuscript owner. Often the *waqf* statement prohibits the manuscript from being sold, purchased, or moved from the institution (Eddé and Geoffroy 2002).

44 The TEI's <acquisition> element includes a prose description of the acquisition within a <p> element. This structure encourages a wordy description, which makes creation of an authority list difficult. To overcome this problem, we can use the @ana attribute that allows marking the element to be interpreted. This as shown by the following example:

```
<acquisition ana="#Ms01">
…
<interp xml:id="Ms01">loan</interp>
```

45 The <interp> element is used to indicate the acquisition type of the manuscript. Within the <interp> element, we can find one of the following: "purchase," "gift," "loan," and "*waqf*."

3.7. Writing Style

46 The calligraphic appearance in TEI-ms is described by the <handNote> element, which is a sub-element of <physDesc>. <handNote> contains few attributes directly concerning the writing style: just @scribe, @script, @scope, and @medium. None of these attributes provide information about either the writing's clarity or its quality.

47 Moreover, calligraphy is considered an art, so particular interest is accorded to the style of calligraphy. Since TEI-ms does not include any appropriate elements to describe the calligraphy, we can use the @ana attribute on the <handNote> element and the <interp> element to define the quality of the writing style. Within the <interp> element, we can find one of the following: "bad", "medium", "good", "readable", and "unreadable".

4. Conclusion

48 In this work we show that the Manuscript Description module, with modifications, can be used for both Western medieval manuscripts and ancient Arabic manuscripts. It accurately respects protocols for describing Arabic manuscripts proposed by the various specialized agencies and consortiums. It will make it possible to generate bibliographic records adapted to the needs of users.

49 Thus, it facilitates access to digitized manuscripts which are not in a machine-readable form. TEI-ms, plus our additions, offers several elements and attributes allowing the description of some complex specific aspects of ancient Arabic manuscripts.

50 However, we note complexity and semantic weakness of some aspects of our suggested solution. Indeed, using the <ref> element for the transliteration and using the @ana attribute to specify the manuscript acquisition type or the writing style quality are cumbersome and not especially transparent to readers of XML.

51 We think that the longevity of a system is related to its adaptativeness to an evolving environment. Consequently, it is important to be able to create new elements within the TEI. With this in mind, we propose the following:

- Dealing with the transliteration in an implicit way;
- Creating new elements close to semantic reality of the studied field (for example, a <transliteration> element and a copyist element which may be used in the same terms like the author element);

- Adding new attributes to some elements (for example, the `@type` attribute on the `<acquisition>` element to assign a term from a taxonomy of manuscript acquisition types.

52 The real purpose of these suggestions is to enlarge the semantic field encompassed by the TEI Guidelines and, in particular, the Manuscript Description module.

BIBLIOGRAPHY

Prince Al-Saud Foundation. 1990. *al-Makhṭūṭāt al-ʿArabīyah fī al-Gharb al-Islāmī: waḍʿīyat al-majmūʿāt wa-āfāq al-baḥth*. al-Dār al-Bayḍāʾ: Wallādah: Muʾassasat al-Malik ʿAbd al-ʿAzīz.

Troupeau, Gérard. 2002. “Les actes de waqf des manuscrits arabes chrétiens de la Bibliothèque nationale de France.” In: “La tradition manuscrite en écriture arabe,” ed. Geneviève Humbert. Special issue of *Revue des mondes musulmans et de la Méditerranée*, 99–100 (November 2002) 45–51. http://remmm.revues.org/document1173.html.

Hassoun, Mohamed. 1987. “Conception d’un dictionnaire pour le traitement automatique de l’arabe dans différents contextes d’application.” PhD diss., Université Lyon 1.

Eddé, Anne-Marie, and Marc Geoffroy. 2002. “Livret de stage d’initiation au manuscript medieval en langue arabe.” AEilis, Publications Pédagogiques, 3. http://aedilis.irht.cnrs.fr/stage-arabe/

Khairy, Iman. 2006. “Le contrôle d’autorité des noms arabes de la période classique à la Bibliotheca Alexandrina.” Paper presented at the annual conference of MELOM International, Alexandria, Egypt, May 23–25, 2005. Translated by Annick Bernard. http://www.sant.ox.ac.uk/ext/melcomintl/melcom/khairy_french.pdf.

Guesdon, Marie-Genvière. 2008. “Bibliothèque nationale de France: Manuscripts catalogue ‘Archives et manuscrits’”. Paper presented at the Fourth Islamic Manuscript Conference, Cambridge, England, July 6–9. http://www.islamicmanuscript.org/files/GUESDON_Marie_2008_TIMA.pdf.

Bosc, Orélie. 2006. “Du catalogue médiéval à l’EAD: 10 siècles de catalogage des manuscrits dans les bibliothèques...” Paper presented at Manuscrits dans tous leurs états, National Library of France (BnF), September 27. http://www.bm-orleans.fr/userfiles/file/portail/manuscrit_2b.pdf.

Gavin, Pierre. 2003. “Catalogage du livre ancien.” Course presented at Sistema bibliotecario ticinese, Bellinzona, Italy, October 13–14. http://www.pierregavin.ch/cours/cours/,

Ghédira Dakhli, Sihem. 2010. “L'usage d'UNIMARC à la Bibliothèque Nationale de Tunisie.” Paper presented at 3rd UNIMARC Users Group Meeting, March 31, Villeurbanne, France. http://www.enssib.fr/bibliotheque-numerique/notice-48449/

Ziedan, Youssef. 2010. *Manuscript Libraries Catalogs*. http://ziedan.com/English/index.asp/

NOTES

1. Available at the following URL: http://library.wellcome.ac.uk/node273.html.

2. OCIMCO: *Oxford and Cambridge Islamic Manuscript Catalogues Online* project, available at URL: http://www.bodleian.ox.ac.uk/bodley/library/specialcollections/projects/ocimco

---

ABSTRACTS

It is incumbent upon libraries holding Arabic manuscripts to provide access to digitized surrogates of their holdings. Users require access both by authority list and by content. Thus, an exhaustive cataloguing method is essential. The TEI P5 Manuscript Description module is a suitable tool for manuscript cataloguing, but it lacks certain features that would allow for exhaustive description of ancient Arabic manuscripts. In this article we make several suggestions that would augment the TEI P5 Manuscript Description module allowing for a richer and more accurate description and cataloging of ancient Arabic manuscripts.

INDEX

Keywords: Arabic manuscripts, cataloguing, digitization, manuscript description, P5

AUTHORS

MOHAMMED OURABAH SOUALAH

Equipe de recherche de Lyon en Information et Communication (ELICO), Lyon, France

MOHAMED HASSOUN

Professeur des universités at ENSSIB Lyon ELICO, France