Library Hi Tech

Dynamic cataloguing of the old Arabic manuscripts by automatic extraction of metadata

Mohammed Ourabah Soualah, Yassine Ait Ali Yahia, Abdelkader Keita, Abderrezak Guessoum,

Dynamic cataloguing of the old Arabic manuscripts by automatic extraction of metadata

Mohammed Ourabah Soualah and Yassine Ait Ali Yahia
Ecole Supérieure en Informatique, Algiers, Algeria
Abdelkader Keita
University of Mali, Bamako, Mali, and
Abderrezak Guessoum
Université Saad Dahlab de Blida, Blida, Algeria

Received 16 April 2016
Revised 8 February 2017
Accepted 16 February 2017

Abstract

Purpose – The purpose of this paper is to obtain online access to the digitised Arabic manuscripts images, which need to use a catalogue. The bibliographic cataloguing is unsuitable for old Arabic manuscripts, and it is imperative to establish a new cataloguing model. In the research, the authors propose a new cataloguing model based on manuscript annotations and transcriptions. This model can be an effective solution to dynamic catalogue old Arabic manuscripts. In this field, the authors used the automatic extraction of the metadata that is based on the structural similarity of the documents.

Design/methodology/approach – This work is based on experimental methodology. The whole proposed concepts and formulas were tested for validation. This, allows the authors to make concise conclusions.

Findings – Cataloguing old Arabic manuscripts faces problem of unavailability of information. However, this information may be found in another place in a copy of the original manuscript. Thus, cataloguing Arabic manuscript cannot be done in one time, it is a continual process which require information updating. The idea is to make a pre-cataloguing of a manuscript, then try to complete and improve it through a specific platform. Consequently, in the research work, the authors propose a new cataloguing model, which the authors call "Dynamic cataloguing".

Research limitations/implications – The success of the proposed model is confronted with the involvement of all actors of the model. It is based on the conviction and the motivation of actors of the collaborative platform.

Practical implications – The model can be used in several cataloguing fields, where the encoding model is based on XML. The model is innovative and implements a smart cataloguing model. The model is useful by using a web platform. It allows an automatic update of a catalogue.

Social implications – The model prompts the user to participate and enrich the catalogue. The user could improve his social status from a passive to an active.

Originality/value – The dynamic cataloguing model is a new concept. It has never been proposed in the literature until now. The proposed cataloguing model is based on automatic extraction of metadata from user annotations/transcription. It is a smart system which automatically updates or fills the catalogue with the extracted metadata.

Keywords Transcription, Digital library, Annotations, Automatic extraction of metadata, Dynamic cataloguing, Structural similarity

Paper type Research paper

1. Introduction

Arabic manuscript cataloguing has become an exciting field (Feodorov, 2006), and there are several projects on the digitisation of Arabic manuscripts where the images were accessed based on the catalogue. These projects aim at providing efficient online access to the digitised resources. For example, the excellent works produced by the Bibliothèque Nationale de France and by the Bodleian Library (Oxford University) in collaboration with the Bibliotheca Alexandrina (Egypt) are well known. These solutions present comfortable

Library Hi Tech
Vol. 35 No. 2, 2017
pp. 251-270

0737-8831
DOI 10.1108/LHT-07-2016-0076

and intuitive interfaces, but all of them present some weaknesses related to the access mode. For instance, their notices fail to describe the manuscript content.

These institutions catalogue the digitised manuscripts using the following access modes: free access dealing with the catalogue content and authority-list access.

The second access mode poses several problems caused by the lack of either the authors' information or specification of the manuscripts. Sometimes, the same manuscript presents several subjects which make using the headings-subject access mode difficult. These aspects make it difficult to catalogue the manuscript using the authority list (Soualah *et al.*, 2012). This access mode is inspired from a classical cataloguing model, which is called bibliographic cataloguing:

> The bibliographic cataloguing of the old manuscripts consists of describing the manuscript in three aspects: codicological aspect (material), palaeographic aspect (content) and historical aspect (manuscript possessions).

Bibliographic cataloguing is based on the principle that the described document is stable. With regards to the stability of the document, the document perception stays relatively the same for various users over time. Thus, there is no new information added to the bibliographic record of a catalogued document. Consequently, the cataloguing model is related to the nature and stability of the document.

The manuscript is unique, and its cataloguing process is based on well-defined metadata (codicological, palaeographic and historical). The philologist is interested in the manuscript content and its authenticity (Auerbach, 1961), and his investigation may require a long period to perform the research. Moreover, the manuscript cataloguing may be incomplete; for example, the some folios may be missing the manuscript title or its author. However, this information can be present in other copies of the manuscript preserved elsewhere.

Therefore, the problem is determining how this information can be integrated into the catalogue. To solve it, we propose an annotation and transcription method that uses online-digitised manuscripts. An interface is provided to the user who can make his/her comments, complete the metadata or transcribe the manuscript. Once the mediator validates the inputted data, the system extracts information from the annotation documents or the transcription documents and automatically updates the catalogue. We call this procedure "dynamic cataloguing". This procedure shows that the cataloguing of old Arabic manuscripts is scalable and gives the catalogue a dynamic aspect.

1.1 The use of annotations and transcriptions for cataloguing manuscripts

Technology has become very developed and information has become available everywhere at every time. However, Arabic manuscript cataloguing faces many problems regarding information unavailability resulting from the manuscript description based on dynamic metadata.

In our paper, we propose a new cataloguing model of old Arabic manuscripts. This cataloguing model requires a permanent update of the catalogue, and the implementation of such model requires the following:

- online publishing of images from a database of the digitised manuscripts;
- implementing the access and navigation tools for the manuscripts images database;
- categorising annotators and transcribers (such as experts, researchers and scholars);
- validating annotation strategy by the mediator; and
- implementing the extraction model of the metadata for improving the catalogue.

Our paper sequentially discusses the items above and shows the need to overcome the difficulties of the bibliographic cataloguing model and proposes the dynamic cataloguing model.

2. Which digital library model for digitised Arabic manuscripts?

The digitised manuscripts require organisational and technical strategies. It is very important to establish a physical structure that offers both enhanced security and storage scalability because of the permanent acquisitions of the manuscripts.

The solution is based on a digital library where the resources are the digitised manuscripts images. The classical structure of digital library is shown in Figure 1.

The valuable task is determining which image-storage technology to use: is using the file management system (FMS) technology better than using the database management system (DBMS)? Each solution presents advantages and disadvantages. For instance, the FMS provides a storage space that tends to infinity and the DBMS presents a reliable security system. Moreover, the problems of portability, compatibility of versions and fragmentation do not encourage the use of DBMS.

There is no storage system standard for implementing a digital library which is focussed on images (Deschatelets and Dufour, 1997). Generally, digitised resources are stored in file servers and in removable storage systems (UKOLN (Université de Bath), 2004).

Moreover, the concept of a digital-manuscript library should be defined, because implementing a digital library requires a radical change in the existing organisational structure. In addition, constraints imposed by using the classic catalogue (manual or automatic catalogue) are surpassed by information-retrieval technology. This property encourages reconsidering the concept of cataloguing. Consequently, we propose to establish a new cataloguing system in this work.

2.1 The digital library for digitised manuscripts

After the apparition of the internet, several works have attempted to provide a suitable definition of a digital library. We only mention those adapted to our case. For example, Ferrand (1996) states that an electronic library is needed to computerise the catalogue and to digitise the documentary fund.

There is no standard model of a digital library. For instance, Deschatelets and Dufour (1997) and Gonçalves *et al.* (2004) proposed two different models; these models present different forms but a very close background.

The first model is based on data, metadata, services, users and management structure. The second one is based on streams, structures, spaces, scenarios and society (5S model).

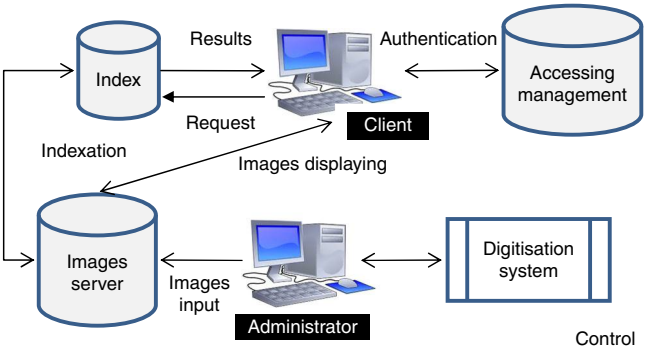

Figure 1.
Classical structure of a manuscript's digital library that uses a database management system

The difference between the two models is the contribution of the second model that resides in the spaces that are used to specify the content of the library.

We are inspired through these two models to design our own digital library model. It is based on three principle elements: digitised resources, access tools and actors. Additionally, these three concepts are defined as follows:

(1) Digitised resources (spaces): digitised manuscripts stored on file servers.

(2) Access tools (metadata, services): first, we mention the catalogue which is the backbone of the solution and then the other tools such as index, access procedures, interfaces and coordination programmes.

(3) Actors (such as users, librarians, administrators, researchers and scholars): different persons do specific tasks to make the system operational. For example, the librarian is a mediator and responsible for the cataloguing, the administrator manages the system and the users, and the researchers and the experts annotate or transcribe manuscripts and enrich the catalogue content.

2.2 Accessibility of digitised images

When cataloguing, a link is created between the bibliographic record and the described manuscript. In contrast, when indexing, terms are extracted from the bibliographic record and stored in the inverse file.

The indexation is done through two modes:

(1) Headings list indexing consists of browsing the catalogue, which is in XML format, and then extracting the terms corresponding to the headings list (author, title, and subject).

(2) Free indexing provides users with a full-text access mode. Thus, all the catalogue's terms are extracted and stored in the inverse file.

Once a word is selected in the index, it permits access to the bibliographic record. Afterwards, the hyperlink contained in the selected bibliographic record is used to load the manuscript image. The browsing of images is done because of a specific viewer which offers enlargement and scrolling methods and annotation interfaces.

3. Problems met in cataloguing old Arabic manuscripts

Because of the cataloguing concepts, the catalogue usage seems sufficient to identify and access digitised manuscripts. However, during the cataloguing process of the old Arabic manuscripts, the cataloguer faces several problems. For example, artisanal artwork manuscripts are many centuries old and have had many owners. Therefore, they are subject to several damage threats such as physical erosion and content erasure. We can also include other damages caused by humans such as the removal of parts or the storage of the manuscript in inappropriate environments.

Therefore, the manuscript can be found incomplete by not having an author, title or edition date. These problems are made more complex by the anarchic gathering of the manuscript. Thus, the main problem of cataloguing the old Arabic manuscripts is the unavailability of information; this work presents a solution to this difficult problem.

4. The genesis of dynamic cataloguing

The missing information may be found in copies of the manuscripts which can be found in other places (such as country and library). When this occurs, it is imperative to upgrade the catalogue content. Thus, cataloguing the old Arabic manuscripts cannot be made at once, as

it is always upgraded following new or updated information which is integrated into the catalogue, which produces dynamic cataloguing.

The results of our research have shown the instability of old Arabic manuscripts, and this concept will be developed in a future paper. The dynamic cataloguing model is spawned by the instability of a catalogued document. However, the dynamic cataloguing requires a collaborative environment that allows qualitative input information, which is based on the annotation and transcription of the images of old Arabic manuscripts.

5. The structure of collaborative platform

The collaborative platform of the annotations and transcriptions of digitised manuscripts is established by several actors: annotators, transcribers, scholars and a mediator. The latter is the main animator of the platform, and she or he validates the inputs, contacts other actors and manages the whole system.

The collaborative platform aims at obtaining useful information to automatically upgrade the catalogue. The architecture of the solution is based on two systems:

- System 1 – collaborative platform: it manages the annotation and the transcription documents relating to a specific metadata.
- System 2 – automatic extraction of metadata: the role of this second system is to automatically upgrade the catalogue. The system compares the annotation/transcription document to the catalogue content. To do this, it uses the structural similarity between two documents.

Figure 2 shows a general view of the dynamic cataloguing model:

(1) Annotation and transcription of a digitised manuscript. The annotation/transcription document is stored inside a specific database.

(2) The mediator validates the annotation/transcription document, which will be the candidate for updating the catalogue.

(3) Comparison of the annotation/transcription document with the catalogue content by using the structural similarity between two documents.

(4) Extraction of the specific metadata which will be integrated into the catalogue.

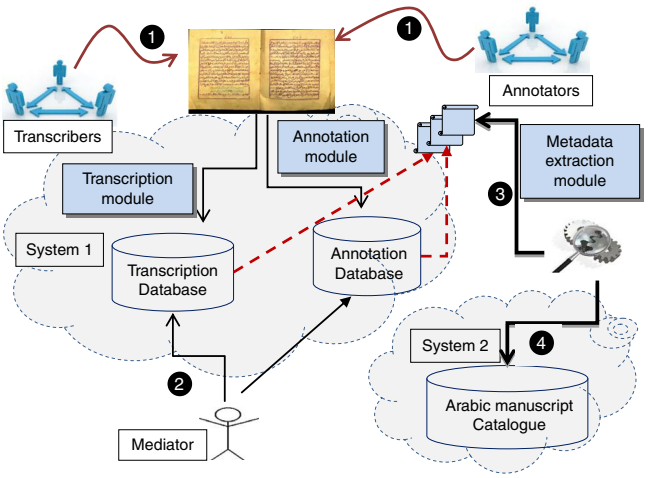

Figure 2.
Dynamic cataloguing model

5.1 The annotation of digitised old Arabic manuscripts

There is no formal definition for the annotation concept (Azouaou *et al.*, 2003), because the annotation definition depends on its assigned objective. In our work, the annotation of the manuscripts' images is defined as a textual block associated to a specific area of the selected image, and librarians define the annotation as a reader activity, which consists of graphical or textual marks on a document following objectives (Azouaou *et al.*, 2003). The annotation may be a comment, a translation, a transcription, a correction or additional information regarding a selected zone.

This definition can be extended to the digitised documents, in which the annotation objective provides a method for extracting the metadata and the optimising access to the electronic resources.

Deschatelets and Dufour (1997) defined the annotation as graphical or textual information attached to a document which the annotation is often placed within the same document. This definition highlights the need to link the annotation to the annotated document. However, generally, the annotation is used as an indexing media without any semantic aspect.

Thus, the annotation management system requires annotation formalisation. This formalisation aims at defining the annotation purpose and structure, which the latter is used to locate the annotation position on the digitised document and to specify the content and the annotation form.

The annotation objective is the main part of the annotation activity, because the objective defines the annotation rules. Thus, it becomes possible to specify the annotation restrictions, and what introduces the semantic to the annotation model. For example, it becomes possible to help and guide the annotator through the restrictions by displaying specific input fields. This aspect makes classifying the annotation easy in the encoding model.

5.1.1 Various usages of annotations in digitised manuscripts field. Several works use annotations to manage digitised manuscripts for different goals. The main objective of those works is to provide an efficient system to access the manuscript database. For instance, Al-Rajebah and AL-Khalifa (2012) implemented an annotation system for Arabic manuscripts based on the linked data. The authors' solution is based on the conversion of metadata into a linked metadata model. Afterwards, the manuscript is connected to Wikipedia through a specific field. The project is called Makhtota +, which aims to use a social network to access the digitised Arabic manuscript.

Agosti *et al.* (2005) proposed another annotation model of the illuminated manuscripts. The aim of the work was to manage image digital archives in methods that offer the professional final user support for annotating their content. The authors saved the relations between the whole annotations to keep the trace of the evolution of scientific representations.

Sanderson *et al.* (2011) used the digitised manuscript annotations to create a single shared model for description of manuscripts, and the model is called the sharedCanevas. The objective was to provide every manuscript repository various digital-manuscript collections. To reach this goal, the authors used the linked data concept that facilitates the implementation of interoperability solution between manuscript repositories. Such solution allows users to navigate within the institutional silos of manuscripts images and text rather than to use cross-collection capabilities.

There are other works that try to increase access to manuscripts by using annotation. For example, El Makhfi *et al.* (2011) associated the annotation to the manuscript images. The process of identification and accessing to the manuscript database is provided by a search engine based on metadata and annotations.

5.1.2 Position of our research work. Our work is based on the annotation that is used to complete the empty fields (metadata) within the bibliographic record and to input additional information within the catalogue to improve access to the manuscripts.

These aspects show the existence of two types of annotations:

(1) the annotation related to the metadata of manuscript cataloguing; and

(2) the annotation related to the manuscript content.

We note that in our case, the old Arabic manuscripts are encoded in XML format using the TEI P5 manuscripts description standard (TEI-ms). Thus, the manuscript is described through seven meta-metadata categories:

(1) < msIdentifier > : it defines the needed information to identify a manuscript.

(2) < head > : it is the header.

(3) < msContents > : it describes the intellectual content of the manuscript.

(4) < physDesc > : it contains the physical description of the manuscript.

(5) < history > : it contains elements describing the history of the manuscript.

(6) < additional > : it contains additional information about the manuscript (e.g. bibliography).

(7) < msPart > : it describes other manuscripts assembled into a single one.

Each level contains sub-elements which define a specific aspect of the manuscript.

The main idea of our research work is to exploit annotations and transcriptions of Arabic manuscripts to automatically update the catalogue. The process seems to be easy to perform, but its implementation uses several concepts. For instance, the process begins by storing the annotations and the transcriptions in specific documents which are linked to the manuscript images. Additionally, every annotation or transcription document is compared to the bibliographic records. Thanks to the concept of structural similarity, specific metadata is extracted from the annotation or transcription documents and automatically added into the catalogue.

5.1.3 Arabic manuscripts particularities. A manuscript is defined as an artisanal artwork which is manually done and handwritten. The Arabic manuscripts present some particularities in their physical structure and their logical organisation. For instance, concerning the physical structure, Scheper (2015) indicated in a recent publication some differences that exist between Arabic manuscripts and western manuscripts. These differences are in the binding structure and in the paper quality. The author says "tactility has been the most intriguing aspect of the Arabic manuscript in the Oriental Collection; they are very different from Western books". Furthermore, she adds "The leather which is used to cover the bindings is quite different from the leather used on the Western bindings".

In addition, concerning the logical organisation aspect, we showed the particularities of the Arabic manuscripts during the cataloguing process in Soualah and Hassoun (2012). For example, metadata cataloguing protocol has specific features compared to western manuscripts such as the following:

- the cataloguing mode can be done in two ways: specimen cataloguing mode (Soualah and Hassoun, 2011) or volume cataloguing mode (Soualah and Hassoun, 2012);
- the author's name;
- the copyist specifications;
- the logical structure;
- the manuscript history; and
- the writing style.

These different specificities have led to the need of adapting the TEI-ms for the description of ancient Arabic manuscripts.

5.1.4 Annotation-system workflow. Each annotation is associated to a specific metadata. This provides the annotation with a semantic that is related to the metadata. However, the annotations associated to the manuscripts contents may be pertinent, but they are not associated to any explicit metadata. In addition, they are classified per their category (such as transcription, translation and comment).

Some annotations are not associated to any explicit metadata of the cataloguing protocol, and in this case, they are classified per their category. In our model, each annotation represents a specific document related to a specific bibliographic record. The annotation must be validated by the cataloguer (the mediator) before its integration into the catalogue. Thus, a collaborative platform is necessary to create interactions between the different actors. For example, the librarian can ask for details or invite another annotator to give his or her opinion regarding an annotation. Therefore, a dynamic relationship is created between the different actors, and this makes the information more efficient and more qualitative.

Figure 3 shows the role of each actor and the system workflow.

5.1.5 The modelling of the annotation system. The objective is to associate semantics to the index. This becomes possible because of the catalogue structure and the direct link existing between the catalogue and the annotations.

For instance, on one hand, the catalogue structure is provided with a semantic content, and each element corresponds to a specific metadata that defines an aspect of the manuscript. On the other hand, the annotation is directly related to a tag representing an informational category which provides meaning for each annotation.

Therefore, each annotated zone of the digitised manuscripts is well defined. This is possible because of an interactive interface between the annotating system and the user. In addition, the appropriate interfaces assist the user when he or she enters the annotations into the system. Once finished, a suitable tag is selected and associated to the annotation.

The assistant facilitates the annotation classification and allows the mediator to control the inputs. Additionally, additional functionalities can be used with the annotation such as the following:

- the transcription;
- the trace of the user navigation in the manuscript images; and
- the classification of the annotated documents per the user profile.

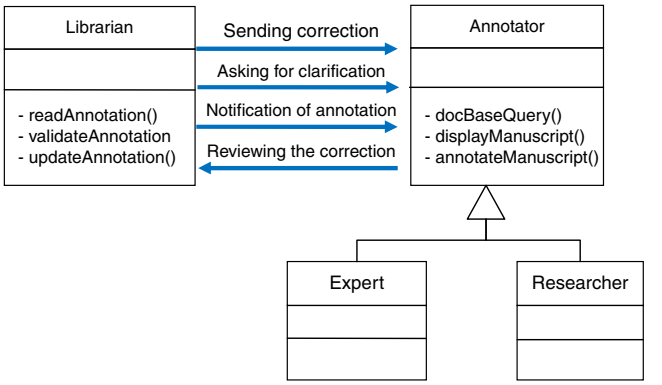

Figure 3.
Annotation-system workflow

The annotations constitute a cluster related to a specific bibliographic record and the updates of the annotation form a chronological chained list. This structure allows the system to perform the search through the concepts associated to the annotations.

Our annotation modelling is based on the association of the annotation to the entire information system. This task is performed thanks to the catalogue. Consequently, the modelling strives to improve the catalogue functionalities. The catalogue is no longer a straightforward descriptive object of the manuscript, but it becomes a real metadata repository of the manuscripts.

Figure 4 shows the knowledge structure which associates the catalogue to the annotations.

Old Arabic manuscripts are sought by scholars and are studied for several reasons such as material aspects, manuscript content and critical edition. This generates additional information regarding the manuscript and represents a remarkable advantage for increasing access points to the digitised manuscripts and for the evolutionary catalogue.

It seems that the cataloguing procedure is performed just once for any document. This is true when the document is stable, for example, the cataloguing of monographs. Unfortunately, this cannot be applied for Arabic manuscripts. Since the manuscript is an archaeological document, it can be studied per several aspects over time, and the viewpoints vary from a scholar to another, which they can even be contradictory.

Thus, the contents of bibliographic records change over time: information is added, deleted or corrected. Thus, the manuscript is an unstable document. In addition, the cataloguing of the same manuscript is continually performed, thus introducing the concept of dynamic cataloguing.

The system knowledge space contains the whole annotations that are entered by different users over time.

Figure 5 shows the model of cataloguing old Arabic manuscripts.

5.1.6 The users categorisation. The relevance of the annotation depends on the nature of the user. The annotator profile is a very important parameter, because it can be used to the manuscripts classification. Thus, we classified the users into two categories:

(1) The visitors: persons who have occasional access to the database of the manuscript. The users of this category have no right to use the annotation tools.

(2) The subscribed users: persons registered by the system. They may be members or remote users that have previously completed the application form. The following user information is required:

- identification of the annotator (unique reference);
- name, surname, address, e-mail, country, birth date and civility;

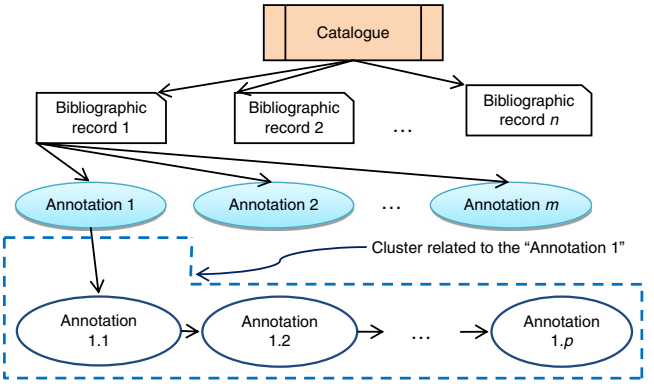

Figure 4. Annotation-system structure

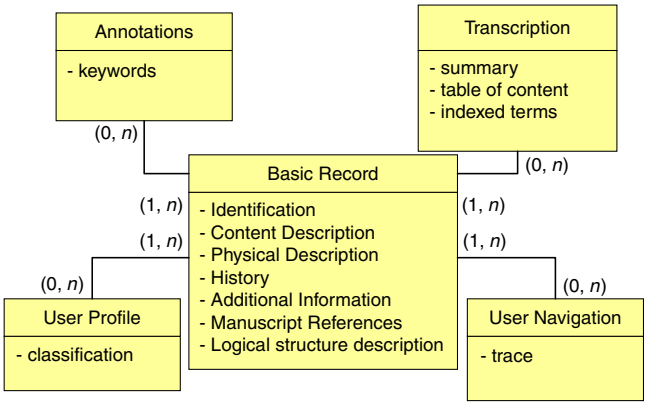

Figure 5.
Digitised Arabic-manuscripts cataloguing model

Source: Soualah *et al*. (2012)

- activity (such as scholar, student and expert); and
- personal preferences (such as anthropology, history and theology).

After the subscription, the system creates the user account that must be used in every communication on the platform.

The librarian is responsible for the annotation validation, and each e-mail contains the manuscript identification and the image references. Thus, the system considers all the information referring to the manuscript as an annotation.

The user profile is a parameter that can greatly improve the quality of the information-retrieval system, but this paper does not give more details of this parameter.

5.1.7 The validation of the annotations. The librarian is the main actor of the system. He or she is the responsible for the integration of the annotations into the system. Moreover, he or she coordinates the annotators' actions and manages their status. Thus, his or her role is to moderate the annotation system, and once the annotator has created the annotation or updated it, the librarian will automatically be informed by the system.

The librarian is the unique actor who can list the annotation or the user trace. In addition, he or she can delete a specific annotation, but the system keeps a copy of it, so that it could be restored if required. This annotation history improves the tacit knowledge of the system.

5.1.8 The annotations extraction and their use. The objective of the annotation platform is to provide efficient access to the manuscripts images database. Each annotation is related to a specific bibliographic record (see Figure 4). Thus, selecting an annotation provides access to its bibliographic record, which automatically gives access to the image of the manuscript.

5.1.9 The TEI/XML use for the annotations encoding. Chapter 11 of the TEI Guidelines (TEI P5: Guidelines for Electronic Text Encoding and Interchange) presents efficient tools for the description of primary sources. However, several elements are available to encode the annotations, such as < facsimile >, < sourceDoc >, < surface >, < surfaceGrp > and < zone >. These elements use some attributes such as @facs, @ulx, @ur and @ury to define the abstract and coordinate space which can be used to address parts of the manuscript image (Poibeau, 2003). Figure 6 defines the TEI/XML structure of an encoded annotation:

- < teiHeader > contains the annotation references (such as annotator, date and annotation number).
- < zone > defines the coordinates of the area of the annotated image.

```
<TEI xmlns = "http://www.tei-c.org/ns/1.0>
<TeiHeader>
    <!-- …….. -->
</TeiHeader>
<facsimile>
    <surface start="#PB49R"
          <zone xml:id="B49RW457" ulx="45" uly="125" lrx="60" lry="130"/>
    </surface>
    <graphic url="IbnuTumert-49r.png"/>
    <ref url="notice22.txt">
</facsimile>
<text type="Transcription">
    <body>
        <div>
            <!-- …… -->
            <pb xml: id="B49RW457/>
            <fw> <!--Text of the selected area --> </fw>
        </div>
    </body>
</text>
```

Figure 6.
Example of annotation encoding by using TEI-ms

- < text > contains the text of the annotation. However, we added @type attribute to specify the annotation category (such as transcription and comment).
- < graphic > defines the URL of the annotated image.

The association between the annotation and the bibliographic record is defined in another file. The latter specifies the corresponding annotation document for each bibliographic record. In contrast, XML/TEI encoding specifies full information about the annotation in the same document, which makes indexing of such document easier.

5.2 Arabic manuscripts transcription

Bilane (2010) said that "the transcription is an operation which tends to substitute a phoneme or a sound in any language by a grapheme or a group of graphemes of a writing system in the same language". More simply, transcription is the act of converting any signal towards a grapheme. Following this definition, the Arabic manuscript transcription is the operation of passing from the digital-image form of the manuscript to its textual form (ASCII or Unicode format).

The manuscript transcription aims at making its content readable and understandable to the readers. The aim of manuscript transcription is to generate a text which respects the whole original structure of the manuscript, and any changes should be designed to facilitate the reading of the text. The transcription cannot cause any loss of information that could harm the study of the original manuscript.

5.2.1 The transcription objectives. First, the ability to read the old Arabic manuscript is not possible for all users. Only some experts and scholars can do it. Thus, the transcription is the best way to help different categories of readers and researchers who are interested in the manuscript content.

Second, the transcription provides the possibility of a full-text mode to access the digitised manuscripts. In addition, the transcribed documents can be used for the automatic extraction of metadata, which is useful for cataloguing the manuscripts.

5.2.2 Description of the manuscripts transcription process. Figure 7 shows the schema that describes the transcription process of Arabic manuscripts:

- Image pre-processing: the manuscript image is filtered to reduce its noise, which can disturb the analysis of the document.

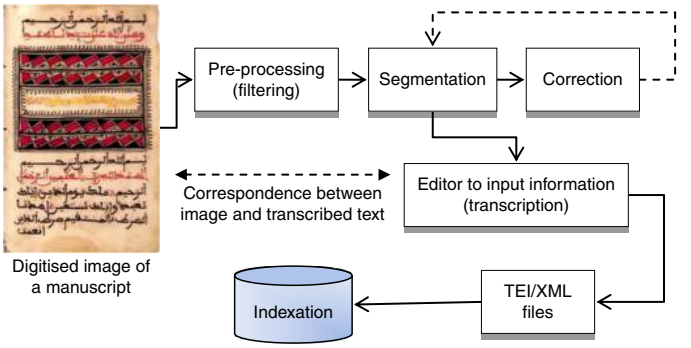

Figure 7.
Transcription process of old Arabic manuscript

- Segmentation: in this step, the system extracts every line of the original text and then isolates each term (word) of the line. As output, the system obtains the whole words of the documents.
- Correction: it is the critical step of the system. It consists of a manual regulation of the erroneous words. That operation allows localising each term in its correct aspect.
- Edition step: the text editor generates fields to enter the text. Each field is associated to a specific bounding box, which allows the creation of a link between the transcribed text and the image of the manuscript by affecting the image coordinates of each word in the digitised manuscript.
- XML encoding: the transcription process ends by generating a XML file for each page of the manuscript per the TEI Manuscript Description (TEI-ms).
- Indexation: the indexation of the XML file provides direct access to the terms of the manuscript image.

5.2.3 Transcription interface. The manuscripts transcription system requires a simple and intuitive interface which provides all the tools that the transcriber needs. Additionally, the cataloguing metadata (such as named entities and dates) should be clearly identified by a tag in a transcription document. That is why the system must offer tools which allow such information to be specified.

Furthermore, the transcriber faces several problems. We can quote the following difficulties:

- the poor readability of a manuscript;
- the deterioration of some folios; and
- the presence of deletions, additions and marginal annotations.

Thus, the transcription system should consider these aspects and provide tools to the transcriber to manage them.

5.2.4 The correspondence between the transcription terms and the words of images. The transcription objective is not only to provide the possibility of metadata extraction but also the possibility of linking the terms of images and their textual transcription.

Figure 8 shows the schema that describes the correspondence principle.

The correspondence process is divided into two phases:

(1) The elaboration of the dictionary: every term in the image is associated with a unique reference. This latter is defined through the manuscript identification, the folio number, the line number and the sequence number of the term in the line.

Once the image term is correctly localised, the system reports its references into the dictionary, and Figure 9 describes the reference term structure.

(2) The correspondence establishment: in the second phase, the system associates each transcription element to a specific reference in terms of a dictionary.

The correspondence between the transcribed text and the manuscript digitised image is performed by using TEI-ms elements such as < facsimile >, < surface >, < zone > and some attributes which define the Cartesian coordinates of the selected segment in the image.

Figures 10-12 show examples of encoding the correspondence in the transcription document.

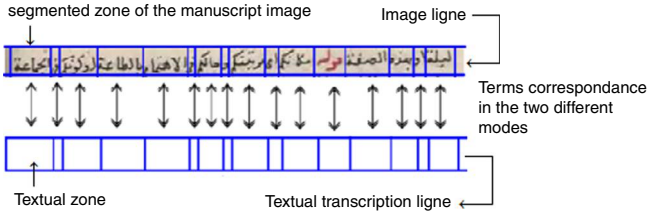

Figure 8.
The correspondence between the terms of image and the transcription

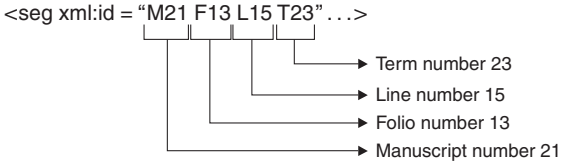

Figure 9.
Reference term structure

```
<TEI>
    <teiHeader/>
    <facsimile>
        <surface>
            <graphic url = "M21.jpg" width="393" height="1100" xml:id="M21"/>
            <zone xml:id="zone-1" ulx="213" uly="3" lrx="342" lry="67"/>
            <zone xml:id="zone-2" ulx="29" uly="31" lrx="214" lry="67"/>
            ....
        </surface>
    </facsimile>
</TEI>
```

Figure 10.
Definition of the image segment and their Cartesian coordinates

```
<div type="transcription" xml:lang="ara">
    <p>
        <seg xml:id="transc-0001">أمير</seg>
        <seg xml:id="transc-0002">المؤمنين</seg>
        ....
    </p>
</div>
```

Figure 11.
Definition of the content of each segment

6. The metadata extraction from the annotation document or the transcription document

The dynamic cataloguing is based on the concept of metadata extraction which stems from the notion of information extraction. Pazienza (1997) defined the information extraction as an activity that aims at automatically filling a form from text written in natural language. This definition presents an understanding of text contents, which is a complex problem that has been classified as a difficulty by researchers (Poibeau, 2003).

The similarity of the information extraction and the metadata extraction is that these two research fields look for relevant information. For example, in both cases, the problem is to find within textual document-specific data which are required to fill the database.

The information extraction depends on an automatic processing stage of the language. Therefore, the problem is transformed into understanding and recognition of a text written in the Arabic language. Consequently, the information extraction is often considered as a complex domain.

The problem becomes easier in the metadata extraction domain by using XML documents. For instance, this kind of document provides the necessary tools by having a specific semantic. Thus, the problem is reduced to a simple instantiation of XML elements, and the system extracts the useful textual content.

The implementation of the metadata extraction model requires constraints when the user annotates or transcribes the manuscript. This makes the metadata spotted by tags. For example, the logical structure of the author's name must be encoded, as shown in the schema presented in Figure 13.

This example shows the contribution of the XML format to facilitating the metadata identification, which makes its extraction easier.

6.1 The principle of the automatically updating the catalogue

Ideally, the logical structure of the annotation document or the transcription document (annotation/transcription document) would be the same as the related metadata structure available within the bibliographic record.

Therefore, any structural difference that appears between the two structures induces the integration of the annotation/transcription document content within the bibliographic record.

```
<linkGrp>
        <link targets="#trans-0001" "#zone-1"/>
        <link targets="#trans-0002" "#zone-2"/>
        …
</linkGrp>
```

Figure 12.
Correspondence between the image references and the text references

```
<author>
  <persName>
    <forename xml: id="Tr03" type="ism" xml:lang="ara">
      الحسين
    </forename>
    <addName xml:id="Tr04" type="kunyah" xml:lang="ara">
      إبن سينا
    </addName>
  </persName>
</author>
```

Figure 13.
The author's name encoding within the annotation text or the transcription text

To do this, we used the concept of structural similarity between two documents. Thus, the system assesses the necessity to integrate the content of the annotation or transcription document into the catalogue. We can face two different cases mentioned as follows:

(1) The lack of the annotated or the transcribed element in the bibliographic record: the whole content of the annotation/transcription document should be copied into the bibliographic record.

(2) Just a part of the annotation or transcription document content is present in the bibliographic record: this is a typical updating case. The system should integrate the inexistent metadata into the bibliographic record.

The structural similarity between two XML documents aims at studying the correspondence of their elements and focusses on the element position within the document hierarchy.

The structural similarity can be evaluated by using the distance between the trees generated by the documents structure (De Francesca *et al.*, 2003). It can also be determined by the detection of the most frequently used sub-trees (Termier *et al.*, 2002). This method is based on the direct relationship of the nodes of the trees.

6.2 The structural similarity measurement

In the literature, several formulas assess the structural similarity between two XML documents. For example, Aitelhadj *et al.* (2009) proposed a formula which considers the ontological constraints and the context of the XML documents nodes.

The structural similarity between two trees T_1 and T_2 is given by the following formula:

$$\mathrm{Sim}(T_1,\ T_2) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\mathrm{sim}(e_{1i},\ e_{2j})}{\mathrm{Max}(|T_1|,\ |T_2|)}$$

$|T_1|$ and $|T_2|$ are the number of nodes (called size) of trees T_1 and T_2, respectively, and they are represented by n and m, respectively. In addition, e_{1i} and e_{2j} are the elements of trees that belong to T_1 and T_2, respectively.

The division by Max ($|T_1|$, $|T_2|$) normalises the result of the summation, and Sim (e_{1i}, e_{2j}) represents the similarity between the nodes e_{1i} and e_{2j}.

In this formula, every e_{1i} node of the T_1 tree is compared with every e_{2j} node of the T_2 tree in any hierarchical position. Thus, this model generates many processing operations, and processing time is estimated to be equal to ($n \times m$).

6.3 The structural similarity between the bibliographic record and the annotation/transcription document

In our model of dynamic cataloguing, the structural similarity is performed between the bibliographic record and the annotation or transcription document.

The encoding format of the annotation or transcription document is well formed according the catalogue Document Type Definition, and each annotation/transcription document is associated to a specific metadata. Thus, there is an implicit equivalence between the two trees representing the bibliographic record and the annotation/transcription document.

These specifications show that the previous formula can be simplified. The comparison between the cited documents is performed through the hierarchical level. Each node of the tree T_1 is compared only with its equivalent in the tree T_2, and we optimise the formula to produce the following equation:

$$\mathrm{Sim}(T_1,\ T_2) = \frac{\sum_{i=j=1}^{n}\mathrm{sim}(e_{1i},\ e_{2j})}{\mathrm{Max}(|T_1|,\ |T_2|)}$$

Remarks:

(1) In our case, the structural similarity measurement can take two values:
- "0" if the node does not exist within the bibliographic record (no associated metadata in the catalogue); and
- "1" if the node exists in the bibliographic record (case of an update of the metadata).

(2) The structural similarity measurement between the annotation/transcription document and the bibliographic record is performed by comparing the e_{1i} and e_{2j} nodes that should have the same position in the respective sub-trees T_1 and T_2. That is why the indices i and j are equal when they go through these sub-trees.

6.4 The use of structural similarity for updating the bibliographic record

This research aims to implement a new cataloguing model where the traditional cataloguer will be less active.

To reach this goal, the focus remains the bibliographic record, which may be incomplete and in a fragmentary state of information. Therefore, it is imperative to assess the necessity of including the annotation or transcription contents in the bibliographic record. The following algorithm enables the implementation of this model:

Algorithm
Input:

– *Annotation/transcription Document* (T_1)
– *Bibliographic notice* (T_2)

Output:

– *Updated bibliographic record*

Similarity assessment {
Calculate the similarity $Sim(T_1,T_2)$;
if $Sim(T_1,T_2)=0$
then
include the T_1 *content into* T_2
else
replace the metadata content of T_2 *by* T_1 *content;*
endif
}

6.5 Metadata extraction process

The metadata extraction process is mainly based on the logical structure of the XML document associated to the annotations or to the transcription of the digitised manuscript. It is based on the instantiation of the tag name which represents the metadata.

The content of every element has been thoroughly studied and validated by a specialised community. Thus, the extraction of the metadata from the annotation/transcription document is based on several phases that we describe in this section:

- calculation of the structural similarity measurement between the entered documents;
- determination of the operation type (such as update and integration);
- mapping the annotation/transcription document;
- instantiation of the tag name of each metadata;
- element (node) content extraction; and
- updating or integrating the metadata tag and its content into the bibliographic content.

Figure 14 shows the general procedure of the metadata extraction.

Phase 1: XML document mapping. In this phase, the annotation/transcription document has gone through every node by starting from the root to the sheets. The elements (nodes) relationship (parent-children) and their path are stored. Thus, each element is associated to several items (content, parent node, attributes and children nodes).

Phase 2: items instantiation. After the calculation of the structural similarity measurement, the system identifies all the missing metadata and those which have been already modified in each bibliographic record. Then, it updates the metadata directory (dictionary).

In this phase, each missed element (metadata) of the bibliographic record is retrieved within the file of the extracted metadata. This is performed by using the instantiation technique.

Phase 3: metadata integration in the bibliographic record. Once the system identified and extracted the content of the metadata, it assigns this content to the related bibliographic record.

This final phase enables updating the bibliographic record, which is the basic resource for the digitised manuscript online access.

7. Implementation and experimental results

The implementation of dynamic cataloguing system requires a transcription module, an annotation module and an automatic metadata extraction module (cf. Figure 2).

The experimental corpus of manuscripts contains ten digitised manuscripts, which means that the catalogue contains ten bibliographic records as well.

The groups of annotators and transcribers were composed by students in the information science field and two researchers. All of them are specialists of cataloguing old Arabic manuscripts.

In total, 17 annotations (comments and metadata) were validated and five pages of various manuscripts were transcribed.

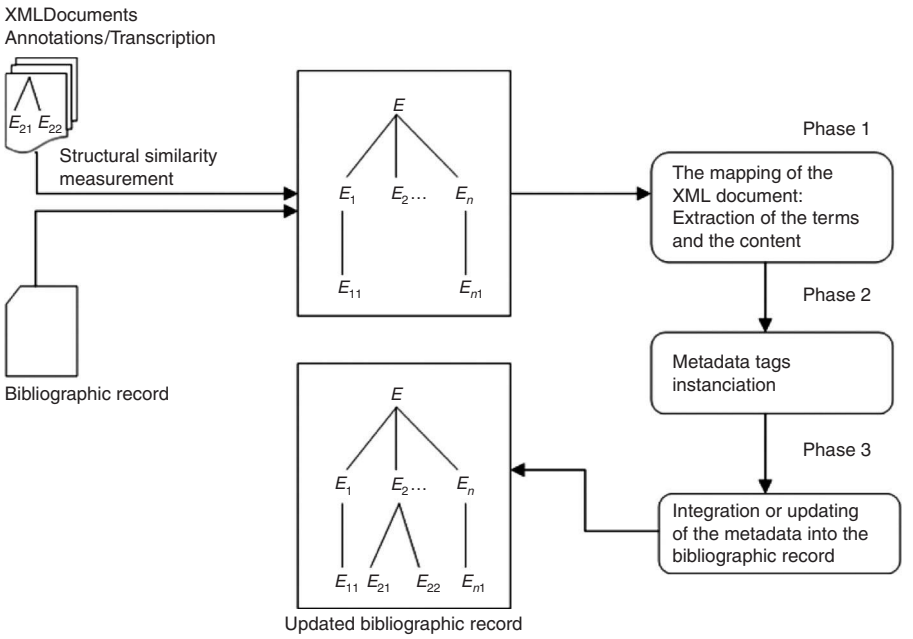

Figure 14. General procedure of the metadata extraction

Experiments were focussed on two cases:

- Case 1: adding a new metadata into the catalogue.
- Case 2: updating the catalogue content. In this case, we faced two different situations. On one hand, the bibliographic record may be updated by making corrections; on the other hand, it can be completed by missed information.

The tests were conclusive and they have produced efficient results. However, experiments have shown the reliability of the dynamic cataloguing model. This latter is a revolutionary tool for the libraries world.

8. The extension of the dynamic cataloguing model

The dynamic cataloguing model of the digitised old Arabic manuscripts is based on the resources relationship. Thus, every library should have access to timely information regarding each shared resource. Therefore, links must be established between the various catalogues of the different libraries.

The challenge is to establish a new structuration catalogue model which links the metadata of the catalogues of the various libraries. This aspect needs to introduce the concept of interoperability between catalogues. Therefore, the extension of our research work is based on the semantic web concept.

9. Conclusion

Through our paper, we introduced a new concept of evolutionary cataloguing which can be updated at any time per the scholars' remarks. This new concept results from the instability of the old Arabic manuscript, and we called this cataloguing model "the dynamic cataloguing".

The dynamic cataloguing model depends on the information quality provided by the scholars and the experts who annotate and transcribe the digitised manuscripts. To do this, we used a powerful tool: "The automatic extraction of metadata from XML document".

In addition, our metadata extraction approach is based on the knowledge of the XML document logical structure. That is why we focussed our effort on the tag names of XML documents.

The XML document characteristics make implementing our approach easy, because the annotation or transcription document is data oriented; this avoids the analysis step of the XML document content.

In our work, we propose an original cataloguing model based on the transcription and the annotation to catalogue the digitised manuscripts. The solution allows the dynamic cataloguing model to be established. In addition, it highlights the dynamic aspect of metadata of old Arabic manuscripts cataloguing model. However, the implementation of this solution requires collaboration of actors from different backgrounds such as cataloguers, experts and scholars through a collaborative platform.

References

Agosti, M., Ferro, N. and Orio, N. (2005), "Annotating illuminated manuscripts: an effective tool for research and education", *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, ISBN: 1-58113-876-8, ACM, New York, NY, pp. 121-130, doi: 10.1145/1065385.1065412.

Aitelhadj, A., Mezeghiche, M. and Souam, F. (2009), "Classification de structure arborescente: Cas de documents XML", *Proceedings of the Sixth Francophone Conference on Information Retrieval and Applications, CORIA, Université du Sud Toulon-Var*, pp. 301-317.

Al-Rajebah, N.I. and AL-Khalifa, H.S. (2012), "Makhtouta +: enhancing old Arabic manuscripts with linked data", *Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services*, ISBN: 978-1-4503-1306-3, ACM, New York, NY, pp. 323-327, doi: 10.1145/2428736.2428794.

Auerbach, E. (1961), *Introduction aux études de philologie romane*, V. Klostermann, Frankfurt am Main.

Azouaou, F., Desmoulins, C. and Mille, D. (2003), "Formalisme pour une mémoire de formation à base d'annotations : Articuler sémantique implicite et explicite", *Actes de la conférence EIAH, INRP*, pp. 43-54.

Bilane, P. (2010), "Contribution à l'indexation et à la reconnaissance des manuscrits syriaques", IT doctoral thesis, INSA de Lyon, available at: https://tel.archives-ouvertes.fr/tel-00499537/document (accessed 8 February 2017).

De Francesca, F., Gordano, G., Ortale, R. and Tagarelli, A. (2003), "Distance-based clustering of xml documents", *Workshop Proceedings of the First International Workshop on Mining Graphs, Trees and Sequences, ECML/PKDD'03, Cavtat-Dubrovnik, Croatia*, Vol. 3, pp. 75-78.

Deschatelets, G. and Dufour, C. (1997), "Étude de faisabilité pour la mise en place d'une bibliothèque virtuelle en sciences de l'information dans le cadre du projet SLISNET", École de bibliothéconomie et des sciences de l'information, Université de Montréal, available at: https://web.archive.org/web/20080520084852/; http://enssibal.enssib.fr/autres-sites/SLISNET/bibelec/index.html (accessed 8 February 2017).

El Makhfi, N., El Bannay, O., Benslimane, R. and Raïs, N. (2011), "Search engine of ancient Arabic manuscripts based on metadata and XML annotations", 2011 Colloquium in Information Science and Technology (CIST), IEEE, Fez 2011, p. 10.

Feodorov, I. (2006), *A New Standard in Arabic Manuscript Cataloguing*, revised ed., Etudes Sud-Est Europ.

Ferrand, N. (1996), "Les bibliothèques virtuelles", Magazine littéraire, No. 349, December, pp. 38-39.

Gonçalves, M.A., Fox, E.A., Watson, L.T. and Kipp, N.A. (2004), "Streams, structures, spaces, scenarios, societies (5s): a formal model for digital libraries", *ACM Transactions on Information Systems*, Vol. 22 No. 2, pp. 270-312.

Pazienza, M.T. (Ed.) (1997), *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, Vol. 1299, Springer, Heidelberg.

Poibeau, T. (2003), *Extraction Automatique D'information: du texte brute au web sémantique*, Lavoisier, Paris.

Sanderson, R., Albertitton, B., Schwemmer, R. and Van De Sampel, H. (2011), "SharedCanevas: a collaborative model for medieval manuscript layout dissemination", *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, ISBN: 978-1-4503-0744-4, ACM, New York, NY, pp. 175-184, doi: 10.1145/1998076.1998111.

Scheper, K. (2015), *The Technique of Islamic Bookbinding Methods, Materials and Regional Varieties*, ISBN: 9789004290921, Vol. 8, Islamic Manuscripts and Books, Leiden University Library (UBL), Brill, Leiden and Boston, MA, p. 428.

Soualah, M.O. and Hassoun, M. (2011), "Which metadata for ancient Arabic manuscripts cataloguing?", *DCMI: International Conference on Dublin Core and Metadata Applications – The Hague Proceedings, DC-2011*, pp. 126-136, available at: http://dcpapers.dublincore.org/pubs/article/view/3629 (accessed 8 February 2017).

Soualah, M.O. and Hassoun, M. (2012), "A TEI P5 manuscript description adaptation for cataloguing digitized Arabic manuscripts", *Journal of the Text Encoding Initiative*, No. 2, February, doi: 10.4000/jtei.398, available at: http://jtei.revues.org/398

Soualah, M.O., Ait Ali Yahia, Y., Hassoun, M. and Guessoum, A. (2012), "Multilingual online access to digitised Arabic manuscripts by using metadata", *International Journal of Metadata, Semantics and Ontologies*, Vol. 7 No. 4, pp. 237-253.

Termier, A., Rousset, M.C. and Sebag, M. (2002), "Tree finder: a first step towards XML data mining", *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02), IEEE Computer Society, Washington, DC*, pp. 450-457.

UKOLN (Université de Bath) (2004), "Recommandations techniques pour les programmes de création de contenus culturels numériques", Projet européen Minerva, p. 48, available at: www.enssib.fr/bibliotheque-numerique/documents/63467-recommandations-techniques-pour-les-programmes-de-creation-de-contenus-culturels-numeriques.pdf (accessed 8 February 2017).

Further reading

Desmontils, E., Jacquin, C. and Simon, V. (2003), "Vers un système d'annotation distribué", Rapport de recherche No. 03.01 Institut de Recherche en Informatique (IRIN), Université de Nantes, available at: www.desmontils.net/Documents/DJS03a.pdf (accessed 8 February 2017).

Doumat, R., Egyed-Zsigmond, E. and Pinon, J.M. (2008), "Un modèle d'une bibliothèque numérique collaborative – ARMARIUS", *Proceedings of the Fifth Francophone Conference on Information Retrieval and Applications, CORIA, Trégastel*, pp. 417-424.

Corresponding author
Mohammed Ourabah Soualah can be contacted at: m_soualah@esi.dz