New protocol design for wordspotting assistance system: Case study of the collaborative library model - ARMARIUS

Abir CHAARI, Fadoua DRIRA, Adel M. ALIMI
University of Sfax, REGIM-Lab, ENIS
3038 Sfax, Tunisia
abir.chaari@gmail.com, fadoua.drira@ieee.org,
adel.alimi@ieee.org

Előd E.-ZSIGMOND, Franck LEBOURGEOIS
University of Lyon, LIRIS, INSA-Lyon
69621 Villeurbanne Cedex, France
elod.egyed-zsigmond@insa-lyon.fr,
franck.lebourgeois@insa-lyon.fr

Abstract—The cultural heritage is full of important manuscript collections preserved in digital libraries. The need to annotate and enrich the scanned documents is claimed by some users to keep traces in the system for a further use. Moreover, the reuse of annotations could help other users to accomplish repetitive tasks in a semi-automatic way. One manuscript annotation technique is the wordspotting. It is a process that seeks in a document for all the fragments that are similar to the one specified by the user.

The main focus of this research work is to propose a solution integrating and encapsulating the wordspotting algorithm in digital libraries. This solution involves, in particular, the specification and the implementation of an architecture to integrate the image processing tool using Restful Web services. The proposed prototype is tested on the ARMARIUS digital library. This library is one of the collaborative digital archiving models that stores ancient digitized manuscripts.

Keywords - digital library, manuscripts annotation, assistance system, wordspotting, Web service.

I. INTRODUCTION

Digital libraries include collections of documents in a database. They manage a variety of digitized manuscripts. They offers different services to the users such as searching information, securing collections, creating personal spaces, saving search links, etc. The need to annotate and enrich the scanned documents is claimed by some users to help the search and exploitation. Document annotation facilitates the exchange and the share of knowledge among a group of users. Indeed, annotation is defined as an “added value” that adds information judged pertinent for an optimal retrieval of the stored data. An annotated document is much more interesting and easy to handle than a set of images. Whatever the field of research (physics, biology, geography, sociology, etc), generally characterized by a huge amount of data, it’s natural that researchers are looking for tools to support annotation, and reduce the effort required to produce an annotated corpus. In fact, the annotation of handwritten manuscripts is impossible with OCR technique originally defined for printed documents. Furthermore, these documents are not structured in XML [1, 2]. For this reason, the manuscripts must be manually transcribed and annotated. Therefore, this solution is very expensive and tedious to establish.

A possible solution to these observations is to set up a collaborative library model incorporating assistance systems. These systems offer to their users the possibility to save their opinions, enrich permanently their documents with texts or graphical forms and share it in a wide collaborative space. Thus, they offer the semi-automatic annotation of manuscripts. Nevertheless, the annotation of scanned images of manuscripts requires the use of image processing tools such as wordspotting, figure or region detection...

This paper proposes the integration of a communication protocol between a digital library and image processing tools to manage annotations. The use of this protocol is not confined to a specific architecture of a digital library; it could be integrated to any one.

The paper is organized as follows.

Section 2 presents a state-of-the-art of some annotation systems. We reveal the limits of these systems mainly in supporting image processing tools. Section 3 gives the specification and the implementation of an architecture to integrate the image processing tool using Restful Web services. We focus in this study on the wordspotting process as an example of image processing tool. A brief review of this process is thus introduced. A case study is given in Section 4. The proposed architecture is thus tested on the ARMARIUS digital library [1, 16]. This library is one of the collaborative digital archiving models that stores ancient digitized manuscripts.

II. STUDY OF EXISTING ANNOTATION SYSTEMS

This section presents a study of some annotation projects and research prototypes that have been developed to annotate Web documents, multimedia or personal documents.

Annotea, a W3C project, offers a shared annotation system for Web documents. Annotations could be attached to both HTML and XML based documents. This project uses the RDF schema for describing annotations as metadata and XPointer for locating them in the document [3]. Amaya is the Web browser supporting collaboration via shared

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.242

metadata, bookmarks and text annotations on Web pages using Annotea.

MADCOW [4] (Multimedia Annotation of Digital Content Over the Web) is a system for multimedia annotation over the Web. It permits to add information to a document or part of it.

Critlink [5] is an annotation tool that allows the individual or public annotation of the Web pages and permits the implementation of critical discussions around these documents. It stores the annotations with the HTML standard in an autonomous proxy server. It requires neither plugins nor specific client nor server software.

Yawas [6] (Yet Another Web Annotation System) is a Firefox plugin for annotating any text or HTML document from the Web or the local machine.

The above systems offer different approaches to support collaborative work. Nevertheless, they are web-based ones where both the sources as well as annotations are HTML/XML documents. This represents the most important inconvenience of these systems which limits their application on while it is not able with scanned document images. For both of Annotea and Critlink, the Web pages and the contents of their objects (images, texts, hyper links, etc.) are identified by URLs where as scanned images of the manuscripts are identified by IDs [6].

Web-based image management systems such as Flickr [7] and Picasa [8] give their users the ability to annotate manually the images. Their annotation service allows the introduction of user information related to a given image as a comment or tag. The major drawback of such systems is that annotations are generally restricted to one or two words describing the image. This restriction limits enormously the exploration of textual documents.

This study reveals the need for a semi-automatic annotation tool that could be integrated into a digital library. Information retrieval systems dedicated to the images based document have been mainly based on Optical Character Recognition (OCR) systems which are specific to a limited category of documents: mainly printed or typewritten contemporary ones. The ancient manuscripts are in general very damaged documents. Thus, the poor quality of writing and even the variability of writing styles require the application of specific tools useful for indexing and searching information in documents. The wordspotting [9], a possible solution, facilitates the search for information by identifying occurrences of a word as image fragment in all the pages of manuscripts. The integration of this tool into a digital library will be the aim of the next section.

III. Proposed solution: Web service-based session oriented architecture for wordspotting assistance system

The incorporation of the wordspotting assistance system in a digital library model can be achieved only through a communication protocol. The difficulty is to find out how communication can be made between heterogeneous applications? Which architecture to choose or design? Which format to use for data exchange?

The first idea to resolve the problem of interoperability could be the development of a protocol that allows applications to communicate over the Web. The proposed solution was to integrate Web services. In fact, Web services are software programs that provide the communications and information exchange among heterogeneous systems and applications. They are based on both protocols and Web languages especially HTTP and XML [10]. These services define a standard way to ensure a remote application and retrieve the results through the Web. They are used to interconnect different applications by combining a set of resources that can be deployed on a server independently of platforms and languages. There are two main approaches to the implementation of Web services SOAP (Simple Object Access Protocol) and REST (Representational State Transfer).

REST is rather new, it is neither a protocol nor a technology, but it is a style of architecture. It was described by R. Fielding in his dissertation "Architectural Styles and the Design of Network-based Software Architectures". This concept assumes that the Internet is not a set of pages but a set of resources accessible from an URI [11]. REST is an architectural style defined as an alternative to TCP and most cases of using SOAP. It provides data exchange, as SOAP and XML-RPC, often with XML. Web services that use the principles of REST architecture are often called Restful. The REST architecture is based on the infrastructure of the Web that is defined by the states and functions of a remote application as resources. Each resource is available as a unique one and has a standard address format. Under HTTP, it will be an URI with a universal syntax for addressing resources. These have a standard interface in which the operations and data types are precisely defined in XML or JSON or HTML format of data exchange. Data exchange is done on a client-server protocol, without states, characterized by a proxy cache server and multiple software layers [11, 12].

Another issue of the wordspotting based annotation assistance is the time, the image processing algorithm takes. When it has to treat several hundreds of high resolution images, wordspotting can take more time than a real-time annotation assistant can authorize.

To overcome this issue, we have created a session based approach, where users can select the collections, define and annotate or transcribe the candidate words and launch the wordspotting creating a new job session. Users can view the state of each session and validate the finished ones.

The role of a REST Web service is to initialize the wordspotting algorithm with data encapsulated in the client request. It creates files of treatment sessions and checks for the collections of manuscripts that are the subject of the research. If the document images have not been found, the Web service will download the missing images of manuscripts. After downloading them, they will be copied to the correct location at the runtime of the wordspotting.

When the wordspotting sessions launch the image processing tool, it extracts the result into an XML file as the coordinates of the occurrences of the original annotation. These coordinates will be returned, used and processed by the Web service. Transcripts will be stored and displayed to the client Web application that uses this service. The user have the choice between validating the transcripts displayed, in which case they will be visible in the collection, or deleting them.

IV. Experimentation: Case of the library ARMARIUS

In this section, we describe the evaluation of some digital library models in order to choose the best one to integrate and validate the function of our wordspotting protocol. The evaluation of digital libraries must be based on essential criteria. The requested criteria are to have a graphical interface, a search engine, a security system and especially a support system and personal space. All these services allow users to take advantage of them and realize their needs in a short time.

Some models of digital libraries including support systems and monitoring annotations exist like ETANA-DL [13], an archaeological digital library, which manages a huge amount of manuscript collections and provides various services such as browsing, searching and annotating. GALLICA [14] is a European digital library that facilitates to its users different services like searching the documents, annotating and saving their preferences. Google Books Search [15] is a digital library that provides a personal space to search, add the user favorite books and even to comment them. ARMARIUS [16] is a digital archiving model that stores several manuscripts. It offers a personal space, a security system to manage the access rights of users, tracing system, support system and supports several types of annotations. Consequently, we are choosing ARMARIUS the most rewarding library and which includes the most requested features to facilitate the creation of a collaborative space.

The prototype of the wordspotting Web service was tested with the collaborative digital library ARMARIUS. The latter is an “open source” prototype that manages a corpus of manuscripts and annotations. This model is structured in several collections which are organized into several sub-collections containing images. This model allows the manual annotation of manuscripts. Fig. 1 illustrates the content of ARMARIUS digital library. This library is characterized by annotations which are of several types (keywords, transcripts, comments, questions, answers, digital signatures, messages). They are anchored to the images independently of the pixels which allow having the same notation for different versions of the image [16]. The library users identify themselves by their personal accounts to search into the collections, retrieve pages and view annotations. The generation of new annotations is allowed only to users who have the necessary rights.

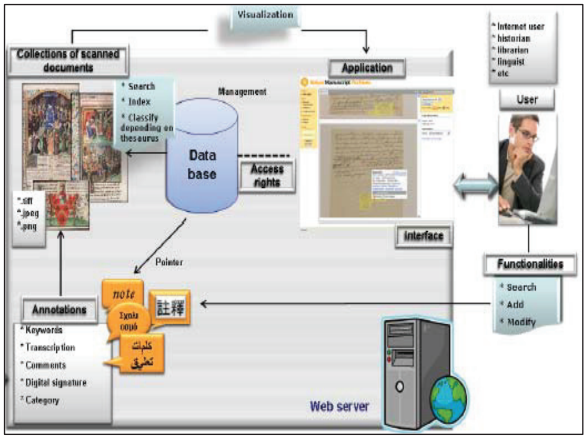


Figure 1. The content of ARMARIUS digital library

ARMARIUS is chosen as the case study to validate the proposed Web-service based architecture for the assistance system.

We must notice that the wordspotting algorithm tested here is that proposed by Leydier et al [5]. This algorithm is based on differential features that are compared using a cohesive elastic matching method, based on zones of interest in order to match only the informative parts of the words. It aims to facilitate the search for information by identifying occurrences of a word image in all the pages of manuscripts. This word query is selected by the user who defines the contours of an instance in any page of the document. Then, the system sends back the ordered list of the similar words for the image instance. Next, the list of answers is returned and sorted in order of similarity. The operator can by the way select the correct answers and eliminate the false detections.

The Fig. 2 illustrates the possible communication scenarios between ARMARIUS and the wordspotting tool

(AlgoWOS) after the integration of the proposed Web service prototype.

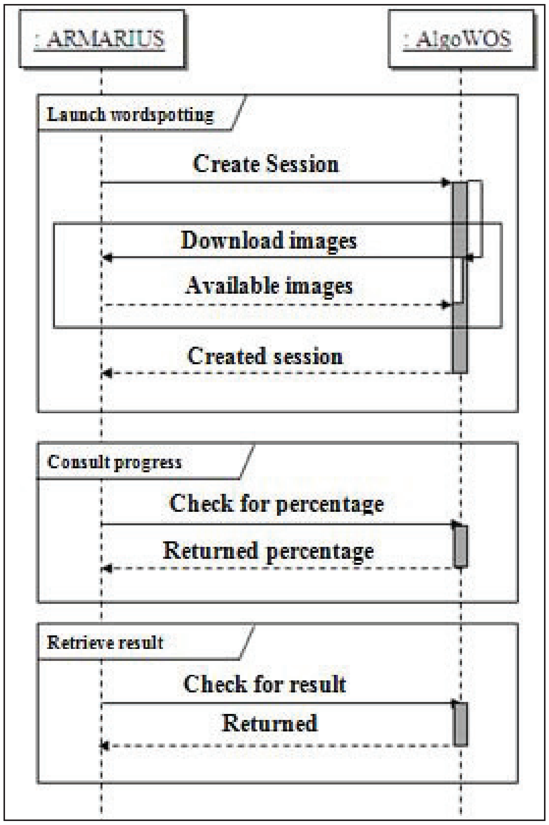


Figure 2. Communication scenarios between ARMARIUS and Wordspotting

Once the Web service is installed, the ARMARIUS logged user can launch a wordspotting session. This step is illustrated by the Fig. 3. In fact, the user browses the document of the selected collection on which he adds either manually or with the assistance system of the application new keywords or transcript. Then, he saves the transcript of the annotation to display the link “Launch wordspotting”.

When the user clicks on the link “Launch wordspotting” on the annotation popup, the interface of the Fig. 4 appears to choose the writing direction necessary to be specified for the wordspotting process (Left->Right (in French), Right->Left (in Arabic)). At the same time, the Web service replies with the needed images to be downloaded if they are not already present in the local machine of the wordspotting. Once the images are available, the session file will be created into an XML file as it is shown in Fig. 5.

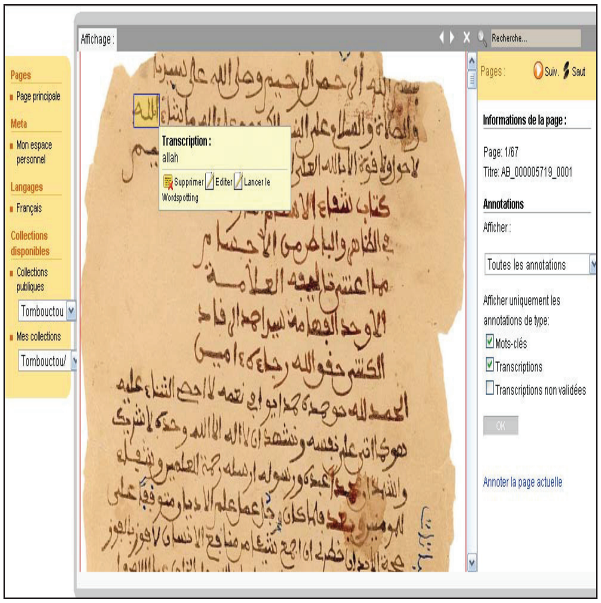
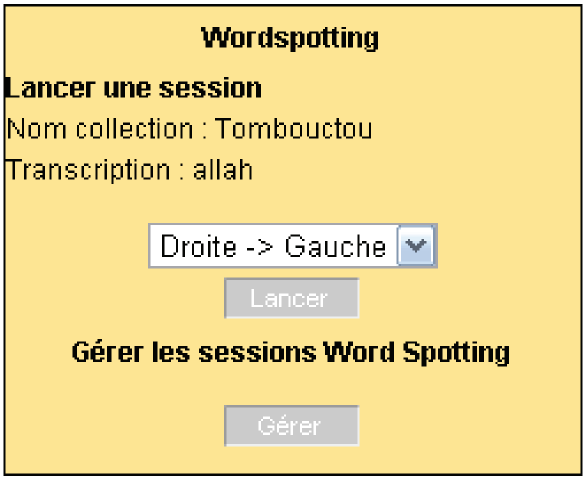


Figure 3. Launch of the wordspotting tool through the ARMARIUS interface



Wordspotting

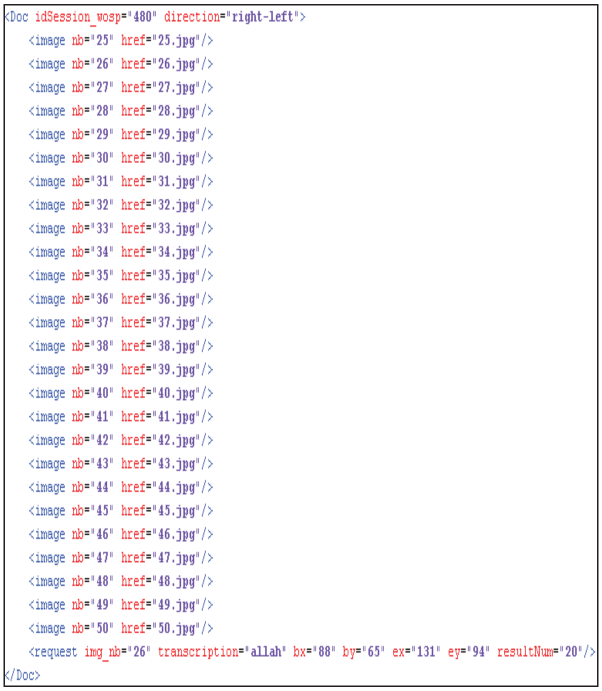
Lancer une session

Nom collection : Tombouctou

Transcription : allah

Gérer les sessions Word Spotting

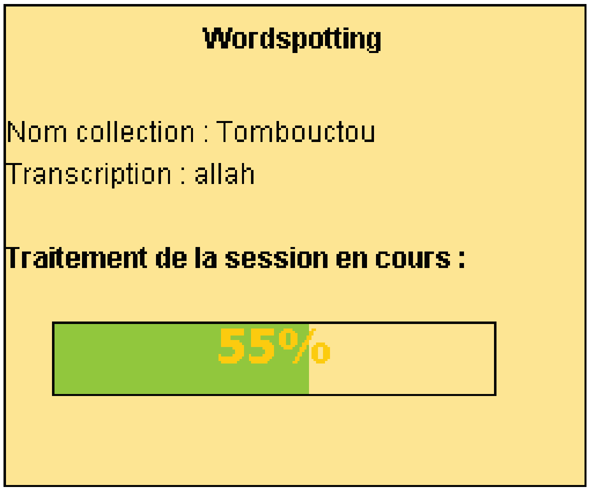
Figure 4. ARMARIUS interface for the choice of the writing direction (Left->Right (in French)) here for Arabic texts



```
<Doc idSession_wosp="480" direction="right-left">
    <image nb="25" href="25.jpg"/>
    <image nb="26" href="26.jpg"/>
    <image nb="27" href="27.jpg"/>
    <image nb="28" href="28.jpg"/>
    <image nb="29" href="29.jpg"/>
    <image nb="30" href="30.jpg"/>
    <image nb="31" href="31.jpg"/>
    <image nb="32" href="32.jpg"/>
    <image nb="33" href="33.jpg"/>
    <image nb="34" href="34.jpg"/>
    <image nb="35" href="35.jpg"/>
    <image nb="36" href="36.jpg"/>
    <image nb="37" href="37.jpg"/>
    <image nb="38" href="38.jpg"/>
    <image nb="39" href="39.jpg"/>
    <image nb="40" href="40.jpg"/>
    <image nb="41" href="41.jpg"/>
    <image nb="42" href="42.jpg"/>
    <image nb="43" href="43.jpg"/>
    <image nb="44" href="44.jpg"/>
    <image nb="45" href="45.jpg"/>
    <image nb="46" href="46.jpg"/>
    <image nb="47" href="47.jpg"/>
    <image nb="48" href="48.jpg"/>
    <image nb="49" href="49.jpg"/>
    <image nb="50" href="50.jpg"/>
    <request img_nb="26" transcription="allah" bx="88" by="65" ex="131" ey="94" resultNum="20"/>
</Doc>
```

Figure 5. The session file created by the Web service after the data client request

The second scenario "Consult progress" is demonstrated by the following screen (Fig. 6) which permits the user to consult the completion percentage of the current session of the wordspotting. Indeed, when the session file was created, the wordspotting starts the processing of the images. The operator can always send a request to consult the progress of the current session treatment.



Wordspotting

Nom collection : Tombouctou

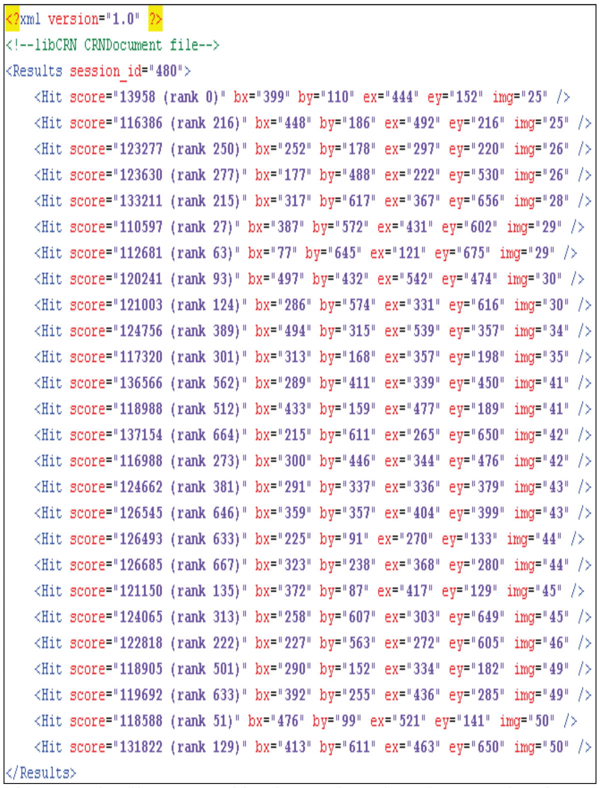
Transcription : allah

Traitement de la session en cours :

55%

Figure 6. Illustration of the wordspotting session progress

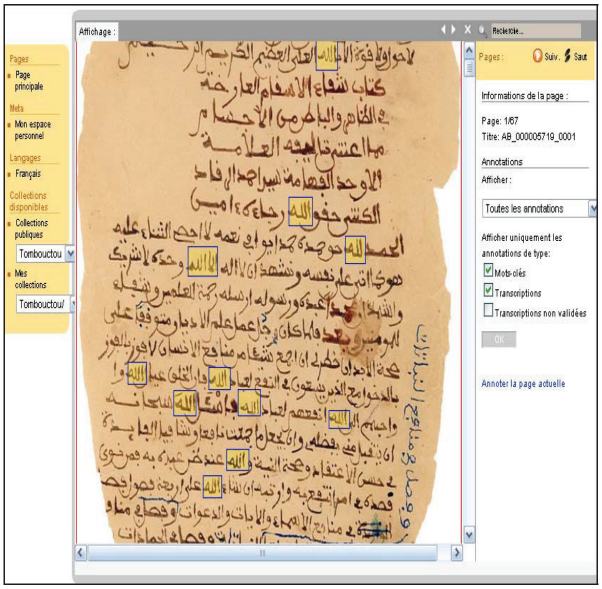
Finally, the user can retrieve the result of the worspotting session. For instance, the wordspotting daemon regroups in a file the different coordinates of all the occurrences similar to the original annotation (Fig. 7). When the percentage reaches 100%, the ARMARIUS user can be informed that the session is finished and he can check the result. The Web service converts and transmits it in a JSON format to the ARMARIUS application. The results can then be validated by the connected user to make them visible in his collection. These annotations can be used for other users who are allowed to access to the collection. The Fig. 8 shows the annotations found by the wordspotting tool and validated by the user.



```
<?xml version="1.0" ?>
<!--libCRN CRNDocument file-->
<Results session_id="480">
    <Hit score="13958 (rank 0)" bx="399" by="110" ex="444" ey="152" img="25" />
    <Hit score="116386 (rank 216)" bx="448" by="186" ex="492" ey="216" img="25" />
    <Hit score="123277 (rank 250)" bx="252" by="178" ex="297" ey="220" img="26" />
    <Hit score="123630 (rank 277)" bx="177" by="488" ex="222" ey="530" img="26" />
    <Hit score="133211 (rank 215)" bx="317" by="617" ex="367" ey="656" img="28" />
    <Hit score="110597 (rank 27)" bx="387" by="572" ex="431" ey="602" img="29" />
    <Hit score="112681 (rank 63)" bx="77" by="645" ex="121" ey="675" img="29" />
    <Hit score="120241 (rank 93)" bx="497" by="432" ex="542" ey="474" img="30" />
    <Hit score="121003 (rank 124)" bx="286" by="574" ex="331" ey="616" img="30" />
    <Hit score="124756 (rank 389)" bx="494" by="315" ex="539" ey="357" img="34" />
    <Hit score="117320 (rank 301)" bx="313" by="168" ex="357" ey="198" img="35" />
    <Hit score="136566 (rank 562)" bx="289" by="411" ex="339" ey="450" img="41" />
    <Hit score="118988 (rank 512)" bx="433" by="159" ex="477" ey="189" img="41" />
    <Hit score="137154 (rank 664)" bx="215" by="611" ex="265" ey="650" img="42" />
    <Hit score="116988 (rank 273)" bx="300" by="446" ex="344" ey="476" img="42" />
    <Hit score="124662 (rank 381)" bx="291" by="337" ex="336" ey="379" img="43" />
    <Hit score="126545 (rank 646)" bx="359" by="357" ex="404" ey="399" img="43" />
    <Hit score="126493 (rank 633)" bx="225" by="91" ex="270" ey="133" img="44" />
    <Hit score="126685 (rank 667)" bx="323" by="238" ex="368" ey="280" img="44" />
    <Hit score="121150 (rank 135)" bx="372" by="87" ex="417" ey="129" img="45" />
    <Hit score="124065 (rank 313)" bx="258" by="607" ex="303" ey="649" img="45" />
    <Hit score="122818 (rank 222)" bx="227" by="563" ex="272" ey="605" img="46" />
    <Hit score="118905 (rank 501)" bx="290" by="152" ex="334" ey="182" img="49" />
    <Hit score="119692 (rank 633)" bx="392" by="255" ex="436" ey="285" img="49" />
    <Hit score="118588 (rank 51)" bx="476" by="99" ex="521" ey="141" img="50" />
    <Hit score="131822 (rank 129)" bx="413" by="611" ex="463" ey="650" img="50" />
</Results>
```

Figure 7. The file generated by the wordspotting after grouping the different occurrences similar to the original annotation

The above study of the feasibility of the wordspotting Web service assistance system on the digital library ARMARIUS is carried out for an extract of a manuscript provided by the "Tombouctou" archive. Other document images are also tested but for lack of space, we give just one example. Overall acceptable results are obtained and we can consider that we reached the aim of our research work.



Informations de la page :

Page: 1/67

Titre: AB_000005719_0001

Annotations

Afficher : Toutes les annotations

Afficher uniquement les annotations de type:

☑ Mots-clés

☑ Transcriptions

☐ Transcriptions non validées

OK

Figure 8. Wordspotting result validated by the user

This section proves the efficiency of the proposed architecture once integrated to ARMARIUS. Other digital libraries could also be subject to this integration as the proposed architecture is useful for any Web application. This concern of supporting heterogeneity is argued by the main principal advantage of the REST-based architecture. For instance, each resource of the Restful Web service is identified by an URI. Thus, any Web application could access the service.

V. Conclusion

The integration of image processing tools in Web-based applications especially in digital libraries is being a promising issue. For instance, many users could share, exchange and take benefits from several information in a short duration and in a collaborative environment. With the REST-based Web service that considers the Web as a resource application, interoperability and communication among heterogeneous applications is simple, flexible and easy to realize.

This paper emphasizes the fact that a digital library such ARMARIUS can take advantage from the wordspotting Web service through a REST-based architecture. We also show that long processing times can be handled through a session based approach. As a perspective to this study, we tend to enrich and enhance other digital libraries by integrating wordspotting and even other image processing tools. Further research will investigate the efficiency of a protocol based on the cloud computing...

References

[1] Reim Doumat, E. Egyed-Zsigmond, J.M. Pinon. A web 2.0 archive to access, annotate and retrieve manuscripts. Multimedia Tools and Applications ():1-21, Springer. ISSN: 1380-7501, 1573-7721, 2011.

[2] Reim Doumat, E. Egyed-Zsigmond, J.M. Pinon. "User Trace-Based Recommendation System for a Digital Archive". Dans ICCBR 2010, Isabelle Bichindaritz, Stefania Montani ed. Alexandria, Italie. pp. 60-74. Lecture Notes in Computer Science 6176. Springer. ISBN 978-3-642-14273-4,2010.

[3] J. Kahan, M. Koivunen. "Annotea: an open RDF infrastructure for shared Web annotations". In Proceedings of the 10th International World Wide Web Conference, Hong Kong, pp. 623-632, May 2001.

[4] P. Bottoni, R. Civica, S. Levialdi, L.Orso, E.Panizzi, R.Trinchese. "MADCOW: a multimedia digital annotation system". In Proceeding of the working conference on advanced visual interfaces (AVI'04), pp.55-62, Italy 2004.

[5] Ka-Ping Yee. "CritLink: Advanced Hyperlinks Enable Public Annotation on the Web", ACM Conference on Computer Supported Cooperative Work, 2002.
http://crit.org/critlink.html

[6] L. Denoue, L.Vignollet. "Yawas: un outil d'annotation pour les navigateurs du Web". In IHM'99, Montpellier, France, Novembre 1999.

[7] W. Jill. "Feral hypertext: when hypertext literature escapes control". In Proceeding of the sixteenth ACM conference on Hypertext and hypermedia, Salzburg, Austria, pp. 46-53, September 2005.

[8] Available at: http://picasa.google.com/

[9] Y. Leydier, F. Lebourgeois, H. Emptoz. « Text search for medieval manuscript images". Pattern Recognition n°40 (12), pp. 3552–3567, Décembre 2007.

[10] H. Kadima, V. Monfort. "Les Web Services : techniques, démarches et outils", France. Dunod. ISBN 978-2-10-006558-5, 2003.

[11] R. Fielding. "Architectural Styles and the Design of Network-based Software Architecture". PhD thesis, University of Califormia, USA, 2000.

[12] R. Fielding, R. N. Taylor. "Principle design of the modern web architecture". In Proceeding of the 22nd international conference on Software engineering, pp. 407–416, New York, 2000.

[13] U. Ravindranathan., S. Rao, M.A. Goncalves, W. Fan, E.A. Fox, J.W Flanagan. "ETANA-DL: a digital library for integrated handling of Heterogeneous archaeological data". Joint Conference on Digital Libraries, pp.76-77, USA, 2004.

[14] Available at: "http://gallica.bnf.fr".2000.

[15] Available at: "books.google.fr/".

[16] Reim Doumat, E. Egyed-Zsigmond, J.M. Pinon, E. Csiszar. "Online ancient documents: Armarius". Dans ACM DocEng'08, Sao Paulo, Brésil. pp. 127-130, ISBN 978-1-60558-081, 2008.