A data model and a cataloguing, storage and retrieval system for ancient document archives

Pasquale Savino, Anna Tonazzini, and Franca Debole

***Abstract*—— Digitalization of ancient manuscripts is becoming a common practice in many archives and libraries, mainly for preservation purposes. This opens many new opportunities for the diffusion of these precious cultural assets, since several scholars and researchers, as well as the general public, may access and use them for research purposes, for study, and for general information. This is made possible if the documents, their descriptions, and the result of all processing activities are acquired at a good level of quality, and can be easily accessed by using simple and powerful retrieval mechanisms.**

Acquired manuscripts suffer from degradations that may require different types of elaboration on the digital images, to improve their visual quality and legibility, or to discover hidden text that is not visible. Natural Language Processing technology presupposes the availability of transcriptions of the text contained in the manuscript, as well as encoding of the document structure and the creation of user annotations.

This paper presents a document management system and a metadata schema that make possible the storage and content-based retrieval of original documents, elaborations performed to improve their readability, textual transcriptions, and linguistic annotations. The archive offers the possibility of describing, storing and accessing all the available manuscript versions, document transcriptions and annotations, and to search and retrieve documents based on all this information.

Index Terms*— *Ancient manuscript preservation and accessibility, Metadata schema for multispectral images, Metadata Editor tool, Digital Library of multispectral images

I. INTRODUCTION

THE availability of high quality digital acquisition equipment with affordable costs is making the digitalization of old manuscripts a common practice. Currently, several cultural institutions have undertaken or even completed the digitalization of their rich documental collections. Frequently, these data are of high quality in terms of resolution, levels of detail, modalities (e.g., multispectral digital representations and 3D representations), and content description. These collections are, or could be, accessible on-line, thus offering to experts and scholars the possibility to study previously unavailable pieces of our cultural heritage.

Unfortunately, natural ageing, usage, poor storage conditions, humidity, molds, insect infestations and fires often have produced degradations that make complex the reading and interpretation of these manuscripts. In addition, the materials used in the original production of the manuscripts, i.e. paper or parchment and inks, are usually highly variable in consistency and characteristics. These problems are common to the majority of the governmental, historical, ecclesiastic and commercial archives, so that seeking out for remedies to restore the digital images of ancient manuscripts and documents would have an enormous social and technological impact.

Furthermore, experts and scholars such as philologists and linguists may need a textual transcription of the document content, in order to study and comment the document from the semantic, linguistic, lexical, and philological point of view. The result of their analysis is then associated with the document as multi-layered annotations, one for each analysis performed (philological, linguistic, lexical, semantic) [16].

The proposed data model permits the detailed description of document content, document structure, annotations, and all types of digital restoration processing performed to improve readability.

In the last decades, an increasing number of techniques have been proposed to digitally restore ancient degraded documents through sophisticated digital image processing tools, sometimes making also possible to reveal undisclosed content of great historical interest [1, 2].

Transcription of handwritten documents can be performed manually by an expert, semi-automatically with the computer support to the expert, or automatically by an HTR (Handwritten Text Recognition) system. Experts can also add annotations regarding the linguistic analysis of the transcription. These activities require that a relationship is maintained between the original digital document, the restored version used for linguistic analysis, the transcription, and user annotations.

As soon as manuscript images are processed with this plurality of digital restoration and content analysis tools, a major challenge is the creation of structured digital archives that enable the proper conservation and simplify the access and retrieval to the wealth of data produced. This includes, on the one side, all the acquisition channels available and the subsequent elaborations performed on them, together with the corresponding parameters, when required. On the other side, the document content description, in terms of document structure, text transcription, and annotations containing text analysis is included in the data as well.

Application services supporting users to analize document content, to improve document quality through digital

The Authors are with the Institute of Information Science and Technologies – Italian National Research Council, Pisa, (e-mail: pasquale.savino@isti.cnr.it).

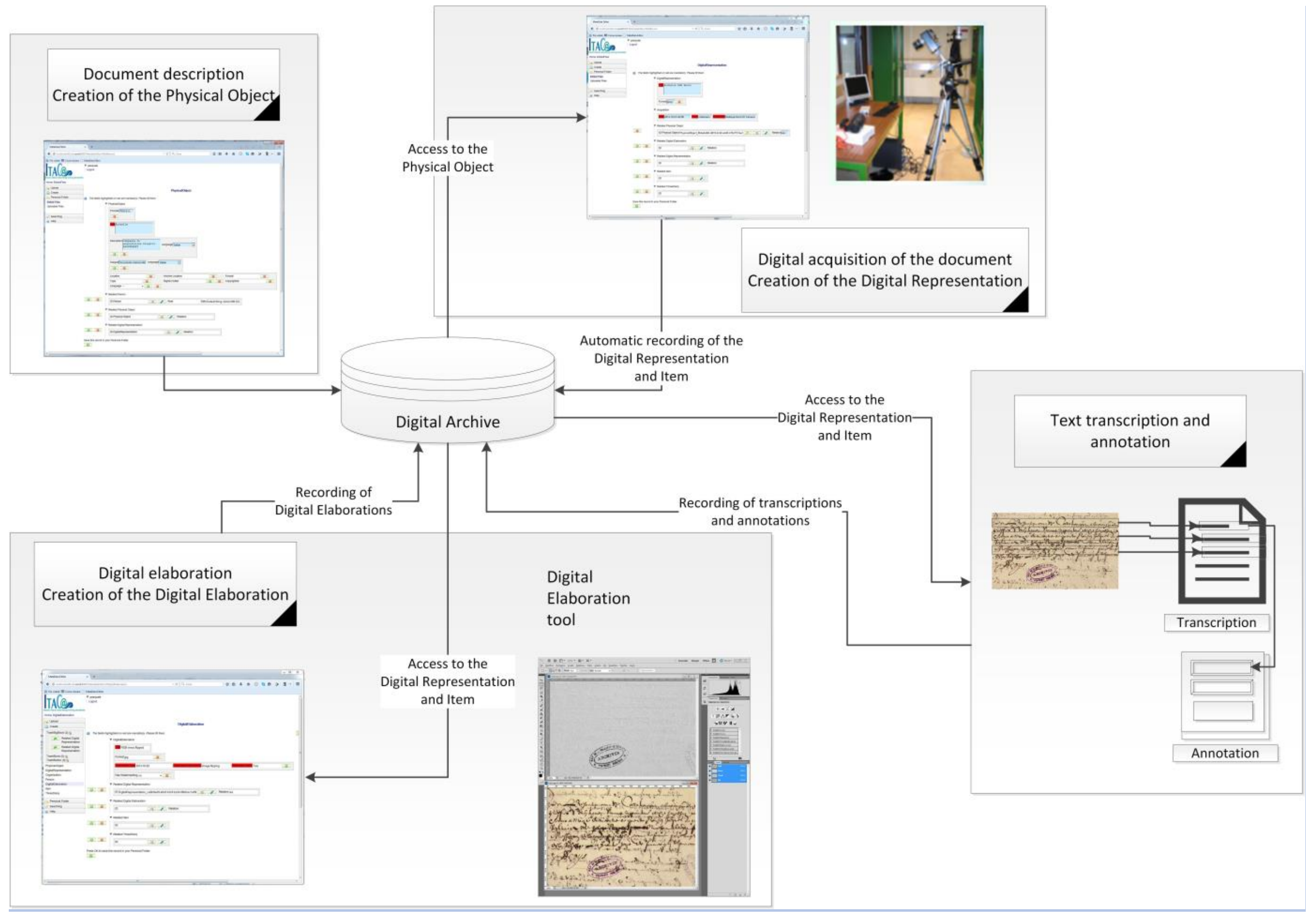


Fig. 1 The Ancient Document Processing life cycle

restoration, to search, retrieve, and visualize documents and elaborations performed on them should be provided. These services are usually tailored to needs of different user categories: linguistic experts, digital restoration experts, general public.

Linguistic experts should be able to access the documents, either in their original form or after digital restoration, and add to them a textual transcription as well as linguistic annotations. During this activity, experts should have the possibility to search other documents with similar content, both from a visual point of view and from a content point of view.

Digital restoration experts could take advantage from an application that describes the degradations that the document suffered over time, and from the possibility to keep track of all procedures adopted and the parameters used to achieve any specific virtual restoration result. Maintaining a documentation of the virtual restoration activities would enable the application of the same process to other documents with similar damages, and the comparison of the results achieved when different parameters are used.

An application for the general public should support the search of a document based on its content, and the visualization of the original digital representation together with the best restored version and the document transcription. Simple linguistic annotations should be shown on user request.

The life cycle of an ancient document processing system supporting all the functionalities described above is shown in Fig. 1. This life cycle, described more in detail in section II, will be used through the entire paper to illustrate the details of the activities performed to process and use ancient manuscripts.

In the last decades, great emphasis was given to the preservation of Cultural Heritage and to make easier the access to Cultural Heritage artefacts for purposes of work, learning or leisure. Many projects and research initiatives invested in the development of software tools and services for the creation of digital archives of Cultural Heritage objects. However, none of them supports all phases of the document life cycle. For example, the e-codices [20] and Europeana [3] initiatives are mainly dedicated to catalogue documents with descriptive metadata used for document browsing and retrieval. In particular, Europeana, an initiative of the European Union for the creation of Digital Libraries for Cultural Heritage, defined a metadata model for descriptive metadata (EDM – Europeana Data Model [11]). Mapping rules and tools from other existing metadata schemas and EDM have been defined. Europeana provides a web application to search and visualize the cultural heritage objects. However, it does not provide any support to the description of user annotations and elaborations performed on the objects. Other research projects invested in supporting annotation and transcription of digital manuscripts (see, for example, *Pelagios* [21], *Transkribus* [22]). In particular, Transkribus is a software platform dedicated to the automatic transcription of handwritten documents. Users may provide the system with their manual transcriptions that are used for training. However, limited support is provided to archiving and search, and digital image processing activities are not managed. Finally, there are initiatives dedicated to support archiving and interoperability of digital image repositories. An example is IIIF (International Image Interoperability Framework) [23], which offers APIs for image representation, presentation, and content search.

In this paper, we propose a metadata schema model to describe a manuscript and its entire life cycle, focusing on the liaison between the digital document and its digital elaborations. The proposed metadata schema extends existing metadata representations, describes the semantic content of a document as a whole, and mainly introduces the possibility to relate the digital elaborations with the digital document. Moreover, we illustrate a Metadata Editor Tool (MET) that supports the creation, editing, and search of metadata records, exploiting the Digital Archive, based on the MILOS Multimedia Content Management System [12], which supports archiving and content-based retrieval of digital document representations and metadata associated with them. The paper also includes a complete running example of document elaboration and archiving, based on the manuscripts acquired and elaborated in the national project Itaca [4].

II. The Ancient Document Processing Life Cycle

The digital processing life cycle that ancient manuscripts must undergo in order to ensure their conservation, legibility, accessibility and interpretation includes the following steps (Fig. 1):

1. Document digitalization, mainly based on Multispectral Imaging or even X-Ray Fluorescence. Acquiring manuscripts in digital form guarantees their preservation from further degradations, and can be considered as a preliminary tool to enhance their legibility in case of severe damages (e.g. erased texts in palimpsests) [5].

2. Creation of descriptive metadata. This is available in many existing digital archives (e.g. e-codices, Europeana). Metadata, such as Title, brief textual description, conservation place, etc. are associated with the document.

3. Digital enhancement and restoration. This is the process of removing or attenuating in the digital representations of the documents all degradations due to ageing or mistakes of the human intervention during conservation or physical restoration.

   The most typical degradations are ink diffusion and fading, blurred or low-contrasted writings, seeping of ink from the reverse side (bleed-through), spots, and noise. In addition, it may be necessary to correct the distortions introduced by the acquisition system, such as an incorrect setting of the equipment, or the effect of transparency from either the reverse side or from subsequent pages (show-through) often occurring during the scan process. The correction of these degradations may require the application of several different and sophisticated image processing techniques and the comparison of their results. Thus, it may happen that many interventions are required on the same document before a significant result is achieved. Sometimes, none of the results is perfect but each approach may produce a specific type of improvement (e.g., enhance the contrast for improving text legibility and remove complex backgrounds or spots) [6, 7, 8].

4. Digital analysis of the manuscript contents. This is the process of generating descriptions of the manuscript related to specific features (e.g., colour, textures) or in simplified forms (e.g., text binarization), in order to facilitate annotation, transcription, interpretation, etc.

5. Document structure recognition. Simple layout structures such as pages, paragraphs, lines of text, figures, etc. can be automatically detected. However, even complex logical structures can be recognized.

6. Textual content extraction and analysis. Experts frequently need a transcription of the text contained in the manuscript, in order to facilitate the philological and linguistic analysis and to enable the comparison with other manuscripts.

   Textual transcription can be performed either manually or by using automatic and semi-automatic tools (such as Transkribus). The possibility to search for similar document parts may help the expert during this process. Then, philologists and linguists may add several layers of annotations containing semantic, lexical, and philological interpretation of the document content. The transcript and the annotations must be connected to the corresponding parts of the digital document.

7. Archiving of digitized documents and all their elaborations, by also recording the possible multiple processing steps that produced each result along with the parameters used. This allows taking advantage of previous experience gained on a document for the processing of other documents with similar degradations and/or characteristics. The storage and retrieval support in text transcription, as well as the text annotations, may be used to aid further linguistic analysis performed on the document, through a comparison with other documents.

8. Support to content-based document retrieval, based on all metadata available. Descriptive metadata as well as metadata describing elaborations performed can be used. As an example, it should be possible to search for all documents written by a certain author or in a given period (by using descriptive metadata), or for document lines containing a certain text (by using the annotations or transcriptions) or, finally, for the original acquired document together with a digital restoration and the various related elaborations.

9. Development of end user applications, such as tools for document annotation, document processing, document

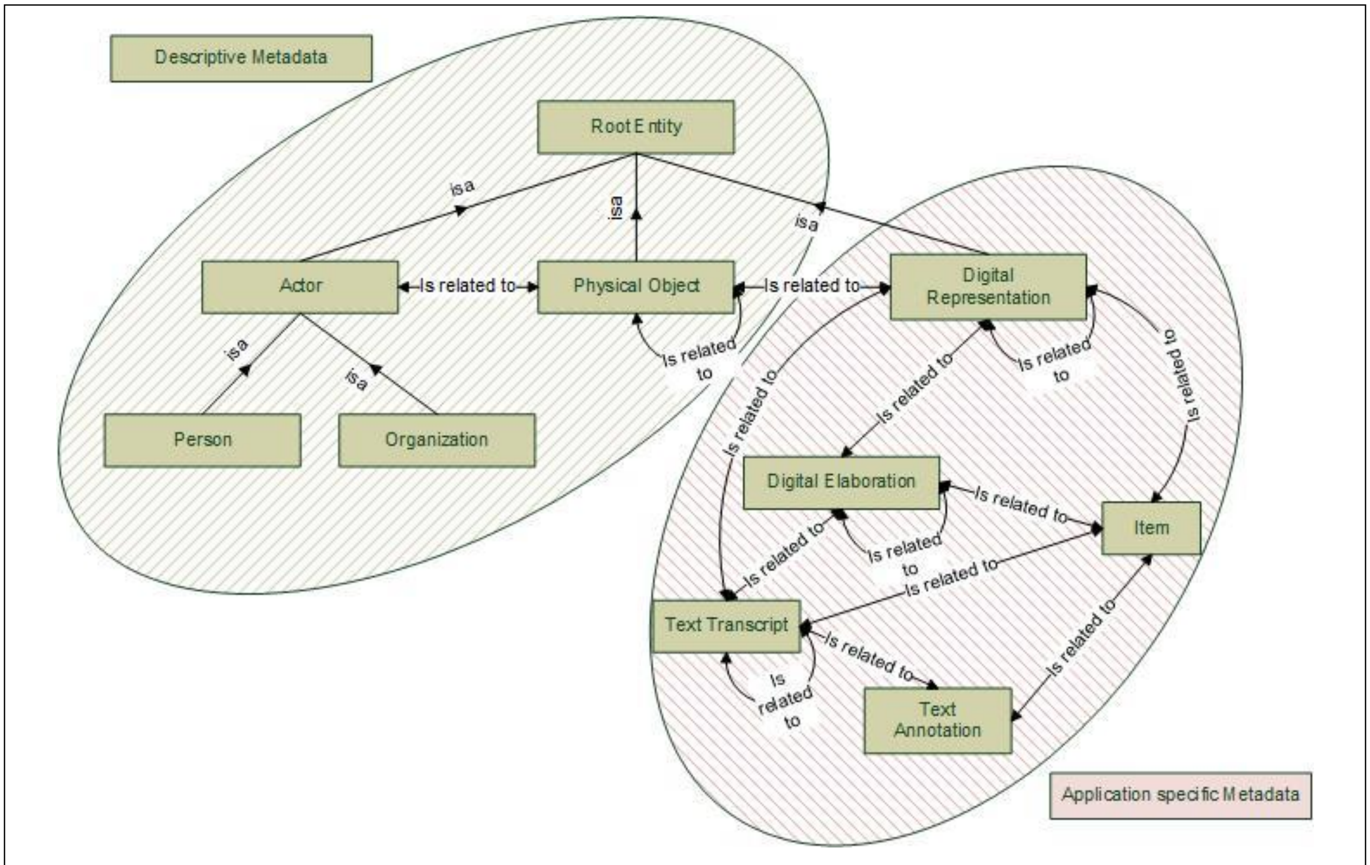


Fig. 2 Graphical representation of the metadata schema

brwosing and retrieval. These tools typically access documents and document elaborations stored in the archive and present them to the user.

The complete life-cycle is supported by a specific metadata schema that permits the description of the original document together with all elaborations performed and all relationships between the elaborations and parts of the document. It also requires a tool, the Metadata Editor (MET), which helps the user to create the metadata values associated with the document. Finally, it requires a document archiving system that enables the storage of digital representations and all the metadata values. The Document Archive will also enable document retrieval based on document content and on metadata values.

III. The Metadata schema

In the field of Cultural Heritage several metadata schemas were defined and used by different institutions [9, 10, 11, 17]. Usually, document content description is limited to the use of descriptive metadata of the digital artefact, while the digital elaborations performed on the digital object are not recorded.

The proposed metadata schema is structured, so that it can support the inclusion of (i) descriptive metadata, such as for example bibliographic metadata, and (ii) application specific metadata, in order to describe all elaborations performed on the document. The first part is typically based on a standard metadata schema: we consider the adoption of the Dublin Core standard (DC) [9], in order to maintain the compatibility with existing archives. However, the MILOS archiving system and the metadata editor described in the following enable the adoption of any schema having an XML representation. For example, it is possible to use the FRBR (Functional Requirements for Bibliographic Data) standard [17] for bibliographic records. Thus, institutions that have already performed the digitalization by using a specific metadata schema for bibliographic records may save the work already done. The extensions carried out enable the description of the complete acquisition process, and the description of the different processing activities performed on the digital representation of the object.

In many digital libraries archiving cultural heritage objects, the description of an object is composed of metadata associated with the Physical Object – a unique man-made object stored in the Museum or Archive, such as a photograph, a printed document or a manuscript, a painting, a sculpture, a vase – and a Digital Representation (DR) of the object, i.e., the visual surrogate or reproduction of a Physical Object. Usually, the DR consists of a single image, with few attributes that describe its physical characteristics (e.g. format, resolution, etc.), the acquisition parameters, such as acquisition date, equipment used, etc. The retrieval is mainly performed by using the attributes of the Physical Object.

The proposed metadata schema includes these descriptions, maintains the compatibility, and supports the interoperability with other existing metadata schema by using all DCMI

Metadata Terms [9] for the Physical Object description. Where possible, the metadata element names directly match the DCMI element names. The main difference is in the introduction of more entities than just the Physical Object, as well as in the qualification of the DC Relation that is expected to be refined within the DC standard. By modifying the relationship between the Physical Object and the DR it is possible to integrate the metadata schema with other existing schemas used to describe the Physical Object, instead of Dublin Core. It is also possible to integrate other types of metadata, such as administrative or technical metadata.

Fig. 2 shows the set of entities of the metadata schema involved in the example illustrated on Sec. V. The root entity of the schema is composed of a Physical Object, a DR, and an Actor entity, which describes the creator of the Physical Object and the cultural organization holding it. This first level of the schema is comparable to what is usually provided by existing digital library schemas for cultural heritage objects. However, the proposed metadata schema has many extensions: i) it enables the recording of complex DRs; ii) it supports the description of all elaborations performed on the DR, iii) it records the complete procedures followed to achieve the virtual restoration or the content analysis of the digital object, iv) it enables the storage of the text transcriptions and annotations, and v) it allows to record the document structure.

Indeed, a DR can be composed of a single image, as in traditional digital archives, or it can include more complex structures. For example, a painting or a photograph are digitally represented by a set of images, one for each acquired spectral band. In case of a document in the form of a book or an envelope its digital representation includes also the various acquisitions of each page. This plurality of possible content is described through the Item entity. It consists of the digital object plus metadata describing its format. The result of image processing performed on a DR (or jointly on several DRs) is stored in a new entity type: the Digital Elaboration (DE), which contains the metadata describing the characteristics of the elaboration, whereas the digital objects obtained are described by the Item element. Similarly, the text transcription performed on a DR or on a DE is stored in the Text Transcription Entity that specifies who did the transcription, how it was performed (manually, automatically, etc.), when it was performed. The Item contains the value of the transcript. The Text Transcription Entity is related to the Text Annotation Entity that contains the type of annotation performed (e.g. philological, lexical, semantic), and the author of the annotation. The Item contains the actual value of the annotation.

Of particular importance are the relationships among different entities, as they enable the description of the processes performed on each cultural heritage object, from its acquisition in digital form up to all the digital elaborations performed. All these entities can be related to each other, so that we may have that a Physical Object can be related to Actors, e.g. the author of the manuscript, and to the Organization that maintains it. At the same time, we may have relationships among different Physical Objects. For example, we may have a relationship among documents belonging to the same collection. Similarly, the relationships among DRs may be used to relate the pages composing the document, whereas relationships among DEs may support the description of the document structure. A Physical Object may also have a relationship with one or many DRs. They are, for example, the digital scans of a manuscript, either composed of a single image or several multispectral images. Different DRs can also be related, exactly as the Physical Objects are. Digital processing techniques can be applied to each DR, which then results linked to a DE. It is also possible that several DRs are used as the inputs to a single DE, and we may have that the elaboration of an image may be used as an intermediate step for further processing, so that we may have that a DE is linked to other DEs.

The model supports typed relationships that are used to specify how two objects are related, by using a user defined controlled vocabulary. By changing the relation type it is possible to adapt the model to specific application domains. For example, we may specify that the person related to a Physical Object is the creator, or the researcher that analyzed its content. Similarly, we may specify the type of relation between different Physical Objects, e.g. recto/verso, part-of, etc.

Such a rich description of acquisitions and their processing results is what is archived into the Digital Archive. This model was experimented in different projects [4, 18].

Although we do not describe here in detail the search capabilities offered by the Digital Archive, it is worth mentioning that it supports efficient content-based similarity retrieval on image content, and searches on the metadata. This means that it could be possible to express queries requiring to retrieve all DRs processed through a certain software tool, those that still require a specific processing, those with acquisitions in certain spectral bands, or having image Items similar to those of a given example.

IV. The metadata editor

The Metadata Schema described in Sec. III has been exploited in the Metadata Editor Tool (MET)[1] [19] supporting the creation, modification, and search of documents with their different digital elaborations.

MET is a web-based cataloguing tool that allows to add, edit and delete new metadata records for cultural objects, persons, organizations, digital objects and the elaborations performed on them, create text transcriptions and annotations, as well as establish relationships among them. After editing, the record is stored in a Digital Archive, currently based on the MILOS Multimedia Content Management System [12]. The User Interface of the MET is form-based: the user can write the value for a specific element or choose the correct value among those suggested by the tool, using a drop-down list conforming to controlled vocabularies.

[1] Users interested to use the MET may contact the paper authors. Publicly available at http://multimatch01.isti.cnr.it/MetadataEditor

MET is a general-purpose metadata editor that permits the management of records complaint with any preloaded XML schema. As soon as the metadata schema is loaded, the MET creates an interface supporting the user in the creation and update of metadata values associated with the documents. Records can be created and modified, and relations among them can be established. The MET is a web application developed by using the Google Web Toolkit (GWT). A GWT application consists of two parts: the server and the client part. While the client is turned into an Ajax-JavaScript application running inside the browser, the server part is standard Java code running inside the web container (Fig. 3).

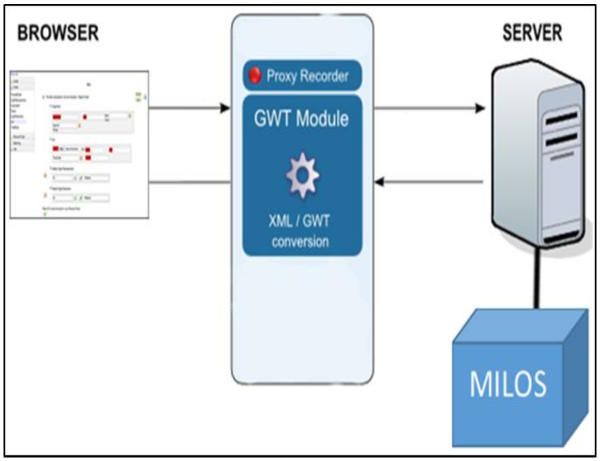


Fig. 3.Overview of the MET Architecture

MET uses the MILOS (Multimedia dIgital Library for On-line Search) system to support the storage and the retrieval of any multimedia element whose descriptions are provided by using arbitrary metadata models represented in XML.

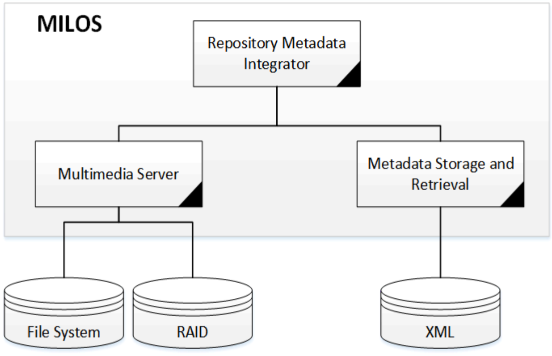

MILOS
Repository Metadata Integrator
Multimedia Server
Metadata Storage and Retrieval
File System
RAID
XML

Fig. 4. Milos Architecture

A. The MILOS Multimedia Content Management System

MILOS is Multimedia Content Management System (MCMS) supporting the storage of multimedia documents described by any XML compliant metadata schema. Data can be distributed over different storage devices, while searches on multimedia document content (text, and images) as well as XML attributes are supported. MILOS MCMS is composed by three main components, as highlighted in Fig. 4:

- the Metadata Storage and Retrieval (MSR),
- the Multi Media Server (MMS) and
- The Repository Metadata Integrator (RMI).

MSR manages the metadata and the different storage and search functions of heterogeneous metadata. MMS enables to manage documents composed of different media and registered with appropriate archiving strategies. Finally, RMI supports the integrated access to MMS and MSR and it maintains the mapping between the metadata schemas "seen" by the services of the Digital Library and those used internally by MCMS. Many different tools can be created on top of MILOS MCMS, such as a Metadata Editor Tool, a Content Based Search Tool, a Content Browsing Tool, etc. The MILOS system can be used in different types of Digital Libraries, such as indexing services, document insertion, search, navigation, etc...

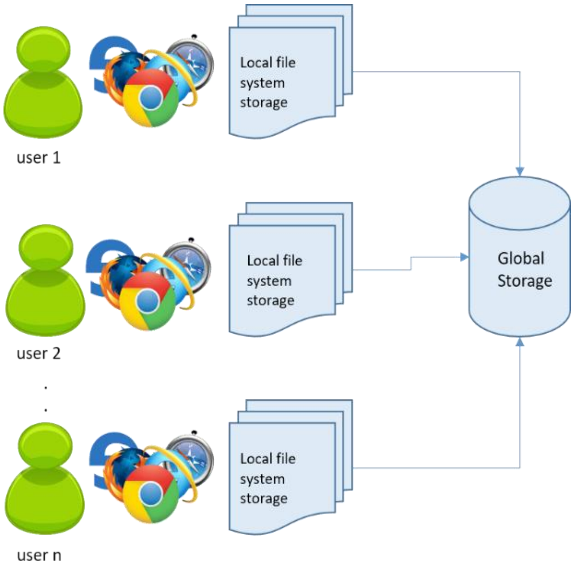
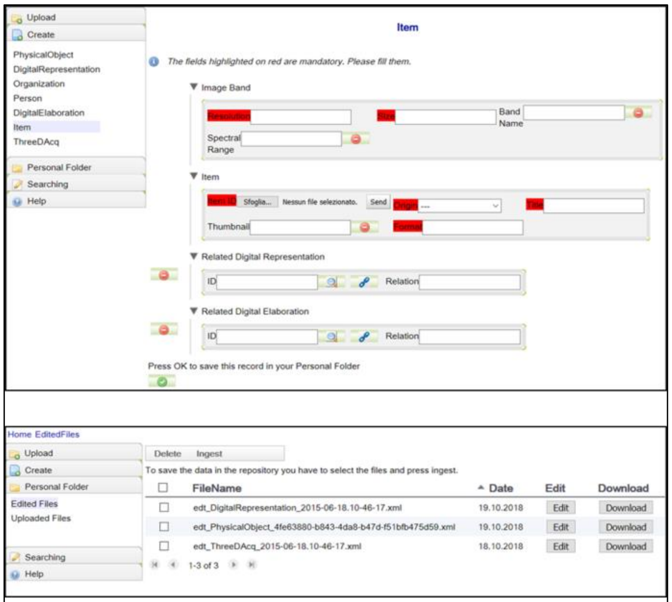

Local file system storage
user 1
Local file system storage
user 2
.
.
Local file system storage
user n
Global Storage

Fig. 5. Local and Global Storage



Item

The fields highlighted on red are mandatory. Please fill them.

▼ Image Band

Resolution
Size
Band Name
Spectral Range

▼ Item

Item ID Sfoglia... Nessun file selezionato. Send
Origin ---
Title
Thumbnail
Format

▼ Related Digital Representation

ID
Relation

▼ Related Digital Elaboration

ID
Relation

Press OK to save this record in your Personal Folder

To save the data in the repository you have to select the files and press ingest.

□	FileName	▲ Date	Edit	Download
□	edt_DigitalRepresentation_2015-06-18.10-46-17.xml	19.10.2018	Edit	Download
□	edt_PhysicalObject_4fe63880-b843-4da8-b47d-f51bfb475d59.xml	19.10.2018	Edit	Download
□	edt_ThreeDAcq_2015-06-18.10-46-17.xml	18.10.2018	Edit	Download

1-3 of 3

Fig. 6. From the top: an example of Edit form, an example of Personal Folder Interface

MILOS was developed by the NeMIS laboratory at ISTI-CNR of Pisa, and it has been exploited as back-end service of different digital libraries (*ECD-Enhanced Content Delivery*, *EFG-European Film Gateway*, *Itaca-Innovative Tools for cultural heritage Archiving and restoration*, *@mmira-Acquisizione Multispettrale, Miglioramento, Indicizzazione e Ricerca di artefatti artistici*).

Of particular importance for the storage system is the document search function, typically based on metadata associated with documents and document multimedia content, such as text and images. MILOS search is based on indices that can be created on user defined metadata elements.

B. MET Functionalities

MET provides a web interface to manage metadata records. It also supports user management by providing user logging and authorization access control. Users may be authorized to read only, or to the creation/update of metadata records.

As illustrated in Fig. 5, there is a single central archive containing digital objects and metadata, which is managed through MILOS. The data present in the central archive are visible to all the authorized users. Every user registered in the system has an own personal archive in which the data can be stored during processing. Each user can operate only on the data present in the personal archive. Therefore, if a user wants to modify the metadata in the central archive, this must be first transferred to the personal archive. When the user completes the creation or modification of a set of metadata, these can be transferred to the central archive and made public to all authorized users.

In addition, MET allows the creation of all entities in the uploaded metadata schema. An example of the interface built for the metadata schema described in Sec. III is shown in Fig. 6. In the left menu, a list of the entities that can be created is available. For each entity, the interface proposes a manifold form that allows the insertion of the values associated with the entity based on the preloaded XML schema. The user can write the value for a specific element or choose the correct value among those suggested by the tool using a drop-down list. Furthermore, the MET highlights the mandatory fields as specified on the preloaded schema.

When a user creates a metadata record, this is available only locally in the personal folder of the user. The user can manage the local version of the records using the entry on the left side menu called Personal Folder.

By selecting the "Edited files" field of the "Personal Folder" it is possible to manage the metadata records present in the Personal Folder. For each record, the name and date of creation are displayed, and it is possible to download, delete, ingest or edit each metadata record.

By choosing to modify a record, a new window opens with the record data displayed (Edit button). At the end of the editing, the user can decide whether to register the folders and save the record in the personal folder.

In this way, the user has the possibility to make subsequent changes until ingestion into the public archive. After the ingestion, the object will no longer be visible in the personal folder and will be accessible by all users of the system.

MET also offers simple search capabilities, allowing to select the records stored in the central archive. The search module

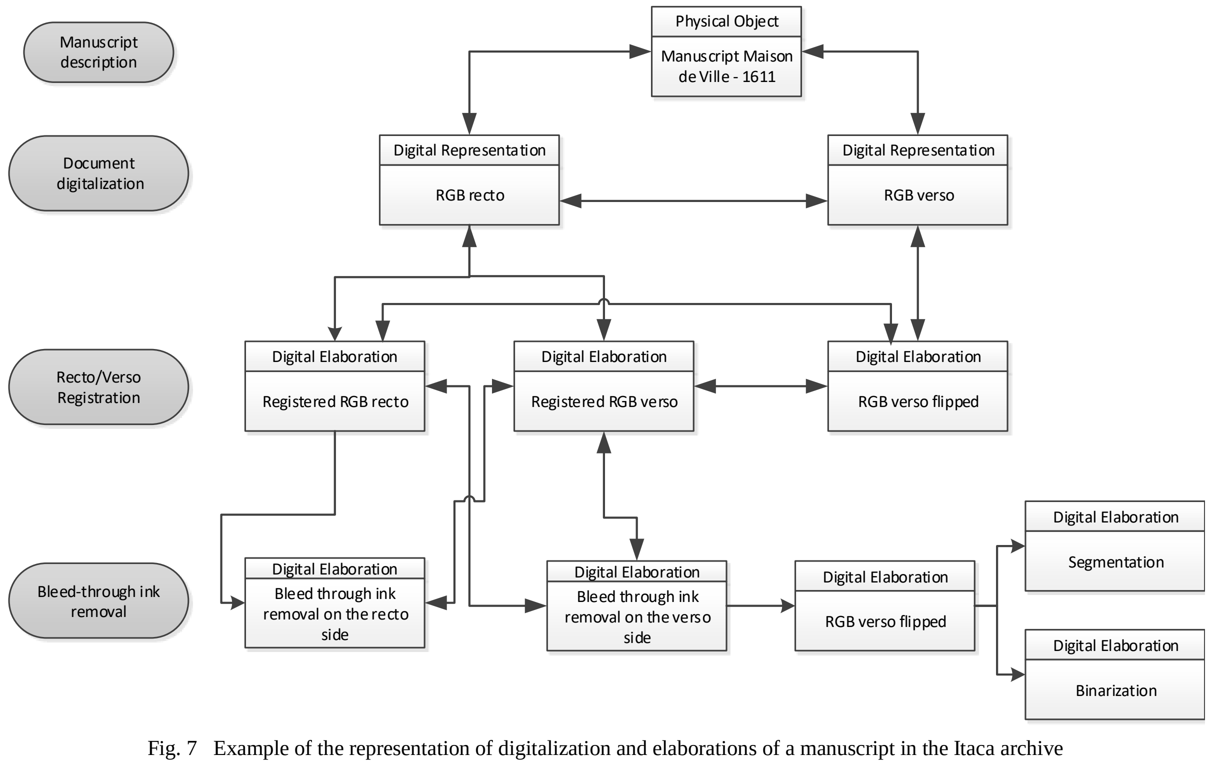


Fig. 7 Example of the representation of digitalization and elaborations of a manuscript in the Itaca archive

uses a form similar to the one used for ingestion, where the user can specify query conditions on all metadata elements. Free text searches are also possible on all metadata schema entity values.

V. A Complete Example

The main phases of acquisition, digitization, and processing of digital representations of an ancient manuscript are illustrated in Fig. 1.

The first phase provides the description of the cultural heritage object, with the creation of the Physical Object performed through the Metadata Editor Tool (MET).

The digital acquisition of the document and the creation of the Digital Representation is initiated by MET by accessing the Physical Object description from the archive. The acquisition process is performed through a multispectral camera [5], which can produce several digital images in different spectral bands, and automatically generates associated Digital Representations (DR) containing all the acquisition parameters used. These Digital Representations are then stored in the archive.

The subsequent phase is devoted to the elaboration of the Digital Representation by using image processing functions developed within the Itaca Project. These functions were integrated as plug-ins into the GIMP [13] image processing tool. A Digital Elaboration (DE), containing all the parameters used, is automatically generated and stored in the Digital Library. The elaborations performed depend on the type of images, their degradation, the desired results, etc.

Fig. 7 illustrates the phases above through a specific example, and it details the metadata elements related to the creation, acquisition and processing of a given manuscript. To simplify

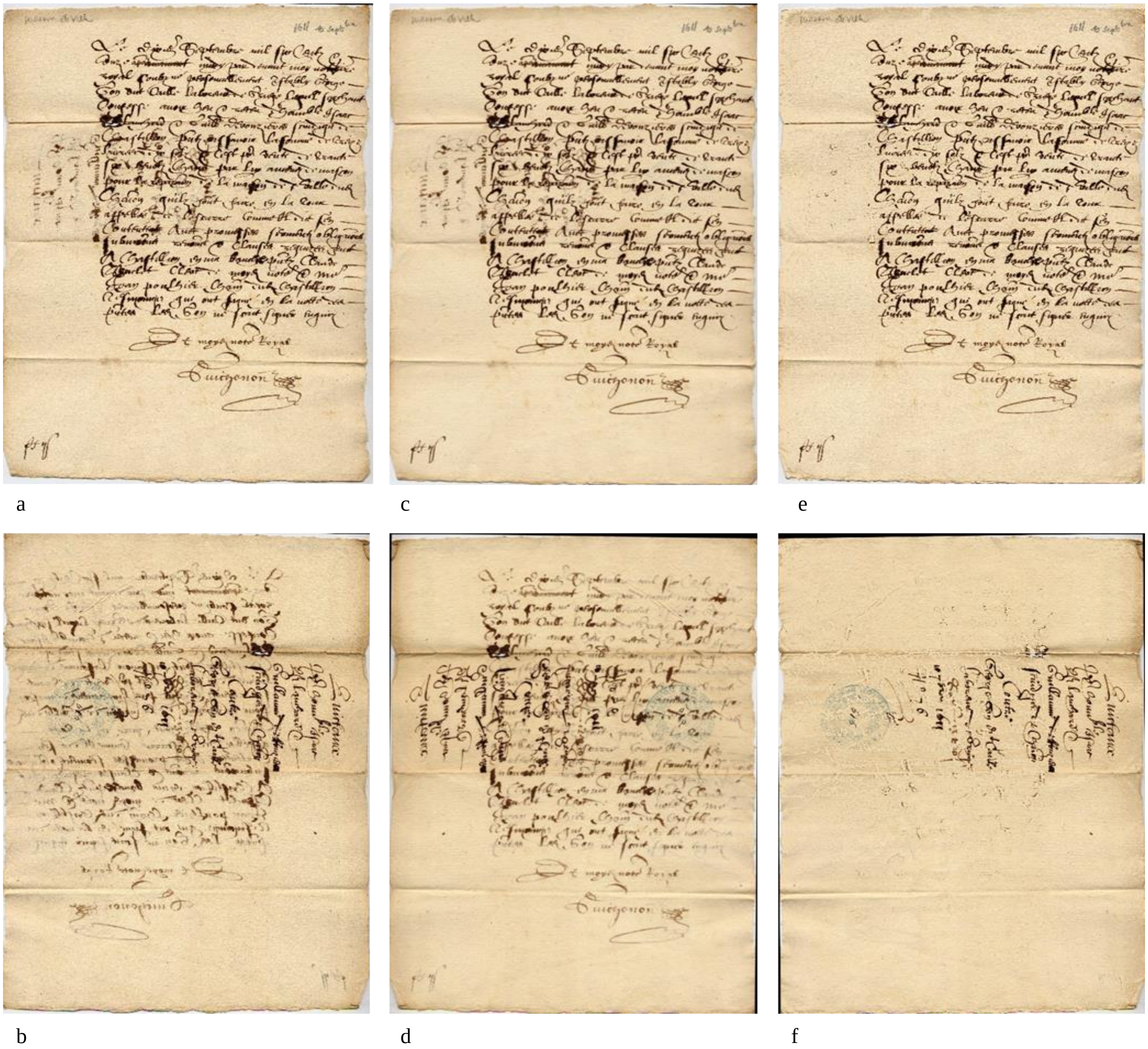


Fig. 8 Example of some of the results achieved when processing a manuscript according to the metadata schema of Fig. 2: (a)-(b) original recto-verso pair; (c)-(d) recto and verso after registration of the verso on the recto; (e)-(f) recto and verso after ink bleed-through removal.

the figure and the presentation, we do not show the Items, which are associated with each DR and DE.

For this manuscript, the recto and verso scans are first processed to correct possible geometric distortions introduced during the scanning process, and to align the verso image with the recto image (or vice-versa) [14]. Subsequently, the registered pair is processed in order to remove bleed-through [6, 15].

The phases depicted on Fig. 7 are:

- *Manuscript description*, with the creation of the Physical Object;
- *Document digitalization*, with the creation of two DRs, one for the recto and one for the verso;
- *Recto-verso registration*, which first produces a horizontally flipped version of the verso side and then registers the recto and the flipped verso;
- *Bleed-through ink removal*, which results in two DEs, one for the recto and one for the verso. The restored verso is then put back to its original asset. Fig. 7 also illustrates the relationships among different metadata entities.

Fig. 8 shows the images produced during the elaborations of the considered manuscript. In particular, Figs. 8 (a)(b) show the originally acquired images of the recto and the verso, and Figs. 8 (c)(d) show recto and verso after flipping of the verso and its alignment on the recto. The algorithm used to perform recto-verso registration is described in [14]. Finally, Figs. 8 (e)(f) show the recto and the verso after automatic ink bleed-through removal performed on the registered recto-verso pair. The algorithm used is based on a model of text overlapping, specifically designed for recto-verso pairs affected by bleed-through [15].

The complex metadata structure shown in Fig. 7 can remain completely hidden to the end user. However, it is useful to build specific applications providing the users with detailed information about the processing performed on the objects. For example, the application may present to the user the original images after acquisition, together with the results obtained by ink bleed-through removal. In addition, it is possible to perform further elaborations on the results achieved so far, for example in order to produce a segmented or binarized version of the restored document.

As shown in Fig. 1, it is also possible to create a text transcription of the document and provide linguistic annotations of the transcript. Fig. 9 illustrates the phases of this process through a specific example, and details the metadata elements related to the creation of text transcriptions and annotations.

In the example, to simplify the presentation, we assume that a single document page is transcribed and annotated. The document is initially subdivided into sections and then, for each section a transcript is produced and described by a Text Transcription. A single transcript may have several annotations provided by experts, containing the linguistic analysis of the text. We assume that the transcript is derived from the DE containing a single document page restored by using a bleed through ink removal algorithm. This page is composed of two sections, represented by two different DEs. Each DE contains a reference to the part of the page containing the section. Then, the description of the text transcription of each section is given in the Text Transcription Entity, containing information about the author of the transcription, the method used (manual, semi-

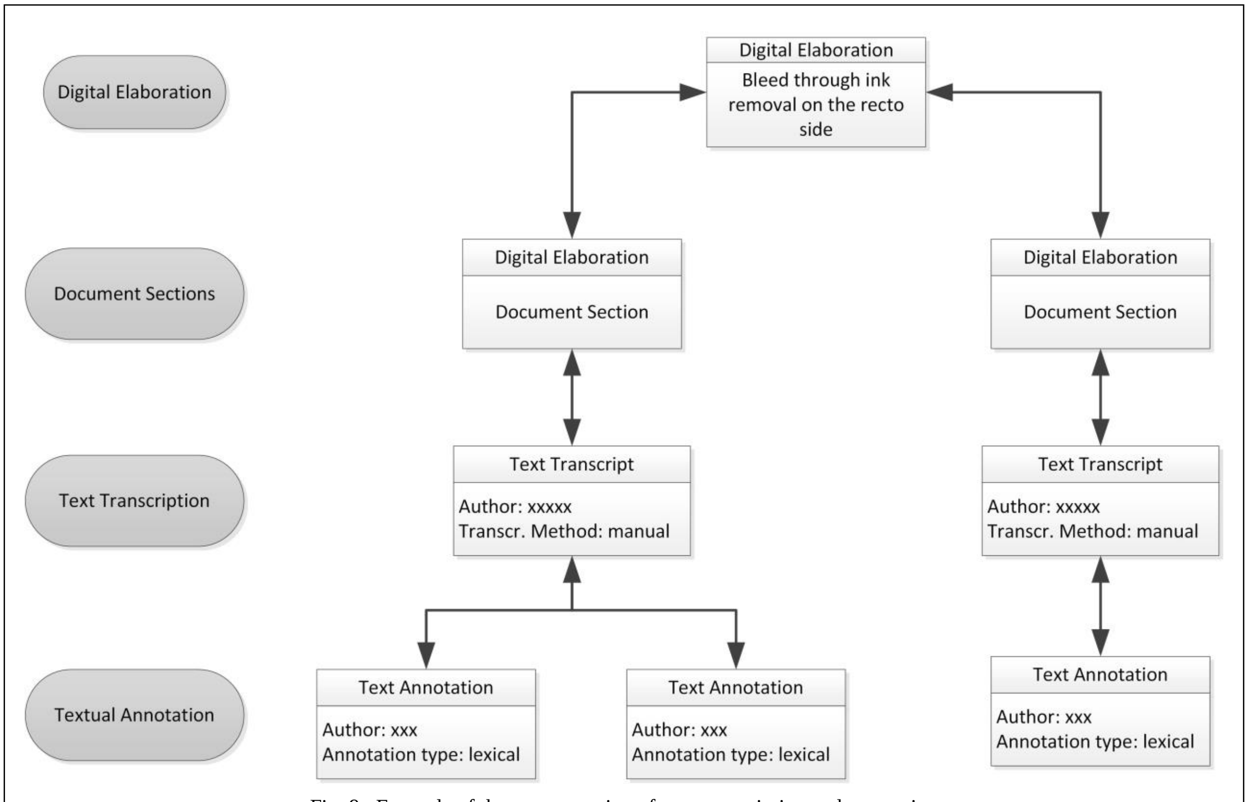


Fig. 9 Example of the representation of text transcription and annotation

automatic, automatic), and the time and date of the transcription. A single section may have multiple transcriptions performed by different authors and different methods. Finally, to each transcription we may associate multiple lexical annotations containing philological, linguistic, lexical, and semantic annotations. Each annotation is represented through a Text Annotation containing the annotation type, annotation date, and annotation author. The actual content of annotations is contained in an Item linked to the Text Transcription (not shown in Fig. 9).

This structure is completely transparent for the end user, since it is managed by the application supporting transcriptions and annotations. The advantages deriving from the recording of all this information can be appreciated when we try to search for the document. In order to give a (limited) idea of the searches that can be performed, in the following we list some typical searches: a) search for documents processed with certain digital restoration algorithms, b) search for documents and document parts containing a given list of terms, c) search for documents where the text contains a certain lexeme, and d) search for documents annotated by a certain expert. This list is not exhaustive, and we must emphasize that combinations of the above searches are possible as well.

VI. CONCLUSIONS

The paper describes the processing life cycle that ancient manuscripts undergo in a specialized digital archive: from their acquisition to their digital processing aiming at improving legibility, up to their annotation and archiving. A metadata model, compatible with those currently used when describing cultural heritage assets, and supporting the description of all phases of the life cycle, is described, together with a complete example of the use of the model. The model supports the description of cultural heritage objects in all possible representations, from the physical object to its various digital representations. The model also supports the complete description of all processing activities performed on the digital object to improve its quality, to extract hidden information, etc. A metadata editor, combined with a multimedia content management system, enables the creation, editing, archiving, and content based retrieval of the metadata elements. The metadata editor is fully integrated with the acquisition and processing components, so that an automatic generation of metadata element values is possible with a limited user intervention. These modules have been experimented in the context of the Itaca Project [4], and an experimental Digital Library has been created.

Further research will include the development and experimentation of the proposed model and archiving system into a tool that offers new innovative and integrated computational and philological instruments for supporting all phases needed to arrive at the production of reliable transcriptions and text-critical editions of ancient degraded manuscripts. Furthermore, the integration of innovative solutions for handwritten text recognition will be considered.

REFERENCES

[1] Knox, K. and Easton, R. "Recovery of lost writings on historical manuscripts with ultraviolet illumination"'Proc. of Fifth International Symposium on Multispectral Color Science (Part of PICS 2003 Conference)', Rochester, NY, 2003, pp. 301--306.

[2] Salerno, E., Tonazzini, A. and Bedini, L. "Digital image analysis to enhance underwritten text in the Archimedes palimpsest," *Int. Journal of Document Analysis and Recognition (IJDAR)* (9:2), 2007, pp. 79--87.

[3] Europeana initiative. https://www.europeana.eu/

[4] Itaca project. http://www.teaprogetti.com/itaca/

[5] E. Console, A. Tonazzini, E. S. P. S. and Bruno, F. "Integrating optical imaging and digital processing for nondestructive diagnosis of artifacts"'Proc. of TECHNART', 2015.

[6] E. Salerno, F. M. and Tonazzini, A. "Nonlinear model identification and seethrough cancellation from recto-verso data," *Int. Journal on Document Analysis and Recognition* (16:2), 2013, pp. 177-187.

[7] Tonazzini, A., Bianco, G. and Salerno, E. "Registration and Enhancement of Double-Sided Degraded Manuscripts Acquired in Multispectral Modality"'2009 10th International Conference on Document Analysis and Recognition', 2009, pp. 546-550.

[8] Tonazzini, A., Salerno, E., Mochi, M. and Bedini, L. "Blind Source Separation Techniques for Detecting Hidden Texts and Textures in Document Images", *in* Campilho, A. and Kamel, M., ed.,'Image Analysis and Recognition', Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 241--248.

[9] DCMI (2008) Dublin Core Metadata Element Set, Version 1.1: Reference Description, http://dublincore.org/documents/dces/

[10] The CIDOC Conceptual Reference Model. http://www.cidoc-crm.org/technical_papers.html

[11] Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C. and Van De Sompel, H. "The Europeana Data Model (EDM)," *World Library and Information Congress: 76th IFLA General Conference and Assembly* (), 2010, pp. 10-15.

[12] Amato, G., Gennaro, C., Rabitti, F. and Savino, P. "Milos: A Multimedia Content Management System for Digital Library Applications"'Proc. of the 8th European Conference ECDL', 2004, pp. 14--25.

[13] GIMP: GNU Image Manipulation Program, http://www.gimp.org/

[14] Savino, P. and Tonazzini, A. "Digital restoration of ancient color manuscripts from geometrically misaligned recto-verso pairs," *Journal of Cultural Heritage* (19), 2016, pp. 511-521.

[15] A. Tonazzini, P. S. and Salerno, E. "A non-stationary density model to separate overlapped texts in degraded documents," *Signal, Image and Video Processing* (9:1), 2015, pp. 155--164.

[16] A.M. Del Grosso, A. Bellandi, E. G. S. M. and Nahli, O. "Scanning is Just the Beginning: Exploiting Text and Language Technologies to Enhance the Value of Historical Manuscripts"'Proc. IEEE 5th International Congress on Information Science and Technology', 2018, pp. 214--219.

[17] IFLA-FRBR (Functional Requirements for Bibliographic Records). https://www.ifla.org/publications/functional-requirements-for-bibliographic-records

[18] Artini, M., Bardi, A., Biagini, F., Debole, F., La Bruzzo, S., Manghi, P., Mikulicic, M., Savino, P. and Zoppi, F. "Data Interoperability and Curation: The European Film Gateway Experience", ed.,'Digital Libraries and Archives', Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 33--44.

[19] F. Debole, E. Salerno, P. Savino, and A. Tonazzini, "Editing metadata to support the content analysis, storage and retrieval of ancient documents ", Proc. 5th International Congress on "Science and Technology for the Safeguard of Cultural Heritage in the Mediterranean Basin" Proceedings, vol. III (2nd Part) pp. 180 - 185. Valmar, 2012.

[20] e-codices . https://www.e-codices.unifr.ch/it

[21] *Pelagios.* https://pelagios.org/

[22] *Transkribus.* https://transkribus.eu/Transkribus/

[23] IIIF (International Image Interoperability Framework). https://iiif.io/