Archiving and retrieving digital elaborations of ancient manuscripts

Pasquale Savino, Anna Tonazzini, Franca Debole, Emanuele Salerno
ISTI-CNR
Pisa, Italy
name.surname@isti.cnr.it

Abstract— **Digitalization of ancient manuscripts is becoming a standard in libraries and archives. In many cases, manuscripts suffer of degradations that may require performing different types of elaborations on the digital images, in order to improve their readability and analyze their contents. Digital archives containing digital images of manuscripts and all the elaborations performed on these images are thus of primary importance for a complete exploitation of all available information regarding the manuscripts themselves. This paper presents a metadata schema suitable for the management of such an archive. The archive will offer the possibility of describing, storing and accessing all the available manuscript versions, and to search them based on their content.**

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

I. INTRODUCTION

Availability of high quality digital acquisition equipment with affordable costs is making the digitization of old manuscripts a common practice. Currently, several cultural institutions have undertaken or even completed the digitization of their rich documental collections. In many cases, 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 produced degradations that make complicate 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 in Europe, so that seeking out for remedies to restore ancient manuscripts and documents would have an enormous social and technological impact. In the last decades, an increasing number of techniques have been attempted to digitally restore this wealth of data through sophisticated digital image processing tools**,** sometimes making also possible to reveal undisclosed content of great historical interest [1, 2]. Due to the variety of degradations and characteristics of the documents, it might also be necessary to attempt several different techniques on a same manuscript in order to remove all the degradations and extract the entire hidden content.

As soon as manuscript images are processed with this plurality of algorithmic tools, a major challenge is the creation of structured digital archives that enable the proper conservation and simplify the access to the wealth of data produced. These include all the acquisition channels available and the subsequent elaborations performed on them, together with the corresponding parameters, when required. This rich description of acquisitions and of processing results should support the archiving of the digital manuscripts and their retrieval, based on the characteristics of the image processing technique used. At the same time, the availability of traditional descriptive metadata should support content based search, such as those usually done in a Digital Library. Possibly, all these metadata should be provided with limited user intervention by automatizing, as much as possible, the cataloguing process.

This would allow building applications where a document is shown in all its forms, from the originally acquired images up to the results of all elaborations performed on them. As an example, it would be possible to build applications that describe the degradations that the document suffered over time, and to keep track of all procedures adopted and the parameters used to achieve any specific virtual restoration result. This would enable to maintain a documentation of the virtual restoration activities, so that the same process can be easily applied to other documents with similar damages, or to compare the results achieved when different parameters are used.

In this paper, we propose a metadata schema model to support such a combination of classical and new ways of describing a manuscript and its analysis process, and illustrate a Metadata Editor Tool (MET) that supports the creation, editing, and search of metadata records. The proposed metadata schema, extends existing metadata representations (see, e.g., [3]), and describes the semantic content of a document as a whole, and that of the results obtained after its processing. The paper also includes a complete running example of document elaboration and archiving, based on the manuscripts acquired or elaborated in the Itaca Project [4].

II. THE ANCIENT DOCUMENT PROCESSING LIFE CYCLE

The digital processing life cycle that ancient manuscripts must undergo in order to ensure their conservation, readability, accessibility and interpretation includes the following steps:

1. Document digitization, mainly based on Multispectral Imaging or even X-Ray Fluorescence. Acquiring manuscripts in digital form guarantees their preservation form 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. 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 to compare their results. Thus, it may happen that many attempts are performed 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 readability and remove complex backgrounds or spots) [6, 7, 8].

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

4. Archiving of digitized documents and all their elaborations, by also recording the possible multiple processing steps that produced each result and the parameters used. This allows taking advantage of previous experience done on a document for the processing of other documents with similar degradations and/or characteristics.

5. Development of end user applications, such as tools for document annotation, or for production of manuscript transcriptions. These tools typically access documents and document elaborations stored in the archive and present them to the user.

III. THE METADATA SCHEMA

In the field of Cultural Heritage several metadata schema were defined and used by different institutions [9, 10, 11]. Usually, they limit themselves to provide descriptive metadata of the digital artifact, but do not provide a complete description of all the document processing activities.

The proposed metadata schema has been designed as an extension to the Dublin Core standard (DC) [9], in order to maintain the compatibility with existing archives. 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 to 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 of the object, i.e., the visual surrogate or reproduction of a Physical Object. Usually, the Digital Representation 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. 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.

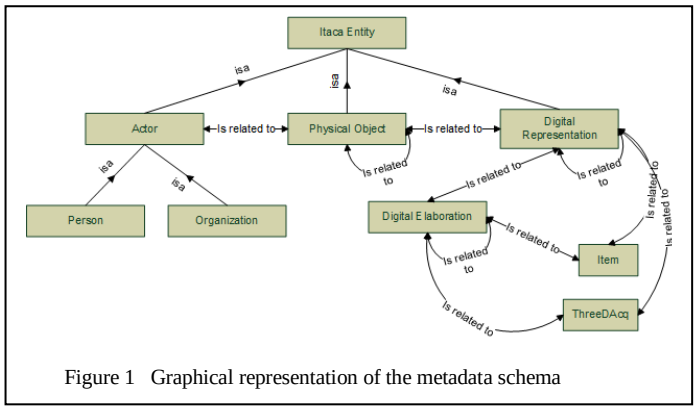

Figure 1 Graphical representation of the metadata schema

Figure 1 presents the complete set of entities of the metadata schema. The root entity of the schema is composed of a Physical Object, a Digital Representation, 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 Digital Representations; ii) it supports the description of all elaborations performed on the Digital Representation; iii) it records the complete procedures followed to achieve the virtual restoration or the content analysis of the digital object.

Indeed, a Digital Representation can be composed of a single image, as in traditional digital archives, or it can include

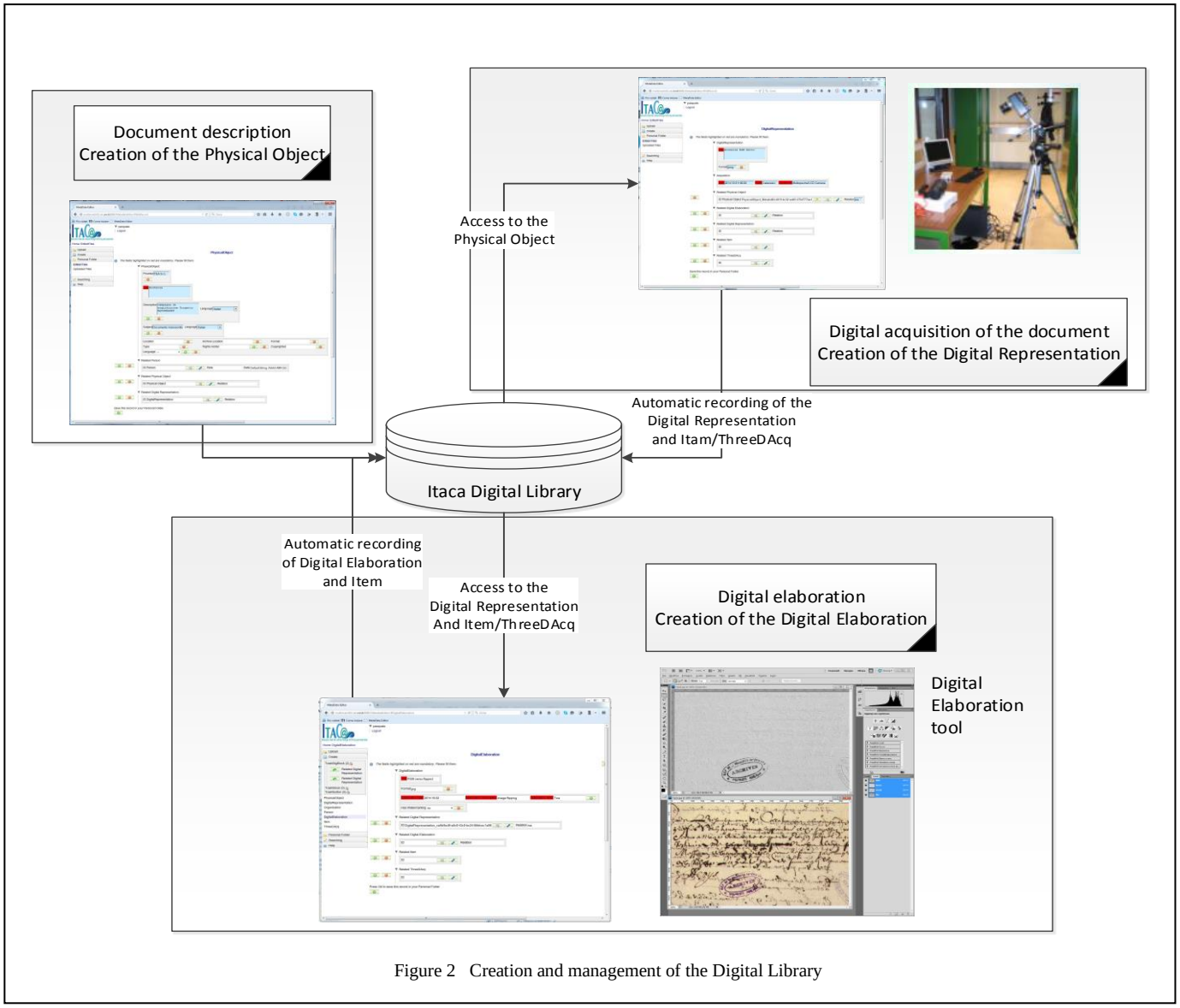

Figure 2 Creation and management of the Digital Library

more complex structures. For example, a painting, a photograph, or a document are digitally represented by a set of images, one for each acquired spectral band, but a sculpture may also be represented by a 3D structure. This plurality of possible acquisitions is described through two specific entities of the model, the Item and the ThreeDAcq. They include the digital object plus metadata describing its format. The result of image processing performed on a Digital Representation (or jointly on several Digital Representations) is stored in a new entity type: the Digital Elaboration, which contains the metadata describing the characteristics of the elaboration, while the digital objects produced are described by the Item and ThreeDAcq elements.

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 artist that created it, and the Organization that maintains the object. At the same time, we may have relationships among different Physical Objects. For example, we may have a relationship among all pages composing a manuscript. A Physical Object may also have a relationship with one or many Digital Representations. They are, for example, the digital scans of a manuscript, either composed of a single image or several multispectral images. Different Digital Representations can also be related, exactly as the Physical Objects are. Digital processing techniques can be applied to each Digital Representation, which then results linked to a Digital Elaboration. It is also possible that several Digital Representations are used as the inputs to a single Digital Elaboration and we may have that the elaboration of an image may be used as an intermediate step for further processing, so we may have that a Digital Elaboration is linked to other Digital Elaborations.

The model supports typed relationships that are used to specify how two objects are related. By changing the relation

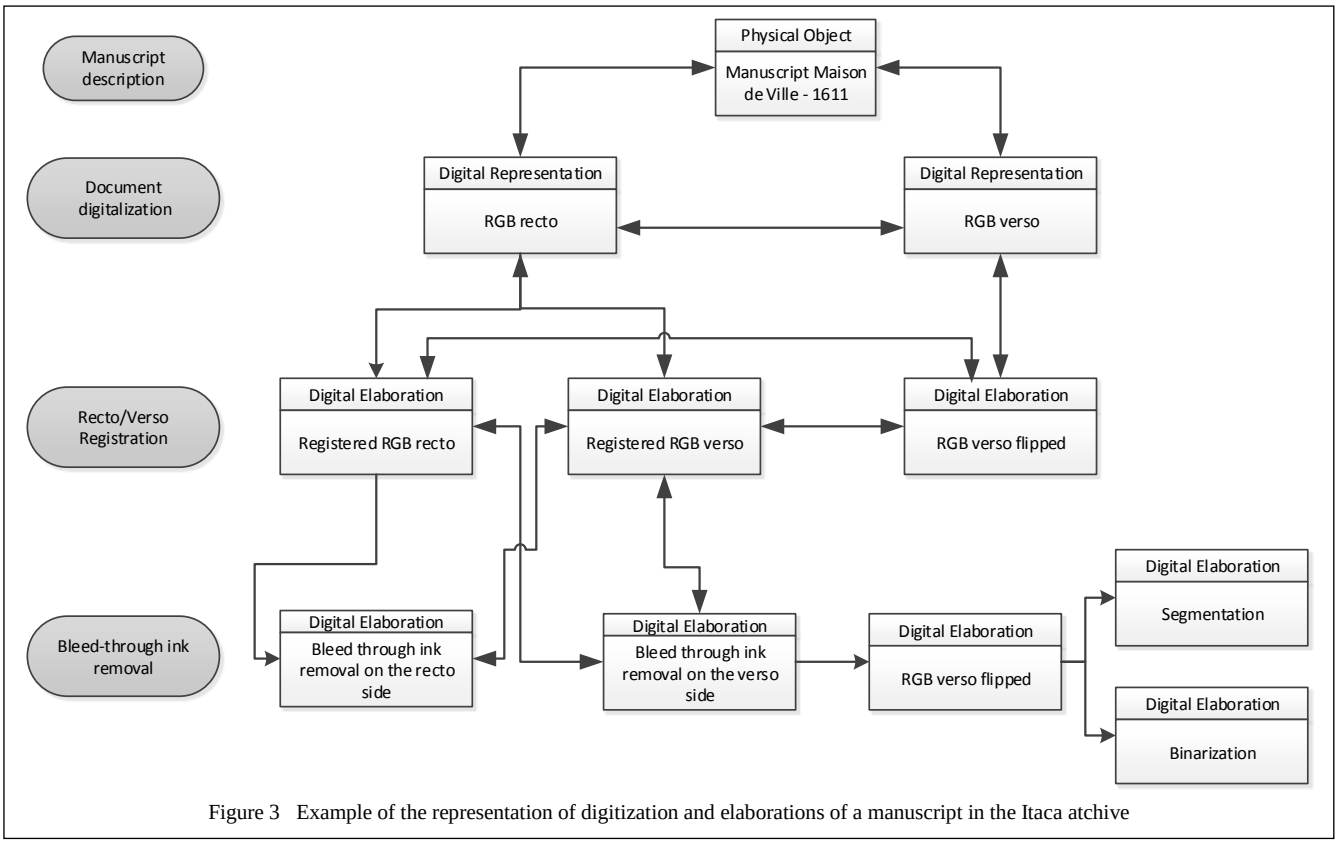

Figure 3 Example of the representation of digitization and elaborations of a manuscript in the Itaca atchive

types 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 person that discovered it. 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, for instance, into the Itaca Digital Library, the digital archive created within the Itaca Project [4], and containing several ancient documents, including those acquired during the Project life.

Although we do not describe here in detail the search capabilities offered by the Itaca Digital Library, it is worth mentioning that it supports efficient content-based similarity retrieval on image content, and searches on the metadata structure and on attribute values. This means that it could be possible to express queries requiring to retrieve all Digital Representations 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 [12].

IV. A Complete Example

The main phases of acquisition, digitization, and processing of digital representations of an ancient manuscript are illustrated in Figure 2.

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 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, as well as to establish relationships among them. After editing, the record is stored in the Itaca Digital Library. 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.

The digital acquisition of the document and the creation of the Digital Representation (Figure 2) is initiated by the 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 ranges and automatically generates associated Digital Representations containing all the acquisition parameters used. These Digital Representations are then stored in the archive.

The subsequent phase is dedicated to the elaboration of the Digital Representation by using image processing functions developed within the Itaca Project. These functions were

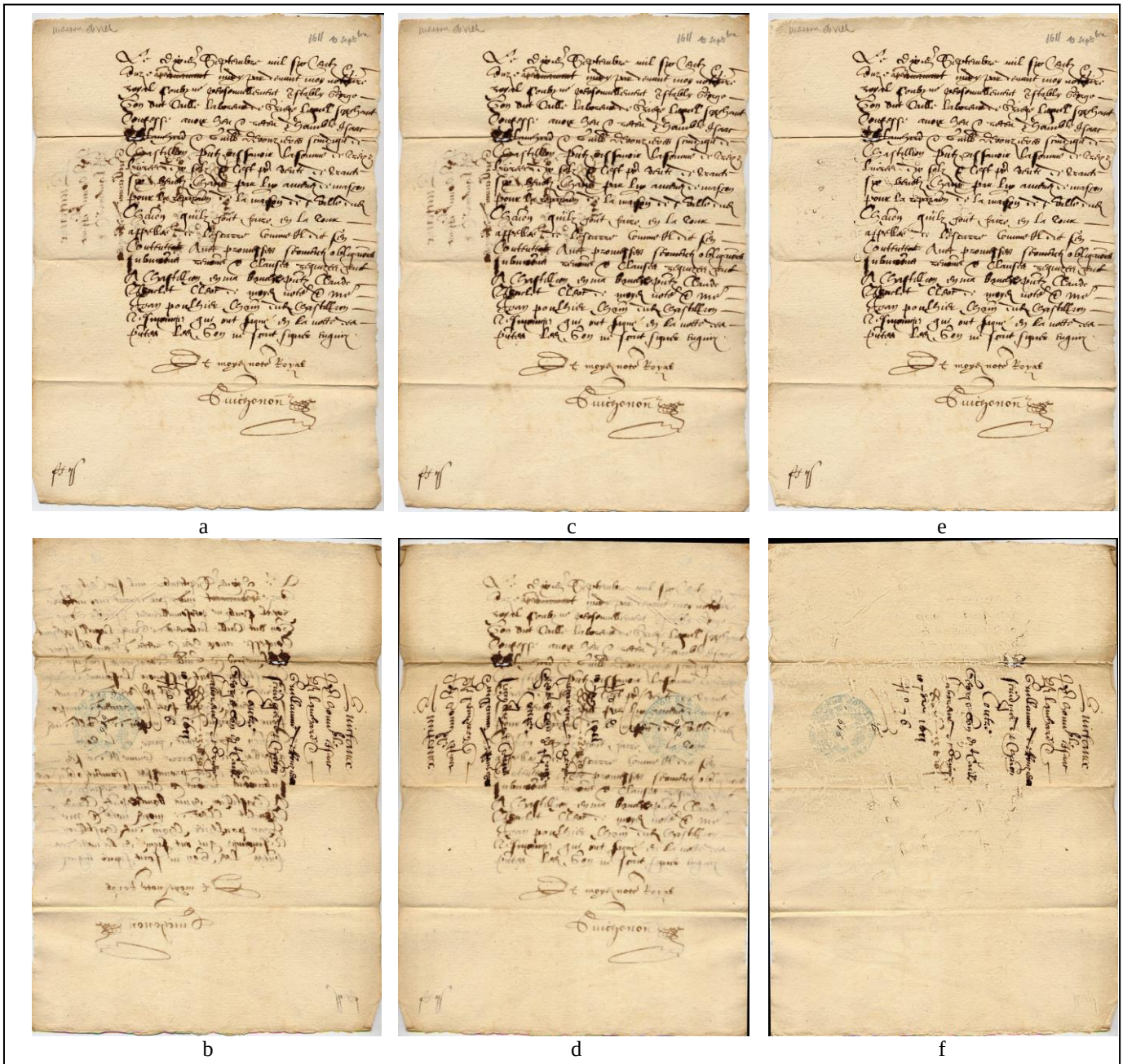

Figure 4 Example of some of the results achieved when processing a manuscript according to the metadata schema of Fig. 3: (a)-(b) original recto-verso pair; (c)-(d) recto and verso after registration of the verso on the recto; (e)-(f) recto and verso after ink bleed-through removal.

integrated as plug-ins into the GIMP [13] image processing tool. A Digital Elaboration, 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.

Figure 3 illustrates the phases above through a specific example, and details the metadata elements related to the creation, acquisition and processing of a given manuscript.

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

The phases shown are: “Manuscript description”, with the creation of the Physical Object; “Document digitalization”, with the creation of two Digital Representations, 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; the “Bleed-through ink removal”, which results in two Digital Elaborations, one for the recto and one for the verso. The restored verso is then put back to its original asset. Figure 3

also illustrates the relationships among different metadata entities.

Figure 4 shows the images produced during the elaborations of the considered manuscript. In particular, Figures 4a-b show the originally acquired images of the recto and the verso, and Figures 6c-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, Figures 6e-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].

It is worth mentioning that the complex metadata structure shown in Figure 3 can be 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.

V. 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 readability, up to their 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 integration 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.

REFERENCES

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

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

[3] Europeana web site. http://www.europeana.eu/portal/

[4] Itaca web site, http://www.teaprogetti.com/itaca/

[5] Console, E., Tonazzini, A., Salerno, E., Savino, P., Bruno, F (2015) Integrating optical imaging and digital processing for nondestructive diagnosis of artifacts, *Proceedings of TECHNART 2015*.

[6] Salerno, E., Martinelli, F., Tonazzini, A.. (2012) Nonlinear model identification and seethrough cancellation from recto-verso data, *Int. Journal on Document Analysis and Recognition*, 16, 177-187

[7] Tonazzini, A., Bianco, G., Salerno, E. (2009) Registration and Enhancement of Double-sided Degraded Manuscripts, *Proc. 10th International Conference on Document Analysis and Recognition*, 546-550

[8] Tonazzini, A., Salerno, E., Mochi, M., Bedini, L. (2004) Blind Source Separation techniques for detecting hidden texts and textures in document images, *Proceedings Int. Conference ICIAR 2004, Porto, Portugal, September 29 - October 1, 2004*, Lecture Notes in Computer Science 3212, 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] EDM – The European Data Model. http://pro.europeana.eu/share-your-data/data-guidelines/edm-documentation

[12] Amato, G., Gennaro, C., Rabitti, F. Savino, P. (2004) Milos: A Multimedia Content Management System for Digital Library Applications, *Proc. of the 8th European Conference ECDL*, Lecture Notes in Computer Science, 3232, 14-25

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

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

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