Building a Digital Library containing digital elaborations of ancient documents

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

***Abstract*— Digital archives containing digitized images and detailed descriptions of cultural heritage objects are of primary importance in order to guarantee the preservation and to foster the fruition of many fragile artifacts of our culture and history. Digital processing of these images is frequently needed in order to improve their readability, to correct degradations and damages, and to analyze their contents. This paper presents a metadata schema and a metadata editor supporting the description and the archiving of all elaboration activities performed. The archive allows one to perform content based searches of the original object's descriptions as well as of the results of the elaboration activities.**

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

I. INTRODUCTION

Collections of unique and exciting cultural content are an important asset of our society and their discovery and understanding has the power to enrich significantly our daily life. Nevertheless, what means do citizens have to access these complex Cultural Heritage (CH) objects and experts to easily discover their content? What are the undergoing activities aiming at their preservation, conservation, and accessibility? Due to recent advances in digital image acquisition methods and the wide range of imaging modalities available, the digitization of our cultural heritage has become common practice. Currently there are several cultural institutions holding rich digital collections of the reproduction of their cultural assets. In many cases, these data are of high quality in terms of resolution, levels of detail, modalities (e.g. we may have multi-spectral digital representations, 3D representations, etc.), and content description. However, conservation, readability, accessibility and interpretation of ancient documents is often compromised by several and different damages that they have undertaken over time, and that continue to cause a progressive decay, so that we undergo the risk to lose much of our past memory during the next years. Furthermore, the fragility of rare or very important historical documents prevents their direct access by scholars and historians. Natural ageing, usage, poor storage conditions, humidity, moulds, insect infestations and fires are the most diffuse degradation factors. In addition, the materials used in the original production of the documents, i.e. paper or parchment and inks, are usually highly variable in consistency and characteristics. All these factors concur to cause document degradations, which, in many cases, prevent from their effective access and understanding. These problems are common to the majority of the governmental, historical, ecclesiastic and commercial archives in Europe, so that seeking out for remedies to preserve and restore ancient documents would have an enormous social and technological impact. Thus far, little attempt has been made to subject this wealth of data to sophisticated digital image processing tools, even though such computational analyses could potentially discover content of great historical interest.

As soon as cultural heritage objects are processed to improve their readability and to extract relevant hidden content, a major challenge is the creation of structured digital libraries that enable the proper conservation and simplify the access to the wealth of ancient documents available. Currently, there are several projects aiming at the creation of digital libraries containing cultural heritage objects [1]. They record the object's description and in some cases support the access to their digital representation, but none of them supports the storage of the plurality of representations and of the result of digital image processing tasks performed. 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 acquired images 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 are provided with limited user intervention by automatizing, as much as possible, the cataloguing process.

This will allow building applications where the document is shown in all its forms, from the originally acquired image up to the result of all elaborations performed on it. As an example, it could be possible to build applications that describe the degradation the document was subject to over time, and to keep track of all procedures adopted and the parameters used to achieve any specific virtual restoration result. This enables to maintain a documentation of the virtual restoration activities that can be used to apply the same process 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 document and its analysis process, and we

illustrate a Metadata Editor Tool (MET) that supports the creation, editing, and search of metadata records. The metadata schema we propose extends existing metadata representations, and describes the semantic content of the documents in its whole, and that of the results obtained after its processing. This work is part of the activities performed in the Itaca Project [2], which aims at the creation of a complete system supporting the acquisition, digitization, image restoration and analysis, archiving and retrieval of digital images of cultural heritage objects [3].

II. DIGITAL PROCESSING OF ANCIENT DOCUMENTS FOR VIRTUAL RESTORATION AND CONTENT ANALYSIS

Document image restoration is the process of removing from the digital representations of documents all degradations due to ageing or to 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 results is achieved. Sometimes, none of the results is perfect but each approach may produce a specific type of improvement (e.g. enhance the contrast and provide a nicer view of the document or improve its readability).

This implies that the Digital Library must allow one to record all elaborations done on the document, by also recording the possible multiple processing steps that produced each result and the parameters used. This allows one to take advantage of previous experience done on a document for the processing of other documents with similar degradations.

Many digital image acquisitions are performed only in grayscale or in the visible range of the spectrum, because of the larger diffusion of the dedicated acquisition equipment. However, owing to specific damages, some documents may be very difficult to read when acquired in grayscale or RGB only. Furthermore, interesting features are often barely detectable in the original visible document, while revealing the whole contents is an important aid to scholars that are interested in dating or establishing the origin of the document itself, or in reading hidden text or features it may contain and that could be the most significant information from a cultural point of view.

It is well known that additional information can sometimes be obtained from images taken at non-visible wavelengths, for instance in the near infrared and ultraviolet ranges. Hence, high resolution, multispectral and multisensory acquisition is becoming an increasingly used practice to help enriching documentation and representation of ancient manuscripts and documents.

Digital image processing techniques can sometimes be alternative to these more costly and specialized acquisition modalities, or, in conjunction with multispectral/hyperspectral acquisitions, they can be used for further improving virtual restoration, for seeking out new information, and for analyzing the document content. For example, pseudocolor images, obtained by suitably composing the available observation channels, are very useful to perceptually enhance and make well distinguishable different texts overlapped in a manuscript, such as faint traces of old erased texts in palimpsests [4]. The creation of pseudocolor images requires a preliminary alignment of the channels to be combined, through image registration techniques.

Still based on registered multiple observations, and assuming that the several distinct information layers contained in a document overlap linearly, statistical blind source separation (BSS) algorithms have proven to be very effective. We applied them to separate those information layers from multispectral single-sided scans, enhance and extract the erased text in palimpsests, reduce the show-through/bleed-through interference in registered recto-verso grayscale and RGB scans, and detect, isolate and extract hidden or masked textures, such as stamps and paper watermarks, or remove, e.g., stains. Constrained maximum likelihood and regularization approaches can be used for the same purposes either in the linear overlapping assumption, or when a more realistic non-linear data model is considered. The above mentioned approaches are described in details in [5,6,7,14,15,16]. Elaborations of the kind described above are or could become standard practice in libraries and archives, thus concurring to enrich the documentation and improve the access and usability of the cultural heritage object.

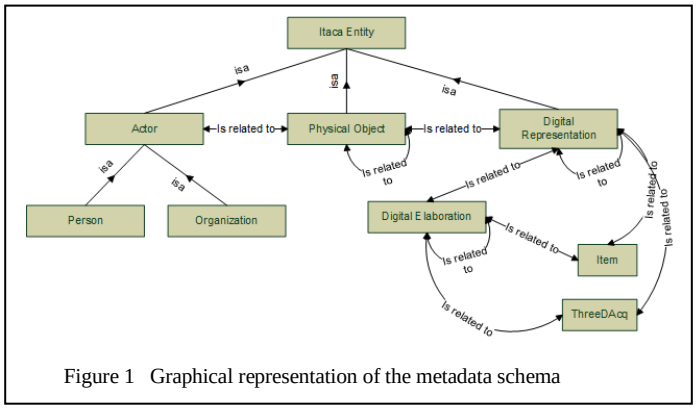


Figure 1 Graphical representation of the metadata schema

III. THE METADATA SCHEMA

In the field of Cultural Heritage several metadata schema were defined and used by different institutions, and here we mention only few of them. Dublin Core [10] is the simplest metadata schema adopted to describe digital objects and currently there are efforts to extend it to better describe Cultural Heritage objects. It is particularly suitable to describe single resources, while it does not easily support representation of relationships among them. The CIDOC CRM Model [17] provides a formal structure for describing concepts and relationships used in Cultural Heritage objects. Finally, the European Data Model (EDM) [18] is the metadata schema used

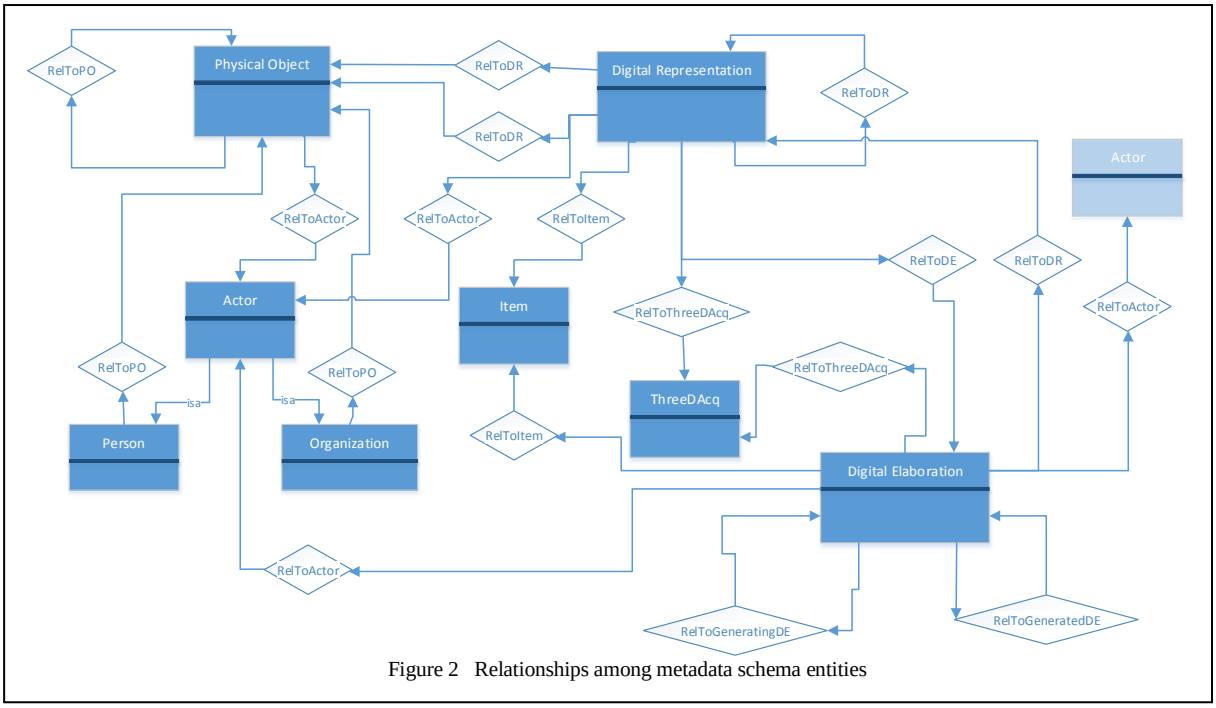


Figure 2 Relationships among metadata schema entities

in Europeana project [1] and it mainly aims at supporting interoperability among different institutions.

The metadata schema we developed for ancient document digital images and their elaborations can be viewed as an extension of existing metadata schemas for digital images of cultural heritage objects and it could be easily integrated with them. Such a schema satisfies two main requirements. On one hand, it supports the representation and description of the cultural heritage object (the ancient document in this specific case), in order to support its retrieval and access. The model also supports the interoperability with other existing metadata representations and reuse of existing resources [8,9]. On the other hand, it supports the description of the complete acquisition process, and the description of the different processing activities performed on the digital representation of the object.

The first requirement is satisfied through the definition of a metadata schema that conforms as much as possible to existing metadata schema standards, and is able to cope with the rich content of cultural heritage digital objects. In addition, the metadata schema includes a description of image content, thus enabling their content-based search.

In many digital libraries archiving cultural heritage objects, their description 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. In many digital libraries, 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 and it maintains the compatibility and supports the interoperability with other existing metadata schema by using all DCMI Metadata Terms [10] 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 Physical Object, as well as in the qualification of the DC Relation that is expected to be refined within the DC standard. Figure 1 shows that 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 when compared to existing metadata schemas for cultural heritage. It enables the recording of complex Digital Representations. It supports the description of all elaborations performed on the Digital Representation. 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 complex structures. For example, a painting, a photograph, or a document is digitally represented by a set of images, one for each acquired spectral band. It also allows one to describe Digital Representations containing a 3D structure. This plurality of possible acquisitions is represented 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 several Digital Representations combined together) 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 enabling the description of the processes performed on each cultural heritage object, from its acquisition in digital form up to all digital elaborations performed (Figure 2). All these entities can be related one 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 scan of a document, either composed of a single image or composed of several images, one for each acquired spectral band. Different Digital Representations can also be related, exactly as 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 input for 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.

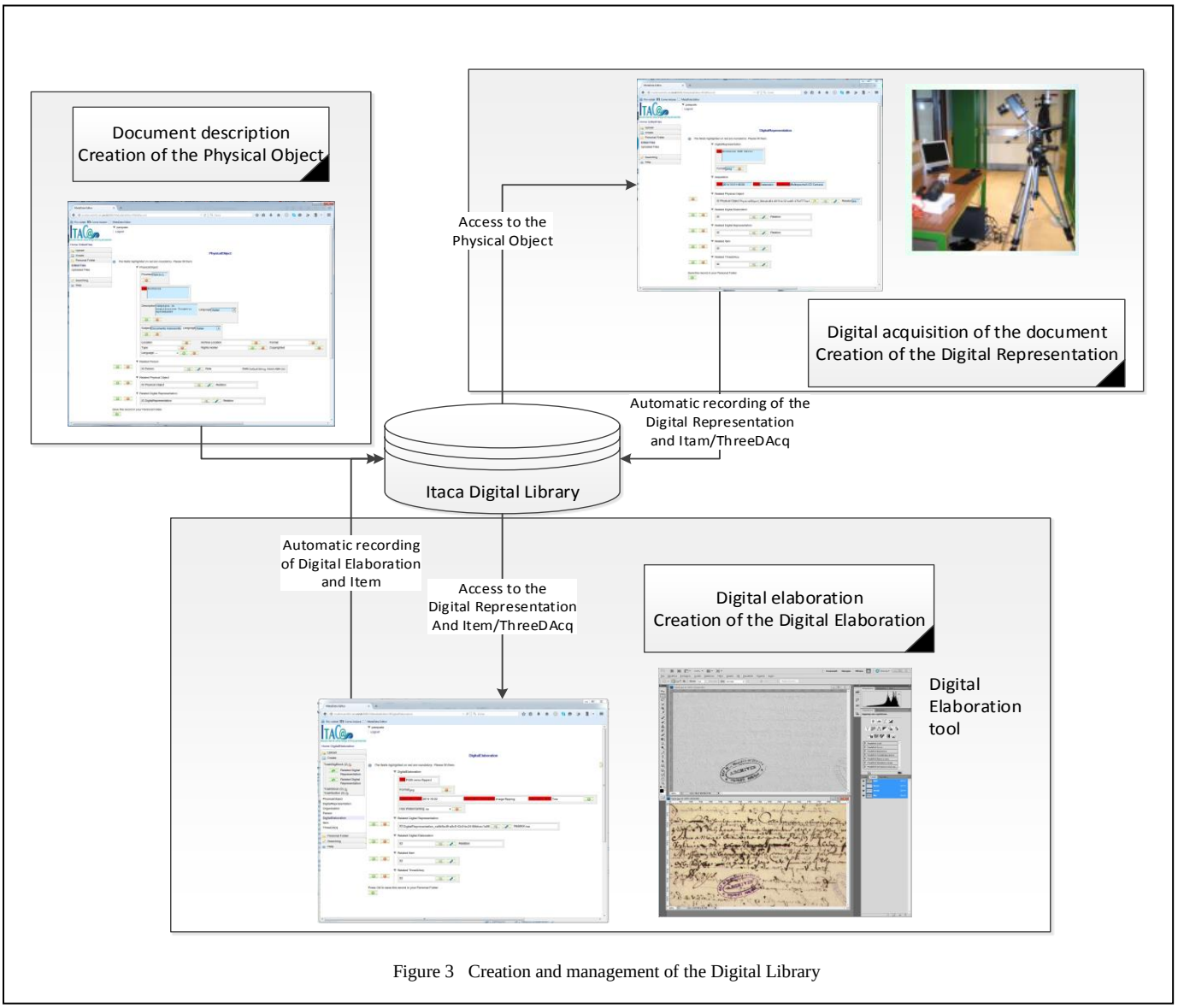
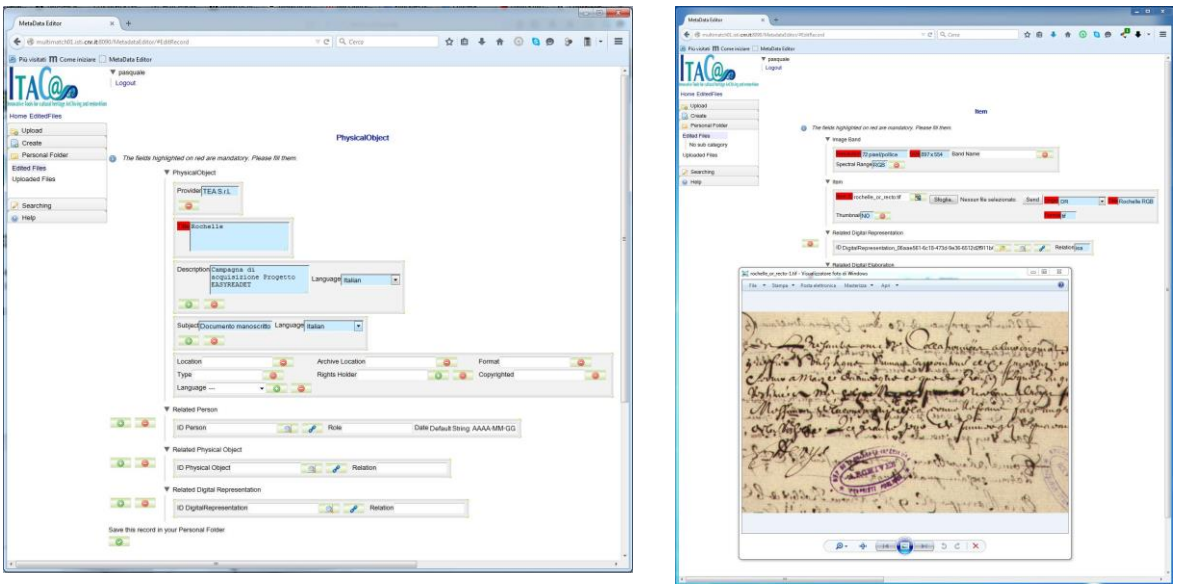


Figure 3 Creation and management of the Digital Library



(a) (b)

Figure 4 Example of the creation of a Physical Object (a) and of a Digital Representation (b)

The model supports typed relationships that are used to specify how two objects are related. By changing the relation 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. This rich description of acquisitions and their processing results are archived into the Itaca Digital Library, the digital archive created in the Itaca Project [2]. It contains all ancient documents acquired during the Itaca Project.

The Digital Library makes use of the MILOS Multimedia Content Management System (MCMS) [11]. MILOS supports simple adaptation to different metadata schemas, schema interoperability and content based retrieval on all metadata

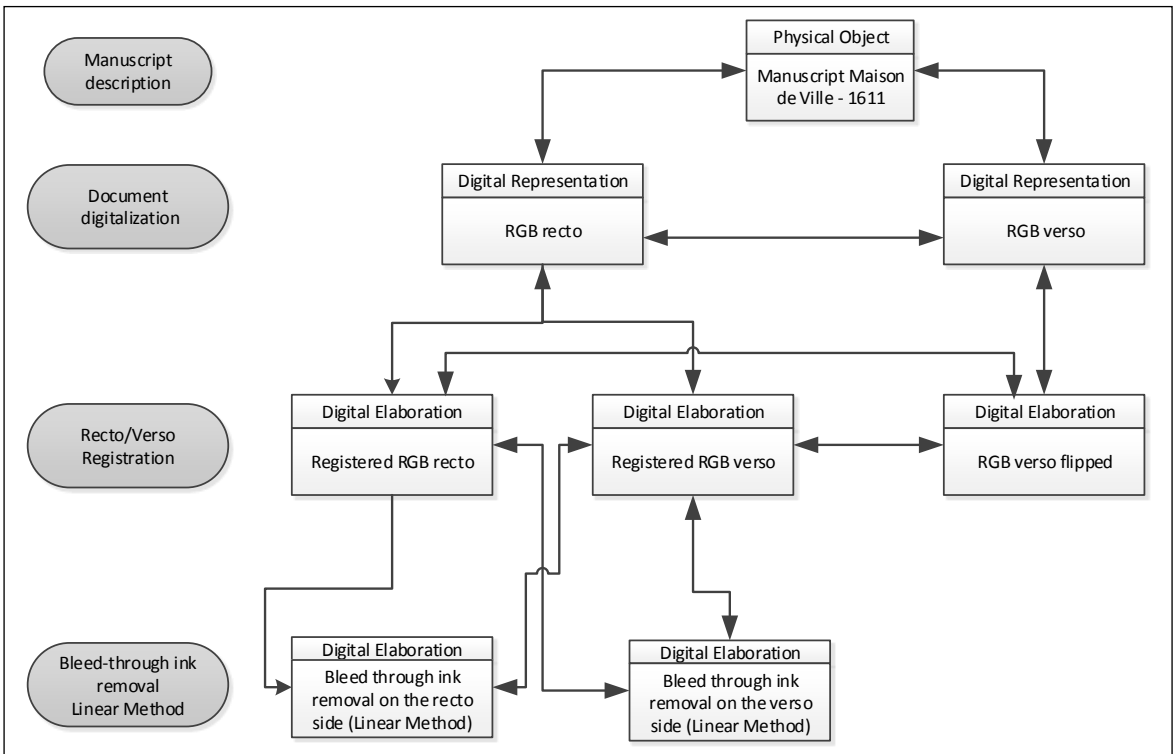


Figure 5 Example of the representation of digitalization and elaboration of a document in the Itaca atchive

elements. It has been customized to manage the metadata schema described so far. MILOS supports searches on all metadata elements of the schema, as well as similarity search on image content. Thus, the availability of traditional descriptive metadata associated to the Physical Objects will support content based search, as usually done in a Digital Library [12].

We do not describe in this paper in detail the search capabilities offered by the Itaca Digital Library. However, it is worth mentioning that is 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 registered through a certain software tool, those that still require a certain processing, those with acquisitions in certain spectral bands, and having image Items similar to a given example. These queries are extremely complex and may be, in principle, quite inefficient. The MILOS MCMS uses specific access structures which allow to represent image objects and metadata values in a unified manner, and to store them by using the Lucene text retrieval system [11], thus achieving efficient retrieval. For example, an image dataset composed of 106 Million images, indexed by

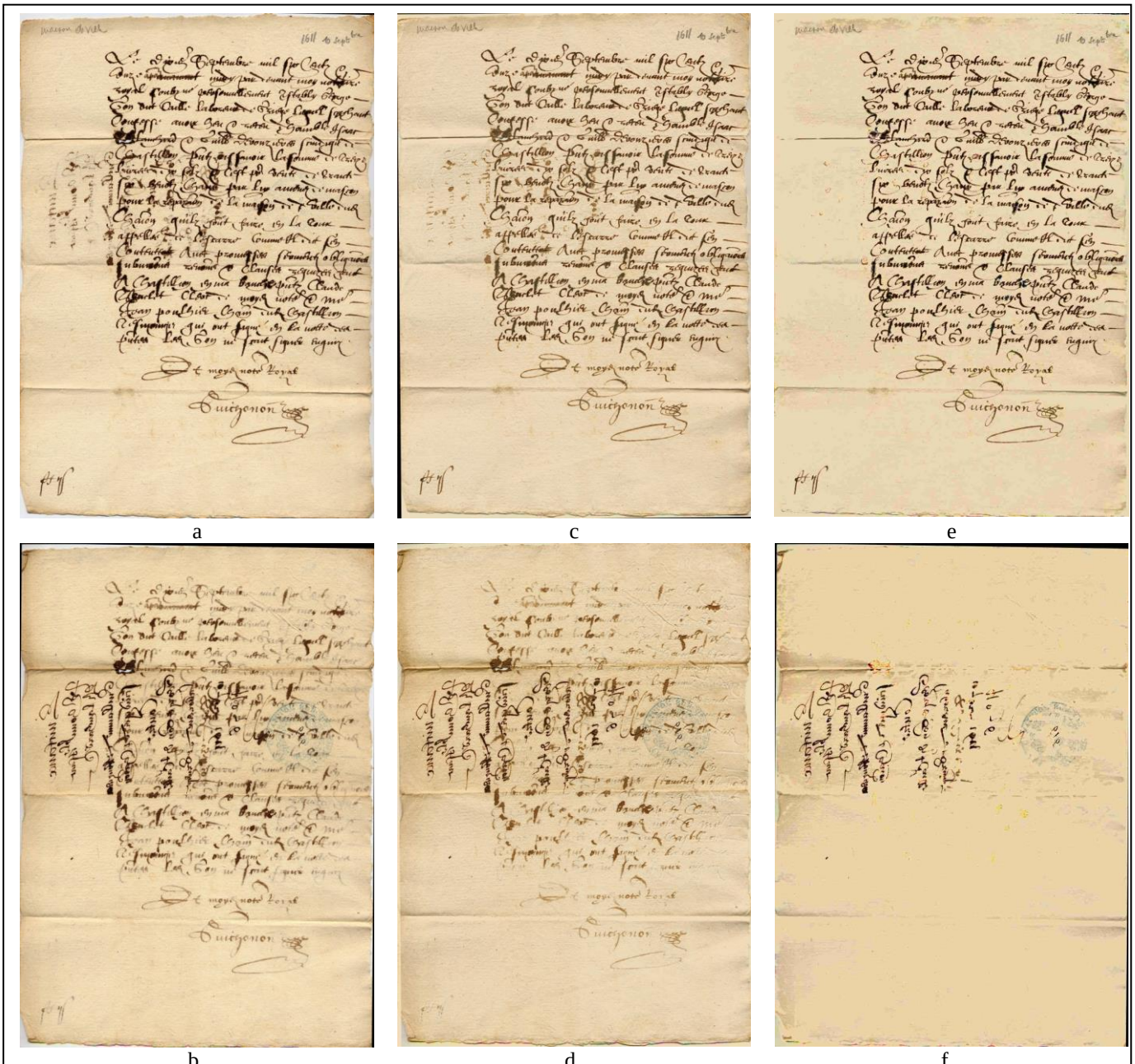


Figure 6 Example of the processing results achieved. (a)-(b) recto/verso of the document after registration of the verso on the recto side. (c)-(d) ink bleed-through removal based on a linear data model. (e)-(f) ink bleed-through removal based on a non-linear data model.

using the MPEG-7 image descriptors, can be searched in less than 2 secs. on a Personal Computer.

IV. METADATA EDITING

The main phases of acquisition, digitization, and processing of digital representations of a cultural heritage object are illustrated in Figure 3. The first phase provides the description of the cultural heritage object, through the creation of the Physical Object performed through the use of the Metadata Editor Tool. The Metadata Editor (MET) is a web-based cataloguing tool which 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 metadata of the Physical Object are then stored into the Itaca Digital Library. An example of the interface of the MET during the creation of a Physical Object is given in Figure 4a.

The digital acquisition of the document and creation of the Digital Representation (Figure 3) is initiated by the MET by accessing the Physical Object description from the archive. The acquisition process is performed through a multispectral camera [3], which produces several digital images in different spectral ranges and automatically generates a Digital Representation containing all acquisition parameters used. The Digital Representation is stored into the Digital Library. Its content can be updated by using the MET, if needed. An example of the interface of the MET during the editing of a Digital Representation is provided in Figure 4b.

The third phase is dedicated to the elaboration of the Digital Representation by using image processing functions developed in the Itaca Project. These functions were integrated in into the GIMP [13] image processing tool. A Digital Elaboration, containing all 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. A complete example of possible elaborations performed on a document in shown in the next section.

V. A COMPLETE EXAMPLE

As an example, let us suppose we acquire a manuscript composed of a single page. The scans of the recto and verso sides of the manuscript are acquired in RGB, and then processed to correct possible geometric distortions introduced during the scanning process [6,14], by registering the verso image on the recto image (or vice-versa). Subsequently, the registered pair is processed in order to remove back-to-front interferences (bleed-through or show-through), by using two different algorithms [5,6].

Figure 5 illustrates the metadata elements related to the creation, acquisition and processing of the document, as they are recorded into the Itaca Digital Library. The phases shown are: the "Manuscript description", with the creation of the Physical Object; the "Document digitalization", with the creation of two Digital Representations, one for the recto and one for the verso of the document; the "recto-verso registration", which first produces a flipped version of the verso side of the document and then registers the recto and the flipped verso of the document; the "Bleed-through ink removal", which results in two Digital Elaborations, one for the recto and one for the verso side, after ink removal. Figure 5 also illustrates the relationships among different metadata entities. Figure 6 shows the images produced during the elaborations. In particular, figures 6a-b show the images of the

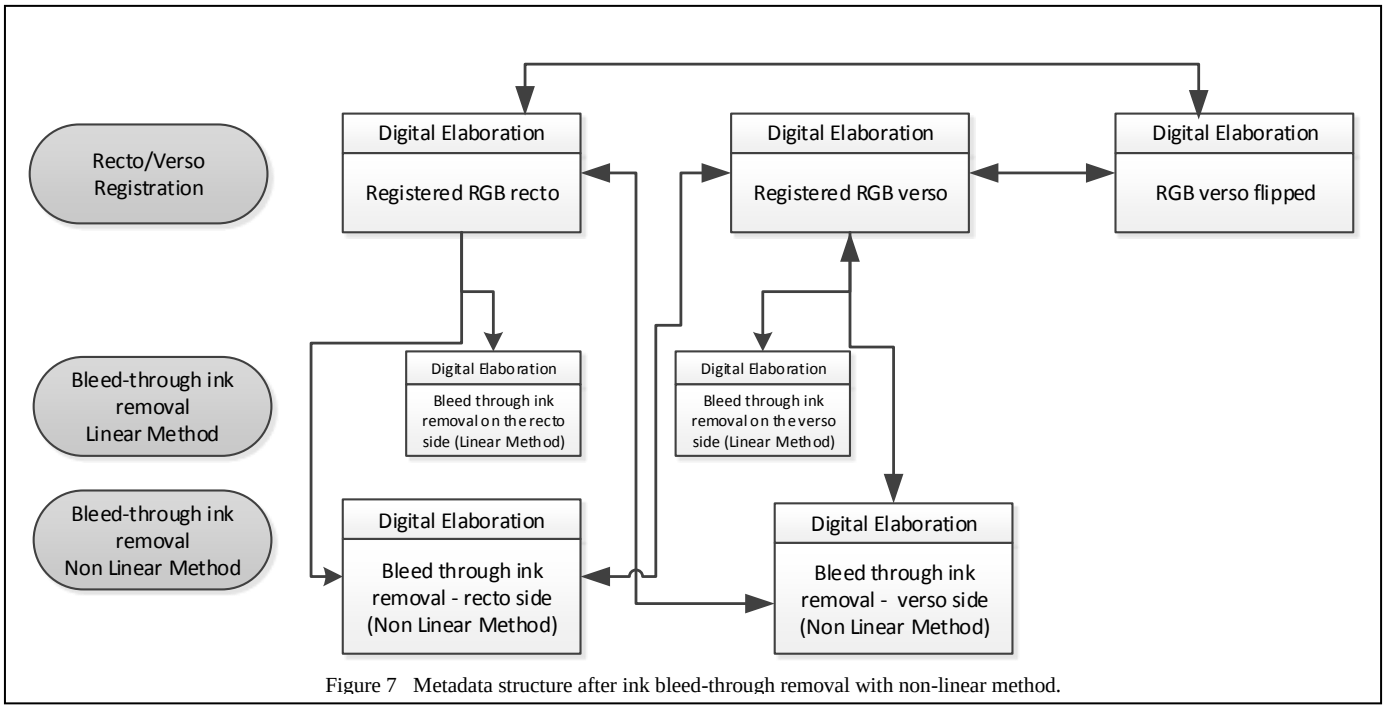


Figure 7 Metadata structure after ink bleed-through removal with non-linear method.

recto and verso of the document after registration, while Figures 6c-d show the document's recto and verso after automatic ink bleed-through removal performed on the registered recto and verso sides, by using the first of the two algorithms. This implements a BSS method [6], based on the decorrelation of the three pairs of RGB channels, assuming a linear overlapping of the two texts.

In this case, the result of the ink bleed-through removal is quite unsatisfactory, as shown in Figures 6c-d. Hence, a more powerful method, based on a non-linear data model [5], is attempted on the previously registered pair. This method has higher computational complexity and requires a minimum user intervention, but the result is of better quality (see Figures 6e-f). Two new Digital Elaborations are created to keep track of the result of this second algorithm as well.

Figure 7 illustrates how the metadata elements are updated to record the new elaboration performed. It can be seen that the Digital Elaborations containing the RGB registered recto and verso are now also linked to the new Digital Elaborations containing the result of the non-linear ink bleed-through removal. The images obtained are shown in Figures 6e-f. It is worth mentioning that the complex metadata structure shown in Figures 5 and 7 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 using both methods adopted for 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 version of the restored document. In this specific case, one might want to isolate and extract the stamp appearing in the verso side. Such a segmentation could be obtained, e.g., by applying BSS algorithms to the image of Figure 6f.

VI. CONCLUSIONS

The paper describes the creation of Digital Library containing digital representations and digital elaboration of cultural heritage objects. A new metadata model, compatible with those currently used when describing cultural heritage assets, is proposed. The model supports the description of cultural heritage objects in all possible representations: from the physical object to the digital representation of the object acquisition. 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 enable 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 [2] and an experimental Digital Library has been created.

ACKNOWLEDGEMENTS

This work has been supported by program POR Calabria FESR 2007-2013 - PIA Regione Calabria Pacchetti Integrati di Agevolazione Industria Artigianato Servizi, project ITACA (Innovative Tools for cultural heritage ArChiving and restorAtion).

REFERENCES

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

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

[3] 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*

[4] 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

[5] 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

[6] 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

[7] 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

[8] Lagoze, C., Van de Sompel, H. (2001) The open archives initiative: building a low-barrier interoperability framework, *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries, ACM, New York, NY*, 54-62

[9] Paepcke, A., Chang, C.K., Winograd, T., García-Molina, H. (1998) Interoperability for digital libraries worldwide, *Communications of the ACM*, 41, 4, 33-42

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

[11] 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

[12] Jagadish, H.V., Mendelzon, A.O., Milo, T. (1995) Similarity-Based Queries, Proceedings of the Fourteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, May 22-25, 1995, San Jose, California, ACM Press, 36-45

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

[14] G. Bianco, G., Bruno, F., Tonazzini, A., Salerno, E., Savino, P., Console, E., Zitova, B., Sroubek, F. (2010) A framework for virtual restoration of ancient documents by combination of multispectral and 3D imaging, *Proc. Eurographics Italian Chapter Conference*, 1–7

[15] A. Tonazzini, I. Gerace, F. Martinelli, "Document image restoration and analysis as separation of mixtures of patterns: from linear to non-linear models", in: Image Restoration: Fundamentals and Advances, Bahadir K. Gunturk and Xin Li Eds., CRC Press / Taylor & Francis Group, Digital Imaging and Computer Vision Series, 2013

[16] A. Tonazzini, P. Savino, E. Salerno, "A non-stationary density model to separate overlapped texts in degraded documents", Signal, Image and Video Processing, Springer, published online 13 December 2014]

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

[18] EDM – The European Data Model. http://pro.europeana.eu/share-your-data/data-guidelines/edm-documentation