ices. The essential internal components of the RI are: (a) the Object Repository (OR), a distributed mass data storage layer, (b) the Metadata Repository (MR), an integrated semantic network layer, (c) the Content Retrieval Indices (CRI) for different search modalities by several content modules, and (d) the Query Manager (QM), which provides a single homogeneous access point to query the three components. The RI offers a central thumbnail database and provides a webservice for handling http thumbnail management requests.

The Object Repository (OR) connects to a (potentially large) number of distributed OR nodes for the physical data storage. The central OR Service provides access security, data integrity and risk-of-loss control, including metadata backup. It consists of a relational database (ORDB) for data file management, the query manager (QM) for mapping the relational database to the RDF format, the DT-Controller module for controlling the data transfer between client computers and OR nodes, and between OR nodes (replica management). The distinctive feature of the OR component is that all data transfers are logged for legal reasons: Not only the acquisition and post-processing of digital 3D assets are expensive, but high-quality 3D models can even be used for creating a high-quality physical replica, eg, through 3D-printing. So it is becoming ever more widely understood that 3D datasets are valuable assets that must be treated carefully, and that their proliferation needs to be faithfully recorded and controlled.

The Metadata Repository (MR) is an RDF triple store that aims at providing a common place to reason on, query, manipulate and export provenance metadata concerning any temporary or permanent object stored in the OR and related metadata about the modelled reality. Metadata are recorded in files in the units of creation, physically backed-up in the OR together with their related content files, while in the MR a semantic network is built with the integrated metadata information. The MR is based on a homogeneous global schema - an extension of the CIDOC CRM (ISO21127) that models provenance metadata (CRMdig). It comprises physical object descriptions, annotations and co-reference information, format and compatibility information of 3D models, historical events, and real world objects. All this information is stored in a coherent semantic network that enables useful and complex inferences to support content management and research, comprising even diverse content indexing and retrieval mechanisms for 3D objects. An integral part of the MR is the Annotation and Co-reference Manager (ACoRM). The Annotation Manager connects links into content segments of any kind and dimension (areas with no modifications on the original object), with semantic information capturing the related scientific discourse. The Co-reference Manager allows for collapsing duplicate URIs in the semantic network without loosing their provenance. This feature acts as a "mending" mechanism of the semantic network and will contribute significantly to the reasoning performance and the future connection into a Linked Open Data (LoD) world.

The responsibility for the RI design and implementation is shared between FORTH-ICS (Greece) and CGV, TU-Graz (Austria), while other partners in the Project deal with the creation of a rich Integrated Visual Browser interface to the RI and an immense spectrum of 3D tools, all integrated via the RI.

Links:
3D-COFORM Project: http://www.3d-coform.eu/
CIDOC CRM: http://www.cidoc-crm.org/
CIDOC CRM v5.0.2 Encoded in RDFS: http://www.cidoc-crm.org/rdfs/cidoc-crm
CRMdig 2.5 Encoded in RDFS: http://www.ics.forth.gr/isl/rdfs/3D-COFORM_CRMdig.rdfs

Please contact:
Katerina Tzompanaki, Martin Doerr, Maria Theodoridou
FORTH-ICS, Greece
E-mail: {katetzob, martin, maria}@ics.forth.gr

Sven Havemann
CGV, TU-Graz, Austria
E-mail: s.havemann@cgv.tugraz.at

Attaching Semantics to Document Images Safeguards our Cultural Heritage

by Elena Console, Anna Tonazzini and Fabio Bruno

Extracting and archiving information from digital images of documents is one of the goals of the project AMMIRA (multispectral acquisition, enhancing, indexing and retrieval of artifacts), led by Tea-Sas, a service firm based in Catanzaro, Italy, with the collaboration of two Italian research teams, the Institute of Information Science and Technologies of CNR in Pisa, and the Department of Mechanical Engineering of the University of Calabria in Cosenza. AMMIRA is supported by European funding, through the Italian regional program for integrated support to enterprises.

Gathering as much information as possible from the documents of our past is essential to preserve our cultural heritage for future generations, especially in an era when the documents are migrating from traditional supports towards the new electronic devices. For many historical documents, this involves capturing their appearance, as well as mitigating distortions and degradations, in order to help human or automatic readers to extract their content. All the information extracted must then be organized to facilitate storage, access and retrieval.

Depending on the type of document, the first step is to choose the most accurate

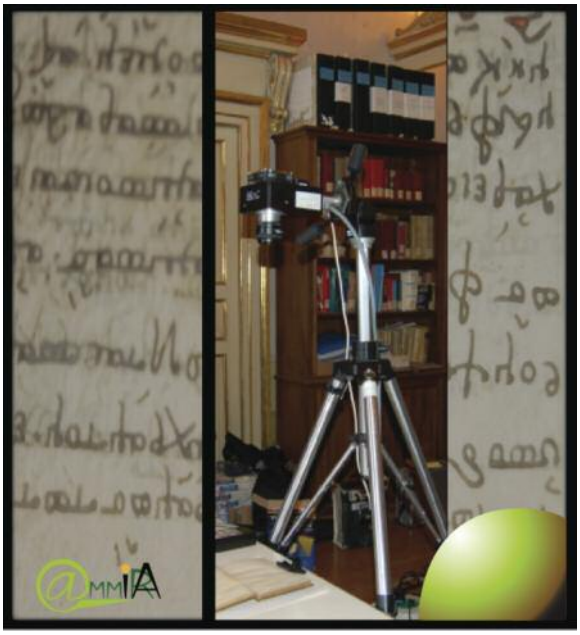

Figure 1: Capturing pages from an ancient book: the AMMIRA DTA Chroma multispectral camera.

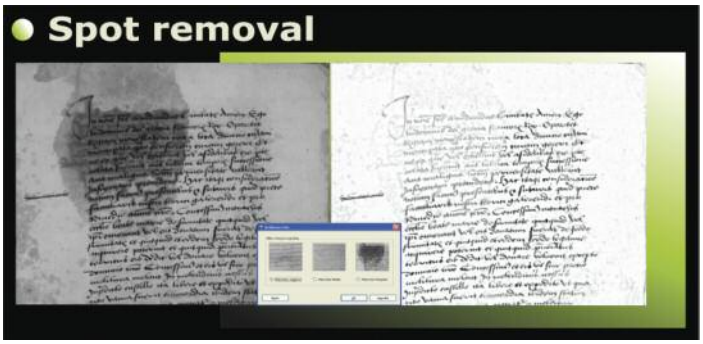

Figure 2: A screenshot from the AMMIRA image manipulation user interface.

and detailed data capture modalities. These modalities include the acquisition geometry and the spatial and spectral ranges and resolutions. The imaging system that we have adopted consists of a high-resolution, multispectral, computer-controlled camera with three visible and two infrared channels, coupled with a visible-ultraviolet lighting system and a structured-light projector for 3D acquisition. Thus, depending on the importance and the state of preservation of a document, we can represent it through panchromatic or colour images in the visible range, infrared reflectance or transmission images, and ultraviolet fluorescence images and, in some cases, a precise 3D shape. This flexibility of representation can help us detect and isolate many kinds of hidden features, and also account for geometrical distortions caused by deformations of the document support. The 3D acquisition augments the level of information that we are able to preserve, because deformations and degradations often describe the history of a manuscript. So, the fruition of the digital replica of a manuscript or an entire book acquires a new dimension as these can be represented as a 3D object in the space.

However, all these features are still raw data, ie a more or less detailed representation of the document appearance with no semantics relating to it. The images must next be freed from interferences and distortions, extracting all the features to which, at some level, semantics can be attached (eg "main text", "footnotes", "images-graphics", etc.). The software system that supports our imaging system is capable of performing many of these tasks. First, we can spatially co-register the available images, after correcting 2D or 3D geometric distortions if necessary. This provides us with a set of data maps with precisely located pixels. Further processing includes virtual restoration, ie removal of distortions and interference (stains, blur, bleed-through), and extraction of features and patterns. The latter task is accomplished through several fully or partially automatic approaches, based mainly on linear data models.

The key to incorporating semantics into the processed data is to adopt an efficient metadata schema that traces all the processing steps applied to any piece of data and all its relationships to other stored material, including all the administrative and descriptive information needed and, eventually, enabling content-based retrieval from a large data repository. The parts classified as text can be translated into machine-readable form through automatic or semi-automatic character recognition systems. The AMMIRA project began in September 2009 and will be completed in September 2011. The partners all have a consolidated experience in this field. The research is expected to continue in the future with the aim of implementing a fully integrated hardware-software system.

Link:
AMMIRA homepage:
http://www.ammira.eu

Please contact:
Elena Console
TEA-SAS, Italy
E-mail : elena@teacz.com

Anna Tonazzini
ISTI-CNR, Italy
E-mail : anna.tonazzini@isti.cnr.it

Fabio Bruno
University of Calabria, Italy
E-mail: f.bruno@unical.it