RESEARCH NETWORKING PROGRAMME
STANDING COMMITTEE FOR THE HUMANITIES

COMPARATIVE ORIENTAL MANUSCRIPT STUDIES

Digital Support for Manuscrip Analysis

Hamburg, 23-24 July 2010

Scientific Report

ANTONELLA BRITA
Università degli Studi di Napoli “L’Orientale”

The purpose of my visit in Hamburg was to attend the international workshop “Digital Support for Manuscript Analysis”, organised in the ambit of the COMSt project, whose aim is to find solutions for a standardization of the studies on Oriental manuscript (Mediterranean and North Africa), also thanks to the contribution of non-Oriental disciplines. The workshop was held at the Asia-Africa Institute (Hamburg University) and it was led by Jost Gippert (Frankfurt), team leader of the Team 3 (Digital Approaches to Manuscript Studies).

The first paper was presented by Manfred Thaller who, within the ambit of the e-Science, has described the Virtual Research Environment and the three basic steps to make a virtual archive (1. acquisition of information; 2. analysis of information; 3. publication). He presented also the project “Monasterium”, a digital archive which collects historical documents (political, economic…) as digital copies. At present it contains 180.000 charters coming from ten different countries. Seventy institutions cooperate to the project. The documents are encoded with XML by the CEI (Charters Encoding Initiative). Thaller has also defined the various phases of the project: 1) digitization; 2. symbol manipulation (tools for basic image improvement; specialized functions for manuscripts image improvement; support for symbol catalogue); 3. transcription (all image improvement capabilities of symbol manipulation layer); 4. editing (EAD/CEI/TEI); 5. research (administration of notes which may become public); 6. publication (PDF export); 7. teach (two years only pilots).

The second paper presented by Andreas Lammer focussed on the project DARE (Digital Averroes Research Environment) whose aim is, in part, bring together the results of the various editorial activities on the commentary of Averroes about *De Anima* (Cicero), preserved (at present) in 143 manuscript copies and make them available in digital form using digital encoding XML/TEI Markup and OWL. The digitalization will take place in two moments: 1) photographic work (grabbing digital images from books) and 2) scanning manuscripts (with high-powered system, as the scanner in aluminium). The segmentation of the text will enable to switch between different textual witnesses and to cross the fragments of text with the occurrence in others manuscripts and incunabula. Lammer showed also the project ARACHNE, an object database created by the German Archaeological Institute and the Archaeological Institute of Cologne University. All the archaeological objects are digitalized and described using TEI. The platform interfaces the digital images and the metadata using the OWL-DL language, which enables to add a semantic value to the documents, to cross the data and to create a complex network of information.

In the third paper Johannes den Hejer presented the project MANUMED, for the safeguard and the valorisation of the Mediterranean written heritage. The sources involved in the project are: 1) original manuscripts, microfilms and photos; 2) inscriptions and documents; 3) old printed books; 4) oral materials (witness of linguistic diversity). The flexibility is the prerequisite of the project. The main guidelines of the project are: 1) copyright; 2) selective digitization; 3) top quality versus accessibility; 4) open access; 5) conservation; 6) digitization of manuscripts (editions, text transmission, codicology…); 7) processing metadata (catalogues).

The fourth paper, presented by Michael Phelps, focussed on the study of the palimpsests (with examples from the Archimedes palimpsest) through the Spectral Imaging which allows to capture image data (non visible to the naked eye) at specific frequencies across the Electromagnetic Spectrum. He argued that the quality of the images captured depends on the conditions of the manuscripts. The use of Electromagnetic Spectrum depends on the way the parchment reflects the light. This method allows to: 1) maximize data collection; 2) create high-quality data archive; 3) strive for increased efficiency and cost-effectiveness; 4) realize high standards for materials. The stages of the work are the Spectral Image Capture and the Spectral Image Processing. In the Spectral Reflectance Imaging the light reflected by the surface of the parchment (through a lens fitted with a filter - blue, red and green) is captured by a sensor. Moreover, the Spectral Fluorescence Imaging has made possible to visualize structures deep to the skin.

The fifth paper also, presented by Jost Gippert, was about the study of the palimpsests. He presented some examples of employment of the MuSIS technology (Spectral Imaging Solutions) during recent researches (2003-2008). He showed that the application of multispectral imaging must concentrate upon two aims: 1) increasing the contrast between the (erased) lower script and the background and 2) exploiting the difference of several images showing the same object to reduce the preponderance of the upper script. He emphasised the extended software functions of this system and its application in many fields. Furthermore, he underlined the importance of light in the sources involved (depending on the material). In particular, for parchment, the use of ultraviolet and, for paper, the use of infrared.

In the sixth paper, Steven Delamarter, after a short introduction concerning the difficulties which can be found in a work involving dealers and owners of manuscripts both in North-America and Ethiopia, presented his project (EMIP) whose aim is to create digital proxies for the manuscripts. He photographed at present ca. 1500 manuscripts.

In the seventh paper, Ira Rabin, emphasised the necessity to create a writing material database. She explored, according to different areas and periods, the chemical composition of the inks, the processing of the parchment, the techniques of dating materials and the state of preservation (Radiocarbon, Shrinkage temperature method - based on the hydrothermal denaturation of the collagen protein molecules -, Chemometrics - chemical discipline that uses mathematical and statistical analysis). Moreover, she showed the techniques of analysis destructive for materials (shrinkage temperature, scanning electron microscopy…) and the techniques non-destructive (µ-XRF – x ray fluorescence –, Raman spectrometry…).

In the eighth paper Steven Delamarter presented some results of the analysis carried out on the Ethiopic manuscript Walters Codex. In general, about the composition of the inks and of the binding, he showed that: 1) black ink contains traces of gallic acid, Arabic gum and iron sulphate; 2) red ink contains traces of cinnabar and Arabic gum; 3) blue ink contains traces of barium and aluminium. FT-IR exam confirmed the presence of synthetic binders in paintings.

Straight afterwards a forum on tools and techniques took place. Among other things, a list of information for the creation of digital projects has been elaborated: 1) project name; 2) web address / contents; 3) short description; 4) images online (yes or not); 5) images type (complete manuscripts / samples); 6) images quality (size / format); 7) viewers; 8) metadata catalogues available (yes or not); 9) encoding; 10) languages used (multiple); 11) keyword; 12) text available; 13) transcription available / edition available; 14) searchable (search type: google…).

In the ninth paper Bernd Neumann has explored the development of innovative image processing methods. The contents of his speech can be summarized as follows: 1) image restoration (with digital systems) and segmentation (a. segmentation based on pixel through segmentation based on subpixel accuracy, which is much more precise and can be applied in particular to images with low resolution; b. segmentation of the degraded text using shape priors); 2) writer verification with CEDAR-FOX (Buffalo University) though manual processing which allows to remove non text and major noise. Reprocessing as automatic character recognition and manual correction. Comparison of corresponding based on GSC features - Gradient-based; Structure-based (corners, lines…); Concavity based (holes, strokes…). Similar approach extended to comparison of words. The handwriting recognition technology is adapted to requirements in forensics and is not immediately applicable to other writing systems (incomplete prototypical systems for palaeographical applications); 3) the content-based image retrieval determines the occurrences found in manuscripts in a large database. Object recognition using SIFT (Scale Invariant Feature Transform) features.

In the tenth paper, Torsten Schassan proposed an overview on the trends in codicological and palaeographical analysis and description. The topics broached by Schassan are the following: 1) codicology methods whose task is to integrate different types of resources and different types of information: a. portals - local, regional, national (ex. Gallica, Manuscripta Medievalia…), European (Europeana, Manuscriptorium, CERL…); b. data structures - TEI (which develops from textual view to the consideration of the object or to a full representation of the real world objects, including descriptions, facsimile and text), EAD, OPAC's…; c. research environments: representation of semantic structures and relationship (RDF, topic maps) and structures for specialized information; 2) materials: digital manuscripts; single/representative pages of manuscripts; single pages digitized upon request; 3) methods for digitization: a. multispectral digitization; b.

thermographic imaging (esp. for watermarks); c. segmentation; d. OCR; e. neuronal networks; f. linking text and image; 4) palaeography tasks: a. algorithm which drives interpretation of manuscripts (count lines, measure lines, distinguish words, interpret distribution of the text); b. models for letter (software BIT Alpha). The last paper was presented by Matthew Driscoll and Eric Haswell It was about the computing systems linking text and images and, in particular, about TEI. TEI provides mechanism which enable one to link sections within and between texts or between text elements and elements in the TEI header or TEI point to external resources, through an interface. The <facsimile> element contains one or more <surface> elements. The <surface> element defines in terms of a rectangular space any written surface (this may be a piece of manuscript, papyrus…). The <tone> element defines a rectangular area within a <surface>. Coordinates may be provided to locate the surface. Markup is not an end goal (University of Victoria by Martin Holmes). Limitations: 1) windows only; 2) web view is not dynamical generated from XML; 3) one image is equivalent to one XML file. Other projects not yet fully realized: 1) TEI P5 XML (open source); 2) TILE (Text-Image Linking Environment) is under development; 3) IMT (an excellent starting point which can be used for demarcating areas and establishing pixel-based boundary values for projects requiring non complex encoding).