

Grid Archiving for Ancient Texts

S. Scifo, Consorzio COMETA - Catania, Italy (salvatore.scifo@ct.infn.it)
 G.Foti, Dipartimento di Fisica, Università di Catania, Italy (gaetano.foti@ct.infn.it)
 F. Portuese, IR&T engineering s.r.l. - Catania, Italy (f.portuese@irt-engineering.it)
 S.Parisi, IR&T engineering s.r.l. - Catania, Italy (s.parisi@irt-engineering.it)

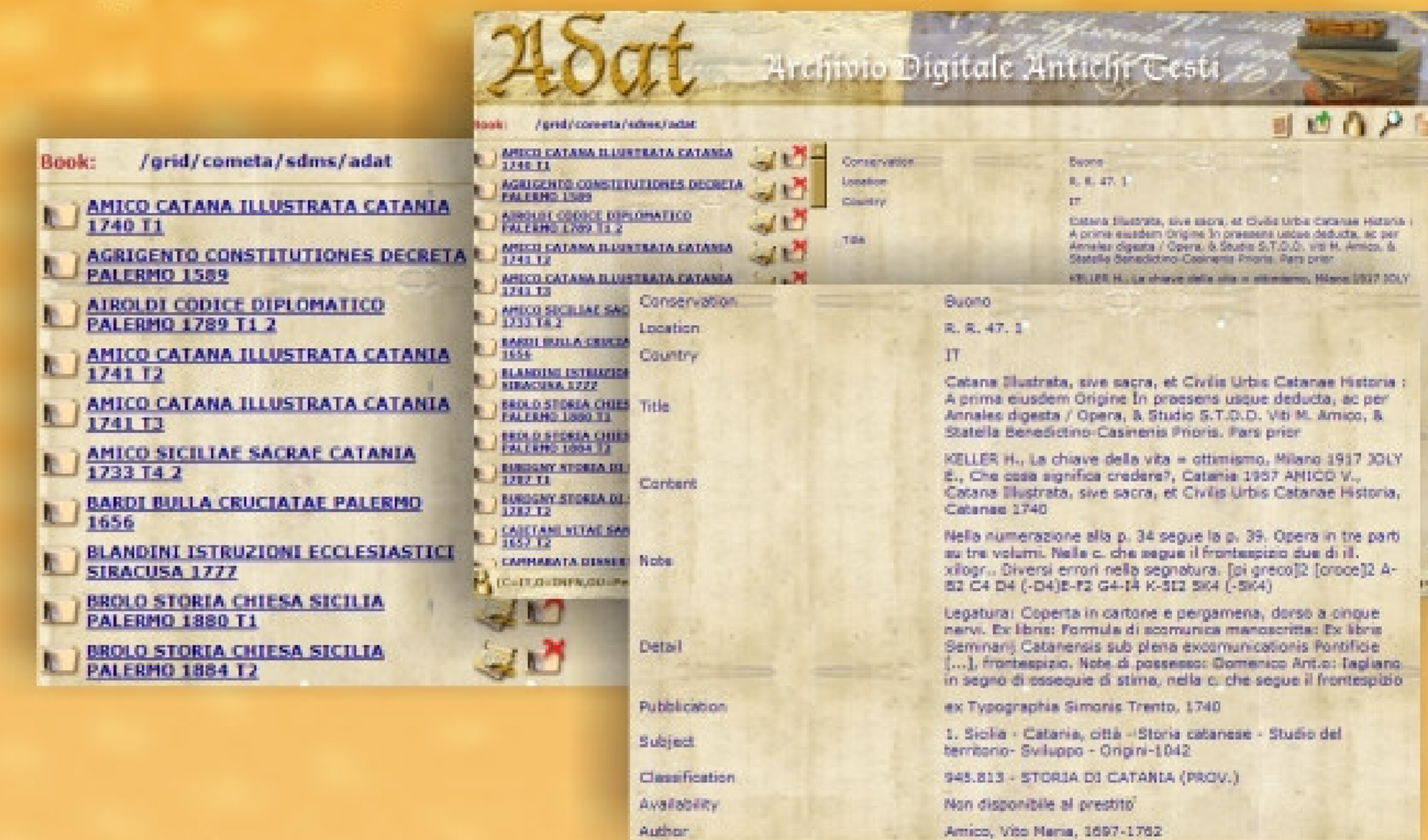
The preservation of cultural heritages is becoming more and more important nowadays. Unfortunately, several patrimonies has been destroyed due to atmospherical agents, natural disaster and/or human faults. Digitalization is a convenient way to preserve these heritages from every kind of alteration including that one derived from physical restoration. The fast growing of digital data for ancient texts, stresses the storage system in order to handling such large amount of information. The Grid seems the optimum solution for several reasons: Storage scalability, Security, High Availability, Accessibility. This work regards the implementation of a full qualified process based on the digitalization of several terabytes of ancient texts from the Biblioteca Agatina of Catania (Italy); their cataloguing, their storing and their fruition through an innovative Grid Web interface.

Grid Digital Archive

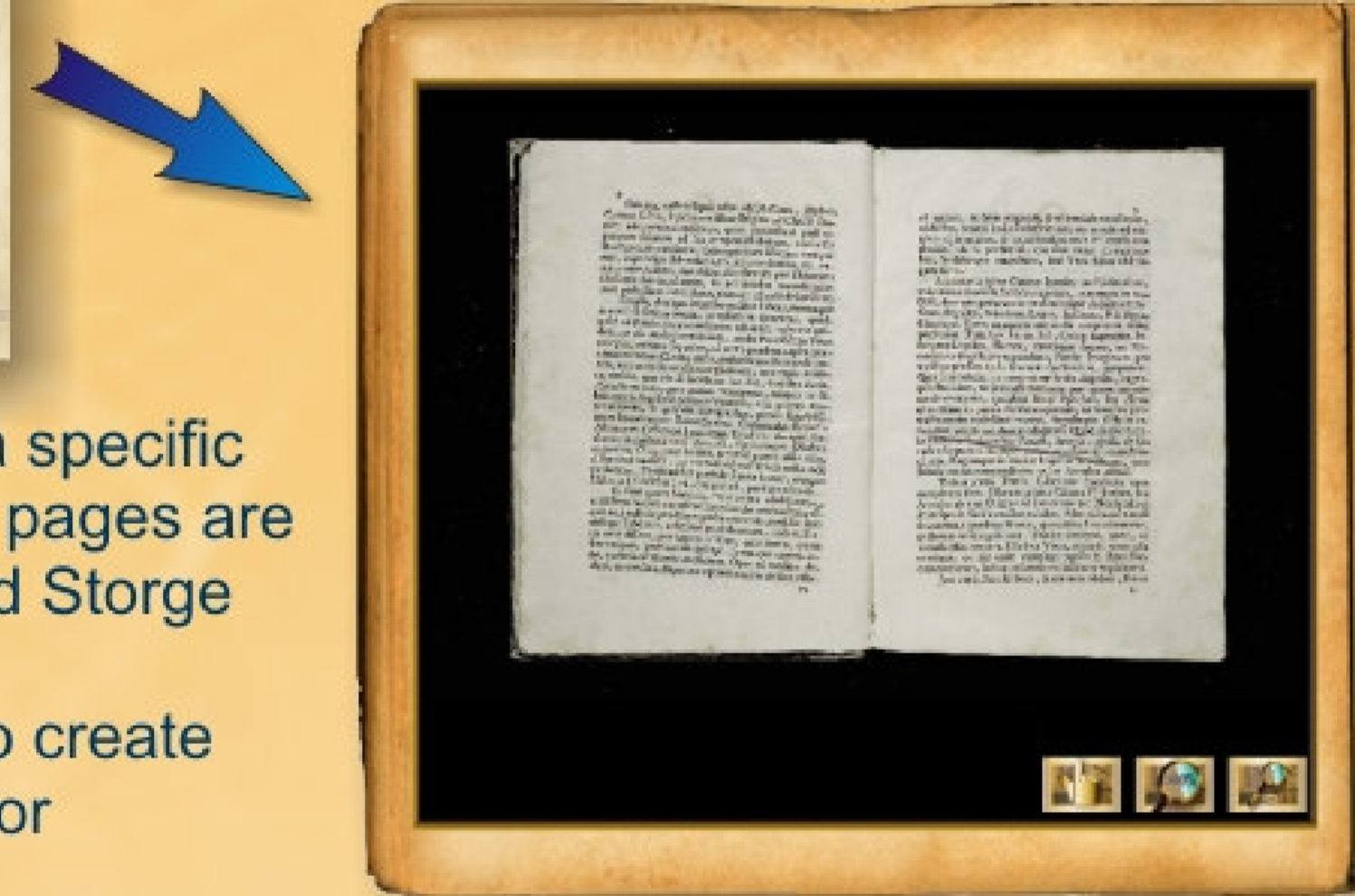
- Handling and managing large amount of data (Tera/Peta Bytes)
- Geographical distributed storage
- Net access (web oriented) for several functionalities (administrative,operational, consultative)
- Centralized access control mechanism based on Virtual Organization roles that users belong to
- Providing indexing and cataloguing services
- Delegating management aspects for both net infrastructure and storage system (maintenance and security) to the Grid Site Management

ADAT Web Interface

A web oriented digital archive as been developed using a specific Grid API called GSAF (Grid Storage Access Framework).



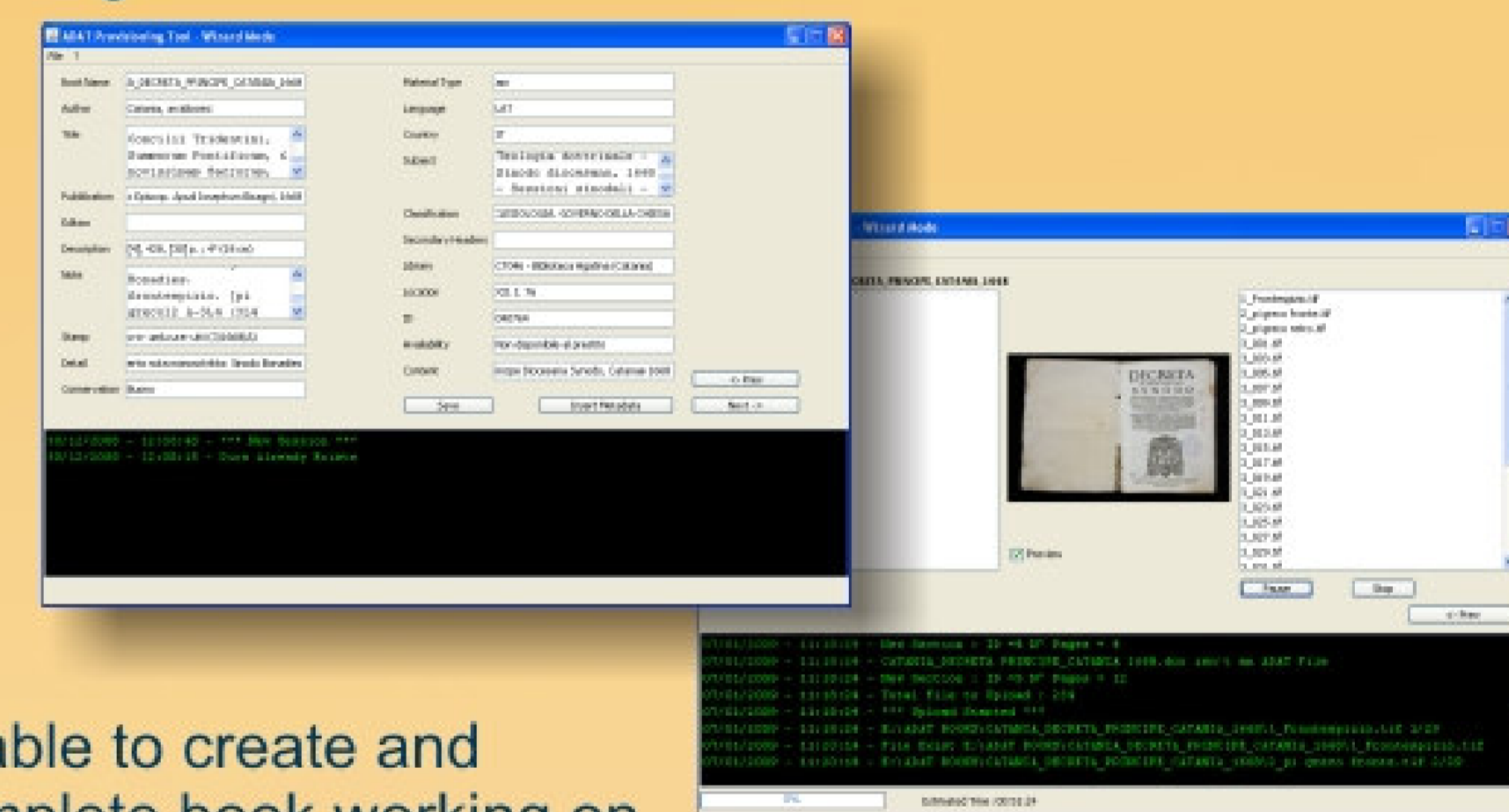
ADAT Web Interface



- Books are classified using a specific bibliographic card and their pages are collected on the remote Grid Storage
- For each page is possible to create several replicas for backup or furthermore analysis
- Each digital page (master or replica) can be displayed using a specific web

Authoring Tool

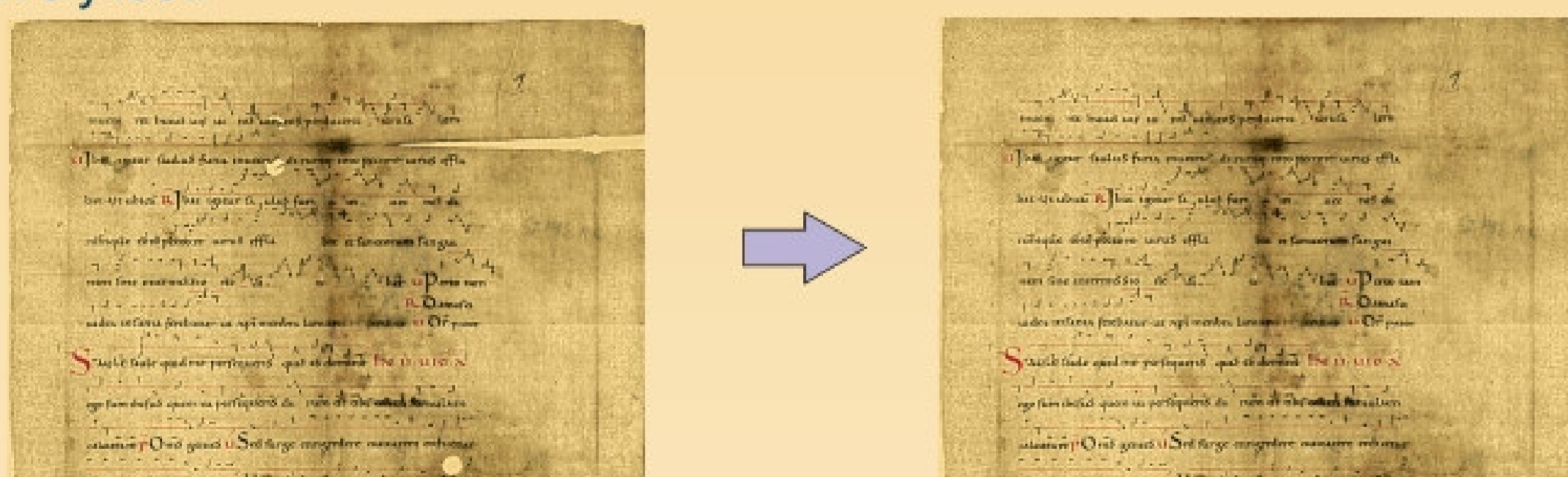
A dedicated software system as been developed in order to speed up the full process to automatically transfer every pages and metadata from the camera to the Grid storage.



The user is able to create and upload a complete book working on its own workstation.

Digital Restoration: mechanical alteration

Torn can dramatically compromise the document because it can interest an huge portion of the paper. In latter case the information couldn't be reconstructed and is definitively lost.

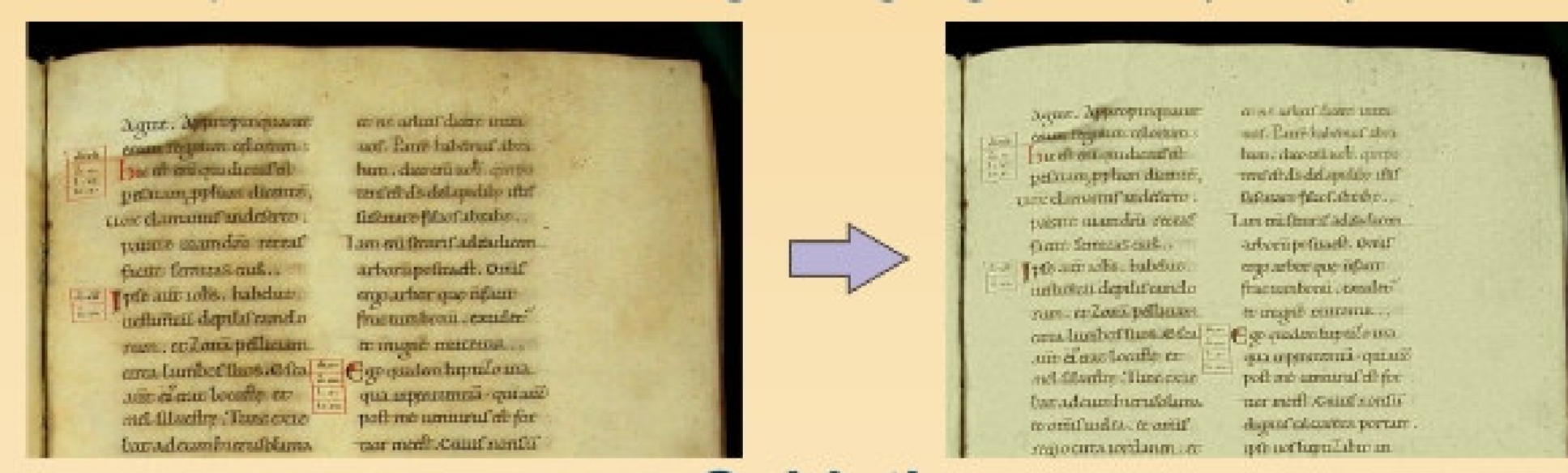


Torn papers

The detection of this defect could be done by using an uniform predetermined background color during paper digitalization. The reconstruction of the missing parts could be done through inpainting algorithms.

Digital Restoration: chemical alteration

The oxidation of the paper is a natural process that affect every old paper. This effect is characterized from a yellowing of the paper. The removal of this defect, also make the background more homogeneous improving the readability of the document for optical character recognizing algorithms (OCR).

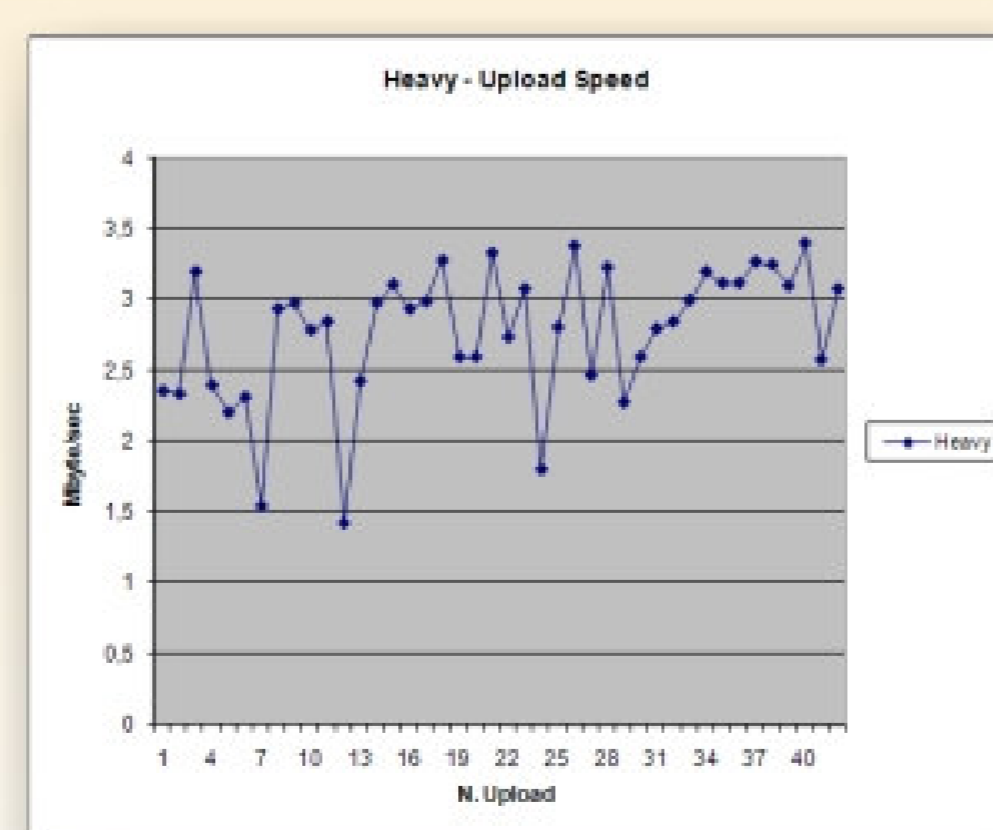
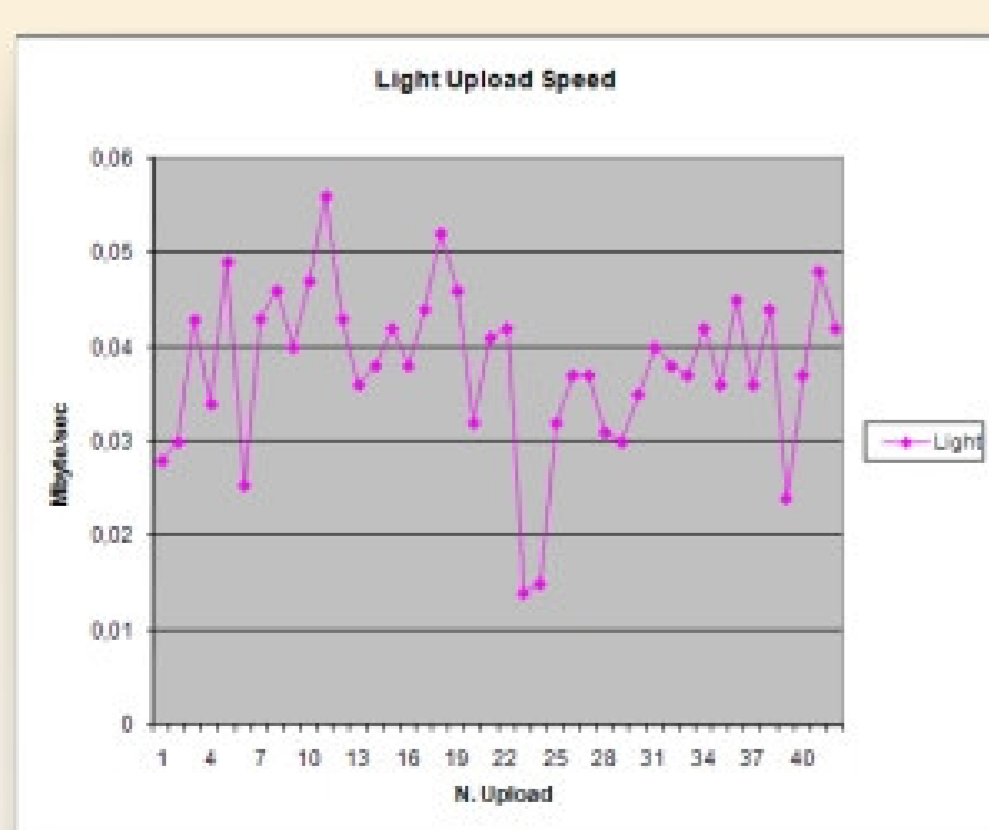


Oxidation

The algorithm try to separate the text from background partitioning the luminance histogram. The restoration occurs by adjusting the color of pixels that belong to background with their median.

Speed Measurements

Our experiences show a fast input of data for TIFF imaging up to 50 Megabytes for each sheet without any practical limitations. The measured data rate transfer on COMETA grid depend on the file size: for upload process of 300KB JPEG images we found a rate **0,04MB/sec**, and for the 50MB TIFF images **2,7MB/sec**



Conclusion

Future Objectives

- Development of high added value applications oriented to e-commerce, e-learning and last but not least **multimedia cultural tourism**
- Digital resources accessible towards the net and addressed to both **specialized user** and cultured and curious user
- **User/archive interface** that emphasizes not only the cultural heritage knowledge but also, encourages research and experimentation promoting creative