ITSERR

ITSERR Scientific Kick-off

WP5

Table of Content

- Subject of the WP
- Technologies involved
- Research disciplines/skills needed (ie from PhDs, RTDs, specialists, etc.)
- Outcomes
- Impact to the Research Community
- Team members involved

ITSERR

Subject of the WP

- The management and cataloguing of documentary heritages has been the subject of study in Europe since 1700
- New demand for systems and procedures for managing and sharing cultural heritages also in supranational and multiliterate contexts
- DigitalMaktaba (in Arabic "maktaba", "library": the "place where books are found" and "where you write") is a project that aims to build a digital library to preserve cultural heritage.
- Goal: provide a system for creating, managing, and cataloguing historical heritage in non-Latin alphabets
- Project test-case: large collection from the "Giorgio La Pira" library in Palermo

Technologies involved

- **Knowledge Extraction & Supervised Cataloguing techniques**
  - Optical Character Recognition (OCR) (libraries, online services, …)
  - Linguistic / knowledge resources
  - Machine learning libraries
- **Data management**
  - relational databases, noSQL databases, graph databases
- **Web application cataloguing software**
  - Python
  - Web application framework
  - …

Research disciplines/skills needed

- Big data management and analytics
- Natural language processing and knowledge extraction
- Big data integration
- Machine learning with human in the loop
- Digital Humanities
- Linguistic and librarianship skills
- Religious studies
- Libraries and archives

Outcomes (1/4)

T1 - Analysis of the operating scenario and materials *[M 1-6]*

- **T1.1** ***[M 1-6]*** Preliminary activity of recognition of the operational scenario, focusing on the problems from both the IT and the historical-critical, linguistic perspective
- **T1.2** ***[M 1-6]*** Preliminary recognition of:
  - OCR tools and technologies for the languages considered by the project linguistic tools (multi-lingual resources such as dictionaries, thesauri, tools for processing text) available in the languages considered by the project:
    - text mining techniques
    - long term preservation of digital documents techniques
    - big data management tools/techniques
    - (interpretable) machine learning techniques

Outcomes (2/4)

T2 - Development of algorithms for automatic text recognition, metadata and knowledge extraction *[M 4-14]*

- **T2.1** ***[M 4-9]*** Definition of text acquisition/OCR techniques for assisting/automating text extraction
  - exploiting object fusion techniques based on fuzzy matching for aligning/merging the outputs of different OCR tools
- **T2.2** ***[M 6-12]*** Definition of techniques for extracting syntactic metadata
- **T2.3** ***[M 6-13]*** Definition of knowledge extraction techniques for linguistic / semantic metadata:
  - exploiting multilingual resources
  - exploiting techniques for title and author automatic recognition
- **T2.4** ***[M 13-14]*** Validation of the developed recognition and extraction techniques, both on samples of materials found in WP1 and on larger corpora, available in the literature and provided by partner institutions.

Outcomes (3/4)

T3 - Data Management, Interactive Search and Supervised Cataloguing *[M 14-24]*

- T3.1 ***[M 14-18]*** design of a database for storing the extracted catalogue data and metadata, including data management techniques for interfacing / interchange with catalogue data from other libraries and exploiting:
  - Long term preservation practices
  - Big data management / distributed
- T3.2 ***[M16-20]*** definition of advanced search techniques (including approximate and full-text search) for searching archive data
- T3.3 ***[M16-22]*** design of a web user interface for cataloguing new documents and searching the archive
- T3.4 ***[M16-24]*** definition of intelligent assistance techniques based on similarity search and supervised (incremental and interpretable) ML algorithms:
  - to assist data entering also based on suggestions from user feedback and previously entered data
  - to automate publication type recognition
  - to ensure that the tool can "learn" and become more and more automated and effective with use
- T3.5 ***[M 25-26]*** Validation of algorithms on samples of materials found in WP1 and on corpora existing in the literature.

Outcomes (4/4)

T4 – Integration of the proposed solutions [M 27-34]

- **T4.1** ***[M 27-30]*** Integration of the solutions developed in WP2 and WP3 in the system prototype
- **T4.2** ***[M 32-33]*** Validation in terms of accuracy and completeness of the information extracted.

Impacts for the Research Community

Benefits are expected on several fronts of innovativeness, from a broader standpoint:

- **Advancement of studies** on cataloguing in multi-literate environments (without leaning exclusively on confusing transliteration systems)
- **Exchange of IT,** humanist and library personnel, enhancement of professional skills, training activities extended to realities with similar needs
- **Strengthening of library services** thanks to shared international standards, increasing library heritage, databases integration, maximum access to the heritage, possibility of using the language of the document without mediation of other languages
- **Shared knowledge tools** between different cultural, linguistic and religious realities. The multidisciplinary and multicultural nature of the project and a new typology of services can make effective contributions by affecting, in a broad perspective, the dynamics of interaction

Impacts for the Research Community

From a more technical point of view many foreseen advantages are auspicable in the scientific domain:

- *Overcoming the limitations of current text extraction tools*
- *Faster cataloguing pipeline*
- *Greater consistency and less errors*
- *Consistently better system output through time*
- *Intelligent features for further user assistance*
- *Exploitation of available libraries catalogues*
- *Flexibility of data output/exchange*
- *Efficiency and Explainability*

Team Members involved

- Professor **Riccardo Martoglia**
- Professor **Sonia Bergamaschi**
- Professor **Federico Ruozzi**

- PhD student **Riccardo Amerigo Vigliermo**
- PhD student **Matteo Vanzini (CDS)**
- PhD student **Luca Sala (ICT)**

- 3 more PhD students (one on 1st January 2023, the other two on November 2023)

ITSERR

Thanks for your kind attention