Vol. 5 - N° 1 - June 2021

iJIST

International Journal of Information Science and Technology

Special Issue on Digital Preservation of Written Heritage and Text Processing Technology

GUEST EDITORS
Dr. Ouafae Nahli
Dr. Anna Tonazzini
Dr. Stefano Legnaioli

PAPERS

ISSN 2550-5114

PREFACE

The notion of "written heritage" concerns the immense field regarding written culture. It is difficult to attempt a definition of written heritage. In very general terms, it can be said that it includes any surface containing something written. Different writing supports can be involved, such as papyri, parchments, manuscripts, books, but also contemporary type of medium, such as posters, newspapers, magazines, etc. The written cultural heritage such as general written records are means of communication through space and time of costumes, culture, and beliefs. They constitute the archive of the evolution of mentality, knowledge, sciences, and arts and are considered guardians of the language [1].

Any country has the right and duty of the long-term preservation, the efficient dissemination, and the effective promotion of written heritage. As a matter of fact, the same time it is important not only the preservation of written objects as it was traditionally conceived, but also considers the possibility to access them to researchers and citizens.

Since ancient times, this has been achieved through the creation of new transcripts. However, the conversion of the ancient written heritage poses many problems and challenges. The preservation effort is a crucial requirement since the antiquity to the extent that it has been conducted by using any means, even if making copies of the primary source or transferring the initial content from one medium to other results in a loss of original quality. The desire to keep documents in their original form has become pressing only in the modern time. Digital technologies seem to have opened new perspectives allowing digital preservation and processing of written heritage. They offer the possibility to describe, store, and access - in digital format- hundreds of thousands of manuscripts that survived through time, increasingly considering their preservation, investigation, and dissemination.

Once restored, the original text remains fragile and must be protected and preserved. However, the cultural, scientific, and historical content, i.e., the immaterial heritage, is an integral part of the public heritage and, accordingly, it should be always available to scholars and general audience. Thanks to digital technologies, the data dissemination and communication are experiencing a real large-scale revolution, allowing for the consultation of written heritage to become common.

The article "Challenges in the digital preservation of historical laminated manuscripts" describes the workflow that underlies the steps to be performed to achieve this end. The authors discuss the possible architecture and feasibility of a complete system for the digital safeguarding of historical manuscripts. In their own opinion, an effective computational system should follow this workflow:

- digitization of the manuscript;
- digital image processing to have a clearer version, facilitating reading;
- digital transcription of the text;
- critical edition of the corrected version with comments that explain the text;
- linguistic analyses of the text;

– for a global diffusion, phases of translations and ontological studies are necessary.

The authors choose as their example the digital safeguarding of historical manuscripts degraded by the process of lamination. Since the lamination process is an irreversible operation, the concrete risk is a complete loss of text readability. Consequently, urgency imposes a digital preservation of such manuscripts, together with a philological study and a textual analysis of the content. As a case study, the authors make reference to a Moroccan manuscript of major significance, the longest poem of medieval Islamic medical literature, the ``Poem in Rajaz on medicine". Specialized digitization techniques and subsequent digital elaborations improve the readability of the text and support the process of scholarly transcription, text encoding, and linguistic analysis, performed through specialized semi-automatic computational tools.

Multiple manuscripts of the same text, made by different copyists, contain variants due to the way the text was copied, interpreted, and transmitted by the copyists themselves. The challenge for a philologist is to get as close as possible to the original text. One of the most delicate phases of philological activity consists in the collection and examination of documents - called witnesses - attesting a literary or historical work. During this phase, philologists compare the text of many witnesses, such as manuscripts and printed editions, and record the significant differences (variant readings) in a special area, called the critical apparatus. Manually encoding variant readings is a difficult and time-consuming task.

In the article “Document analysis and textual philology: a formal perspective”, the authors illustrate a mathematical and formal model that describes the dynamics of text transmission. This implies the formalization of phenomena that are common to the various textual traditions (e. g. unconscious or mechanical corruption) as well as of phenomena which are different in each domain but are the result of general scribal habits and approaches (conscious innovations such as exegetical reworkings, interpolations due to moral or theological concerns, etc.). Such a process of formalization must identify the entities, properties, and relations characterizing the textual-philological domain in a strict sense, but it must also take into account the relationship existing between textual philology and other forms of criticism, such as literary criticism, source criticism, and redaction criticism.

From general theories to specific case studies, Luigi Bambaci illustrates an empirical approach in the article “Critical apparatus as Domain-Specific Languages: a rule-based parser for encoding an eighteenth-century collation of Hebrew manuscripts”. He shows how it is possible to exploit the structured language of critical apparatus as a means of automating processing and encoding. He discusses the advantages deriving from the adoption of a parsing system over a manual encoding, advantages which range from speed in data acquisition to the possibility of automatically detecting misprints or inconsistencies in the printed source, correcting errors originated after OCR processing, and better controlling the generation of semantic errors while encoding into XML file format. As a relevant case study, he chooses the digitization of a collation of Hebrew manuscripts and printed editions realized by the English scholar Benjamin Kennicott in the second half of the XVIII century.

The digital age has allowed multicultural and multilingual globalization. For an international dissemination of written heritage, the translation of textual content is necessary. This new and vast sharing raises linguistic, cultural, and philosophical reflections. Translation has always played an important role throughout history in conveying thoughts and knowledge from one nation to another. However, the act of translating is not simply to change a message from the source to the target language. Language and culture are closely linked and sometimes it is difficult to translate some cultural aspects related to the beliefs, customs, history, literature, and shared knowledge of the linguistic community, in which the original text was created. Thus the notion of untranslatability emerges. It can be said that a text is untranslatable when no equivalent text or statement can be found in another language [2] [3]. In addition, the technologies developed provide tools and resources that enable automatic translation. The latest study "Machine Assisted Human Translation: The Dichotomy of Translatability and Untranslatability of Terminology" aims to highlight the advantages, limitations, and failure of human-assisted translation in providing equivalences of terminology from different domains.

In summary, the contributions to this special issue discuss the use of technologies in the context of the specific task of safeguarding and disseminating written heritage. All the articles point out the enormous potential of these technologies in assisting the completion of repetitive tasks or in providing solutions to problems hitherto insurmountable. The increasing availability of developed technologies will make human performance significantly smoother, more reliable, and more consistent.

References

[1] P.M. Debiasi, "Le patrimoine écrit". Encyclopaedia Universalis, 2000. Encyclopædia Universalis [online], accessed September 8, 2021. URL : https://www.universalis.fr/encyclopedie/manuscrits-le-patrimoine-ecrit/

[2] M. Banafsheh and A. Keshavarzi, "Cultural translatability and untranslatability: a case study of translation of "rostam and sohrab". Journal of Global Research in Education and Social Science, 2016, 6(3): 138-147.

[3] J. Cui, "Untranslatability and the method of compensation". Theory and Practice in Language Studies, 2012, Vol. 2, No. 4, pp. 826-830.

GUEST EDITORS BIOGRAPHIES

Ouafae Nahli - Institute for Computational Linguistic, Italian National Research Council, Pisa, Italy - is full-time Researcher at the CNR Institute for Computational Linguistics "Antonio Zampolli" in Pisa. Her research interests and contributions cover a number of aspects of both Classical and Standard Modern Arabic, including morpho-syntactic and lexico-semantic issues in Arabic Natural Language Processing, computational analysis of literary texts, and lexico-ontological modelling.

Anna Tonazzini - Institute of Information Science and Technologies, Italian National Research Council, Pisa, Italy - is a senior researcher at the Signals and Images Lab of the Institute of Information Science and Technologies, CNR, in Pisa. She has coordinated national and international projects on neural networks and learning, computational biology, and document image processing, and has been supervisor of various postdoctoral fellowships. Her current research interests concern image analysis for cultural heritage, and computational methods for structural genomics.

Stefano Legnaioli - The Institute for the Chemistry of OrganoMetallic Compound,Italian National Research Council, Pisa, Italy - is researcher at the Italian National Council for Research - Institute for the Chemistry of OrganoMetallic Compounds since 2008, working in several research activities of the Applied Laser Spectroscopy group in Pisa, particularly in the field of spectroscopic techniques applied to Cultural Heritage, Environmental, Biomedicine and material characterization such as monitoring of industrial processes. He is co-author of more than 150 peer reviewed papers (h-index: 36 source Scopus). Based on the classification of the European Research Council, its research activity can be classified within the following sectors: PE2, PE4, SH5_1, SH6_1