

Pinakes Text. A tool to compare, interoperate, distribute and navigate among digital texts

Andrea Bozzi – ILC CNR (Pisa)

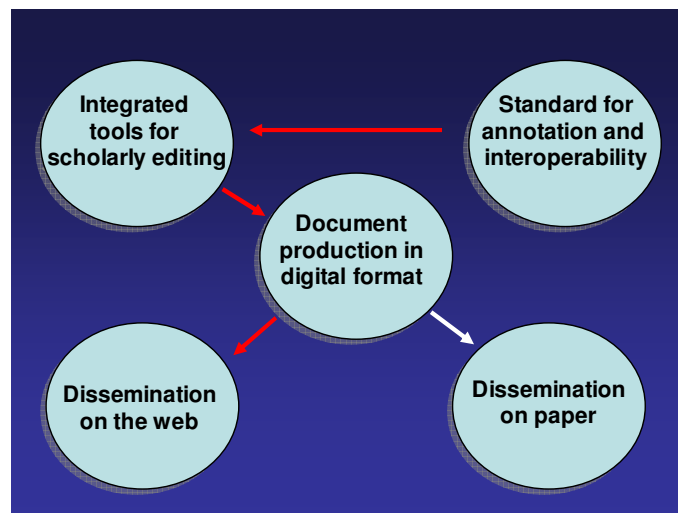
andrea.bozzi@ilc.cnr.it

Abstract

Over the past years text processing systems have become part of the daily language of many scholars working in the field of the Humanities, despite some objections raised against this type of technology which still seems to be distant in terms of simplicity of usage, appropriateness, and flexibility.

Usage requires particular attention as concerns the interface between the information system and the user, while appropriateness and flexibility have not been sufficiently taken into consideration even due to the fact that they almost seem to be in contradiction. Therefore, it is not easy to plan and implement a text processing system which is suitable for specific types of research and at the same time flexible to operate in various sectors of study.

The project Pinakes Text (PKT) that I am presenting here, is aimed at achieving this ambitious target through an architecture based on interconnected modules. In other words, the system works with a nucleus of components for the treatment of both text files and digital image files, which form the core of the system. According to the specific needs, from time to time a number of programs are added both for the management of images (enhancement, segmentation, pattern recognition, etc.) and of text (natural language processing, information extraction, data mining, electronic editing, ecc.). The simplified scheme which lies behind PKT could be represented as follows:



1. The first element is represented by the respect of internationally shared standards, so that the information managed by PKT is interoperable with other data produced in the humanistic field. The standards are also followed when not only primary data (texts, images, etc.), but also secondary information, such as annotations, variants, comments and/or information produced by computational systems (e.g. morphological, syntactic,

semantic analyses) are introduced. The software development tools are totally open source in order to avoid any fees for end-user licences.

2. The information system is entirely web-based and the tools for the production or search for information are oriented towards the sector of critical and textual scientific editing. At present, the target of PKT is represented by specialist users. However, the structure of the system also envisions a number of operations, in particular those connected to the phases of search and query, which can be further developed so as to meet the needs of a non-specialist-user.
3. PKT allows to produce on a web server data that have been labelled and annotated in collaborative form, as long as all the members of the same community (e.g. mediaeval philologists, Greek papyrologists, Egyptologists, Latin epigraphists, historians and science philosophers, etc.) agree with the same standards, as evidenced in point 1.
4. Some experiments are in course to check whether PKT meets the needs requested by a community of scholars working on documents of Egyptian archaeology. The documents and annotations are produced in digital format and are classified according to a domain ontology agreed upon by the same members of the community. This semantic-conceptual structure can be replicated not only to classify the documents, but also part of their content. In this way, it is possible to retrieve information both at the level of forms (words, strings of characters, lemmas), and at the level of concepts expressed in the single parts of the texts.

The main form of dissemination envisioned by PKT is the one on-line on a server permanently connected to the internet. The encoding of the data, entirely performed in XML language, also allows distribution in paper format. As concerns this particular aspect, for the next phases of development of the system, the modes of production of the information managed by PKT on e-book will be taken into consideration. The introduction of e-book on the market should provide considerable medium-term increase percentages.

With regard to the current use of PKT:

- it is one of the text and image management components participating in a COST Action of the European Science Foundation (<http://www.interedition.eu/>);
- it has been considered the suitable technological basis for the project ERC (Advanced Grant) “Greek into Arabic”, approved and financed in the first days of November 2009 by the International SH5 Panel of ERC;
- it manages the corpus of the National Edition of the works of Galileo Galilei (<http://pinakes.imss.fi.it:8080/pinakestext/home.jsf>);
- two European project proposals are underway which in case of success could consolidate the position of PKT as infrastructure of research in the sector of sciences of the text.

Implementation of the system is the result of a collaboration between the Institute for Computational Linguistics “A. Zampolli” of the National Research Council of Pisa, the “Fondazione Rinascimento Digitale” and the Institute and Museum of the History of Science in Florence.