The Corpus dei Manoscritti Copti Letterari

Tito Orlandi

Università degli Studi di Roma, La Sapienza, Dipartimento di Studi Storico-Religiosi, Piazzale A. Moro 7, 00185 Roma, Italy

Abstract: The Corpus dei Manoscritti Copti Letterari is a project whose original aim was to reconstruct the Coptic codices from the White Monastery in Upper Egypt. The project was later expanded to include all Coptic literature. In 1980 a new project was launched to transfer the data into machine-readable form and make the information available, in as generic a format as possible, to scholars throughout the world.

Key Words: Coptic, Coptic literature, databases, text manipulation, non-Roman characters.

The project, which is now called Corpus dei Manoscritti Copti Letterari, started in 1968 as a conventional archive, mainly of photos of Coptic literary manuscripts. Its aim was the reconstruction of the Coptic codices coming from the library of the White Monastery in Upper Egypt, which were dispersed in libraries throughout the world, and the study and publication of their contents.

The program was later expanded to include the whole field of Coptic literature, and in the process unique archives of data and photographs were gathered. The work was concentrated upon the following items: (i) photographic archives; (ii) catalogue of manuscript collections; (iii) history of the manuscripts; (iv) catalogue of Coptic literary texts; (v) reconstruction of the White Monastery Library; (vi) bibliography of the Coptic literature, then general Coptic bibliography; (vii) publication of texts with introduction and translation.

In 1980 a new project was started, to transfer the data of the archives into electronic form, in order to allow for automatic processing. We used a database management program OMNIDATA working in the SPERRY 1100/80 mainframe of the Centro di Calcolo of the University of Rome, and several datafiles were created: (i) descriptions of manuscripts; (ii) Coptic bibliography; (iii) inventory of the collections of manuscripts; (iv) catalogue of literary works in Coptic.

After the experiences of the following years, with different database and (later) text manipulation programs, the organization of the work was changed. Presently, the Corpus dei Manoscritti Copti Letterari is an enterprise whose aim is to treat in a comprehensive but also articulate and flexible way, and — where convenient — automatically, the reproduction of the coptic literary manuscripts and all information on them and on subjects related to them (scribes, authors, texts, production, readers, collections, scholars), and to disseminate the scientific results thereof by the most appropriate technical means.

The functions of the project are organized around two main tasks: description and reproduction, of manuscripts, works, etc. The description is obtained by establishing connections between the information about Coptic manuscripts and literature, in order to form a consistent and possibly complete picture of the Coptic literary world. For
this database management techniques are used. The reproduction is obtained in "analogical" form through different photographic systems (mainly microfilm and microfiche); and in encoded (digital) form suitable for various automated text processing possibilities.

The problems posed by this approach have by now been partially answered, accounting for the fact that technical and theoretical progress keeps matters in movement. We should especially mention the recent choice of UNIX (but especially of the UNIX "philosophy") as the privileged environment in which the computerized work is done. This has brought a series of invaluable improvements in all levels of the organization.

The CMCL has to cope with a number of procedural problems: the identification of the different archives, through the definition of uniform characteristics of the objects taken into consideration; the organization of the information to be put in the records forming the archives; and the relations between the archives and cross references. In addition we have to concern ourselves with certain technical problems such as: the portability of the files that may be processed in different machines and by different programs; central updating of the files with simultaneous correction of data wherever necessary; visualization of the files convenient for the different steps in the management activity (screen or paper or microfiche; Coptic or Latin characters; etc.).

Because of these considerations, the work of the Corpus dei Manoscritti Copti Letterari is now carried out using many kinds of machines, from portable computers to mainframes (with their peripherals), and five basic kinds of programs: editor, word processor, text formatter, text analyz-
er (mainly concordance producer) in the different steps of information and text processing.

On the other hand, all information (texts, bibliography, archival data, etc.) is stored only in the form of pure ASCII files, without any form of interspersed codes produced by certain packages (notably word-processors and data-managers), but unintelligible to others.

The existing packages are used insofar as they neither require nor insert such codes, except for certain particular (generally final) purposes, or at particular points in the process, after which the texts are again stripped of non-ASCII codes. The products of the Corpus will therefore be available to scholars, not only in the more conventional forms of printed texts and microform, but also in files suitable for manipulation by most of the machines and software they normally use.

Textfiles
The files containing Coptic texts do not "repro-
duce" the physical shape of the text in any given manuscript, but are seen as a "kilometric" text (using the ASCII character decimal 12 for practical reasons in the visualization) in which the ASCII character will be adapted to encode all phenomena found in the manuscript in question and which are relevant to the preservation of the text in electronic form.

The codification ratio, i.e. the correspondence between Coptic and other necessary special char-
acters on the one hand, and the numbers (sequences of bits) stored in the memory of the computer, on the other, will be independent from those actually used in the keyboards and in the printing devices. This has been done because the keyboards and printing devices generally in use do not share exactly the same systems. It is nevertheless true that the systems in use are somewhat similar.

The codification system has been studied in order to facilitate the input of the texts by the scholars through the keyboards normally in use today. It is understood that the encoder of one manuscript, or part of it, will in no way change the Coptic text, not even separate words. He will only read what is surely readable, encoding the Copic text as it appears in the manuscript, and encoding other relevant information according to the chosen rules.

For encoding purposes, each Coptic letter will correspond to one ASCII character which on normal keyboards and printing devices corresponds to one alphanumeric or special character. Inside the text, the following other relevant information will appear: end of line; end of column; beginning of page with eventual numeration; punctuation (in several forms); capital in the margin; physical lacunae; illegible letters; separator (a special Coptic orthographic feature); raised dot; apostrophe; blank; marginal glosses.