TABLE OF CONTENTS

2. Text Encoding in the Humanities...................................................................................... 1
2. Text Encoding in the Humanities

Since the earliest uses of computers and computational techniques in the humanities at the end of the 1940s, scholars, projects, and research groups had to look for systems that could provide representations of data which the computer could process. Computers, as Michael Sperberg-McQueen has reminded us are binary machines that ‘can contain and operate on patterns of electronic charges, but they cannot contain numbers, which are abstract mathematical objects not electronic charges, nor texts, which are complex, abstract cultural and linguistic objects.’

This is clearly seen in the mechanics of early input devices such as punched cards where a hole at a certain coordinate actually meant a 1 or 0 (true or false) for the character or numerical represented by this coordinate according to the specific character set of the computer used. Because different computers used different character sets with a different number of characters, texts first had to be transcribed into that character set. All characters, punctuation marks, diacritics, and significant changes of type style had to be encoded with an inadequate budget of characters. This resulted in a complex of ‘flags’ for distinguishing upper-case and lower-case letters, for coding accented characters, the start of a new chapter, paragraph, sentence, or word. These ‘flags’ were also used for adding analytical information to the text such as word classes, morphological, syntactic, and lexical information. Ideally, each project used its own set of conventions consistently throughout. Since this set of conventions was usually designed on the basis of an analysis of the textual material to be transcribed to machine readable text, another corpus of textual material would possibly need another set of conventions. The design of these sets of conventions was also heavily dependent on the nature and infrastructure of the project, such as the type of computers, software, and devices such as magnetic tapes of a certain kind that were available.

Although several projects were able to produce meaningful scholarly results with this internally consistent approach, the particular nature of each set of conventions or encoding scheme had lots of disadvantages. Texts prepared in such a proprietary scheme by one project could not readily be used by other projects; software developed for the analysis of such texts could hence not be used outside the project due to an incompatibility of encoding schemes and non-standardization of hardware. However, with the increase of texts being prepared in machine-readable format, the call for an economic use of resources increased as well. Already in 1967, Michael Kay argued in favour of a ‘standard code in which any text received from an outside source can be assumed to be.’

This code would behave as an exchange format which allowed the users to use their own conventions at output and at input.