

The Encodability of Historic Scripts in ISO/IEC 10646 / Unicode®

Carl-Martin BUNZ

0. During the last decade the international encoding standard ISO/IEC 10646-1:1993 / Unicode has been and still is being elaborated as to comprise all important scripts used to represent the languages of the world.¹ Only a few historic scripts, however, have been taken into account so far, since the primary goal of the standard is to provide an encoding for industrial and commercial purposes. This is not to say that ISO and Unicode deviate from their aim to support historic scripts as well. Rather, the standardization committees postponed the consideration of the academic problems involved when historic scripts are prepared for a script based abstract character encoding, obeying to the character/glyph operational model². In principle, the Unicode designers understand the encoding standard as an overall area where virtually the entirety of all characters ever developed by man should find their proper code points. This idealistic concept had of course to be tailored so as to meet the practical needs of the standard. Therefore the 16-bit encoding space of ISO/IEC 10646-1 / Unicode must of necessity be reserved for living scripts currently used in written communication. The encoding of other scripts is planned for Plane 1 and higher planes of the 31-bit encoding space of ISO/IEC 10646³.

1. ISO has not been, however, idle with respect to historic scripts. Recently, these efforts culminated in an encoding proposal for Egyptian hieroglyphs,

¹ The documentation volume of version 3.0 of the standard is currently in the printing process, distribution is expected for the beginning of 2000. Provisionally reference can be made to Unicode 3.0 Beta exhibited at <http://www.unicode.org/unicode/standard/versions/Unicode3.0-beta.html>. The fully documented version of the standard is 2.1, the book available is a description of Unicode 2.0 (Unicode Standard).

² ISO TR 15285.

³ On the architecture of Unicode and ISO/IEC 10646 and the relationship between the codespaces cf. BUNZ 1998, 45 figure 1, FREYTAG 1999, KSAR 1999 and, in future, BUNZ in prep. The issue of universality, including historic scripts, is dealt with in BUNZ 1997.

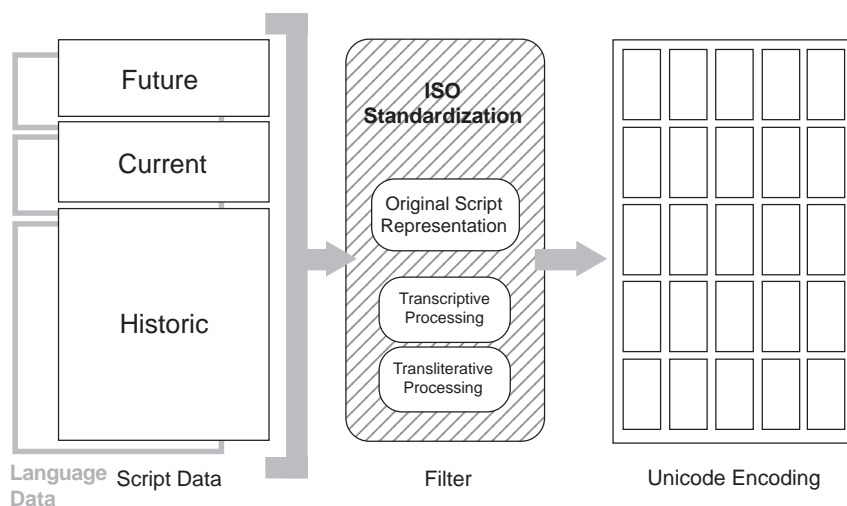


Figure 1: The Unicode Idea: Universal Character Set

a detailed documentation of more than 3 MB in size (see below). Except for a few instances, the encoding proposals for historic scripts have been compiled without participation of the scholarly community. Necessarily the complexity of the problems has been widely underestimated. In most cases, the information material which the proposals are built upon is outdated and does not match the current stage of linguistic and philological research. The situation is aggravated by the fact that the authors of the proposals, being non-specialists in the field, do not even raise the question of the encodability, but take the script units documented and described in the selected literature, as given entities which could enter as such into an abstract character encoding. Under this condition, the scientific value of an encoding cannot be estimated.

2. From the scientific point of view, the state of transmission of ancient texts and text corpora is very disparate. There are well described character repertoires ready for standardization, but also inventories that are taken from a dozen extremely fragmentary texts. Often the phonetic value of the graphic units has not yet been established. In such a case a character encoding would rely on graphic distinctiveness only without the function of the units with respect to the language they represent being understood. Of course, the notion

of standardization is not applicable at all to these encodings. In the center of the scale of encodability we may locate historic scripts that could be standardized under certain conditions but at any rate with considerable restrictions.

Linguistics, philology and the other disciplines concerned with historic languages and scripts are strongly appealed to comment on the proposals put forth so far.⁴ In several cases this means to articulate even fundamental methodical reservations suggesting that for certain scripts encoding projects have to be abandoned altogether. Otherwise standardizing bodies and scholars might feel alienated from each other; then, the international encoding standard, unchangeable by definition, would contain encodings of historic scripts which only the amateur who is not interested in scrutinizing the script data on the basis of the state of transmission can profit from, but who is dealing with the script units as if they were fantasy symbols like alphabets created in fiction (e.g. Tolkien's characters).

3. Bearing in mind the intention to mediate between engineers and scholars, I would like to propose, by means of a categorization of the historic script material, a strategy that might help to make up a reasonable and realistic roadmap for the encoding efforts of historic scripts in ISO/IEC 10646 / Unicode. Cooperation between both groups can be achieved only if they have sufficient mutual understanding for the different task and role the other has to fulfill. The national bodies (ANSI, DIN, BSI, AFNOR, etc.) as well as the standardization bodies on the European (CEN) and the international level (ISO) produce and approve standards in information technology and many fields of engineering, which the industry implements so that the worldwide data exchange between different locales can be performed on a unique basis. Scholars want to exchange data which do not claim any commercial and/or industrial interest. Especially in the case of historic scripts, the scope of data handling differs fundamentally from the point of view of the engineer, in that the scholar, for a given encodable character set, constantly questions and discusses the encoding process itself. Only exceptionally, he will be able to work with a certain encoding once agreed upon for a long

⁴The list of documents and Web addresses given in BUNZ 1998, 57f. is not yet completely outdated. Unfortunately, there is no room for an updated version of the list in this paper.

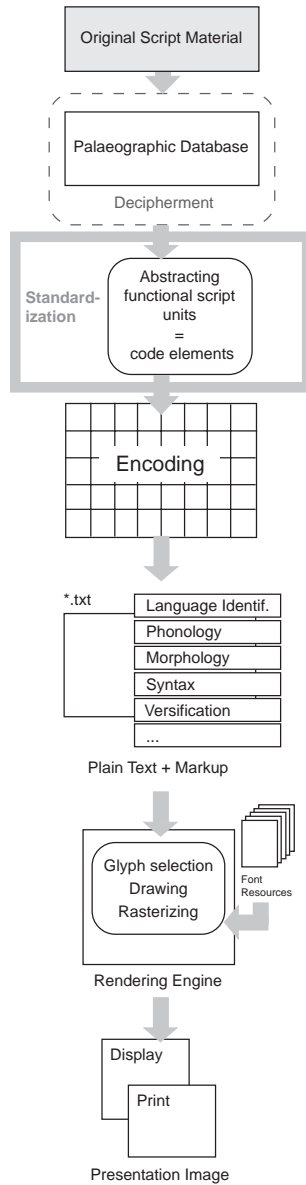


Figure 2: Standardized Representation of Historic Scripts

period. As research proceeds, the encoding becomes not only obsolete but also academically unusable.

Therefore, this paper has to describe briefly the general methodical difficulties which one encounters when putting forward scientifically based encoding proposals for historic scripts. What is the most problematic point is the notion of standardization. We are not talking here about facsimile encoding. This is clearly outside the scope of standardization. In case of a given ancient text document (stone inscription, manuscript, etc.) it cannot be the aim of an encoding standard to provide means to encode, in the sense of depicting electronically, all the features of philological and linguistic interest. Facsimile encoding may of course use an abstract character encoding as a basis, but then document specific features have to be indicated by a mark-up as e.g. SGML conformant tags.

Therefore, within the ISO and Unicode environment, the treatment of historic scripts is reasonable exclusively under the condition that an abstract character encoding can be designed. The flow chart given in figure 2 illustrates the stages of original script encoding, processing and rendering. Especially the starting point is different from that required by scripts with a living practice of writing and printing. While in the case of currently used scripts, tradition itself defines the encodable character repertoire, the acceptable letter forms etc., historic script data first have to be collected in a palaeographic database. Building up the palaeographic database normally is,

but need not necessarily be performed along with decipherment. Of course the palaeographer will take into account the phonetic and/or lexical value of a script unit once the meaning is established. In many cases he will have to reassign certain letter / symbol forms which have come out to be used distinctively on the level of language representation.

The next step towards an encoding norm may be called *standardization*. It includes the abstraction of functional units which is equal to the definition of possible code elements. According to the Unicode design principles, these units are functional primarily on the script level, but one has to bear in mind that such a reduction to script functionality is achieved by a retrograde abstraction from the language dependent use of the script units. Therefore this does not mean that establishing an encoding *standard* – in contrast to a *working version* – of a yet undeciphered script is feasible. On the contrary, it is pointless to define a *standard* for such repertoires.

As mentioned above, the plain text made up by code elements is liable not only to *language* specific but also to *document* specific mark-up.

The Rendering Engine interprets the tagged plain text, looking for appropriate font resources installed on the system. Font resources in the sense of outline data collections may be very sophisticated and adaptable to a special rendering purpose.

4. This is background against which the following categories of encodability have to be considered.

Category 1 (cf. figure 3) comprises historic scripts for the use of which there exists a public interest. The historic scripts in question serve as communication tools between social groups of the present day world. Examples are Ogham and (Germanic) Runes already encoded in the BMP in Unicode 3.0.⁵ There may exist secondary text material written in the scripts of this category, i.e. modern, not ancient, texts the processing of which will be based directly on the encoding derived from the palaeographic investigation.

Social groups include religious communities as well. In the case of Syriac, churches, scholars and standardization bodies achieved a consensus.

⁵ Ogham U+1680 through U+169F; Runic: U+16A0 through U+16FF.

Category 1: Scripts used in intercultural communication between social (ethnic, religious etc.) groups **of present day**; example: Avestan

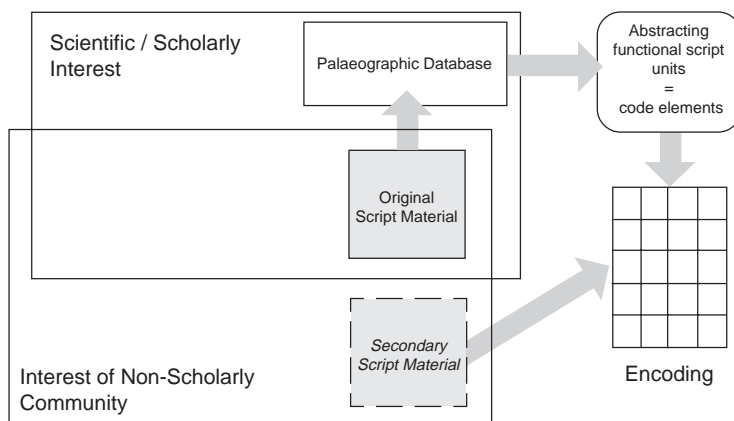


Figure 3: Historic Scripts Category 1

The encoding proposal compiled by all the parties involved⁶ has already been approved by the Unicode Technical Committee and subsequently by the relevant ISO committee; Syriac is now part of Unicode 3.0.⁷ A similar procedure would be possible for the Avestan script, if Iranian scholars and Zoroastrian religious communities cooperate.⁸ Although the palaeographic investigation of this script has not been carried out completely so far, an encoding can be designed. The establishment of standard glyphs must, of course, not be confused with the definition of abstract character codes. What is important is the fact that the functional units of the Avestan script are almost entirely recognized.⁹ The letter forms postulated as standard can be corrected afterwards in subsequent versions of the ISO documentation.

⁶ <http://www.unicode.org/pending/syriac/default.htm>.

⁷ U+0700 through U+074F.

⁸ Cf., for the Zoroastrian religious life, <http://www.avesta.org/avesta.html> with its extensive collection of links.

⁹ The most recent and at the same time most detailed work on the Avestan script and the problems of its development and its interpretation is HOFFMANN/NARTEN 1989. The ISO proposal (author: Michael Everson, date: 1998-01-18) is to be found at <http://anubis.dkuug.dk/jtc1/sc2/wg2/docs/n1684/n1684.htm>. The scientific sources of this proposal are FAUL-

Category 2a: Scripts of interest within cultural (and educational) policy (not communication), **encodable** from the scientific point of view; example: Old Persian

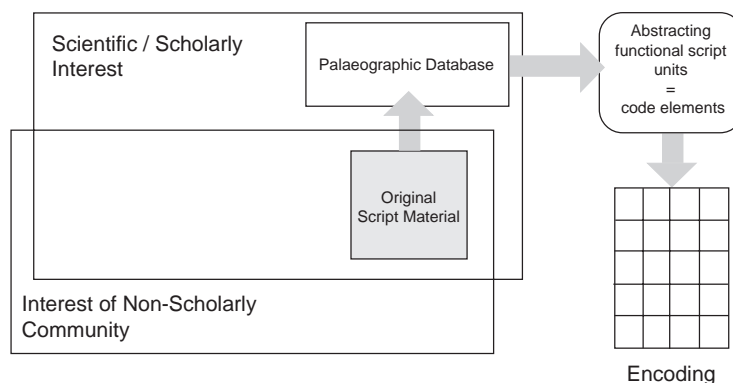


Figure 4: Historic Scripts Category 2a

Historic scripts of category 2 certainly are of politico-cultural interest, but due to the state of their transmission and of the text corpus preserved from antiquity they are in the first place an object of scientific research. This category has two subdivisions:

There are historic scripts which from the palaeographic point of view are ready for an abstract character encoding, but designing normalized glyphs is not feasible since the attested forms differ considerably both diachronically and diatopically (category 2a, cf. figure 4). Often this situation occurs with scripts that are known from small text corpora only, e.g. the Old Persian cuneiform script. Principally, in such a case the scientific profit of an encoding standard is to be questioned. The administration of script data, however, would be more efficient if the functional units, clearly discernable after all, were assigned individual code positions.¹⁰ In his comment on the ISO

MANN 1880 and HAARMANN 1990, which means that neither special literature has been consulted nor competent researchers in the field have been contacted.

¹⁰ The relevant scientific literature on the Old Persian script is given in SCHMITT 1989, 56-85, esp. 61-65. A proposal was put forward by M. Everson again (<http://anubis.dkuug.dk/jtc1/sc2/wg2/docs/n1639/n1639.htm>; date: 1997-09-18) to ISO/IEC JTC1/SC2/WG2. A revised code chart has been published by the same author at <http://www.indigo.ie/egt/standards/iso10646/opc2.pdf>. The resources of information used are not specified.

Category 2b: Scripts of interest within cultural (and educational) policy (not communication), **not yet encodable** from the scientific point of view; example: Egyptian Hieroglyphs

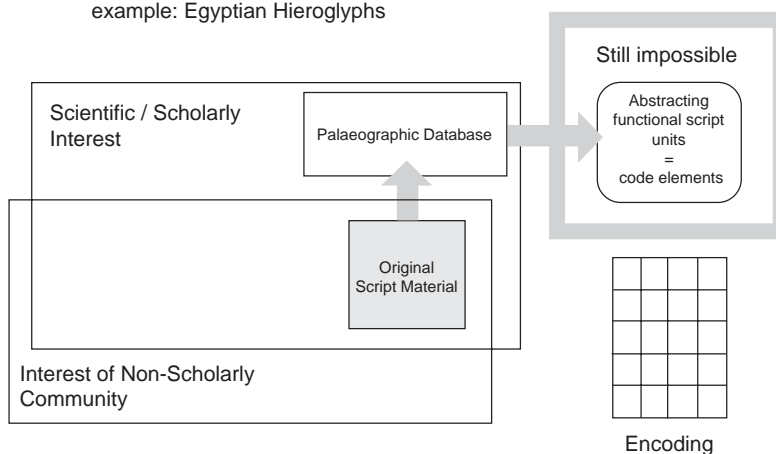


Figure 5: Historic Scripts Category 2b

encoding proposals of historic scripts, Wolfgang Röllig, specialist of cuneiform and semitic alphabetic scripts, characterizes the encoding of the Old Persian script as realizable though relatively unimportant in practice.¹¹

Category 2b (figure 5) corresponds to category 2a with the exception only that the standardization process cannot be performed for the time being because the palaeographic evaluation of the script material has not been carried out to such an extent. A script of widespread and lively interest in the non-scholarly world are the Egyptian Hieroglyphs. A great amount of printed texts is currently used inside and outside the academic world, but the lead types as well as the digital outlines do *not* reflect up-to-date palaeographic findings. Wolfgang Schenkel, one of the leading egyptologists, commented on the amateur proposal exposed on the Web in the beginning of this year,¹² stating that “at this point in time and on the basis of registers of hieroglyphs currently available a standardization within Unicode cannot be recommend-

¹¹ Cf. RÖLLIG 1999, paragraph 1 end: “... I would like to stress the fact that such standardized forms are merely useful for specific, very limited purposes, possibly for the editions of specific collections of inscriptions that are not concerned with palaeographic questions”.

¹² <http://www.indigo.ie/egt/standards/iso10646/pdf/N1944.pdf>.

Category 3a: Scripts of scholarly interest only, **not encodable for systematic reasons**; example: Major Cuneiform Script of Ancient Mesopotamia

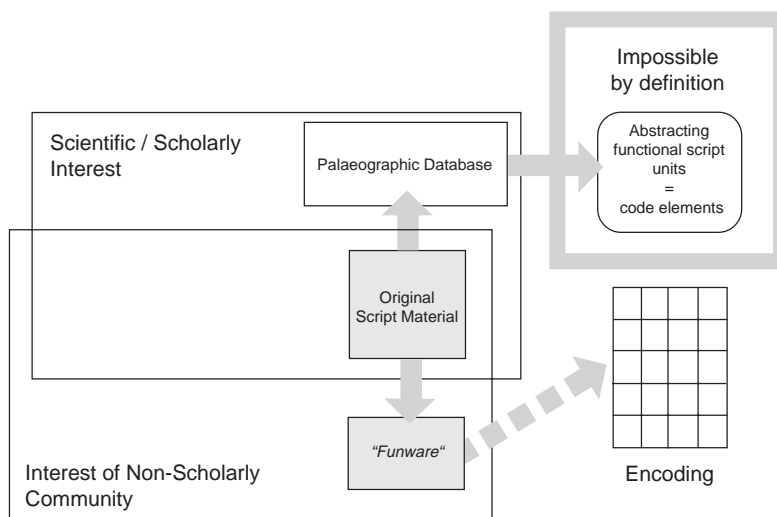


Figure 6: Historic Scripts Category 3a

ed”.¹³ Existing lists of hieroglyphs are derived from these printing types and therefore inappropriate as a basis for abstract code elements. In the near future, however, an encoding standard of a basic set of hieroglyphs used in classical times of the ancient Egyptian civilization can be envisaged.¹⁴

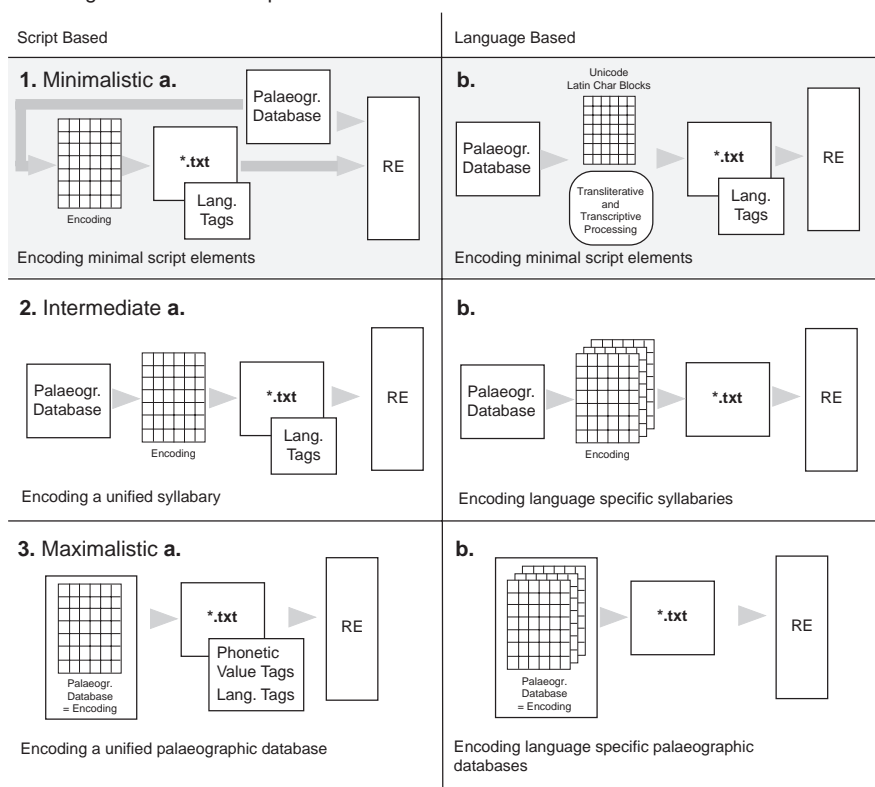
Category 3 is constituted by historic scripts dealt with almost exclusively in the academic world. The interest of the non-scholarly community may exist, but by virtue of the complex structure of the scripts and/or their fragmentary transmission, any non-expert application, including fonts, must be designated as funware.

Subdivision 3a (figure 6) contains scripts which are *unencodable* by definition, as hard and uncompromising such a verdict may seem to anyone not involved in the scientific research. The cuneiform scripts of Ancient Mesopotamia (Sumerian, Akkadian, Babylonian and, geographically shifted westwards but historically in the same tradition, Hittite cuneiform script)

¹³ SCHENKEL 1999, 2 (“Summary”).

¹⁴ SCHENKEL 1999, 1 paragraph 6, and 2, Summary.

Encoding of Cuneiform Scripts



RE = Rendering Engine

Figure 7: Encoding of Cuneiform Scripts

are preserved in large text corpora, but at no stage of their 3000 years' tradition, any kind of standardization has been attempted by their users. Rather, all the documents are to be considered as manuscripts written in a personal cursive variant, the so-called “ductus” of the scribe.

Figure 7 illustrates the possible approaches to the encoding of cuneiform scripts. In principle, one has to distinguish between script based and language based encoding methods, since the script represents several languages some of which are genetically and typologically different. ISO/IEC 10646 / Unicode being a script based encoding, a conformant treatment of language dependent cuneiform script variants would be a unified character block anyway. Then,

three degrees of explicitness could be considered when the encoding is designed: A minimalistic approach, in case of the script based method, would define the minimal graphical components of the cuneiform signs as code elements, which have to be composed by the Rendering Engine (RE). As part of the RE, look-up tables are derived from the palaeographic database so that the character composition is working according to templates. On the other hand, when language is taken into account, minimalistic processing means that the original script material is converted into a representation by Latin characters; the mark-up of the resulting plain text introduces the language specification. An intermediate approach would consist in the encoding of a unified syllabary, i.e. of precomposed syllabic characters, or separate encodings for different syllabaries would have to be defined in case the encoding is expected to convey language information as well. A maximalistic encoding method would project the palaeographic database as such, either unified or language specific, onto the encoding space.

The nature of the text documents described above forbids the intermediate as well as the maximalistic approach. Only the minimalistic methods are reasonable; the encoding of minimal graphical elements, however, is of very limited use in academic practice, i.e. in didactic programs for teaching the principles of cuneiform script. The majority of electronic text processing in the field of cuneiform writing has been so far and will in future be performed by transliteration and transcription, whereas palaeographic discussion continues using so-called “autographs”, i.e. hand-written pen-and-ink copies, or, electronically, three-dimensional images of the original text carriers.

In spite of all this, many efforts have been and still are being made to create cuneiform fonts, even in the academic world, although printing of texts with the help of cuneiform lead types, which was practised for a short period in the beginning of this century, has been abandoned long since. Specialists of cuneiform script agree upon the fact that any kind of standardization introduces an intrinsically foreign element to cuneiform script which does not contribute anything to its analysis and description.

Finally, we may differentiate category 3b (figure 8) which comprises those scripts that not only by virtue of the document type but also of the quantity of preserved text material cannot be subject to standardization and

Category 3b: Scripts of scholarly interest only
Encoding not reasonable: attested material not yet explored / explorable sufficiently; example: Phaistos Disk Script

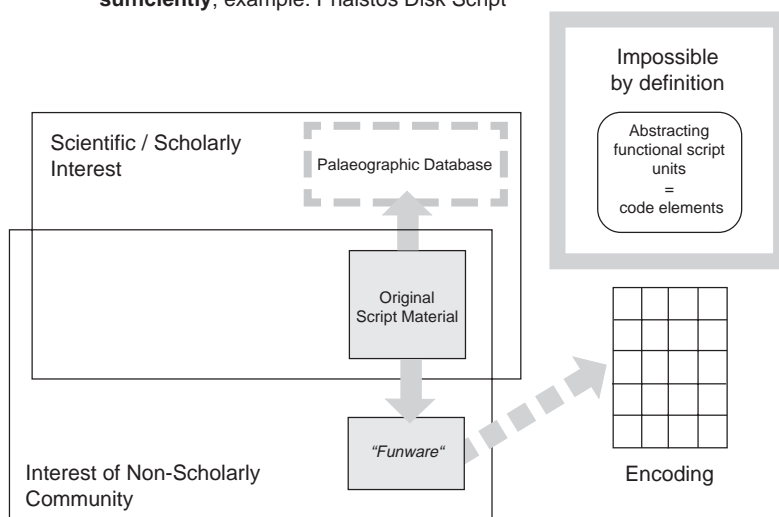


Figure 8: Historic Scripts Category 3b

abstract character encoding. Any consideration of including, e.g., the script of the Phaistos Disk, Old South Arabian, Phoenician etc. in ISO/IEC 10646 is of almost no use for scientific purposes proper. In these cases, encoding would mean either to project the palaeographic databases onto code points, or to set up standard character forms that are not justified by the data available.¹⁵

5. Summing up I would like to express the hope that the proposed classification of encodability will help to decide what priorities have to be taken into account in the encoding debate with regard to historic scripts. Both scholars and amateurs must understand that an international encoding *standard* cannot represent an archive which contains the entirety of the scripts that have appeared during the history of writing. The notion of standardization itself

¹⁵ Cf. RÖLLIG 1999. In ISO and Unicode environment, encoding proposal for scripts of this category are exhibited at <http://www.unicode.org/pending/pending.html> and <http://www.indigo.ie/egt/standards/iso10646/>. See also BUNZ 1998, 57f.

excludes historic script material, from which code elements cannot yet be derived in the specific sense of abstract characters. If ISO and Unicode, without consulting the scholarly community, proceed past the planning stage and approve encodings of scripts of category 2 and 3, then the standardization bodies will have acted to the detriment of the standard. On the other hand, if the scholarly community refuses to participate, it prevents itself from storing and transmitting script data to be encoded according to the international standard. The disadvantage might not be evident immediately, but multiplied proprietary encodings, in cases where conformance with the standard could be achieved, severely hinder electronic communication.

Unfortunately this paper exposes rather theoretical considerations and does not illustrate them sufficiently. This as well as detailed bibliographic documentation of the scientific research in question, must be the subject-matter of a somewhat larger article.

References

- BUNZ, C.-M. (1997): Browsing the Memory of the World. In: *11th International Unicode Conference*, San Jose, California, 2.-5.9.1997, Proceedings, 1, A 7.
- BUNZ, C.-M. (1998): Unicode® and Historical Linguistics. In: *Studia Iranica, Mesopotamica et Anatolica* 3, 41-65.
- BUNZ, C.-M. [& N.N.] (in prep.): Einführung in Unicode® (working title; revised version of *GLDV-Herbstschule 1998 "World Wide Web & Linguistik", Kurs 5: Unicode®, Teil 1: Einführung (unter Berücksichtigung sprachwissenschaftlicher Gesichtspunkte)* – publication in preparation [planned for 2001]).
- FAULMANN, C. (1880): *Das Buch der Schrift, enthaltend die Schriftzeichen und Alphabete aller Zeiten und aller Völker des Erdkreises. Zusammengestellt und erläutert von Carl FAULMANN.* Wien, 2nd ed. (reprints: Nördlingen, Greno 1985; Frankfurt/M., Eichborn, 1990).
- FREYTAG, A. (1999): Introduction to Unicode 3.0. In: *15th International Unicode Conference*, San Jose, California, 30.8.-2.9.1999, Proceedings, Pre-Conference Tutorials, TA 2 (abstract: <http://www.unicode.org/unicode/iuc15/a013.html>).
- HAARMANN, H. (1990): *Universalgeschichte der Schrift.* Frankfurt/M./New York.
- HOFFMANN, K. & J. NARTEN (1989): *Der Sasanidische Archetypus. Untersuchungen zur Schreibung und Lautgestalt des Avestischen.* Wiesbaden.
- ISO/IEC TR 15285:1998: Information technology — An operational model for characters and glyphs. ISO, Geneva, 1998-12-15.

- ISO/IEC 10646-1:1993: International Organization for Standardization. Information Technology – Universal Multiple-Octet Coded Character Set (UCS) – Part 1: Architecture and Basic Multilingual Plane. Geneva 1993.
- KSAR, M. (1999): Unicode and ISO 10646: Achievements and Directions. In: *15th International Unicode Conference*, San Jose, California, 30.8.-2.9.1999, Proceedings, 1, B 9 (abstract: <http://www.unicode.org/unicode/iuc15/a313.html>).
- RÖLLIG, W. (1999): Comments on proposals for the Universal Multiple-Octet Coded Character Set. Translation from German: Marc Wilhelm Küster. Document No. ISO/IEC JTC 1/SC 2 N 2097 (cf. <http://www.dkuug.dk/jtc1/sc2/wg2/docs/n2097.pdf>).
- SCHENKEL, W. (1999): Comments on the question of encoding Egyptian hieroglyphs in the UCS. Translation from German: Marc Wilhelm Küster. Document No. ISO/IEC JTC 1/SC 2 N 2096 (cf. <http://www.dkuug.dk/jtc1/sc2/wg2/docs/n2096.pdf>).
- SCHMITT, R. (ed., 1989): *Compendium Linguarum Iranicarum*. Wiesbaden (pp. 56-85 “Altpersisch” by R. SCHMITT).
- Unicode Standard: The Unicode Standard Version 2.0. The Unicode Consortium, Reading, Mass. etc. 1996.