Encoding Scripts from the Past: Conceptual and Practical Problems and Solutions

Carl-Martin Bunz, M.A.
Universität des Saarlandes, Saarbrücken, Germany

This paper outlines a strategy how to tackle the encoding of historic scripts in Unicode and ISO/IEC 10646. By means of a categorization of the historic script material, this strategy might help to make up a reasonable and realistic operative roadmap for the encoding efforts of historic scripts.

Introduction

At the IUC's 10 and 11, March and September 1997, Historical and Comparative Linguistics presented itself (cf. Bunz/Gippert 1997, Bunz 1997) to the designers and implementers of the Unicode® standard, soliciting understanding for the scientific purposes in multiscript and multilingual text processing from the standardizers. Historical and Comparative Linguistics felt that more could be done in order to integrate the needs of the scholarly community so that the international encoding standard be not an engineering product only but also a means for data transport and storage for research on script and language, especially in the field of "script archeology", i.e. the handling of historic scripts and the texts preserved.

In the course of the subsequent three years the dialogue between Unicode and Historical and Comparative Linguistics continued, and both sides learned from each other what are the crucial points which have to be treated if good compromises are to be settled.

This talk was given on the IUC 16 in Amsterdam (March 2000). I am grateful for being allowed to present the issue of historic scripts encoding once again in San Jose. In the discussions subsequent to the Amsterdam conference, I learned very much about the complexity of the problem, not so much as to the factual aspect, but rather regarding the different layers of interest which have to be taken into account whenever a historic script is being prepared for encoding in the standard. I will try to include, in this new edition of my contribution, all what I got aware of during the last months.

Historical and Comparative Linguistics were not and in many domains still are not ready to deliver encoding proposals for inclusion in the international standard. There are even instances where, in all probability, researchers will never be able to design an encoding for a historic script, nor will they be interested in doing so. On the other hand, especially on the World Wide

Web, the layman community is eagerly waiting for text processing facilities supporting the popular historic scripts such as Egyptian Hieroglyphs and Cuneiform, scripts which ever since have fascinated the learned public. Like this, their use, i.e. their reproduction very soon exceeded the domain of the study of the original texts. Consequently, these major historic scripts, but also a number of the minor ones, have been gaining a new dynamism, being used by a modern user community that feels no longer bound by the analysis of the original documents preserved from antiquity. Researchers have to acknowledge that this is a real user interest of the present day world. But they, too, have the right to point out what is required when an encoding of a given historic script should equally meet the scientific needs.

Researchers may not, however, forget that in their own domain presentation of ancient script material is an essential part of academic teaching. Moreover, researchers contribute to e.g. encyclopedias, manuals, reference works, where ancient scripts have to be presented. in this case the printers would be best served with a normalized encodings, since font vendors could supply special products with this encodings. This kind of script presentation should not affect or even force, of course, the encodability status of a given script.

As long as the different interests appear to be compatible, however, the groups involved should continue looking for a compromise, since an international encoding standard by definition is meant to be an encoding for general, not particular, reference. There will be instances where a compromise cannot be worked out. No compromise can be forced without running the risk that both parties refuse the outcoming standard. In these cases it is better to serve one than not to serve anybody. Then, language science either will design its proper encoding, constructing to-Unicode converters to enhance data exchange, or will continue working without any normalization.

This proposal does not concur with the roadmaps elaborated by the Ad hoc group on Roadmap of ISO/IEC JTC1/SC2/WG2 as exposed on the Web (Cf. `http://www.egt.ie/standards/iso10646/bmp-roadmap-table.html` and `http://www.egt.ie/standards/iso10646/plane1-roadmap-table.html`). Therefore, what is discussed here is called expressly *operative*, because it concerns the schedule for treating the encoding problems, not the allocation of the character blocks in question. Similarly, no claim is implied that a given historic script be encoded in Plane 0 / Basic Multilingual Plane, unless the principles held by Unicode and ISO favor the inclusion in the BMP anyway. Future OS and text processing software will support UTF-16 character values, so that Plane 1 becomes easily accessible.

The expert standardizers on the ISO and Unicode side have not been, however, idle with respect to historic scripts. In the course of the last decade, a considerable number of encoding proposals have been drawn up. Recently, these efforts culminated in a proposal for Egyptian hieroglyphs, a detailed documentation of more than 3 MB in size (see below).

Except for a few instances, these encoding proposals for historic scripts have been compiled without participation of the scholarly community. Necessarily the complexity of the problems has been widely underestimated. Often, 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 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.

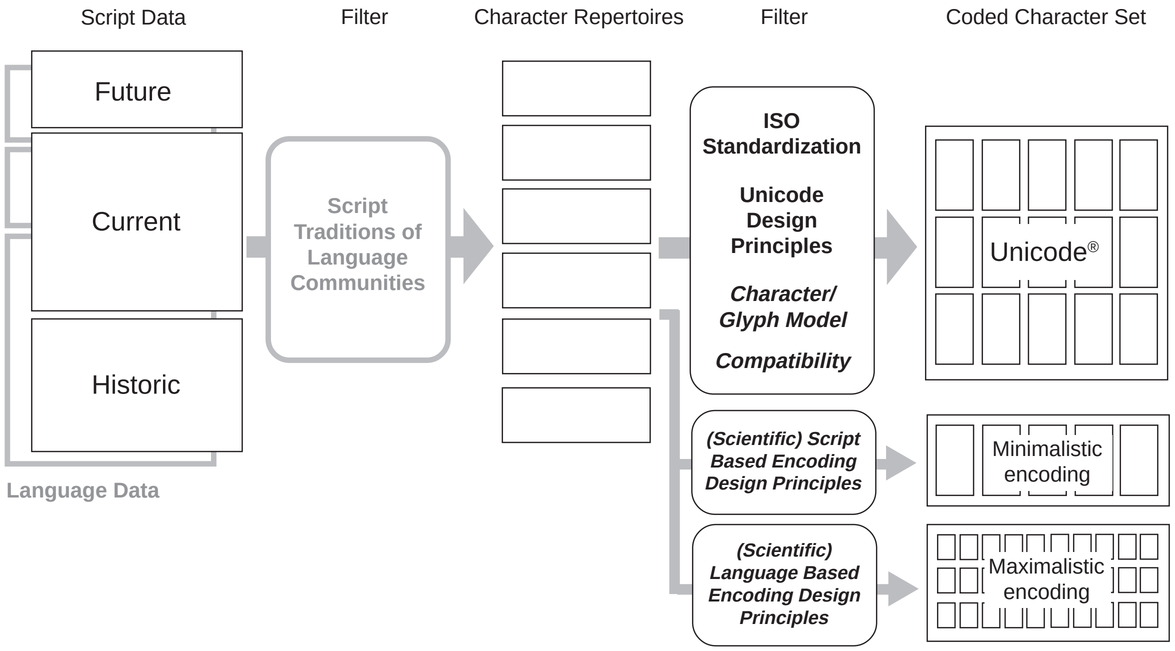
Besides the factual information to be contributed to the discussion, the scope of this talk is to maintain the dialogue between linguistic and philological research and standardization bodies. When the participation of the scholarly community often is felt as an obstacle to progress, then the explanations given here may clarify the conditions. Where scientific work and normalization contradict each other, severe conflicts cannot be avoided. So the scholarly community is hoping for a fair dialogue which has to continue during the 21st century and even beyond, since the work to be done exceeds the manpower of one generation.

The material collected and presented in this paper is not, however, complete in any respect. A thorough investigation on the encodability of all known historic scripts which are still missing in the standard, would require a book or a series of lectures. Certainly the examples chosen to illustrate the different constellations, do not show all aspects of the problems raised. Rather, the examples should highlight the most striking features that make historic scripts as difficult to deal with as they sometimes prove to be. Moreover, the insights into the researcher's daily work may interest the technical public in these studies which at the universities and other research institutions all over the world urgently need to be sponsored in order that they are able to continue their work.

I am indebted to academic teachers and colleagues of mine for discussing with me the relevant problems and supplying me with illustration material. I am grateful as well to Michael Everson, the most active administrator and promotor of historic scripts and other cultural issues within ISO and Unicode, who, in a very fruitful conversation in Berlin in May 2000, went through the diverse problems with me: This exchange repaired a good deal of misunderstandings which dwelled between standardizers and scholars, or, to use the term pair created in a recent Web discussion, expert standardizers and expert reseachers.

Of course I am solely responsable for inaccuracies and errors that remain in the text as well as in the visualizations.

Slide 1: The Unicode Idea – The Universal Character Set



Before entering into the difficulties historic scripts impose upon the code designer, we have to review briefly the general idea of an international encoding standard, in particular the Unicode encoding.

The basic material any character encoding effort which claims to be international and universal, as the Unicode approach does, has to start from is the totality of written documents existing in the world. But it is a serious problem to define what might be called historic and current respectively, when classifying script data. Script, of course, represents language, and so it is natural that script data normally are considered by virtue of the language. Consequently, in the course of writing practice from its beginning onwards, character repertoires have been defined as to represent a specific language, even if different language communities make use of the same script, i.e. the same *basic* inventory of graphic elements. It is the tradition of *language* communities which has described repertoires in use, even if they are aware of the fact that the letters they use are also employed by other communities, often not related linguistically. Like this, script users feel to apply the letters proper to their language. The more abstract consideration of scripts as language independent tools, and of their adaptation to specific graphemic requirements in order to cope with specific linguistic structures, is rather secondary and academic, cf. the script unification discussion in the encoding process.

Moreover, there is an important divergence between historic *language* data and historic *script* data. Currently many language communties use scripts which, in the sense of a language independent graphic representation system, have a long tradition going back to antiquity. So a Latin inscription from the Emperor Augustus, from this point of view, does not display a historic script, while an Egyptian inscription does, because the hieroglyphic script of the ancient civilization of Egypt went out of use during the first centuries of the Roman Empire. The Latin inscription represents historic *language* data, as does the Egyptian one. Therefore, if historic script data are those which are available in currently used writing systems, we have to consider, in the case of e.g. the Latin script, written documents from antiquity onwards.

Therefore, when applying the term *historic* to scripts, we designate inventories or writing systems on the whole which at present are out of use or have disappeared until researchers rediscovered them. A script is called *current*, if it is in use nowadays, not regarding the language(s) represented.

Scripts are modified as soon as they are applied to languages other than those the script originally had been designed for. The category *future* in the general classification of script data, is meant to comprise, in the first place, new script creations, but also modifications of existant repertoires. The question of unification will be posed only later.

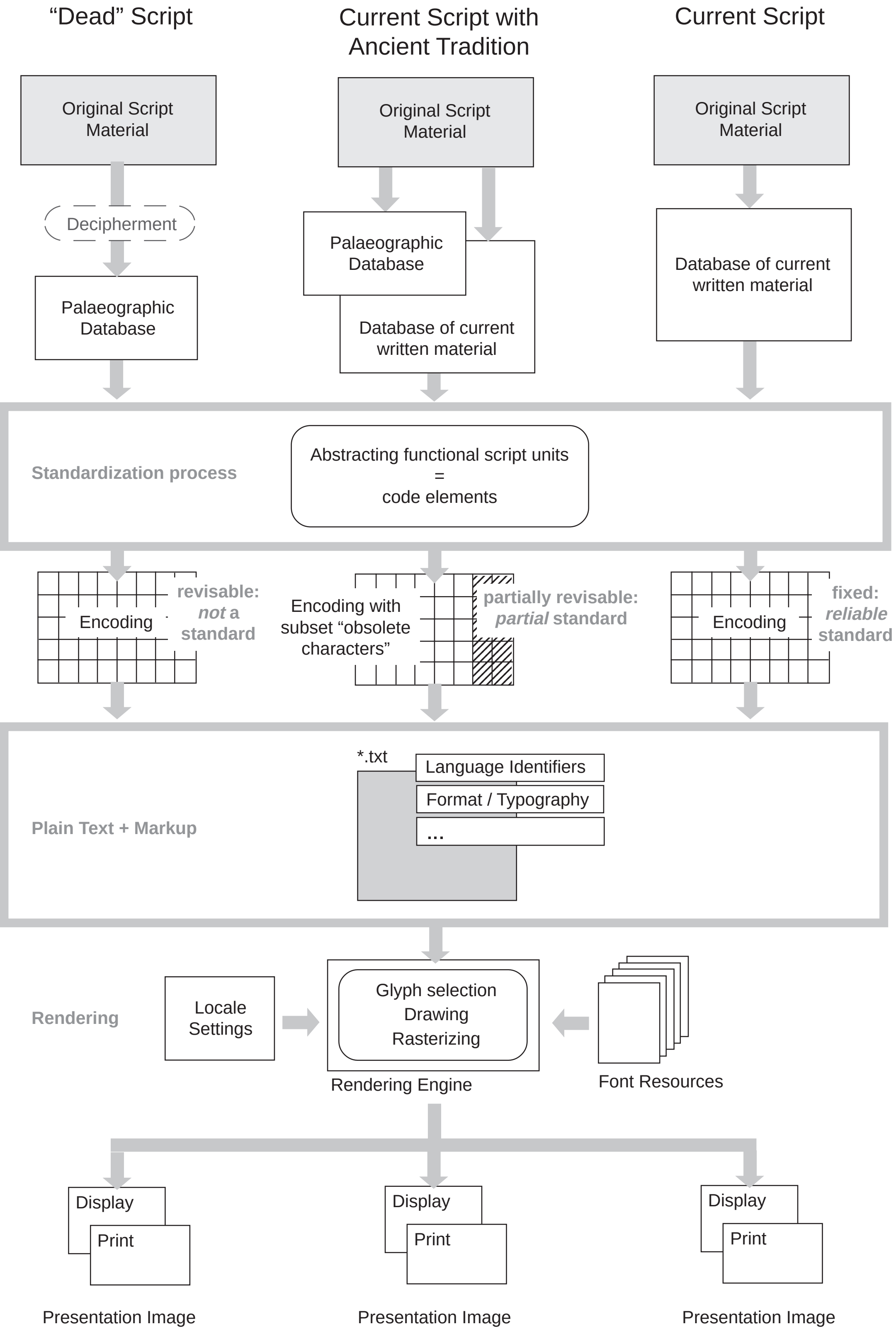
The character repertoires, therefore, are the material the encoding process is acting upon. It should be remembered that there are many concepts according to which an encoding for these repertoires could be designed. In order to achieve a practicable solution, the Unicode designers were bound to compromises, since a consistent encoding method does not yield a structure suitable for industrial and commercial use, i.e. for application in domains where a considerable amount of legacy is to be accounted for, because industry had already defined its way for coping with script and language data.

The fact that Unicode, in this respect, is not perfect, is well-known and often complained about by implementers. In this context, however, we do not simply reiterate this complaint. Rather, the systematic and conceptional questions are raised anew when historic scripts have to be prepared for inclusion in the international standard. When promoting Unicode in the scholarly world, it is precisely the conceptional inconsistency that discourages scholars to consider the standard as a basis for the administration of scientific data. Industrial purposes are outside the scope of research. On the contrary, evidence concerning linguistic structures as well as writing systems is always treated straightforwardly, without regard of implementations which had been made previously for commercial or technical purposes.

So we have to state that the Character/Glyph Operational Model in itself is a sound concept in order to build up a script oriented encoding. In its very consistency, however, it would end up in a minimal number of abstract characters, representing the distinctive values of the major script types (Semitic, Brahmi, Chinese (strokes), individual syllabaries). On the other hand, language oriented encoding would require a maximal encoding space. This is what ISO had been about to realize before its joint venture with Unicode in 1991. Unicode has been and is constantly searching a way in between these extremes, with the intention of best practicability.

For language science, taking part in the Unicode encoding process, compliant script handling often means abandoning methodical exactitude and agreeing to trade-offs researchers normally are not ready to accept. This paper, however, is meant to promote a sense of practicability in the scholarly community too: Why should science not profit from the international standard, even if its architecture does not follow strictly a definite theoretical model?

Slide 2: Designing an encoding for historic scripts



“Dead” Script

Current Script with Ancient Tradition

Current Script

Original Script Material

Original Script Material

Original Script Material

Decipherment

Palaeographic Database

Palaeographic Database

Database of current written material

Database of current written material

Standardization process

Abstracting functional script units = code elements

Encoding

revisable: *not* a standard

Encoding with subset “obsolete characters”

partially revisable: *partial* standard

Encoding

fixed: *reliable* standard

Plain Text + Markup

*.txt

Language Identifiers

Format / Typography

...

Rendering

Locale Settings

Glyph selection

Drawing

Rasterizing

Rendering Engine

Font Resources

Display

Print

Presentation Image

Display

Print

Presentation Image

Display

Print

Presentation Image

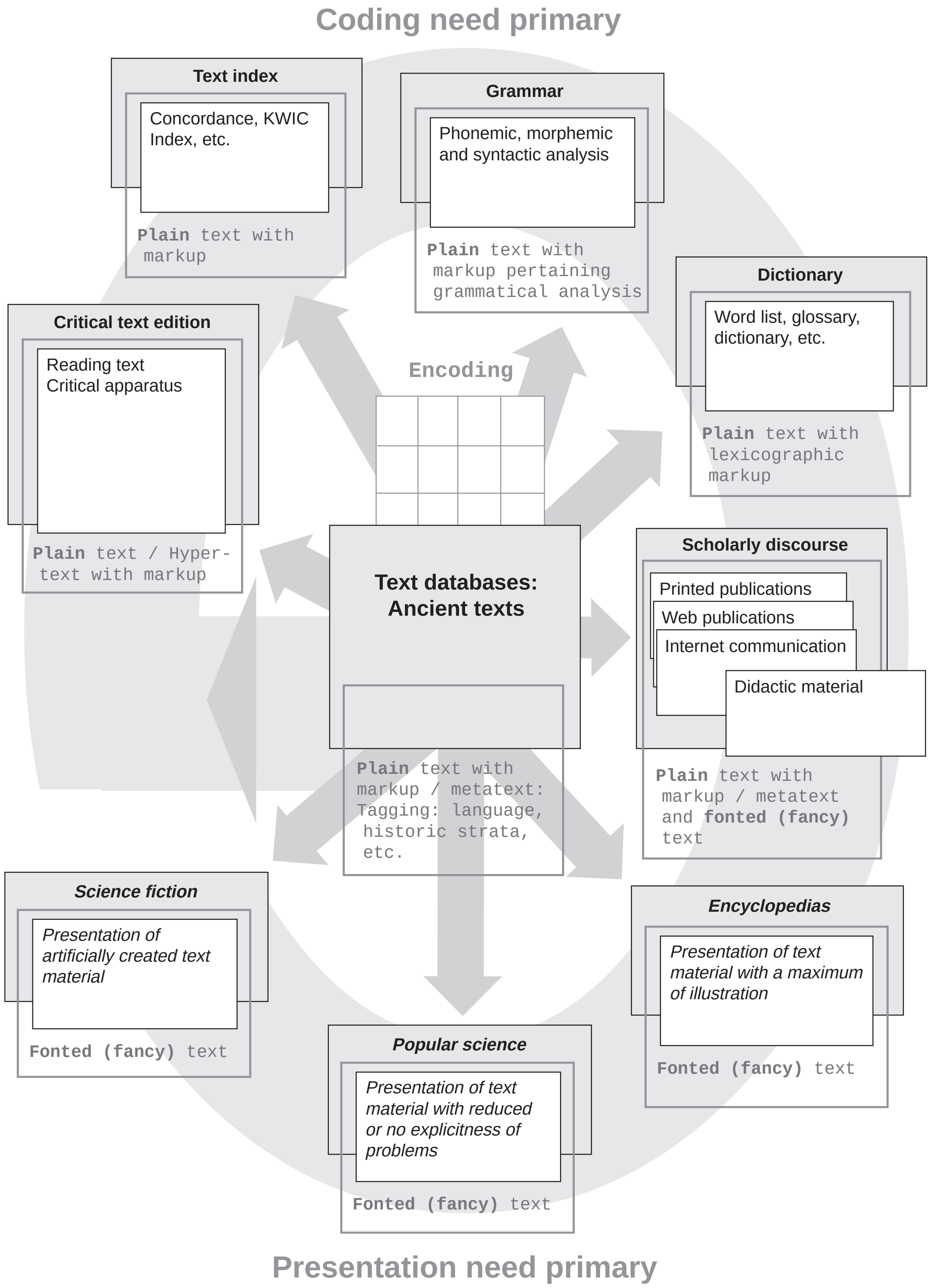
As far as the encoding process of historic scripts is concerned, certain entities do not have the same status in comparison with the treatment of current scripts. The flow chart given in Slide 2 illustrates the stages of original script encoding, processing and rendering, comparing the treatment of (1) "dead", i.e. properly historic, scripts, (2) current scripts with an ancient tradition, differing from its use at present, and (3) current scripts used hitherto without substantial modifications.

When a dead script is to be prepared for encoding, the starting point is considerably different from that required by scripts with a living practice of writing and printing. In the case of current scripts, the handling of letter forms in writing and typography defines what is the encodable material, i.e. the character repertoire, and, moreover, operates a constant preselection of encodable entities, in that the distinctiveness of letter forms is continously challenged in technical script applications as well as in artistic experiments with the graphical shapes, in commercial publicity labels etc. By contrast, historic script data first have to be collected in a palaeographic database, since the whole range of variation of letter forms is not accessible from the first sight, but depends of the number and kind of written documents preserved. Building up the palaeographic database normally is, but need not necessarily be performed along with decipherment. Working on an undeciphered script means grouping and classifying graphically similar units. Once the meaning is established, the palaeographer will take into account the phonetic and/or lexical value of a script unit. In many cases, then, he will have to reassign certain letter / symbol forms which have come out to be used distinctively on the level of language representation. Therefore palaeographic databases of undeciphered scripts are by necessity subject to an even thorough revision. Current scripts with ancient tradition may require palaeographic databases as well.

In order to prepare an encoding, the core task is standardization, i.e. the abstraction of functional units which is equal to the definition of possible code elements. In the case of historic scripts, however, the output of the standardization process is substantially different from what is yielded in the case of current scripts. The abstract character encoding of a historic script is revisable by definition, even every day, if new data have been added to the palaeographic database – therefore this encoding is *not a standard*, it is a mere working basis for character handling. Of course, the degree of revisabililty varies according to the level of analysis of the script data available. It is evidently pointless to claim for any reliability, when an encoding of an yet undeciphered script is proposed. The encoding of a current script, on the contrary, may be static for decades or even centuries, because it has been established upon an all-embracing evaluation of script use. An annex to the standard may contain obsolete characters, but this code range will not be as reliable as the values of the characters labelled "current".

Abstract characters and code values once being available, plain text can be tagged according to the processing requirements in a specific environment. Subsequently, 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.

Slide 3: Databases of ancient texts and their satellites



Coding need primary

Text index

Concordance, KWIC Index, etc.

Plain text with markup

Grammar

Phonemic, morphemic and syntactic analysis

Plain text with markup pertaining grammatical analysis

Dictionary

Word list, glossary, dictionary, etc.

Plain text with lexicographic markup

Critical text edition

Reading text
Critical apparatus

Plain text / Hypertext with markup

Encoding

Text databases: Ancient texts

Plain text with markup / metatext: Tagging: language, historic strata, etc.

Scholarly discourse

Printed publications

Web publications

Internet communication

Didactic material

Plain text with markup / metatext and **fonted (fancy)** text

Science fiction

Presentation of artificially created text material

Fonted (fancy) text

Encyclopedias

Presentation of text material with a maximum of illustration

Fonted (fancy) text

Popular science

Presentation of text material with reduced or no explicitness of problems

Fonted (fancy) text

Presentation need primary

Before approaching the coding issue according to different categories of scripts, we should consider the pivot status the text database assumes when ancient texts are to be processed electronically.

A well-designed text database provides textual information in a most straightforward form. It is quite obvious that formatting is alien to the inner structure of a database. Rather, plain text with precise markup is required, enabling efficient searching and retrieving of the data. According to the complexity of the text in question, the markup may be very complex as well, including several hierarchical layers.

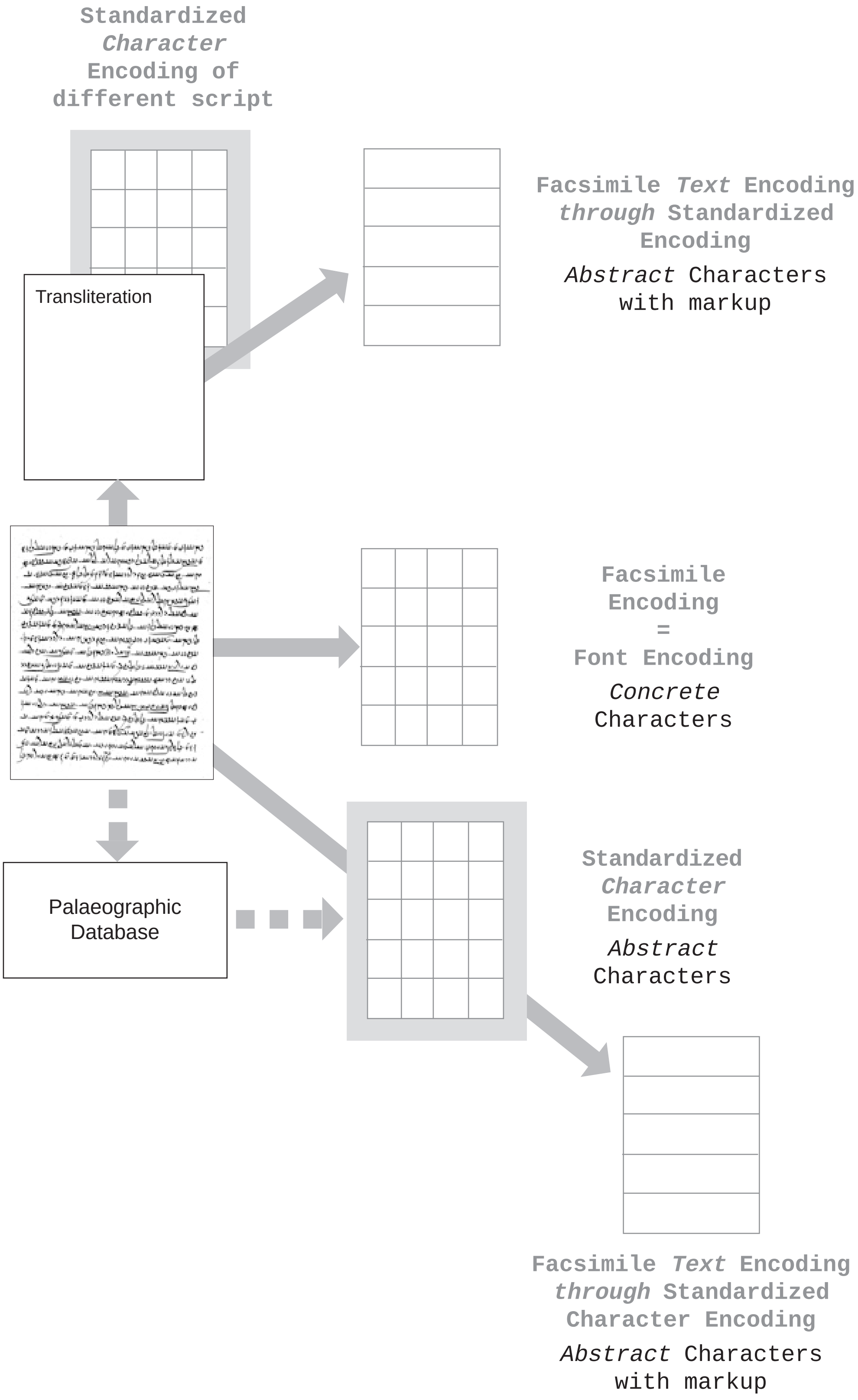
On Slide 3 a number of satellites are arranged around the text database. This is an idealized model, of course, in reality the satellites are often built up separately and independently. A sound framework of electronic processing and administration, however, would have the text databases as a kernel, the other entities being derived from it. The original text materials processed in these entities, dwelling partly in the scientific sphere and partly in the domain of popular science, are in fact nothing but extracts from and rearrangements of text databases. Consequently, the quality of the architecture of the databases is crucial.

The satellites are represented here beginning with the most scientific entity and ending up with the most artistic phenomenon, not controlled by stringent scientific regulations. The critical text edition can be, at a first stage, generated automatically by text comparison. In a second step, the researcher determines the criteria as to what variants are preferable to be included in the reading text the edition wants to establish. Text indexes are built directly upon text databases, with various parameters in order to extract certain information from the texts. A grammar of a historic language is by definition a grammar of the available texts – so it presupposes, in terms of electronic processing, yet another form of indexing, i.e. according to functional elements in the language. Similarly, a dictionary is a sort of categorized index of the ancient text, adding semantic information to each abstract lexical unit derived from this index. Up to this point, the need of *coding* the script material treated is primary.

In the realm of scholarly discourse, precise encoding of the material is essential on the one hand, particularly when original text material is quoted in internet communication or Web publications, but, on the other hand, presentation issues may relegate the coding concerns to the background, especially in case of didactic material which should offer a maximum of illustration to the learner. Therefore the scholarly discourse is a kind of transition area where, as far as the organisation of electronic data is concerned, marked up plain text, i.e. coded text information with additional functional determinants, and fancy text coexist.

When we leave the sphere of science proper, i.e. the definite disciplines that investigate the ancient texts in question, then fancy text is required since the treatment of the data aims primarily at presentation. This is to say that in this domain a character encoding of a certain historic script serves as reference grid for fonts. The recurrence to source text which is the basis of scientific analysis, is no longer necessary.

Slide 4: Text encoding vs. character encoding



Standardized *Character* Encoding of different script

Transliteration

Facsimile *Text* Encoding *through* Standardized Encoding

Abstract Characters with markup

Facsimile Encoding = Font Encoding

Concrete Characters

Palaeographic Database

Standardized *Character* Encoding

Abstract Characters

Facsimile *Text* Encoding *through* Standardized Character Encoding

Abstract Characters with markup

The technically educated reader may reproach me for treating trivia here, but these distinctions between text encoding and character encoding, including the term of facsimile encoding, are badly understood in the academic world, particularly among scholars dealing with ancient texts. So Unicode promotors must be prepared to expound this issue very explicitely to researchers who plan an electronic data processing strategy which might be suitable for their project.

Being faced with an ancient manuscript researchers often imagine a coding method which preserves a maximum of the information contained in the original, on *one* coding level. This would be called facsimile encoding. It corresponds to font encoding since every distinct form in appearance would be assigned a distinct code point. The resulting characters in this kind of encoding might be designated as *concrete* – in contrast to *abstract* characters as e.g. in the Unicode standard. This encoding is proprietary of one single text document. Automatic comparison with other encoded documents is impossible.

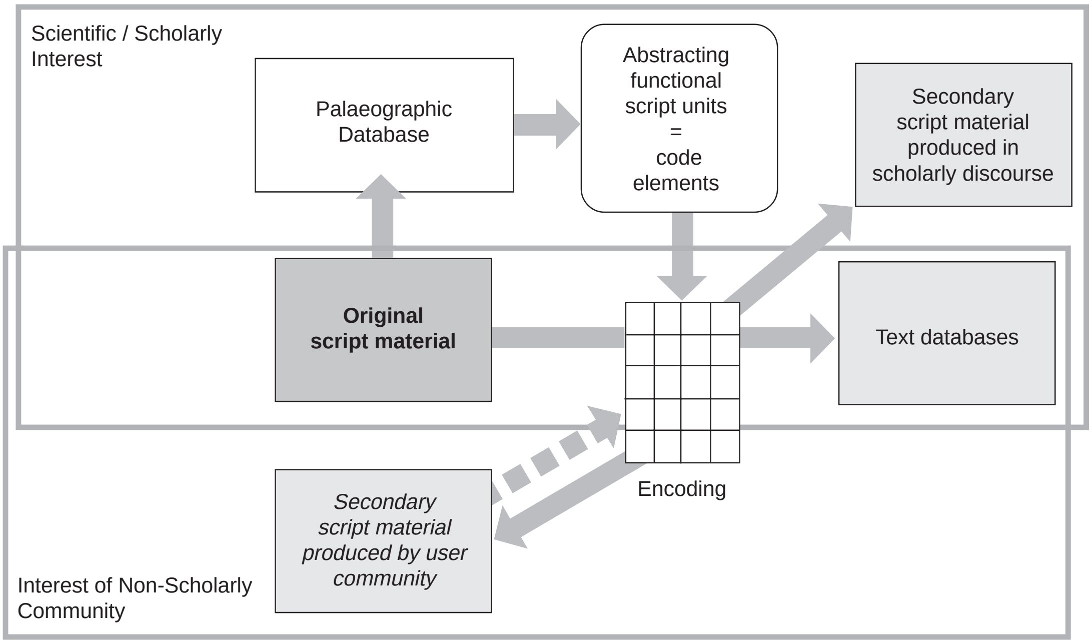
When an abstract character encoding has been established via a palaeographic database, the same goal can be achieved, but by a strategy that ensures full comparability. The registry of palaeographic variants enables the abstraction of characters which represent funtional values on a higher level than the presentation image. Like this, a *standard* can be defined. Using the standardized abstract characters, one can create again a facsimile encoding of the unique text document: then this is a text encoding containing explicit markup where the standardized characters do not convey the information to be preserved.

Evidently, as very often practised in linguistic and philological research, transliteration may precede the coding process so that afterwards no special character encoding is needed at all. Instead, the character encoding of the transliteration medium is used, when the document is treated, and an additional markup supplements the information. This method enabled the electronic processing of complex (historic) texts even in a 5-, 6- or 7-bit environment. But it also provides a means to cope with texts written in a script that cannot be standardized yet.

In the field of computing in the Humanities, the most important guide to text encoding is the Text Encoding Initiative (TEI, cf. `http://www.iuc.edu/orgs/tei/index.html`), recommending a specially regulated use of SGML for scientific text handling purposes.

Slide 5: Category A

Scripts used in intercultural communication between social (ethnic, religious etc.) groups **of present day**


Scientific / Scholarly Interest
Palaeographic Database
Abstracting functional script units = code elements
Secondary script material produced in scholarly discourse
Original script material
Text databases
Encoding
Secondary script material produced by user community
Interest of Non-Scholarly Community

We will now observe a series of scenarios, each illustrating the specific situation of a category of historic scripts. These are categories of encodability, defined with regard to the technical premises made in the previous section (Slide 2).

Category 1 comprises historic scripts for the use of which there exists a public interest on a wider range than a mere educational environment. The historic scripts in question serve as communication tools between social groups of the present day world. Social groups include linguistic and/or ethnic groups, but also religious communities.

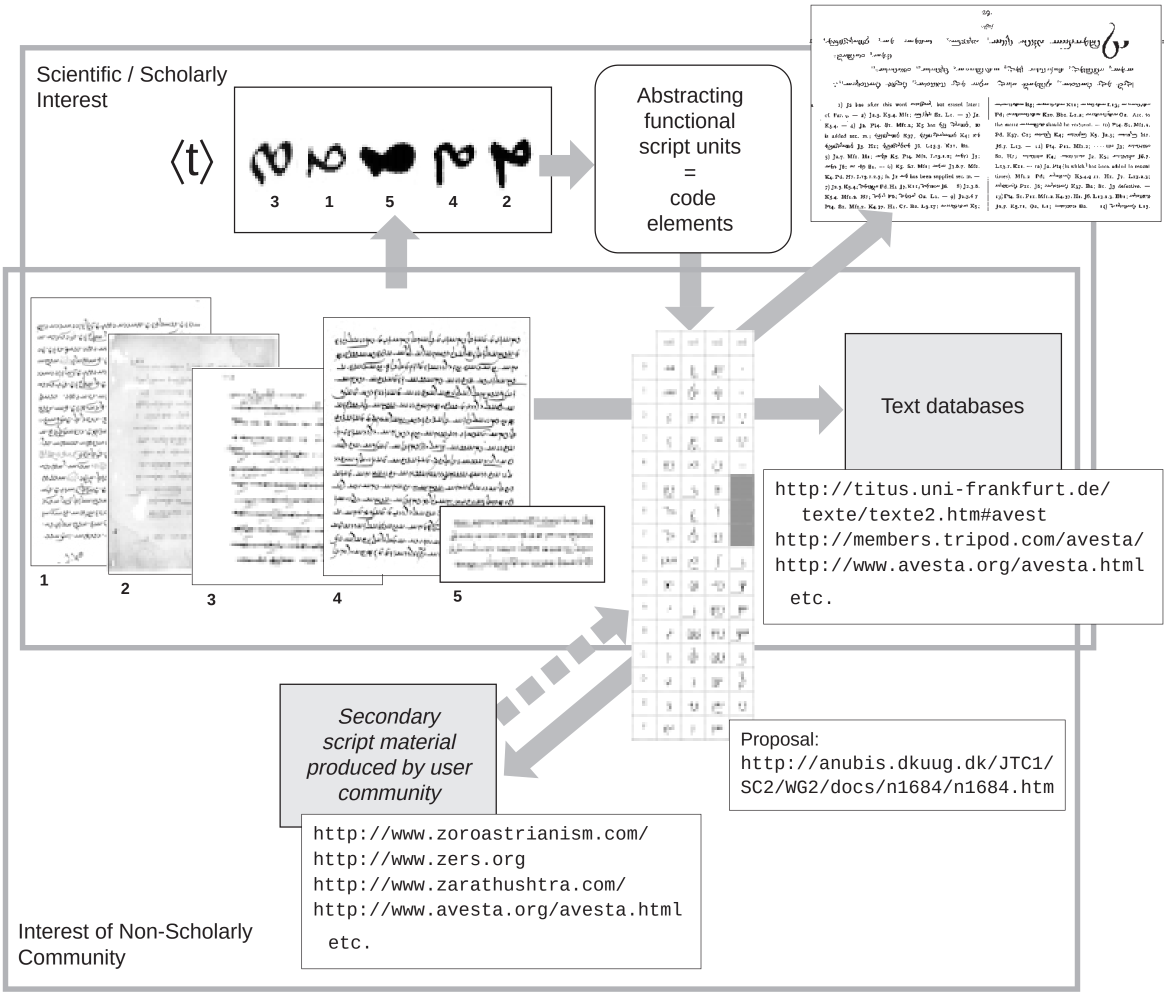
In a scientific context, the primary script material is properly analysed so that a palaeographic database can be established which allows for a quasi-standardization of functional script units. The resulting code values are fairly stable and are used for encoding the texts available in the relevant writing tradition (critical text editions, electronic text databases), but also for presenting secondary texts in this script, e.g. in scientific literature, didactic documents, textbooks etc. On the other hand, there exists secondary text material written or printed in the scripts of this category, *outside* the academic domain, i.e. in a context where the script material is not object of investigation but a means of conveying new information in the language of the original texts, be it rather exegetical or independent in content. It is quite natural that the modern user community prefers to rely on the encoding derived from the palaeographic investigation in order to process their texts. According to the state of scientific investigation as well as to the consciousness of the community, the direct path from the modern writing practice towards a character encoding is taken, which, however, causes the tradition to split into an objective, i.e. scientific, approach and a subjective, even esoteric practice. With religious communities, such a divergence often is inevitable.

Scripts of category A being part of the worldwide communication, Unicode and ISO should include them in the Basic Multilingual Plane. The resulting character blocks are not as large as allocation could be difficult.

Examples for category 1 are Ogham and (Germanic) Runes already encoded in the BMP in Unicode 3.0. (Ogham U+1680 through U+169F; Runic: U+16A0 through U+16FF). Syriac, recently allocated at U+0700 through U+074F, is a perfect example of how a consensus between churches, scholars and standardization bodies can be achieved (cf. `http://www.unicode.org/pending/syriac/default.htm`). The case of Avestan will be presented at some length. The Middle Persian Book Script, also called Pahlavi, the Mandaean script used in Iraq, and Glagolitic are further candidates. Besides, the disunification of Coptic and Old Church Slavonic, at present unified with Greek and Cyrillic respectively, is strongly desired by the scholarly community.

Slide 6: Category A

Example: Avestan



Scientific / Scholarly Interest

Abstracting functional script units = code elements

Text databases

http://titus.uni-frankfurt.de/texte/texte2.htm#avest
http://members.tripod.com/avesta/
http://www.avesta.org/avesta.html
etc.

Secondary script material produced by user community

http://www.zoroastrianism.com/
http://www.zers.org
http://www.zarathushtra.com/
http://www.avesta.org/avesta.html
etc.

Proposal:
http://anubis.dkuug.dk/JTC1/SC2/WG2/docs/n1684/n1684.htm

Interest of Non-Scholarly Community

Avestan is a yet unencoded script of this category. This is a phonetic alphabet created by the Zoroastrians, presumably about 400 B.C., on the basis of the so-called Pahlavi book script, in order to render precisely the pronunciation of the sacred texts as realized in ritual performance. Like this, Avestan is the first known instance of a narrow phonetic transcription. In encoding, unification with Pahlavi is not recommended, since the majority of the characters are specific modifications or new shapes added to the original set. Moreover, distinct code points would be preferable for the processing of complex texts in the Avestan and Middle Persian languages, displaying both scripts simultaneously.

The Avestan text corpus (hymns, prayers, ritual protocols, theological treatises) comprises about 150,000 words, being the remnant of a considerably larger corpus still extant in the 9th century B.C. These are the key texts of the Zoroastrian religion, since its expulsion from Iran in the course of the Islamic conquest, known as the Parsi community all over the world. In Iran itself there live a small number of Zoroastrian people, called Zardushtis, most of whom reimmigrated in modern times. The Websites indicated contain much information on the Parsi identity and religious life, and numerous links point to other Zoroastrian sites. Undoubtedly Zoroastrianism is one of the important religions of the world.

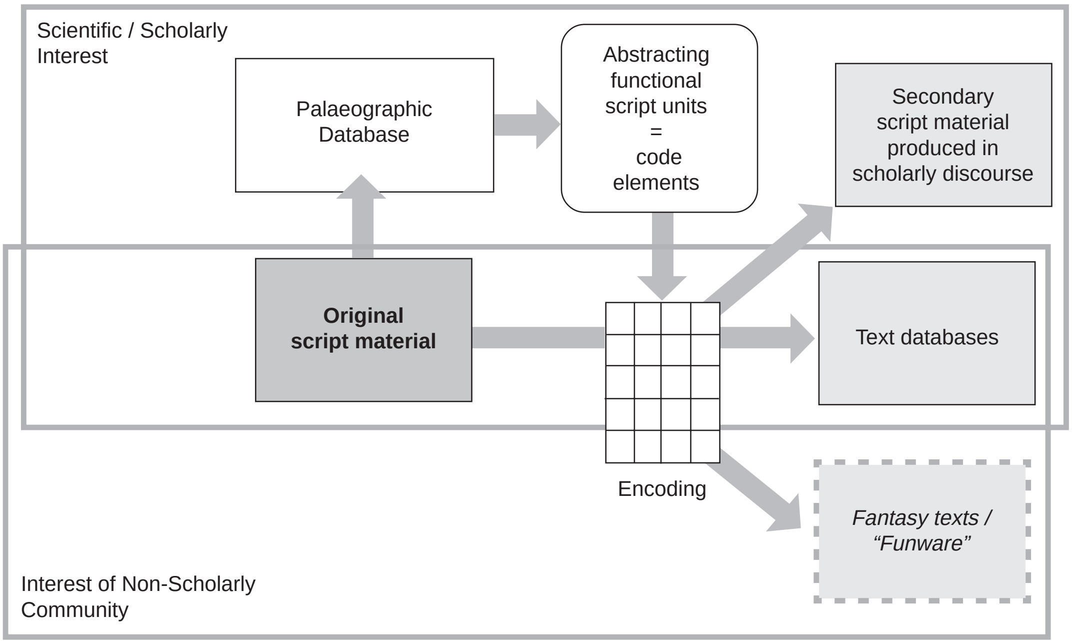
Although Avestan philology, especially the Western scholarship beginning in the 19th century, yielded important results as to the linguistic analysis and textual history, no palaeography of the Avestan script has been worked out so far. While in the view of the non-scholarly user of the script nothing impedes the abstraction of characters, the Iranianist still has to resolve difficult questions, when standard glyphs are to be defined. The example in the flow chart shows the shapes of the Avestan letter ⟨t⟩ picked from selected manuscripts. The decision which shape is more "normal" than the other, is hard to make, even if chronological arguments are taken into account. Nevertheless an Avestan encoding appears to be realizable in near future, since we are able to define abstract characters. Subsequent palaeographic investigations will produce a typology and classification of letter forms on the basis of which normalized shapes can be set up. This, however, will be a secondary and artificial device only whose scientific value remains questionable.

A distinct character block in the BMP would be justified. Iranian philology should now deliver a critical commentary on the Avestan encoding proposal from Michael Everson exposed at `http://anubis.dkuug.dk/jtc1/sc2/wg2/docs/n1684/n1684.htm` which has been drawn up without special literature having been consulted nor competent researchers in the field contacted. The scientific sources of this proposal dating from 1998 are the general handbooks Faulmann 1880 and Haarmann 1990. The author is currently working on a new version of his proposal.

The authoritative study of the Avestan script is Hoffmann/Narten 1989.

Slide 7: Category B1

Scripts of interest within cultural (and educational) policy (not communication), **encodable** from the scientific point of view


Scientific / Scholarly Interest
Palaeographic Database
Abstracting functional script units = code elements
Secondary script material produced in scholarly discourse
Original script material
Text databases
Encoding
Fantasy texts / "Funware"
Interest of Non-Scholarly Community

Historic scripts of category B 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. Scripts of category B are not used in communication, not even in a quite restricted manner as Avestan described in the previous section (Slide 4). Layman's interest in these scripts is of educational or cultural nature, often attracted by the fascination with the – alleged – occult. The category has two subdivisions according to encodability.

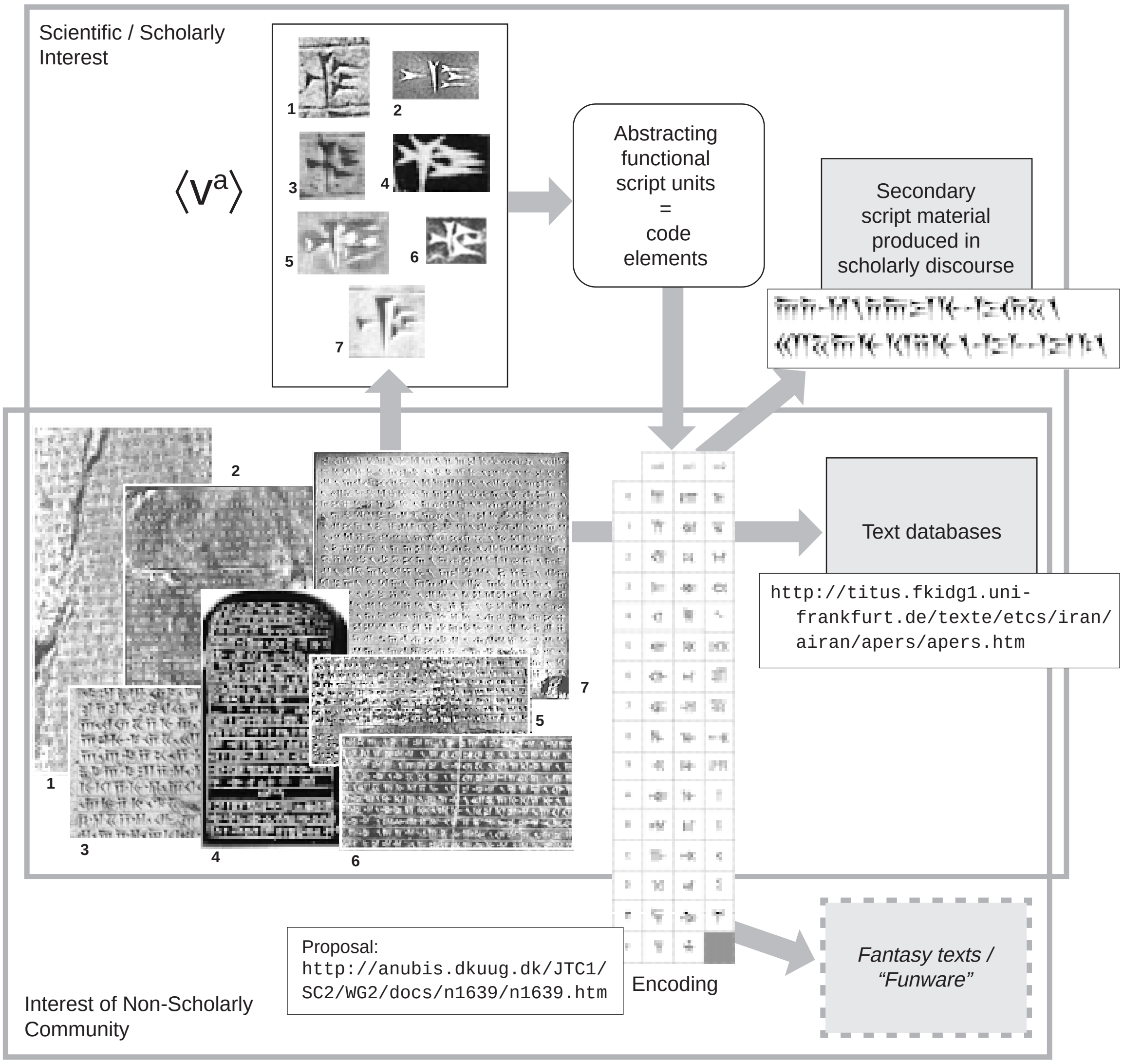
Category B1 comprises historic scripts which, with regard to the functional analysis, are ready for an abstract character encoding, but as far as the palaeographic treatment is concerned, designing normalized glyphs may be not feasible since the attested forms differ considerably both diachronically and diatopically. Often this situation occurs with scripts that are known from small text corpora only. 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.

The interest of the non-scholarly community in the scripts of category B1 varies according to the contents the preserved texts convey, but also to the aesthetic recognition the letter shapes have in certain civilizations or other social, religious etc. communities. The modern texts being produced with these scripts belong to art or fantasy.

A good example for category B1 is the Old Persian cuneiform script which I want to expound now in more detail. Another example is Old Italic which currently is under ballot in the ISO process (cf. below p. 32).

Slide 8: Category B1

Example: Old Persian


Scientific / Scholarly Interest
⟨v^{a}⟩
1
2
3
4
5
6
7
Abstracting functional script units = code elements
Secondary script material produced in scholarly discourse
1
2
3
4
5
6
7
Text databases
Encoding
Fantasy texts / "Funware"
Interest of Non-Scholarly Community

http://titus.fkidg1.uni-frankfurt.de/texte/etcs/iran/airan/apers/apers.htm

Proposal: http://anubis.dkuug.dk/JTC1/SC2/WG2/docs/n1639/n1639.htm

The inscriptions of the Achaemenid kings from Darius I (the Great, 521–486 B.C.) through Artaxerxes III (i.e. 359/8–338/7 B.C.) show a cuneiform script designed especially for the purpose of engraving these official texts in stone, mostly part of buildings, in metal, rarely in clay. It is quite natural that Old Persian cuneiform texts are not attested on papyrus or parchment. The whole corpus counts not more than 8,000 words. The language represented is Old Persian, a south-western dialect of the Old Iranian language family. There exist no other sources of Old Persian besides the cuneiform inscriptions.

The content of the inscriptions is of considerable importance for the ancient Near and Middle East. Especially the largest text, the inscription of Darius the Great on the rock of Bīsutūn (5 colums, 414 lines), contains a detailed report on the beginning of his reign.

The system of the Old Persian cuneiform script has been perfectly analysed, so abstract characters can be derived from the material. Regarding the logograms (word symbols for “king”, “land”, “god” etc.), we are not sure whether they were part of the repertoire originally, since the most ancient inscriptions do not use them. As normalized forms, one would take the Bīsutūn forms, but there are no instances of the logograms in this text.

Evidently, the importance of an Old Persian encoding for research in the field is restricted, because there is no urgent need to depict the characters which the original monuments generally exhibit in good, i.e. readable quality. Nevertheless Iranian philology and linguistics would appreciate to encode the texts in the original script, parallel with transliteration and transcription.

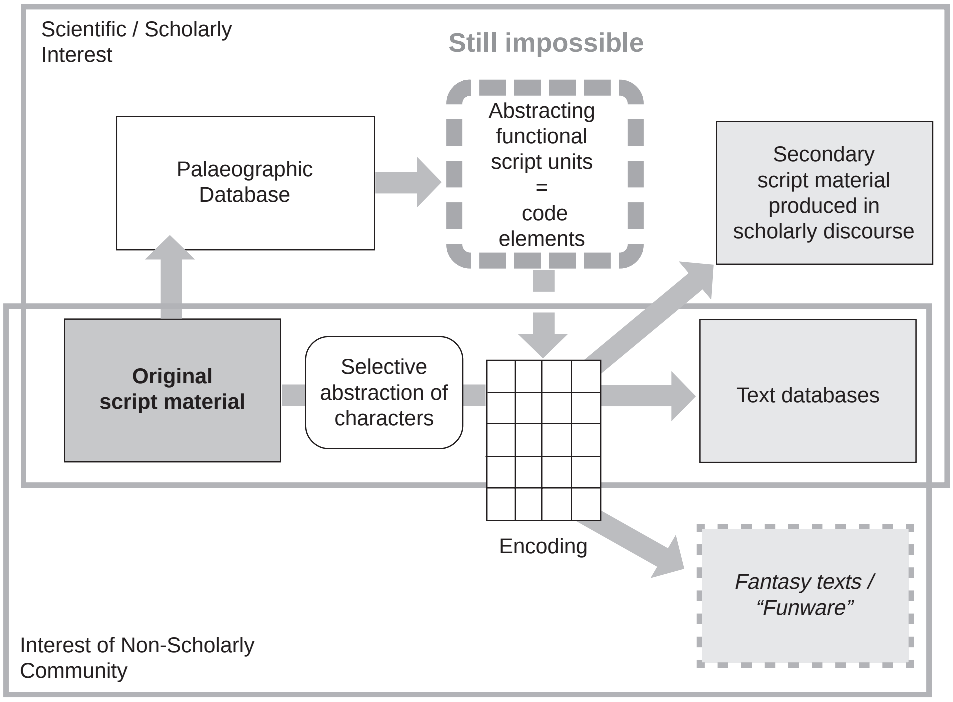
As a sample of the palaeographic situation, let us collect several instances of the character <v^{a}> from texts the creation dates of which are spread over the entire period of 150 years when the script was in use (Bīsutūn: no.1). It is evident from the first sight that there is a high degree of regularity and balance in the script, independently even of the material of the text carrier (stone, metal, clay). In electronic format, the variants are best accounted for in terms of typefaces (different widths), since the position of the graphic elements is remarkably fixed, i.e. there is no variance especially of the three horizontal wedges. This is due to the particular conditions of the script’s creation and use. Being an exclusively monumental writing, Old Persian cuneiform characters never had been crushed quickly into a piece of soft clay, in order to fix a note on sales, accountancy, etc. Like this, the opportunity never occurred for a cursive writing to be derived from the geometrically balanced wedge lengths and positions the script designer once had determined.

Therefore, by virtue of the definite function of the script in history, a normalization for electronic processing can in fact be carried out, i.e. the normalization already done by the very creation of the script can be adopted. Old Persian cuneiform fonts, then, would be styled in accordance with these normalized shapes. In the academic practice, these fonts will be used to print exercises for introductory courses, as well as descriptions of the writing system in scientific or more general reference works.

Cf. Schmitt 1989 for an up-to-date account of the Old Persian script.

Slide 9: Category B2

Scripts of interest within cultural (and educational) policy (not communication), **not yet encodable** from the scientific point of view


Scientific / Scholarly Interest
Still impossible
Palaeographic Database
Abstracting functional script units = code elements
Secondary script material produced in scholarly discourse
Original script material
Selective abstraction of characters
Text databases
Encoding
Fantasy texts / "Funware"
Interest of Non-Scholarly Community

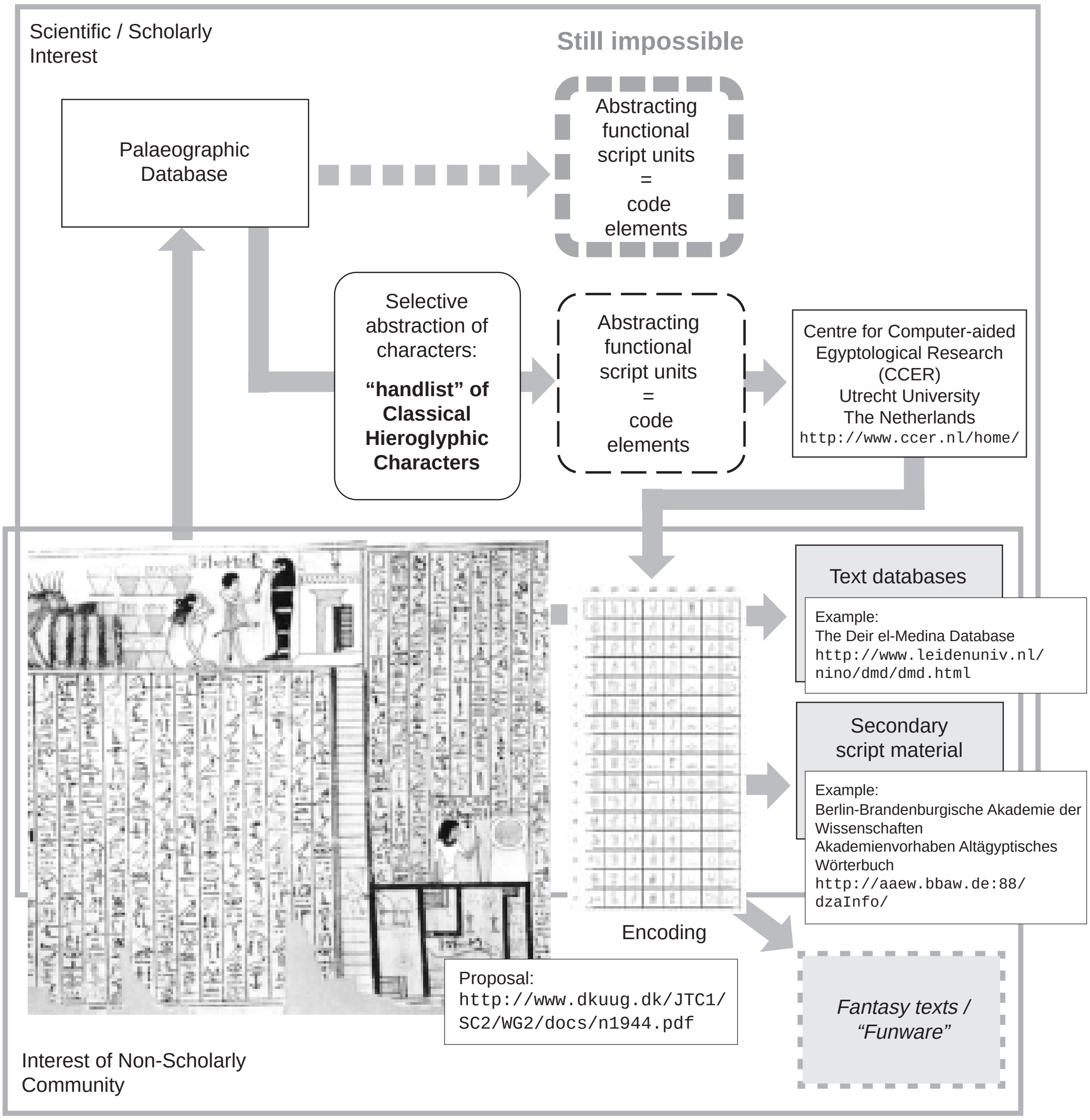
Category B2 differs from category B1 only in that an abstract character encoding cannot be drawn up at the present stage of research. Further palaeographic investigations are indispensable before abstract entities are derived from the attested script material.

The layman communities, if interested in a particular script of this category, often are not willing to wait for the researchers to give the go-ahead for an encoding. As far as the writing system in question is understood, amateurs wish to use the letters or symbols, playing with them and even creating new texts. It is up to the standardization committees to decide whether the international standard can afford an amateur encoding of theses scripts in order to satisfy such a user group. If this will happen, language science and philology will not be able to profit from it.

The situation sometimes is rather complex because computerized typeware in fact is used in the academic world. The glyphs have been derived from selected items of the repertoire, prematurely of course, palaeographic variants being depicted as well as encodable units. Enthusiastic laymen like to extend the glyph collection ad infinitum as many researchers do as well in case they need the glyphs in order to describe certain phenomena. This is to say that the border line between scientific correctness and pseudo-scientific play may be difficult to draw, especially in the view of the outside observer.

Slide 10: Category B2

Example: Egyptian Hieroglyphs



Scientific / Scholarly Interest

Still impossible

Palaeographic Database

Abstracting functional script units = code elements

Selective abstraction of characters:

“handlist” of Classical Hieroglyphic Characters

Abstracting functional script units = code elements

Centre for Computer-aided Egyptological Research (CCER)
Utrecht University
The Netherlands
`http://www.ccer.nl/home/`

Text databases

Example:
The Deir el-Medina Database
`http://www.leidenuniv.nl/nino/dmd/dmd.html`

Secondary script material

Example:
Berlin-Brandenburgische Akademie der Wissenschaften
Akademienvorhaben Altägyptisches Wörterbuch
`http://aaew.bbaw.de:88/dzaInfo/`

Encoding

Proposal:
`http://www.dkuug.dk/JTC1/SC2/WG2/docs/n1944.pdf`

Fantasy texts / “Funware”

Interest of Non-Scholarly Community

A popular example of a category B2 script are the Ancient Egyptian Hieroglyphs. Hieroglyphic writing has been practised during a period lasting from about 3000 B.C. until about 400 A.D. Since the famous decipherment in 1822 by Jean François Champollion, the modern world, researchers and educated public in general, keeps fascinated by this civilization. The decipherment marks the birth of a new discipline in Ancient Studies, Egyptology. Detailed script analysis, observing the modifications of character shapes in the course of the three millenia, forms a core matter in Egyptological studies, because only the knowledge of the script finally allows us to make the impressive architectural and artistic monuments speak. The textual heritage is enormous, its contents covering all aspects of life.

It is quite natural that the advent of the computer enabled Egyptologists to work on the texts in a far more efficient manner. This does not mean, however, that the textual information had to be depicted by graphic outlines, drawn in accordance to the template of the hieroglyphic symbols. Rendering the electronic text encoded in transliteration, had been secondary. The first step is to build up a computerized palaeographic database, where the different shapes and compositions of complex symbols are classified. This is a painstaking enterprise, carried through by several researchers simultaneously, the most advanced project being the database maintained by the Centre for Computer-aided Egyptological Research (CCER) in Utrecht, The Netherlands. Egyptologists do not yet control the entire repertoire of symbols recovered up to this day. Especially the earliest texts still are not understood perfectly.

On the other hand, the numerous educated public does not have much understanding for the hesitation the researchers practise, nor do printers who produce books in which hieroglyphs have to be represented. All the more so because lead types had been made since the middle of the 19th century in order to print manuals, grammars, dictionaries and text editions, and, following the model of these lead types, computerized fonts very soon have been designed. The most famous set of types is Gardiner's (Gardiner 1957). Why should the Egyptian hieroglyphs not be encodable, when academics as well as interested laymen type them quite easily on their PC? It is in this mood that the well-supplied proposal from ISO has been composed.

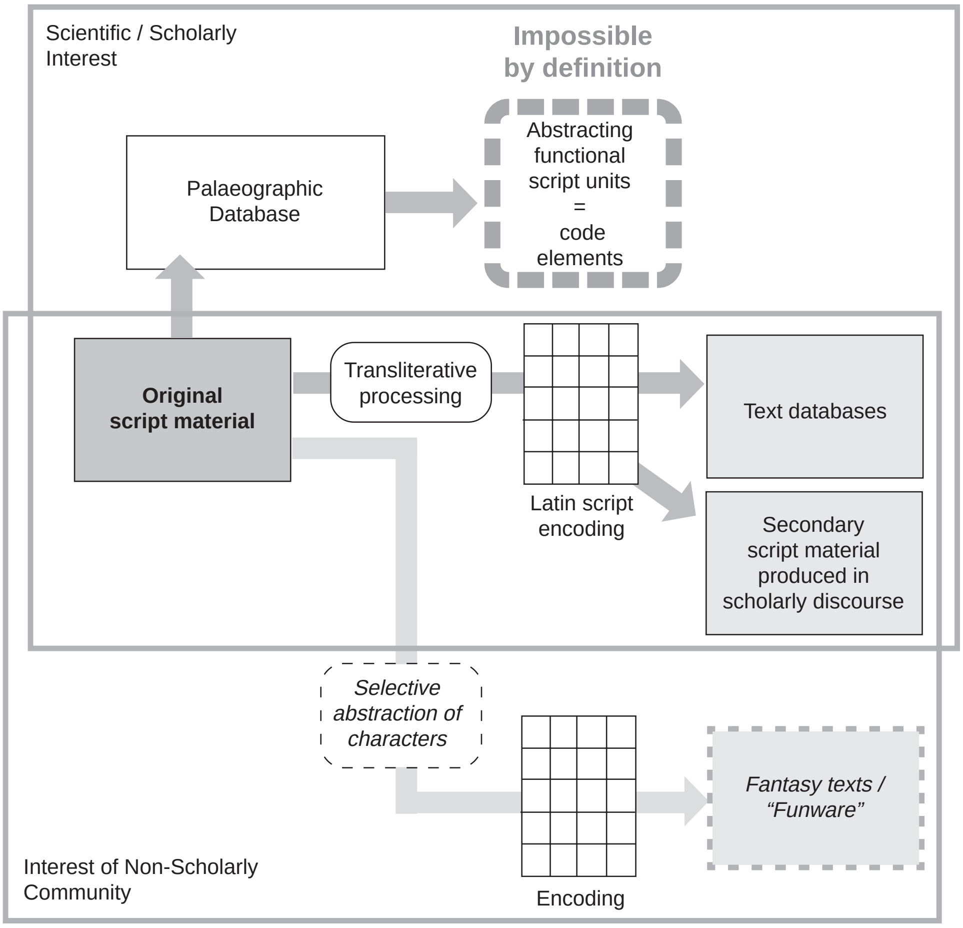
In this paper, there is not enough room for a thorough treatment of the hieroglyphic script and the problems that arise when an abstract character encoding is to be designed. In fact, on a Unicode Conference, Ancient Egyptian may claim a separate time slot for its own, since the matter is so complex. Therefore, in the following, only the most important features can be outlined.

One of the leading specialists in the field, the Egyptologist Wolfgang Schenkel, has formulated his comments on the recent proposal to encode a basic set of hieroglyphs in ISO/IEC 10646 / Unicode (Schenkel 1999). What he says there may be taken for the common opinion among Egyptologists. Both the lead types and the electronic typefaces have *not* been designed as a result of palaeographic investigations. They simply represent current shapes which are required to print a useful reader, grammar or dictionary. A new and methodically exact approach is being made at Basle University, Switzerland, where a so-called handlist of hieroglyphic characters used during the classical period is being drawn up. These characters can be defined on an abstract level, so they are encodable according to the Unicode design principles, and may constitute the core code points for an extensive encoding of hieroglyphs in the future. Before any further attempts of encoding are envisaged, the completion of this work must be waited for – otherwise the coded set would be nothing more than a printer's inventory, but not a researcher's tool.

On the whole, at the present stage of investigation, Egyptology does not feel ready at all for an abstract character encoding. Researchers like to see the repertoire open for additions, since new texts and new character shapes are constantly being discovered, questioning at every instance definitions of abstract characters made previously.

Slide 11: Category C1

Scripts of scholarly interest only, **not encodable for systematic reasons**

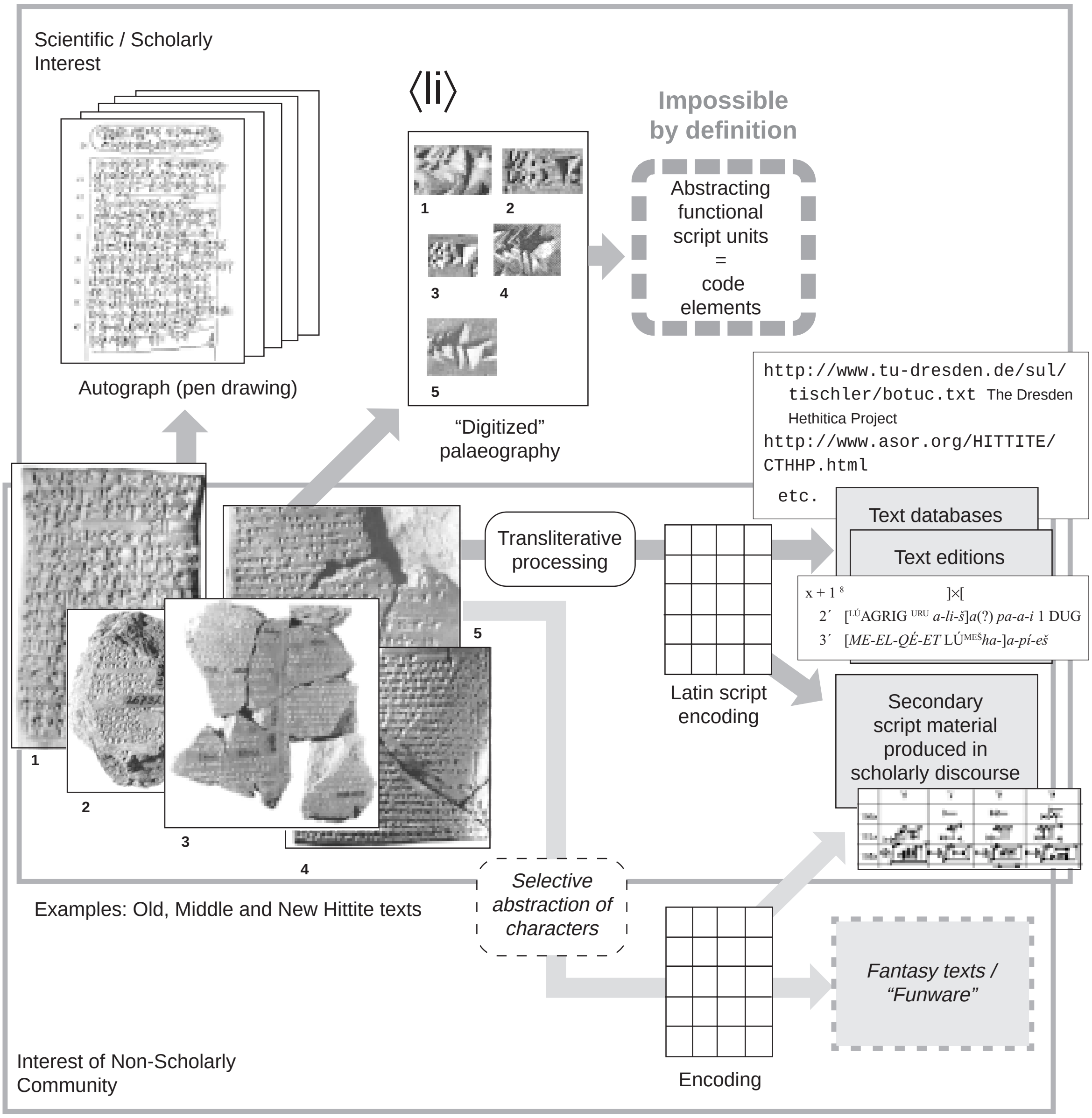

Scientific / Scholarly Interest
Impossible by definition
Palaeographic Database
Abstracting functional script units = code elements
Original script material
Transliterative processing
Text databases
Latin script encoding
Secondary script material produced in scholarly discourse
Selective abstraction of characters
Fantasy texts / "Funware"
Interest of Non-Scholarly Community
Encoding

Historic scripts which are dealt with almost exclusively in the academic world, are grouped here in category C. The interest of the non-scholarly community may exist, and in some cases even is very strong, 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 – except for manuals designed for beginners. All scripts of this category are unencodable by definition, i.e. any attempt to derive abstract characters will necessarily fail, since the repertoires do not support it. To anyone not involved in the scientific research, this might seem an arrogant and uncompromising attitude of the experts. The standardizers, then, are to decide whether the need of the non-scholary community is as urgent as to include an encoding in the international standard that does not meet the requirements of those working scientifically on the scripts and the texts in question. For a researcher's statement on the electronic processing of these scripts, see the comment of a very experienced scholar in the field, Wolfgang Röllig (Röllig 1999); the response is Everson 1999b.

The non-encodability of certain scripts is the result of their internal structure as well as of the mode of their use, although a considerable amount of text is available: this is subcategory C1. There is one very important example, namely the major cuneiform writing.

Slide 12: Category C1: Encoding Cuneiform Scripts

Example: Major Cuneiform Script of Ancient Near East


Scientific / Scholarly Interest
⟨li⟩
Impossible by definition
Abstracting functional script units = code elements
1
2
3
4
5
Autograph (pen drawing)
“Digitized” palaeography
Transliterative processing
Latin script encoding
Text databases
Text editions
Secondary script material produced in scholarly discourse
1
2
3
4
5
Examples: Old, Middle and New Hittite texts
Selective abstraction of characters
Encoding
Fantasy texts / “Funware”
Interest of Non-Scholarly Community

`http://www.tu-dresden.de/sul/tischler/botuc.txt` The Dresden Hethitica Project
`http://www.asor.org/HITTITE/CTHHP.html`
etc.

x + 1 ^{8}]×[
2′ [$^{\text{LÚ}}$AGRIG $^{\text{URU}}$ *a-li-š*]*a*(?) *pa-a-i* 1 DUG
3′ [*ME-EL-QÉ-ET* LÚ$^{\text{MEŠ}}$*ha-*]*a-pí-eš*

The cuneiform script of Ancient Near East (Mesopotamia, Armenia, Syria, Anatolia) has an unbroken tradition of 3000 years. The earliest documents date from about 3000 B.C., showing pictographic symbols which afterwards had been graphically abstracted by a prismatic pencil in the typical ‘cuneiform’ way producing wedges, i.e. straight strokes positioned horizontally, vertically, or diagonally, and ‘hooks’, as the minimal graphic elements of the script. Cuneiform writing had been adopted by many language communities in the region. The language first written in cuneiform characters is Sumerian: it was the syllabic, morphological and lexical material of this language the writing system had been built upon. By a further genial step, pictographic symbols came to be used as syllabic, i.e. phonetic, symbols to represent either just an isolated, abstract syllable, apt to form part of a word, or a different lexical item by virtue of the phonetic correspondence. Moreover, pictograms could be accompanied by such syllabic values as complements representing inflexion, derivation etc. (the ideogram "to go", e.g., thus could be converted to "I go", "I went", etc.). From 2500 B.C. onwards, Akkadian, a Semitic language, came to use cuneiform writing (dialects: Babylonian, about 2000 B.C. – 150 A.D., Assyrian, about 1800 – 625 B.C.), and later on Hittite (about 1600 – 1200 B.C.), so exhibiting the most ancient written documents in an Indo-European language. Both the Akkadian and Hittite writing systems integrate elements of the previous system as heterographs, being interpreted phonetically as word etc. of the language represented. As a result of this, the Hittite system is particularly complex, incorporating Sumerian and Akkadian heterograms.

According to the Unicode design principles, one would establish a unified encoding for cuneiform. But the attempt to draw up an abstract encoding even for one single language dependent writing system is useless, because in the course of its long history no standardization has ever been made. What has come down to us from the extensive text production of the Ancient Near East, are exclusively manuscripts in the very sense of hand-writings, showing up features of date, writing school, office, but also the particular features of the scribe's personal manner of handling the pencil. Deriving standard shapes from more than a sixscore of ductus of different scriptoria as well as of individual and often abbreviated graphic shapes, would mean to introduce something intrinsically alien to cuneiform writing. In order to confine the area, let me examine five instances of one syllabic character in five Hittite texts, dating from the 16th (1), 14th (2 and 3) and 13th (4 and 5) century B.C. There is a sharp contrast to the situation with Old Persian cuneiform, seen previously on Slide 7. In our case the range of variance is considerable.

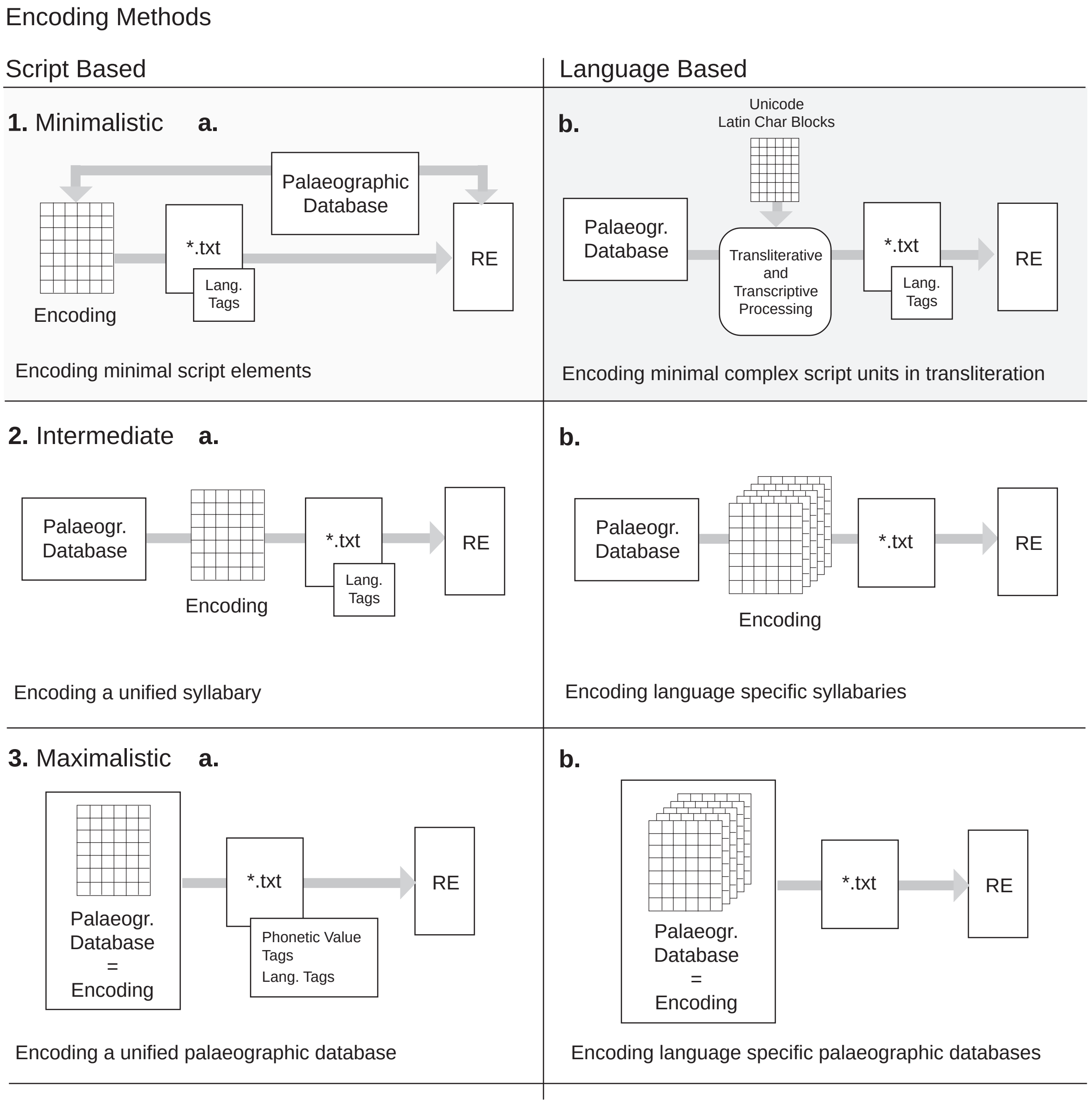
Therefore, cuneiform philology relies on transliteration of the texts. This is the only reasonable method for dealing with cuneiform texts, since the transliteration contains, by virtue of clearly defined formatting conventions, as much script related information as needed for a linguistic analysis of the text. Closer investigation of the written form is the realm of palaeography proper, which, traditionally, uses so-called autographic editions of the texts, i.e. copied drawings by pen and ink, and, recently, digital images of glyphs cut off from digitized or digital photographs of the clay tablets. Computer-aided palaeography certainly will evolve in the next years.

In the beginning of the 20th century, lead types had been created for editing cuneiform texts, but, after a short while, this editing technique had been abandoned. The capacity of electronic data processing, once again, has tempted laymen but also researchers to construct devices to generate cuneiform symbols in print and on the screen. As useful as such very elaborate fonts can be for printing introductory material for students, they contribute, strictly speaking, fairly little if nothing (but illusion) to the palaeographic analysis.

Slide 13 gives a synopsis of the encoding methods for the major cuneiform script, that are possible theoretically. In principle, one has to distinguish between script based and language based encoding methods. ISO/IEC 10646 / Unicode being a script based encoding, a conformant treatment of language dependent cuneiform script variants would be a unified character block

Slide 13: Category C1: Encoding Cuneiform Scripts

Example: Major Cuneiform Script of Ancient Near East



Encoding Methods

Script Based

Language Based

1. Minimalistic **a.**

Palaeographic Database

Encoding

*.txt

Lang. Tags

RE

Encoding minimal script elements

b.

Unicode Latin Char Blocks

Palaeogr. Database

Transliterative and Transcriptive Processing

*.txt

Lang. Tags

RE

Encoding minimal complex script units in transliteration

2. Intermediate **a.**

Palaeogr. Database

Encoding

*.txt

Lang. Tags

RE

Encoding a unified syllabary

b.

Palaeogr. Database

Encoding

*.txt

RE

Encoding language specific syllabaries

3. Maximalistic **a.**

Palaeogr. Database = Encoding

*.txt

Phonetic Value Tags

Lang. Tags

RE

Encoding a unified palaeographic database

b.

Palaeogr. Database = Encoding

*.txt

RE

Encoding language specific palaeographic databases

RE = Rendering Engine

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 (1a). As part of the RE, look-up tables are derived from the palaeographic database so that the character composition is working according to templates (generating certain presentation styles, i.e. cuneiform “typefaces”). 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 (1b). An intermediate approach would consist in the encoding of a unified syllabary, i.e. of precomposed syllabic characters (2a), or separate encodings for different syllabaries would to have to be defined in case the encoding is expected to convey language information as well (2b). A maximalistic encoding method would project the palaeographic database as such, either unified (3a) or language specific (3b), onto the encoding space.

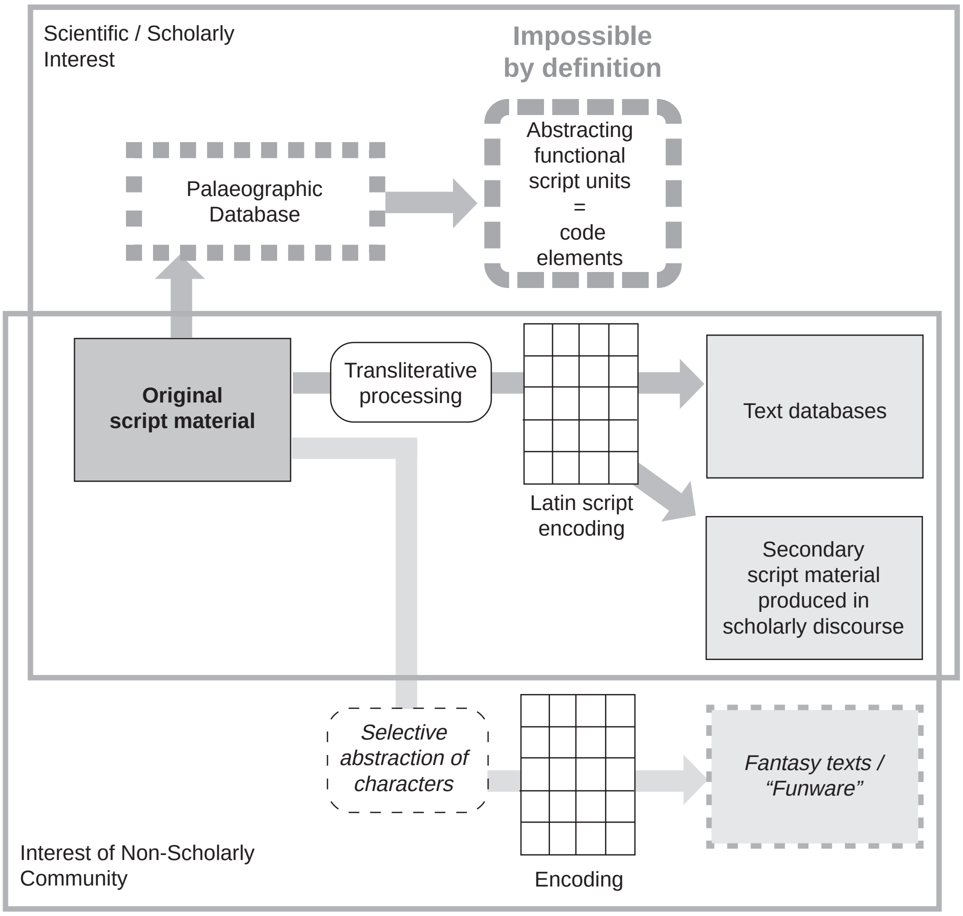
As demonstrated above, method 1b is the only scientifically practicable one. For the purpose of experiments or demonstration in a didactic context, method 1a may be of some use, but a highly sophisticated rendering engine would be required in order to generate the attested variants. Would it be worth while miming the individual scribes' hands? From the scientific point of view, certainly not. Method 2 is implicitly applied when computerized fonts are created nowadays: They are based on a selection of characters *and* glyphs, since the status of a character is not and cannot be proved by the evidence available. Finally, encoding the database itself, is quite absurd. An enormous encoding space would be required, and it would be an immense task to build up a Rendering Engine that processes data nearly as complex as digitized images scanned directly from the original text carrier.

Cuneiform philology, therefore, will continue to use transliteration alongside with digitized or digital photographs of the clay tablets, preferably 3D photographs, so that characters carved into the edges can be read by simply rotating the virtual object on the screen. The pen-and-ink drawing will not lose its importance either, as a means to become aware of the positioning of the wedge elements. Of course, the statement that the major cuneiform script cannot be included in any character encoding standard, is a verdict for all those who are devoting their energy to cuneiform fonts. Perhaps one day Unicode will accept an encoding proposal put forward by amateurs. Then this would be a concession to the non-scholarly user community, which scholars will not criticize as such, since it is not in their interest.

Cf. again, for a statement of a specialist of the cuneiform script, Röllig 1999, and, for the detailed response from the standardizer's viewpoint, Everson 1999a.

Slide 14: Category C2

Scripts of scholarly interest only. **Encoding not reasonable: attested material not yet explored / explorable**


Scientific / Scholarly Interest
Impossible by definition
Palaeographic Database
Abstracting functional script units = code elements
Original script material
Transliterative processing
Text databases
Latin script encoding
Secondary script material produced in scholarly discourse
Selective abstraction of characters
Fantasy texts / "Funware"
Interest of Non-Scholarly Community
Encoding

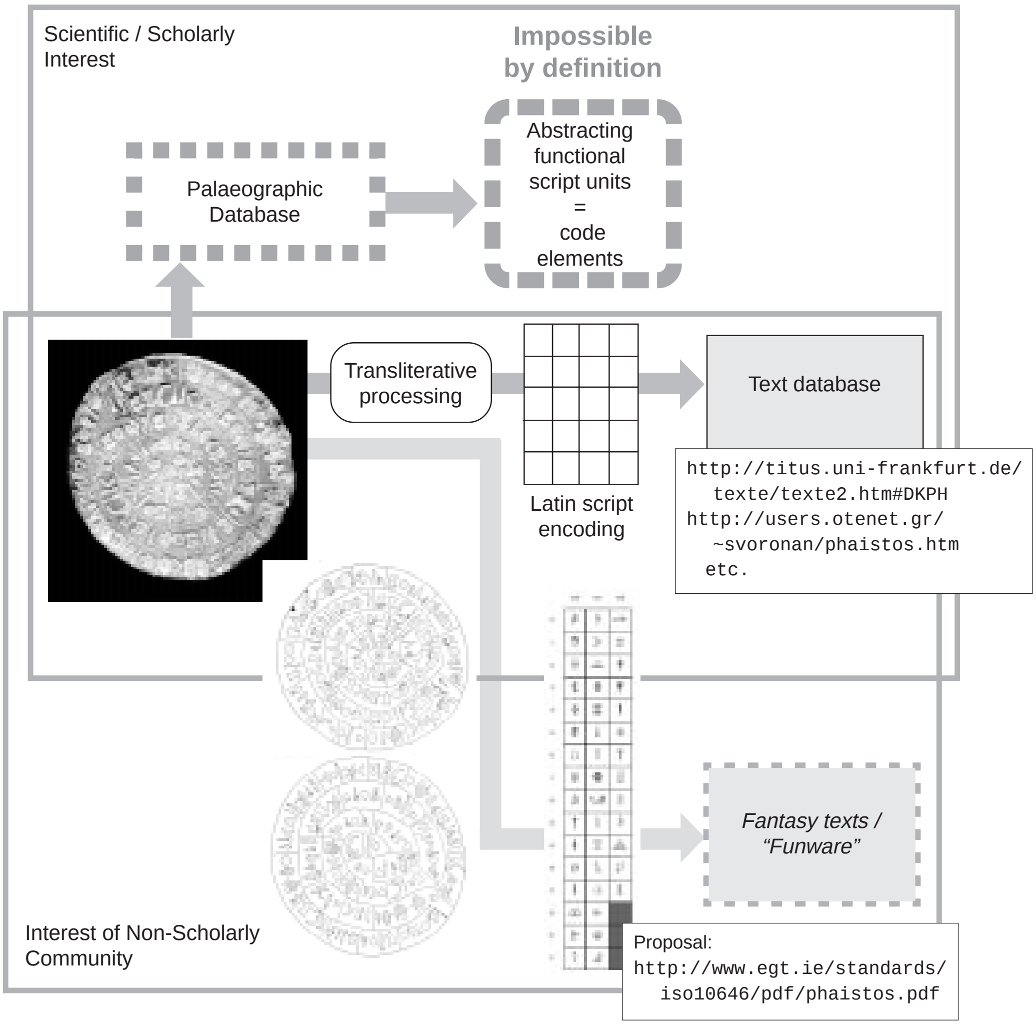
Finally, we differentiate category C2 comprising those scripts that not only by virtue of the document type and the state of analysis, but also of the quantity of the preserved text material cannot be subject to standardization and abstract character encoding. Cf. again Röllig 1999.

Naturally, all yet undeciphered scripts fall into this category. To store script units that are identified provisonally by numeric values only, in an encoding, certainly is a means of handling the data efficiently, but has nothing with normalization. An encoding of this kind may be considered as an intermediate, i.e. operative working base. The question is whether an international encoding standard can serve as clipboard for repertoires that are not yet analysed, or the analysis of which still is much disputed in research.

Another group of scripts belongs here, too: There are writing systems preserved by an extremely small corpus of texts, e.g. a dozen inscriptions consisting of less than one sentence each. Methodically, encoding such scripts would require to project the palaeographic database onto code points. But this, again, is operative only, since any normalization is excluded.

Slide 15: Category C2

Example: Phaistos Disk Script


Scientific / Scholarly Interest
Impossible by definition
Palaeographic Database
Abstracting functional script units = code elements
Transliterative processing
Latin script encoding
Text database
Fantasy texts / "Funware"
Interest of Non-Scholarly Community

http://titus.uni-frankfurt.de/texte/texte2.htm#DKPH
http://users.otenet.gr/~svoronan/phaistos.htm
etc.

Proposal:
http://www.egt.ie/standards/iso10646/pdf/phaistos.pdf

In order to give a particularly telling example of a category C2 script, let us have a look on the Phaistos disk script.

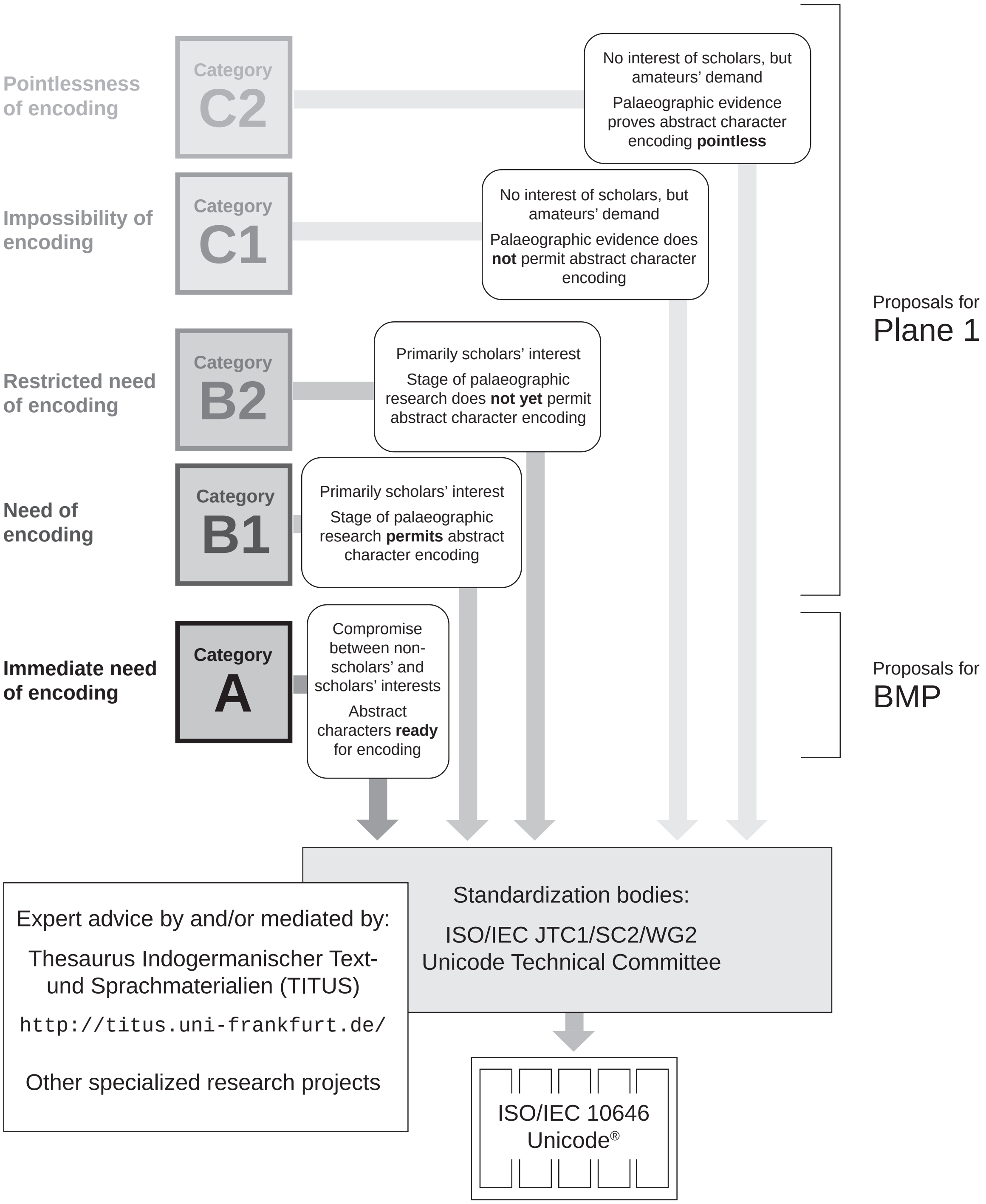
The inscription on this disk of clay, unearthed in the Minoan palace of Phaistos in Crete, is still undeciphered. According to the archaeological evidence, the object dates from the period of 1700–1550 B.C. On both sides of the disk symbols are stamped into the clay, apparently by a kind of dies for each of the symbols, used like lead types in printing. The sequence of the stamped characters runs spirally – the script direction cannot be determined with certainty. Groups of up to five symbols are separated by vertical strokes which must be interpreted as word dividers. The inscription on the disk is an isolated monument in that this type of symbols or hieroglyphs are found nowhere else in the repertoires of the other Aegean scripts of the 2nd millenium B.C.

The inventory counts 45 different items. Naturally it is not probable that this relatively short text exposes all or at least the majority of the characters of the script. Rather, we have to expect dealing with a small random selection of units taken from a more extensive repertoire.

On the Phaistos Disk, there exists a huge bibliography. Again and again the final decipherment is announced, but what has been put forward is nothing more than yet another attempt to apply a certain theory to the inscription and to harmonize this theory with what can be observed on the disk.

The character repertoire of this singular text is fixed in the sense that up to this day no other document showing this script has been recovered. Should an international encoding standard contain this repertoire, like exposing it as a script material not yet interpreted? Evidently, an encoding of the Phaistos Disk symbols as proposed by ISO would not hurt anyone. Whether to include it or not, is rather a question of principle.

Slide 16: Conclusion



Pointlessness of encoding — Category C2 — No interest of scholars, but amateurs' demand. Palaeographic evidence proves abstract character encoding **pointless**

Impossibility of encoding — Category C1 — No interest of scholars, but amateurs' demand. Palaeographic evidence does **not** permit abstract character encoding

Restricted need of encoding — Category B2 — Primarily scholars' interest. Stage of palaeographic research does **not yet** permit abstract character encoding

Need of encoding — Category B1 — Primarily scholars' interest. Stage of palaeographic research **permits** abstract character encoding

Immediate need of encoding — Category A — Compromise between non-scholars' and scholars' interests. Abstract characters **ready** for encoding

Proposals for Plane 1

Proposals for BMP

Standardization bodies:

ISO/IEC JTC1/SC2/WG2
Unicode Technical Committee

ISO/IEC 10646
Unicode®

Expert advice by and/or mediated by:

Thesaurus Indogermanischer Text- und Sprachmaterialien (TITUS)

`http://titus.uni-frankfurt.de/`

Other specialized research projects

Category C2	Linear A Eteocyprian Carian Old Phrygian Lycian Lydian	Raetic Lepontic Messapian Siculan Venetic	Gaul Iberian North-Eastern Iberian Southern Espanca Alphabet Tartessian Alphabet
Category C1	Major Cuneiform		
Category B2	Egyptian Hieroglyphs Mayan	Linear B Brahmic (about 300 historic types at least)	
Category B1	Old Persian Cuneiform Middle Persian Epigraphic Parthian (Epigraphic) Sogdian Manichaean Uyghur	Old Turkic Runes (Orkhon Script) Old Permic Old Hungarian Asokan Brahmi (Epigr.)	Cyprian Syllabary Ugaritic Cuneiform Old Italic: Etruscan, Oscan, Umbrian (currently under ballot) Gothic (currently under ballot)
Category A	Avestan Middle Persian Books (Pahlavi)	Coptic (disunified) Old Church Slavonic (disunified) Glagolitic	Mandaean

The categories of scripts made up in this paper should not be understood as clear-cut classes of repertoires that show a set of specific features each. Rather, they are classes sharing certain features that qualify or disqualify them for an abstract character encoding. The categories are exclusively operative and should, on the level of the standardization debate, facilitate the task to determine a schedule as to which historic script may be considered next.

It has been demonstrated that in many instances there are no evident solutions, and that it is worth consulting the researchers who are dealing with the scripts in question every day. In case of an insurmountable problem, the interest of the non-scholarly community may justify a normalization outside scientific exactitude.

Consequently, standardization efforts should start with category A scripts, calling for participation the relevant user groups existing in the modern world as well as the specialists undertaking scientific investigations on the written tradition and the writing system itself. With category A scripts, the need of inclusion in the international standard is immediate because otherwise social and especially religious communities are prevented from expressing and transmitting their cultural and spiritual message on an international platform. Technically, this platform is the BMP of ISO/IEC 10646 / Unicode.

Next, category B1 scripts may be considered in view of standardization. These repertoires are ready for encoding, the character blocks to be allocated in Plane 1. Then, the resulting standards can be applied immediately to scientific work, especially to data collections that should be available on the Web, e.g. for encoding text databases, dictionaries etc. On the initiative of Michael Everson, ISO recently lauched an effort the encode Etruscan and Gothic which meanwhile have reached Committee Draft status (CD ISO/IEC 10646-2 = SC2 Document No. 3393). The contribution of specialists in the field broadened the approach to the Italic branch of the Phoenician alphabet so that a unified Old Italic encoding is considered to suit best the needs.

The research activities concerning category B2 scripts are worth being following by standardizers, since positive results are to be expected in the near future. Therefore, if laymen currently are pushing for an encoding of a script of this category, the standardization bodies are well advised to stop these efforts until the scholarly community puts forward its proposal.

Designing abstract character encodings for category C1 and C2 scripts is impossible on the basis of scientific methods. I hope to have expounded here the reasons why scientific research cannot profit from normalizations which are contrary to the nature of the script material preserved.

The Unicode Technical Committee and Working Group 2 of ISO/IEC JTC1/SC2 are invited to contact the TITUS project in order to get expert advice. This project is maintaining the largest text database of ancient Indo-European languages as well as the most comprehensive Website on Indo-European and neighbouring languages, and is able to forward questions to competent scholars in case the problem is outside the range of the project’s own activities. At any rate, in cooperation with highly specialized experts, researchers in Historical and Comparative Linguistics are good consultants of the standardizer, since by the very nature of their interest they are all the way analysing writing systems in contrast with each other. They are ready to help the Unicode designers to enhance the universality of the standard.

References

Bunz 1997 — Carl-Martin Bunz, Browsing the Memory of the World, in: 11th International Unicode Conference, San Jose, California, 2.-5.9.1997, Proceedings, 1, A 7.

Bunz 1998 — Carl-Martin Bunz, Unicode® and Historical Linguistics, in: Studia Iranica, Mesopotamica et Anatolica 3, 1998, 41-65.

Bunz /Gippert 1997 — Carl-Martin Bunz / Jost Gippert, Unicode, Ancient Languages and the WWW, in: 10th International Unicode Conference, Mainz, 10.-12.3.1997, Proceedings, 2, C 11 (cf. `http://titus.uni-frankfurt.de/personalia/jg/unicode/unicode.htm`).

ISO/IEC CD 10646-2 — Information technology – Information technology – Universal Multiple-Octet Coded Character Set (UCS) – Part 2: Secondary Multilingual Plane for scripts and symbols; Supplementary Plane for CJK Ideographs; Special Purpose Plane. ISO/JTC1/SC2 N 3393 (cf. `http://www.dkuug.dk/jtc1/sc2/open/dr.htm`).

Everson 1999a — Michael Everson, Response to comments on encoding Old Semitic scripts (N2097). Document No. ISO/IEC JTC 1/SC 2 N 2132 (cf. `http://www.dkuug.dk/jtc1/sc2/wg2/docs/n2133.htm`).

Everson 1999b — Michael Everson, Response to comments on encoding Egyptian hieroglyphs (N2096). Document No. ISO/IEC JTC 1/SC 2 N 2132 (cf. `http://www.dkuug.dk/jtc1/sc2/wg2/docs/n2132.htm`).

Faulmann 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. 1880 (reprints: Nördlingen, Greno 1985; Frankfurt/M., Eichborn, 1990).

Gardiner 1957 — A.H. Gardiner, Egyptian Grammar Being an Introduction to the Study of Hieroglyhs. Oxford, third edition 1957 (first edition 1927).

Haarmann 1990 — Harald Haarmann, Universalgeschichte der Schrift. Frankfurt/M./New York 1990.

Hoffmann/Narten 1989 — Karl Hoffmann/Johanna Narten, Der Sasanidische Archetypus. Untersuchungen zur Schreibung und Lautgestalt des Avestischen. Wiesbaden 1989.

Röllig 1999 — Wolfgang Röllig, Comments on proposals for the Universal Multiple-Octed 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/n2197.htm`).

Schenkel 1999 — Wolfgang Schenkel, 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/n2196.htm`).

Schmitt 1989 — Compendium Linguarum Iranicarum, hrsg.v. Rüdiger Schmitt, Wiesbaden 1989 (pp. 56-85 “Altpersisch” von Rüdiger Schmitt).