Hieroglyphic text corpus

Towards standardization

Vincent Euverte & Christian Roy

Paris

1. Introduction

The I&E meeting in Liège was the eighteenth of a long series. Looking back to these 26 years since the first meeting of this group in the College de France in Paris (1984), three great periods can be recognised:

(1) Two events actually happened 15 years earlier, in 1969, with the appearance of Arpanet (the ancestor of the Internet) and of the first Glyph program by Jan Buurman on a mainframe in Algol language. In the late seventies, information technology has been moving from the domain of industry to that of personal computing. So Jan Buurman migrated his program to Fortran, in order to be transferrable to PCs. The I&E Computer group then contributed at that time by launching the Manuel de Codage, the 3rd edition of which was presented in Cairo at the 1988 Congress of the International Association of Egyptologists. At the same time, the first electronic dictionary of Ancient Egyptian was launched with the *Wörterbuch*.

(2) A second period led this group to Bordeaux, for its 10th anniversary in 1994. Among the major achievements, we will cite the creation of the Multilingual Egyptological Thesaurus (MET, see §2) which was supplemented by a list of 14 "minimum requirements" named "passeport".[1] The publication of the "Beinlich wordlist"[2] was also a significant outcome of this meeting. In his retrospective of these ten years, Dirk van der Plas also raised the idea of a fourth edition of the Manuel du Codage, which was never realised. As technology was progressing quite fast and Internet usage was rocketing up to 10 million users, several hieroglyphic text processors appeared in parallel. We cannot cite them all, but the best known are WinGlyph and MacScribe, or Hierotext migrating to TKSesh (see Gozolli's overview in the present volume).

(3) The third major period is less clear to us as we have not been able to retrieve all the proceedings from these 15 years. Major outcomes seem to be the very long debate on including hieroglyphs in Unicode, finally endorsed in the Fall 2009, thanks to the persistence of Michael Everson and Bob Richmond. During this period, several online dictionaries also appeared, and large developments in computer technology encouraged the appearance of new software, as JSesh for example. Various communication tools, such as AEL, EEF, ThotScribe fora and Internet letters, as well as the multiplicity of databases on mastabas, shawabtis, pyramid texts, etc. (including the *Ramsès* Project) can be added to figure 1.

1 In: *Actes des Rencontres "Informatique et Égyptologie" 1993*, Informatique et Égyptologie 9, 1994, p.4

2 Beinlich, Horst & Friedhelm Hoffmann. 1994. Ägyptische Wortliste, in: *Göttinger Miszellen* 140, p. 101-103.

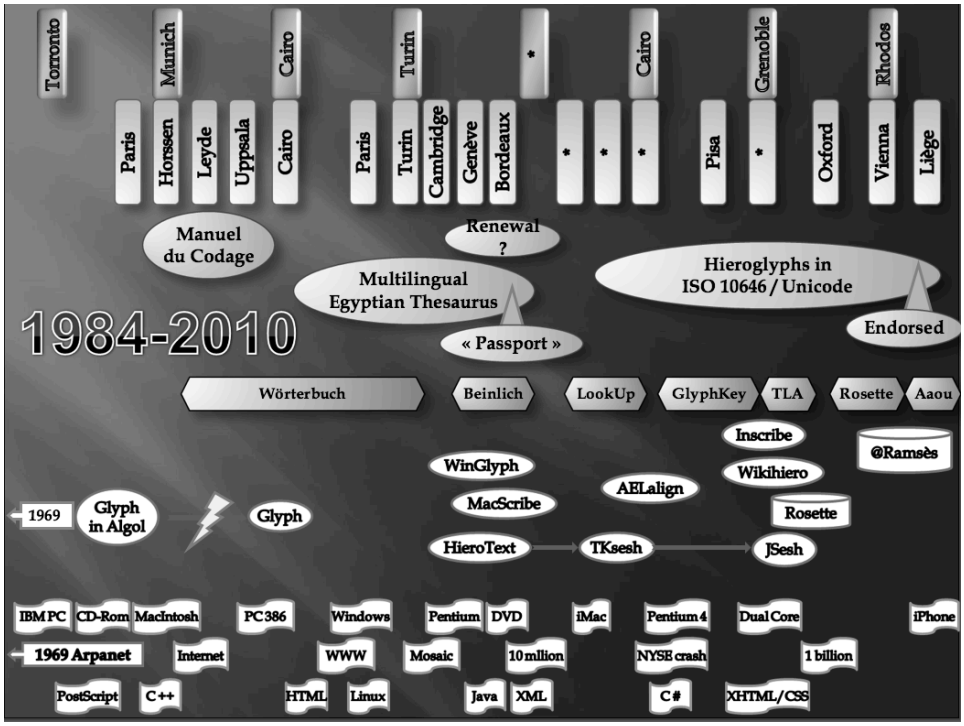

Figure 1. Informatique & Égyptologie timeline

A quick look at figure 1 makes it obvious that sharing the heritage of Ancient Egyptian written production is facing numerous technical challenges in a constantly evolving environment. We will focus in this paper on two specific topics related to the standardization of Ancient Egyptian text corpora: the 'Multilingual Egyptological Thesaurus' (MET) and the 'Manuel de Codage' (MdC).[3]

2. THE 'MULTILINGUAL EGYPTOLOGICAL THESAURUS' (MET)

2.1. *What is an ancient Egyptian object?*

In our field, the "Objects" of study are artefacts with many facets seen from epigraphic and philological viewpoints. To characterize such objects, it is essential to describe their contexts and environments: Where were they discovered? Where are they now? Who found them? What are they made of? To which period do they belong? In addition, things like pictures of the artefact and associated bibliography are necessary to enrich the description.[4]

An excellent base for communality of the discipline was defined in the nineties with the Multilingual Egyptian Thesaurus[5] (MET).

According to the MET definition, an 'object' is any Museum artefact. In a broad sense, it includes objects such as the White Chapel of Senousret in the open air museum at Karnak. It also encompasses any papyrus from any period, even if in a private collection. But what about the objects which are still sitting on their original site, such as an Amarna Boundary Stela, the Famine Stela on Sehel island, or (a bit more complicated) the Qasr Ibrim stela relocated to the new site of Kalabsha? We may also want to characterize similarly the context of reliefs and paintings in temples and tombs or even petroglyphs

3 We would like to highlight that both projects originated many years ago from groups represented at this conference, and both are still in use in various capacities, regardless of weaknesses, obsolescence, or hiatus.

4 More specifically for a text corpus, the content includes several categories specific to the hieroglyphic writing: the characters must be identified, their graphic representation may vary according to the period, the support, the type of text, and the layout may even include scenes which are an integral part of the meaning. These are all aspects addressed by the Manuel de Codage (see section 3 below).

5 http://www.ccer.nl/apps/thesaurus/index.html

located everywhere in the Egyptian deserts: boats and serekhs in Wadi el Shott, ostriches in Aswan quarry, or quarryman's marks in Gebel Silsileh.

2.2. *MET usage: Who?*

Who are the current and potential users of the MET? Of course museums at first, as the MET was created for them; but from previous examples, one can imagine dozens of other applications of this method of cataloguing.

Who may contribute, propose amendments or new data? To our mind, anybody interested in Egyptology has this potential. However these contributors are not necessarily legitimately authorised to endorse and to publish. So we see three major steps in the Thesaurus management:

- *Collection*. As suggested by Reem Baghat, responsible for the MET at CULTNAT,[6] a tool could be implemented, for instance on the GEM (Global Egyptian Museum) website,[7] to allow identified/registered users to record their information/proposals.
- *Validation*. A committee of professionals is definitely required, and it must be international to ensure proper perspectives and translations in each of the agreed languages.
- *Distribution*. The most up-to-date approved version of the Thesaurus could be made publicly available in an exchangeable format (e.g. PDF). The GEM website should then probably be the most appropriate medium.

2.3. *MET completion*

As we saw with the timeline (see §1), no official update has been made since the 1995 publication, which raises several issues:

- The "Provenance" characteristic is missing quite a number of locations, either neglected initially such as Lower Nubian sites now under the Nasser Lake or recently excavated as Tell-Herr in Sinai.
- For the "Current Location" field, several new museums have opened in the past 15 years, such as the Imhotep site museum in Saqqara. Some others have been forgotten in the initial list such as the small Tessé museum in Le Mans (France).
- The Thesaurus details precisely the different types of support material; but how to indicate a David Robert's painting or a 19th century facsimile describing an object no longer available to us, because it has been eroded, robbed or destroyed?
- The philologists could make suggestions in order to expand the current "Language" and "Writing" characteristics of the Thesaurus, for instance to allow the description of the "state of the language" depending on the period and the type of text.
- Last but not least, among the 7 languages already defined, some translations are either incomplete or inconsistent, in particular in Portuguese.

2.4. *MET expansion*

There may also be a lot of enrichment to the current 15 dimensions of the Thesaurus:

- For the sake of clarity, the existing characteristics could be enriched, for instance with dating criteria and with the Ancient Egyptian and Greek names added to the Arabic names (as much as we know them).

6 Center for Documentation of Cultural & Natural Heritage; see http://www.cultnat.org/

7 http://www.globalegyptianmuseum.org/

- The Global Egyptian Museum made a tremendous effort to expand the original MET with interesting characteristics, such as Colour, Culture, Titles, Dimensions, etc. (some of them been integrated in the "Passeport" definition of 1993; see n. 2 above).
- The Global Egyptian Museum also launched the translation of the Thesaurus in Arabic. One can even envisage adding other languages, like Japanese and Chinese, as these countries are becoming more and more involved in archaeology of Egypt.
- Regarding the bibliographic references, the I&E group is already supporting the AEB/OEB as the most international standard.

2.5. *MET Revival Project: Who and When?*

- Since ownership of the MET was transferred from the CCER to CULTNAT, it is clearly the new leader and also has the authority and resources for the web deployment of a MET revival project using the Global Egyptian Museum Internet facilities.
- The I&E Computer Group could serve as a facilitator to identify professionals in each country to complete/validate/translate. A clear message from us, eventually supported by the IAE and its President James P. Allen, could convince CULTNAT to go ahead.
- With an open but controlled Internet interface, many professional contributors may simplify the collection task; one may even envisage the contribution of benevolent amateurs with regard to data collection, pending a final review by a validation team composed of international professionals. Such an approach may significantly reduce the necessary budget for this project.

3. MANUEL DE CODAGE (MDC)

Looking back at the third edition of the MdC, published in 1988, it appears that there are essentially three parts in this document, and we are not convinced that they all belong to the same matter:

- The phonetic values should be part of the user-interface, so something to be managed at the software level, rather than concerning the language itself.
- The sign list is perhaps an endless debate and is not within the scope of this paper. Let's just notice that the acceptance by Unicode of a basic list of 1100 signs is already a significant step on the long path to communality.
- The third part refers to the MdC coding aspect: how to identify a sign and to represent it in the appropriate position. This will be the main focus of the present discussion.

3.1. *The MdC coding aspect*

This part of the paper focuses on the 'syntax' and associated 'semantics' used to code hieroglyphic texts with the aim to display or print them. The text coding is entered by the users:

- either through specialised graphical interfaces
- or through standard text editors.

This results in an 'External viewable coding',[8] where:

- 'External' means 'easily exchangeable between computers and between software';
- 'Viewable' means that 'it can be directly read and understood by human users'.

8 Some rendering engines may choose to use an internal coding. Here we will not consider any internal coding which constitutes specific implementation details.

Our subject here is the syntax and the semantics used in this 'External viewable coding'. So far this coding is based on a formal description elaborated in 1988 and named 'Manuel de Codage 88'.[9]

3.2. *Basic Requirements*

MdC requirements must be considered together from a user's viewpoint and from the rendering engine's perspective:

From a user's view point:

– coding must be easily understood and learned: a modern approach allows for a simpler syntax than the original MdC88;

– syntax must not be too verbose to input easily into standard text editors

– two sets of functionalities should be distinguished:

  – *basic* functionalities are fulfilled by any rendering software with a standardised syntax;

  – *extended* functionalities are not mandatory but, when supported, are based on a standardised syntax;

– the current MdC88 syntax must be supported for a reasonable period of time.

From a rendering engine's perspective:

– a 'regular' syntax affords the above benefits and makes it possible to produce efficient software using fewer system resources, and can be more economically developed and maintained (based on standard tools);[10]

– software does not display errors when an unsupported extended functionality is met;

– software should provide a tool analysing the code and diagnosing 'deprecated' (see below), i.e. unsupported, functionalities and syntax errors.

3.3. *Basic and Extended Functionalities*

Functionalities may be categorized on an axis 'basic / extended'. Below are several examples of functionalities:

- fragments vs facsimile: the simpler (more basic) functionality in this respect is to render fragments of hieroglyphic texts one by one without assembling them like in a facsimile (more extended) where fragments are combined with their relative positions, orientations, directions of writing, etc.
- simple vs complex cadrats: a complex cadrat requires a precise control of the size and positioning of inner subcadrats.
- simple vs complex alignments: a simple alignment is setting a position relative to the current position when the alignment is already specified (rather similar to word processing tabulation). A complex alignment takes into account actual size of the components being aligned.
- no text vs integrated texts — we refer here to texts for comments, transliteration, etc.: as an extended functionality, such text may be embedded within hieroglyphic texts.

3.4. *Developments*

To describe MdC developments, we will cover the two following 'directions' of development:

9 Many extensions without effective standardization have been made to the MdC88 by various software programs.

10 Precise definition of a 'regular' syntax is out of the scope of this paper (see http://en.wikipedia.org/wiki/Backus-Naur_Form)

- new functionalities (§3.4.1),
- new updated syntax (§3.4.2).

3.4.1. New Functionalities

We will give three examples of potential new functionalities:

- Vertical alignments. In some circumstances, it may be very useful to vertically align hieroglyphs and transliteration, for example.[11]

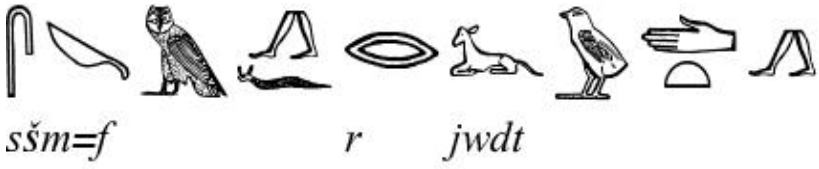

Figure 2. Example of vertical alignments

- Horizontal alignments. Fig. 3 displays a facsimile of the top part of the south face of the Luxor obelisk in the Place de la Concorde (Paris). It is composed of one fragment with three columns: horizontal alignments of the corresponding sections (for example cartouches) are very likely a new functionality desirable for the new MdC.

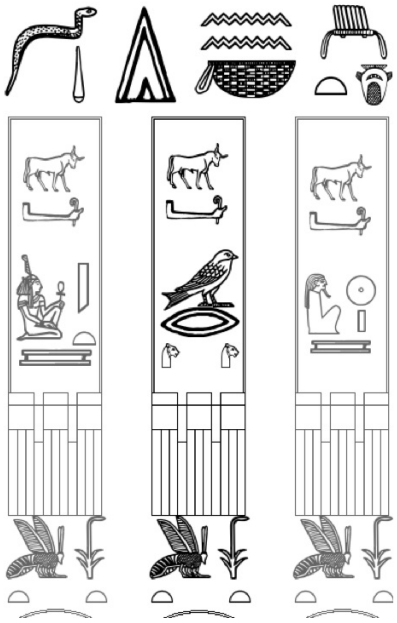

Figure 3. Example of horizontal alignments

- browsable facsimile: several text fragments with different size, orientation, etc. are combined like in the original artefact. In addition the facsimile is browsable.

11 See Nederhof, Mark-Jan. 2009. Automatic creation of interlinear text for philological purposes, in: *Traitement automatique des Langues* 30/2, 237-255.

Figure 4. Facsimile of the Nefertiabet Stela

When initially displayed, brown rectangles are drawn to delimitate 'elements' of the stela. For instance, to the right of the offering table are three horizontal sub-elements. On the left top corner of each rectangle is drawn a brown 'down' arrow which is clickable to explore the corresponding element. Browsing the table on the right side of the stela follows three steps:

Step 1: the full stela is displayed → click on the down arrow of the right-most table.

Step 2: the right most table is displayed. Note that the picture of this table is shown to the right (Hieroglyphic colouring could be an extended functionality) → click on the down arrow of the middle register.

Step 3: the middle register is displayed with transliteration, translation, comments, etc.

At steps 2 and 3, an 'up' arrow is displayed to return to the previous step. These three levels are defined in the text/facsimile MdC coding. The arrows used for navigation are interface elements not in the scope of the MdC coding.[12]

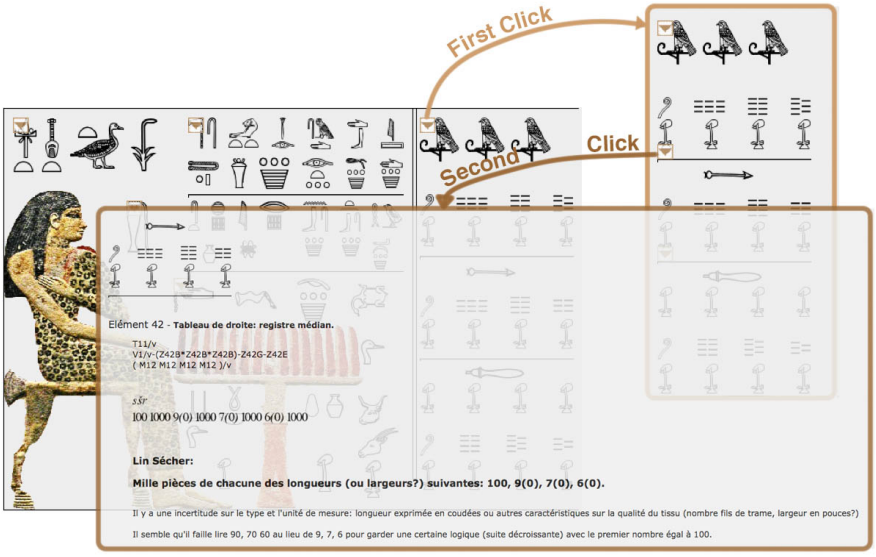

Figure 5. Browsing in a facsimile

12 This static printed presentation cannot demonstrate the whole process. A live example is available at: http://projetrosette.info/page.php?Id=799&TextId=134&line=1&nbrElts=1

Quite complex facsimiles may also be produced using this technique, as in the example of the astronomic ceiling of the Ramesseum in fig. 6.

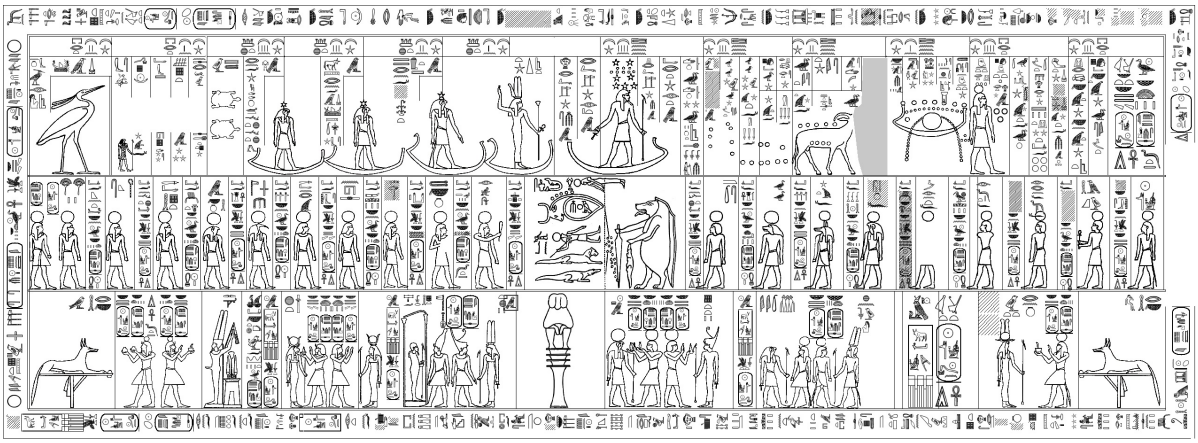

Figure 6. Astronomic ceiling of the Ramesseum

Facsimiles can be exported to external files with high definition and common graphic formats for inclusion in printed documents (facsimile of the fig. 6 has been printed on a poster 2.5 meters wide).

3.4.2. Move to an updated syntax

The basis of MdC88 is combining:

- symbols for hieroglyphs (Gardiner codes, phonetic equivalents)
- operators for positioning ('-', ' ', '*', ':', …)

This complies with all our 'basic requirements' described above. For 'modification' of hieroglyphs, however, MdC88 uses a 'chaotic' syntax, unable to fulfil these same requirements. For instance:

- the same character may signify totally different functions (e.g. '#' is used for both superposition and hashing);
- any new functionality requires the choice of a new 'character' in a more and more limited set. For instance, to colour a hidden or partially erased glyph in grey we use `$g`, which will limit further possible colours, is ambiguous with green, and does not allow us to cover the full colour space;
- this 'irregular' syntax requires more system resources and practically prevents the use of standard tools for the MdC syntax analysis (see http://fr.wikipedia.org/wiki/Lex_et_yacc)

It makes then sense to move to an updated syntax. We suggest four guidelines for evolution to such syntax:

- keep the 'foundation' syntax (described above): an important part of the existing coding remains valid;
- mark inappropriate syntax elements as 'deprecated': MdC88 syntax elements like '#', '$b', '$r', and many others may still be used but for a limited period of transition;
- implement a new and consistent syntax;
- fully support MdC88 during the same transition period.

3.5. *Implementation of these principles in the Rosette Project*

We will now give additional information about the implementation of the above principles on the Rosette Project web site.[13] A few new elements are added:

- 'modification' operators allow us to modify the rendering of affected hieroglyphs. Inserted just after the modified hieroglyph, they combine a '/' with a letter determining the modification:
  - 'c' for colour
  - 'r' for rotation
  - 'a' for hashing
  - etc.

  The letter is followed by relevant parameters like `/cr` for colour red or `/c255,0,0` for an RGB colour. These operators may be 'factorised' to several hieroglyphs enclosed between parentheses, for example: `( A1 A2 A3 )/cr/r45`: the three hieroglyphs will be drawn in red and rotated by 45°.
- '# tags' modify the drawing state. For instance `#poV;x=5;y=5 A1 A2 A3` draws A1 A2 A3 vertically starting from x=5 and y=5. A few other # tags are available. For example `#ssr;x2;y2;h3;w4` will draw a rectangle from x=2, y=2 with an height of 3 and a width of 4.[14]
- 'texts' may be mixed with hieroglyphs and can be amended by modification of operators and # tags.

The two following characteristics should also be mentioned:

- A syntax checker signalling deprecated elements is available.
- Integrated support of Unicode 5.2.

As an illustration, we will show the coding used for two examples:

(1) Stela of king Kamose: on top of this stela, an 'ankh' sign has been overwritten. To render this overwriting, we will use the following code: `M4 (t:3)*anx/y25/s50/x18/c100 G5 xa:a Hr:1 g:f`. Four modification operators are applied to the ankh sign:

`/y25` to set the hieroglyph at 25% in y direction of the cadrat

`/s50` to reduce the size to 50%

`/x18` to set the hieroglyph at 18% in x direction

`/c100` to set colour to grey (three RGB components = 100)

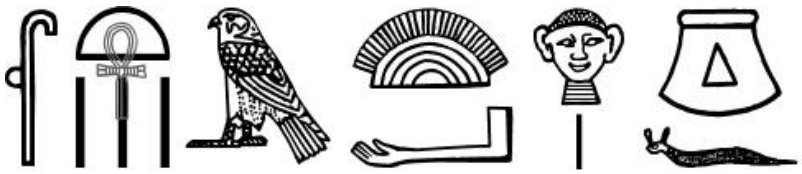

Figure 7. Inscription on the Stela of the king Kamose

(2) Astronomic ceiling (Ramesseum). The full ceiling has been coded and the obtained rendering was given in fig. 6. Let us explain the coding used for one element in the top left register:

13 The MdC88 syntax is fully supported.

14 The values refer to the 'base unit', which is a parameter expressed in pixels.

Figure 8. Fragment of astronomic ceiling in Ramesseum

The underlying code is as follows `#poH;d='r' U28 G1 P34/ar25:pt #poV D58*(N35:W24) G31 D4 Q1*A40`. From left to right we see:

`#poH;d='r'`

`oH`: sets orientation to horizontal

`d='r'`: sets direction of writing to right to left

`P34/ar25`: `P34` with `/a` modification operator and `r25` parameter.

`/ar25`: 25% of the hieroglyph is hashed from right side

`#poV`: sets orientation to vertical. Direction coded above is maintained.

3.6. *Conclusion*

The Manuel de Codage, last updated in 1988, needs to evolve but must remain a standard to allow exchanges between Egyptologists to be conducted as easily as possible.

This paper suggests directions for development (new functionalities and updated syntax), proposes principles for these developments, and finally presents how the Rosette Project implements those principles. It now seems relevant to setup a working group commissioned to:

- define new principles of syntax and, as a consequence, list deprecated elements of the MdC88 syntax;
- list 'basic' and a first set of 'extended' functionalities (see §3.3);
- determine the associated syntax elements;
- determine appropriate milestones for the implementation of above elements.

Abstract

Sharing the heritage of Ancient Egyptian written production means facing numerous technical challenges. The goal of this paper is to build a preliminary inventory of these challenges and to propose some possible solutions. After a quick overview of the topics that are possible candidate to an international standardization, the paper focuses on two aspects. (1) The 'Multilingual Egyptological Thesaurus' (MET), initiated in 1996 by Dirk van der Plas, has not changed since 2003. It could be updated and expanded with minimal effort under the coordination of an official body such as the Center for Documentation of Cultural and Natural Heritage (CULTNAT). (2) The 'Manuel de Codage' (MdC) has not benefited from developments in computer science since the third edition was published under the *Informatique et Égyptologie* (I&E) mandate in 1988. Over time, each hieroglyphic software program has developed its own specific syntax to satisfy emerging needs, making it difficult for users to share ancient Egyptian texts. For these two topics, we will suggest a plan for improvement based on the Rosette Project's experience, though the input of the Egyptologists' community at large is appreciated to refine various concepts and identify the best route forward.