On extending the Egyptian Hieroglyphic repertoire in Unicode

Bob Richmond,2019-09-10

A contribution to the Tournan en Brie meeting on Egyptian Hieroglyphs, September 12-13th 2019

Preamble

"… the forms of the signs chosen to be the staple of the new fount are taken from the Eighteenth Dynasty tombs at Thebes. These are admirably fitted for reproducing texts in Middle Egyptian, the classical period of the language ranging from Dyn. XI down to the end of Dyn. XVIII. The propriety of using these forms for Old Egyptian or for texts of the latest period is far more questionable, and we have even heard seriously canvassed the advantages of a fount so palaeographically untrue to any particular epoch that it is exactly equally suited to all periods alike! To this thesis we cannot, of course, subscribe, and hold that it is best, since a special fount cannot be devised for each separate period, to employ a fount which shows great accuracy for the central and most important stage, and a larger or smaller degree of distortion for the other periods." Preface to Gardiner (1928).

The Gardiner font was developed alongside his Egyptian Grammar (1927). His List of Hieroglyphic signs took the joint font/grammar enterprise beyond being simply a font catalogue. This practical approach probably accounts for the longevity of the system. Hieroglyphs to Unicode (in version 5.2, 2009) used the Gardiner font and sign list as its primary source and reference.

Gardiner cannot have predicted the information revolution with advances in technology which make it possible to envisage different specialist fonts to be created for different epochs and purposes far beyond the printing of books. Nevertheless, his central thesis of placing Middle Egyptian as the central place from which treatments of earlier and later periods could be developed appears to be generally regarded as valid today.

The most recent major development in hieroglyphic fonts is the *Hieroglyphica* font catalogue (2nd and final Edition, 2002). This collection of over 7000 hieroglyphs uses an alphanumeric naming scheme for the hieroglyphs introduced in 1988 by *Inventaire des signes hieroglyphiques en vue de leur saisie Informatique* at the start of the project that yielded the first *Hieroglyphica* release in 1993. *Inventaire* attempted to combine signs from the Gardiner font, the Lepsius/Theinhardt font and the IFAO font using IFAO classification – echoes of which survive in the *Hieroglyphica* systematical lists. The *Inventaire* naming scheme is adapted from Gardiners sign list and catalogue with which it is mostly compatible. This naming scheme is carried over into the *Hieroglyphica* catalogue and has been adopted by much modern work over the last 30 years. *Hieroglyphica* spans the whole historic period hieroglyphic was used from Early Dynastic through to the Greco-Roman era, including substantial additions within the Gardiner Middle Egyptian scope. The font uses normalized forms of glyph in a line drawing graphical style. There is no equivalent to the Gardiner sign list for *Hieroglyphica* so identification of hieroglyphs tends to involve detective work at present. Nevertheless, it is the closest thing to a working standard in Egyptology.

In 2015 work began on extending the capabilities of Egyptian Hieroglyphic in Unicode with discussion of what became the Unicode 12 Egyptian Hieroglyph Format Controls and a document L2/15-240 *Preliminary draft of the Ptolemaic repertoire (A: Man and his occupations)* by Michel Suignard. This was updated to L2/16-079 *Preliminary draft for the encoding of an extended Egyptian Hieroglyphs repertoire* and Michel presented his work at the I&E Cambridge, Summer 2016 meeting. The consensus was that the approach to his database (which I call **Unikemet2** here), using Hieroglyphica and other sources, was worth progressing.

Three year later, Michel's the latest draft is available as L2/19-220 *Revised draft for the encoding of an extended Egyptian Hieroglyphs repertoire / First Tranche: Human, God and Goddess / Second Tranche: Human parts, Mammals, Mammal parts (WG2 N5063).*

In my opinion, we are ready to discuss what remains to be done to begin extending the repertoire.

The Road to Repertoire Extension

There is plenty to be done and various issues to be resolved on how to figure where to go from where we are now to a formal proposal or proposals for an extension or extensions. I've fleshed some of the prerequisites in the next section.

Timescales. If a formal proposal is accepted by the Unicode Technical committee during 2020 there are procedures to be followed, such as scrutiny by National Standards bodies for comment then ballot, before it can be accepted into the standard. This process would normally complete in time to be included in the Unicode Standard release 15 expected in 2022. Implementation of Unicode 15 such as ensuring key software recognises the new characters as Egyptian Hieroglyphs in text shaping engines would normally be expected to be released within several months of the release date.

More than one formal proposal can be in the pipeline at the same time, so it is possible to expand the repertoire in parts without waiting 2/3 years between parts. However, there needs to be a clear rationale for a multi-part strategy both for administrative and logistical reasons as well as consideration of how staging might benefit the hieroglyphic user base.

The current version of the **Unikemet2** database is a good starting point in my opinion and provides a coherent basis for looking at staged expansions. Additions to the database can be made as new hieroglyphs are established as needed. Hieroglyphica as a major source links the database to the practical aspects of working with hieroglyphs as practiced by the user base.

However, as noted previously, Hieroglyphica expands on Gardiner over all periods, from Early Dynastic though to Greco-Roman. Modern technology does not fully invalidate Gardiners 1928 comments on the propriety of a one size fits all font. Development of the current extension work began with a focus on Ptolemaic but the extensions being discussed now are at least as much about expansion of the Middle Egyptian treatment in Unicode, Late Egyptian and earlier periods as the Greco-Roman era.

It is important for non-specialists to understand that almost all hieroglyphic from all periods is dominated by signs already encoded in Unicode. This includes the Dendera Ptolemaic Temple reliefs as published by IFAO, the encoding of which was a major contributor to the many Ptolemaic entries in the Hieroglyphica font catalogue. Furthermore, ultimately hieroglyphic is a handwritten script and there will always be applications where plain text encoding is not enough, and higher-level protocols or facsimile will be necessary to capture an ancient text to the degree deemed satisfactory to the modern writer for a given purpose.

In the next section I've identified some specific topics. The "Actions" are specific pieces of work that can ease the extension process.

Some prerequisites for extensions

Action: Detail use of EGYPTIAN HIEROGLYPH FORMAT CONTROLS in relation to Unikemet2 database

So far, I have only considered two-character combinations of hieroglyphs using the EGYPTIAN HIEROGLYPH OVERLAY MIDDLE format control. Not any of the more exotic overlays from font catalogues. See the draft

There are over 200 such combinations in Unikemet2 at present that are overlays of signs already encoded. Over 100 additional combinations are available if currently unencoded components are included in a repertoire expansion.

Traditional font catalogues (including Lepsius-Theinhardt, Gardiner, IFAO and Hieroglyphica) treat combinations as atomic, including overlays, so combined forms should remain listed in Unikemet2 but marked as implemented using format controls and excluded from consideration for repertoire extension.

It is important that a list of attested overlays (and maybe other combinations) is maintained outside the Unicode standard itself (overlaying random hieroglyphs is rarely useful or pretty and should not be done using font-specific quirks). That way font and software developers can ensure they do a good job with overlays that are used. By keeping the list separate, new items can be added to the attested list when identified without need to go through

the standardization process. It is desirable a consensus is achieved here among interested parties so such a list (with documentation) can be actioned and consequences factored into Unikemet2. For the sake of ongoing work on implementing format controls it is highly desirable this is produced before end of the year. Discuss.

Other format controls could be added such as the mooted INSERT CENTRE that could be used for arrangements such as one or more hieroglyphs enclosed in a large D028 (there are 58 D058 enclosures in HG though their sources are undocumented so it's hard to take a view until their significance is well-understood). There would need to be a clear cost/benefit analysis that show any such addition is worthwhile. Discuss.

Action: Document compatibility issues with existing Unicode hieroglyphs (Gardiner set +)

See https://www.unicode.org/Public/5.2.0/charts/CodeCharts-noHan.pdf for what is encoded. The current code chart (12.1) https://www.unicode.org/Public/UCD/latest/charts/CodeCharts.pdf contains divergences and should not be used as guidance until corrected.

For instance:

- A043 has gained a beard in current code chart. This should be corrected and bearded version (HG.A43) added to Unikemet2 for Hieroglyphica compatibility.
- Hieroglyphica HG.N37A uses a different variant of N037 to N037A (HG has 3 strokes not 2) and the latest Unicode code chart incorrectly uses the Hieroglyphica variant. This should be corrected, and HG version added to Unikemet2.
- The latest D031A is different to Gardiner. Fix.
- The latest D027A seems much too large. Fix.
- Cartouche begin/end signs V011A, V011B and V011C have shrunk in current code chart (important: these are supposed to be full height). Fix.

There are more fixes needed.

V012A and HG27 look suspiciously alike. So, is the glyph a rope (Gardiner) or part of a bird (Hieroglyphica)? Needs resolution for Unikemet2.

Some orientation and other variants given by Gardiner didn't make it to the Unicode release so these should be checked all present in Unikemet2 for inclusion in the next expansion.

It is important that data coded relative to Gardiner hieroglyphs is not broken so these and any similar issues should be documented and reviewed, and any new forms or points of note added to Unikemet2 if necessary. That way we set firm ground for repertoire extensions. Discuss.

Action: Review sources referenced in Unikemet2

Where can I find the *Hornung & Schenkel (2007, last modified in 2015), Zeichenlist* as referenced?

Question is what additional references or changes are necessary/desirable to be added to Unikemet2 before proceeding to a formal proposal for an encoding?

Note it is important that reviewers and users can potentially access the source material and we don't leave people in the future in the situation of Hieroglyphica where unpublished sign lists are stated as sources. Discuss.

The Thot Sign List sounds promising when suitable and ready for addition - https://tsl.philo.ulg.ac.be/ .

The *Inventaire* makes mention of a sign list in preparation for Volume 8 of the Wörterbuch series by the Ancient Egyptian Dictionary Project (Berlin) which was involved in the Hieroglyphica work This would be useful to help dissect Hieroglyphica, if it still exists or any relevant data survives. Is there anything suitable for use available from its successor, the *Thesaurus Linguae Aegyptiae* project (http://aaew.bbaw.de)?

Any others, especially "essentials"?

My *AnalysisOfSomeMdCCodedTexts.pdf (2016, source for Unikemet2)* is now relocated to github https://github.com/HieroglyphsEverywhere/Docs/blob/master/Archive/AnalysisOfSomeMdCCodedTexts.pdf. If there is any more material encoded in MdC-type format I can look at doing some more analysis for reference.

I have a Lepsius/Theinhardt font concordance with Unikemet2 in preparation. I am also preparing an Early Dynastic hieroglyphic draft profile based on Jochem Kahl *Das system der ägyptischen in der 0.-3. Dynastie* (Wiesbaden 1994) and later publications by Kahl. Likewise, a draft Middle Egyptian profile (11th to 18th Dynasties), including signs given in James Allens *Middle Egyptian* Second Edition, 2010 (relevant to his *Middle Egyptian Literature* (2014)). I'm also including relevant Moeller hieratic material. These profiles are intended to be useful for font and software applications and are not tied to Unicode expansions or meant to be a primary source but may prove useful. I hope to share this material on GitHub this year to allow progress to continue in an open, collaborative, spirit. I also hope to add draft profiles for other periods when time permits and would welcome contributions.

Action: Unikemet2 Taxonomy

The taxonomy index based on IFAO for the Unikemet2 database seems to me useful with the index using the ((A-IK-Z){1}|AA)-(0-9){2}-(0-9){3} format. This approach was taken by *Inventaire* although the categories used in the Hieroglyphica Systematical list are broader e.g. 5 subcategories of category A rather than the 34 sub-categories of L2/19-220.

I suggest dwarf be treated as a man (or woman if applicable) with no separate category and man in ship moved to varia or some broader sub-category. It's worth looking through all sub-categories to look for any similar simplifications.

N5O63 "Current Issues" lists some important point for review and other issues may be identified by specialists to add to this list.

Consensus over the full taxonomy list over all categories would be an important step in enabling possible expansions to be based on Unikemet2.

Action: Define transitional PUA system

With the fruits of expansion maybe 3 or more years away from release, it would be valuable to agree a specification of use of the Unicode Private Use Areas for codepoints during the transitional period. Such a specification will simplify work on expansions and, if managed and defined properly, deliver benefits to specialist users of hieroglyphic.

I've been using an unpublished PUA system in Unicode Plane 15 compatible with Aegyptus 5.1 where applicable. Aegyptus is a hieroglyphica-like font, by George Douros. Note Aegyptus 8.0 latest release, http://users.teilar.gr/~g1951d/, has reassigned codepoints so is no longer compatible with work done using the earlier version.

It is important that code points in such a PUA list are not changed, once specified, for stability both during the transitional period to Unicode extensions release and later when users still may want to refer to what is then obsolete material and old fonts.

I suggest the transitional PUA system be organized by taxonomy index with gaps in the code-space for wiggle room, once the Unikemet2 taxonomy is agreed. This to be deliberately defined so that the exact Unicode organization decided on for expansion is not linked to the PUA values. Software can just use a published mapping table to switch between the PUA and released standard. Font developers can use some automated process (e.g. open-source Python script) to build a release font from a PUA version.

Suggestions appreciated.

I'm doing work on fonts anyway so I can edit the specification and make an implementation based on an Aegyptus fork when the full taxonomy list is stable. If Michel and George are ok with that? Although more than happy if somebody else want to take on the task!

Discussion: Content of first extension

Michael Everson proposed an extension in **N4741 L2/16-250** Preliminary proposal to encode Möller's Egyptian Hieroglyphs in the SMP of the UCS. This is specifically oriented at helping with hieratic transliteration to hieroglyphic.

N5063 L2/19-220 Revised draft only details categories A to F as the first and second tranche of the Unikemet2 database but with others to follow.

If there is consensus on using Unikemet2 (with any agreed additions of changes) as the basis for future extensions, there are many ways one could proceed and stage proposals.

N5063 proposes the essential criterion for considering a hieroglyph for encoding be that the weight be 1 or greater. The question, looking towards a formal proposal, is what other criteria might be used to help with selection.

From the point of view of the user base, staging formal extensions to Unicode by category is of limited value. For instance, to release an expansion containing just Anthropomorphic signs (Categories A-C) only makes a marginal improvement to Unicode for work concerning Middle Egyptian hieratic texts. Nevertheless, from a technical perspective it may make sense nevertheless even if it means a slower uptake of Unicode among specialist users for work with hieroglyphic.

The availability of Hieroglyphica and the related Aegyptus font means font availability is assured from the onset even for 1000s of hieroglyphs. However, it is important to recognize that most fonts will not be expected to cover all parts of the extended code space. A font concerned with the long Classical period would not and probably should not attempt to provide glyphs relating to the Greco-Roman era. An application or font focussed purely on Middle Egyptian need not be concerned with the Late Egyptian expansion either and could attempt to better represent 18th dynasty palaeography following the Gardiner approach. Fonts aimed at the general public may be much less ambitious.

My own thinking is that is essential to derive some top-level analysis of the Unikemet2 database before deciding on a strategy. Answer questions such as how many and what signs are currently attested only from Greco-Roman sources?, what is early dynastic coverage like?, likewise Middle Egyptian and Ramesside period? This should help inform Egyptologists and others gauge benefits or potential issues with what is being proposed. It may help with deciding on exactly how to partition the Unicode code space.

I'm personally interested in reading what Egyptologist thinking is on these matters before we lock down details of content and staging.

Discussion: Unicode Naming Convention

The current Unicode naming convention for hieroglyphs is of the form ((A-IK-Z){1}|AA){3}{A-Z} i.e. category letter, a number padded to left by 0s to 3 digits followed by an optional upper case letter denoting graphic variant This is based on the Hieroglyphica/Gardiner system from which it is easy visually distinguished e.g. A001 vs. A1.

N5O63 proposes hieroglyph names in all extensions are based on the Unikemet2 taxonomy index, so A001 would have been named A-01-001 if it was not already named. This allows the Unicode name to communicate more of its graphical form than the current system. There is no attempt to communicate the variant concept in the name, variants are given their own code and any treatment of variants is done outside of Unicode.

For most technical purposes the name is not terribly important. Software works with Unicode code points and can map these into whatever name space e.g. catalogue number as required e.g. map U+13000 to A001, A-1-001, HG.A1 (Hieroglyphica) or LT.A89 (Lepsius-Theinhardt) or whatever. Hyphen '-' is not allowed in an Adobe Feature file glyph name but a simple substitution like A-1-001 -> A.1.001 could be used. Hyphen is not allowed in MdC glyph names but some workaround could likewise be used or more likely MdC dialects would continue with Hieroglyphica style aliases.

Names are the main user-facing feature of Unicode so it is instructive to ask what publications like Wikipedia might look like.

Users are used to the Hieroglyphica form of sign names in modern Grammars and other uses of hieroglyphs over the last 30 years. The taxonomy index as name would make sign lists look more complex so it seems unlikely these names will prove popular for less technical publications and with the users referring to them.

There is a reasonable case to be made for an extension which encodes variants of existing signs using names based on the current system. For instance the Late Egyptian Hieroglyph variant HG.G17A would be encoded as G017A. This avoids introducing unnecessary unfamiliarity into the name space. One might also consider other factors.

Analysis of Unikemet2 would help characterize the possibilities and enable informed discussion on this topic.

I'm interested what Egyptologists make of this. Is there any mileage in using a separate name space for post-classical Egyptian e.g. for Ptolemaic temples and later?