Document numérique

Sommaire Volume 16 – n° 3/2013

GESTION INFORMATISÉE DES ÉCRITURES ANCIENNES

Linguistic issues and intelligent technological solutions in encoding Sanskrit

Peter Scharf

The Sanskrit Library
scharf@sanskritlibrary.org

ABSTRACT. *Contemporary uses of information technology demand higher standards of encoding than the inherited systems prevelant today. Guided by visual factors, current encoding systems reproduce deficiencies inherent in traditional writing systems. The contemporary use of computers for the manipulation of linguistic and textual data demands more relevant information-processing principles. The fundamental issue in encoding natural language texts concerns the relation information selected bears to natural language structure. Selecting appropriate information to encode demands addressing whether to encode written characters or speech sounds, whether to encode segments or features, and what criteria to use to contrast items selected for encoding. These issues are considered in relation to Sanskrit, the principal culture-bearing language of India, which is characterized by an extensive oral tradition, a highly phonetic orthography, and a copious literature.*

RÉSUMÉ. *Aujourd'hui, l'utilisation de technologies de l'information demande des normes de codage de plus haut niveau que les systèmes actuels. Guidés par des facteurs visuels, les systèmes de codage actuels reproduisent des lacunes inhérentes aux systèmes d'écriture traditionnelle. L'utilisation contemporaine de l'ordinateur pour la manipulation de données linguistiques et textuelles demande des principes de traitement de l'information plus pertinents. La question fondamentale du codage des textes en langue naturelle concerne la relation entre les informations choisies pour le codage et la structure du langage naturel. Choisir les informations appropriées au codage demande qu'on réponde à trois questions : coder des caractères écrits ou les sons de la parole ? Coder les segments ou les caractéristiques ? Quels critères utiliser pour bien distinguer les articles choisis pour le codage ? Ces questions sont examinées dans le cadre du sanscrit, la principale langue porteuse de la culture en Inde, qui se caractérise par une vaste tradition orale, une orthographe très phonétique et une abondante littérature.*

KEYWORDS: *Sanskrit, linguistics, encoding, phonetics, segments, features, contrast.*

MOTS-CLÉS : *sanscrit, linguistique, encodage, phonétique, segments, caractéristiques, contraste.*

DOI:10.3166/DN.16.3.15-29

Dictionnaire hiéroglyphique, inventaire des hiéroglyphes et Unicode

Dimitri Meeks

CNRS. Directeur de recherche honoraire
19 rue Gamay, 34980 Saint-Clément-de-Rivière. France
dimitri.meeks@wanadoo.fr

RÉSUMÉ. Le standard Unicode pour les hiéroglyphes égyptiens a été validé en 2009 après des années de débats sans que pourtant l'ensemble de la communauté des égyptologues ait été associée au processus. L'Unicode actuel valide quelque 1200 caractères, les plus usités de la période classique. Son extension n'est pas envisagée dans l'immédiat. Si, toutefois, on souhaite travailler sur l'ensemble des textes hiéroglyphiques, de l'époque archaïque aux époques hellénistique et romaine, c'est de plus de 10 000 caractères qu'il faudrait disposer, un objectif que l'Unicode ne se propose en aucun cas d'atteindre. Je travaille actuellement à l'élaboration d'un dictionnaire de l'égyptien ancien, sous forme d'une base de données lexicales, ainsi qu'à une base de données inventoriant l'ensemble des hiéroglyphes. Ces deux bases se fondent sur la documentation lexicale et paléographique que j'ai rassemblée au cours des quarante dernières années, la seconde ayant vocation à dresser un catalogue aussi exhaustif que possible des caractères hiéroglyphiques actuellement connus à partir des monuments publiés en photographie ou en facsimilé. C'est un travail qui n'a jamais été entrepris par l'égyptologie depuis le déchiffrement des hiéroglyphes par Champollion en 1822. Une fois achevé, il servira de base à la constitution d'une fonte numérique couvrant tous les besoins de la recherche, quelle que soit la période de l'écriture envisagée. C'est aussi sur cette fonte que s'appuiera le dictionnaire pour la rédaction de ses notices.

ABSTRACT. The Unicode standard for Egyptian hieroglyphs was validated in 2009 after years of debate, though the whole community of egyptologists was not associated to the process. As to the present time, Unicode validates the most common 1200 characters of the classical period. No extension is planned in a near future. However, if one needs to work on hieroglyphic texts as a whole, conspicuously more than 10 000 characters are needed, a goal Unicode certainly do not intend to achieve. I am currently working on a dictionary of ancient Egyptian, taking on the form of a lexical data base, and on another data base inventorying all the hieroglyphs. Both are based upon the lexical and palaeographical documentation I collected over the past fourty years, the last being intended to produce a catalogue as exhaustive as possible of all hieroglyphs presently known from publications in photographs or facsimiles. This work was never undertaken by egyptology since the decipherment of hieroglyphs by Champollion in 1822. Once completed it will serve as a source for a new digital font meeting all the requierements of the research, no matter which period of the writing is taken into account. This font will ultimately be used for the editorial work on the dictionary.

Document numérique – n° 3/2013, 31-44

Réviser le codage de l'égyptien ancien

Vers un répertoire partagé des signes hiéroglyphiques

Stéphane Polis[1], Serge Rosmorduc[2]

1. Fonds National de la Recherche Scientifique – Université de Liège
Département des Sciences l'Antiquité
Place du XX Août, 7
B-4000 Liège, Belgique
s.polis@ulg.ac.be

2. Cédric/ILJ/Conservatoire National des Arts et Métiers
Rue Conté, 2
75003 Paris, France
serge.rosmorduc@cnam.fr

RÉSUMÉ. *Nous proposons de réviser le codage de l'égyptien ancien qui repose sur un standard nommé « manuel de codage » (1988) ne répondant pas aux besoins actuels dans la création de corpus hiéroglyphiques. Une analyse des 60 000 graphies du corpus Ramsès nous permet de faire deux propositions concrètes concernant, d'une part, les principes présidant à l'encodage des graphies hiéroglyphiques dans les corpus annotés et, d'autre part, la nécessaire refonte du répertoire des signes hiéroglyphiques.*

ABSTRACT. *Based on an analysis of the 60 000 spellings found in the Ramses corpus, we show that the encoding scheme that is used for hieroglyphic texts, known as the* Manuel de Codage *(1988), is problematic for the development of text corpora in general. We consequently (1) argue in favour of basic principles that should be followed when encoding hieroglyphic texts in the framework of corpus projects, and (2) suggest guidelines for a new encoding scheme, especially with respect to the structure of the sign-list.*

MOTS-CLÉS : *hiéroglyphes, encodage, Unicode, liste de signe, égyptien ancien.*

KEYWORDS: *hieroglyphs, encoding, Unicode, sign-list, Ancient Egyptian.*

DOI:10.3166/DN.16.3.45-67

Polices de caractères et inscriptions monétaires

Le projet PIM

Florence Codine

Département des Monnaies, médailles et antiques
Bibliothèque nationale de France
5 rue Vivienne, 75002 Paris, France
florence.codine@bnf.fr

RÉSUMÉ. *Le contenu sémantique des légendes monétaires n'est en général pas leur unique intérêt. Les particularités graphiques des lettres ou symboles qui les composent sont porteuses de sens, et fournissent des indications essentielles à la compréhension du contexte géographique, chronologique, technique et culturel de la monnaie, ainsi que des pistes pour la lecture et l'étude linguistique des légendes. Les études sur ces questions n'ont pas manqué ces derniers siècles, mais les nouvelles possibilités offertes à la recherche par le numérique font aujourd'hui clairement ressentir le manque d'un outil satisfaisant pour transcrire, publier et analyser ces inscriptions. Le projet de recherche Polices pour les Inscriptions Monétaires (PIM) se veut une réponse à cette situation.*

ABSTRACT. *Coin inscriptions are interesting for more than just their semantic content. The graphic peculiarities of the letters and symbols are just as significant, and provide the researcher with information regarding the geographical, chronological and cultural context of the coin. Their study is also an asset for reading and understanding the inscriptions themselves. These questions have raised interest in the past, but the new opportunities now offered by digital technologies have made patent the need for new means of transcribing, publishing and analysing these inscriptions. The PIM research project on fonts and coin inscriptions is an attempt to provide such a tool.*

MOTS-CLÉS : *épigraphie, numismatique, polices de caractères.*

KEYWORDS: *epigraphy, numismatics, fonts.*

DOI:10.3166/DN.16.3.69-79

Ontologie des formes et encodage des textes manuscrits médiévaux

Le projet ORIFLAMMS

Dominique Stutzmann

Institut de Recherche et d'Histoire des Textes, CNRS (UPR 841)
40 avenue d'Iéna, 75116 Paris, France
dominique.stutzmann@irht.cnrs.fr

Résumé. L'approche historique des écritures du Moyen Âge soulève des questions complexes, (1) pour rendre compte, décrire et représenter le substrat graphique par le « codage », (2) pour analyser les formes en tant que telles par Vision par ordinateur, (3) pour procéder à l'analyse des phénomènes graphiques par le « balisage » ou « encodage ». Les sept partenaires du projet ANR ORIFLAMMS (ANR-12-CORP-0010)*, proposent des solutions nouvelles, en particulier pour l'analyse et l'encodage des abréviations et pour l'alignement texte-image afin de constituer une ontologie des formes écrites du Moyen Âge.*

*Abstract. Medieval scripts are a challenge to historical analysis, as for (1) describing et representing the graphical evidence through "coding"; (2) analyzing and clustering letter forms and their features through Computer Vision; (3) analyzing historical phenomena through "encoding". All seven partners of the ANR funded research project ORIFLAMMS (*ANR-12-CORP-0010*) develop new solutions, esp. for encoding abbreviations and aligning text and image, as a first step to build an ontology of medieval written forms.*

Mots-clés : paléographie, encodage, TEI-P5, ontologie.

Keywords: palaeography, encoding, TEI-P5, ontology.

DOI:10.3166/DN.16.3.81-95 © 2013 Lavoisier

Dealing with glyphs and characters

Challenges in encoding medieval scripts

Odd Einar Haugen

Department of Linguistic, Literary and Aesthetic Studies
University of Bergen
odd.haugen@lle.uib.no

ABSTRACT. *This article investigates the meaning of the key concepts 'characters' and 'glyphs' in the Unicode Standard with reference to medieval script, especially the runic alphabet. Two problems of font encoding are discussed in some detail: (1) the disambiguation of characters sharing the same glyphs, and (2) the delimitation of variant glyphs. Criteria are suggested for specifying the character–glyph relationship, and a minor area of the runic code chart is analyzed on the basis of these criteria. The article suggests that better documentation of variants in medieval script is needed, not only in order to identify suitable candidates for proposals to the official Unicode Standard, but also in order to explain the usage and meaning of these variants for other scholars and for font designers.*

RÉSUMÉ. *Cet article interroge la définition des concepts de caractères et de glyphe définis par le standard Unicode, en se référant aux écritures médiévales et en particulier à l'alphabet runique. Deux problèmes liés au codage des polices y sont analysés en détail : (1) la distinction entre différents caractères partageant le même glyphe, et (2) la définition des variantes graphiques. On propose de définir plusieurs critères permettant de caractériser cette relation glyphe–caractère, et l'on donne un exemple de caractérisation, en appliquant ces critères d'analyse à une zone étroite de la grille des caractères runiques. L'article conclut sur la nécessité de bénéficier d'une meilleure documentation sur les variantes allographiques des écritures médiévales, non seulement pour identifier les caractères manquant à Unicode, mais également pour en expliquer l'usage et la signification aux chercheurs et aux dessinateurs de caractères.*

KEYWORDS: *Unicode Standard, font encoding, characters, glyphs, runes.*

MOTS-CLÉS : *Unicode, polices, codage, caractères, glyphes, runes.*

DOI:10.3166/DN.16.3.97-111

Transcription et codage des imprimés de la Renaissance

Réflexions pour un inventaire des caractères anciens

Jacques André[1], Rémi Jimenes[2]

1. *Inria-Rennes, rédacteur en chef honoraire de* Document numérique
   Jacques.Andre35@gmail.com

2. *Centre d'études supérieures de la Renaissance*
   remi.jimenes@univ-tours.fr

RÉSUMÉ. *Conservant le plus grand nombre possible d'informations du document-source, une transcription de texte imprimé ancien devrait pouvoir servir de base non seulement à des analyses littéraires, mais également à des études « paléotypographiques ». Pour ce faire, il faudrait disposer d'un codage normalisé permettant d'assurer une correspondance univoque entre les caractères de la transcription numérique et ceux de la source originale. Le terme « caractère » pouvant prêter à confusion, nous introduisons un nouveau concept : celui de « typème », intermédiaire entre le caractère et le glyphe tel qu'Unicode les définit. Nous proposons d'utiliser le codage MUFI, une extension d'Unicode, augmentée des typèmes attestés dans les imprimés anciens, afin de produire une transcription dite « typémique », reproduction fidèle de la composition typographique du document original. Nous concluons sur la nécessité de réaliser l'inventaire des typèmes attestés dans les imprimés anciens, qui fera l'objet d'un Projet d'Inventaire des Caractères Anciens (PICA) actuellement à l'étude.*

ABSTRACT. *Preserving as many informations as possible from the original document, a transcription of ancient printed text should serve as a basis not only for literary analysis, but also for palaeotypographic studies. With this aim, we require a standardized encoding able to preserve a unequivocal link between the characters of the digital transcription and those of the original source. We define here the new concept of typem, a transitional element between the notion of character and glyph as defined by Unicode. It is proposed here to use MUFI, an extension to the Unicode standard, by adding new code points dedicated to "typems", in order to produce what we call "typemic transcriptions", reproducing all the characters of the original document. Finally, a project of a census of all the typems, named PICA (Projet d'Inventaire des Caractères Anciens), is described.*

MOTS-CLÉS : *typographie, MUFI, Unicode, codage, documents anciens, inventaire, caractères, typèmes, imprimés, Renaissance, PICA.*

KEYWORDS: *typography, MUFI, Unicode, encoding, ancient document, inventory, types, typems, printed material,* Renaissance, *PICA.*

DOI:10.3166/DN.16.3.113-139