Les systèmes d'information dans le domaine des sciences de l'Antiquité : Le cas de la publication des documents épigraphiques et papyrologiques

Pascal Lemaire

▶ **To cite this version:**

Pascal Lemaire. Les systèmes d'information dans le domaine des sciences de l'Antiquité : Le cas de la publication des documents épigraphiques et papyrologiques. 2009. hal-00583325

HAL Id: hal-00583325
https://hal.science/hal-00583325

Preprint submitted on 5 Apr 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Sommaire

Sommaire 1

Résumé....3

Remerciements....5

Introduction....7

Chapitre 1 : Brève introduction à l'épigraphie et à la papyrologie....10

Epigraphie, l'univers éclaté 10

La papyrologie, domaine de la collaboration amicale 13

Chapitre 2 : XML, TEI, EpiDoc et les autres, tour d'horizon technologique 15

Stocker l'information 15

La problématique de l'encodage des caractères non latins....17

eXtended Markup Language, la souplesse dans la rigueur 20

Text Encoding Initiative, un projet de tous pour tous 27

EpiDoc, de l'universel à la spécialisation....30

Chapitre 3 : La technologie en œuvre, quelques projets sous les projecteurs 32

Informatique et sciences de l'Antiquité, une histoire déjà ancienne 32

Des projets de nature différente 34

Numérisation versus digitalisation 39

Quelques exemples de projets 41

Searchable Greek Inscriptions 41

Electronic Archive of Greek and Latin Epigraphy (EAGLE)....47

Inscriptions of Aphrodisias Project (InsAph)....50

Duke Databank of Documentary Papyri (DDbDP) - Advanced Papyrological Information System (APIS) – Heidelberger Gesamtverzichnis der griechischen Papyrusjunden Ägyptens (HGV) 61

Perseus 64

Trismegistos....71

Papyrological Navigator (PN) et Son of Suda Online (SoSOL) 80

Chapitre 4 : Opportunités, contraintes et interrogations 85

Le choix du marquage : une décision influencée tant par les objectifs du projet que par son environnement. 85

L'automatisation du marquage 87

Mashups : l'exemple du Lexicon of Greek Personal Names et des projets « Inscriptions of Aphrodisias », « Inscriptions of Roman Cyrenaica » et « Inscriptions of Roman Tripolitania » 88

Quelle politique de mise à jour des bases de données de documents digitalisés ? 95

Le concept d'infrastructure 98

Chapitre 5 : La base de données des inscriptions coptes documentaires 101

Conclusions 106

Bibliographie 109

Bibliographie imprimée 109

Webographie 111

- Annexe 1 : tableau de représentation du Beta Code 113

Résumé

Ce mémoire a pour visée l'étude de l'impact des technologies de l'Internet sur le domaine des sciences de l'antiquité, et plus particulièrement les domaines de l'épigraphie et de la papyrologie. Au travers d'une série d'études de projets menés tant en Europe qu'aux USA, l'auteur a cherché à identifier les principales caractéristiques des réalisations publiées et à examiner leurs conséquences épistémologiques sur le travail quotidien des chercheurs.
La synthèse des leçons apprises a ensuite été appliquée à l'étude du site de la banque de données des inscriptions coptes documentaires. Utilisant les contacts qu'il a eu l'opportunité de se faire au cours d'une série de voyages à l'étranger et étudiant les principaux sites web du domaine, l'auteur chercha également à remettre l'information dans son contexte historique afin de souligner les dynamiques sous-tendant à l'évolution de la discipline.
A l'issue de ce travail, l'auteur à identifié quatre grandes conclusions liées respectivement à :

- La vie des projets
- L'accessibilité aux nouvelles technologies
- La qualité des données et des métadonnées
- La nature des projets et leur adéquation avec les attentes des utilisateurs

Remerciements

Il est coutume de remercier celles et ceux qui, par leur assistance pratique ou leur soutient moral, ont contribué à la rédaction d'un ouvrage. Dans le cas présent ce n'est cependant pas tant la coutume qui me pousse à écrire ces quelques lignes qu'un réel souhait de rendre hommage à des gens sans lesquels le travail que vous tenez entre les mains n'aurait tout simplement pas vu le jour.

J'adresse donc mes plus vifs remerciements

- à Alain Delattre (Université Libre de Bruxelles), dont une question innocente allait me plonger dans un univers nouveau
- aux participants à l'*Epidoc Summer School 2008* et en particulier Gabriel Bodard (*King's College London Centre for Computing in Humanities*) et Paul Iversen (*Case Western Reserve University*) pour nos discussions sur EpiDoc, le TLG, les Searchable Greek Inscriptions, la DDbDP et les projets d'inscriptions d'Aphrodisias
- aux participants du colloque « Les historiens et l'informatique : un métier à réinventer » et en particulier à J.-P. Genet (Université Paris I Sorbonne), A. Ghignoli (Universita di Firenze), Ph. Rygiel (Université Paris I Sorbonne) et S. Noiret (European University Institute, Firenze) pour nos discussions sur les grandes évolutions de la discipline
- aux participants à l'Oxford TEI Summer School 2009 et en particulier Lou Burnard (Oxford University Computing Services), Sebastian Rahtz (Oxford University Computing Services) et Edouard Chiricat (projet Lexicon of Greek Personal Names) pour nos discussions sur la TEI et son histoire ainsi que sur vos différents projets
- à Damien Tournay et Myriam Francis-Lemaire, relecteurs attitrés du présent mémoire
- à vous, lecteurs, qui faites que ce travail ne fut pas vain

Introduction

Les sciences de l'Antiquité ne semblent pas à première vue être le domaine dans lequel les nouvelles technologies d'information et de communication sont les plus actives. L'idée même d'associer l'Antiquité aux dernières innovations technologiques, de conjuguer des termes comme épigraphie et web 2.0 semble absurde aux yeux des non initiés.

L'image traditionnelle du chercheur plongé dans ses livres poussiéreux, compulsant de volumineux index et d'imposantes compilations de sources reste prédominante, renforcée en bonne part la formation telle qu'elle est encore dispensée dans nombre d'universités, en ce compris à l'ULB.

Pourtant le métier connait depuis de nombreuses années une véritable révolution liée à l'avènement de l'informatique. Analyse statistique automatisée de larges corpus, constitution automatique d'index, diffusion numérique de textes littéraires ou documentaires, travail collaboratif d'édition de sources conduit via internet, croisement de banques de données internationales sont des réalités auxquelles le chercheur contemporain est sans cesse confronté.

Le présent mémoire a pour objectif d'étudier la situation actuelle dans ce domaine et d'en tirer quelques leçons générales susceptibles d'assister les chercheurs souhaitant se lancer dans des projets de cette nature. Bien sûr le domaine des sciences de l'Antiquité est vaste, couvrant l'étude de domaines aussi différents que les œuvres d'art peintes, sculptées, gravées ou fondues, les textes conservés par la tradition manuscrite, les papyri ou l'épigraphie, les structures bâties, etc.

Archéologie, histoire, philosophie, philologie, il fallait donc faire un choix. Un choix qui s'est fait naturellement, dans la continuité d'une aventure commencée en 2004 et qui se poursuit aujourd'hui encore : celle de la Base de Donnée des Inscriptions Coptes Documentaires, aussi connue sous le nom de *Brussels Coptic Database*. Partie d'une simple question sur comment réaliser un site web, cette collaboration de maintenant 5 ans a débouché sur la mise en ligne d'une base de donnée reprenant les méta-informations disponibles sur 7090 textes et fragments de textes coptes. Ce travail de compilation de longue haleine fut entrepris au début des années 2000 par Alain Delattre, alors jeune doctorant, et se poursuit aujourd'hui encore, s'insérant dans une démarche plus large d'étude de l'Egypte ancienne.

Après une première phase durant laquelle le principal objectif fut de mettre en ligne les métadonnées présentes dans le fichier FileMaker Pro de A. Delattre, le souhait de mettre en ligne le texte même des documents et de permettre des échanges de données avec d'autres systèmes a entraîné une recherche quant à l'existence d'éventuels standards dans ce domaine. La dynamique devant conduire au mémoire que vous tenez entre vos mains était lancée.

Faire un mémoire sur l'histoire du seul projet du Pr. Delattre n'eut pas été suffisant pour répondre aux attentes d'un jury de mémoire et n'eut sans doute pas apporté grand-chose. Le choix fut donc fait très tôt d'élargir le champ de nos investigations aux domaines de l'épigraphie et de la papyrologie en général afin de voir comment les chercheurs ont aujourd'hui adapté l'outil informatique à leurs besoins spécifiques, la manière dont se sont développés les systèmes et les perspectives d'avenir.

Au-delà de l'aspect technique, ce mémoire va également se pencher sur des considérations plus épistémologiques : en quoi les nouveaux outils changent-ils la manière de travailler des chercheurs, en quoi influencent-ils éventuellement leur regard sur leur travail et sur leurs sources ? Quels sont les pièges à éviter, les risques d'erreurs induits pour le chercheur par l'utilisation de ces outils ?

Pour réaliser cette étude, différents types de sources ont été utilisés : la littérature scientifique, bien entendu, a servi de base première à la démarche. Néanmoins elle s'est rapidement révélée insuffisante pour notre propos et ce pour plusieurs raisons.

Tout d'abord le caractère périmé de nombre des rares ouvrages publiés dans le domaine. Mais aussi parce que, dans ce domaine en constante et rapide évolution, les contacts personnels jouent un rôle essentiel.

Beaucoup de projets sont nés, ont débouché (ou non…) sur des résultats, et n'ont jamais rien publié sur leur méthodologie de fonctionnement, leurs choix technologiques, leurs échecs ou leurs réussites, leurs initiateurs considérant que le seul délivrable de leur projet était la publication des textes anciens. Dans ce contexte les conférences, colloques et autres formations jouent donc un rôle essentiel.

C'est donc en nous rendant à des formations dispensées au *King's College Centre for Computing in the Humanities* et à l'*Oxford University Computing Services* ainsi qu'au colloque « Les historiens et l'informatique : un métier à réinventer », organisé en décembre dernier à Rome par l'Ecole Française de Rome, *l'Istituto Storico Italiano per il Medio Evo*,

Reti Medievali et le *LAMOP* de la Sorbonne, que nous avons eu l'opportunité de rencontrer de nombreux chercheurs dont les expériences et les réflexions ont joué un rôle important dans la maturation du présent mémoire. Citons ainsi à titre d'exemple les professeurs Jean-Philippe Genet et Lou Burnard, deux des pères fondateurs du formalisme XML d'édition et de publication de textes TEI, formalisme dont nous reparlerons ultérieurement. Les entretiens qu'ils m'ont accordés nous ont ainsi été d'un grand secours.

Ce mémoire se structure en cinq parties. La première, assez brève, est plus historique et évoquera la manière dont s'est structurée la recherche en épigraphie et en papyrologie, des origines de ces disciplines jusqu'aux premiers efforts de digitalisation et de numérisation, l'ensemble étant replacé dans le cadre plus large de l'évolution de la philologie. La seconde posera quelques bases théoriques sur la technologie la plus utilisée à l'heure actuelle par les chercheurs, fondée sur le langage de programmation XML et sur le formalisme de publication électronique de documents TEI et ses variantes, dont nous étudierons plus en détail la version baptisée EpiDoc.

Viendra ensuite un examen de la mise en œuvre pratique de ces technologies dans divers projets internationaux. Au travers de ces exemples seront étudiées les possibilités de synergies et d'échanges offertes par ces technologies, montrant comment se construit une nouvelle manière de travailler aux chercheurs. Ces nouvelles méthodes seront analysées et critiquées dans la quatrième partie du présent mémoire, une analyse qui nous fournira une grille sur base de laquelle nous étudierons le projet de la Base de Donnée des Inscriptions Coptes Documentaires du Pr. Delattre et proposerons des pistes pour son avenir avant de tirer des conclusions plus générales susceptibles d'aider les initiateurs d'autres projets.

Chapitre 1 : Brève introduction à l'épigraphie et à la papyrologie

Epigraphie, l'univers éclaté

Les inscriptions grecques et latines ont toujours fait partie de la vie quotidienne des européens méridionaux. Les vestiges laissés par Rome dans ses provinces étaient bien souvent couverts d'inscriptions qui ne pouvaient manquer d'attirer les regards des curieux.

La Renaissance, avide de connaissances sur les lointains peuples d'avant le Moyen-âge, vit fleurir les premiers florilèges d'inscriptions à l'image des recueils de Pogge ou de Cyriaque d'Ancône. L'arrivée de l'imprimerie accélérera le mouvement comme en témoigne pour nos régions l'*Inscriptionum antiquarum liber* du flamand Martin de Smedt, publié en 1588 par Juste-Lipse : ce seul ouvrage reprenait déjà 4000 inscriptions.

C'est au début du XVIIème siècle que les idées de grands corpus apparaissent mais c'est le dix-neuvième siècle qui verra leurs réalisations. L'un des premiers est sans aucun doute le *Corpus Inscriptionum Graecae* (CIG) de Boeckh, publié à Berlin à partir de 1815 et se basant sur les meilleurs recueils disponibles. Les allemands s'affirment déjà comme des champions de la discipline : en 1840 apparaissent les *Elementa Epigraphices Graecae* rédigés par Joh. Franz qui livre avec cet ouvrage un véritable manuel d'épigraphiste.

D'autres pays essayent également de produire des corpus épigraphiques, notamment la France de Napoléon III, dans le cadre de la rivalité régnant entre les nations d'Europe à cette époque.

La Commission Epigraphique de France va ainsi concevoir un plan général pour une publication reprenant toutes les inscriptions latines connues mais le manque de spécialistes fait que ce seront les Allemands qui mettront en chantier le Corpus Inscriptionum Latinarum (CIL), un ouvrage qui reste aujourd'hui encore la référence principale.

A la même époque le CIG de Boeckh, pourtant pas encore entièrement publié, apparait comme nécessitant un remplaçant en raison du nombre très important de nouvelles inscriptions découvertes depuis l'indépendance de la Grèce (1830).

Les grandes nations ont en effet toutes obtenu des chantiers de fouille dans le pays nouvellement indépendant et les découvertes majeures se succèdent sur l'Acropole, à Delphes, Delos, Olympie, Dodonne et bien d'autres lieux encore.

Les épigraphistes vont publier dans des collections spécifiques à chaque site, grands ouvrages participant à la lutte pour le prestige scientifique des nations. La guerre franco-prussienne de 1870 exacerbe encore les tensions entre Français et Allemands.

La structure de classement des inscriptions épigraphiques est donc avant tout géographique et liée à l'ordre de découverte des documents. Les pratiques éditoriales sont par ailleurs très diversifiées, chaque ouvrage ou série ayant son propre code.

Ce n'est qu'en 1932 qu'une réunion est organisée à Leyden pour définir un système unifié de représentation graphique des éléments relatifs à l'état des textes épigraphiques et papyrologiques, une décision résumée dans plusieurs articles[1].

Ce travail est nécessaire car les documents épigraphiques sont plus fréquemment endommagés que les manuscrits contenant les grands textes littéraires édités depuis la Renaissance : pierres brisées, inscriptions rédigées sur plusieurs pierres dont certaines sont perdues, etc.

On peut résumer les conventions adoptées à Leyden par la liste suivante :

[...] lacune d'une taille déterminée indiquée par le nombre de points

[— — —] lacune d'une taille indéterminée

[abc] lettres disparues mais restaurées par l'éditeur

a(bc) abréviation et sa résolution marquée entre parenthèses

<ab> faute ancienne corrigée par l'éditeur

{ab} lettres considérées comme surnuméraires par l'éditeur

ạḅ lettres endommagées dont la lecture n'est permise que par le contexte

... traces de lettres indéchiffrables (en usage en épigraphie grecque et en papyrologie)

[1] van Groningen B.A., « De signis criticis in edendo adhibendis », *Menemosyne* 59 (1932), pp. 362-365 ; Hunt A.S., « A note on the transliteration of papyri », *Chronique d'Égypte* 7 (1932), pp. 272-274

+++ traces de lettres (en usage en épigraphie latine)

ABC lettres lisibles mais ne pouvant être interprétées

[[abc]] lettres effacées

vac. espace laissé vide par le graveur/rédacteur (*vacat*)

Ces conventions facilitent la lecture des éditions de textes et n'utilisent que des caractères faciles à imprimer, raison pour laquelle elles ont été choisies. Elles furent néanmoins légèrement remaniées par H. Krummrey et S. Panciera en 1980[2]. C'est dans cette version de 1980 qu'elles connaîtront quelques 20 ans plus tard leur passage au numérique dans le formalisme EpiDoc.

De tout cela on retiendra surtout la difficulté historique qu'ont les épigraphistes à travailler ensemble : encore aujourd'hui les disputes entre épigraphistes sont plus fréquentes et plus bruyantes que celles entre papyrologues…

[2] Krummrey H., Panciera S., « Criteri di edizione e segni diacritici », *Tituli* 2 (1980), pp. 205-215.

La papyrologie, domaine de la collaboration amicale

L'histoire de la papyrologie est très différente de celle de l'épigraphie. Il faut dire que la discipline est beaucoup plus récente, sa date de fondation étant souvent rattachée à celle de la publication de la Constitution des Athéniens du pseudo-Aristote, un papyrus découvert en 1879 et publié en 1891, deux ans après son acquisition par le British Museum.

Bien entendu ce n'était pas le premier papyrus connu en Europe, des découvertes plus anciennes ayant déjà eu lieu, en ce compris celles de la Villa des Papyri à Herculanum où une série de papyri calcinés avait été découverte au dix-huitième siècle.

Les circonstances dans lesquelles les papyri avaient été extraits du sol d'Egypte durant les dix-huitièmes et dix-neuvièmes siècles furent loin d'être idéales : les paysans égyptiens vendaient des kilos de papyri, tranchés à la pelle pour en égaliser le poids, à des marchands qui n'hésitaient pas à déchirer leurs marchandises pour en tirer plus d'argent : il en résultat une dispersion phénoménale des collections

Ce n'est que dans les années 1890-1900 avec les fouilles de W. Flinders Petrie que des fouilles plus scientifiques furent menées. C'est aussi à cette période que sont inventés les premiers outils bibliographiques et publiés les premiers recueils, principalement en Allemagne sous l'impulsion de pères fondateurs comme U. Wilcken.

La première guerre mondiale ralentit fortement les études papyrologiques mais celles-ci redémarrent dés le retour de la paix. En 1930 une initiative belge conduit à la fondation d'un Comité international de Papyrologie basé à Bruxelles. En 1933 apparait l'expression d'Amicitia Papyrologorum, appelée à faire fortune. Ce n'est pas qu'une simple formule creuse comme le prouve dés 1948 la reprise des relations internationales, y compris avec les scientifiques allemands dont le comportement n'avait pas forcément toujours été des plus recommandable.

En réalité cet esprit de collaboration est avant tout dû à la nécessité pour les chercheurs d'échanger entre eux afin de donner du sens aux fragments de papyrus conservés dans les musées du monde entier. Comment en effet faire autrement quand un même document à pû être vendu, fragmenté, à quatre ou cinq institutions ?

L'entre deux-guerres a également vu la papyrologie se doter d'autres outils de la plus grande importance, notamment trois séries lancées par Fr. Preisigke, un ancien responsable du service des postes allemandes qui se lança en amateur dans la papyrologie[3].

La *Berichtigungsliste der griechischen Papyrusurkunden aus Ägypten*, recueil de toutes les corrections apportées aux éditions antérieures de papyri , le *Sammelbuch Griechisher Urkundend aus Ägypten* liste de tous les papyri publiés hors des grandes collections et le *Namenbuch*, index de tous les noms de personnes identifiés dans le matériel grec d'Egypte.

Des outils exceptionnels nés de l'esprit d'un homme habitué à classer minutieusement des montagnes de courrier qui devaient rester sans équivalent dans les autres disciplines, des outils qui vont aussi contribuer à renforcer l'*amicitia papirologorum* par le degré de collaboration qu'ils imposent aux chercheurs.

Une situation très différente des rivalités existantes dans le domaine de l'épigraphie…

[3] On lira sa biographie dans H.I. Bell, « Friedrich Preisgke » in The Journal of Egyptian Archeology vol.10 n°2, juillet 1924, pp. 172-173

Chapitre 2 : XML, TEI, EpiDoc et les autres, tour d'horizon technologique

Après avoir posé quelques bases historiques sur les domaines de l'épigraphie et de la papyrologie avant tout destinées aux non-spécialistes de ces matières, il nous a semblé important de procéder de la même manière en replaçant dans leur contexte historique les grands concepts techniques auxquels nous ferons appel tout au long du présent mémoire afin que des lecteurs moins au fait des questions informatiques puissent tirer profit de sa lecture.

Stocker l'information

L'information digitale est stockée dans des fichiers pouvant avoir été conçus dans différents formats et susceptibles d'être ouverts dans des logiciels aux fonctionnalités très différentes. Cette évidence doit cependant conduire le concepteur d'un projet de digitalisation à une réflexion à-priori sur la manière dont il veut exploiter l'information encodée et sur les ressources et compétences nécessaires à la bonne marche du projet.

Ce choix est évidemment conditionné à l'état de la technologie au moment du lancement du projet et l'initiateur de ce dernier doit garder à l'esprit que les progrès de l'informatique auront certainement des implications lourdes sur le cycle de vie de son projet : nous examinerons d'ailleurs plusieurs projets où cette problématique est particulièrement visible.

Cependant l'on peut dire que deux grandes options existent aujourd'hui dans le domaine de la digitalisation de documents épigraphiques et papyrologiques :

- les bases de données (généralement du type relationnel) gérées à l'aide d'un logiciel appelé *Database Management System (DBMS)*, à l'image de MySQL ou de Oracle
- les systèmes basés sur le langage XML.

A vrai dire les deux solutions ne sont pas mutuellement exclusives puisque de l'information encodée en XML peut être stockée dans une base de données. Nous nous concentrerons cependant sur les formalismes d'encodage d'information TEI et EpiDoc, des dérivés d'XML qui sont aujourd'hui utilisés par certains des plus importants projets de digitalisation en cours dans le domaine des sciences de l'Antiquité.

La question du stockage n'est pas anodine dans un projet car elle peut avoir un impact direct sur les performances d'un système. Les DBMS et XML ont chacun leurs points forts et s'avèrent plus efficaces pour certains types d'applications et le choix de l'une ou l'autre solution doit être basé sur une étude préalable du type de données que l'on souhaite encoder et des traitements que l'on souhaite leur faire subir.

La taille du projet peut également avoir un impact sur l'architecture à adopter : dans la demande de fond pour le projet *Integrating Digital Papyrology 2* Roger Bagnall et ses collègues indiquent que le nombre d'inscriptions grecques et latines publiées sur papier se situe autour de 800 000[4] : un tel nombre d'entrées nécessite une bonne préparation en amont de l'architecture si l'on ne veut pas se retrouver avec un système inopérant, un travail qui doit se faire bien avant d'atteindre ce seuil comme le montre l'attention portée à ces questions par l'équipe de la DDbDP, dont la base de données ne comprend pourtant que de 75 à 80 000 entrées.

Par ailleurs dans le cas où un projet fait le choix d'encoder ses données en XML il n'est pas exclu que le système d'interrogation et d'exploitation des données comprenne des index générés automatiquement par l'application et stockés par un DBMS.

Ces questions d'architecture, très complexes, ne seront pas approfondies ici mais doivent être à l'esprit du papyrologue ou de l'épigraphiste souhaitant se lancer dans un projet de digitalisation ou de numérisation afin qu'il n'oublie pas de l'évoquer avec son conseiller technique ou l(es) informaticien(s) avec le(s)quel(s) il collabore.

Bagnall R. et all., [4] http://www.duke.edu/~jds15/IDP2-FinalProposalRedacted.pdf, 2008, p.7

La problématique de l'encodage des caractères non latins

L'entrée dans la base de données gérée par le DBMS ou le fichier XML formeront le support de stockage mais, de la même manière qu'un mot latin est parfois écrit en lettres grecques sur une stèle, un papyrus ou un ostracon, l'alphabet dans lequel l'entrée est rédigée peut également être une source de problèmes.

C'est particulièrement vrai dans le cadre du domaine qui nous intéresse car les études classiques concernent une série de langues qui n'utilisent pas l'alphabet latin à la base des premiers standards de représentation graphique de caractères.

Le grec n'était ainsi pas représenté dans le standard ASCII (*American Standard Code for Information Interchange*), un ensemble de 128 caractères représentés par des combinaisons de sept chiffres codés en binaire, l'une des premières normes d'impression de caractères.

Cet ensemble très limité inventé en 1961 par Bob Berner comprenait les chiffres arabes, les lettres latines majuscules et minuscules (non accentuées), quelques signes basiques de ponctuation et des règles pour communiquer avec les terminaux. En réalité seules 96 des 128 combinaisons représentaient effectivement des chiffres, des lettres ou des signes de ponctuation.

Cet ensemble limité fut rapidement élargi et déboucha sur deux normes : ISO-8859[5] (qui introduisit l'encodage de l'information relative aux caractères sur 8 bits, première publication en 1987) puis, quelques années plus tard, UTF-8[6] (Unicode et ISO/CEI-10646, première publication en 1991)[7]. Ces normes définissent la combinaison binaire correspondant à un caractère spécifique et permettent la création de polices de caractères, tableau corrélant ces combinaisons binaires avec des représentations graphiques susceptibles d'être exploitées par les systèmes d'impression[8].

Cependant les projets de digitalisation de sources grecques commencèrent bien avant que ces nouvelles normes ne soient mises en œuvre. Il fallut donc utiliser diverses astuces pour

[5] http://anubis.dkuug.dk/JTC1/SC2/WG3/docs/n411.pdf
[6] Il faut noter qu'UTF-8 est en réalité un sous-ensemble de la norme Unicode qui comprend également la capacité d'enregistrer un caractère sur 16 ou 32 bits (UTF-16, UTF-32)
[7] Pour plus de détails sur ces questions on se réfèrera notamment à Andries. P., *Unicode 5.0 en pratique*, Paris, 2008
[8] Système d'impression désigne ici tant les écrans que les imprimantes

contourner le problème. L'une dont l'usage se révéla le plus courant fut le *Beta Code*, mit au point fin des années 70 par le Dr. David W. Packard et notamment utilisée par le projet du *Thesaurus Linguae Graecae* (TLG), ce qui assura sa diffusion. Une combinaison de lettres et de chiffres était définie pour chaque lettre de l'alphabet grec[9] dont le glyphe n'était de ce fait pas affiché.

Cette solution, quoique pratique, était néanmoins assez limitée en ce sens qu'elle ne permettait pas de rendre l'accentuation ou de marquer l'esprit, contrairement à la pratique usuelle. Il faut cependant noter que ce système est encore fortement utilisé de nos jours malgré ses limites car il reste d'une grande simplicité.

L'arrivée de la norme ISO-8859 changea la situation, permettant non seulement d'encoder directement les caractères grecs mais aussi d'y ajouter un certain nombre de signes diacritiques : esprits, accents et plus généralement divers caractères spéciaux pouvaient désormais être directement encodés. Quelques combinaisons furent laissées libres pour que les utilisateurs avancés puissent s'en servir pour des usages qui leurs seraient propres.

Cependant ce progrès entraîna immédiatement un nouveau problème : les lettres imprimées sur les claviers des utilisateurs ne correspondaient plus aux lettres encodées lors de la frappe de la touche. Deux solutions s'offrirent alors aux chercheurs : apprendre par cœur la configuration du nouveau clavier ou utiliser un petit logiciel permettant soit de reprogrammer son clavier (pour accéder plus facilement à certains caractères) soit de cliquer à l'écran sur les lettres souhaitées (un processus laborieux et peu ergonomique…).

En outre, la liberté laissée aux créateurs de polices de caractères d'associer certains codes à des représentations graphiques de leur choix rendit l'interopérabilité entre polices plus qu'aléatoire.

Ces problèmes s'amplifièrent avec l'arrivée de l'Unicode et ses dizaines de milliers de combinaisons binaires car ce qui était sous ISO-8859 des combinaisons de lettres et de signes diacritiques fut transformé dans Unicode en glyphes disposant de leurs combinaisons individuelles. Si cela simplifiait effectivement l'encodage de certaines combinaisons, cela élargissait aussi le nombre de codes à connaître.

[9] Cf. Annexe 1

En outre les innovations contenues dans Unicode le rendirent partiellement incompatible avec la norme ISO-8859, ce qui cause toujours de fréquents problèmes aux informaticiens travaillant dans le domaine du web. En effet, si l'usage de Unicode est recommandé depuis plus de dix ans il n'en reste pas moins que nombre de sites sont encore codés en ISO-8859.

Par ailleurs le concepteur web ne peut avoir la certitude que l'utilisateur dispose bien de la bonne police de caractères permettant d'afficher la représentation graphique du caractère encodé, particulièrement pour les langues moins fréquentes comme le copte…

Il faut noter que l'évolution que nous venons de retracer ici est avant tout vraie pour les langues d'Europe d'Ouest et d'Afrique : l'Asie et l'ancienne URSS avaient en effet adopté leurs propres systèmes et ce n'est qu'avec l'arrivée de l'Unicode que tous les caractères employés dans toutes les langues ont enfin pu être représentées dans un système unique d'encodage.

Les combinaisons non utilisées par les langues pratiquées sur Terre ont même été accaparées par des amateurs de langues imaginaires comme le Klingon (issu des films Star Trek) ou le Cirth et le Tengwar, alphabets elfiques inventés par J.R.R. Tolkien pour son univers du Seigneur des Anneaux.

eXtended Markup Language, la souplesse dans la rigueur

A la fin des années 60 IBM décide, sous l'impulsion de Charles Goldfarb, d'adopter un langage descriptif appelé *Generalized Markup Language*. Ce projet était issu de recherches fondamentales menées dés 1967 par William Tunnicliffe, lequel avait proposé de séparer l'encodage du fond du document de celui de sa forme.

Goldfarb, assisté par E. Mosher et R. Lorie, reprit le concept de Tunnicliffe dans le but d'offrir aux juristes un outil de recherche performant capable d'extraire les textes légaux correspondants à leurs requêtes. IBM comprit très vite l'utilité du système pour gérer des ensembles documentaires et décida de commercialiser un logiciel baptisé *Document Composition Facility*.

Il fallu près de dix ans de recherche à Goldfarb et son équipe pour concevoir ce produit, ancêtre de plusieurs langages dont notamment SGML, XML et HTML.

Le principe de GML était de marquer dans le texte la structure de ce dernier à l'aide d'une série de balises : chapitres, titres, paragraphes, tableaux, listes étaient ainsi identifiés dans le texte et une série de macro était ensuite appliquée au document lors de son impression. Ces macro étaient enregistrées dans une bibliothèque qui était appelée à chaque impression du document afin de produire la mise en page recherchée[10].

Courant des années 80 Goldfarb décide de quitter IBM pour lancer son propre produit. Ce sera SGML, qui apparait en 1986 et connait très vite un succès considérable grâce notamment à un support important de la Communauté Européenne qui cherche alors un standard pour la masse considérable de documents qu'elle produit annuellement.

SGML se distingue notamment de GML par la manière dont sont stockées les informations relatives à la forme du document. Les concepts de Définition de Type de Document (DTD) et de feuille de style font leur apparition dans SGML.

Le DTD sert à définir la structure des documents en décrivant les relations entre chacun des éléments constitutifs du texte et en permettant de nommer chaque élément par rapport à la structure générale tandis que la feuille de style à pour objectif de décrire toutes les opérations

[10] Pour plus de détails on se référera au manuel de GML : http://publibfp.boulder.ibm.com/cgi-bin/bookmgr/BOOKS/dsm04m00/CCONTENTS

nécessaires pour assurer une impression correcte du document. Ces feuilles de style permettent ainsi de fixer les règles d'affichage pour une configuration donnée ou pour exporter l'information dans un nouveau format.

Une partie des concepts de SGML sont repris par Tim Berner-Lee fin des années 80 pour créer un nouveau langage appelé HTML, une version très simplifiée de SGML mais offrant la possibilité de créer des hyperliens : c'était là l'acte fondateur de ce que nous connaissons aujourd'hui sous le nom de *World Wide Web*.

Mais c'est là une autre histoire, revenons donc à SGML. Le langage SGML est riche et complexe, sans doute un peu trop d'ailleurs puisque quelques 10 ans après son introduction va débuter un projet de simplification du format destiné notamment à faciliter son utilisation sur le jeune web comme l'indique le synopsis de la 1ère recommandation XML :

« *The Extensible Markup Language (XML) is a subset of SGML that is completely described in this document. Its goal is to enable generic SGML to be served, received, and processed on the Web in the way that is now possible with HTML. XML has been designed for ease of implementation and for interoperability with both SGML and HTML.*"[11].

Abandonnant un certain nombre des contraintes ou de limitations fixées par SGML, les pères fondateurs du XML se concentrent sur trois objectifs majeurs :

- Le passage à Unicode : SGML avait été créé aux USA dans un environnement ou ASCII était la norme, sa conception remontant à avant la première publication de la norme ISO-8859. De ce fait SGML était d'une utilité quelque peu limitée en dehors du monde anglo-saxon. Aussi lorsqu'apparait, en 1996, la seconde version de la norme Unicode, l'équipe en charge de la définition de XML décide d'en faire son standard afin de permettre un usage universel.

- Une délimitation explicite du contenu de chaque balise : la puissance informatique à l'époque du développement de SGML était infiniment plus restreinte que celle disponible au milieu des années 90 : les Amiga, Atari, Commodores 64 et autres Appel Macintosh ont en effet cédé la place aux premières générations de PC équipés de processeurs 3.86, 4.86 et même, à partir de 1994, Intel Pentium. Aussi certaines des limitations de SGML purent être levées, notamment les règles assez souples de

[11] http://www.w3.org/TR/1998/REC-xml-19980210

fermeture des balises, furent supprimées afin de permettre un meilleur contrôle des résultats tout en simplifiant la norme : là où certaines balises devaient être clôturées et d'autres laissées ouvertes il faudrait désormais toujours fermer ses balises afin de fournir une information sans ambiguïté.

- Une plus grande souplesse : SGML était fondé autour du concept de validation du document par rapport au DTD associé. Pour rendre XML plus souple que SGML il fut donc décidé de lui ajouter la possibilité de travailler avec plusieurs vocabulaires de balises (via plusieurs DTD) dans un même document.

En conséquence une liste de dix objectifs est formulée, objectifs qui seront repris dans le chapitre 1.1 de la définition d'XML[12] :

1. *XML shall be straightforwardly usable over the Internet.*
2. *XML shall support a wide variety of applications.*
3. *XML shall be compatible with SGML.*
4. *It shall be easy to write programs which process XML documents.*
5. *The number of optional features in XML is to be kept to the absolute minimum, ideally zero.*
6. *XML documents should be human-legible and reasonably clear.*
7. *The XML design should be prepared quickly.*
8. *The design of XML shall be formal and concise.*
9. *XML documents shall be easy to create.*
10. *Terseness in XML markup is of minimal importance.*

C'est en 1998, deux ans après le lancement du projet, que la définition du XML est publiée par le consortium W3C. La vitesse du projet tient notamment dans le fait que XML est plus un sous-ensemble de SGML qu'une création ex-nihilo.

Le succès sera rapide : comme on le verra dans la suite de ce mémoire, de nombreux projets vont évoluer de SGML vers XML et le langage sera utilisé pour créer de nombreux formats dérivés comme les flux RSS et ATOM ou les formalismes Dublin Core, RDF ou TEI dont nous reparlerons dans quelques instants.

[12] http://www.w3.org/TR/2008/REC-xml-20081126/

Chaque balise XML présente trois caractéristiques : l'élément, le (ou les) attributs de l'élément et le contenu de l'élément. L'élément est simplement le nom que l'on donne à la balise, chaque utilisation de la balise devenant une instance de l'élément.

L'attribut est une donnée fournissant une ou des précision(s) sur l'élément : il peut s'agir d'un identifiant (pour différencier les instances de l'élément), d'une métadonnée précisant la manière d'interpréter le contenu de la balise (ainsi en XML-TEI <measure type="cm">25</measure> indique que la valeur numérique 25 représente une mesure exprimée en centimètres) : l'attribut permet d'ajouter du sens d'une manière que peut interpréter l'ordinateur.

Ainsi dans l'exemple donné la présence de l'élément <measure></measure> autour de toute mesure identifiée dans un texte permet d'effectuer une recherche automatique destinée à récupérer toutes les mesures exprimées dans ce texte, ce qui serait impossible sans ce marquage.

Le contenu est tout ce qui se trouve entre les balises d'ouverture et de clôture de l'élément : cela peut être une valeur spécifique, un texte entier, etc.

Une des grandes forces de XML est de pouvoir servir de métalangage pour définir d'autres langages : il est autodescriptif et extensible, l'utilisateur pouvant créer n'importe quelle balise tant qu'il respecte les règles de base de langage (on parle de *document bien formé*) et recommande la validation du document à partir d'un DTD (on parle alors de *document valide*).

Si l'on ne doit retenir que deux règles de XML alors on gardera à l'esprit l'obligation de toujours fermer les balises et de veiller à ce qu'aucun chevauchement de balises ne se produise. Les autres règles spécifient que tout document XML doit être inclus dans un élément unique, appelé racine, que les noms des éléments sont toujours sensibles à la casse et que la valeur des attributs doit toujours se retrouver entre guillemets : ces cinq règles suffisent à faire d'un document XML un document bien formé.

Un petit tableau permet d'illustrer facilement certaines des erreurs les plus courantes rendant un fichier mal formé :

XML bien formé	XML incorrect	Explication
<seg>some text</seg>		Un simple élément
<seg><foo>some</foo> <bar>text</bar></seg>	<seg><foo>some <bar></foo> text</bar></seg>	Chevauchement de balises
<seg type="text">some text</seg>	<seg type=text>some text</seg>	Oubli des guillemets autour de l'attribut
	<seg type="text">some text<seg/>	Mauvais emplacement du / de clôture
<seg type='text'>some text</seg>		Toute forme de guillemets peut faire l'affaire
	<seg type="text">some text< /seg>	Espace entre < et /
	<seg type="text">some text</Seg>	Le nom des éléments est sensible à la casse
<seg type="text">some text<gap/></seg>		<gap/> est un élément auto-fermant car il n'a aucun contenu (il marque simplement la présence d'un espace vide)

Ces règles permettent notamment de donner une structure hiérarchique claire au document, raison pour laquelle le code est souvent présenté de manière indentée, facilitant pour un lecteur humain la compréhension de la structure du document. On utilise d'ailleurs souvent la métaphore de l'arbre pour décrire la hiérarchie de XML, avec un tronc central unique dont partent de nombreuses branches.

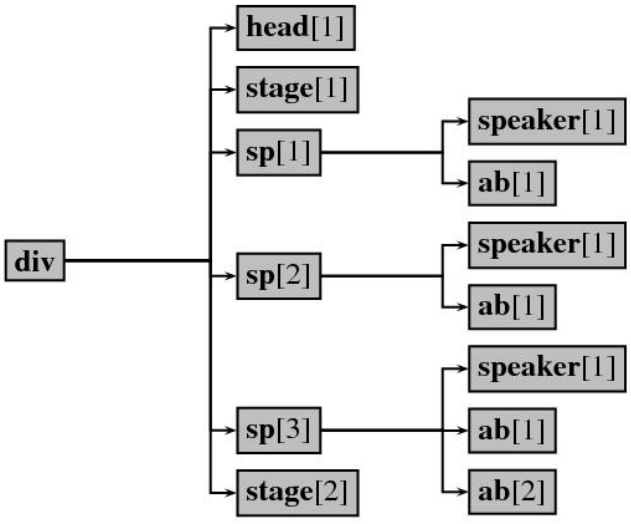
Un autre avantage de XML est qu'il est très facile à transmettre et peut donc servir à des échanges de données automatisés. Cela nécessite cependant de le déclarer en début de fichier afin que l'ordinateur interprète bien le contenu comme étant de l'XML. Cette déclaration fournit la version d'XML utilisée et indique à l'ordinateur quelle variante d'Unicode est utilisée pour l'encodage du document : le plus souvent il s'agira d'UTF-8.

Un exemple pouvant ici encore s'avérer bien plus clair qu'un long discours, voici la retranscription d'une petite scène de théâtre marquée à l'aide d'un extrait simplifié de la norme TEI P5 :

```
<?xml version="1.0" encoding="utf-8" ?>
<div n="1">
        <head>SCENE I. Un étudiant présente son mémoire</head>
        <stage>L'étudiant entre dans la salle de défense</stage>
        <sp>
                <speaker>Professeur</speaker>
                <ab>Ah voici l'étudiant Lemaire !</ab>
        </sp>
        <sp>
                <speaker>Etudiant Lemaire</speaker>
                <ab>Bonjour professeur</ab>
        </sp>
        <sp>
                <speaker>Professeur</speaker>
                <ab>Prêt pour votre défense ? </ab>
                <ab>Car, je dois vous le dire, nous avons de nombreuses remarques à vous
                faire !</ab>
        </sp>
        <stage>La marche funèbre de Chopin se fait entendre</stage>
</div>
```

On remarque la déclaration XML à la première ligne puis une succession d'éléments qui sont tous regroupés au sein de l'élément racine . Ce bref extrait est donc un document XML bien formé et sera reconnu comme tel par un ordinateur.

On notera que pour la commodité de la lecture humaine les balises sont indentées et colorées afin de mettre en lumière la structure du document. Cela cependant n'a aucune influence pour l'ordinateur. On peut également représenter l'information sous forme de schéma pour avoir une meilleure représentation de la structure :



Il faut toutefois noter que certains des atouts d'XML forment l'un de ses principaux handicaps : le balisage des éléments augmente considérablement le poids des fichiers car augmentant d'autant le nombre de caractères contenus dans le fichier.

Cela a une incidence non seulement sur le stockage (quantité de données à stocker) mais cela peut aussi influer sur la vitesse de transfert des données, avec des implications en termes de réactivité des applications basées en ligne et de coûts de bande passante, une considération à garder à l'esprit pour les programmes de plus grande ampleur.

Néanmoins on peut dire au final que XML s'avère un langage aux règles strictes mais permettant une grande souplesse d'utilisation.

Text Encoding Initiative, un projet de tous pour tous[13]

En 1987 un groupe de chercheurs parmi lesquels Lou Burnard et Jean-Philippe Genet se réunit pour chercher à résoudre le challenge représenté par la disparité des formes d'encodage de documents, disparité qui causait de nombreuses incompatibilités et d'importantes pertes de temps, chaque projet semblant créer sa propre infrastructure, le résultat étant souvent incomplet, comprenant des bugs, d'un réemploi ardu et très difficile à maintenir.

Cette situation était un véritable frein à l'évolution de la discipline et était due tant à des obstacles technologiques qu'à l'imitation des pratiques courantes dans le monde informatique à l'époque : les systèmes étaient alors rarement interopérables et le concept de transfert de données en réseau loin d'être aussi courant que de nos jours.

A titre d'exemple rappelons que l'université d'Oxford ne mis en place son premier service interne d'emails et de transfert bidirectionnel de fichiers qu'en 1981[14], profitant du premier projet national britannique de standardisation des réseaux informatiques académiques JANET pour développer son infrastructure.

Un des problèmes causant les incompatibilités était la volonté (existant encore aujourd'hui…) des entreprises de fournir des logiciels utilisant des formats de fichiers propriétaires afin de garder captive leur clientèle.

C'est en novembre 1987 que se tint le colloque de Vassar College (USA) durant lequel les bases intellectuelles devant conduire à la TEI furent posées. Un élément important du projet fut la diversité d'origine de ses premiers contributeurs : spécialistes des bibliothèques ou de l'archivistique, représentants des principales compagnies privées et de nombreux projets en cours des deux côtés de l'Atlantique purent en effet donner leur opinion et veiller à ce que le produit final corresponde à leurs besoins.

Une première proposition de norme (TEI P1) fut proposée en 1990, fruit du travail de plus de 50 chercheurs. Elle fut rapidement suivie des versions P2 et P3. C'est cette dernière qui fut, en 1994, la première version « officielle » recommandée pour une utilisation pratique: basée sur le standard SGML, elle offrait un large champ de possibilités et fut rapidement adoptée par

[13] Il est suggéré de procéder à la lecture de ce point en écoutant les chansons consacrées à la TEI P3 téléchargeables à l'adresse http://www.tei-c.org.uk/Vault/Songs/

[14] http://www.oucs.ox.ac.uk/internal/history/

différents projets. Néanmoins la P3 comprenait encore un certain nombre de défauts et ne couvrait pas l'ensemble des cas de figures pouvant se présenter lors de la digitalisation d'un document.

De plus la diffusion de plus en plus large du formalisme TEI assura un retour toujours plus important de commentaires, remarques et suggestions venues des quatre coins du monde : fin des années 90 près de 200 chercheurs issus de toutes les disciplines contribuaient activement au projet, nécessitant une nouvelle structure de gestion du projet

1999 vit donc la naissance du Consortium TEI, dont les objectifs étaient de fournir une infrastructure stable au projet, d'en assurer une gestion transparente et démocratique et de contribuer à la création d'une large communauté d'utilisateurs pour en assurer la maintenance sur le long terme.

Pendant ce temps le travail sur la norme elle-même continuait avec pour objectif la transition de SGML à XML afin de profiter du support d'Unicode et de faciliter le portage sur internet. La version P4 de la TEI fut publiée en juin 2002 et fut essentiellement une conversion simple de la P3, l'intégration de nouvelles fonctionnalités étant repoussée à la version P5.

Le succès de la TEI continua de croître, la norme étant adoptée par diverses institutions gouvernementales ou scientifiques. La possibilité d'échanges avec certains des principaux standards d'échange d'informations bibliographiques (MARC, Dublin Core, …) est une de ses grandes forces, tout comme sa modularité.

En effet la TEI avait surtout été conçue à l'origine pour décrire des textes imprimés et ne comprenait pas les éléments spécialisés nécessaires pour l'encodage de manuscrits, de papyri ou d'inscriptions épigraphiques.

Des projets parallèles au projet TEI comme le projet MASTER (*Manuscript Access through Standards for Electronic Records*)[15] lancé en 1996 et remplacé depuis 2007 par le projet ENRICH (*European Networking Resources Onformation Concerning Cultural Heritage*)[16] ou encore *EpiDoc*[17], lancé en 1999 (et dont nous reparlerons un peu plus loin) se penchèrent pour leur part sur ces besoins spécialisés et développèrent des modules additionnels venant

[15] http://www.tei-c.org.uk/Master/Reference/oldindex.html
[16] http://enrich.manuscriptorium.com/
[17] http://epidoc.sourceforge.net/index.shtml

s'ajouter aux recommandations de base de la TEI, certains d'entre elles étant directement intégrées comme éléments « canons » de la version P5 de la norme.

La norme P5 fut officiellement publiée en novembre 2007 et comprends près de 1500 pages de recommandations concernant plus de 500 éléments et leurs attributs que l'on retrouve sur le site de la TEI[18].

Ces éléments sont classés dans un certain nombre de modules dont on peut sélectionner lesquels sont nécessaires à son projet et produire le DTD approprié par simple sélection dans le logiciel ROMA[19] des modules et éléments que l'on souhaite conserver.

Cette solution extrêmement souple permet d'adapter la TEI à n'importe quel type de document écrit ou même cinématographique ou sonore : on pourrait même envisager d'utiliser la TEI pour effectuer la description d'objets même si là beaucoup d'éléments spécifiques seraient à rajouter.

[18] http://www.tei-c.org/Guidelines/P5/
[19] http://www.tei-c.org/Roma/

EpiDoc[20], de l'universel à la spécialisation

Nous l'avons vu, la norme TEI était initialement destinée à la digitalisation de documents imprimés et personne n'avait conçu de module spécifique pour la description de l'écriture manuscrite ou à fortiori des textes gravés.

C'est en 1999 que le Pr. Panciera prononce devant l'Association Internationale d'Epigraphie Grecque et Latine un discours qui sera suivi de la publication d'un manifeste[21] appelant à la création d'une base de données couvrant l'ensemble de l'épigraphie grecque et latine, ce qui va entrainer la naissance d'EpiDoc[22].

En effet T. Elliott, alors étudiant à l'Université de Chapel Hill, va entendre l'appel de Panciera et se mettre à la recherche d'une solution technique pour atteindre l'objectif fixé par ce dernier.

Très vite, il suggère d'adopter le modèle de la TEI qui sera adapté aux besoins spécifiques des épigraphistes, notamment par l'addition d'éléments permettant de marquer les signes diacritiques définis par la convention de Leyde.

Panciera ne suivra pas la proposition de T. Elliott et lancera le projet EAGLE, dont nous reparlerons ultérieurement. Cependant cela ne signifie pas la fin de l'adaptation de la TEI à l'épigraphie et une première version de la norme est publiée en janvier 2001.

Le document attira l'attention de plusieurs épigraphistes de renom dont Ch. Roueché qui cherchait alors une solution pour une publication électronique de son ouvrage *Aphrodisias in Late Antiquity*.

Elle décida d'adopter le nouveau formalisme et obtint un financement pour la réalisation d'un prototype : nous verrons le résultat de cet effort dans la suite du présent mémoire. Le projet fut rapidement mené à bien, entraînant une série de modifications de la norme pour mieux répondre aux besoins des chercheurs.

Certains projets comme celui des tablettes de Vindolanda ont également modifié le DTD tout en conservant une totale compatibilité avec le standard. EpiDoc était lancé.

[20] http://epidoc.sourceforge.net/
[21] http://www.edr-edr.it/Documenti/Document1_it.html
[22] http://www.edr-edr.it/Documenti/Document2_it.html

Le standard EpiDoc est fondé sur la norme TEI P4 et le travail de mise à niveau vers une compatibilité avec la version P5 est envisagé pour bientôt.

Les modifications apportées par EpiDoc au standard sont principalement liées à la description de la forme du document et à l'implémentation des règles de la convention de Leyden, une implémentation qui nécessita l'ajout de certains éléments et attributs mais surtout des modifications des feuilles de style afin d'assurer une impression optimale des résultats.

Cependant les changements ne se firent pas que dans une direction puisque la norme TEI P5 inclut un certain nombre d'éléments et d'attributs développés dans le cadre d'EpiDoc, tout comme ils l'ont également fait pour le projet européen de description de manuscrits ENRICH.

Aujourd'hui EpiDoc est utilisé par plus d'une vingtaine de projets dont la Duke Databank of Documentary Papyri, que nous évoquerons plus loin, ou le projet du palimpseste d'Archimède.

La capacité d'Epidoc à décrire aussi bien des pierres gravées d'Asie Mineure que des tablettes de bois inscrites du nord de l'Angleterre ou des papyri égyptiens montre toute la souplesse du système qui sert aussi bien à encoder des inscriptions rédigées en étrusque qu'en grec ou en latin.

Notons également que si le projet de Panciera, EAGLE, n'adopta pas EpiDoc car cela eut demandé trop de modifications aux bases de données initiales des participants à ce projet, cela ne signifie pas forcément que les deux projets soient rivaux. Leurs équipes ont en effet assuré une compatibilité limitée et développé des outils de conversion automatique des données partielles.

Chapitre 3 : La technologie en œuvre, quelques projets sous les projecteurs

Informatique et sciences de l'Antiquité, une histoire déjà ancienne

L'utilisation de l'informatique dans le domaine de l'Antiquité n'est pas une nouveauté absolue. Comme on l'a vu au chapitre précédent des projets comme la TEI ont débuté il y a plus de 20 ans sur base du constat des limitations des projets antérieurs.

En fait des projets informatiques en lien avec les sciences de l'Antiquité sont attestés dés les années 60. Le Lexique de base du latin produit par J. Michel et M. Gester[23] et publié en 1967 est issu de travaux de digitalisation sur cartes perforées qui permirent des études statistiques de fréquence d'apparition de mots dans la langue latine classique.

Le Laboratoire d'Analyse Statistique des Langues Anciennes (L.A.S.L.A.) est fondé à l'ULG dés 1961[24] : le résultat du travail conduit depuis lors est aujourd'hui disponible sur un site internet[25] dont une description détaillée a été publiée il y a quelques années[26]. Ce site web est lui-même l'héritier d'un cd-rom publié dans le courant des années 90 décrit dans un article publié en 1995[27].

Ces premiers essais concernaient avant tout les textes littéraires et avaient des visées essentiellement lexicographiques et statistiques. Dans le domaine de l'épigraphie nous n'avons pu trouver de projet antérieur à celui mené à partir de 1979 par l'Université de Cornell et le *Packard Humanities Institute* et connu sous le nom de *Searchable Greek Inscriptions*[28]. Ce projet visant à rassembler les inscriptions grecques fut d'abord publié sur le cd-rom PHI-7 en 1996 avant d'être porté sur le web[29].

[23] Michel J. et Gester M., Lexique de base du latin, Anvers et Paris, 1967
[24] http://www.cipl.ulg.ac.be/Lasla/ , cf. également **E**vrard E., « Le laboratoire d'analyse statistique des langues anciennes de l'Université de Liège », dans *Mouvement scientifique en Belgique***,** 9, 1962, p. 163-169
[25] http://www.cipl.ulg.ac.be/Lasla/descriptionop.html
[26] Denooz J., « *Opera latina* : une base de données sur internet », dans *EVPRHOSYNE*, 32, 2004, pp.79-88
[27] Purnelle G., « Opera Latina : la banque de donnée des textes latins du L.A.S.L.A. sur cd-rom » in Revue Informatique et Statistique dans les Sciences humaines XXXI 1 à 4, 1995, pp. 173-182
[28] http://www.packhum.org/phi/phi-info.html
[29] http://epigraphy.packhum.org/inscriptions/

Les quelques 80 000 inscriptions digitalisées représentaient 65mo de données et étaient livrées aux côtés de quelques 50 000 papyri alors digitalisés par le projet de la *Duke Databank of Documentari Papyri* (DDbDP)[30].

A côté de ces projets de digitalisation existaient d'autres projets visant à rendre disponible en ligne des ressources consacrées aux sciences de l'Antiquité : la numérisation de livres anciens et/ou épuisés se développe particulièrement dans les années 2000, à l'exemple de la mise en ligne du *Corpus Vasorum Antiquorum* (CVA)[31], projet entamé en 2000 qui se conclut en 2004. Nous rediscuterons d'ailleurs de ces numérisations dans quelques instants.

Les années 90 et 2000 voient également fleurir les sites internet équipés de moteurs de recherche permettant d'accéder aux métadonnées relatives aux sources ou même aux sources elles-mêmes, les dernières années voyant l'apparition de nouveaux projets venant se construire par-dessus les bases de données actuelles afin d'offrir de nouvelles possibilités aux chercheurs.

Des revues bibliographiques comme l'Année Philologique (APH)[32] ont également lancé au début des années 2000 leurs sites en ligne permettant d'explorer les ressources qu'elles offrent.

Notons toutefois que cet aspect de l'usage des nouvelles technologies dans le domaine des sciences de l'Antiquité ne sera pas abordé dans le présent mémoire, pas plus que la problématique plus générale du passage en ligne de nombreuses revues scientifiques.

Ce tableau est bien sur très succinct mais donne une bonne idée de la tendance générale des projets conduits au cours des 50 dernières années, des premiers usages à visées statistiques aux développements contemporains de mashup dans la lignée des évolutions du web 2.0.

[30] Plus d'informations sur le contenu du cd-rom PHI-7 sont reprises sur le site http://132.236.125.30/content.html de l'Université de Cornell
[31] http://www.cvaonline.org/
[32] http://www.annee-philologique.com/aph/

Des projets de nature différente

Un problème majeur lié aux ressources digitalisées est de savoir ce qui a été digitalisé, particulièrement dans le cadre de vastes corpus de plusieurs (dizaines de) milliers de textes.

La référence de la source ayant servit à l'encodage est en effet indiquée de manière parfois peu claire ou même complètement omise, ce qui limite l'utilité de la version électronique pour le chercheur.

C'est d'ailleurs l'une des critiques adressée à l'actuelle vague de digitalisation par A. Ghignoli lors du septième atelier Athis[33] lorsqu'elle dénonce la publication en ligne d'une version digitale des Monumenta Germaniae Historica[34], ouvrage du dix-neuvième siècle présentant des états souvent périmés des textes médiévaux d'Allemagne.

Ce monument de l'érudition germanique a en effet bien vieilli en raison des développements de la recherche et de nombreux textes qu'il contenait ont depuis été ré-édités et re-commentés par des chercheurs qui y ont découvert d'importantes erreurs ou lacunes.

Or le projet de digitalisation des MGH ne procède qu'à la copie de cet ouvrage et de son contenu sans tenir compte de toute la science accumulée depuis et ce alors que bien souvent aucune autre version en ligne des textes repris dans les MGH n'existe : les chercheurs souhaitant étudier ces textes à l'aide d'outils informatiques doivent donc travailler sur des versions périmées des documents ou procéder personnellement à la digitalisation de la ressource.

Dans ce contexte le projet des MGH peut donc apparaître comme un gaspillage de ressources ou, pire encore, comme une menace pour la qualité du travail scientifique et un facteur devant être pris en compte par le chercheur exploitant la ressource, ce que certains ont tendance à oublier.

Dans le domaine des sciences de l'Antiquité également on retrouvera ce genre de problèmes. Lors de la même présentation A. Ghignoli évoquait un autre exemple qui touche directement au propos du présent mémoire, en l'occurrence le projet de la Duke Databank of Documentary

[33] Ansani M. et Ghignoli A., « Testi digitali : nuovi media e documenti medievali », prochainement publié dans les Mélanges de l'Ecole Française de Rome
[34] http://www.mgh.de/dmgh/

Papyri (DDbDP). La principale critique de Ghignoli à l'égard de ce projet (sur lequel nous reviendrons plus longuement dans un autre chapitre du présent mémoire) est le caractère… scientifiquement inutile de la base de données.

En effet, selon elle, l'information critique entourant les textes repris dans la base de données n'ayant pas été encodée, seule une information périmée est offerte aux utilisateurs. Une information offerte à jamais et pour le plus grand malheur de la recherche puisqu'elle se basera sur des fondations incertaines.

La situation qu'elle décrit dans son intervention est cependant elle-même partiellement dépassée, comme nous le verrons lorsque nous étudierons plus en détail la DDbDP, mais il n'en reste pas moins qu'elle soulève là un véritable problème que l'auteur du présent mémoire avait déjà soulevé il y a cinq ans auprès des enseignants des cours de grec et de latin lorsqu'il s'était rendu compte que les éditions digitalisées des textes littéraires PHI 5 (latin) et TLG E (grec) étaient en parties basées sur de vieilles éditions allemandes de ces textes et que nombre de textes grecs et latins disponibles sur le web étaient des versions tombées dans le domaine public par l'expiration des droits d'auteurs, marque certaine de leur âge.

Le problème ne concerne d'ailleurs pas uniquement les sources premières puisque divers projets ont numérisé des sources secondaires, notamment des encyclopédies du XIXème ou du début du XXème siècle, qui sont devenus des outils courants pour les étudiants préparant leurs travaux ou les curieux cherchant à se renseigner.

La version électronique du *Dictionnary of Greek and Roman Antiquities (DGRA)*[35] de W. Smith, ouvrage datant de 1870, est ainsi devenu une source majeure d'information pour le monde anglophone alors que bien des données contenues dans cet ouvrage remarquable sont périmées.

Il en va de même pour la version numérique du *Dictionnaire des Antiquités grecques et romaines* de Ch. Daremberg et Edm. Saglio (DAGR)[36], monument de la science française paru à partir de 1877 et ayant connu un regain massif de consultation depuis sa mise en ligne (un compteur de visites indiquait près de 95 000 visites lors de notre dernière consultation de ce site en août 2009).

[35] http://www.ancientlibrary.com/smith-dgra/
[36] http://dagr.univ-tlse2.fr/sdx/dagr/index.xsp

Mais les projets ayant mené à cette numérisation avaient-ils vocation à servir d'assise à la recherche en sciences de l'Antiquité ? La question mérite que l'on s'y attarde quelque peu car en vérité il apparait à la lecture des descriptifs de ces projets publiés sur leurs sites internet que l'objectif de leurs créateurs était parfois de mettre en évidence l'ouvrage lui-même pour ce qu'il révèle sur son époque de rédaction. Il suffit de lire un extrait d'une critique du DAGR pour se rendre compte que ce genre d'ouvrage jouait un rôle majeur dans la politique de prestige des nations : *«No other nation as yet possesses anything approaching to it in beauty and completeness; it is absolutely necessary to every classical library and to every public library frequented by scholars or artists. »*[37]

D'outil de travail, l'ouvrage se transforme donc en source sur une époque méritant d'être étudié pour lui-même et son contexte. On peut le voir clairement en consultant les notices explicatives des sites du DAGR et du DGRA.

Ainsi la notice descriptive accompagnant la version en ligne du Dictionnaire indique :

« *Présentation sommaire du Dictionnaire des Antiquités grecques et romaines*

A partir de 1877 et avec régularité jusqu'en 1919, paraissent chez Hachette les dix forts volumes du Dictionnaire des Antiquités grecques et romaines, le fameux "Daremberg", du nom de son premier auteur et concepteur, Charles Daremberg. L'entreprise, à laquelle collaborent tous les plus grands noms de la science française, se situe dans un contexte européen de rivalités autour des "Sciences de l'Antiquité", alors considérées comme référence pour juger de la qualité de l'Université. Or en ce domaine, les maîtres sont les Allemands, depuis qu'ils ont refondé, en 1810, l'Université de Berlin, pièce essentielle à leurs yeux du redressement de la Prusse contre l'invasion napoléonienne. La "science de l'Antiquité", l'Altertumswissenchaft, telle qu'elle est élaborée par les grands philologues et historiens allemands comme Welcker, Boeckh, Droysen ou Mommsen, est même considérée comme une arme qui fait la supériorité des élites d'outre-Rhin et qui explique, pour une part, la défaite française de 1870-1871. C'est pourquoi les Français feront le "voyage d'Allemagne", dans les années qui suivent, pour se mettre à l'école de la science allemande de l'Antiquité et

[37] Mayor, John E. B., « Review of Dictionnaire des Antiquités Grecques et Romaines d'après les textes et les monuments, contenant l'explication des termes qui se rapportent aux mœurs, aux institutions, à la religion, aux arts, aux sciences, au costume, au mobilier, à la guerre, à la marine, aux métiers, aux monnaies, poids et mesures, etc. etc., et en général à la vie publique et privée des anciens » in *The Classical Review* 1 (7), pp 201–202

contribuer ainsi à préparer la revanche. Le Dictionnaire des Antiquités grecques et romaines trouve sa source dans ces luttes politiques et scientifiques. Il possède donc une incontestable valeur patrimoniale, et doit être mis en regard de tous les autres grands Dictionnaires de l'Antiquité qui paraissent en Europe au même moment. Les Allemands lui donnent en effet la réplique dès 1894 avec la monumentale Realencyclopädie der Altertumswissenchaft, dont quatre-vingts volumes sont parus à ce jour. En Italie paraît, à partir de 1886 le Dizionario epigrafico di antichità romana, dont la publication se poursuit. Les Anglo-saxons dans le même temps font paraître plusieurs dictionnaires concernant les champs de l'Antiquité classique. Le "Daremberg", pour toutes ces raisons, se présente comme un monument d'une très haute tenue scientifique. Chaque article est rédigé par le meilleur spécialiste de la question ; la référence à toutes les sources figure en bas de page ; les débats les plus actuels ne sont pas ignorés. Dans le registre des institutions, des techniques, de l'histoire des religions en particulier, il constitue alors une référence. Pour la maîtrise des sources textuelles, il reste inégalé. Bien évidemment, son contenu mérite d'être mis à jour en raison des découvertes et des progrès dans les domaines de l'archéologie, de l'épigraphie, de la numismatique. Par ailleurs, les problématiques se sont renouvelées profondément grâce aux apports de la sociologie et de l'anthropologie, dont les spécialistes de l'Antiquité ont intégré les méthodes. »[38].

Ce cours texte montre bien que, dans l'esprit des concepteurs du projet, la mise en évidence d'un document témoin d'une époque et de ses enjeux politiques apparait comme essentiel, au moins aussi important que les objectifs affichés :

« *Le projet de mise en ligne du Dictionnaire des Antiquités Grecques et Romaines est issu d'un partenariat entre des enseignants d'Histoire de l'Université du Mirail et la Mission Innovation Technologique et Multimedia qui assure son développement. Ce projet prévoit l'accès Internet au contenu de l'ouvrage et la possibilité de commenter et d'enrichir les articles.* »[39].

Celle du *Dictionnary* est plus explicite encore puisque le projet fut réalisé dans le cadre du projet *The making of America*[40], un projet dont on apprend sur le site principal que :

[38] Payen P., http://dagr.univ-tlse2.fr/sdx/dagr/index.xsp
[39] http://dagr.univ-tlse2.fr/sdx/dagr/index.xsp
[40] http://www.ancientlibrary.com/smith-dgra/

« *Making of America (MoA) is a digital library of primary sources in American social history from the antebellum period through reconstruction. The collection is particularly strong in the subject areas of education, psychology, American history, sociology, religion, and science and technology*. »[41]

Point question dans ce projet d'offrir un accès à un outil de travail en sciences de l'Antiquité ! La valeur du document est tout autre et ne pas le comprendre serait commettre une grossière erreur. Cela signifie également qu'il faudra sérieusement réfléchir avant d'envisager la mise en place de liens entre son projet et une ressource de cette nature.

[41] http://quod.lib.umich.edu/m/moagrp/

Numérisation versus digitalisation

Nous venons de le voir, les projets peuvent être très différents en ce qui concerne leurs objectifs. Ils peuvent également l'être sur les moyens utilisés : l'encodage complet du contenu d'un livre ou la prise de simple photographie de chacune de ses pages ensuite réunies dans un fichier n'ont pas grand-chose à voir.

Bien sur la numérisation peut s'accompagner de la création d'un index permettant de naviguer plus facilement dans le document mais elle ne permet pas des recherches en profondeur comme celles autorisées par la digitalisation du contenu, c'est-à-dire la copie dans un format intelligible pour l'ordinateur de l'information.

Les deux techniques ne sont pas mutuellement exclusives et il peut s'avérer très utile de combiner les deux approches pour offrir à l'utilisateur une consultation optimale du document : il est ainsi possible de créer des liens entre des zones identifiées sur une image et des contenus digitalisés, par exemple pour mettre en regard une édition critique et la ligne correspondant dans un manuscrit, une inscription ou un papyrus.

On peut même associer ensemble plusieurs documents de façon à afficher par exemple le texte, sa traduction et une image du document original sur laquelle le passage concerné est mis en évidence.

Cependant de tels projets demandent d'importantes ressources car ils impliquent un marquage de l'image, de la version digitalisée du texte et de sa traduction. En outre les images numérisées en haute résolution occupent souvent plus d'espace mémoire que des textes digitalisés, entrainant notamment des temps de chargement plus longs.

L'usage du format .pdf pour rassembler plusieurs images dans un même document, pratique courante dans les projets de numérisation, rencontre les mêmes problèmes de temps de chargement même si ils sont usuellement un peu plus brefs.

Notons enfin que les projets de numérisation se heurtent parfois à des intérêts financiers car certaines institutions détentrices de manuscrits rares n'hésitent pas à charger des tarifs exorbitants pour des copies de leurs documents comme le montre l'exemple de la Boldleian Library : 0.29£ pour une photo noir et blanc à 600dpi, 3.87£ pour une photo en échelle de

gris et 17,20£ pour une photo couleur ![42]. Eloignons nous un bref instant de notre sujet pour remarquer que de telles pratiques sont de véritables freins à la recherche et ne devraient plus avoir lieu à notre époque mais restent relativement courantes comme le montrent des exemples provenant de la Bibliothèque Nationale de France (30$ par page pour des images issues d'un mauvais scan de microfilms[43].

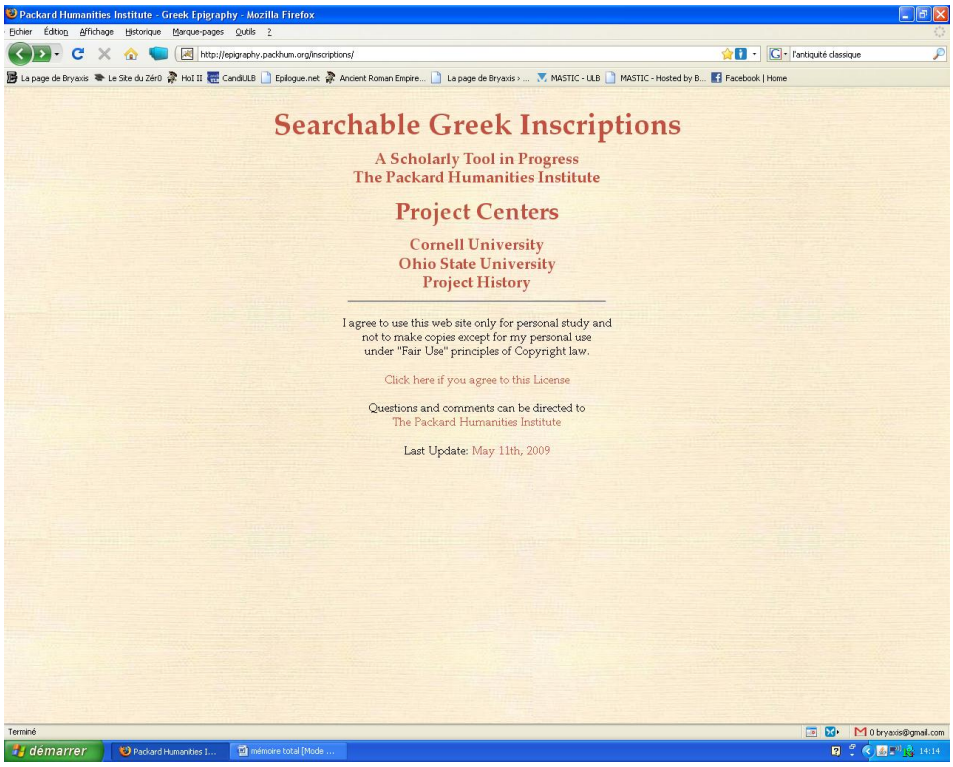
Pour quels revenus financiers ? L'information sur la question est rare mais un débat actuellement en cours au Royaume Unis[44] permet d'avoir quelques indications : une question adressée à la National Portrait Gallery de Londres à révélé que cette institution avait retiré entre 10 et 20 000 £ par an des droits de reproduction sur le net des portraits qu'elle conserve[45].

[42] http://www.roger-pearse.com/weblog/?p=2277
[43] http://www.roger-pearse.com/weblog/?p=1326
[44] http://blog.kennisland.nl/knowledgeland/2009/07/26/uk_national_portrait_gallery_vs_wikimedia_vs_the_public_domain/
[45] http://www.roger-pearse.com/weblog/?p=2113

Quelques exemples de projets

Nous allons maintenant nous pencher sur un certain nombre de projets sélectionnés pour leur rôle historique, leurs fonctionnalités ou la technologie qu'ils utilisent. Ce chapitre comprendra de nombreuses captures d'écran utilisées pour illustrer le propos et éviter de trop longues descriptions.

Searchable Greek Inscriptions[46]



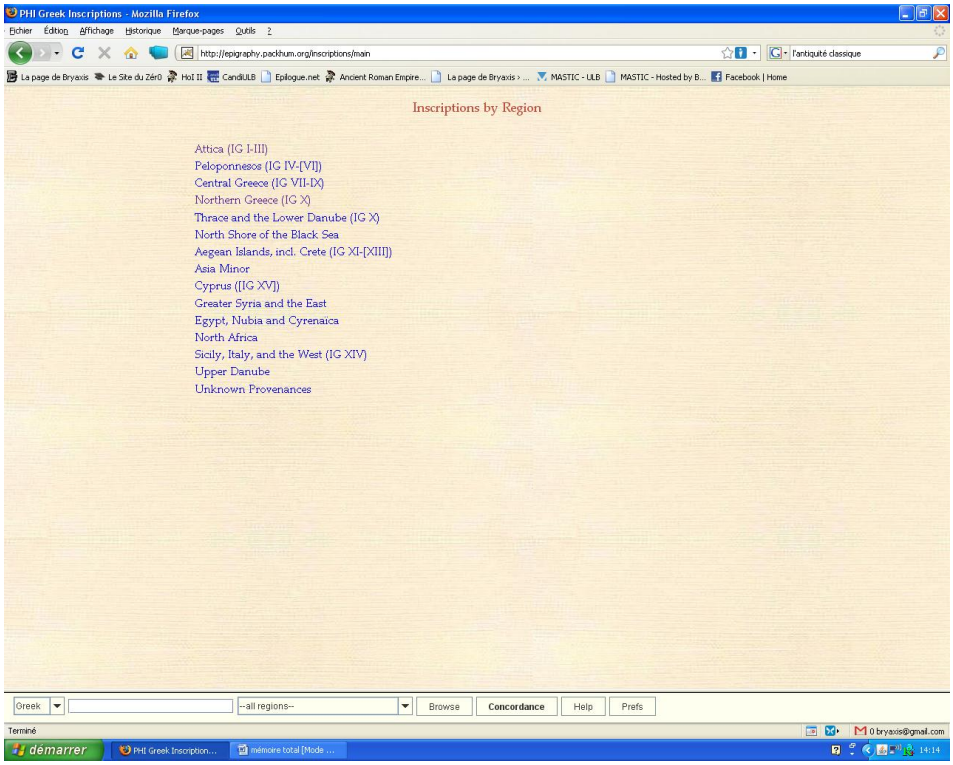
Le projet démarra dés 1979 à l'Université de Cornell avec pour objectif la digitalisation des inscriptions grecques antiques de toutes les périodes et de tous lieux. Comme nous l'avons déjà indiqué le projet fut d'abord livré aux côtés d'une série d'autres ressources sous la forme d'un cd-rom publié en 1996, le site internet apparaissant quelques temps après.

Ce projet de grande ampleur reçu le support de nombreux chercheurs du monde entier dont deux professeurs de l'ULB, A. Martin et P. Heilporn, qui y contribuèrent en fournissant le fruit de leurs recherches sur l'Egypte greco-romaine.

Le lien « project history »[47] présent sur la page d'accueil conduit à quelques informations sur le projet et notamment à une liste des personnes y ayant contribué. Néanmoins on constate

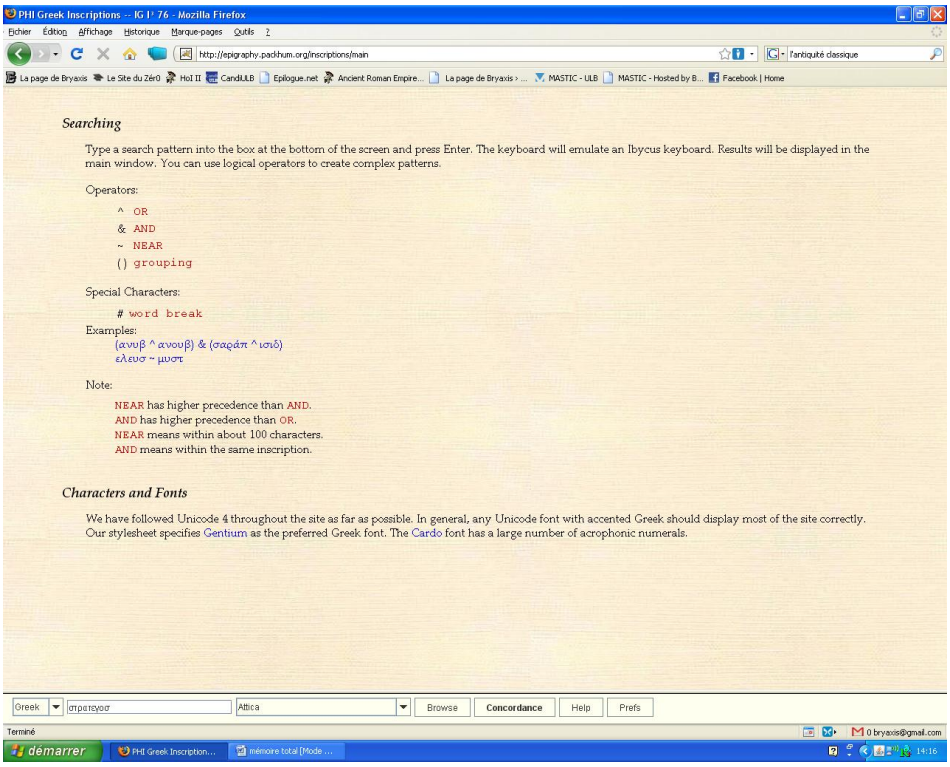
[46] http://epigraphy.packhum.org/inscriptions/
[47] http://132.236.125.30/history.html

qu'elle n'a plus été mise à jour depuis près d'une quinzaine d'années, donnant clairement une impression d'abandon d'une partie du site, lequel présente par ailleurs un design d'une sobriété toute spartiate.



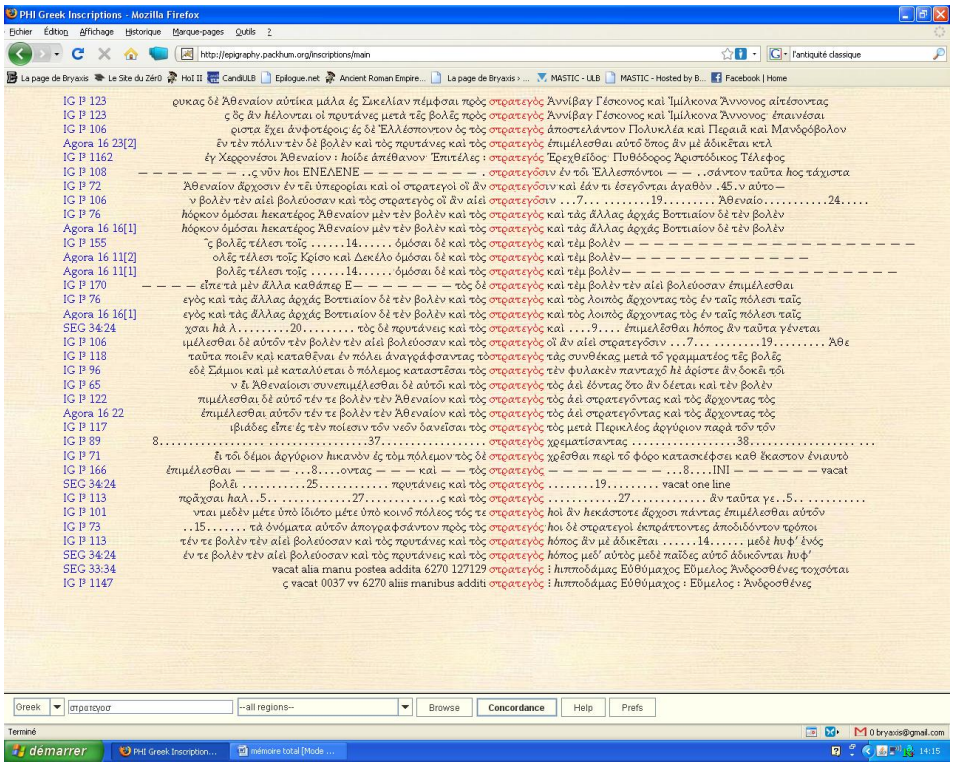
L'interface, particulièrement dépouillée, fournit un moteur de recherche et une liste des régions couvertes par la base de données, permettant une exploration géographique et un accès direct à chaque corpus ou une recherche plus complexe sur un terme spécifique pour l'ensemble de la base de données ou pour un sous-ensemble particulier.

Cette structure montre bien la manière de procéder du projet, qui travaille avant tout à mettre sous forme digitale le texte des premières éditions des inscriptions, sans travail de mise à jour de ces éditions : l'état des connaissances ainsi digitalisé fait parfois l'impasse sur des décennies de recherche.

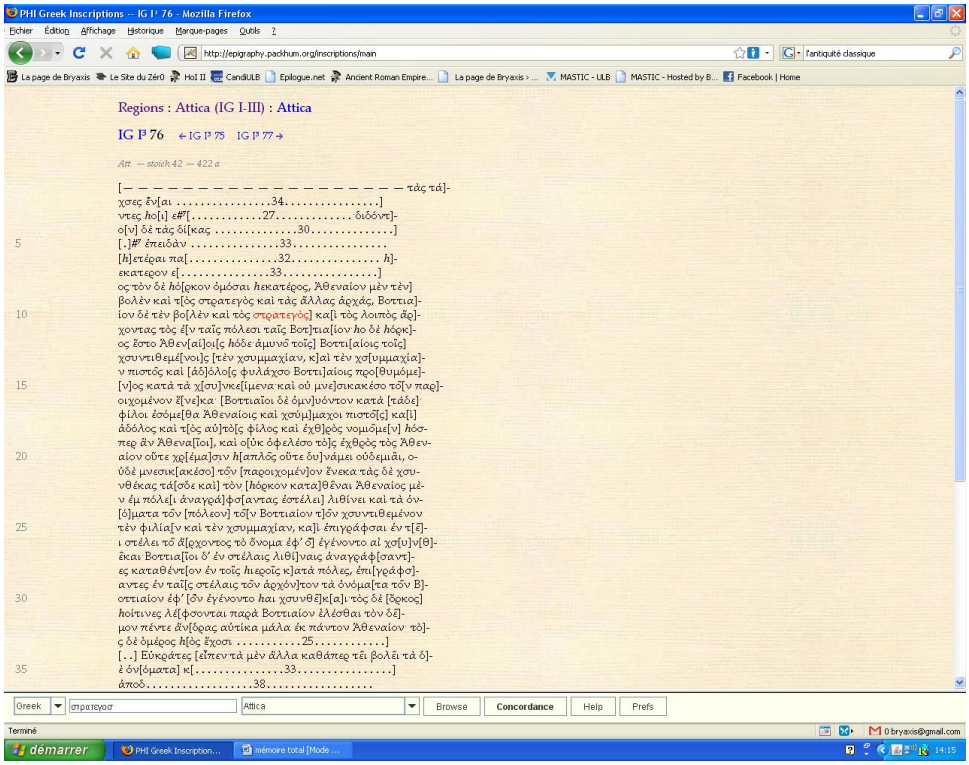


Le moteur de recherche du site permet de faire appel à un certain nombre de fonctionnalités avancées décrites dans un petit guide qui donne également quelques informations relatives aux problèmes de polices de caractères.

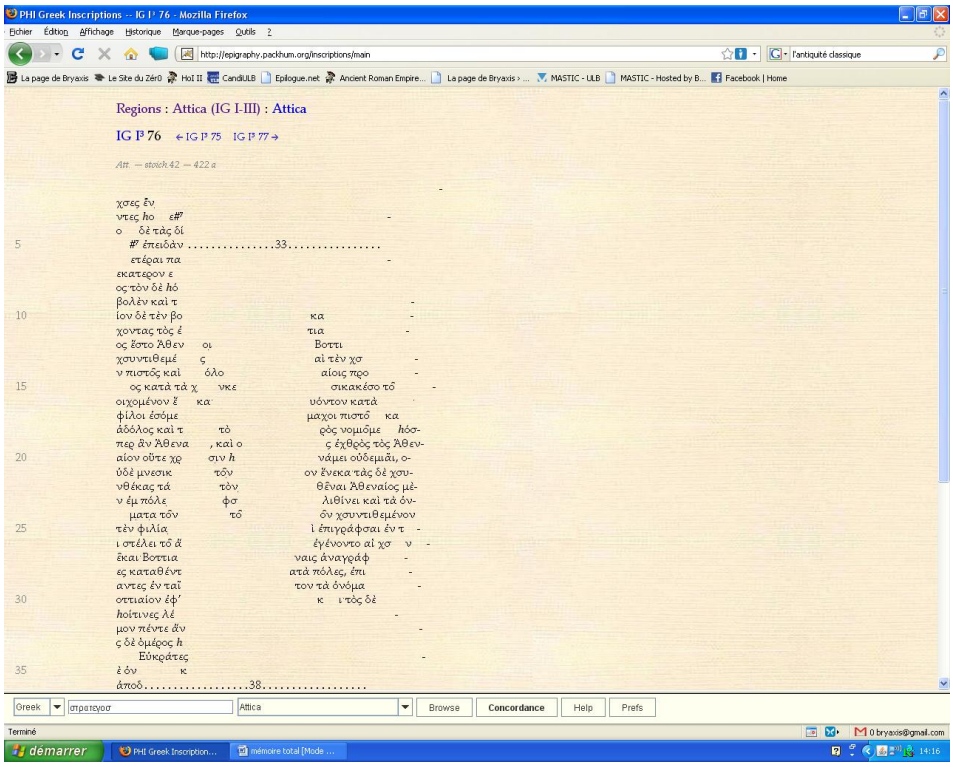
L'utilisateur peut concevoir des équations relativement complexes combinant des éléments de proximité (NEAR) et des éléments booléens (AND et OR). Néanmoins on note l'absence d'index autre que celui des ouvrages dépouillés.



Nous avons procédé à une recherche sur le terme grec strategos et demandé d'afficher le résultat en mode concordance : l'affichage des résultats, très dépouillé, est néanmoins efficace. On y retrouve le terme recherché replacé dans son contexte, avec la référence au numéro de la première publication du document.



En demandant l'affichage d'un texte spécifique on obtient sa retranscription intégrale, le mot recherché étant mis en évidence. Il est possible d'obtenir un aperçu de l'état réel de la pierre en cliquant sur le texte, ce qui fait disparaître toutes les restitutions et conjonctures.



On note que le texte fournit comprend très peu d'éléments de paracritique et que l'information bibliographique relative à l'inscription n'est pas fournie. En réalité on ignore tout à fait la nature de ce que nous lisons : est-ce la première lecture publiée du document, une retranscription plus récente, … ? Seule un entretient avec le Pr. P. Iversen, ancien membre de l'équipe du projet, nous a permis d'obtenir quelques précisions sur la manière de procéder de ses collègues.

L'exploitation de l'information fournie ici n'est donc qu'un premier tri au sein des corpus et nécessite d'approfondir la recherche par de nombreuses heures en bibliothèques pour dépouiller les revues comme le *Supplementum Epigraphicum Graecum* (SEG). S'il y a effectivement un gain de temps pour le chercheur il n'est en réalité pas aussi important que l'on ne pourrait le penser.

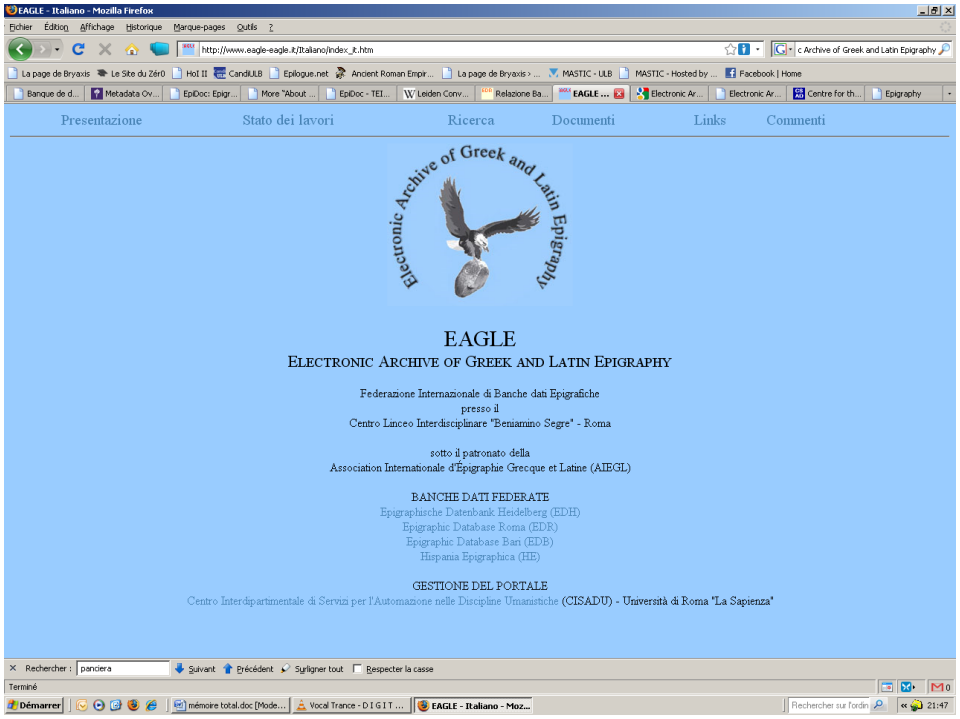
En revanche un point très positif est la possibilité (non visible sur les captures d'écran) de récupérer une adresse URL fixe pointant vers une inscription, ce qui permet d'envisager la création de liens vers d'autres systèmes.

Aucune information n'est donnée sur la technologie utilisée pour fournir le service, le système d'enregistrement ou la politique de mise à jour des enregistrements (seule la date de la dernière mise à jour et le nombre d'inscriptions ajoutées sont fournis, indiquant ainsi l'ajout de 6422 inscriptions depuis le début de l'année 2009).

L'absence d'index et de concordances empêche par ailleurs de connaître la politique en cas de présence d'une inscription dans deux recueils digitalisés.

On note également que certains corpus plus récents et disponibles sur internet ne sont pas repris dans la base de données. Ainsi le corpus des *Inscriptions of Aphrodisias* n'est pas repris dans la liste des corpus repris alors qu'il comprend près de 500 inscriptions inédites.

Electronic Archive of Greek and Latin Epigraphy (EAGLE)

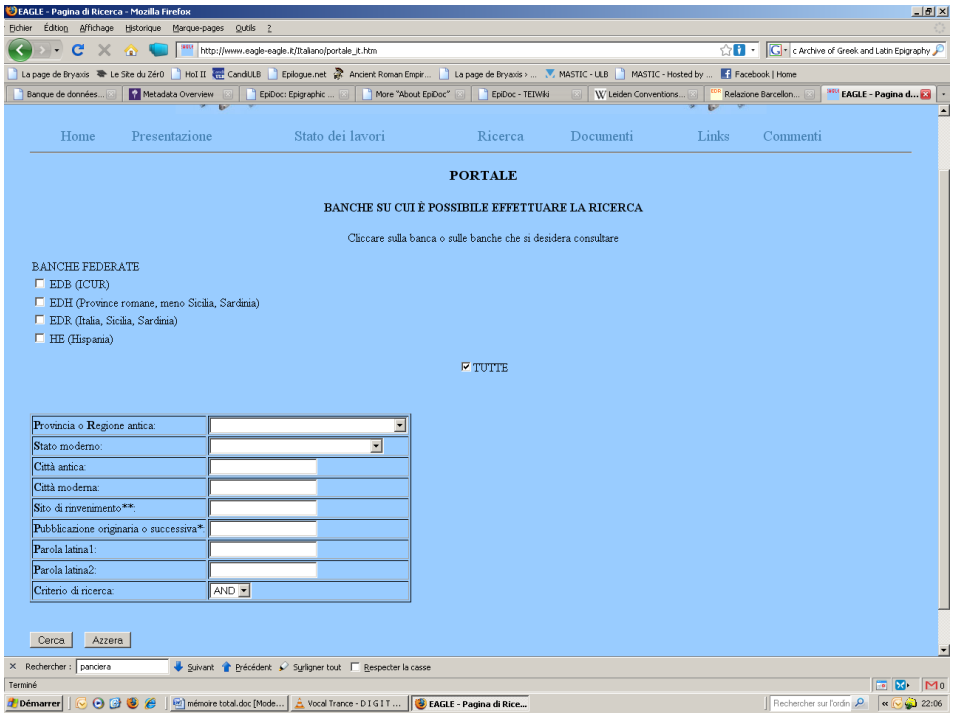


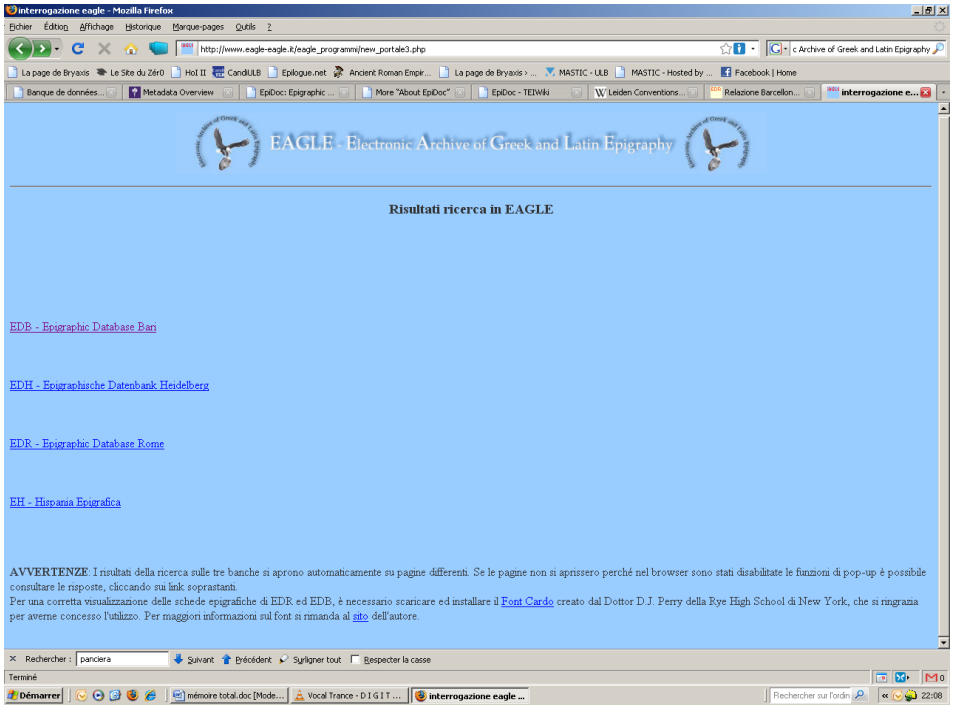
Le projet EAGLE[48] est été initié par le Pr. Panciera et lancé en 2003 sous les auspices de l'Association Internationale d'Epigraphie Grecque et Latine. Il s'agit en réalité plus d'une fédération de projets que d'une œuvre unique. En effet, il s'agit du regroupement autour d'un standard minimal commun de quatre projets plus anciens dont les données furent adaptées : *l'Epigraphische Datenbank Heidelberg*, *l'Epigraphic Database Roma*, *l'Epigraphic Database Bari* et l'*Hispania Epigraphica*.

Quoique EpiDoc ait déjà été présenté lors du lancement du projet, le formalisme de description issu de la TEI ne fut pas retenu, principalement parce qu'il eut nécessité trop de travail sur les bases de données existantes alors que le premier projet à en utiliser le standard n'était pas encore paru.

Depuis le projet progresse très lentement et son site n'est pas remis à jour de manière très régulière : la plupart des informations qui y sont disponibles remontent à 2006 alors même qu'un colloque fut organisé à son sujet en novembre 2008, colloque dont les minutes sont disponibles sur le site de l'*Epigraphic Database Roma*[49].

[48] http://www.eagle-eagle.it
[49] http://www.edr-edr.it/Download/EAGLE-Testi.pdf





Malgré sa date le moteur de recherche est peu évolué et montre la subsistance d'incompatibilités, certains champs n'existant pas dans certaines des bases de données. Plus gênant encore, le résultat de la recherche n'est pas affiché sur le site : il faut suivre des liens vers chacune des bases concernées si l'on souhaite connaître le résultat de sa requête.

En outre le site qui devait être international n'est disponible qu'en Italien.

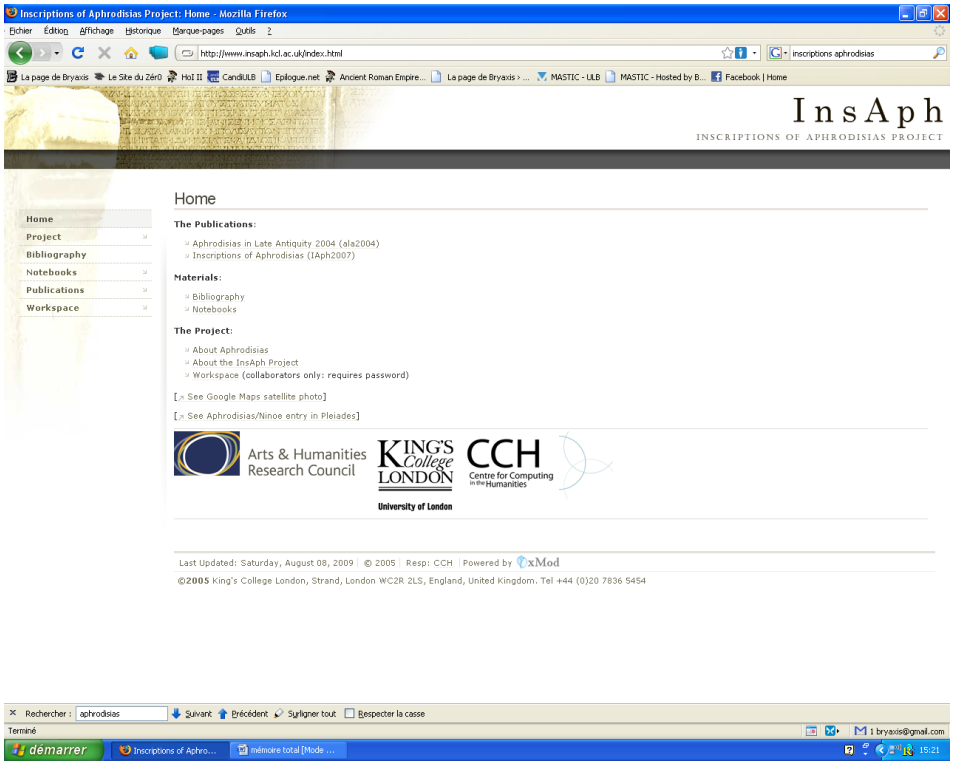
Cependant tout n'est pas mauvais dans ce projet. On note, notamment, que contrairement à la façon de procéder de la *Searchable Greek Inscriptions*, le projet EAGLE a prévu dés le départ de mettre à jour ses notices en fournissant à ses collaborateurs une interface d'encodage permettant notamment d'éditer les champs bibliographiques.

Il n'en reste pas moins que le projet EAGLE apparait comme techniquement dépassé et peu performant, surtout lorsqu'il est comparé aux sites (travaillant il est vrai sur des corpus plus modestes) mis au point par le Computing Center for Humanities du London's King's College dont nous allons maintenant parler.

EAGLE reste en effet fondamentalement au niveau de consultation offert par les grands corpus papier et n'exploite absolument pas les possibilités offertes par XML pour fournir de nouvelles informations ou de nouveaux modes de consultation.

Inscriptions of Aphrodisias Project (InsAph)[50]

Le projet des inscriptions d'Aphrodisias est en réalité composé de deux bases de données publiées respectivement en 2004 et 2007. Conçu initialement comme un démonstrateur technologique et un test grandeur nature pour la nouvelle norme EpiDoc, il se fixa comme première mission la publication de la version révisée d'un ouvrage rédigé fin des années 80.



La page d'accueil du projet permet d'accéder aux deux bases de données ainsi qu'à une information assez riche sur les participants au projet, les enjeux de ce dernier, et présente de manière générale la ville d'Aphrodisias. Une présentation assez grand public qui tient en partie de l'attention que portent toujours les britanniques (et plus généralement les anglo-saxons) à la vulgarisation du travail scientifique.

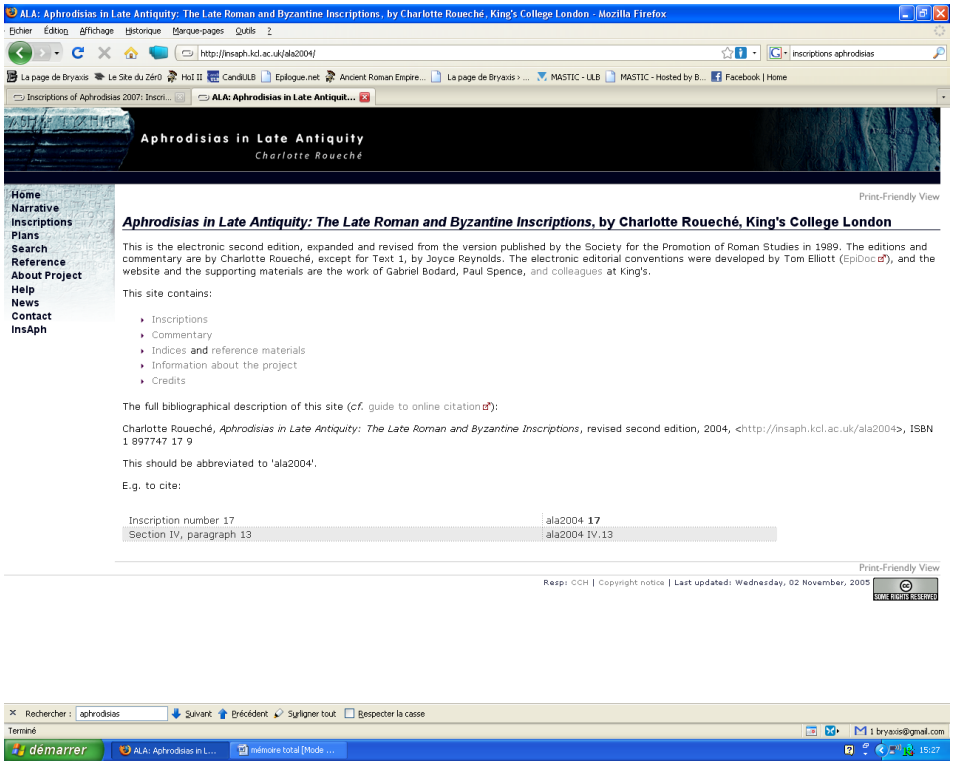
L'information assez détaillée relative à l'organisation donne une idée de la taille de l'équipe ayant travaillé sur le projet : en plus d'une équipe principale composée de cinq chercheurs disposant à la fois de compétences en épigraphie et en informatique, cinq épigraphistes, deux archéologues et deux informaticiens contribuèrent au projet.

[50] http://www.insaph.kcl.ac.uk/index.html

Les inscriptions reprises dans les deux banques de données du projet sont encodées en XML grâce au formalisme EpiDoc et peuvent être exploitées via deux sous-sites disposant chacun de leur propre URL.

Aphrodisias in Late Antiquity (Ala2004) [51]

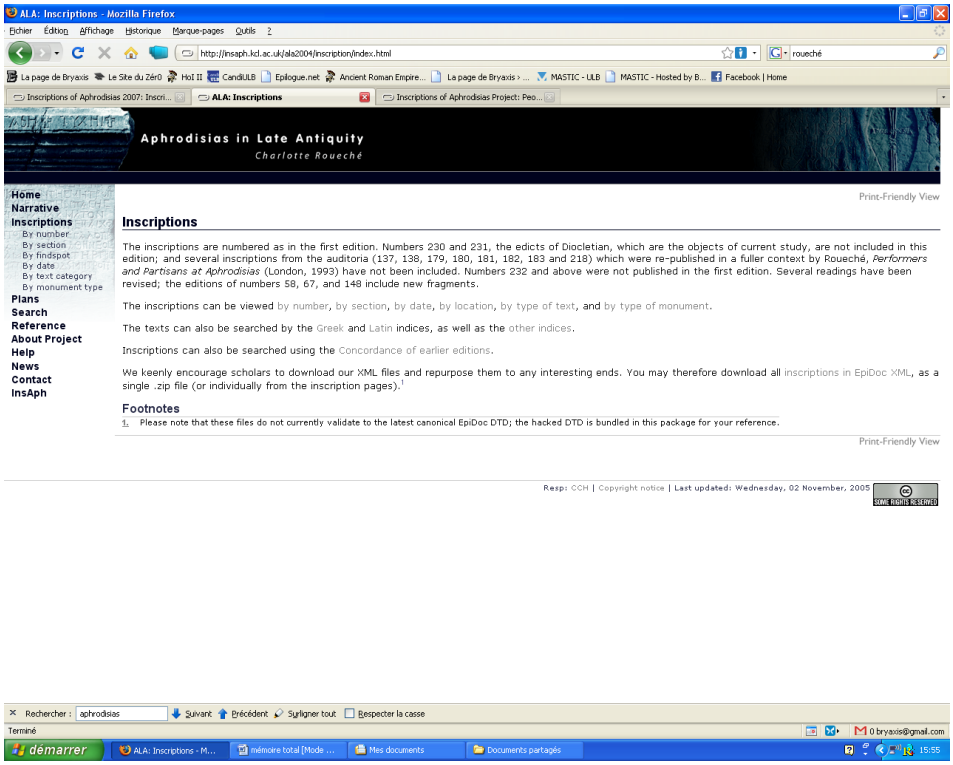
Le premier lot d'inscriptions encodé est celui des inscriptions relevées par Ch. Roueché dans son ouvrage Aphrodisias in Late Antiquity : The Late Roman and Byzantine Inscriptions[52], présentée ici avec divers ajouts et révisions par rapport à l'ouvrage d'origine.



Avec un corpus raisonnable de seulement 254 inscriptions, l'objectif pouvait sembler modeste mais était d'une taille raisonnable pour la première implémentation d'un nouveau standard et permit d'atteindre l'objectif en un peu plus de deux ans entre la phase de conception et la publication du site.

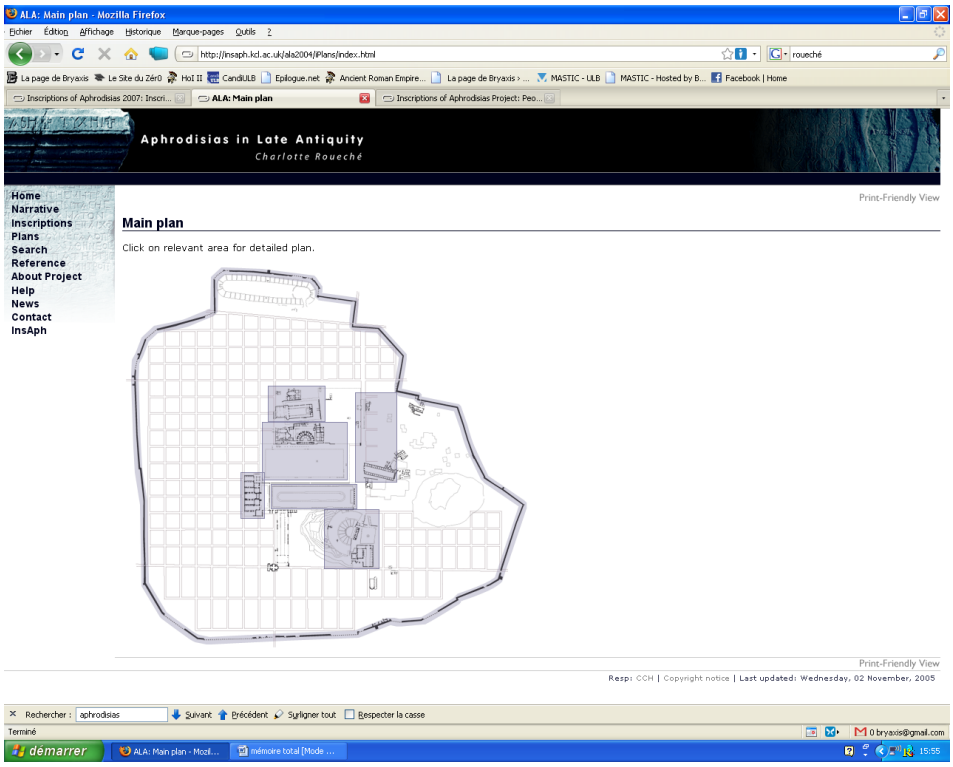
[51] http://insaph.kcl.ac.uk/ala2004/index.html

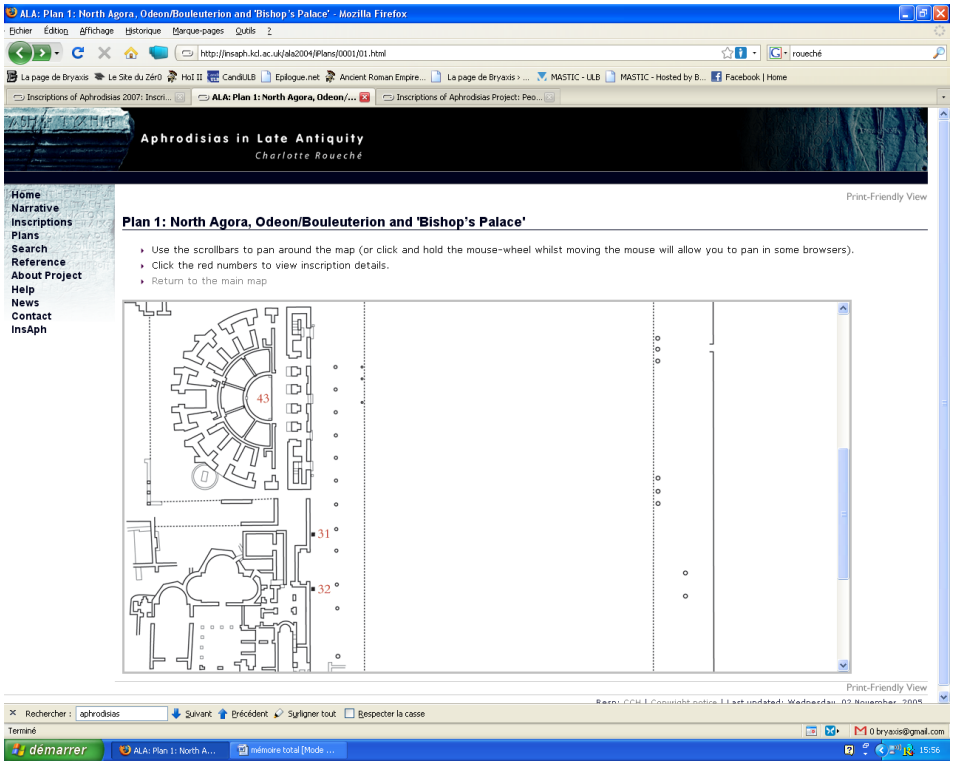
[52] Roueché Ch., Aphrodisias in Late Antiquity, Londres, 1989



Les inscriptions peuvent être atteintes de différentes manières, notamment au moyen d'index ou de recherches plein texte sur le corpus. Pour faciliter la recherche, il est également possible de consulter une table de correspondances entre les codes de la base de données et les références de publication dans d'autres ouvrages ou sites internet.

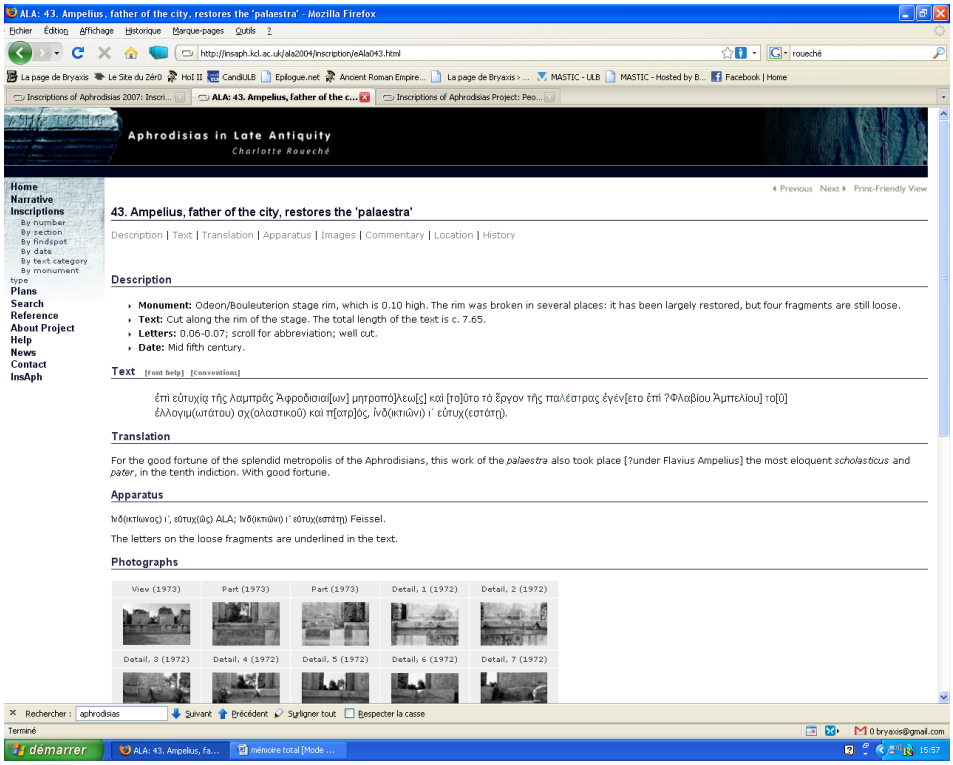
Cependant ces tables de concordances sont, comme d'ailleurs toutes les autres informations du site, livrées dans leur état au moment de la publication : l'approche des concepteurs est celle d'une édition comparable à celle d'un livre, un état stable de l'information sans réactualisation.

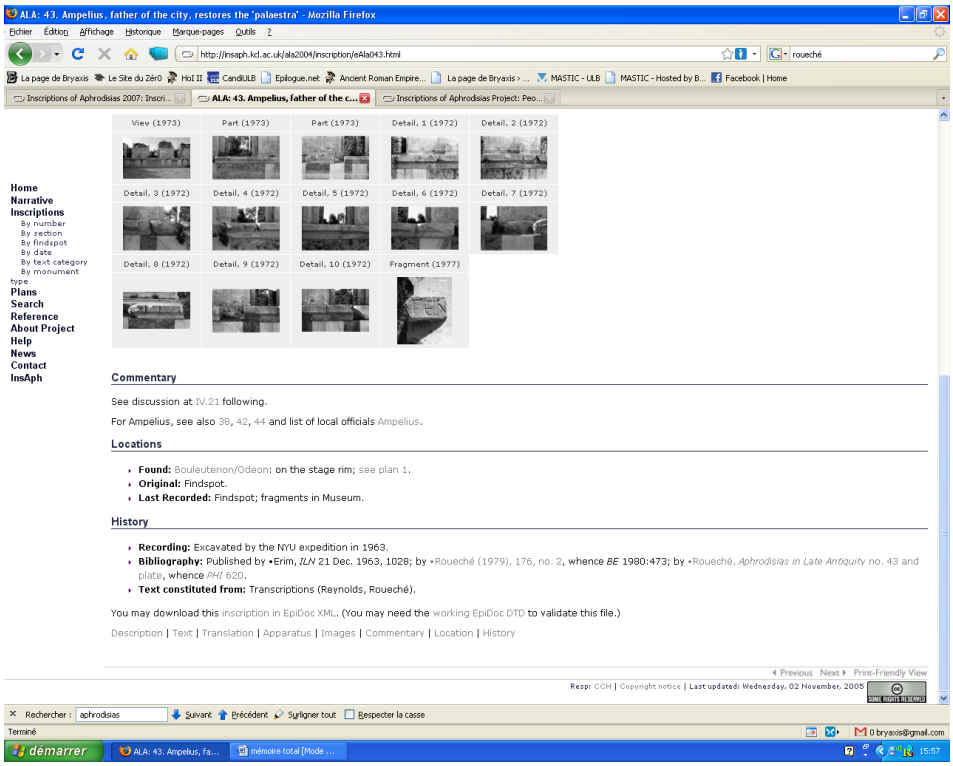




Une des fonctionnalités parmi les plus impressionnantes du site est la possibilité qu'il offre de localiser les inscriptions sur le plan du site. Une série de zones principales a été définie et donne accès à des plans des vestiges fournis par les archéologues fouillant Aphrodisias. Cliquer sur les numéros indiqués sur le plan permet d'accéder directement à la fiche de l'inscription.

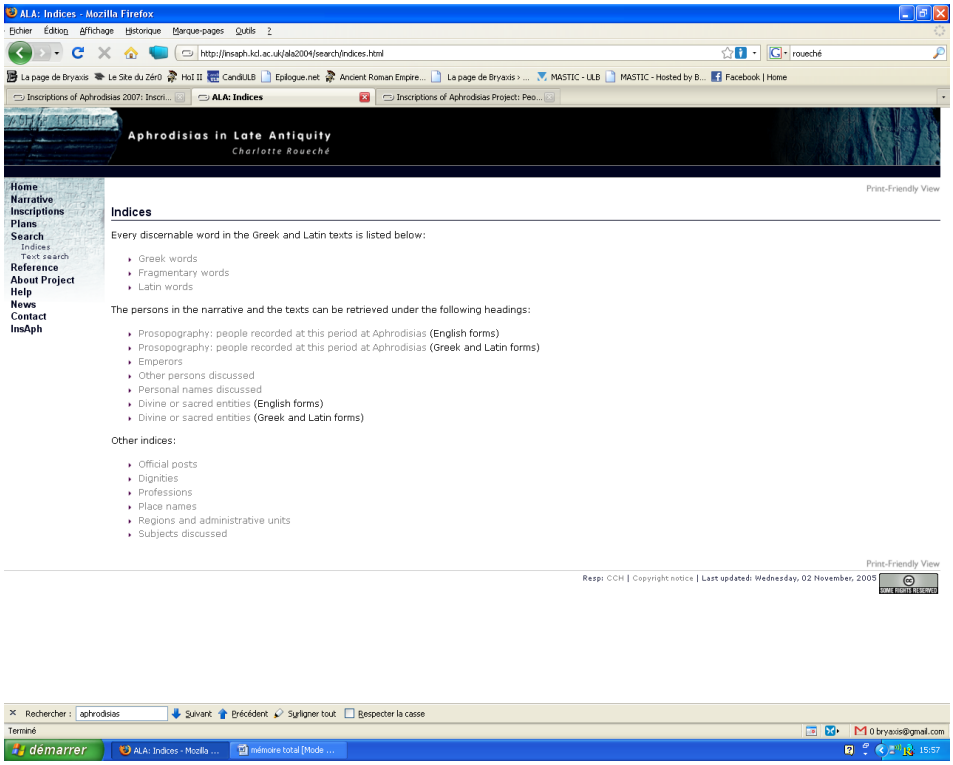
Avec un tel système on se rapproche des rêves de plans interactifs globaux de sites de fouilles dont parlent les archéologues depuis plus d'une décennie sans que beaucoup d'exemples n'aient encore vu le jour. Rien n'empêche en effet d'ajouter sur cette carte des photographies des vestiges et des objets trouvés sur le site.





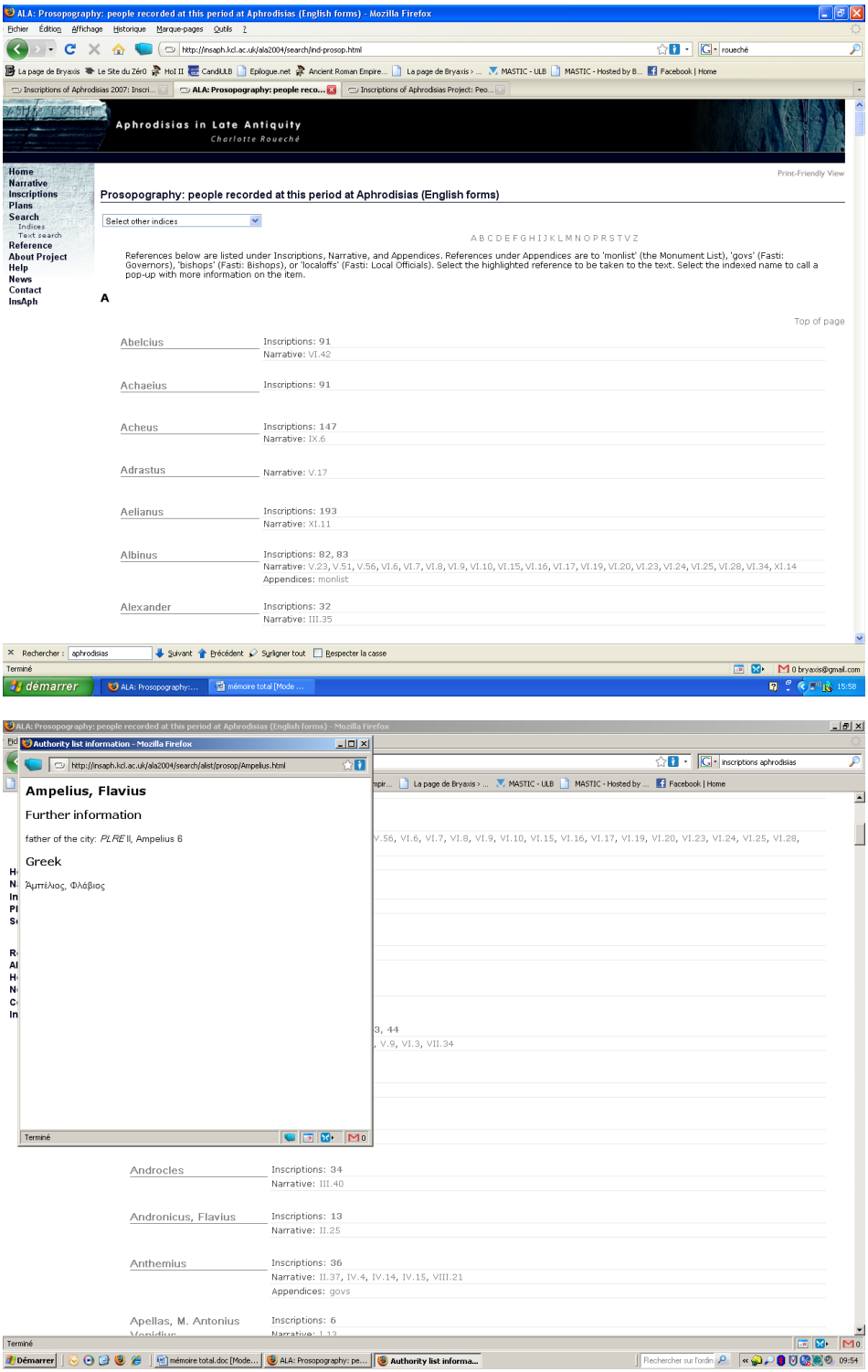
Ces deux captures d'écran montrent l'information fournie pour chaque inscription : description du monument portant l'inscription, description du texte, informations relatives à la

gravure, datation, texte de l'inscription et traduction, appareil critique, photographie de l'inscription dans son contexte, commentaire, informations sur sa localisation actuelle et notice historique comprenant la bibliographie de l'inscription. La fiche est téléchargeable dans le format EpiDoc pour réemploi[53].



La liste des index consultables est impressionnante car elle représente à chaque fois un marquage spécifique de l'information ainsi indexée : on peut donc obtenir très rapidement des listes d'anthroponymes ou de théonymes dans leur forme ancienne ou anglaise, des informations prosopographiques, la liste des professions ou des magistratures attestées dans la cité, etc.

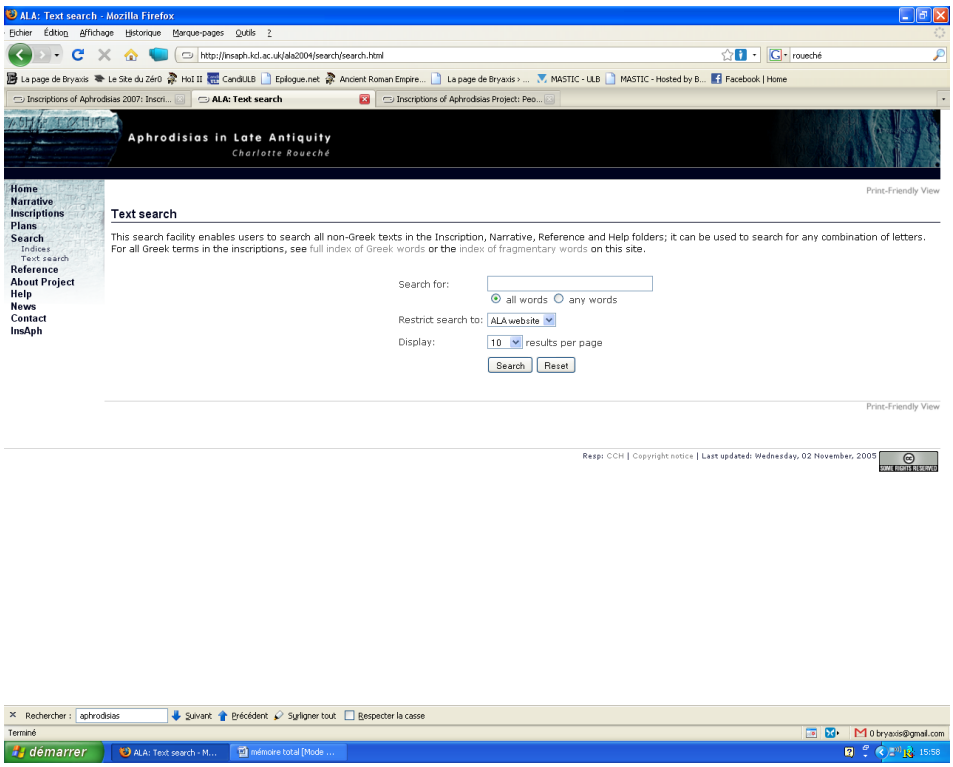
[53] http://insaph.kcl.ac.uk/ala2004/redist/inscriptions/eAla043.xml



L'information géographique et prosopographique est enrichie de références externes comme montré ici avec un personnage également repris dans la *Prosopography of the Late Roman Empire*. Les noms de lieux sont eux associés à leurs références dans le *Barrington Atlas of the Ancient World* ce qui permet d'envisager leur association avec des projets extérieurs comme

Pleiades[54] afin de compléter les données de ce projet ou d'importer de nouvelles données relatives aux inscriptions : toutes les informations étant accessibles il suffirait qu'une personne intéressée crée un logiciel faisant le lien entre ces informations pour obtenir un outil d'une grande richesse.

Une fonctionnalité d'un tel logiciel pourrait être d'afficher à l'écran l'ensemble des inscriptions concernant une personne tout en marquant leur emplacement sur une carte du bassin méditerranéen. Nous reviendrons d'ailleurs sur cette possibilité au chapitre 4 avec un autre projet ayant mis en place de telles fonctionnalités.



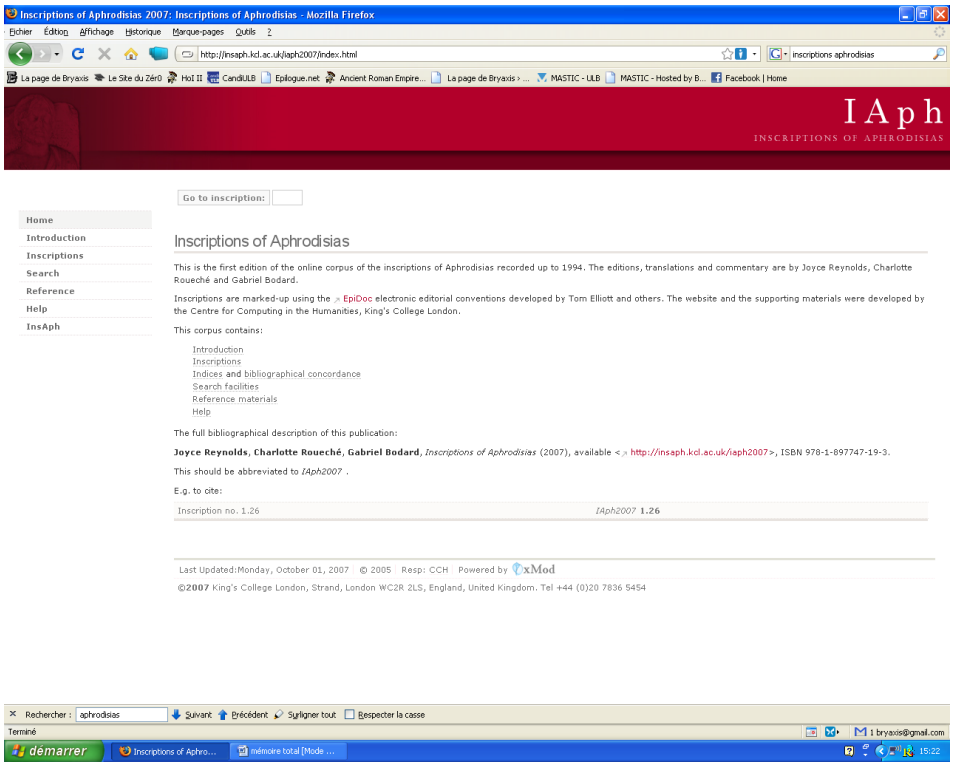
Un moteur de recherche est également fourni pour explorer les données mais il ne permet pas de faire une recherche dans le texte grec, l'index complet des mots grecs répertoriés dans les inscriptions remplissant cet usage. Ce moteur de recherche limité est un peu une déception car l'utilisateur peut souhaiter rechercher une séquence spécifique de mots, ce qui n'est pas possible dans le cas présent.

[54] http://pleiades.stoa.org/

Inscriptions of Aphrodisias (IAph 2007)[55]

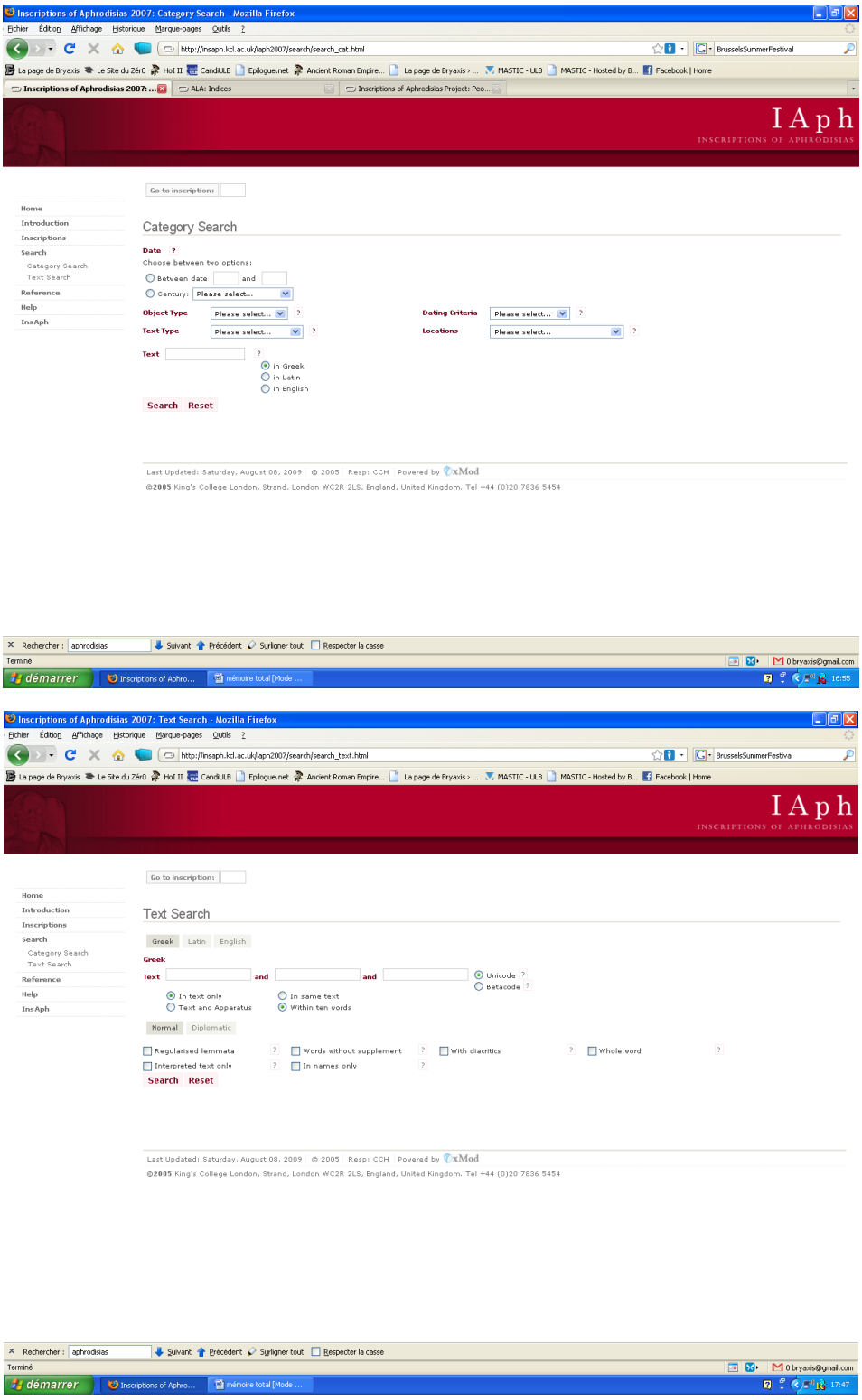
Premier projet d'encodage d'un corpus ancien réalisé à l'aide d'EpiDoc, le site Late *Antiquity in Aphrodisias* est une belle réussite qui conduit bientôt à un nouveau projet : la numérisation de tous les textes d'Aphrodisias publiés avant 1994.

Ce projet aboutira en 2007 à la publication du site Inscriptions of Aphrodisias : près de 1500 inscriptions (dont un tiers d'inédits) mises en ligne sur un site dédié.

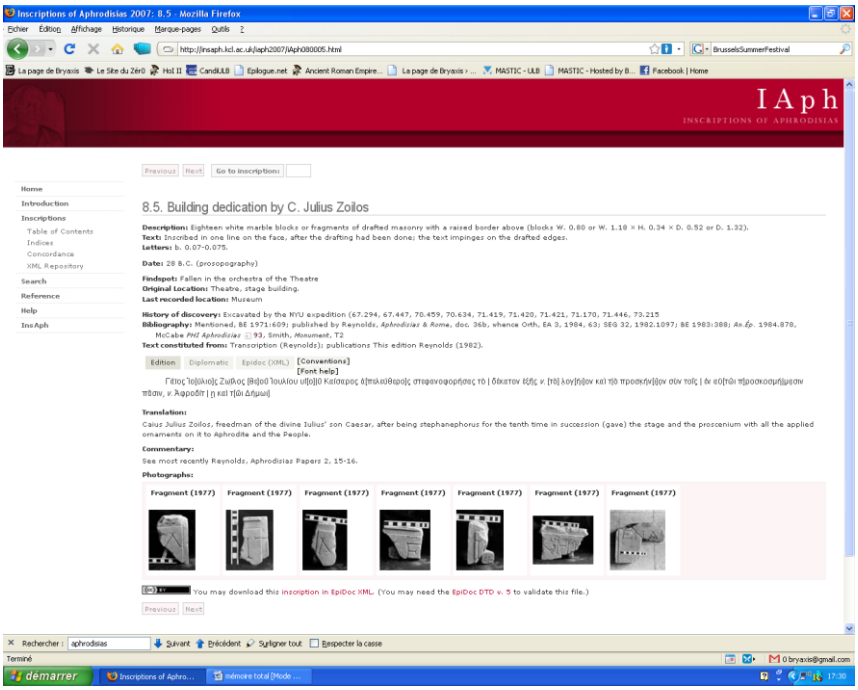


Le projet présente un certain nombre d'améliorations par rapport à son prédécesseur, principalement dans le nombre d'index disponibles. On note toutefois que le repérage des inscriptions sur une carte du site n'est plus intégré, les inscriptions restant toutefois classées par zone géographique.

[55] http://insaph.kcl.ac.uk/iaph2007/index.html



Contrairement à son prédécesseur, le site des IAph2007 comprend un moteur de recherche complet. Ce dernier est puissant et permet d'effectuer des requêtes approfondies sur la base de données, y compris sur le texte même des inscriptions, comblant ainsi une lacune de son prédécesseur.



Par rapport au projet précédent on peut dire que le site fournit une information de qualité similaire et utilise globalement la même présentation. Les données sont toujours récupérables en format XML pour les chercheurs qui souhaiteraient les ré-utiliser ou simplement examiner le balisage. On note toutefois que les fiches des inscriptions semblent un peu moins riches, le projet ayant encodé six fois plus de documents en autant de temps qu'il en avait fallu pour mettre au point le démonstrateur. Par ailleurs la même politique de non réactualisation du matériel publié semble en vigueur.

Duke Databank of Documentary Papyri (DDbDP) - Advanced Papyrological Information System (APIS) – Heidelberger Gesamtverzichnis der griechischen Papyrusjunden Ägyptens (HGV)

La *Duke Databank of Documentary Papyri* (DDbDP) est un projet déjà ancien visant à digitaliser l'ensemble des papyri rédigés en grec ou en latin. C'est en 1982 que le projet est lancé avec l'aide notamment du Dr. David W. Packard, fils de David R. Packard, fondateur de la firme informatique Packard-Bell.

A la fois ingénieur informaticien et philologue, le Dr. Packard a financé au moyen de la *David and Lucile Packard Foundation*[56] et du *Packard Humanities Institute* (PHI) une série de projets de digitalisation et ce dès le début des années 80. Le résultat de ce travail était fourni sous la forme de cd-rom : nous avons déjà noté plus haut que la DDbDP fut publiée sur le cd-rom PHI-7 en 1996, aux côtés de la base de données des *Searchable Greek Inscriptions*.

Le Dr. Packard avait une vision bien à lui de la manière dont devait progresser la digitalisation des textes anciens, allant jusqu'à concevoir un ordinateur spécialement dédié au projet *Thesaurus Linguae Graecae (TLG)*[57], l' Ibycus[58], qui offrait des facilités pour l'encodage du grec sans commune mesure avec ce qu'offraient les systèmes informatiques de l'époque, utilisant notamment le Beta-Code que Packard avait inventé à la fin des années 70.

Le système comprenait également une imprimante créée pour le projet et était capable de lire des cd-rom bien avant que les lecteurs ne se diffusent auprès du grand public le premier cd-rom produit par l'équipe du TLG étant publié en 1985.On peut d'ailleurs trouver sur le site internet du centre liégeois de papyrologie CEDOPAL une vidéo montrant la consultation du cd-rom TLG-C (1988)[59].

La machine resta d'ailleurs opérationnelle au sein du projet TLG jusqu'en 1999 même si les données avaient été rendues compatibles avec les systèmes UNIX dés le milieu des années 90[60] : l'outil étant clairement devenu obsolète, nombreux étaient ceux souhaitant cesser

[56] http://www.packard.org/home.aspx
[57] http://www.tlg.uci.edu/
[58] http://opuculuk.blogspot.com/2009/05/ibycus-mainframe.html
[59] http://promethee.philo.ulg.ac.be/cedopal/video/film_ibicus.htm
[60] http://www.tlg.uci.edu/about/ibycus.php

d'utiliser ce matériel ne respectant aucune norme (notamment pour ses ports de connexion de périphériques) et pour lequel les pièces détachées étaient devenues pratiquement introuvables.

La publication du cd-rom PHI-7 marqua donc la fin de la collaboration entre Packard et le projet et entraina la migration des données vers le site internet du projet Perseus[61] (dont nous reparlerons dans un instant).

La migration sur Perseus impliqua la conversion des données du format de stockage utilisé par Packard vers le format SGML alors utilisé par le projet de l'université de Tuft. Le projet continua en vitesse de croisière pendant près d'une décennie, étant financé dans le cadre de l'*Advanced Papyrological Information System* (APIS) conduit par l'université de Columbia[62].

Là où la DDbDP était avant tout un projet visant à transcrire le texte des papyri, APIS avait été lancé dans le but de fournir des métadonnées sur les papyrus conservés dans les collections universitaires américaines. Cet objectif s'est progressivement élargi pour couvrir quelques institutions supplémentaires.

En mai 2008 le projet annonçait près de 30000 enregistrements comprenant un peu plus de 20000 photos. Environ 6000 traductions de papyrus étaient également disponibles à cette date et 5724 entrées étaient liées à la transcription du texte contenue dans la DDbDP.

Parallèlement un projet était mené en Allemagne pour répertorier toutes les métadonnées relatives à l'ensemble des papyri et ostraca publiés en grec et en latin, qu'ils proviennent d'Egypte ou des régions voisines.

Ce projet, baptisé *Heidelberger Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens* (HGV)[63], utilise le logiciel FileMaker Pro pour enregistrer les seules métadonnées car ses concepteurs avaient estimé que la transcription ayant déjà lieu dans le cadre de la DDbDP il était inutile de dépenser des ressources dans ce travail long et fastidieux, particulièrement dans le cadre des technologies de l'époque.

Le projet HGV se concentra donc sur la validité des métadonnées afin d'offrir aux chercheurs un moyen de faire une recherche préliminaire dans les corpus et de déterminer les papyri pertinents pour leurs recherches. Un soin particulier fut apporté à la validation des

[61] http://www.perseus.tufts.edu
[62] http://www.columbia.edu/cu/lweb/projects/digital/apis/
[63] http://www.rzuser.uni-heidelberg.de/~gv0/gvz.html

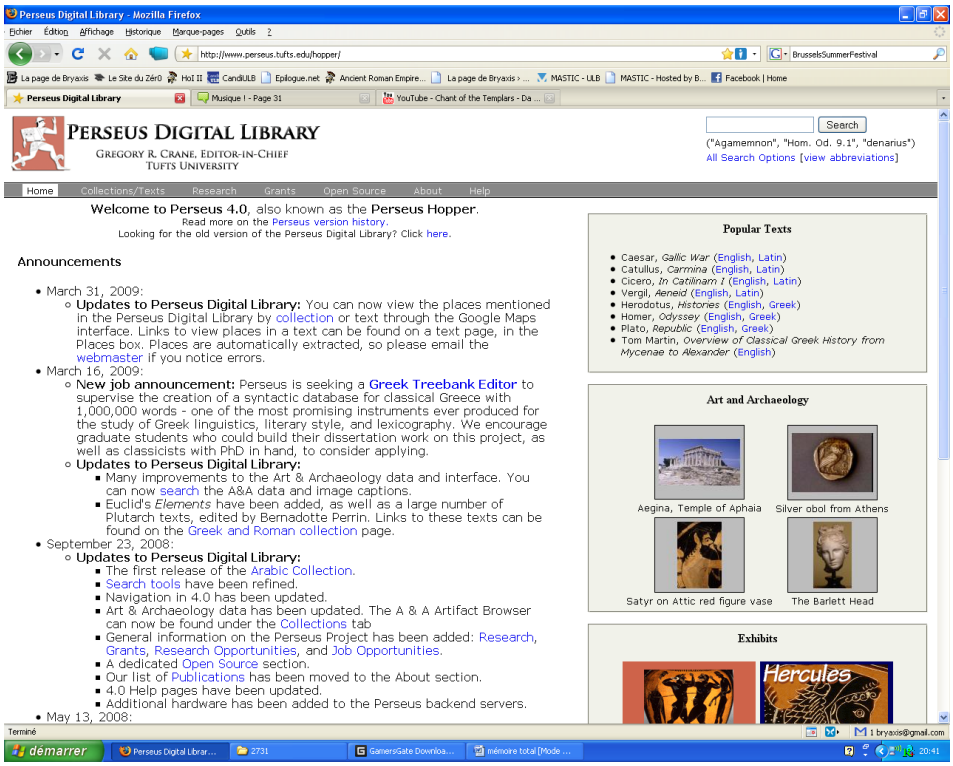
informations de datation des papyri et des listes chronologiques de papyri furent publiées à partir de 1993.

C'est en 1997, alors que la DDbDP passait en ligne sur Perseus, que l'équipe du HGV rendit sa base de données accessible par internet, apportant de nouvelles possibilités d'interrogation de la base de données : la recherche multi-champs fut en effet offerte à partir de cette époque.

Néanmoins les bases de données des projets DDbDP et la HGV n'étaient pas liées entre elles et un chercheur ne pouvait obtenir l'ensemble de l'information disponible dans le cadre d'une seule requête.

C'est pour cette raison que les deux équipes se réunirent à partir de 2004 pour procéder à une évaluation de la faisabilité de la normalisation de leurs données, un projet finalement entamé en 2007 et dont nous reparlerons lorsque nous aborderons le navigateur papyrologique.

Perseus



La *Perseus Digital Library*[64] est un des plus importants sites dans le domaine de la vulgarisation de l'information sur les sciences de l'Antiquité. Le planning des opérations devant conduire au site actuel débuta en 1985.

Digitalisant des centaines de ressources portant non seulement sur le monde greco-romain mais s'intéressant également à d'autres civilisations comme les peuplades germaniques, la renaissance européenne, le monde arabe, l'histoire américaine du 19ème siècle, etc.

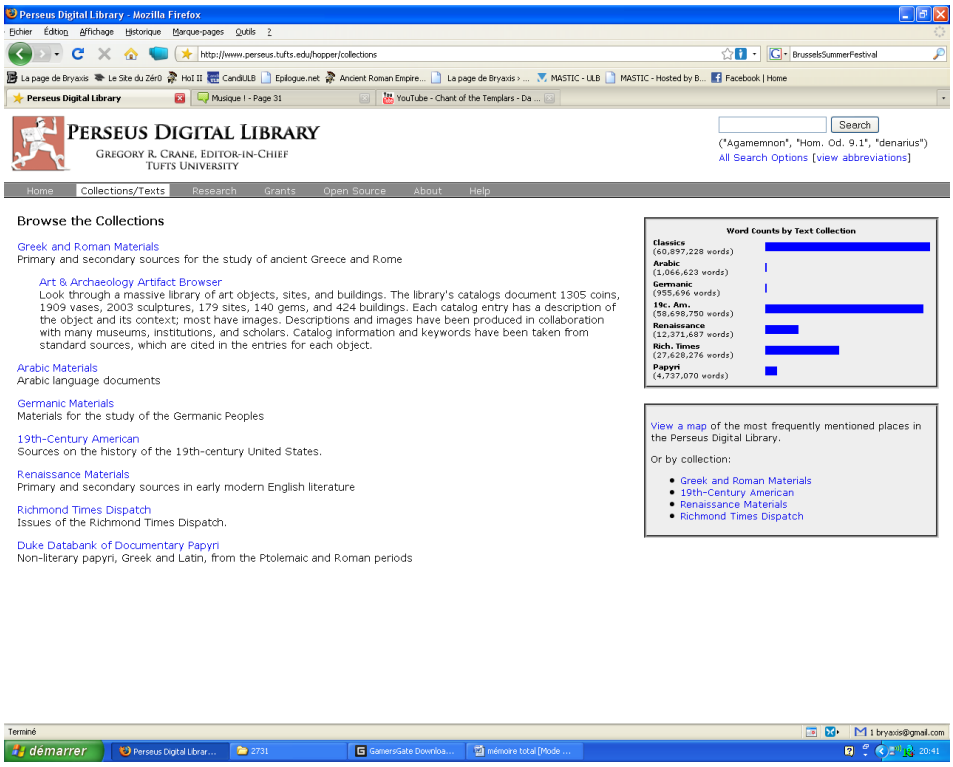
Quoique plutôt centré sur la publication de textes littéraires grecs et latins, le site Perseus a également été de 1996 à 2008 l'hôte de la DDbDP et a eu une certaine influence sur nombre d'autres projets dont le *Papyrological Navigator*.

C'est pour cet ensemble de raisons que nous avons choisi d'inclure ce site dans notre étude, afin de voir en quoi son influence se ressent encore aujourd'hui sur les projets les plus à l'avant-garde.

[64] http://www.perseus.tufts.edu/hopper/

La première version publique de Perseus fut fournie en 1992 par les Yale University Press, suivie en 1995 de la première version en ligne, une version bientôt remplacée par Perseus v2.0, un logiciel ne tournant alors que sous Mac. Le projet tournait alors sous SGML.

Il faut attendre 2000 avant que Perseus v2.0 ne devienne disponible pour les utilisateurs de PC et que le site ne connaisse une refonte, donnant naissance à la version 3.0 du projet. Il faudra attendre 5 ans avant que la version actuelle de Perseus n'apparaisse sur le web, les collections grecques, latines et anglaises étant publiées au format TEI l'année suivante[65].

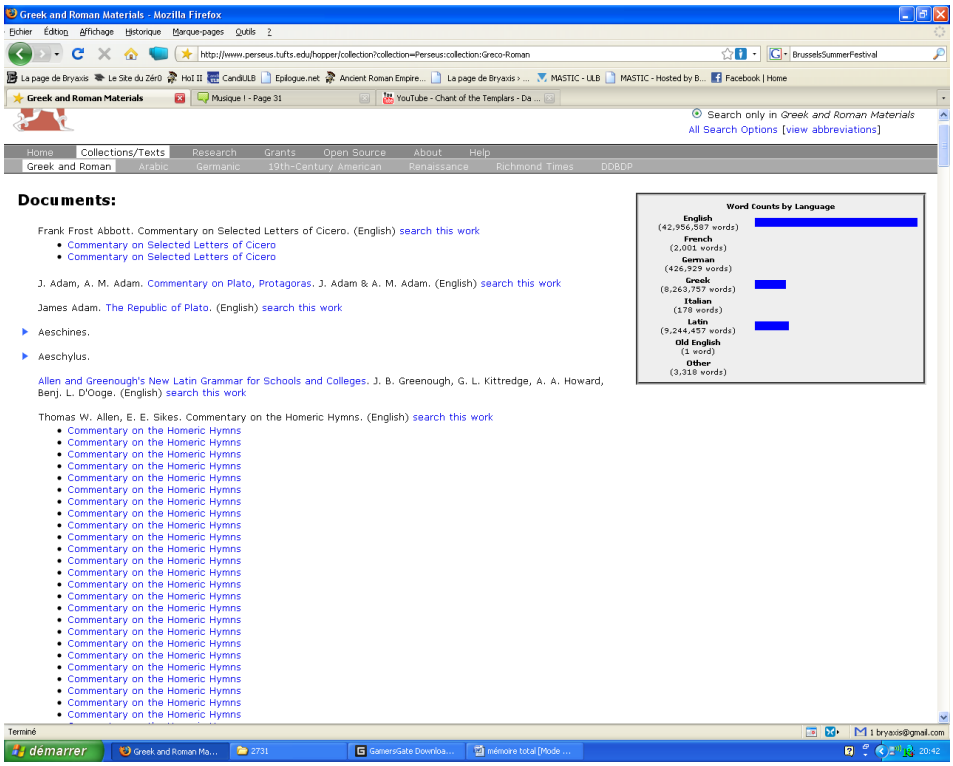


Nous l'avons déjà dit, Perseus héberge diverses collections. Non seulement offre-t-il l'accès à des sources premières mais il propose également des introductions sur différents sujets historiques ou archéologiques ainsi qu'une importante galerie d'images de sites et d'objets anciens. Il propose également quelques expositions virtuelles sur des thèmes comme les jeux olympiques ou le mythe d'Hercules.

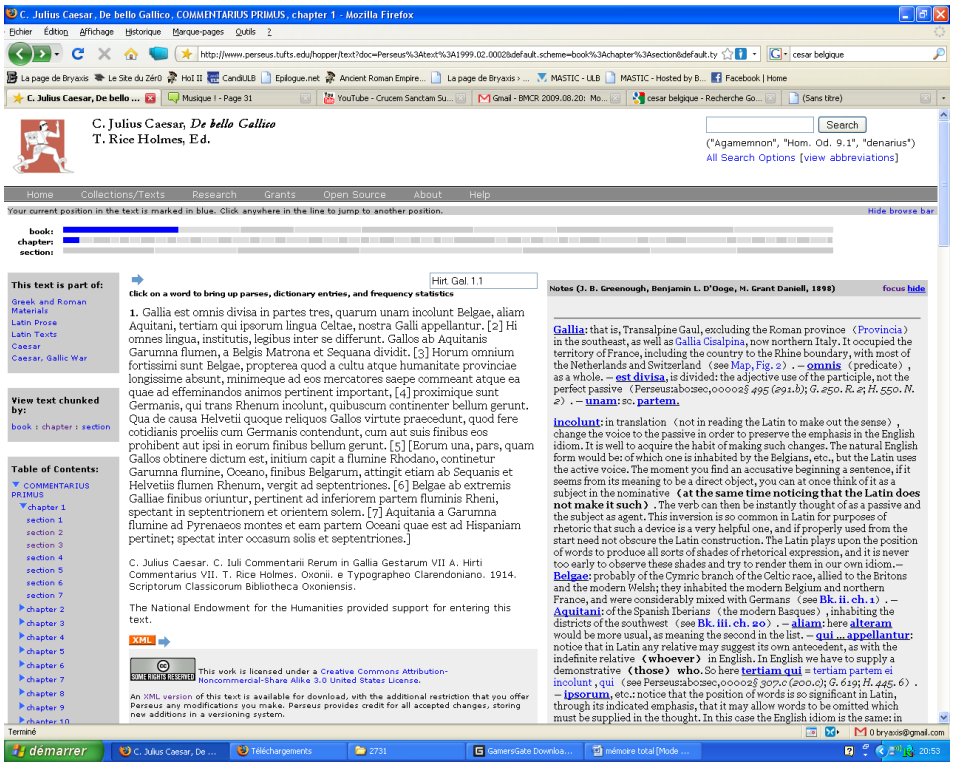
La quantité de matériel digitalisé est impressionnante : près de 61 millions pour la seule collection de textes greco-romains (exception faite des papyri). En conséquence la navigation se fait par collections, même si les moteurs de recherche permettent de transcender ces limitations.

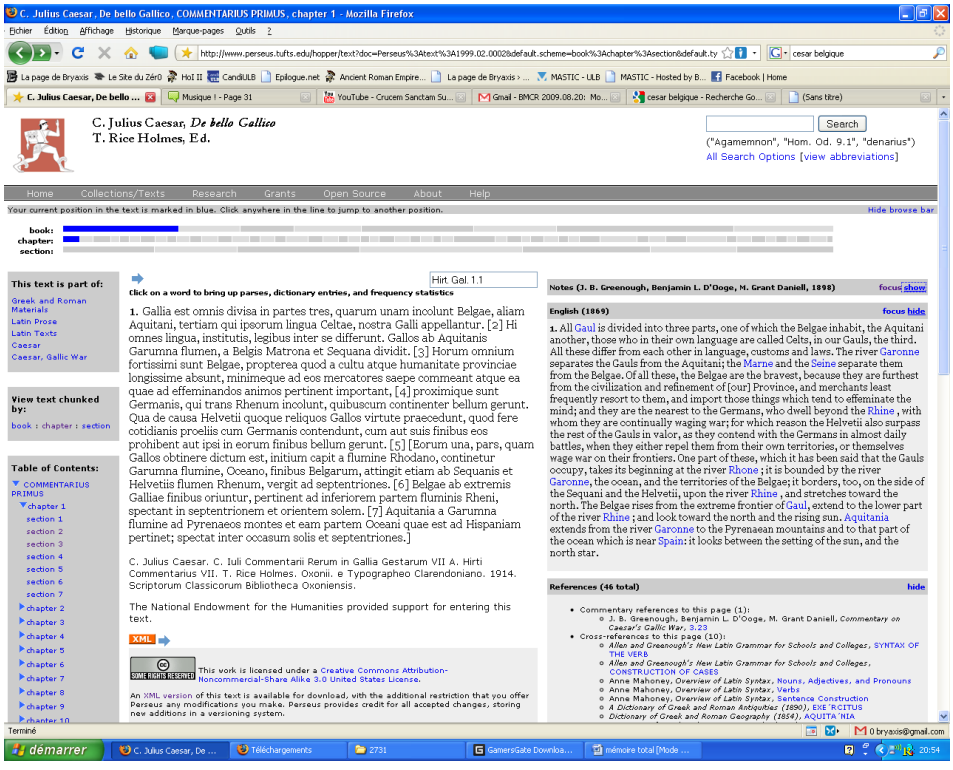
[65] http://www.tei-c.org/Activities/Projects/pe01.xml

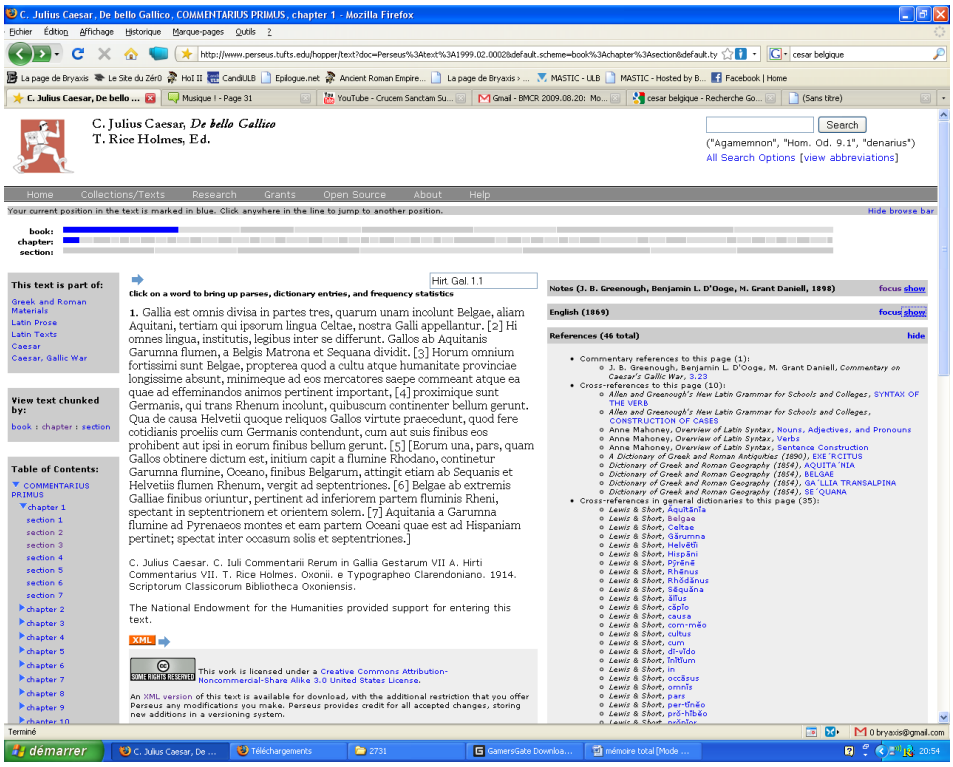
Par ailleurs, les collections ne rendent pas bien l'immense quantité de documents digitalisés : on ne trouvera ainsi aucun lien direct vers le dictionnaire *Liddel, Scott, Jones* (LSJ) : ses informations sont directement accessibles en cliquant sur n'importe quel mot grec : le système affichera l'entrée du LSJ (comprenant la traduction, la forme normalisée du mot et ses cas, genre, nombre, temps, etc.), les statistiques de fréquence calculées par corpus,…

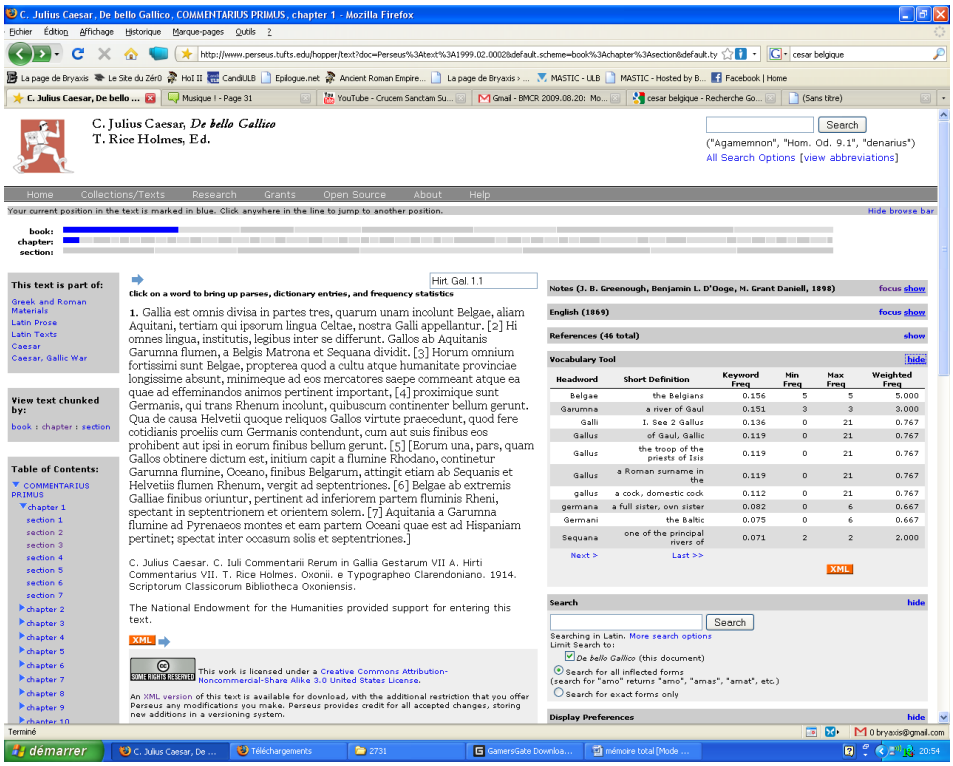


Les collections se présentent comme de longues listes de contenus. Comme on le note sur cet exemple, le contenu est classé par nom de l'auteur du texte antique : sous son nom sont classés la transcription de son texte, une ou plusieurs traduction(s) et divers ouvrages complémentaires faisant partie de la tradition moderne. Ces ouvrages ne sont pas de simples numérisations mais bien des digitalisations comprenant un marquage permettant d'accéder aux ressources adéquates sur le site (par exemple les extraits de textes anciens cités dans un commentaire).


C. Julius Caesar, De bello Gallico, COMMENTARIUS PRIMUS, chapter 1 - Mozilla Firefox
C. Julius Caesar, De bello Gallico
T. Rice Holmes, Ed.
Search
Home
Collections/Texts
Research
Grants
Open Source
About
Help
This text is part of:
Table of Contents:
Notes (J. B. Greenough, Benjamin L. D'Ooge, M. Grant Daniell, 1898)


C. Julius Caesar, De bello Gallico, COMMENTARIUS PRIMUS, chapter 1 - Mozilla Firefox
C. Julius Caesar, De bello Gallico
T. Rice Holmes, Ed.
Search
Home
Collections/Texts
Research
Grants
Open Source
About
Help
This text is part of:
Table of Contents:
Notes (J. B. Greenough, Benjamin L. D'Ooge, M. Grant Daniell, 1898)
English (1869)
References (46 total)


C. Julius Caesar, De bello Gallico, COMMENTARIUS PRIMUS, chapter 1 - Mozilla Firefox
C. Julius Caesar, De bello Gallico
T. Rice Holmes, Ed.
Search
("Agamemnon", "Hom. Od. 9.1", "denarius")
All Search Options [view abbreviations]
Home
Collections/Texts
Research
Grants
Open Source
About
Help
Your current position in the text is marked in blue. Click anywhere in the line to jump to another position.
Hide browse bar
book:
chapter:
section:
This text is part of:
Greek and Roman Materials
Latin Prose
Latin Texts
Caesar
Caesar, Gallic War
View text chunked by:
book : chapter : section
Table of Contents:
COMMENTARIUS PRIMUS
chapter 1
section 1
section 2
section 3
section 4
section 5
section 6
section 7
chapter 2
chapter 3
chapter 4
chapter 5
chapter 6
chapter 7
chapter 8
chapter 9
chapter 10
Hirt. Gal. 1.1
Click on a word to bring up parses, dictionary entries, and frequency statistics
1. Gallia est omnis divisa in partes tres, quarum unam incolunt Belgae, aliam Aquitani, tertiam qui ipsorum lingua Celtae, nostra Galli appellantur. [2] Hi omnes lingua, institutis, legibus inter se differunt. Gallos ab Aquitanis Garumna flumen, a Belgis Matrona et Sequana dividit. [3] Horum omnium fortissimi sunt Belgae, propterea quod a cultu atque humanitate provinciae longissime absunt, minimeque ad eos mercatores saepe commeant atque ea quae ad effeminandos animos pertinent important, [4] proximique sunt Germanis, qui trans Rhenum incolunt, quibuscum continenter bellum gerunt. Qua de causa Helvetii quoque reliquos Gallos virtute praecedunt, quod fere cotidianis proeliis cum Germanis contendunt, cum aut suis finibus eos prohibent aut ipsi in eorum finibus bellum gerunt. [5] [Eorum una, pars, quam Gallos obtinere dictum est, initium capit a flumine Rhodano, continetur Garumna flumine, Oceano, finibus Belgarum, attingit etiam ab Sequanis et Helvetiis flumen Rhenum, vergit ad septentriones. [6] Belgae ab extremis Galliae finibus oriuntur, pertinent ad inferiorem partem fluminis Rheni, spectant in septentrionem et orientem solem. [7] Aquitania a Garumna flumine ad Pyrenaeos montes et eam partem Oceani quae est ad Hispaniam pertinet; spectat inter occasum solis et septentriones.]
C. Julius Caesar. C. Iuli Commentarii Rerum in Gallia Gestarum VII A. Hirti Commentarius VII. T. Rice Holmes. Oxonii. e Typographeo Clarendoniano. 1914. Scriptorum Classicorum Bibliotheca Oxoniensis.
The National Endowment for the Humanities provided support for entering this text.
XML
SOME RIGHTS RESERVED
This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 United States License.
An XML version of this text is available for download, with the additional restriction that you offer Perseus any modifications you make. Perseus provides credit for all accepted changes, storing new additions in a versioning system.
Notes (J. B. Greenough, Benjamin L. D'Ooge, M. Grant Daniell, 1898)
focus show
English (1869)
focus show
References (46 total)
hide
Commentary references to this page (1):
J. B. Greenough, Benjamin L. D'Ooge, M. Grant Daniell, Commentary on Caesar's Gallic War, 3.22
Cross-references to this page (10):
Allen and Greenough's New Latin Grammar for Schools and Colleges, SYNTAX OF THE VERB
Allen and Greenough's New Latin Grammar for Schools and Colleges, CONSTRUCTION OF CASES
Anne Mahoney, Overview of Latin Syntax, Nouns, Adjectives, and Pronouns
Anne Mahoney, Overview of Latin Syntax, Verbs
Anne Mahoney, Overview of Latin Syntax, Sentence Construction
A Dictionary of Greek and Roman Antiquities (1890), EQUITES
Dictionary of Greek and Roman Geography (1854), AQUITA'NIA
Dictionary of Greek and Roman Geography (1854), BELGAE
Dictionary of Greek and Roman Geography (1854), GA'LLIA TRANSALPINA
Dictionary of Greek and Roman Geography (1854), SE'QUANA
Cross-references in general dictionaries to this page (35):
Lewis & Short, Aquītānĭa
Lewis & Short, Belgae
Lewis & Short, Celtae
Lewis & Short, Gărumna
Lewis & Short, Helvētĭi
Lewis & Short, Hispāni
Lewis & Short, Pȳrēnē
Lewis & Short, Rhēnus
Lewis & Short, Rhŏdănus
Lewis & Short, Sĕquāna
Lewis & Short, ăllus
Lewis & Short, căpĭo
Lewis & Short, causa
Lewis & Short, com-mĕo
Lewis & Short, cultus
Lewis & Short, cum
Lewis & Short, di-vĭdo
Lewis & Short, Initium
Lewis & Short, in
Lewis & Short, occāsus
Lewis & Short, omnis
Lewis & Short, pars
Lewis & Short, per-tĭnĕo
Lewis & Short, prō-mĭnĕo
démarrer
Terminé


C. Julius Caesar, De bello Gallico, COMMENTARIUS PRIMUS, chapter 1 - Mozilla Firefox
C. Julius Caesar, De bello Gallico
T. Rice Holmes, Ed.
Search
("Agamemnon", "Hom. Od. 9.1", "denarius")
All Search Options [view abbreviations]
Home
Collections/Texts
Research
Grants
Open Source
About
Help
Your current position in the text is marked in blue. Click anywhere in the line to jump to another position.
Hide browse bar
book:
chapter:
section:
This text is part of:
Greek and Roman Materials
Latin Prose
Latin Texts
Caesar
Caesar, Gallic War
View text chunked by:
book : chapter : section
Table of Contents:
COMMENTARIUS PRIMUS
chapter 1
section 1
section 2
section 3
section 4
section 5
section 6
section 7
chapter 2
chapter 3
chapter 4
chapter 5
chapter 6
chapter 7
chapter 8
chapter 9
chapter 10
Hirt. Gal. 1.1
Click on a word to bring up parses, dictionary entries, and frequency statistics
1. Gallia est omnis divisa in partes tres, quarum unam incolunt Belgae, aliam Aquitani, tertiam qui ipsorum lingua Celtae, nostra Galli appellantur. [2] Hi omnes lingua, institutis, legibus inter se differunt. Gallos ab Aquitanis Garumna flumen, a Belgis Matrona et Sequana dividit. [3] Horum omnium fortissimi sunt Belgae, propterea quod a cultu atque humanitate provinciae longissime absunt, minimeque ad eos mercatores saepe commeant atque ea quae ad effeminandos animos pertinent important, [4] proximique sunt Germanis, qui trans Rhenum incolunt, quibuscum continenter bellum gerunt. Qua de causa Helvetii quoque reliquos Gallos virtute praecedunt, quod fere cotidianis proeliis cum Germanis contendunt, cum aut suis finibus eos prohibent aut ipsi in eorum finibus bellum gerunt. [5] [Eorum una, pars, quam Gallos obtinere dictum est, initium capit a flumine Rhodano, continetur Garumna flumine, Oceano, finibus Belgarum, attingit etiam ab Sequanis et Helvetiis flumen Rhenum, vergit ad septentriones. [6] Belgae ab extremis Galliae finibus oriuntur, pertinent ad inferiorem partem fluminis Rheni, spectant in septentrionem et orientem solem. [7] Aquitania a Garumna flumine ad Pyrenaeos montes et eam partem Oceani quae est ad Hispaniam pertinet; spectat inter occasum solis et septentriones.]
C. Julius Caesar. C. Iuli Commentarii Rerum in Gallia Gestarum VII A. Hirti Commentarius VII. T. Rice Holmes. Oxonii. e Typographeo Clarendoniano. 1914. Scriptorum Classicorum Bibliotheca Oxoniensis.
The National Endowment for the Humanities provided support for entering this text.
XML
SOME RIGHTS RESERVED
This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 United States License.
An XML version of this text is available for download, with the additional restriction that you offer Perseus any modifications you make. Perseus provides credit for all accepted changes, storing new additions in a versioning system.
Notes (J. B. Greenough, Benjamin L. D'Ooge, M. Grant Daniell, 1898)
focus show
English (1869)
focus show
References (46 total)
show
Vocabulary Tool
hide
Headword
Short Definition
Keyword Freq
Min Freq
Max Freq
Weighted Freq
Belgae
the Belgians
0.156
5
5
5.000
Garumna
a river of Gaul
0.151
3
3
3.000
Galli
1. See 2 Gallus
0.136
0
21
0.767
Gallus
of Gaul, Gallic
0.119
0
21
0.767
Gallus
the troop of the priests of Isis
0.119
0
21
0.767
Gallus
a Roman surname in the
0.119
0
21
0.767
gallus
a cock, domestic cock
0.112
0
21
0.767
germana
a full sister, own sister
0.082
0
6
0.667
Germani
the Baltic
0.075
0
6
0.667
Sequana
one of the principal rivers of
0.071
2
2
2.000
Next >
Last >>
XML
Search
hide
Search
Searching in Latin. More search options
Limit Search to:
De bello Gallico (this document)
Search for all inflected forms
(search for "amo" returns "amo", "amas", "amat", etc.)
Search for exact forms only
Display Preferences
hide
démarrer
Terminé

Cette longue série de captures d'écran montre l'ensemble des types d'information accessibles pour chaque texte antique digitalisé : outre un commentaire descriptif, le site offre des traductions (parfois plusieurs versions), des informations bibliographiques et lexicographique, des aides pédagogiques (notamment des aides de vocabulaire pour ceux souhaitant traduire le texte sans s'aider de la traduction fournie), et un puissant moteur de recherche permettant notamment de retrouver toutes les occurrences d'un mot sans tenir compte de sa flexion.

Cela demande évidemment, en amont de la digitalisation, un travail considérable de compilation de données qui seront exploitées pour accomplir ce type de recherches, quelque chose que l'équipe du projet avait compris très tôt puisque des ressources comme le LSJ avaient été digitalisées dés 1995.

Par rapport aux autres sites on note que Perseus fournit une information complète sur les sources qu'il utilise pour ses projets de digitalisation : on ne lit pas juste le texte de César mais bien l'édition que T. Rice Holmes en a fait en 1914.

En revanche cet exemple montre également l'une des limitations de Perseus : si le nombre de textes fournis est impressionnant et si le marquage qui leur a été appliqué fournit une quantité exceptionnelle d'informations sur chacun d'entre eux, il n'en reste pas moins que des questions de droit d'auteurs font que beaucoup de ces éditions sont relativement anciennes.

Par ailleurs il faut noter que sur le plan technique le site de Perseus s'avère très lent, tant lors des recherches que pour le simple affichage des pages.

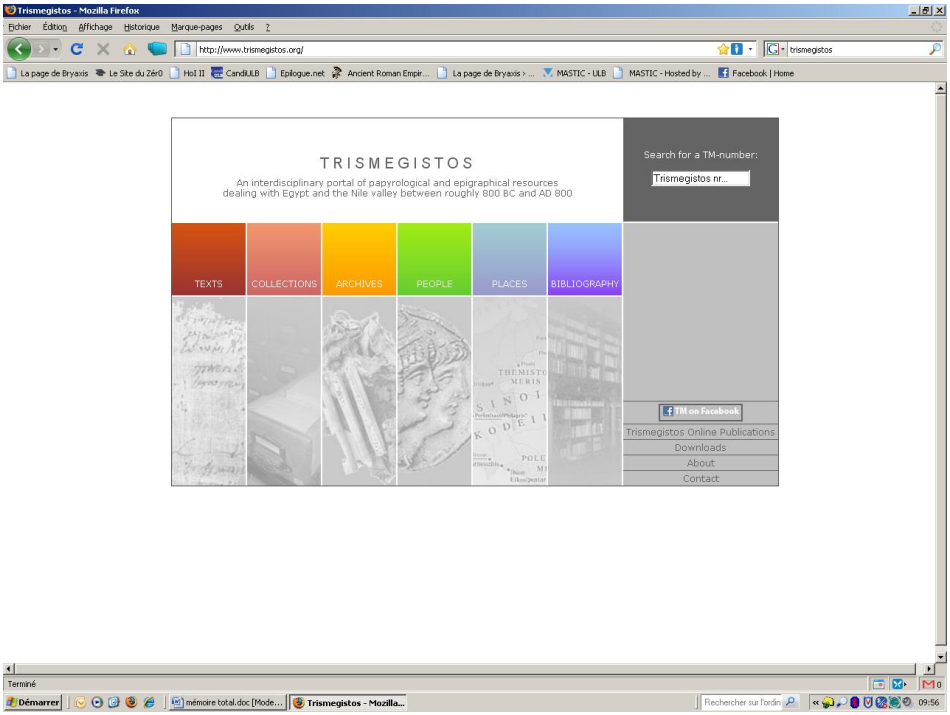
Deux raisons principales à ce phénomène : tout d'abord la masse énorme de marquages que porte chaque page et que le serveur doit traiter à chaque appel afin d'assurer un rendu adéquat

Ensuite l'immense popularité du site, notamment auprès des étudiants du secondaire ou des bacheliers qui, en Angleterre ou aux USA, se servent de Perseus pour faire leurs traductions juste avant d'entrer en cours… Cette utilisation sans doute non prévue par les initiateurs du projet a pour conséquence que le site est parfois inutilisable à certaines heures en Europe.

Perseus présente une interface moderne, quoique parfois un peu surchargée, composée de différents modules qui peuvent être agrandis ou réduits selon l'intérêt de l'utilisateur. Ces modules sont en fait des widgets attachés à chaque page qui vont puiser l'information qui les intéresse.

Ce design de Perseus fut implémenté en 2005 lors de la parution de la v4.0 du projet et rencontra rapidement un large succès, devenant l'une des sources d'inspiration du projet Papyrological Navigator, projet dont nous reparlerons dans quelques instants.

Trismegistos



Nous l'avons vu dans l'historique, la Belgique francophone n'est pas la dernière des nations en terme d'utilisation des ressources informatiques pour l'étude des langues anciennes même si l'épigraphie et la papyrologie n'y ont pas bénéficié des mêmes attentions. Le site des philologues classiques de l'Université Catholique de Louvain[66] fournit avant tout des retranscriptions de quelques auteurs et la traduction de leurs œuvres avec accès[67] à une base de données fournissant un service d'analyse de chaque mot similaire à celui offert par Perseus.

Si la Wallonie est en état intellectuel de se servir d'un ordinateur, il en va de même pour la Flandre : comme l'avait dit César les Belges sont les plus forts[68].

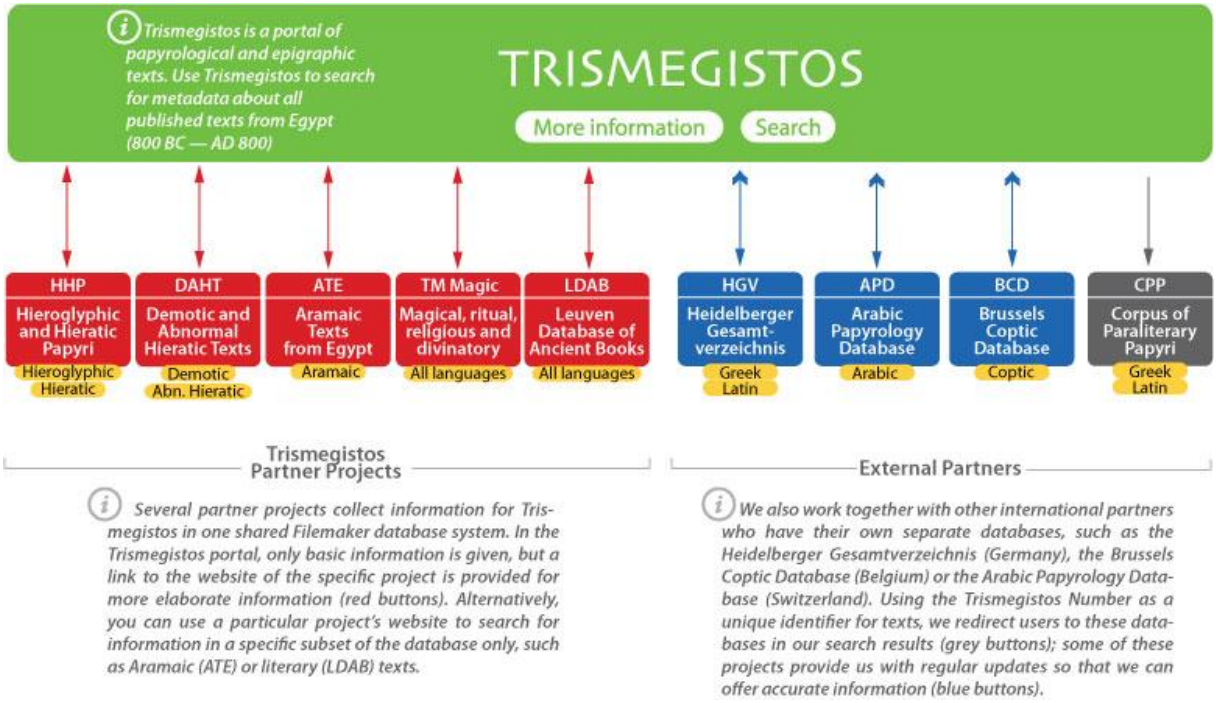
En effet la Katholiek Universiteit Leuven est à l'origine d'un impressionnant projet du nom de Trismegistos[69] visant à rassembler sur la toile le plus grand nombre possible de ressources relatives à la papyrologie, plus particulièrement celles touchant à l'étude des documents remontant à la période greco-romaine.

[66] http://bcs.fltr.ucl.ac.be/
[67] Non opérationnel lors de notre dernière visite du site le 9 août 2009
[68] Les philologues de la salle me pardonneront d'avoir un peu forcé la traduction du passage I.3 de la Guerre des Gaules de César…
[69] http://www.trismegistos.org/index.html

Au cœur du projet se trouve une banque de données reprenant tous les textes pertinents rédigés dans plus d'une demi-douzaine de langues anciennes : plus de 110000 entrées initialement encodées dans de nombreuses bases de données et rassemblées afin d'être interrogées par un moteur de recherche unique.

Ainsi les chercheurs peuvent, pour la première fois, parcourir en une seule opération la documentation issue d'Egypte, que son auteur ait été d'expression grecque, latine, hiéroglyphique, hiératique, démotique, copte, méroïtique, araméenne, arabe, nabatéenne ou encore carienne. Le site indique même la présence d'autres langues dans sa base de données…

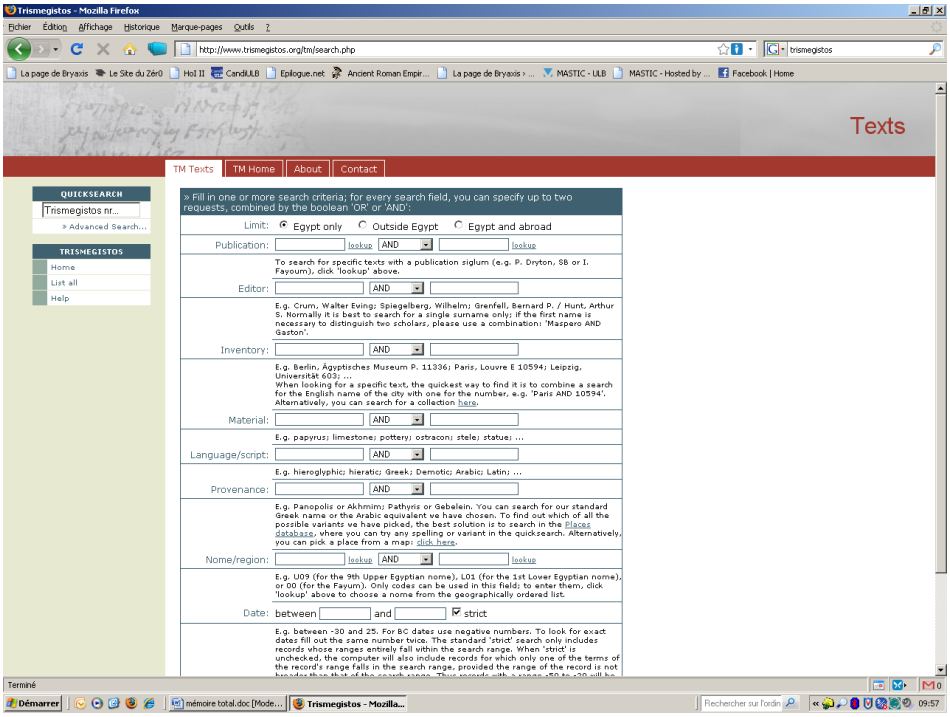
Un schéma disponible sur le site du projet en décrit bien l'architecture[70] :

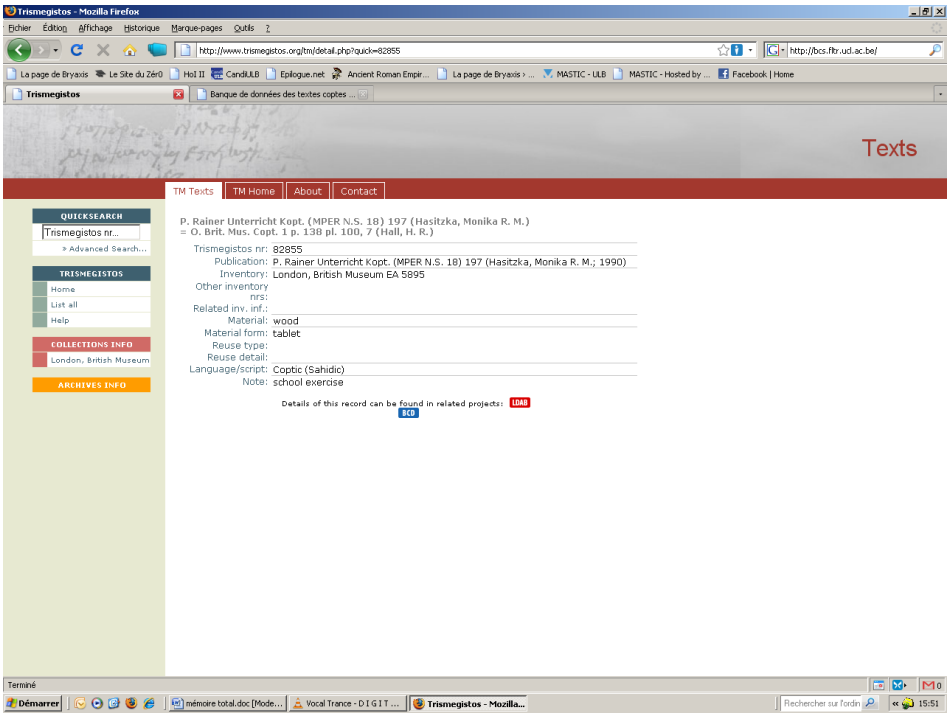


On retrouve parmi les partenaires de Trismegistos des ressources que nous avons déjà évoquées ou que nous devons encore étudier : le HGV bien sûr, et à travers lui la DDbDP et APIS, et la Banque de Données des Inscriptions Coptes Documentaires d'Alain Delattre.

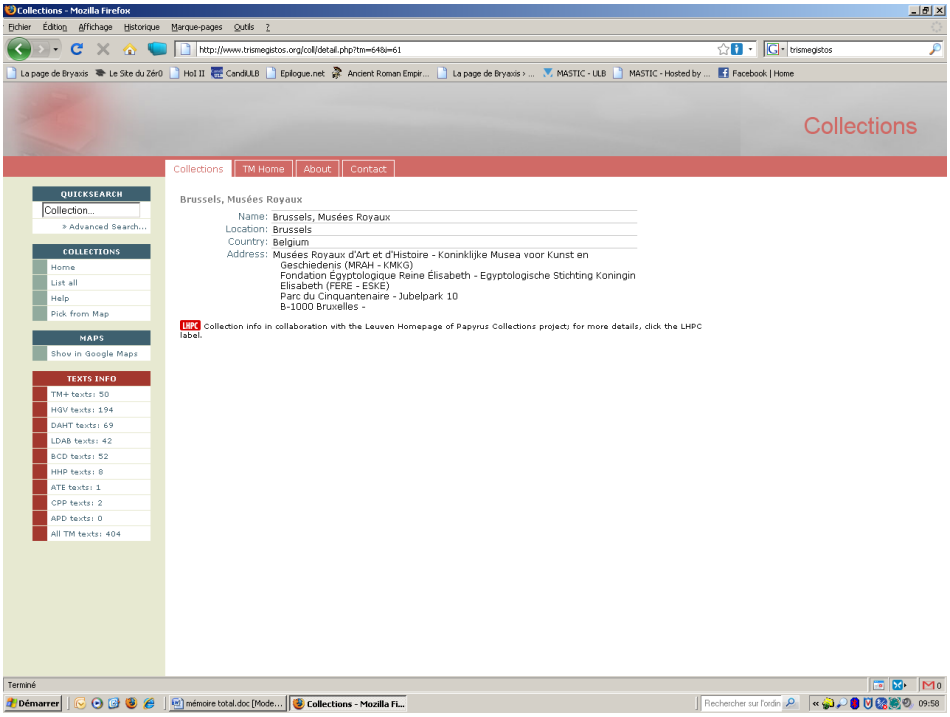
La page d'accueil permet d'accéder aux différentes ressources du site : recherche sur les textes, recherche sur les collections, recherche sur les archives anciennes (regroupées indépendamment du lieu où les papyrus les composant sont conservés), recherches sur des éléments géographiques ou prosopographiques, recherches bibliographiques.

[70] http://www.trismegistos.org/index2.php





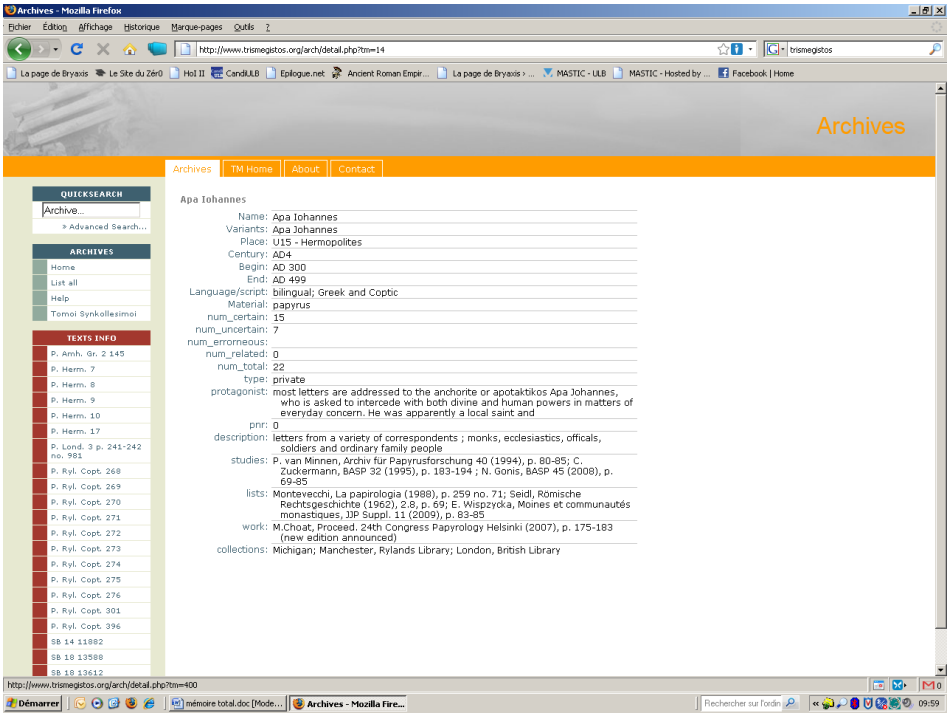
La recherche de base est évidement la recherche sur les textes eux-mêmes. Le formulaire de recherche est très complet et fournit des précisions quant à l'usage de chaque champ. Des listes pré-compilées de valeurs possibles sont également disponibles pour certains champs (notamment ceux relatifs aux noms des collections ou des revues). Elle fournit une fiche sur laquelle se trouve les informations disponibles et des liens vers les bases de données liées à Trismegistos et susceptibles de contenir plus d'informations, ici notamment la banque de données d'Alain Delattre.



La seconde catégorie de recherches permet de connaître le contenu des collections conservées aux quatre coins du monde. Mais la requête permet également de voir les sites internet disposant d'informations sur des textes contenus dans ces collections : ainsi 194 textes de la collection du Musée du Cinquantenaire sont répertoriés dans la base de données HGV et 52 dans celle de Alain Delattre, pour un total de 404 occurrences dans Trismegistos.

Bien sûr le système n'indique pas ici les doublons mais cela n'est pas important car chaque base de données peut avoir un contenu différent. On note aussi que l'information contenue dans Trismegistos comprend également l'adresse postale de la collection, permettant ainsi aux chercheurs de prendre contact avec les conservateurs des documents.

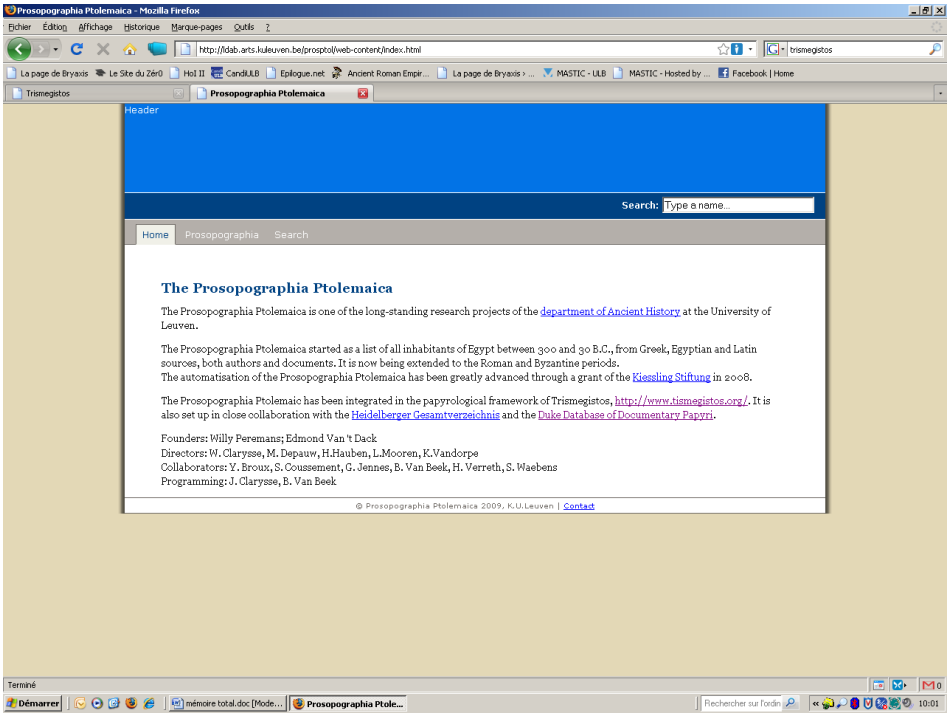
Cette information peut sembler banale mais présente un grand intérêt pour les chercheurs car elle permet d'éviter de longues recherches. L'option d'afficher l'information sur Google Maps est aussi une de ces petites innovations qui ne semblent pas importantes mais peuvent faire gagner du temps.

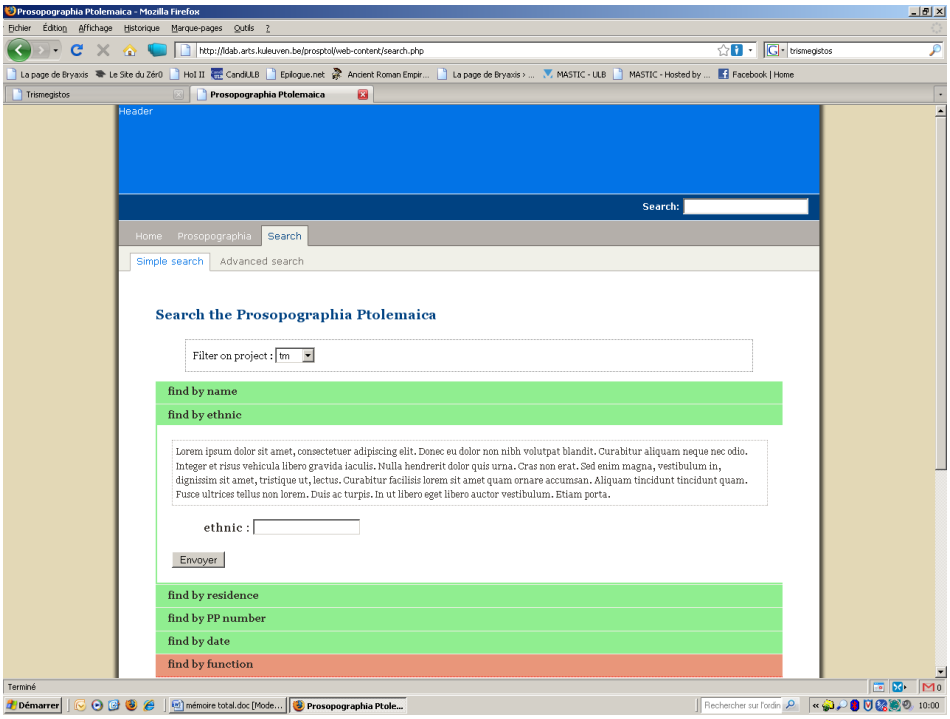


Troisième section du site, celle consacrée aux archives. Par archives, les papyrologues décrivent des ensembles documentaires relatifs à une personne unique ou à un lieu spécifique : les archives du comptable d'un grand domaine, les courriers officiels adressés par un notable d'Alexandrie, les « post-it » utilisés dans un couvent sont des exemples de ces archives qui peuvent être conservées dans plusieurs lieux différents en raison de leurs conditions de découverte, de leur histoire après excavation ou de toute autre facteur.

Cette forme d'indexation des documents est évidement très pratique pour le chercheur qui n'a pas à rechercher dans la littérature toutes les références à cette archive car l'information bibliographique pertinente est fournie dans la fiche Trismegistos qui lui est dédiée.

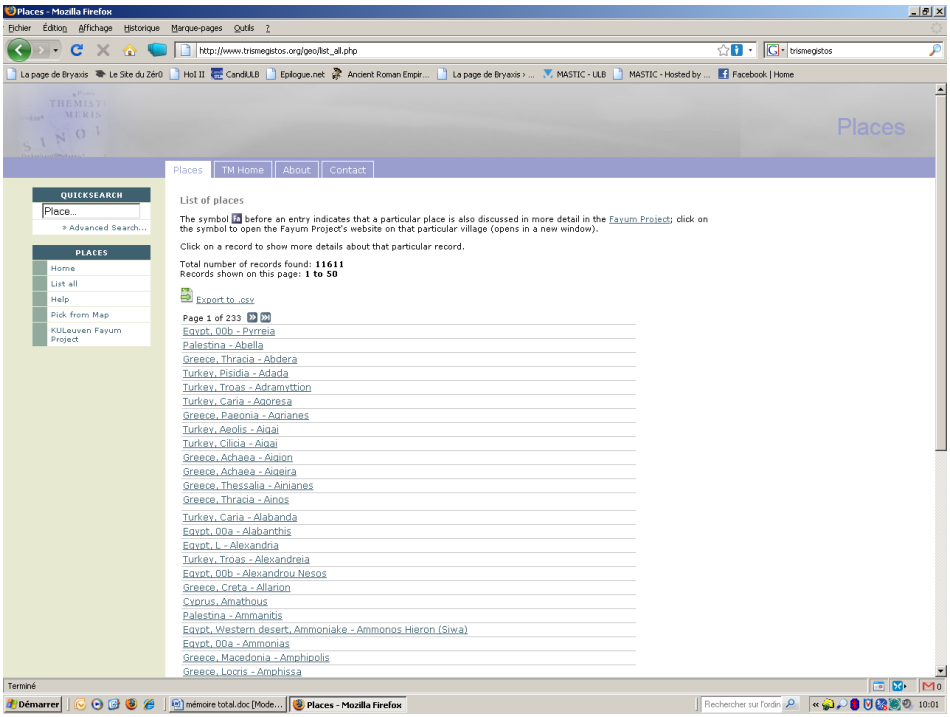
Ces archives sont accessibles principalement au moyen d'une grande liste même si la possibilité de procéder à une recherche précise est également possible.

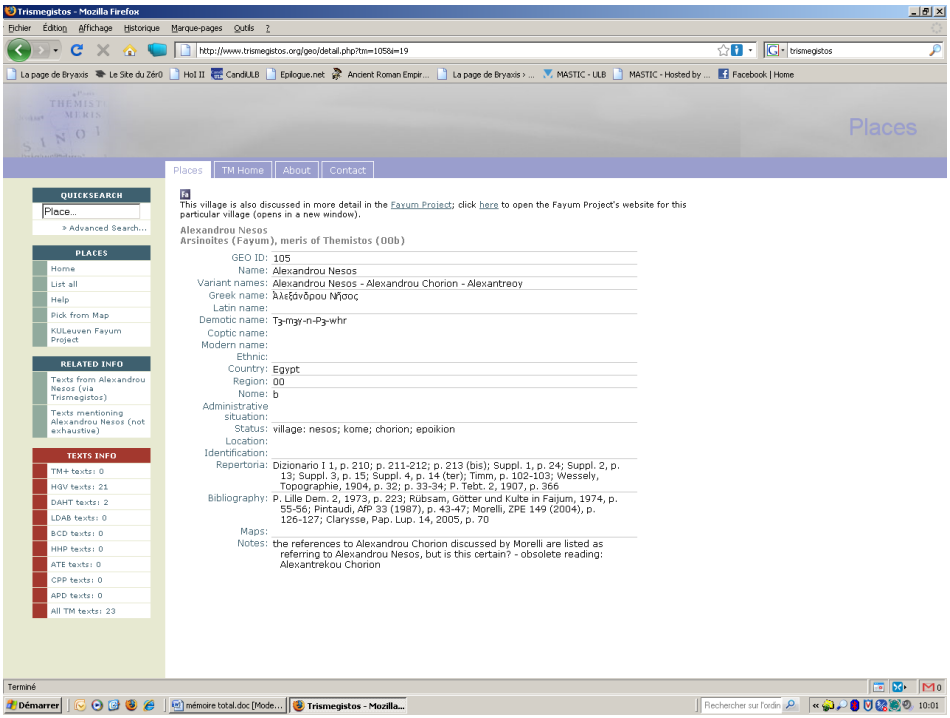




Encore en chantier, la section prosopographique du site permet d'identifier les individus et de retrouver toutes les informations qui leurs sont relatives. Avec ce genre d'outils et la quantité de données incorporées dans le système on approche du concept d'anthologie et ce d'autant plus que cette partie du projet Trismegistos est menée en coopération étroite avec les projets HGV et DDbDP : un marquage commun de l'information permet une interopérabilité entre

systèmes ouvrant de nouvelles perspectives de recherche et permettra sans doute de créer de nouvelles archives par le rapprochement de papyri dont on ne s'était pas encore rendu compte qu'ils parlaient de la même personne…

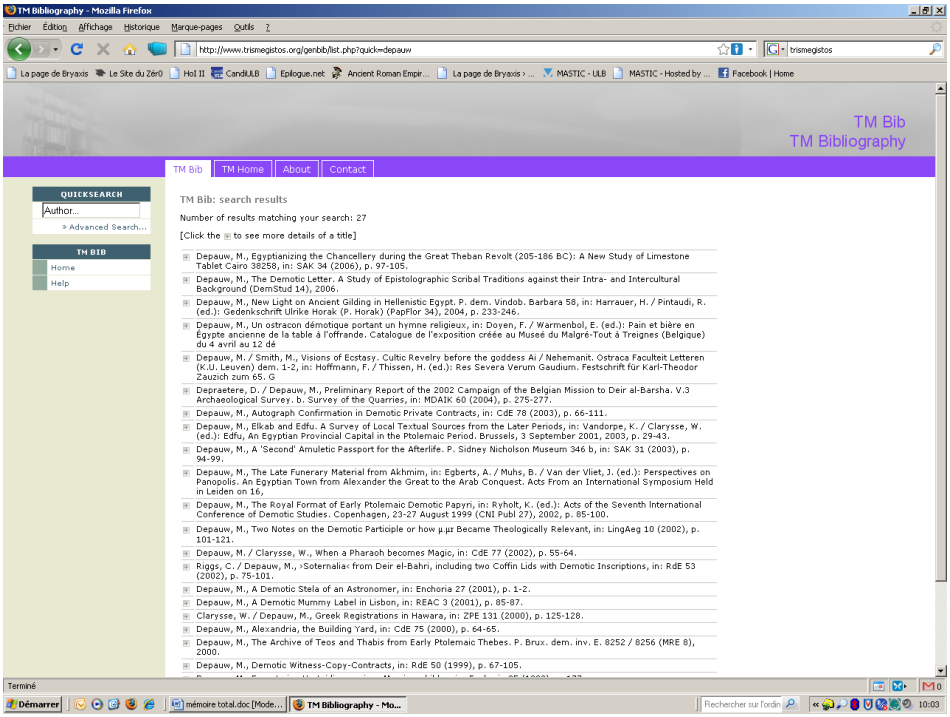




Dans le même ordre d'idée que pour l'information prosopographique, Trismegistos propose une information détaillée sur les toponymes identifiés dans la documentation. On note cependant que ce projet ne semble pas incorporer les données de localisation issues de la digitalisation du *Barrington Atlas of the Ancient World* et ne pas encore incorporer de module

Google Maps dans la fiche elle même. Il est toutefois possible d'accéder à une carte sur laquelle sont reportés les sites identifiés.

L'état actuel du module de prosopographie ne permet pas encore de savoir si l'équipe Trismegistos envisage de croiser les données des deux modules à l'image de ce que fait le projet LGPN[71]. Néanmoins on note que la donnée temporelle manque sur les fiches consacrées aux emplacements géographiques ce qui limite un peu les possibilités de fonctionnalités d'un outil automatique d'affichage des sites géographiques.



Dernier élément que nous présenterons au sujet de Trismegistos, le module de recherche bibliographique est aussi un outil très précieux sans doute destiné à devenir incontournable dans le domaine. L'enregistrement séparé de la bibliographie et des données relatives aux objets (lieux, personnes, textes, …) mentionnés dans les articles permet en effet des usages plus nombreux que le simple encodage des références bibliographiques dans la fiche de chaque texte, offrant notamment la possibilité de réutiliser l'information au lieu de la recopier.

Au final, avec Trismegistos, on observe l'un des outils les plus puissants du domaine des « Digital Humanities ».

On en viendrait même à se demander si cette concentration d'outils n'est pas une des premières représentations du web sémantique dans le domaine des sciences de l'Antiquité. Paul Bertrand prédisait en décembre que l'on verrait bientôt arriver des « ensembles structurés

[71] Cf. chapitre 4

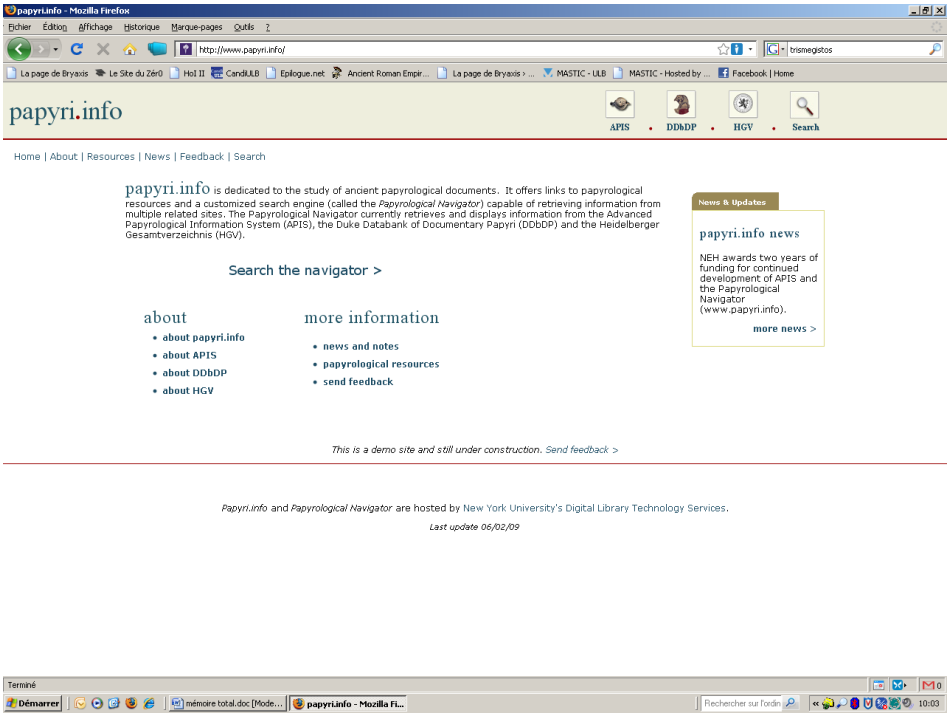
de termes et de concepts représentant le sens d'un champ d'informations »[72]. Avec Trismegistos le futur est peut-être déjà là.

En effet Trismegistos permet à la fois de représenter les concepts présents dans le domaine des études de l'Egypte gréco-romaine mais aussi les relations entre elles : certes cette structure n'est pas encore modélisée et l'on ne peut donc encore parler d'ontologie mais l'on peut considérer que les éléments d'information nécessaires sont dors et déjà présent.

Nous n'approfondirons cependant pas notre réflexion sur ce point car ce n'est pas là notre domaine d'expertise et cela nous semble sortir du cadre du présent mémoire. Il faudrait sans doute un mémoire ou une thèse pour déterminer si les concepts du web sémantique et des ontologies peuvent être appliqués au domaine des sciences de l'Antiquité, travail que nous laissons à nos successeurs.

[72] Bertrand P., « Autour de l'édition électronique et des digital humanities : nouvelle érudition, nouvelle critique ? » in MEFRA, à paraitre

Papyrological Navigator (PN) et Son of Suda Online (SoSOL)



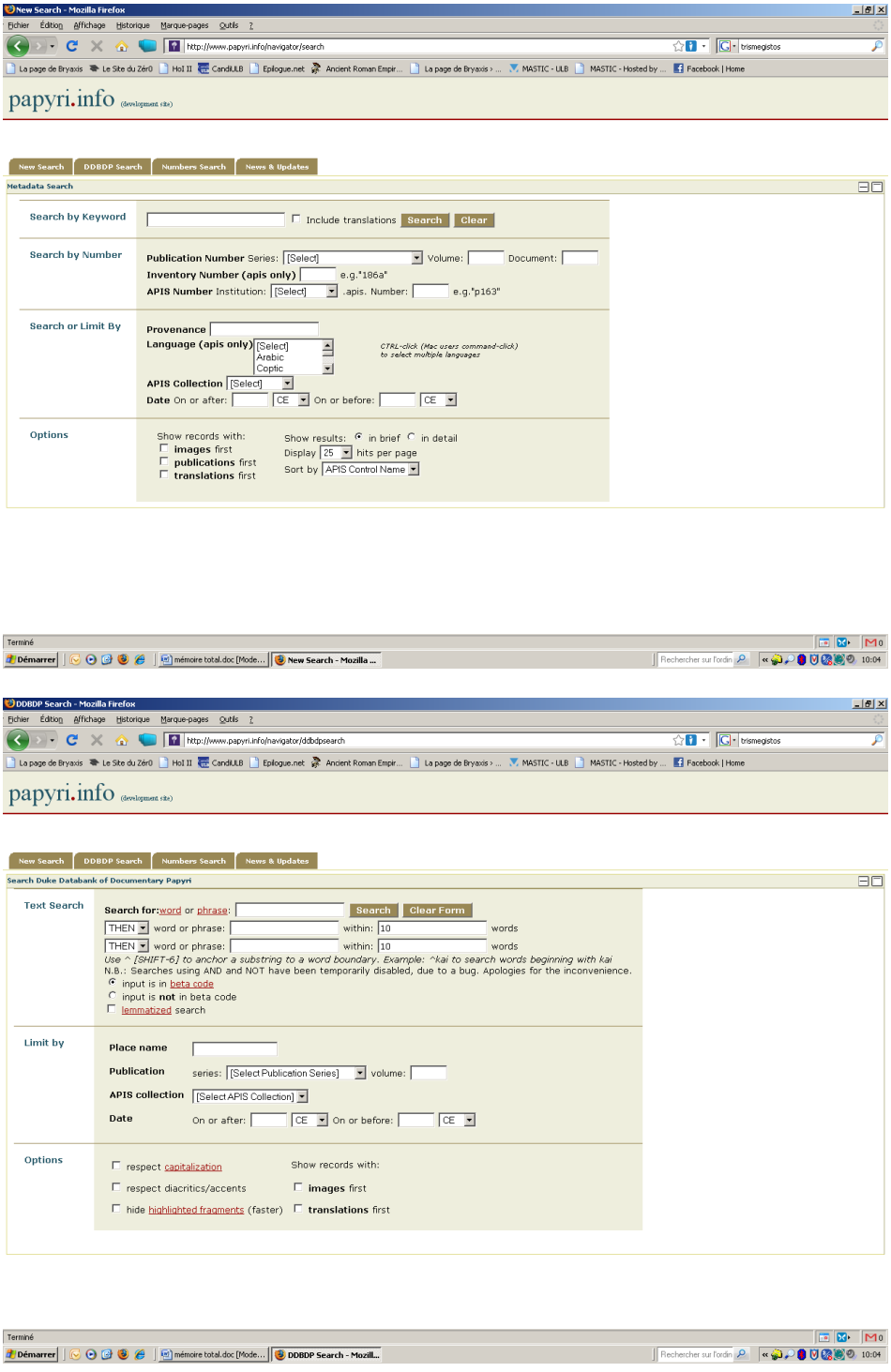
Dernier projet que nous présenterons dans ce chapitre, le Papyrological Navigator[73] est également le plus récent. Il est issu de la volonté d'intégrer de manière plus approfondie une série de ressources existantes dans le domaine de la papyrologie, en l'occurrence les textes de la DDbDP, les métadonnées et les photographies d'APIS et les métadonnées du projet HGV.

C'est en 2006 que débuta cette synergie qui visa dans un premier temps à la création d'une interface unique d'interrogation de ces ressources tout en modernisant l'infrastructure de la DDbDP. En effet, c'est aussi à cette époque que ce projet quitte le cadre de Perseus pour se joindre de manière plus efficace au projet APIS.

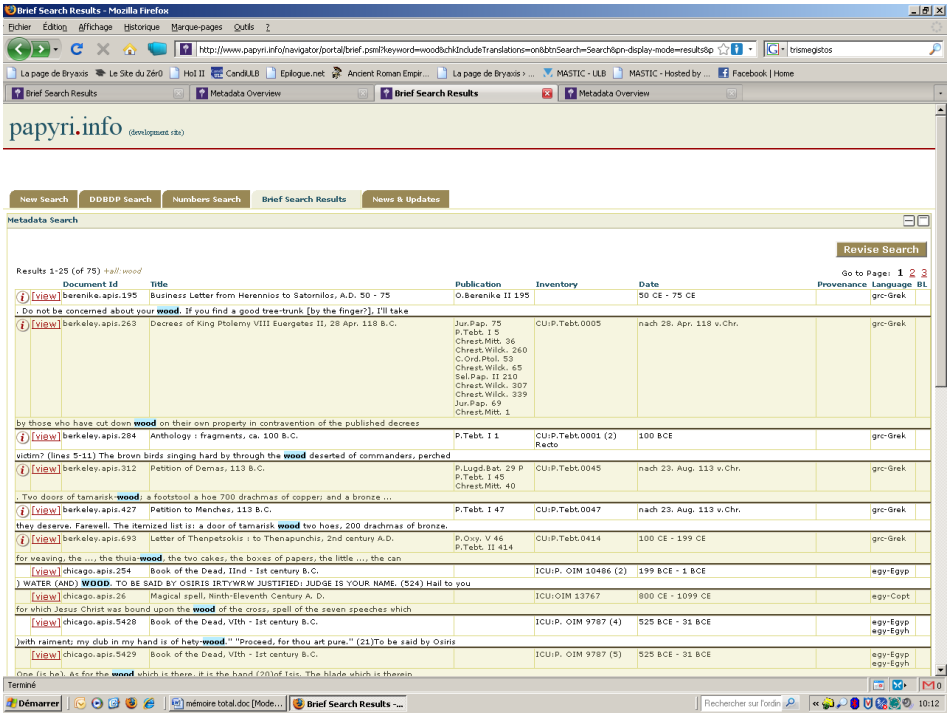
C'est dans ce contexte que l'équipe de la DDbDP décida d'adapter le standard de publication des documents épigraphiques EpiDoc à la publication de papyri et de convertir l'ensemble de sa base de données SGML vers ce format XML qui venait juste de faire ses preuves sur le projet Aphrodisias in Late Antiquity.

Le projet débouche rapidement sur un prototype capable d'afficher les données des différentes sources dans un environnement unique sur base d'une requête introduite dans le moteur de recherche du site.

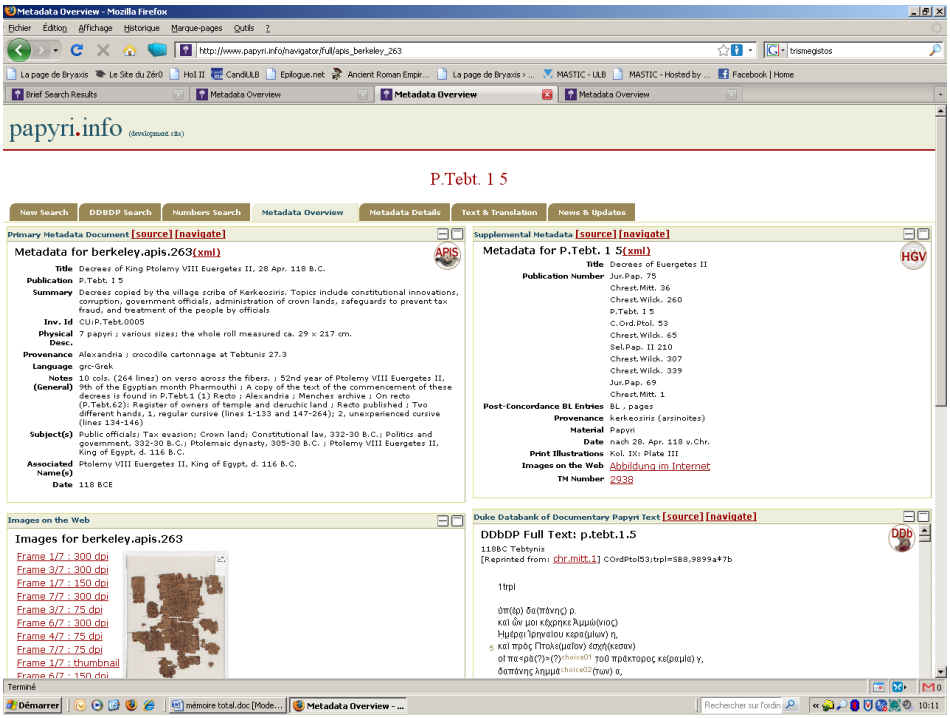
[73] http://www.papyri.info/

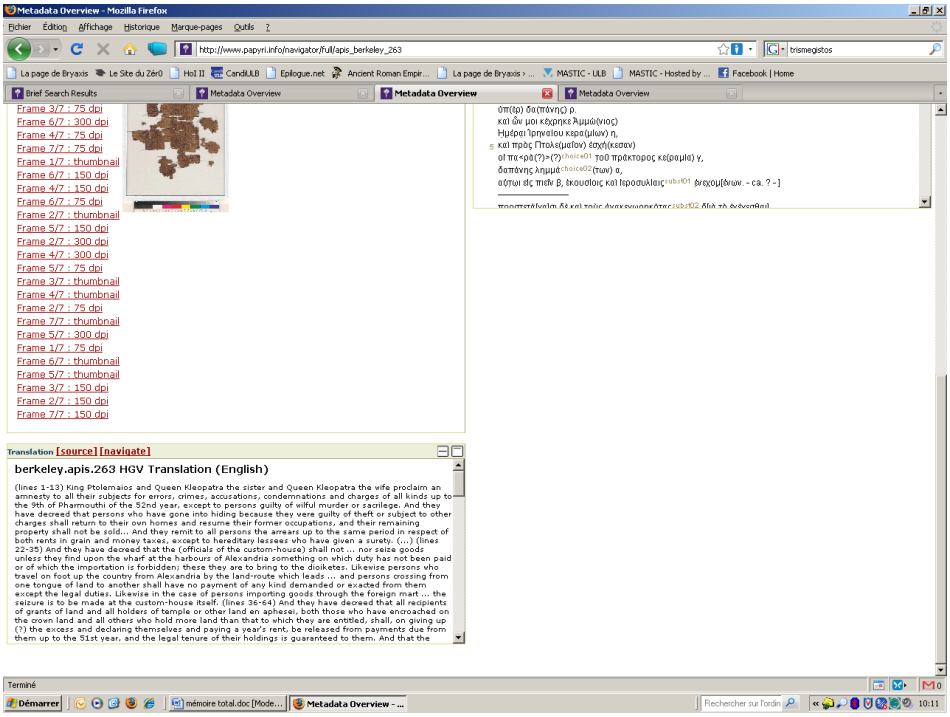


Le résultat de l'intégration des données est un moteur de recherche relativement complexe en raison des *caveat* présents sur certains champs qui ne sont pertinents que pour les données de certaines sources. Néanmoins l'outil est efficace et répond assez rapidement, bien plus vite en tout cas que ne le fait le moteur de recherche de Perseus.



Les résultats d'une recherche sont fournis sous la forme de tableaux donnant une information de base relativement importante sur chaque réponse.





Mais c'est surtout la fiche de résultats qui impressionne par la masse d'informations qu'elle délivre. Chaque information est livrée dans un portlet que l'utilisateur peut réduire au besoin, comme sur Perseus v4.0. Le module de visualisation des images est également très impressionnant, permettant d'agrandir l'image de manière à pouvoir décerner le moindre détail du document numérisé afin de le comparer avec son édition digitale.

Nous avons déjà évoqué au chapitre 3 les critiques que formulent M. Ansani et A. Ghignoli à l'encontre de la DDbDP, principalement au sujet de son manque de mise à jour des données encodées. Le problème est évidement reproduit ici, les données n'ayant pas été modifiées et/ou mises à jour. De plus il semble que le passage au format XML EpiDoc fut effectué à minima.

Pourtant il existe dans le domaine de la papyrologie un outil reprenant toutes les corrections apportées aux éditions de papyri, la *Berichtigungsliste der griechischen Papyrusukunden aus Aegypten* fondée en 1922 par Fr. Preisigker. Ce formidable outil de travail n'ayant cependant jamais été digitalisé, il n'existe pas de moyen de procéder à une mise à jour automatique de l'information par croisement des données.

Face à cette situation il a été décidé de lancer un nouveau projet basé sur le Papyrological Navigator : Integrated Digital Papyrology 2[74] a été proposé courant 2007 et lancé officiellement en 2008.

L'objectif principal du projet est de créer une interface de travail collaboratif permettant à tout contributeur volontaire d'introduire une édition d'un document qui sera ensuite validé par un comité éditorial. L'espoir est évidemment que la nouvelle structure de fonctionnement permette de rattraper le retard accumulé dans la digitalisation de nouveaux documents et de permettre l'affichage d'éditions nouvelles des textes déjà encodés.

En d'autres termes la vision des initiateurs du projet est de voir les notions du web collaboratif (ou du web 2.0) appliquées à un projet majeur des sciences de l'Antiquité. Ce ne serait pas une première : le nom même du projet, Son of Suda Online, montre clairement la filiation avec un projet entamé il y a plus de dix ans, la Suda Online[75], projet visant à digitaliser une énorme encyclopédie byzantine comptant près de 30 000 entrées et datant du Xème siècle.

[74] Bagnall R. et al., http://www.duke.edu/~jds15/IDP2-FinalProposalRedacted.pdf , 2008
[75] http://www.stoa.org/sol/

Chapitre 4 : Opportunités, contraintes et interrogations

Le choix du marquage : une décision influencée tant par les objectifs du projet que par son environnement.

L'utilité des fonctionnalités de marquage offertes par les formalismes TEI et EpiDoc doit être évaluée à la lumière du contexte dans lequel le projet s'exécute et des objectifs recherchés par le dit projet. En effet un marquage sémantique complet de données comme les noms propres identifiés au sein d'un document ou d'un corpus peut avoir pour objectif de créer des index et/ou de permettre l'ajout automatique de contenus, ce qui nécessite que ce contenu additionnel soit disponible par ailleurs.

Un tel marquage consomme également énormément de temps et peut aussi se révéler incompatible avec le marquage d'informations d'une autre nature. Ainsi un marquage orienté sur les caractéristiques grammaticales d'un ensemble de texte peut s'avérer difficile à faire cohabiter, sur une même copie du texte, avec un marquage destiné à faire ressortir la structure du discours ou des informations relatives aux noms propres identifiés dans le texte.

Un exemple de ce genre de textes susceptibles de nécessiter plusieurs balisages est celui des textes théâtraux : le formalisme XML permet de marquer les éléments purement linguistiques (orthographe, langue(s) utilisée(s), …), des éléments relatifs au sens des mots (identification de personnages masqués sous un pseudonyme, synonymes de termes archaïques pour aider à la lecture,…) ou la manière dont la pièce est jouée (intonations, évènements scéniques, …) : marquer tout cela peut s'avérer impossible en raison des règles du langage XML.

Pourtant les trois types d'informations peuvent être utiles à des chercheurs. Le formalisme TEI est ainsi utilisé pour retranscrire et marquer des enregistrements vidéo et audio, les rendant ainsi explorables par des moteurs de recherche[76].

Ces différents types de marquage impliquent également, en amont, le référencement des informations nécessaires. Or un tel travail peut s'avérer être une tâche considérable susceptible d'entrainer d'importantes hausses des coûts du programme, surtout lorsqu'il n'existe aucun ouvrage de référence. Et encore, dans le cas où de telles ressources existent,

[76] Cf. les guidelines TEI P5 http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DR.html

faut-il également qu'elles aient été digitalisées sous peine d'ajouter un travail important de digitalisation à la charge de l'équipe conduisant le projet.

Même dans le cas où existent des ressources digitales appropriées sur le plan de leur contenu, il n'est pas garanti que leur forme permette un usage aisé dans le cadre d'autres projets : absence d'identifiant unique permettant de pointer vers les données de manière unéquivoque, formats non échangeables ou technologies d'encodage périmées sont autant d'obstacles qui peuvent affecter la vie d'un projet.

Il faut donc également étudier l'environnement dans lequel le projet évoluera avant de choisir le type de marquage que l'on appliquera aux textes digitalisés. Mais encore faut-il pour ce faire, avoir fixé au projet des objectifs clairs qui aideront à identifier les ressources dont l'ajout est susceptible d'apporter une véritable plus value aux données digitalisées par le projet.

Si l'environnement présente de telles ressources, alors on pourra mettre en œuvre les opérations techniques permettant à terme d'obtenir ce que l'on appelle un *mashup* ou une application composite, une combinaison de plusieurs contenus ou services offerts par des logiciels hétérogènes dans le but d'en offrir de nouveaux.

Nous aborderons plus en détail la question de ces *mashups* lorsque nous aborderons l'exemple de la version digitale du Lexicon *of Greek Personal Names*.

L'automatisation du marquage

Il est évident que des raisons de coût poussent le chercheur à rechercher un marquage le plus automatisé possible des données. Cela est, dans une certaine mesure, envisageable. Néanmoins une telle automatisation est forcément limitée et va dépendre des objectifs du projet et/ou de la disponibilité d'informations externes accessibles par l'ordinateur.

Ainsi un marquage automatique de la grammaire d'un texte est-il envisageable en se servant de caractéristiques unéquivoques de chaque mot pour lui appliquer un certain nombre de caractéristiques.

Ainsi peut-on très simplement appliquer automatiquement la balise <w></w>[77] à tous les ensembles de caractères comprenant entre 1 et x unités séparées par un espace, sans faire appel à la moindre ressource externe. Cela ne demande en effet qu'un peu de programmation et des outils spécifiques ont d'ores et déjà été développés pour effectuer ce genre de marquage.

On peut également envisager de se référer à une concordance pour donner à chaque mot sa racine, sa traduction, etc. Mais cela implique de disposer d'une telle concordance ou de grammaires permettant l'application de ces règles, parfois complexes, au corpus étudié.

S'il est parfois facile d'identifier des éléments grammaticaux permettant ce marquage automatique, par exemple pour certaines formes de japonais archaïque, cela n'est pas forcément possible pour toutes les langues : la grammaire grecque ancienne par exemple rend ce genre d'opérations très difficile.

Enfin il faut bien admettre que la programmation de ce genre de balisage peut vite devenir relativement complexe et nécessiter un bon niveau en programmation. Si les gros projets bien financés peuvent accéder à ce degré de compétences, il n'en va pas de même pour des projets individuels conduits par des chercheurs peu ou pas soutenus par leur institution.

[77] Cf. TEI P5 Guidelines : http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-w.html

Mashups : l'exemple du Lexicon of Greek Personal Names et des projets « Inscriptions of Aphrodisias », « Inscriptions of Roman Cyrenaica » et « Inscriptions of Roman Tripolitania »

Un attribut identifiant de manière unique une donnée peut être soit un identifiant récupéré dans une autre base de données soit un identifiant propre au projet et renvoyant à une fiche qui comprendra par exemple des données bibliographiques ou une description extensive de la donnée.

Dans les deux cas un tel identifiant permet également de réaliser automatiquement des index ou des tables de concordances. Cet identifiant permet donc de réutiliser, dans différents contextes, la donnée.

Un exemple d'illustrant ce concept est celui du *Lexicon of Greek Personal Names*[78]. Ce projet fut entamé au Royaume-Uni sous forme papier dans les années 70 et rassemble un panel international de chercheurs dont l'objectif est :

« To collect and publish with documentation all known ancient Greek personal names (including non-Greek names recorded in Greek, and Greek names in Latin), drawn from all available sources (literature, inscriptions, graffiti, papyri, coins, vases and other artefacts), within the period from the earliest Greek written records down to, approximately, the sixth century A.D. »[79]

Ce projet a d'abord consisté en une base de données consultable uniquement de manière locale et destiné à la préparation d'ouvrages papier, puis connu diverses évolutions devant conduire fin des années 90 à l'adoption du formalisme XML TEI, une solution toujours en œuvre à l'heure actuelle.

Lancé en 1972 et informatisé pour la première fois à partir de 1975, le projet fut contraint par les progrès de la technologie et l'obsolescence des machines utilisées de remplacer à plusieurs reprises son infrastructure : en 1984, le projet décida d'utiliser un DBMS pour créer la première véritable banque de données du projet (le projet de 1975 ayant produit des données stockées sur bandes que l'on chargeait lorsque l'on souhaitait consulter une partie de

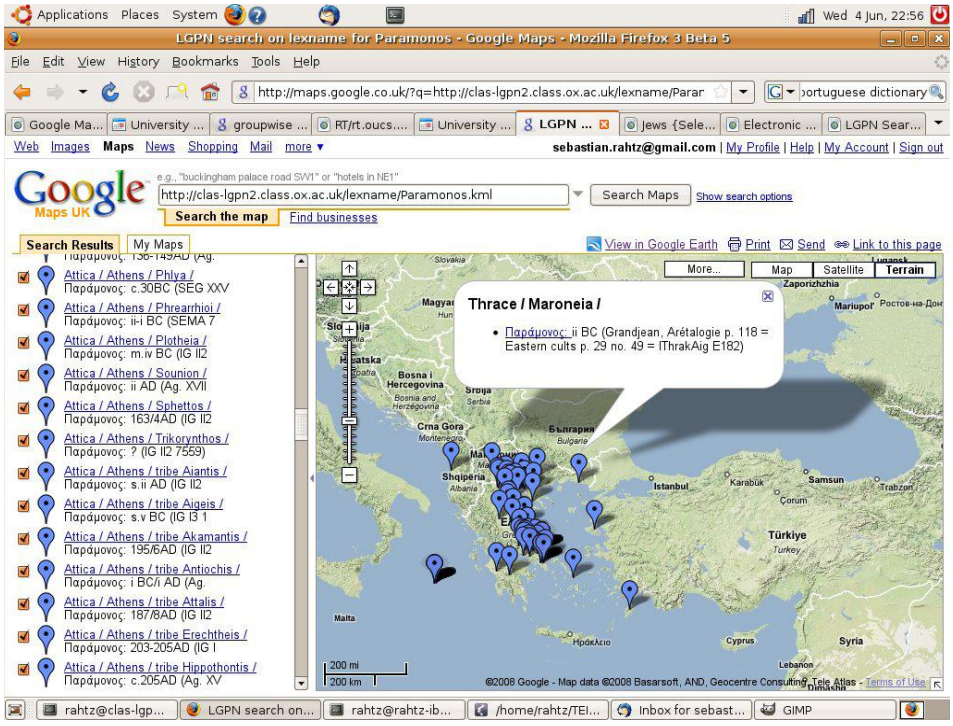
[78] http://www.lgpn.ox.ac.uk/
[79] http://www.lgpn.ox.ac.uk/project/index.html

l'information stockée) mais dû changer de système dés 1987 suite à la mise au rebut d'un ordinateur ICL 8859[80].

Ce changement hardware entraina également un changement software au niveau du DBMS. Dix ans plus tard le VAX 6620 sur lequel la base de données avait été transférée fut lui aussi retiré, entrainant un passage sous MS Access et le début du projet de conversion au format XML-TEI. C'est ce nouveau projet qui a conduit à la conversion des données au format UTF-8 et à la mise en œuvre du site web actuel.

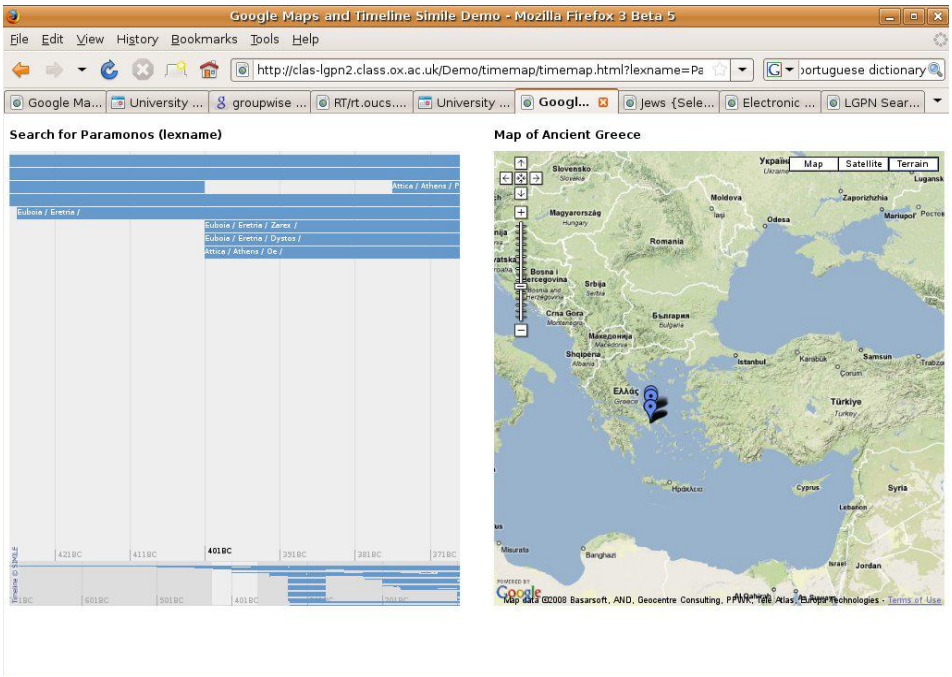
Cette nouvelle infrastructure a également permis le lancement de divers projets de *mashups* qui utilisent à leur tour différentes sources d'information en cours de digitalisation.

L'équipe du LGPN a ainsi créé des modules d'affichage, dans Google Maps, des lieux où les noms sont attestés, des lignes du temps montrant les périodes d'attestation de chaque nom, instance par instance, et finalise actuellement un nouveau *mashup* unissant toutes ces informations de manière dynamique comme l'a récemment présenté Sebastian Rahtz lors d'une conférence donnée dans le cadre de la TEI@Oxford Summer School [81].



[80] http://www.oucs.ox.ac.uk/internal/history/

[81] S. Rahtz, http://tei.oucs.ox.ac.uk/Oxford/2009-07-oxford/talk-23-tei_mashups.pdf , 2009



Les informations géographiques sont par ailleurs consolidées grâce à un travail parallèle conduit sur le *Barrington Atlas of the Ancient World* par l'équipe du projet *Pleiades*[82], sur base d'une idée développée par les équipes du projet *Concordia*[83].

Mais de la même manière que le projet LGPN commence à utiliser d'autres programmes pour améliorer l'information qu'il dispense, divers projets ont commencé à se servir du LGPN pour augmenter sémantiquement l'information contenue dans leurs bases de données.

L'information de base contenue dans la base de données du LGPN est assez limitée comme le montre un exemple tiré de la présentation de S. Rahtz[84] :

Fiche XML d'une entrée dans le LGPN

```
<person xml:id="V2-60057">
        <sex value="2"/>
        <persName type="full" nymRef="#Nani1s">
                <forename>Ναvίς</forename>
        </persName>
        <birth notAfter="-0350" notBefore="-0399">
                <placeName key="Athens" cert="low">Athens</placeName>
        </birth>
        <floruit>f.iv BC</floruit>
        <persName type="namevariant" xml:lang="grc">
                Να<seg type="orth">νείς</seg>
        </persName>
```

[82] http://pleiades.stoa.org/

[83] http://concordia.atlantides.org/

[84] S. Rahtz, http://tei.oucs.ox.ac.uk/Oxford/2009-07-oxford/talk-23-tei_mashups.pdf , 2009, page 5

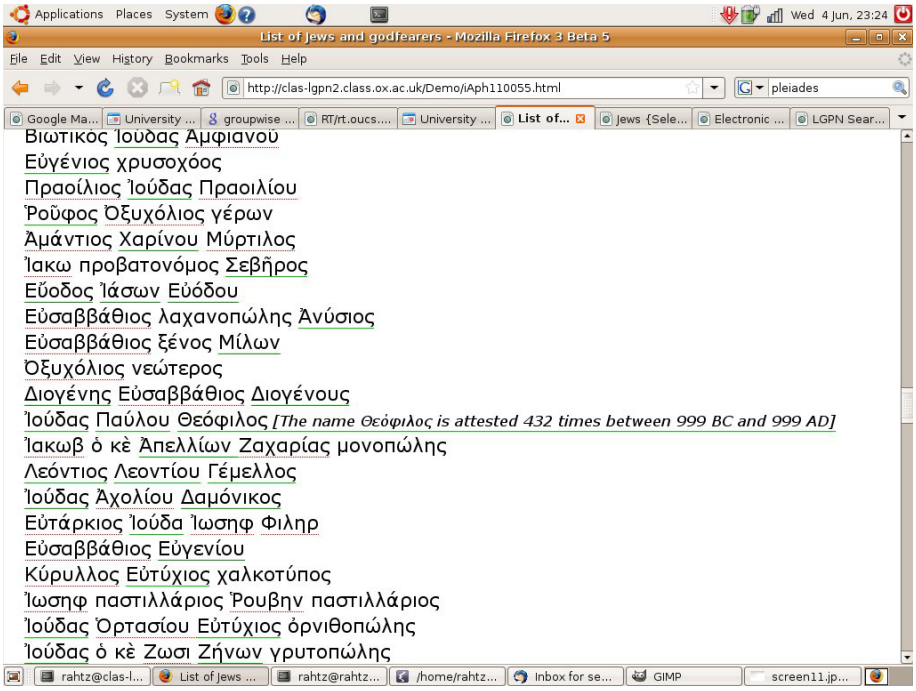
```
        <bibl>
                <title>IG</title> II<hi rend="sup">2</hi> 12229
        </bibl>
</person>
```

L'information ici fournie est donc à la fois biographique (sexe, nom, date de naissance, période d'activité), linguistique (variations linguistiques du nom), chronologique (date de naissance et période d'activité) et bibliographique (référence de l'inscription ayant livré le nom), le tout agrémenté de différents attributs (références, degré de certitude d'une information, information relative à la langue utilisée, commentaire sur la nature d'une information voir même informations sur le rendu d'un caractère lors d'une impression).

Le formalisme selon lequel cette information est structurée étant défini et communiqué, un programme n'a plus qu'à appeler les informations dont il a besoin pour son propre projet. La présence d'identifiants (par exemple, dans la fiche ci-dessus, « placeName key="Athens" ») permet ensuite de sélectionner la bonne information.

C'est ainsi qu'ont procédé les chercheurs du *London Computing Center for Humanities* en charge des projets *Inscriptions of Aphrodisias*[85], *Inscriptions of Roman Cyrenaica*[86] et *Inscriptions of Roman Tripolitania*[87] : en ajoutant dans leurs fichiers une balise autour des noms propres rencontrés dans leurs inscriptions, les chercheurs de Londres ont pu mettre en place un système qui recherche automatiquement l'information disponible sur ce nom dans le LGPN comme le montre l'exemple ci-dessous[88] :

[85] http://insaph.kcl.ac.uk/
[86] http://ircyr.kcl.ac.uk/
[87] http://www.tei-c.org/Activities/Projects/in04.xml
[88] S. Rahtz, http://tei.oucs.ox.ac.uk/Oxford/2009-07-oxford/talk-23-tei_mashups.pdf , 2009, page 34



Le document à l'écran est une inscription d'Aphrodisias reprenant la liste des juifs de la ville, présentant de ce fait un grand nombre de noms propres : le système vérifie dans le LGPN tous les noms et souligne en vert ceux présents dans le système et fournit par ailleurs une information complémentaire à l'utilisateur lorsque ce dernier passe la souris au-dessus de l'un de ces noms.

Cela n'est possible qu'en raison du balisage préalable du texte d'Aphrodisias selon une convention permettant l'échange de données entre les systèmes : la décision d'appliquer ce marquage fut donc une décision importante qui aurait pu ne pas être prise si l'approche des chercheurs avait été différente.

Cependant la richesse des ressources digitales fournies dans un format ouvert et libre comme l'est le formalisme TEI est aussi leur adaptabilité et leur réutilisabilité : une fois le texte encodé dans un fichier un chercheur peut le modifier selon ses besoins tout en économisant énormément de temps.

Construire une base de données en utilisant cette technologie signifie donc produire une infrastructure pour la recherche, une démarche similaire à celle qui inspira les grands compilateurs du CIL ou des IG au dix-neuvième siècle.

Comme on le voit également très clairement dans cet exemple, des centaines d'heures de travail ont été nécessaires au cours des trente-cinq dernières pour assurer la préservation de

l'information et élargir son cadre d'exploitation mais le résultat de ces efforts est une véritable ressource susceptible d'être exploitée et intégrée de manière décentralisée par n'importe quel chercheur. Cependant seule la disponibilité d'importants moyens financiers, humains et matériels a permis d'atteindre ce niveau : le projet LGPN représente en effet plusieurs millions de livres sterling d'investissement.

Pourtant, jusqu'il y a peu ce travail ne représentait que la simple mise en ligne du contenu d'un certain nombre de livres, n'apportant pas grand-chose de plus que la consultation d'une version numérique de ces ouvrages. Ce n'est que dans une seconde phase qu'est survenue la possibilité d'exploiter les capacités du média numérique pour enrichir l'information.

Enfin se pose la question de la pérennité de ces services : on l'a vu, obsolescences matérielles et logicielles ont eu un impact majeur sur le développement du projet et rien ne garantit la stabilité de l'URL d'accès aux données ni, dans le cas des *mashups* dérivés, celle des URL des applications ou des données sollicitées par le service. Une panne, une coupure du financement, un simple désintérêt des fondateurs du projet ou de leurs successeurs et toutes les données peuvent disparaître.

Si un tel cas de figure peut sembler improbable dans le cas d'un projet comme celui du LGPN il faut en revanche le prendre très au sérieux pour des projets plus modestes comme celui de la *BCD* d'A. Delattre. Nous reviendrons d'ailleurs sur ce point lorsque nous étudierons plus en détail ce projet.

Ces questions soulèvent également, outre des problèmes de continuité du service qui sont le lot de tous les projets de *mashup* du web, de graves interrogations spécifiques au milieu scientifique, en l'occurrence l'accès aux données de base permettant la vérification du travail scientifique, un principe fondateur de la recherche moderne.

En effet comment assurer cet accès si les données peuvent disparaître ou être déplacées en un clin d'œil ? Le nombre d'URL fournies dans des revues scientifiques et aujourd'hui inactives est ainsi un bon exemple de ce genre de problèmes. Et les projets d'archivages du web comme « *The internet archive* »[89] ne sauvegardent généralement pas le contenu des banques de données des sites. Une recherche sur le site de « *The internet archive* » pour le site de la *BCD*

[89] http://www.archive.org/index.php

ne donne ainsi que des copies datant de novembre 2005 et mai 2006[90], une preuve des limites de ces solutions.

[90] http://web.archive.org/web/*/http://dev.ulb.ac.be/philo/bad/copte/

Quelle politique de mise à jour des bases de données de documents digitalisés ?

La question du maintien à jour de l'information est complexe et va pour une bonne part dépendre de la nature du projet et de ses objectifs. Une collection de livres numérisés n'a en effet pas le même objectif qu'une base de données reprenant tous les textes d'une catégorie donnée et une collection de sources secondaires ne requiert pas le même traitement qu'un ensemble de sources primaires sans cesse relues et commentées sur lesquelles s'accumule une masse de sources secondaires qui entrainent parfois une modification de la lecture de la source première.

Pourtant la question de la mise à jour de l'information et de la digitalisation des progrès de la recherche ne semble pas encore avoir été réellement prise en compte dans les projets que j'ai eu l'opportunité d'observer, la problématique apparaissant seulement maintenant dans la dernière demande de financement rédigée par l'équipe de la *DDbDP*[91].

Un exemple de ce phénomène de non mise à jour des données est le cas de la relecture des papyrus grecs O. Leid. 124 et 129 (repris sous l'identifiant SB XX 14601) présentée par P. Heilporn lors du colloque « Et maintenant ce ne sont plus que des villages... : Thèbes et sa région aux époques hellénistique, romaine et byzantine ».

Dans l'article « A la recherche du clergé thébain à l'époque romaine »[92] il présente le processus qui l'a conduit à relire deux papyrus et à y identifier le fait qu'un apax d'anthroponyme était en fait une mauvaise lecture d'un théonyme, en l'occurrence le nom du dieu Amonrèsonthèr, plus connu en français sous le nom d'Amon Ré.

La relecture transforma un reçu de taxes d'une importance toute relative en une rare indication sur la politique romaine de taxation des lieux de culte et du clergé de la région de Louxor.

Suite à cette relecture , il proposait une nouvelle édition critique du papyrus qui devrait donc remplacer la précédente dans les bases de données. Or l'on constate que la *DDbDP* offre ici la

[91] Bagnall R. et all., http://www.duke.edu/~jds15/IDP2-FinalProposalRedacted.pdf, 2008

[92] Heilporn P., A la recherche du clergé thébain, in Delattre A. et Heilporn P. (ed.), Et maintenant ce ne sont plus que des villages… » : Thèbes et sa région aux époques hellénistique, romaine et byzantine, actes du colloque tenu à Bruxelles les 2 et 3 décembre 2005, Bruxelles, 2008, pp. 125-136

version périmée du texte comme le prouve la comparaison des deux versions, dans lesquelles le nom du dieu et sa mauvaise lecture ont été mis en gras :

Edition présente dans la *DDbDP*[93]	**Edition de P. Heilporn[94]**
Ἀπολ( ) καὶ μ(ἐτοχοι) πρά(κτορες) ἀργ(υρικῶν) μη(τροπόλεως) **Ἀμονρήο(υτι)** **Θαουέτις** διὰ Πετεμού(νιος) Ψεναμ(ούνιος)· ἔσχομεν ὑπ(ὲρ) γεω(μετρίας) δεκ(άτου) (ἔτους) ρυπ(αρὰς) (δραχμὰς) δεκαέξ, (γίνονται) (δραχμαὶ) ις. (ἔτους) δεκ(άτου) Τραιανοῦ τοῦ κυρίου Μεσορὴ - ca. ? - . .[- ca. ? -]	Ἀπολ(λ) καὶ μ(ἐτοχοι) πρά(κτορες) ἀργ(υρικῶν) μη(τροπόλεως) **Ἀμονρήσ(ονθηρ)** **Θεου μέγισ(του)** διὰ Πετεμενώ(φιος) Ψεναμ(ούνιος)· ἔσχομεν ὑπ(ὲρ) γεω(μετρίας) δεκ(άτου) (ἔτους) ρυπ(αρὰς) (δραχμὰς) δεκαέξ, (γίνονται) (δραχμαὶ) ις. (ἔτους) δεκ(άτου) Τραιανοῦ τοῦ κυρίου Μεσορὴ - ca. ? - . .[- ca. ? -]

La traduction du texte se lit dés lors :

« Apol(--) et collègues, receveurs des impôts en argent pour la métropole, (aux prêtres) d'Amonrèsonthèr, très grand dieu, par l'intermédiaire de Petemenöphis fils de Psenamounis. Nous avons reçu, pour la *geômetria* de la dixième année, seize drachmes brutes, soit 16 (dr.). Dixième année de Trajan notre seigneur, Mesorè [jour. Signature.] »[95].

Une non réactualisation de la base de données crée donc de très gênants silences dans les recherches qui peuvent induire en erreur les scientifiques cherchant à obtenir l'information ou être à la source de mauvaises théories ou conclusions. En effet un document comme celui-ci peut avoir un effet important lors d'analyses statistiques quand un corpus n'est composé que de quelques dizaines de textes….

Pourtant le domaine de la papyrologie greco-latine est doté d'outils performants qui, s'ils étaient digitalisés, permettraient de corriger ce genre de problèmes. Une version digitale du *Sammelbuch griechischer Urkunden aus Ägypten* et de la *Berichtigungsliste der griechischen Papyrusurkunden aus Ägypten* serait ainsi un outil remarquable qu'une intégration avec la

[93] http://www.papyri.info/navigator/full/hgv_SB_20:14601
[94] Heilporn P., A la recherche du clergé thébain, in Delattre A. et Heilporn P. (ed.), Et maintenant ce ne sont plus que des villages… » : Thèbes et sa région aux époques hellénistique, romaine et byzantine, actes du colloque tenu à Bruxelles les 2 et 3 décembre 2005, Bruxelles, 2008, p. 127
[95] Heilporn P., A la recherche du clergé thébain, in Delattre A. et Heilporn P. (ed.), Et maintenant ce ne sont plus que des villages… » : Thèbes et sa région aux époques hellénistique, romaine et byzantine, actes du colloque tenu à Bruxelles les 2 et 3 décembre 2005, Bruxelles, 2008, p. 128

DDBDP ou tout du moins le projet du Navigateur Papyrologique[96] rendrait particulièrement intéressante puisqu'elle permettrait de fournir une version toujours à jour de l'information scientifique.

Dans le cas de la BCD d'Alain Delattre la mise à jour des données représente une part non négligeable du travail car le choix fut fait dès le début d'indiquer la bibliographie complète de chaque texte présenté. Les ressources disponibles sont moins riches que pour les papyrus grecs ou latins mais existent tout de même.

Cependant, là encore, on doit déplorer le fait que peu sont digitalisées ce qui limite les possibilités. L'infrastructure nécessaire à ce genre de fonctionnalité doit encore être bâtie, un travail de longue haleine…

[96] http://www.papyri.info/

Le concept d'infrastructure

J'évoquais il y a quelques lignes le concept d'infrastructure. A mon sens il s'agit d'un concept important dans l'étude des *Digital Humanities* mais qui n'est que très rarement pris en compte par les personnes discutant de la pertinence des efforts parfois considérables investis dans les programmes de digitalisation.

En effet, il n'est pas rare de rencontrer des chercheurs se demandant à quoi servent tous ces sites qui coûtent, leurs semblent-ils, plus chers qu'une édition papier, être plus fragile et fournir une information de moindre qualité.

En réalité ce que ces personnes ne comprennent pas est le fait que malgré près de quarante ans d'efforts nous n'en sommes toujours qu'aux débuts du processus.

Que la plupart des projets menés jusqu'à aujourd'hui ont été des essais, pas toujours très réussis par rapport à leurs objectifs, pas toujours à la hauteur de nos attentes d'aujourd'hui, mais toujours à la pointe du progrès.

Et que ces tâtonnements nous mènent maintenant à l'aube d'une nouvelle ère.

En effet, alors que nous regardons derrière nous et observons le labeur accomplis, alors que de nombreux projets convertissent leurs données vers quelques standards interopérables, alors que des moteurs de recherche sont créés pour fouiller à travers de multiples données, nous nous rendons compte qu'en réalité le travail effectué fut celui du bâtisseur posant des fondations.

Les fondations en elles-mêmes ne sont pas jolies, sont rarement vues, n'apparaissent pas dans les revues d'architecture. Et pourtant sans elles rien ne serait possible, les gratte-ciels ne pourraient s'élever.

Les bases de données constituées dans les années 80, 90 et durant la première moitié des années 2000 sont, pour les *Digital Humanities*, ces fondations. Une infrastructure qui n'est certes pas parfaite, qu'il faut encore retravailler et enrichir pour parvenir à fournir aux chercheurs des données stables, à jour et de qualité.

Mais une infrastructure qui nous permet d'ors et déjà d'imaginer de nouvelles manières d'interroger les textes, voir même de nouvelles questions à leur poser.

Les équipes à la pointe de la recherche comme celles du projet *Integrated Digital Papyrology 2* s'en sont bien rendues compte et leur projet s'inscrit dans cette logique d'améliorer les bases en mettant au point de nouveaux outils permettant d'améliorer le contenu de ces « silos d'information ».

Le concept d'infrastructure est aussi associé à l'idée de durabilité, de solidité : l'infrastructure est l'élément le plus dur à détruire, ce qui reste après que tout soit tombé en ruine car cela a été construit dans ce but. Investir dans les infrastructures les plus robustes c'est prendre les meilleures dispositions pour que les données ne disparaissent pas.

Dans le cas qui nous occupe cela signifie adopter des standards ouverts, autoriser la réutilisation des données dans des projets autres que celui ayant produit l'effort initial de digitalisation, adopter une architecture de système comprenant des backups et des serveurs miroirs.

Certains projets sont sur ce point plus en avance que d'autres : Perseus dispose ainsi d'un serveur miroir en Allemagne qui assure la survie du projet en cas d'incident sur le serveur principal. D'autres projets ne disposent pas encore de cette infrastructure, notamment parce que cela à un coût non négligeable, notamment pour obtenir les compétences informatiques nécessaires au maintien de cette infrastructure.

Ce qui nous amène à un autre point des infrastructures que je souhaiterais soulever, celui de la compétence des individus. On observe en effet qu'à leur actuelle les bases de données utilisées par les chercheurs sont souvent gérées par de petites équipes conduisant plusieurs projets et travaillant ensemble mais d'une manière parfois un peu détachée du reste de la communauté des épigraphistes et des papyrologues.

Pourquoi ? Tout simplement parce que la plupart des chercheurs (et même des enseignants) manquent cruellement de connaissances dans le domaine informatique. De ce fait beaucoup sous-utilisent les ressources déjà à leur disposition.

Si l'on ne regarde qu'à l'ULB, aucun cours d'informatique n'est donné aux étudiants de langue et philologie classique. Un cours sera bien introduit à la rentrée 2010-2011 mais il ne sera pas adapté aux besoins des futurs épigraphistes et papyrologues.

Le cas du Pr. Delattre, que nous étudierons dans quelques instants, est également une bonne illustration de ce manque cruel de compétences avancées en informatique : s'il n'avait pas

posé, un jour, la question de comment l'on crée un site internet et si un étudiant de 1ere licence n'avait pas été là pour lui répondre son projet aurait connu au moins un ou deux ans de retard.

Enfin, et ceci est le dernier point que nous aborderons sur cette question des infrastructures, il faut bien avoir conscience du fait qu'elles ont un coût. Un coût assez important même : électricité, bande passante, serveurs, entretiens, développement, des projets comme le Papyrological Navigator coûtent cher.

Ainsi la demande de financement du projet IDP 2 indiquait que près de 900 000 dollars seraient nécessaires pour mener à bien les opérations prévues pour deux années de travail[97].

Cela implique de rassembler de véritables équipes et d'organiser de véritables projets interuniversitaires. Cela demande aussi un soutien de la part des universités auxquelles appartiennent les chercheurs entamant de tels projets, mais aussi le support d'organismes de financement.

Des conditions qui ne sont pas forcément toujours rassemblées dans notre pays, notamment à l'ULB. Aux USA des institutions comme le *Packard Institute for Humanities*, la *Andrew W. Mellon Foundation*[98] ou la *National Endowment for the Humanities*[99] dépensent chaque année plusieurs millions de dollars pour la mise sur pied de cette infrastructure.

En Europe le projet HGV n'a plus de financement propre depuis quelques années et vit grâce aux crédits d'un autre projet. La Commission Européenne finance pourtant elle aussi des projets de digitalisation mais ils concernent plutôt les manuscrits médiévaux, à l'instar du projet ENRICH. Où sont les antiquistes ?

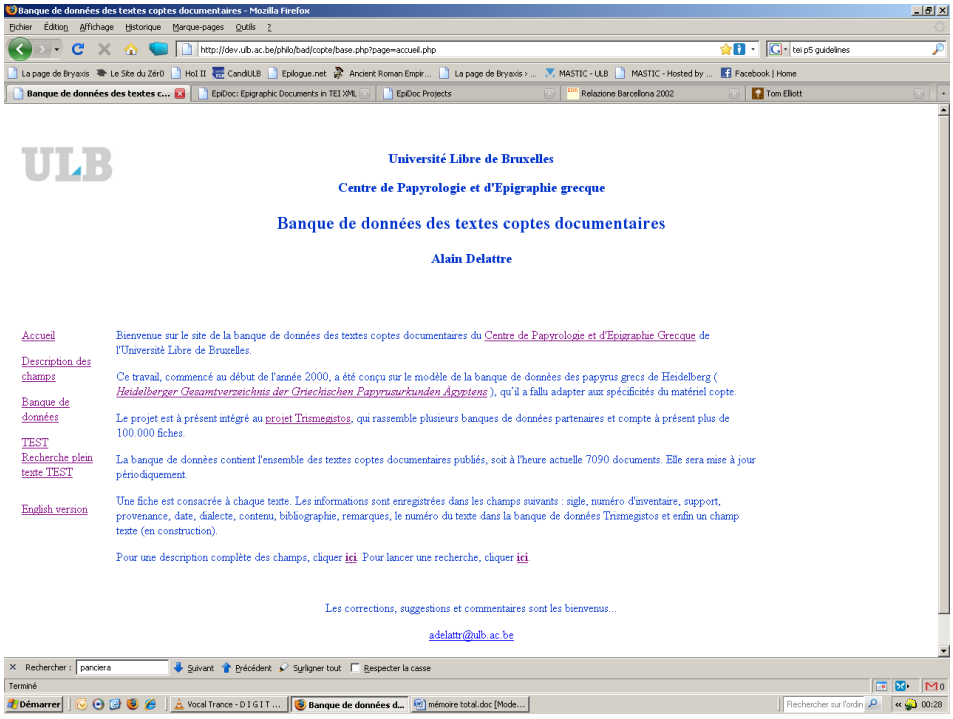
[97] Bagnall R. et al, http://www.duke.edu/~jds15/IDP2-FinalProposalRedacted.pdf, 2008, p. 2
[98] http://www.mellon.org/
[99] http://neh.gov/

Chapitre 5 : La base de données des inscriptions coptes documentaires[100]

Nous avons réalisé un long panorama de la situation générale de l'édition en ligne des sources épigraphiques et papyrologiques telle qu'elle se présente en ce milieu d'année 2009.

Un panorama très diversifié qui permet toutefois de mettre en valeur un certain nombre de tendances et de constater que des différences existent entre les approches adoptées par les papyrologues et les épigraphistes. De mettre en lumière également un certain nombre de problèmes épistémologiques.

Il nous reste encore un projet à étudier avant de conclure ce mémoire, celui conduit depuis le début des années 2000 par le Pr. A Delattre, de l'Université Libre de Bruxelles.



La base de donnée des inscriptions coptes documentaires, aussi connue sous le nom de Brussel's Coptic Database, a pour ambition de mettre à la disposition des chercheurs l'ensemble des papyri et ostraca coptes portant des textes autres que des incantations magiques ou des extraits de textes littéraires ainsi que toutes les métadonnées pertinentes, y compris des données bibliographiques à jour.

[100] http://dev.ulb.ac.be/philo/bad/copte/

A ce titre, on peut dire que la base de données remplit un peu pour les coptisants le rôle que tiennent pour les papyrologues travaillant dans le monde grec la *Berichtingsliste* et le *Sammelbuch* de Preisigke.

Reçus de taxes, plaintes en justice, courriers privés, comptabilité de monastères, contrats de mariages, la nature des textes est extrêmement diversifiée. Près de 10 000 documents sont répertoriés dans le fichier FileMaker Pro abritant la copie principale de la base de données mais seules 9070 sont actuellement en ligne.

Commencé alors qu'A. Delattre était encore doctorant, ce projet progresse lentement par manque de ressources. En effet A. Delattre ne peut jusqu'à présent compter que sur son labeur et l'assistance bénévole de l'auteur du présent mémoire.

C'est en 2004 que l'idée de mettre la base de données en ligne fut sérieusement envisagée par A. Delattre qui se renseigna alors sur la manière dont il pouvait créer un site web. Après quelques entretiens avec l'auteur, alors étudiant de 1ère licence[101] en Histoire de l'Antiquité, ce dernier lui proposa de l'assister dans son projet : c'était le début d'une collaboration qui continue aujourd'hui encore.

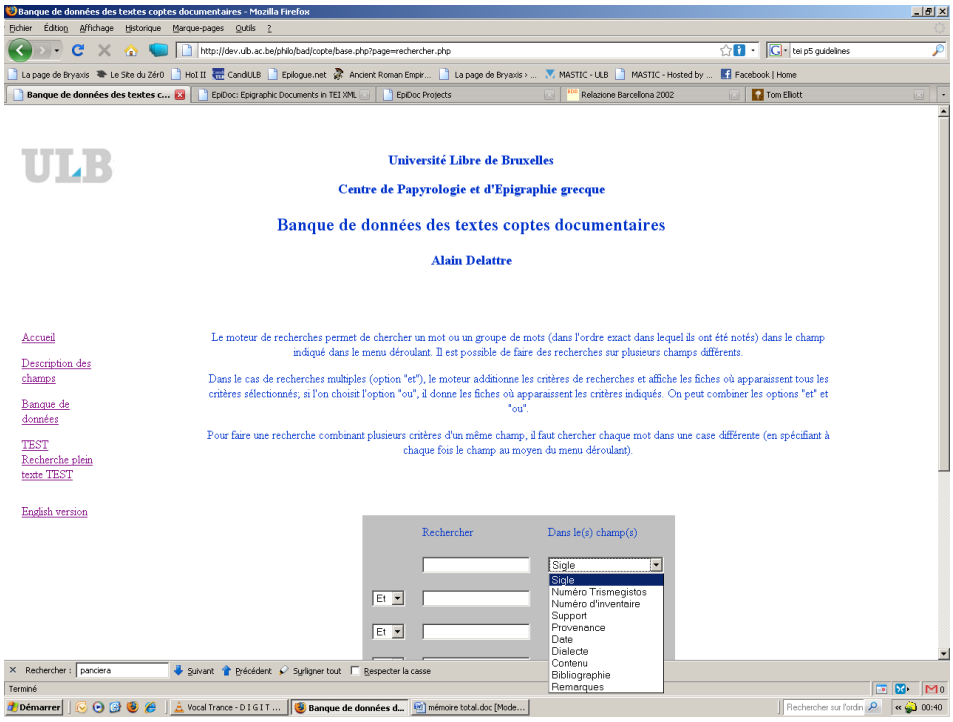
Très vite il fut décidé de transférer la base de données de publication de FileMaker Pro vers MySQL, une interface en PHP étant développée pour permettre des recherches complexes.

Des problèmes d'affichage des caractères coptes, alors encodés à l'aide de la norme ISO-8859, poussèrent l'équipe à se contenter de la publication des métadonnées en attendant de faire passer les données en encodage Unicode, une opération qui a eu lieu tout récemment.

Quelques tests ont déjà été menés pour procéder à l'affichage de textes retranscrits en UTF-8 mais le résultat n'a pas été très concluant.

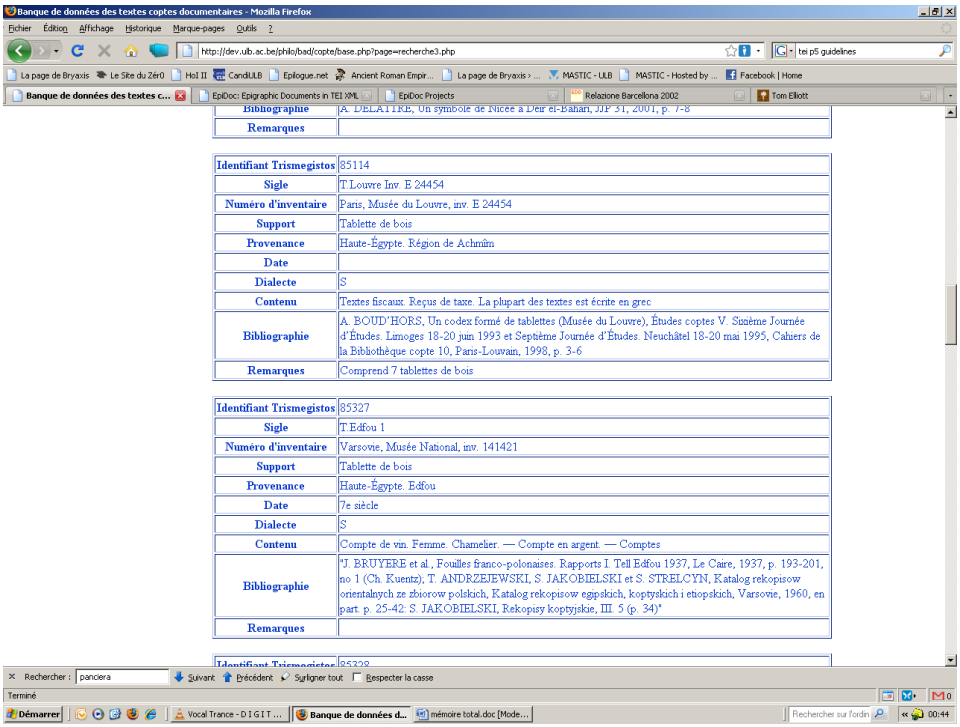
Néanmoins même sans ces textes le site a offert aux chercheurs spécialisés un outil assez efficace

[101] Dans le système pré-Bologne, correspondant au MA-1 actuel



Le moteur de recherche du site permet d'effectuer une recherche sur n'importe lesquels des champs de la base de donnée et offre la capacité de combiner jusqu'à 6 critères de recherche articulés par les opérateurs booléens AND et OR. Il est néanmoins très sensible à la casse, ce qui pose parfois problèmes avec les sigles et les numéros d'inventaire.

Le site est par ailleurs traduit en anglais (même si les données elles-mêmes sont toujours en français, le travail de traduction étant en cours).



Bibliographie	A. DELATTRE, Un symbole de Nicée à Deir el-Bahari, JJP 31, 2001, p. 7-8
Remarques	

Identifiant Trismegistos	85114
Sigle	T.Louvre Inv. E 24454
Numéro d'inventaire	Paris, Musée du Louvre, inv. E 24454
Support	Tablette de bois
Provenance	Haute-Égypte. Région de Achmîm
Date	
Dialecte	S
Contenu	Textes fiscaux. Reçus de taxe. La plupart des textes est écrite en grec
Bibliographie	A. BOUD'HORS, Un codex formé de tablettes (Musée du Louvre), Études coptes V. Sixième Journée d'Études. Limoges 18-20 juin 1993 et Septième Journée d'Études. Neuchâtel 18-20 mai 1995, Cahiers de la Bibliothèque copte 10, Paris-Louvain, 1998, p. 3-6
Remarques	Comprend 7 tablettes de bois

Identifiant Trismegistos	85327
Sigle	T.Edfou 1
Numéro d'inventaire	Varsovie, Musée National, inv. 141421
Support	Tablette de bois
Provenance	Haute-Égypte. Edfou
Date	7e siècle
Dialecte	S
Contenu	Compte de vin. Femme. Chamelier. — Compte en argent. — Comptes
Bibliographie	"J. BRUYERE et al., Fouilles franco-polonaises. Rapports I. Tell Edfou 1937, Le Caire, 1937, p. 193-201, no 1 (Ch. Kuentz); T. ANDRZEJEWSKI, S. JAKOBIELSKI et S. STRELCYN, Katalog rekopisow orientalnych ze zbiorow polskich, Katalog rekopisow egipskich, koptyskich i etiopskich, Varsovie, 1960, en part. p. 25-42: S. JAKOBIELSKI, Rekopisy koptyjskie, III. 5 (p. 34)"
Remarques	

Le moteur de recherche affiche ses résultats à raison de 25 fiches par écran. Parmi les champs affichés figure le numéro d'identifiant Trismegistos qui permet de faire le pont entre les deux projets : si le site d'A. Delattre ne permet pas d'interroger la base Trismegistos, il est en revanche tout à fait possible de récupérer ce numéro pour lancer une requête sur le site du projet louvaniste. C'est aussi la présence dans la base de données de ce numéro qui permet à Trismegistos de renvoyer ses utilisateurs vers la base de données.

Si l'on compare le site avec ceux étudiés précédemment on note, outre son caractère spartiate, qu'il est entièrement dépourvu d'index. Il ne fournit pas non plus beaucoup d'indications sur sa politique éditoriale.

Il serait assez aisé d'améliorer ses métadonnées en créant une table des données bibliographiques qui permettrait de disposer d'un outil comparable à celui fournit par Trismegistos, voir même de créer des liens avec ce site.

Le champ « contenu » pourrait également utiliser une liste contrôlée de valeurs en lieu et place de l'encodage libre qui y a court actuellement. Et les données géographiques pourraient sans doute être améliorées et complétées.

Néanmoins le site soutient la comparaison avec les autres projets que nous avons observés car il offre une mise à jour des données et les textes qu'il mettra en ligne seront entièrement édités selon les normes de la convention de Leyde.

L'idéal serait évidemment d'envisager le transfert des données dans un format XML comme EpiDoc, par exemple dans sa version personnalisée par la DDbDP. Néanmoins il faut pour cela surmonter une forte courbe d'apprentissage et disposer du temps nécessaire au marquage des textes. D'autant plus que, là où il est facile de trouver des étudiants de bachelier ou de master maîtrisant le grec ou le latin, les étudiants sachant lire le copte sont assez rares.

On ne peut donc recruter un lot d'étudiant pour procéder au marquage des noms propres, toponymes, façon de procéder relativement courante dans les projets anglo-saxons.

Les décisions sur ces questions n'ont pas encore été prises même si l'idée d'un prototype réalisé dés le départ en XML est actuellement à l'étude. Il s'agirait d'un projet neuf, parallèle à la base de donnée des inscriptions documentaires et se focaliserait sur l'épigraphie copte. Ce projet ne devrait dans un premier temps pas poser de problème technique majeur puisqu'il ne faudrait pas procéder à des conversions de données.

En revanche la conversion des données stockées dans FileMaker Pro s'avérera sans doute plus difficile. Il n'est en effet absolument pas certain que les outils de conversion conçus pour le grec ou le latin fonctionnent avec le copte.

La banque de données devrait également comprendre un champ indiquant la date de dernière mise à jour de chaque entrée, de manière à fournir une information transparente aux utilisateurs sur l'âge de la donnée qu'ils consultent, une information qui n'est aujourd'hui pas affichée par les systèmes en place même si la norme TEI indique que l'information doit être encodée dans le fichier XML portant l'entrée.

Le soutien d'une personne disposant de meilleurs compétences techniques pour assurer la conversion semble donc impératif mais la Belgique ne semble pas disposer de spécialistes de la TEI dans sa variante EpiDoc : il faudrait sans doute faire appel à des chercheurs du *King's College* de Londres ou de l'*Oxford University Computing Services*.

Dernière question, le maintient à long terme de la qualité des données sera-t-il possible avec l'architecture actuelle ? Sans doute pas. La masse de données est susceptible de surcharger le seul gestionnaire du projet ce qui entrainerait un retard dans les mises à jour.

Conclusions

Voici venu le temps de tirer quelques conclusions. Nous avons rappelé quelques éléments historiques relatifs au développement de l'épigraphie et de la papyrologie, notant que ces deux disciplines, pourtant très proches, avaient suivi des voies différentes. Nous avons ensuite rappelé quelques notions technologiques, remises elles aussi dans leur contexte historique, afin de permettre une meilleure compréhension des différentes analyses conduites ensuite sur une série de projets menés tant en Europe qu'aux USA.

Nous avons identifié à cette occasion quatre points essentiels :

- Un projet informatique de longue durée à une vie mouvementée liée à l'évolution des technologies, soit que ces dernières forcent à apporter des modifications au système, soit qu'elles ouvrent de nouvelles perspectives dont veulent s'emparer les chercheurs

- Les *Digital Humanities* sont une entreprise coûteuse et requérant des compétences spécifiques. En outre leur réelle utilité ne se montrera que lorsque des systèmes d'exploitation des données brutes seront mis en place. Les quarante dernières années ont vu la pose de fondations installées sans réel plan directeur et nous en somme à présent à une phase d'ajustement des ressources pour permettre leur interconnexion et la construction des superstructures qui seront, elles, les vrais éléments porteurs de plus-value.

- Une question particulièrement importante sur le plan scientifique est celle de la qualité des données et des métadonnées : quels sont exactement les éditions affichées, la date de dernière mise à jour, quelle valeur accorder aux données ?

- Le chercheur souhaitant utiliser les ressources en lignes doit toujours se poser la question de savoir quels sont les objectifs du site qu'il consulte et si les questions qu'il souhaite poser au système sont en adéquation avec les réponses que ce dernier peut lui fournir. Il est donc nécessaire pour les sites de fournir des informations plus claires permettant d'aider le chercheur à répondre à ces questions tout en veillant à évoquer ces questions dans le cadres des débats relatifs à la méthode scientifique en cette ère numérique.

Nous avons également constaté qu'une tendance au travail collaboratif s'était lancé dans le domaine de la papyrologie en ligne même si les projets épigraphiques restent bien souvent des projets isolés ou ne se touchant que par leur plus petit dénominateur commun.

Cette tendance au travail collaboratif, inspirée des thèses du web 2.0, est positive car elle permet de répondre à un certain nombre de problèmes des structures actuelles et d'assurer une baisse des coûts de digitalisation en permettant une participation libre sous contrôle éditorial d'un comité scientifique.

Cependant ce même phénomène contribue à une centralisation du travail qui peut avoir des effets pervers. Tout d'abord quelle visibilité pour le travail des chercheurs contribuant à l'encodage ? Le nom des personnes ayant encodé des inscriptions ou des papyri n'apparait en effet nulle part sur les interfaces, même si la norme TEI (et par conséquence EpiDoc) comprend un champ pour encoder cette information.

Et quelle visibilité pour les chercheurs n'encodant pas de documents, simples utilisateurs des systèmes ? N'y a-t-il pas un risque d'exclusion par manque d'expérience ou de moyens informatiques ?

Par ailleurs on connait des exemples d'articles refusés par des comités éditoriaux pour cause de désaccords entre certains de ses membres et l'auteur du texte soumis. Si l'on se dirige vers une centralisation de la publication électronique de papyrus il y a-t-il un risque de voir de tels phénomènes se produire ?

Enfin, en ces périodes d'évaluation de la recherche et d'efforts de quantification du travail mené en sciences humaines, n'y a-t-il pas un risque d'abus d'outils statistiques comme ceux prévus par le projet IDP2 ?

Nous n'avons évidemment pas les réponses à ces questions mais nous avons pu constater au cours de nos recherches qu'elles n'avaient pas toutes été posées avant la présente conclusion. Est-ce à dire que la réponse est évidente ou que les acteurs de terrain, préoccupés par l'aspect pratique de leurs réalisations, ne se sont pas encore posé ces questions plus épistémologiques ?

Bibliographie

Bibliographie imprimée

Les sources imprimées pertinentes pour mon mémoire furent peu nombreuses, en voici la liste :

- Andries. P., *Unicode 5.0 en pratique*, Paris, 2008
- Ansani M. et Ghignoli A., « Testi digitali : nuovi media e documenti medievali », in Mélanges de l'Ecole Française de Rome, à paraître
- Bell H.I., « Friedrich Preisgke » in The Journal of Egyptian Archeology vol.10 n°2, juillet 1924, pp. 172-173
- Bertrand P., « Autour de l'édition électronique et des digital humanities : nouvelle érudition, nouvelle critique ? » in Mélanges de l'Ecole Française de Rome, à paraitre
- Bodel J., Epigraphic evidence, ancient history from inscriptions, New-York, 2001
- Cameron S., Richardson S., Using Computers in History, 2005
- Cocaud M., Histoire et informatique: Base de données, recherche documentaire multimédia : actes du 1er Colloque, 1995
- Denooz J., « *Opera latina* : une base de données sur internet », dans *EVPRHOSYNE*, 32, 2004, pp.79-88
- **E**vrard E., « Le laboratoire d'analyse statistique des langues anciennes de l'Université de Liège », dans *Mouvement scientifique en Belgique***,** 9, 1962, p. 163-169
- Genet Jean Philippe, Standardisation et échange des bases de données historiques, 1998
- Heilporn P., A la recherche du clergé thébain, in Delattre A. et Heilporn P. (ed.), *Et maintenant ce ne sont plus que des villages… » : Thèbes et sa région aux époques hellénistique, romaine et byzantine, actes du colloque tenu à Bruxelles les 2 et 3 décembre 2005, Bruxelles*, 2008, pp. 125-136
- Hunt A.S., « A note on the transliteration of papyri», *Chronique d'Égypte* 7 (1932), pp. 272-274

- Krummrey H., Panciera S., « Criteri di edizione e segni diacritici », *Tituli* 2 (1980), pp. 205-215

- Lassère J.-M., Manuel d'épigraphie latine, 2 vol., Paris, Picard, 2005
- Lewis M.J., Roger Lloyd-Jones, Using Computers in History: A Practical Guide, 1996
- Mawdsley E., Munck T., Computing for Historians: An Introductory Guide, 1993
- Mayor, John E. B., « Review of Dictionnaire des Antiquités Grecques et Romaines d'après les textes et les monuments, contenant l'explication des termes qui se rapportent aux mœurs, aux institutions, à la religion, aux arts, aux sciences, au costume, au mobilier, à la guerre, à la marine, aux métiers, aux monnaies, poids et mesures, etc. etc., et en général à la vie publique et privée des anciens » in *The Classical Review* 1 (7), pp 201–202
- Michel J. et Gester M., Lexique de base du latin, Anvers et Paris, 1967
- Press John, Harvey Charles, Databases in Historical Research: Theory, Methods and Applications, 1995

- Purnelle G., « Opera Latina : la banque de donnée des textes latins du L.A.S.L.A. sur cd-rom » in *Revue Informatique et Statistique dans les Sciences humaines* XXXI 1 à 4, 1995, pp. 173-182

- Roueché Ch., Aphrodisias in Late Antiquity, Londres, 1989
- Townsend S. et all, Digitising History: A Guide to Creating Digital Resources from Historical Documents, 1999
- van Groningen B.A., « Projet d'unification des systèmes de signes critiques », *Chronique d'Égypte* 7 (1932), pp. 262-269
- Waltzing J.P., Le receuil général des inscriptions latines et l'épigraphie latine depuis 50 ans, Louvain, 1892

Webographie

Par son sujet même le présent mémoire fait un large usage de sources internet. Toutes les pages ont été visitées entre le 7 et le 10 aout 2009 et les capures d'écran prises à cette occasion. Voici la liste des pages concernées:

- Bagnall R. et all., http://www.duke.edu/~jds15/IDP2-FinalProposalRedacted.pdf, 2008
- S. Chabert, Histoire sommaire des etudes d'épigraphie grecque en Europe, 1906, http://www.class.ulg.ac.be/enseignement/ressources/chabert.pdf
- Payen P., http://dagr.univ-tlse2.fr/sdx/dagr/index.xsp
- S. Rahtz, http://tei.oucs.ox.ac.uk/Oxford/2009-07-oxford/talk-23-tei_mashups.pdf , 2009
- http://132.236.125.30/content.html

- http://132.236.125.30/history.html
- http://www.ancientlibrary.com/smith-dgra/
- http://www.annee-philologique.com/aph/
- http://anubis.dkuug.dk/JTC1/SC2/WG3/docs/n411.pdf

- http://www.archive.org/index.php
- http://web.archive.org/web/*/http://dev.ulb.ac.be/philo/bad/copte/
- http://bcs.fltr.ucl.ac.be/
- http://blog.kennisland.nl/knowledgeland/2009/07/26/uk_national_portrait_gallery_vs_wikimedia_vs_the_public_domain/
- http://www.cipl.ulg.ac.be/Lasla/
- http://www.cipl.ulg.ac.be/Lasla/descriptionop.html

- http://www.columbia.edu/cu/lweb/projects/digital/apis/
- http://www.cvaonline.org/
- http://dagr.univ-tlse2.fr/sdx/dagr/index.xsp
- http://dev.ulb.ac.be/philo/bad/copte/
- http://www.eagle-eagle.it
- http://www.edr-edr.it/Documenti/Document1_it.html
- http://www.edr-edr.it/Documenti/Document2_it.html
- http://www.edr-edr.it/Download/EAGLE-Testi.pdf
- http://enrich.manuscriptorium.com/
- http://epidoc.sourceforge.net/
- http://epigraphy.packhum.org/inscriptions/
- http://www.insaph.kcl.ac.uk/index.html
- http://insaph.kcl.ac.uk/iaph2007/index.html
- http://insaph.kcl.ac.uk/ala2004/index.html
- http://insaph.kcl.ac.uk/ala2004/redist/inscriptions/eAla043.xml
- http://www.lgpn.ox.ac.uk/
- http://www.lgpn.ox.ac.uk/project/index.html
- http://www.mellon.org/
- http://www.mgh.de/dmgh/
- http://neh.gov/
- http://opuculuk.blogspot.com/2009/05/ibycus-mainframe.html
- http://www.oucs.ox.ac.uk/internal/history/

- http://www.packard.org/home.aspx
- http://www.packhum.org/phi/phi-info.html
- http://www.papyri.info/
- http://www.papyri.info/navigator/full/hgv_SB_20:14601
- http://www.perseus.tufts.edu
- http://www.perseus.tufts.edu/hopper/
- http://pleiades.stoa.org/
- http://promethee.philo.ulg.ac.be/cedopal/video/film_ibicus.htm
- http://publibfp.boulder.ibm.com/cgi-bin/bookmgr/BOOKS/dsm04m00/CCONTENTS
- http://quod.lib.umich.edu/m/moagrp/
- http://www.roger-pearse.com/weblog/?p=1326
- http://www.roger-pearse.com/weblog/?p=2113
- http://www.roger-pearse.com/weblog/?p=2277
- http://www.rzuser.uni-heidelberg.de/~gv0/gvz.html
- http://www.stoa.org/sol/
- http://www.tei-c.org/Activities/Projects/in04.xml
- http://www.tei-c.org/Activities/Projects/pe01.xml
- http://www.tei-c.org/Guidelines/P5/

- http://www.tei-c.org.uk/Master/Reference/oldindex.html
- http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DR.html
- http://www.tei-c.org/Roma/
- http://www.tei-c.org.uk/Vault/Songs/

- http://www.tlg.uci.edu/
- http://www.tlg.uci.edu/about/ibycus.php
- http://www.trismegistos.org/index.html
- http://www.trismegistos.org/index2.php
- http://www.ulb.ac.be/philo/cpeg/#par2
- http://www.ulb.ac.be/philo/cpeg/hombert.htm
- http://www2.ulg.ac.be/facphl/services/cedopal/pages/papyrologie.htm
- http://www.w3.org/TR/1998/REC-xml-19980210

- http://www.w3.org/TR/2008/REC-xml-20081126/

Annexe 1 : tableau de représentation du Beta Code

Alphabet grec et sa représentation en Beta Code

Majuscule	Beta Code	Character name	Minuscule	Beta Code
Α	*A	Alpha	α	A
Β	*B	Beta	β	B
Γ	*G	Gamma	γ	G
Δ	*D	Delta	δ	D
Ε	*E	Epsilon	ε	E
Ζ	*Z	Zeta	ζ	Z
Η	*H	Eta	η	H
Θ	*Q	Theta	θ	Q
Ι	*I	Iota	ι	I
Κ	*K	Kappa	κ	K
Λ	*L	Lambda	λ	L
Μ	*M	Mu	μ	M
Ν	*N	Nu	ν	N
Ξ	*C	Xi	ξ	C
Ο	*O	Omicron	ο	O
Π	*P	Pi	π	P
Ρ	*R	Rho	ρ	R
Σ	*S	Medial Sigma	σ	S, S1
Σ	*S	Final Sigma	ς	S, S2
Ϲ	*S (*S3)	Lunate Sigma	ϲ	S (S3)
Τ	*T	Tau	τ	T

Υ	*U	Upsilon	υ	U
Φ	*F	Phi	ϕ	F
Χ	*X	Chi	χ	X
Ψ	*Y	Psi	ψ	Y
Ω	*W	Omega	ω	W