Se procurer des textes latins numérisés

- **Hommage à Roberto Busa**
- **Définir l'objectif**
- **Définir une stratégie**
- **Domaine public**
- **Copyfraud**
- **Scan et OCR**
- **Corrections**
- **Principaux sites**
- **Techniques de récupération**

Hommage à Roberto Busa

- Roberto Busa S.J. 1913-2011.
- 1949 : accord avec Thomas Watson, fondateur d'IBM, pour la constitution d'un corpus informatique numérisé des œuvres de Thomas d'Aquin.
- 1974 : 56 volumes imprimés, 1989 CD-ROM
- Aujourd'hui, en open-acces...

Définir l'objectif

- L'utilisation que l'on peut faire d'un corpus, tous les résultats de l'analyse, dépendent entièrement de la nature du corpus.
- La taille est un facteur important.
- On peut récupérer tout ou partie d'un corpus existant, ou en créer un de toutes pièces.
- S'agissant de textes anciens (i.e. en gros antérieurs à 1900), il faut réfléchir très attentivement aux rapports, tels qu'on peut les connaître ou les supposer, entre eux et la société qui les a produits. **Quelle que soit son étendue, un corpus donne toujours une vue partielle sur cette société**, il faut se documenter solidement sur ce point.
- Rappel : il faut renverser la perspective.

Définir une stratégie

- Récupérer et organiser un corpus demande du temps ; il faut avoir une idée précise du **temps dont on dispose**, c'est le critère fondamental.
- En fonction de ce temps, il faut examiner les possibilités pratiques, et évaluer au plus juste le temps qui sera nécessaire en fonction des choix de départ.
- S'il s'agit d'un travail de recherche, le temps de préparation du corpus doit être drastiquement limité ; mieux vaut un corpus bref, mais intelligemment analysé, qu'un corpus énorme à peine effleuré.
- ATTENTION : on est toujours trop optimiste quant au temps que l'on croit nécessaire ; il y a toujours une grande quantité de difficultés imprévues ; faire une évaluation et multiplier par 2 !!
- Ordres de grandeur :
  moins de 5M de tokens : petit corpus
  entre 5 et 500M de tokens : corpus ordinaire, moyen
  plus de 500M de tokens : grand corpus
- ATTENTION 2 : relation avec la RAM disponible.

Domaine public

- TOUS LES TEXTES ANCIENS SONT DANS LE DOMAINE PUBLIC ; personne ne dispose sur eux d'un quelconque droit commercial.
- La législation est simple et claire (par exception !), et sanctionnée par des traités internationaux dont elle dépend.
- Il faut lire la **Convention de Berne pour la protection des œuvres littéraires et artistiques** (http://www.wipo.int/treaties/fr/text.jsp?file_id=283699), c'est le texte fondamental. Cette convention prévoit que les droits patrimoniaux de l'auteur durent 50 ans après sa mort.
- Cette convention reçoit des applications plus ou moins restrictives selon les pays. En France, le délai est porté à 70 ans, avec un supplément de 30 ans pour les auteurs « morts pour la France ». En France, il faut lire le **Code de la propriété intellectuelle** (http://www.legifrance.gouv.fr/affichCode.do?cidTexte=LEGITEXT000006069414).
- L'auteur d'une oeuvre de l'esprit jouit sur cette oeuvre, du seul fait de sa création, d'un droit de propriété incorporelle exclusif et opposable à tous. (article L 111-1)
- Ceci est le texte fondamental : **l'auteur se définit par une création**, il n'y a aucun autre critère.
- Aucun procédé d'édition ou de reproduction ne crée de droit. Les « ayant-droits » ne disposent d'un droit patrimonial qu'à titre de délégation.
- Le **droit moral** est inaliénable et imprescriptible.
- En 2014,en France, il est sans objet de discuter de droits patrimoniaux portant sur les textes écrits par des auteurs morts avant 1944.
- Pour tous les autres, il faut respecter le droit d'auteur.

Copyfraud

- En dépit de cette totale clarté, on rencontre fréquemment des entreprises ou des institutions qui prétendent disposer de « droits » sur des textes de périodes anciennes…, en particulier au prétexte du travail et de la dépense qu'elles ont engagés pour assurer une forme ou une autre de publication.
- Cette **manœuvre est délictueuse**, elle porte un nom, c'est le **copyfraud**.
- **Aucune forme d'édition, de diffusion ou de reproduction ne modifie l'auteur d'un texte** !
- Les « éditeurs scientifiques » n'ont aucun droit sur les textes anciens ou leurs variantes, ils détiennent des droits sur les notes, commentaires et traductions. Ils ont un droit moral à la reconnaissance de leurs collègues, que ceux-ci leur refusent trop souvent….
- En dépit d'un intense lobbying, les « ayant-droits » n'ont pas trouvé de moyen de modifier cet aspect fondamental du droit d'auteur. **La jurisprudence est rare, mais sans équivoque**.
- Malheureusement, beaucoup de responsables universitaires, ignorants et lâches, agissent par « peur du procès » et se plient aux injonctions illégales mais comminatoires de quelques éditeurs commerciaux effrontés.
- Les entreprises les plus importantes connaissent d'ailleurs parfaitement cette situation : Google diffuse sans restriction les scans de textes du domaine public, et un éditeur de textes belge bien connu encrypte très solidement les fichiers-textes qu'il diffuse, parce qu'il sait parfaitement qu'il n'a aucun droit sur ces textes.
- La question est simple : les textes de Virgile, de Dante et de Victor Hugo sont-ils dans le domaine public, et si oui quelles sont les limites à leur diffusion ?
- **Il est extrêmement important de faire connaître la notion de copyfraud**, que beaucoup d'acteurs s'emploient à maintenir confidentielle.

Scan et OCR

- S'il n'existe aucun fichier informatique disponible contenant un texte que l'on souhaite intégrer dans un corpus, la solution ordinaire consiste à scanner une édition et à procéder à l'OCR (reconnaissance optique de caractères).
- Si l'on a le choix, il faut employer l'édition la mieux imprimée, car **la netteté des caractères est un critère essentiel** de réussite. Prendre une **loupe** et rechercher les caractères cassés ou collés.
- Il faut ensuite prendre un soin tout particulier au moment du scan, et procéder à de **nombreux essais** en modifiant les réglages du logiciel.
- Lorsqu'on dispose de fichiers-images, ceux-ci doivent être traités par un logiciel d'OCR. A côté d'une offre commerciale (fort) onéreuse, on dispose d'un logiciel libre efficace (de plus en plus efficace avec les années qui passent) **tesseract**. Ce logiciel a été placé dans le domaine public par Google, et le principal développeur travaille dans le cadre de cette entreprise.
- Il s'agit d'un logiciel en ligne de commande, il est plus confortable d'utiliser une interface graphique, plusieurs existent, je suggère **gImageReader**.
- Si la page à océriser comporte des caractères particulier, on peut exécuter une phase préalable d'apprentissage du logiciel.
- Dans les cas les plus résistants (éditions anciennes notamment), il faut utiliser un logiciel fonctionnant uniquement en « mode image », c'est-à-dire à partir d'un apprentissage complet de tous les caractères que le logiciel peut rencontrer dans son travail. Le plus connu et efficace est **gamera** (actuellement développé et maintenu en Allemagne).

Corrections

- Le résultat du scan, quel que soit le logiciel et les conditions de travail, est rarement satisfaisant, il faut donc passer par une étape de correction.
- La solution standard est celle du correcteur orthographique. Pour cela, on a besoin d'un logiciel approprié, mais surtout d'un « dictionnaire » comportant toutes les formes acceptables.
- Le logiciel libre aujourd'hui le plus utilisé est **hunspell**. Il fonctionne vite et bien, en arrière plan d'un logiciel d'édition (geany ou gedit par exemple).
- Il existe de **nombreux dictionnaires pour les langues actuelles**.
- On n'a, pour le moment, rien d'équivalent pour le latin. La difficulté centrale est celle des variations graphiques ; les flottements sont considérables, et l'on n'a pas encore réalisé de répertoire des formes (de mots communs et de noms propres) qui puisse être considéré comme vraiment efficace.
- La solution consiste donc à constituer soi-même un dictionnaire ad hoc. On peut par exemple tokeniser le résultat du scan, établir une liste de fréquences et repérer toutes les formes impossibles (la quantité dépend de la qualité du scan) ; la liste triée peut alors servir à repérer automatiquement toutes les occurrences de formes rejetées.

Principaux sites pour le latin

- De nouveaux sites mettant en libre accès des textes anciens numérisés apparaissent constamment. Il faut donc commencer par une recherche approfondie des possibilités.
- Les grands corpus bien connus des médiévistes (MGH, Acta Sanctorum, PL) sont disponibles.
- Il existe de nombreux sites consacrés à un auteur (Thomas d'Aquin, Nicolas de Cues…)
- Le site www.thelatinlibrary.com est une ressource très largement utilisée. D'autres sites généralistes fournissent de plus en plus de textes : la bibliotheca augustana (Augsburg), wikisource (http://la.wikisource.org/wiki/Pagina_prima). Le site pot-pourri de Louvain est une importante ressource.
- D'autres sont plus spécialisés : epigraphische Datenbank, camena, poeti d'italia, neo-latin colloquia.
- Une catégorie très importante est constituée par les sites qui proposent en libre accès des chartes médiévales. Les CBMA à Dijon, le CDLM à Pavia, monasterium à Köln sont parmi les plus connus, mais il y aussi l'Artem de Nancy et le projet des Chartae Galliae.
- Divers sites proposent des metamoteurs spécialisés, on peut les utiliser mais ils ne sont jamais complets.

Techniques de récupération

- Les mises en ligne sont, au plan technique, des plus variables. Les cas les plus sympathiques sont ceux où un lien est prévu pour le téléchargement (les CBMA à Dijon par exemple). Dans de nombreux cas, on a accès à une « page html » que l'on peut simplement recopier. Tout dépend alors du contenu de cette page : si elle comporte l'oeuvre entière, c'est simple, sinon il faut récupérer les morceaux les uns après les autres, la situation se complique.
- Le cas le moins complexe est celui où les pages sont reliées par des « liens ». Dans ce cas, c'est le site entier qui est constitué par une sorte d'arborescence de pages. Auquel cas existe un outil très simple et très efficace, **la commande wget**. Il suffit d'indiquer l'adresse et de laisser le logiciel travailler, il récupère successivement toutes les pages liées.
- Dans d'autres cas, il n'y a pas de lien, mais les adresses d'une page et de la suivante ne diffèrent que par un numéro, il suffit donc de modifier l'adresse d'une unité ; en pratique, il faut écrire un petit script qui effectue successivement l'appel à toutes les pages.
- Une fois les pages html récupérées, le principal travail commence : il faut nettoyer les fichiers de tous les éléments inutiles et recoller les morceaux autant que nécessaire. Tous les cas sont particuliers, il faut analyser la structure des pages pour identifier les balises (tags) utiles, c'est-à-dire ceux qui marquent les limites des parties à récupérer. Si l'on doit traiter plus de quelques dizaines de fichiers, il faut écrire un script ad hoc, par exemple en perl. C'est un des objets de la présentation suivante.

QUESTIONS ?