Initiation à la chaîne d'édition numérique de corpus en SHS – Les fondamentaux
Session1

Plan

- Intro

Comme notre premier tour de table l'illustre bien, nous venons toutes et tous de disciplines différentes. J'ai donc essayé de trouver des problématiques communes à la diversité de vos approches disciplinaires pour élaborer cette présentation de l'édition numérique de corpus en SHS.

En effet, certains d'entre vous auront pour objectif de publier les brouillons d'un auteur pour peut-être établir la genèse d'un texte ou pour analyser le processus de travail d'un penseur ; d'autres voudront ré-éditer de vieux journaux avec un appareil critique ; d'autre encore voudront simplement mettre en ligne des données collectées dans un long travail de relevés, etc.

Là, déjà, nous pouvons entrevoir différentes approches disciplinaires entre littérature, philo, histoire. Mais aussi différentes traditions éditoriales liées à une discipline : critique génétique, anthropologie de l'écriture, édition critique, bases de données prosopographiques.

Au delà du problème de réussir à se situer dans une tradition éditoriale disciplinaire précise, il y a plusieurs autres problèmes qui vont se poser. Je vais vous donner 2 exemples.

Tout d'abord la définition de votre objectif scientifique lorsque vous aller vous vous lancer dans un projet d'édition numérique pourra être floutée par la multiplication d'outils de publication numérique mis à disposition, comme les documents partagés (framapad, google doc), les forums, les blogs, les logiciel de gestion de contenu comme Wordpress, Drupal, SPIP, etc. qui peuvent laisser penser que les contraintes de publication qui existaient pour le papier n'existent plus pour le numérique et qu'il est somme toute facile de mettre son corpus en ligne. Bonne nouvelle : oui, c'est vrai, on peut tout mettre en ligne ou presque très facilement aujourd'hui, mais mauvaise nouvelle : cela ne donnera pas forcément lieu à une édition savante de qualité.

Le deuxième problème est que la multiplication des données et des fonctionnalités utilisables sur Internet rendent difficile, pour l'utilisateur sans formation documentaire, la distinction des objets numériques : archives, journal, carnet de terrain individuel, carnet de terrain collectif, CV, annonce, etc. et tous les objets numériques nés des traditions éditoriales évoquées plus haut.

L'atelier de ce matin a pour objectif de vous donner quelques outils de réflexion pour la réalisation de votre projet. On apprendra ici à distinguer certains concepts ou objets, comme les documents structurés, les métalangages, l'encodage, la numérisation, le HTML, l'XML, etc. J'espère que ces définitions vous permettront de mieux définir les contours de votre projet et de vous motiver à être des acteurs dans la réalisation de vos éditions numériques.

- Partie I : Document structuré, structuration de l'information et document numérique
    - Le document source et ses zones signifiantes : Commençons avec 3 exemples de documents sources dans lesquels nous allons essayer de repérer les zones signifiantes.
      Exercice : analyse de 3 documents
        - Brouillon manuscrit de M. Foucault
        - Lettre manuscrite du XVIII
        - Edition papier 1ère page de l'Echo de la Fabrique

Transition : ce travail de repérage de zones signifiantes que vous venez d'effectuer ici, dans différents documents types, n'est pas anodine et est à l'origine de plusieurs tentatives de modélisation pour l'informatique depuis le milieu du siècle dernier. Pour vous expliquer comment on en est venu à modéliser la structure des documents je vais commencer par vous donner quelques exemples des premières tentatives d'utilisation de l'informatique pour éditer des textes.

- La fouille de texte et le document structuré
    - Index Thomisticus : En 1949, un prêtre jésuite italien, Père Roberto Busa, démarre son *Index Thomisticus* un index exhaustif papier de tous les mots utilisés dans les œuvres de St Thomas d'Aquin et des auteurs apparentés. Au total cet index représentera 11 millions de mots du latin médiéval. Pour l'aider dans sa tache il entame une collaboration avec Thomas J. Watson à IBM au EU. Cette collaboration marque pour lui le début d'un travail qui durera 30 ans : pendant qu'IBM développe un programme de calcul de concordances lemmatisées, l'ensemble des textes sont peu à peu transférés sur des cartes perforées. A la fin des années 1970, 56 volumes papier.
    - A partir des années 60, beaucoup d'autres chercheurs commencent à voir l'intérêt de travailler sur l'analyse des concordances de mot dans les œuvres qu'ils étudient ou pour déterminer le véritable auteur de certains textes, quelque chose qu'on essayait déjà de faire à la fin du XIXeme, avec les œuvres de Shakespeare.

- Là je vais introduire le concept de marquage informatique ou langage de modélisation ou langage d'encodage. Je vais essayer de vous expliquer ce que j'entends par marquage du texte : ce que vous avez fait de manière assez intuitive tout à l'heure pour distinguer des zones signifiantes, un ordinateur ne pourra pas le faire si on lui donne un document texte tel quel. Pour lui un texte est simplement une suite de caractères où, chaque caractère, a un nombre spécifique de 0 et de 1. Ainsi, chaque caractère se vaut. C'est donc à nous d'indiquer à l'ordinateur le début et la fin de certaines suite de caractères pour qu'il les distingue du reste du texte. Il nous faut donc lui donner une sorte de panneau de signalisation au début et une autre à la fin pour lui signifier nos fameuses zones signifiantes : ces panneaux de signalisation sont appelées des « balises ». Selon les langages elles seront exprimées dans différentes syntaxes. Voici quelques exemples d'invention de langage de structuration du document.
- COCOA : A la fin année 60, début années 70, afin de réaliser *l'Archive of Older Scottish texts*, Paul Bratley développe le modèle d'analyse de texte COCOA. Celui-ci s'appuie sur un processeur de mots composé d'un programme de cartes perforées FORTRAN. Il permettait de compter les mots d'un texte et de créer des cooccurrences. Son originalité est d'introduire un système de marquage du texte permettant de repérer des informations simples dans le document mais également de définir sa propre spécification de la structure du document.
  Voici un exemple de syntaxe Cocoa, ici les balises sont indiquées par des chevrons et une lettre majuscule qui indique une catégorie particulière. Le <T indique le titre, le <C indique le charatère (le personnage), les (()) doubles parenthèse indiquent les didascalies. Chaque catégorie est considérée comme vraie jusqu'à ce qu'une nouvelle instante de la même catégorie soit indiquée.
- GML : En 1969/1979, Charles Goldfrap est juriste et invente le GML (Charles Goldfrap, Edward Mosher, Raymond Lorie ou Generalized Markup Language), un modèle de structuration du document qui doit servir à l'élaboration d'un système de gestion intégré du droit, c'est à dire de faciliter la recherche de textes de loi et de la jurisprudence associée. Le GML sera développé par IBM.

        - Parmi les balises repérant la structure logique des documents de loi, on trouve par exemple l'adresse, le résumé, le corps, la citation, la date, l'auteur, le titre, etc. mais également des informations propres aux textes de loi.
        - SGML : son extension, est le SGML (Standard Markup Language) est modélisation orientée objet des documents et des hyper-documents, à l'origine du XML et HTML.

Transition : je pourrai vous donner d'autres exemples de structuration du document et de repérage de zones signifiantes en revenant sur l'histoire de la géographie, de la paléographie, l'archéologie, etc. On trouverait alors plein d'autres modèles descriptifs, mais l'objectif n'est pas d'être exhaustif mais de vous montrer comment ces différentes tentatives de structuration du document et de l'information on conduit à l'Internet qu'on connaît aujourd'hui, à la naissance d'HTML, d'XML, etc.

Passons maintenant à la structuration de l'information, et à la notion de liens entre les informations, qui semble si évidente aujourd'hui, mais qui a aussi une histoire dont voici quelques fragments.

- Structuration de l'information
    - Un premier exemple de tentative de structurer les connaissances, que vous connaissez bien, c'est *L'encyclopédie Diderot et d'Alembert*, avec ses renvois classiques (vers des définitions de mot), ses renvois de choses (pour confirmer réfuter une idée), ses renvois dits « de génie » (vers des idées pouvant mener à des inventions), et ses renvois satiriques (pour éviter la censure) qui vont relier ainsi 72000 articles. Ces nombreux renvois et la réflexion de Diderot sur leur usage font qu'il est parfois considéré comme l'ancêtre de l'hypertexte.
    - MEMEX : Si on se rapproche de notre époque, en 1945, dans un article intitulé *As We May Think* le Docteur Vannear Bush encourage les scientifiques à développer des outils pacifistes qui rendront les informations plus accessibles et permettront à tous de maîtriser les connaissances accumulées à travers les âges. Dans cet article Docteur Bush décrit un bureau électromécanique futuriste (ou machine analogique fictive), le MEMEX (ou *Memory Extender*), qui permettrait à l'utilisateur de naviguer parmi les informations, grâce à des liens entre des paires d'images de microfilms et des liens linéaires entre des informations. Le Memex permettrait également l'ajout d'information par l'utilisateur (des photos ou du texte) à partir d'un écran tactile transparent).
    - MARC : Dans le monde des bibliothèques c'est le langage MARC (le Machine Readable Cataloging) qui est développé en

1965 par Henriette Avram, chercheure en informatique américaine. Elle travaille pilote le projet MARC à la bibliothèque du Congrès aux Etats Unis qui aboutit en 1968. Ce projet prévoit la création de notices descriptives lisibles par les machines et échangées entre les bibliothèques. En 1969 on lance le premier service d'échange de notices au format MARC. Il s'agit d'utiliser pour chaque notice bibliographique une succession de champs de données. Chaque champ est séparé par un dollar ($).L'article qui décrit le MEMEX est la première description théorique de ce que pourrait être l'hypertexte, et est à l'origine de la création d'internet et du web.

L'article qui décrit le MEMEX est une description théorique de ce que pourrait être l'hypertexte, qui directement inspiré la création d'internet et du web.

- La représentation sur la toile du document structuré
  - En effet, l'hypertexte c'est, comme on le voyait dans la description du Memex, un texte exposé sur une machine qui permet le renvoi vers d'autres références et c'est le concept qui sous-tend la création du web.
  - Alors je vous passe toutes les étapes de l'origine d'Internet (TCP/IP), mais la première mise en réseau de réseaux d'ordinateur remonte à ARPANET au début des années 60. Les ordinateurs pouvaient communiquer entre eux par l'envoi de paquet, mais ça n'avait encore rien à voir avec le web tel que vous le connaissez aujourd'hui.
  - La Naissance du Web ou du World Wild Web, c'est la fusion de 2 concepts : celui d'Internet (mise en réseau d'ordinateurs) et celui du document structuré ou d'hypertexte.
  - Le Web ne représente qu'une des applications d'internet. (autres exemple : mails, ftp, tchat, etc.) et a été inventé par Tim Berners-Lee en 1989 pour permettra aux chercheurs-visiteurs du CERN (Conseil Européen pour la Recherche Nucléaire -centre de physique des particules-) d'échanger des informations scientifiques (articles, rapports) après leur séjour au Centre. Son principe est fondé sur l'extension du concept d'hyperdocument (un document avec du texte, de la vidéo, du son) aux réseaux internationaux (hyper documents répartis) et la création d'un accès plus convivial aux serveurs existants (avant client/serveur). On créer alors les navigateurs web qui permettront l'affichage de document /page (le premier navigateur Mosaic).
  - HTML (liens hypertexte) C'est là qu'on arrive enfin au document HTML. HTML ça veut dire HyperText Markup Language. Il s'agit

d'un modèle de représentation d'hyperdocument pour lequel on définit à la fois :
    - Des nœuds logiques
    - La structure physique et sa représentation
- => objectif : langage utilisé pour représenter/rendre à l'écran les pages que nous allons consulter via un navigateur. C'est le langage utilisé pour présenter le résultat de notre consultation (consultation = requête)
  Le HTML est en constante évolution depuis les années 1989 (
    - HTML 1 1989 (textes, qques styles, liens hypertextes)
    - HTML 2 1994 (html1 + images + formulaires)
    - HTML3 1996 (html2 + graphiques vectoriels + son + applets)
    - HTML4 1998 (html3+ vidéo + CSS + outils pour intranet)
    - HTML5 2014 (html4 + conformité à la syntaxe XML, plein de nouveaux éléments / 8 nouvelles API pour créer des applis web : des dessins 2D, intégrer des vidéos, de l'audio, etc.)
- Ce fichier est sur un serveur, quand je l'appelle dans un navigateur, le navigateur va bâtir ce qu'on appelle un DOM, une sorte d'arbre dans lequel chaque balise va être interprétée comme un noeud.
- *Exercice 1 :*
    - je créé mon premier fichier HTML : exo1.html
    - je l'ouvre dans un navigateur
- *Exercice 2 :*
    - je créé mon premier fichier CSS : exo2-representation.css
    - je fais un lien vers lui depuis mon fichier html en ajoutant dans le *head* la ligne :
      ```
      <link rel="stylesheet" href="exo2-representation.css">
      ```
    - je regarde à nouveau mon fichier sur internet

- L'essor des standards de métadonnées
    - XML : eXtensive Markup Language : au départ il s'agit d'une forme extensive qui permet de définir ses propres balises (aujourd'hui HTML respecte le format XML) Comme HTML il vient de la famille de modélisation des documents SGML.
    - XML et DTD :
      XML permet de repérer une structuration logique d'un document, exactement comme nous l'avons vu tout à l'heure avec le COCOA, le GML et le SGML. Voici un exemple de fiche descriptive d'un brouillon de Michel Foucault extrait du même fonds que tout à l'heure.

        - On voit que là le choix du modèle descriptif est propre au travail archivistique. C'est le format EAD (Encoded Archival Description)
        - En effet comme je vous le disais tout à l'heure, une fois qu'on a choisi un langage avec une syntaxe particulière (entre cocoa, gml, sgml, xml) on peut choisir son propre modèle de balises descriptive : on appelle ça des standards ou format : il s'agit d'une une sorte de dictionnaires de toutes les balises qu'on aura le droit d'utiliser. Les standards de métadonnées sont depuis longtemps utilisés dans le monde des bibliothèques pour le catalogage informatique, comme les standard MARC, UNIMARC etc.) et, dès les débuts de l'internet, (Dublin Core) – pour les images METS – pour les archives EAD – On en aura d'autres pour la géographie, etc
        - TEI : Alors, pour revenir à l'édition numérique : Le standard ou format qui va nous permettre d'encoder des éditions savantes, connait ses prémisses en novembre 1987. Invité par Nancy Ide, au Vassar College in Poughkeepsie, plusieurs chercheurs spécialistes du texte, des formats d'encodage ou d'informatique, commencent à réfléchir ensemble à un encodage pour les humanités numériques. Ils en ont marre de devoir reformater leur texte pour les rendre compatibles avec les différents logiciels existants : Ainsi nait la Text Encoding Initiative et ses *Guidelines for Electronic Text Encoding and Interchange* cad pour l'encodage et l'échange de texte électronique. Vous pourrez en apprendre plus sur l'histoire de la TEI en lisant les articles de Lou Burnard. Voici deux références qui racontent l'évolution de la TEI, au départ simple un projet de recherche jusqu'à devenir aujourd'hui une véritable infrastructure.

    - Imaginons maintenant des sites web avec des centaines de pages et des milliers de lignes de contenu, comme ce que peuvent représenter des corpus textuels dans le cas de réédition ou dans le cas d'édition de manuscrits). Pour éditer numériquement ces gros volumes, on va devoir chercher à automatiser au maximum l'affichage et non pas nous mettre à créer une page web par document. Si on prend l'exemple du corpus de Morand qu'Anne connait bien, on a 350 lettres en ligne. On ne va pas créer une page html à la main par lettre. On va faire en sorte, qu'à partir de la transcription on puisse créer les 350 fichiers html automatiquement, ou à la volée.

- Il nous faut donc stocker les informations/ la transcription, les biographies, les entrées d'index, etc. dans une base de donnée. On écrira ensuite des scripts informatiques qui iront chercher ces informations ou données sur les données ou métadonnées dans la BDD, pour les afficher selon un modèle/pattern/canevas prédéfinit.
- Il existe pour cela 2 types de BDD :
  - Les plus connues et utilisées sont les BDD relationnelles [diapo 31 à 34] On a des tables pour chaque type d'info, et ses tables sont liées les unes aux autres.
  - L'autre type de BDD c'est le XML
    [diapo 36 à 40] Une sorte de super fichier ressemblant à un fichier HTML, arborescent lui aussi, dans lequel chaque nœud élément va représenter une métadonnée à l'intérieure de laquelle on pourra trouver un nœud textuel et des attributs. Ce type de BDD permet de stocker de larges textes et est donc relativement bien adapté à des corpus textuel.
  -

Partie II : Exemple d'une chaîne des opérations d'une édition numériques simple ; Le Roman des Morand

- Les 1ers marquages de zones signifiantes [diapo 43] sont souvent effectuées par le chercheur, dans un logiciel de traitement de texte, comme Word, ou Open office. Mais les étapes de transformations récurrentes vers le XML sont très chronophages et sources d'erreur. C'est pourquoi on vous fera travailler dans le XML directement.
- Voici un exemple de stylage
  Dans cet exemple de lettre chaque personne citée a donné lieu à un renvoi vers une forme normalisée. Par exemple la sœur de l'expéditeur ici appelée "sa femme" (sous entendu "la femme de mon beau frère") est normalisée sous la forme EleonoreDeBesson (plus tard cette forme sera remplacée sous forme d'un identifiant unique numéral). On voit que si une même personne citée peut apparaitre sous différentes formes dans la lettre elle est sera toujours stylée sous la même forme normalisée.
- Par exemple ici Antoinette Morand pourra être interpellée de façon directe "Chère maman"" ou évoquée par l'expéditeur comme la personne qu'il voudrait qu'elle soit : "une bonne mère" de manière à faire culpabiliser son interlocutrice. Par contre, dans le style appliqué, qui plus tard pourra être transformé par exemple en entrée d'index, la personne citée aura toujours la même forme normalisée, AntoinetteMorand par exemple.

    - A partir du fichier texte on va pouvoir récupérer un fichier XML, qui après quelques nettoyages devrait ressembler à quelque chose comme ça
    - Mais ce fichier, s'il est uniformisé au niveau local, c'est à dire entre nous trois, ne sera pas interopérable ou exploitable à l'extérieur.
    - Donc il va falloir transformer une nouvelle fois ce document XML pour qu'il soit interprétable par d'autres universitaires et d'autres machines.
-

- Partie III : Vos projets Tour de table