Mardi 8 novembre 2016

Séance AHN Retour d'expérience *Encodage et DTD*

Introduction

La séance d'aujourd'hui est consacrée au formats d'encodage et aux DTD. Mais avant de parler des formats qu'est-ce qu'un encodage?

Différents métiers sont représentés dans la salle (documentalistes, sociologues, historiens, développeurs),… et chacun.e possède des définitions différentes ou complémentaires du terme « encodage ».

On peut donc redonner ici quelques définitions et repères historiques qui permettront d'expliciter le terme « encodage » dans contexte de recherche en sciences humaines et sociales ; et qui permettront également de rappeler la complémentarité de nos différents métiers.

Gardons en tête que l'encodage pourra nous servir *a minima* à 2 choses complémentaires :

- structurer les informations
- structurer les documents
- représenter les doc et infos structurés sur la toile.

En effet, à partir du moment où on travaille sur un ordinateur ou sur internet, que l'on cherche à décrire des documents, des données prosopographiques, et à les organiser, ou que l'on cherche à créer des index de mots, ou à repérer des zones signifiantes d'un texte afin d'en faciliter l'exploration, l'analyse ou la navigation, on doit indiquer à l'ordinateur quelle est la structure des informations et la structure des documents.

En effet, pour l'ordinateur 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 ».

Encodage pour structurer les informations

Le premier exemple connu de tentative de structuration des connaissances est l'*Encyclopédie* Diderot et d'Alembert avec 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.

— 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 qui pilote le projet MARC à la bibliothèque du Congrès aux Etats Unis qui aboutit en 1968. Ce projet prévoit la création de notices descriptives, avec une succession de champs de données séparés par un dollar ($), notices lisibles par les machines et échangées entre les bibliothèques sous forme de bandes électromagnétiques. En 1969 on lance le premier service d'échange de notices au format MARC.

Encodage pour structurer les documents

Les balises qui sont les marques d'un encodage. Un encodage peut être exprimé dans différents langages :

— 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 de formation. Avec Edward Mosher et, Raymond Lorie il invente pour IBM le GML (le Generalized Markup Language ou Goldfrap, Mosher & 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 pour les juristes.

— SGML : Goldffrapp quitte IBM pour développer l'extension de GML : c'est le SGML (Standard Markup Language) : une modélisation orientée objet des documents et des hyper-documents, à l'origine du XML et HTML.

Encodage pour représenter sur la toile les infos et documents structurés

Avec la création d'internet préfiguré par le Memex, on bénéficie pour la première fois de 2 choses : la mise en réseau des ordinateurs et la mise en relation des informations et documents entre eux avec l'hypertexte.

— HTML :
En effet, en 1989 au CERN on créé l'HTML ou Hyper Text Markup Language 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).

Modèle de représentation d'hyper-document définit à la fois par :
- des noeuds logiques
- une structure physique et sa représentation

HTML = en constante évolution :
1989 : HTML-1 / 1994 : HTML-2 / 1996 : HTML-3 / 1998 : HTML-4 / 2014 : HTML-5

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
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.

Le nom des balises utilisées dans le langage XML est libre. Des groupes d'intérêts se sont donc constitués pour définir leur propre modèle de balises descriptive : on appelle ça des standards ou format ou DTD (Document Type Definition).

Utilisés dans le monde des bibliothèques pour le catalogage informatique, comme les standard MARC, UNIMARC etc.) et,
ou utilisé :
- pour l'internet, (Dublin Core)
- pour les objets numériques METS
- pour les archives EAD
- pour la transcription de corpus : TEI

Nos trois intervenants de ce matin Vincent Ventresque, Vincent Baas et Carole Etienne vont nous donner 3 exemples de formats d'encodage ou DTD XML utilisés dans le contexte de la recherche en SHS.