Réunion du groupe méthodologique dans le cadre du GIS : 23 juin 2017
Présents : *Arnaud Millereux, Sitthida Samath, Agnès Viola, Céline Alazard, Violaine Chaléat-Fonck, Sébastien Jacquot, Serge Wolikow, Maxime Tissier, Aurélie Hess (en visioconférence), Aurelia Vasile*

Le but de la réunion a été d'établir les sujets techniques et méthodologique d'intérêt commun pour les membres du GIS, le type d'activité envisagé pour faire connaître le sujet et un calendrier prévisionnel.

Projets :

1. Chaîne de traitement : de l'océrisation à l'xml-ead pour les revues

Objectifs : la méthodologie pour créer des instruments de rechercher de manière semi-automatique, à partir de l'ocerisation des tables de matières de revues.

Il s'agit de montrer toutes les étapes de la chaîne de traitement qui comprend plusieurs phases:

- la sélection et l'océrisation des tables de matières
- l'identification de tous les modèles de tables de matières. (on s'est aperçu que sur la durée de vie d'une revue la structure du sommaire change) l'ordre des éléments (auteur, titre, numéro de page) étant essentielle pour la modélisation
- établissement d'une règle de traitement des données (quelles sont les conditions que les données doivent remplir pour que l'automatisation fonctionne)
- le traitement des fichiers textes issus de l'océrisation : correction des erreurs issus de l'océrisation, élimination des éléments auxiliaires aux sommaires afin de garder la structure nécessaire pour faire fonctionner le programme informatique.
- Choix à faire pour l'encodage des données.

Forme de présentation : atelier 2h
Equipe : Aurelia Vasile, Arnaud Milleureux
Calendrier: automne 2017

Étude de cas à partit des ressources suivantes :
Revue *La Nouvelle critique*
Revue *L'Ecole et la Nation*

Réutilisation et amélioration

- Application de la méthode à d'autres projets : (par exemple aux rapports des préfets préservés aux AN qui contient une informations déjà structurée et qui s'adapte en conséquence à un traitement semi-automatique).
- Export des données dans une multitude de formats de type xml et pas uniquement l'EAD.
- Utilisation de l'xml-alto pour la compréhension de la table de matière

2. Réflexions sur l'indexation, l'enrichissement et l'interrogation des corpus

2.1 Atelier Correcteur d'index

Objectifs : Corriger de manière semi-automatique une liste d'entrées d'index qui comprend des erreurs de saisies

Forme de présentation : atelier 2h
Equipe : Agnes Viola, Aurelia Vasile, Arnaud Milleureux
Calendrier: automne 2017
Étude de cas à partir des fonds de présents sur Pandor (à préciser)

2.2 Réflexion sur la normalisation des index, l'utilisation de référentiels ou la création de listes d'autorité

- Noms géographiques (noms anciens (Bessarabie, Perse, etc),
- Mots-clé
- Noms d'organismes

Questions et autres pistes de réflexions :

- utilisation de référentiels déjà existants ou création de listes propres,
- échange de jeux de données entre différentes institutions (par ex. les AN ont un référentiel très riche sur les noms géographiques d'Ile de France. Une partie des référentiels pourra être trouvée en ligne sur data.culture.gouv.fr)
- logiciel Analec (dans la plate-forme TXM) qui est un outil d'annotation et d'analyse de corpus écrits. Il permet de faire des « chaînes de correspondance » : permet de « taguer » les mêmes entrées sous le même nom
- logiciel Symogy (base de référentiels en ligne) logiciel utilisé plutôt par les géographes. Du laboratoire LARHRA, unité de recherche en histoire, s'intéresse au web sémantique et la définition d'ontologies depuis plusieurs années avec le développement de la plateforme symogih. Autour de leurs travaux, un groupe de travail est en train de se structurer en formant un consortium : 'Données pour l'histoire numérique'

3. Séminaire web sémantique

Organiser une journée d'études pour connaître les enjeux du WS et le fonctionnement technique. Intervenants possibles

- Sonia Hamdi-Guérin (ISH)
- Emmanuelle Bermès (BNF)
- Viviane Boulétreau (Data Persée)

4. Restituer visuellement des données issues d'instruments de recherche (ex. cartographie, géolocalisation). Réaliser des entrées par zones géographiques.

Étapes du traitement :

- extraction à partir des fichiers l'xml ead
- création de fichier calque
- relever les noms des lieux
- utilisation de logiciels de restitution cartographique libres

Étude de cas :

- dossiers de police judiciaire qui signale tous les meurtres et assassinats commis depuis 1940 aux années 1960. Cette méthode permet de faire des catégories par années, par type de meurtre (de droit commun, liée à la Résistance)
- rapports de l'inspection de camps internement 1940-1944

Équipe: Archives nationales

5. Croiser les données issues d'instruments de recherche pour en faire ressortir des corpus communs

(ex. croiser les listes de personnes concernées par des dossiers individuels d'archives dans les fonds de la DST, de la police judiciaire et de la cour de justice de la Seine).

Ce type de projet peut intéresser aussi l'équipe de la MSH qui se confronte avec le même type de besoins.

Équipe: Archives nationales

Calendrier pour les deux ateliers des AN : Ateliers de 2-3h à organiser plutôt au début de l'année 2018

6. Archives de la recherche

Objets de travail :

a) Réalisation d'un état de l'art
b) Questions juridiques (fonds publics ou fonds privés : fonds de chercheur, fonds thématiques (programme scientifique collaboratif, fonds organiques (de

laboratoire), prise en charge des fonds, conventions, communicabilité, anonymisation, etc.)

c) Question du traitement technique et descriptif avec notamment l'indexation (importance du vocabulaire/index par discipline concernée).
d) Question des fonds « mixtes » : archives et documentation / papier et numérique
e) Archivage pérenne : appui sur institutions et organismes spécialistes de l'archivage à long termes. Travail sur la réalisation d'un cahier des charges en ce sens.
f) Valorisation : fonds avec des données personnelles. Il faut travailler pour la publication à des échelles ouverte / restreinte

Atelier : partir des cas concrets, d'une problématique spécifique

D'autres problématiques de réflexion au sein du GIS:

- La question de la mixité des fonds :
  - dans la nature des documents : papiers, cartes, publications, films, hétérogénéité de l'indexation.
  - les nouveaux fonds des chercheurs (ressources nativement numériques, films, documents audio, etc) ce qui implique une méthodologie différente.
  - Les ressources obsolètes des fonds des chercheurs : comment convertir des anciens formats pour récupérer l'information.
- Proposer une réflexion sur l'indexation : le vocabulaire par discipline (la définition et la réappropriation par les autres disciplines, à choisir ou pas Rameau).
- La question de la collecte, du dépôt et de la conservation. (pensée une chaîne de traitement)

La question de la mixité des fonds concerne aussi le projet sur Jean-Luc Lagarce de Besançon (http://fanum.univ-fcomte.fr/lagarce/).

Calendrier à définir

Équipe coordinatrice : Céline Alazard, Samath Sitthida (ISH), Maxim Tissier, Aurélie Hess, Julie Demange.

7. Le choix des outils

Objectifs : il s'agit plutôt de faire un échange sur les outils existants, de faire une comparaison des outils existants, notamment sur les plate-formes modulaires de type Omeka (qui s'appuie sur des plugin pour apporter de nouvelles fonctionnalités)

Solution : faire un inventaire des outils existants qui sont utilisés de manière systématique par nos partenaires (dans un premier temps)

Démarche : faire un questionnaire/une enquête à envoyer aux utilisateurs et aux fabricants pour avoir plus d'éléments sur les fonctionnalités, les formats pris en charge, les réponses apportées aux besoins des chercheurs, pourquoi un outil a été abandonné etc.

8. Traitement des archives nativement numériques (mail, archives du web).

Aspect qui n'a pas été abordé en détail. Cet objet sera-il traité de manière autonome ou dans le cadre des archives de la recherche ? A réfléchir.

Équipe coordinatrice: Maxim Tissier, Julie Demange, Françoise Blum (?)