Synthèse de la séance de démarrage du MODOAP

1. Présentation générale du projet

Julien Schuh

Le but de la séance est de :

- Présenter Cyril Bruneau, IGR associé au projet, et les premiers résultats obtenus par les outils développés
- Redéfinir les objectifs de chaque sous-projet, discuter des corpus, des possibilités
- Mettre en place un calendrier du séminaire et rappeler les principes de ces séminaires ateliers (2-4 séances par projet, séances de travail et séance terminale de présentation de résultats)

2. Présentation des outils en développement

Cyril Bruneau

Les outils actuellement en développement au sein du ModOAP ont été présentés.

- Détection et extraction d'illustrations dans les périodiques
- Détection et extraction d'objets dans les images
- Calcul de similarité entre images et repérage de doublons

- Classification
- Textométrie
- Topic Modeling

Il a été souligné que certains outils nécessitent un auto-apprentissage, ce qui implique des tâches d'annotation de corpus (réalisables notamment avec le logiciel Annotate présenté plus tard par Gilles Bertin)
Les résultats obtenus sont dans un premier temps stockés dans des dictionnaires ou des fichiers structurés.
Les scripts actuellement fonctionnels sont disponibles sur https://github.com/MODOAP/corpus_test/releases/ au format Google Colab .ipynb accompagnés de mini-corpus test.

3. Sous-projet Kagan

Cécile Tardy, Cyril Burté

Des précisions ont été apportées sur l'exposition Elie Kagan prévue pour Octobre 2021, qui s'articulera autour de 3 pôles :

A. La réception des images d'Elie Kagan comme photographe de presse, à partir de tirages et de presse imprimée, sur les thématiques de la mise en scène dans le reportage et la construction d'un récit à partir d'images parfois réutilisées, recadrées, etc..

B. La réception et la mobilisation des images d'Elie kagan dans les milieux militants, et notamment le traitement de la journée du 17 octobre 1961, dont la plupart des tirages n'ont été diffusés que des décennies plus tard.

C. Un travail sur le quotidien de Paris et les petits métiers. La détection d'objets sur ces photographies peut être intéressante.

Il est également question d'une borne multimédia au sein de l'exposition, qui permettrait de naviguer dans le corpus.

Concernant les corpus :

L'activité d'Elie Kagan en tant que photographe de presse s'étale principalement de 1961 à 1988. Le fonds est constitué de négatifs, de tirages, de planches contact et d'archives, et principalement de négatifs numérisés correspondant à la période 1961 – 1975. D'autres numérisations ne sont pas datées. Un inventaire de ces numérisations communiqué par Cyril Burté est donc désormais en ligne sur : http://www.calames.abes.fr/pub

Le journal Rouge est en cours de numérisation, et pourrait être disponible d'ici fin Novembre.

La numérisation de la presse est un soucis dans l'ensemble. Il faudrait voir si la BNF dispose de certains corpus, et parallèlement s'il est possible de faire intervenir des vacataires pour la numérisation. Les périodiques concernés sont France Observateur (le pré-Nouvel Obs de 1961 à 1965), Libération, le Nouvel Obs et Témoignage Chrétien.

4. Sous-projet Manuels scolaires

Laurence Jung, Xavier Riondet

Le corpus comprend 68 manuels scolaires océrisés dont 52 sur la période 1880-1914. Ils correspondent au niveau primaire et supérieur, jusqu'au certificat d'études. Deux disciplines sont privilégiées : **Histoire** et **Leçons de choses**. L'étude du corpus s'articule autour de deux axes :

A. Questions sur les enjeux de l'école Républicaine et lien avec l'avènement de la presse scolaire :

Etudier la représentation de la société française dans les manuels à travers la Leçon de choses (pédagogie sur les sciences), les éventuelles prises de position des éditeurs, les connotations associées aux métiers : hypothèse d'une vision de la société véhiculée pour des enfants populaires, et donc la surdétermination culturelle par la scolarité.
En plus de l'analyse textuelle, les illustrations pourront être exploitées : analyse de la provenance des images (revues scientifiques, musées).

B. La lutte des références (plutôt pour la discipline Histoire)

Etude du récit national, des choix de traitements de personnages ou d'évènements, en positif ou négatif, et comparaison des auteurs sur des sujets identiques.
Etude des représentations des personnages en images : figure de l'homme à cheval, du soldat…
Il est aussi possible de comparer ces représentations de l'histoire à des manuels destinés spécifiquement aux colonies, et peut-être d'ouvrir la réflexion à des comparaisons internationales.
Comparaison également à la presse généraliste de l'époque : les thématiques de la presse sont-elles transférées dans les manuels ?
Pistes d'analyse : reconnaissance d'entités nommées dans les textes pour cibler les personnnages historiques (en partenariat avec MoDyCo), classification automatique des idéologies des manuels, observation de tendances dans les manuels par l'analyse notamment de segments répétés.

5. Intégration dans Huma-Num et BnF Data Lab

Jean-Philippe Moreux, Nicolas Sauret, Stéphane Pouyllau

La plateforme propose un GitLab communautaire. Elle ne permet pas l'utilisation de notebooks Jupyter, pour des raisons notamment de sécurité. Il est possible cependant de profiter d'une puissance de calcul sur une autre grille, avec une interface Anaconda. Possibilité d'exécuter des notebooks via Binder ou la mise en place d'une machine virtuelle.

Le but du projet Huma-Num Lab est de raccrocher les services d'Huma-Num et resynchroniser ce qu'ils peuvent industrialiser avec les besoins des communautés.
- Réfléchir à des interconnexions de boîtes à outil et faciliter accès au DataLab pour les communautés
- Se dirriger vers une plateforme de template spécialisés SHS comme FloydHub (http://www.floydhub.com)
- Anticiper les usages des services d'Huma-Num

Pour ces raisons, le projet ModOAP est vu comme un terrain permettant d'alimenter réflexion du Lab.
Des questions ont été soulevées concernant l'utilisation de notebooks, l'appropriation de ces méthodes par les chercheurs et la production de connaissances nouvelles.
L'utilisation et la compréhension de ces techniques nécessitent un déclic chez le chercheur, qui impliquerait de repenser la logique bibliothéconomique et d'inverser le rapport aux sources et à l'archive.

Enfin, il est possible d'utiliser Tesseract sur Huma-Num via un environnement sur un serveur dédié. Le service fonctionne avec un système de location de jetons, en raison d'une mauvaise utilisation antérieure.

6. Articulation avec Annotate

Gilles Bertin

L'outil Annotate permet d'annoter des zones dans des images fixes de manière ergonomique, et le projet OPPA IIIF cherche à l'étendre à des collections hétérogènes, comprenant la vidéo.
L'un des usages consiste à annoter des documents pour la phase d'apprentissage des algorithmes. Il sera alors possible d'exporter des formes comme les polygones.
Une partie annotation collaborative est en développement, elle permettra l'annotation dans un navigateur . Il sera également possible d'exporter des manifestes IIIF.

7. Calendriers

Il s'agit de construire progressivement pendant les séminaires des problématiques de recherche comme l'identification des données, et de développer un protocole d'analyse des problématiques.
Chaque sous-projet (Kagan et Manuels Scolaires) pourra être consulté alternativement un mois sur deux.
Les corpus des Manuels Scolaires étant prêts, une prochaine réunion peut être envisagée fin octobre.
Pour le sous projet Kagan, une réunion est préférable fin novembre afin d'attendre la numérisation du journal Rouge.

Un évènement Evento sera adressé aux représentants des sous-projets afin de décider d'une séance aux périodes prévues.