Rapport de Projets : Histoire de l'Art et Coopérations Franco-Américaines

Question de Recherche

L'Histoire de l'Art est un domaine particulièrement riche en échanges transatlantiques, tant sur les sujets étudiés (périodes, mouvements, artistes français) que sur les trajectoires professionnelles et de formation (diplômes, séjours de recherche). Ainsi, dans un domaine foisonnant de circulation intellectuelle, des liens s'esquissent entre chercheurs d'université américaines et champ académique et culturel français.

C'est dans ce contexte que nous nous sommes posés les questions suivantes:

Comment les affiliations thématiques et les collaborations explicites avec la France se structurent-elles au sein d'un corpus de chercheurs en Histoire de l'Art des universités américaines ? Quels domaines de recherche constituent-ils des ponts privilégiés entre les deux scènes académiques ?

I. Constitution et harmonisation du Corpus

1. Constitution du corpus

Le corpus a été initialement constitué à partir de profils publics de chercheurs et d'enseignants en Histoire de l'Art (ou autres domaines équivalents dans six universités américaines: **BYU, Rice, Harvard, Stanford, Williams, Michigan**.

Le travail de départ a consisté à:

- importer et filtrer les données issues de la collecte par la classe (scrapping) de chaque université
- sélectionner uniquement les notices rattachées à un département d'Art History ou un intitulé similaire (par exemple "Comparative Arts and Letters" ou "History of Art")
- nettoyer et normaliser les noms de colonnes pour les rendre uniformes et exploitables (mise en minuscule, suppression des espaces et des guillemets)

2. Harmonisation des données

Puisque nos données brutes, de sources hétérogènes, présentaient des colonnes différentes pour le même type d'information, nous avons procédé à une concaténation de six bases de données filtrées puis harmonisées pour obtenir un jeu de données unique et cohérent.

Cette étape a permis de consolider les informations dans des colonnes standardisées: les informations sur le nom du chercheur, l'université, la page de profil, les domaines de recherche, la biographie, la position professionnelle, les contacts, les prix et les formations

ont été agrégées à partir de multiples colonnes sources (par exemple, on a procédé à la fusion des colonnes research, research_areas, academic_interests, subfields et fields dans la colonne unique expertise). Les colonnes initiales et intermédiaires ont été supprimées ensuite pour ne conserver que la structure finale et nettoyée.

II. Enrichissement et Analyse Thématique

Pour transformer le corpus textuel en données structurées pour l'analyse, deux étapes d'enrichissement ont été réalisées:

1. **Extraction d'informations par Expressions Régulières (Regex)**

Nous avons regroupé le texte de l'ensemble des colonnes pertinentes de chaque profil (expertise, biography_summary, education, etc) dans une colonne unique (__full_text__). Celle-ci a servi de base pour la détection automatique de marqueurs thématiques et géographiques.

Nous avons utilisé des dictionnaires de mots-clés et des expressions régulières (Regex) pour identifier et extraire:

a. **les marqueurs liés à la France**: nom de pays, adjectifs (French, franco-américain), grandes villes (Paris, Lyon, Marseille), institutions artistiques et universitaires françaises (Louvre, Musée d'Orsay, Sorbonne, Beaux-Arts de Paris)
b. **les marqueurs de l'Histoire de l'Art**:
- périodes (*Medieval, Renaissance, Baroque, Modern, 19th century*)
- mouvements artistiques (*Impressionism, Surrealism, Cubism, Dadaism*)
- médiums et supports (*painting, sculpture, architecture, photography, video art*)

Ces extractions nous ont permis de créer des colonnes d'étiquettes (france_markers, art_periods, etc) et surtout une variable binaire pivot, link_france_regex (**Yes/No**) qui indique un lien explicite avec la France, ainsi que d'extraire la phrase exacte qui cause cette détection (france_link_quote), ce qui nous permet d'avoir une validation.

2. **Extraction d'informations par l'API OpenAlex**

Pour affiner l'analyse des thèmes de recherche réels et des collaborations, nous avons enrichi le jeu de données avec l'API OpenAlex (base de données de publications académiques) des manières suivantes:

- **identification OpenAlex**: nous avons associé chaque chercheur à son identifiant OpenAlex à l'aide d'une stratégie de *matching* basée sur le nom et l'université pour garantir la précision
- **corpus de publications**: nous avons collecté les titres de l'ensemble des travaux de chaque chercheur pour créer un vaste corpus textuel de leurs publications
- **collaboration avec la France**: une attention particulière a été portée à l'extraction des titres de publications qui présentaient explicitement un co-signature avec une institution basée en France

3. Processus d'analyse sémantique avancée

Pour aller au-delà de la simple détection par mot-clés, nous avons effectué une étape d'analyse sémantique des titres de publications. Le but était d'utiliser des outils de traitement de langage (Large Language Model - LLM) pour générer des étiquettes sémantiques fines (study_tags) qui décrivent précisément:

- les domaines/thèmes de recherche (exemples: *Art et politique, Histoire culturelle*)
- les objets d'étude concrets (artistes, lieux, institutions, groupes sociaux)
- les liens spécifiques avec la France non détectés par Regex

L'objectif de notre méthode était de structurer les thématiques pour la modélisation en réseau en associant chaque chercheur à des domaines d'expertise déduits de ses travaux.

Puisque le résultat des Regex était peu exhaustif, trop restreint et fermé (risquant de nous apporter le même problème que pour notre premier projet), on a ainsi demandé à un LLM de nous renvoyer des objets que nous ne pouvions pas identifier par Regex avec le prompt ci-dessous:

I will give you the list of titles of publications of one art historian.You ONLY see the titles. Do not assume any other information. TASK (use ONLY what can be reasonably inferred from the titles):

Your goal is to build ONE SINGLE flat list of tags (short phrases) that summarise:

1. Broad domains this scholar works on (e.g.: photography studies, museum studies, early modern art, medieval art, contemporary art, visual culture, postcolonial studies, gender studies, architecture, urban history, visual anthropology, media studies, etc.).

2. More detailed themes or problematics (e.g.: documentary photography, war photography, French colonial art, women artists,memory and heritage, Black visual culture, Mediterranean cities, city planning under empire, curatorial practices, migration and diaspora, exile, trauma and memory, etc.).

3. Concrete objects of study:

- artists or historical figures
- groups or communities (e.g. women artists, migrants, Black communities)
- places or regions (cities, countries, regions, empires)
- institutions (museums, archives, universities, cultural institutions)
- periods or dates (centuries, dynasties, regimes, clearly identified periods)
- art movements or styles (Impressionism, Surrealism, Bauhaus, Gothic, Baroque, etc.).

4. Important keywords:

- key concepts, techniques, genres, materials, critical notions
- anything that looks central to understand the scholar's research.

5. LINKS WITH FRANCE (be very explicit INSIDE THE SAME LIST):

- If the titles suggest any work on France or something French, add very precise tags to the SAME list: * e.g. "French art", "French colonialism", "Paris", "Louvre Museum", "19th-century French painting", "French photography", "Francophone Caribbean", etc.
- Include names of French artists, French movements, French institutions and French places as separate tags whenever you can identify them from the titles.

RULES:

- Be as precise and exhaustive as reasonably possible.
- DO NOT INVENT facts that are not supported by the titles.
- Prefer short, clear noun phrases (2–6 words).
- Avoid duplicates or near-duplicates.
- Return EXACTLY this JSON structure:

{{

"study_tags": ["tag 1",

"tag 2",

"tag 3"

]

}}

Here are the titles (one per line):

III. Modélisation et visualisation en réseaux

Afin de visualiser la structure des coopérations et des thématiques, nous avons modélisé notre jeu de données enrichi (chercheurs et leurs domaines/tags sémantiques) en trois différents graphes, chacun construits pour résoudre les limites du précédent et pour répondre le mieux à notre problématique. Nous les avons tous réunis sur un site web à trois pages:

Visualisations

Graphe 1. ***Visualisation des liens entre chercheurs de différentes université et thèmes de recherche*** (deuxième page du site web)

Nous avons converti notre dataset en format GEXF (Graph Exchange Format) à l'aide de la librairie NetworkX.

Notre réseau le plus simple est composé de deux types de nœuds: les "**auteurs**" (chercheurs) et les "**domaines**" (thèmes extraits par analyse sémantique). Un lien existe entre un "auteur" et un "domaine" si le chercheur étudie ce thème, nous permettant de voir les tendances de recherche des différentes universités en histoire de l'art, sans mettre en évidence quelles universités ou quels auteurs travaillent sur la France.

Graphe 2. ***Visualisation interactive du lien franco-américain entre chercheurs, universités et thèmes de recherche*** (première page du site web)

Pour une exploration plus dynamique et une meilleure mise en évidence du lien franco-américain, nous avons converti le réseau en format JSON afin de le visualiser avec la bibliothèque D3.js (Data-Driven Documents) qui permet une interaction directe avec les données.

Pour montrer la spécificité du lien franco-américain, nous avons utilisé les éléments suivants:

- **domaines france**: nœuds thématiques où au moins un chercheur est lié à la france, colorés de façon bien distincte
- **auteurs france**: chercheurs liés à au moins un domaine france, représentés sous forme de carré pour les différencier des autres auteurs

- **auteurs non-france**: chercheurs liés à aucun domaine france, représentés par des cercles de la couleur de leur université d'origine

Nous avons également choisi des **filtres interactifs** pour nous permettre d'isoler les nœuds selon l'université et les domaines France/non-France et de mettre en évidence les voisins d'un nœud sélectionné (*focus mode*).

Graphe 3. ***Visualisation interactive du lien entre chercheurs et thèmes de recherches obtenus par processus d'analyse sémantique*** (dernière page du site web)

Après avoir enrichi notre dataset à l'aide du LLM comme expliqué plus haut, on a pu obtenir une dernière visualisation pour mieux mettre en évidence les liens entre auteurs et thématiques précises obtenus avec LLM. Les filtres interactifs par université sont conservés, mais le lien franco-américain n'est pas mis en évidence comme dans le graphe 2.

Par la suite, il serait intéressant d'allier les thématiques plus précises à l'accent mis sur les liens entre auteurs et domaines liés à la france pour pouvoir conclure à partir d'un seul graphe.

IV. Résultats

L'analyse de nos différents réseaux nous donne une cartographie des spécificités et des coopérations franco-américaines dans le domaine de l'Histoire de l'art.

Les nœuds de connexion "domaines france" sur le graphe 2 révèlent les thématiques privilégiées du lien académique franco-américain, et notre visualisation nous permet de comparer la spécialisation et la diversification de la recherche selon certains thème de par la densité des liens autour des noeuds "auteurs france" et "domaines france". Le réseau permet ainsi d'observer si les connexions auteur-domaine sont concentrées dans certaines universités ou si elles se répartissent de manière transversale sur l'ensemble du campus.

Si nos réseaux ne présentent pas de thèmes particulièrement centraux, nous envisageons dans la suite de regrouper nos thèmes de manière moins précise pour pouvoir mettre en évidence des hiérarchies dans les domaines français étudiés par les chercheurs américains.

V. Conclusion

Ainsi, nos deux projets regroupés en une seule étude nous ont permis de transformer un ensemble de profils académiques hétérogènes en une base de données structurées, elle-même convertie en plusieurs réseaux interprétables. Après avoir combiné harmonisation de données, extraction de marqueurs spécifiques et analyses sémantiques, nous avons pu cartographier les affinités thématiques et les collaborations franco-américaines entre université et chercheurs en Histoire de l'Art.