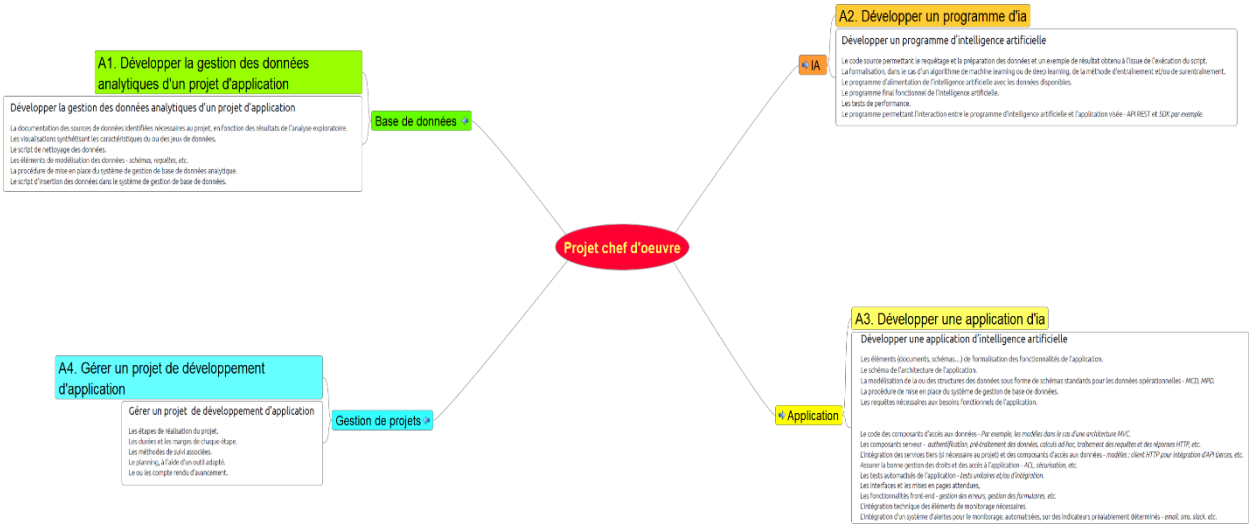
BOREL Bertrand, Ecole Microsoft IA Simplon

Rapport projet chef-d'œuvre

I- Compréhension des attentes et du travail demandé

Exemple d'une représentation graphique :



II – Elaboration d'un sujet :

Une application permettant, à partir d'un ISBN, de récupérer les informations de l'ouvrage (auteur, titre, date, maison d'édition) et offrant la possibilité d'ajouter la référence à une bibliographie déjà existante pour pouvoir par la suite l'exporter.

L'application pourrait fonctionner sur la reconnaissance de l'ISBN inscrit sous le code barre.

L'application pourrait également proposer d'autres ouvrages (NLP) voire de pouvoir les classer par thèmes (en faisant une prédiction).

III-Planification :

La difficulté d'un tel projet réside dans la quantité d'informations nécessaires au bon fonctionnement de l'application. En outre, ce n'est pas la quantité mais bien la qualité de l'information qui prend sens.

La récupération des informations d'un ouvrage grâce à son ISBN est facile à faire, notamment en python car de nombreuses bibliothèques sont disponibles. En revanche, les informations sont souvent lacunaires, voire manquantes, et il faut compter sur le fait que ses ressources exploitent essentiellement des ouvrages en anglais.

A noter qu'il ne semble pas exister de base de données des livres en langue française (en libre accès, générale et non consacrée uniquement à un domaine), tout du moins pas à ma connaissance. Par ailleurs, les données sont dispersées à travers plusieurs bases différentes.

Nous avons commencé par chercher les meilleures sources d'informations possibles, celles présentant le moins d'erreur ou de lacune.

Cette récupération d'informations s'est faite de plusieurs manières :

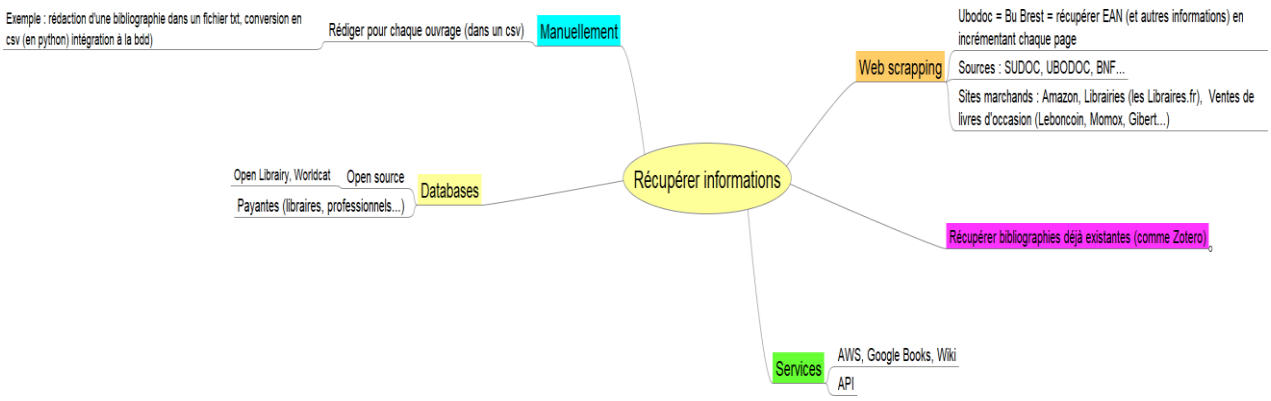
- Manuellement : il faut inscrire les références bibliographiques de chaque ouvrage à la main. Un script python permet d'accélérer le processus, en inscrivant chaque ligne dans un csv. On peut également intégrer des bibliographies déjà existantes. Ainsi un logiciel comme Zotero propose d'exporter ses documents dans un format csv. Nous avons développé un script pour intégrer ces données directement à notre database.

- Librairie python : quelques librairies proposent de récupérer des informations via l'ISBN. Nous avons utilisé pour l'instant *isbnlib* (source : https://pypi.org/project/isbnlib/). Cette bibliothèque permet notamment d'interroger plusieurs databases :
    - Open Librairy
    - Google Books
    - Wikipédia
    - Worldcat

  Nous avons opté pour Worldcat car c'est la source avec le moins d'erreurs. Cependant, en cas de lacune, nous nous tournons vers d'autres modules pour compléter les informations manquantes. Cependant, il est nécessaire d'avoir des ISBN avant de pouvoir effectuer ce travail.

- *Web scraping* : permet d'obtenir toutes les informations nécessaires pour décrire un ouvrage. Dépend cependant de la structure et du « vocabulaire » html qui rend parfois la tâche difficile. Difficulté également pour passer d'un livre à l'autre quand il n'est pas possible d'incrémenter directement l'url. C'est peut-être la meilleure solution vers laquelle se tourner car les données sont généralement fiables et le procédé est assez rapide. Nous avons développé un script pour récupérer les ISBN d'une bibliothèque, interroger des databases, récupérer les informations et les inscrire dans un csv. Ce script peut être largement amélioré, en se focalisant sur la récupération des informations d'un ouvrage, page par page, et d'inscrire le résultat dans chaque ligne d'un csv dédié.
- Enfin, certaines entreprises proposent des API pour parcourir leurs bases de données comme Amazon et Google Books. Nous ne nous sommes pas encore penchés dessus, mais certaines solutions peuvent être intéressantes (comme celle proposée par AWS)

Exemple de représentation graphique sur la récupération d'informations :



Capture du dataframe réalisé :

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1348 entries, 0 to 1347
Data columns (total 5 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   Auteur(s)  1348 non-null   object
 1   Titre      1348 non-null   object
 2   Edition    1348 non-null   object
 3   Date       1348 non-null   int64 
 4   ISBN       1348 non-null   int64 
dtypes: int64(2), object(3)
memory usage: 52.8+ KB
```

Structure et présentation du fichier :

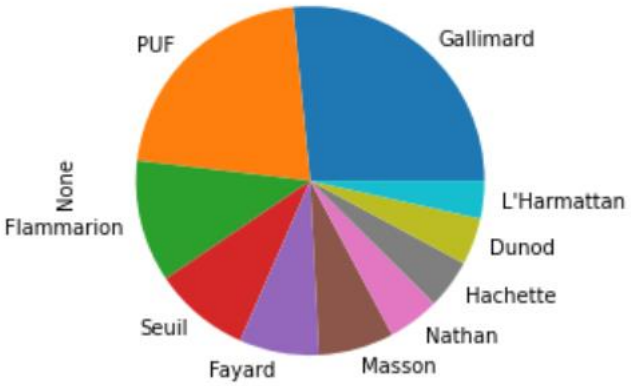
Le fichier est composé de 5 colonnes : le ou les auteurs, le titre, l'édition, la date et l'ISBN de l'ouvrage en question. Il y a pour l'instant 1348 lignes, certaines présentent quelques coquilles/erreurs mais le fichier est exploitable, notamment pour faire des graphiques dans le cadre d'une analyse de données.

Extrait du fichier :

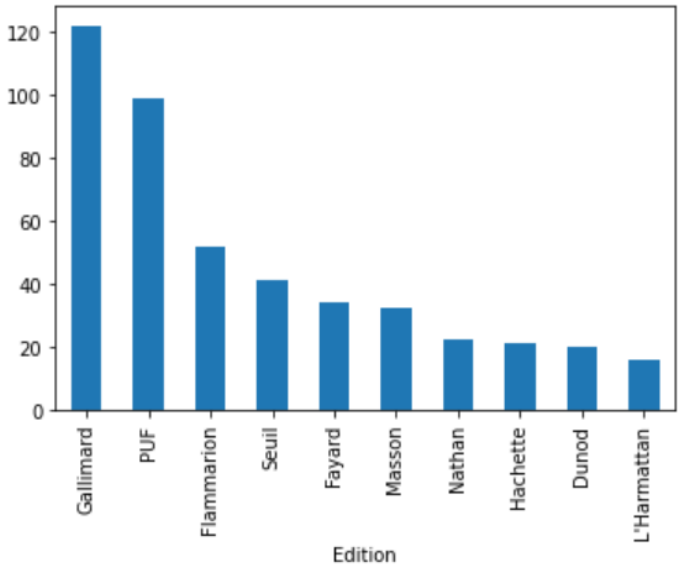
	Auteur(s)	Titre	Edition	Date	ISBN
0	Pierre Vidal-Naquet	Le choix de l'histoire - pourquoi et comment j...	Arléa	2007	9782869597624
1	Ian Kershaw	Choix fatidiques - dix décisions qui ont chang...	Seuil	2012	9782757829707
2	Jean-Claude Barreau	Sans la nation, le chaos	Editions du Toucan	2012	9782810004676
3	François Bousquet	La Droite buissonnière	Rocher	2017	9782268089898
4	Pascal Picq	Il était une fois la paléoanthropologie - Quel...	Odile Jacob	2010	9782738124944
...	...	...	...	...	...
265	Jean-Philippe Couturier	Lorsque mon patron sera une intelligence artif...	VA	2019	9782360930128
266	Luc Julia	L'intelligence artificielle n'existe pas	First	2019	9782412043400
267	Jean-Philippe Desbiolles	L'IA sera ce que tu en feras: les 10 règles d...	Dunod	2019	9782100800438
268	Daniel Crevier	A la recherche de l'intelligence artificielle	Flammarion	1997	9782080814289
269	Laurent Alexandre	La guerre des intelligences: comment l'intelli...	LGF	2019	9782253257417

Exemple de graphiques :

Les 10 maisons d'éditions qui apparaissent le plus souvent dans la base de données :



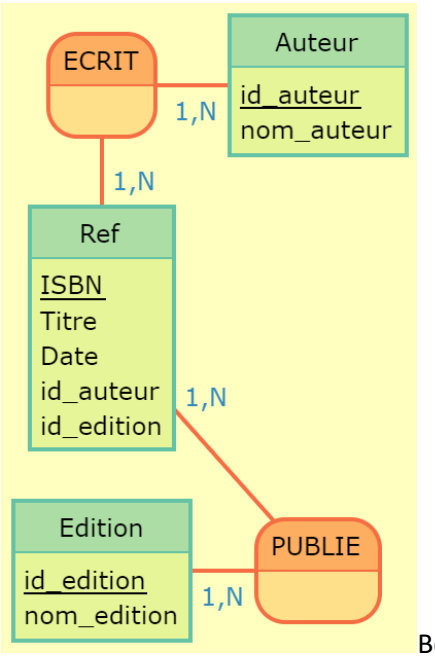
Idem, mais cette fois-ci avec un diagramme en barres :



Choix et structure de la base de données :

Sans doute du relationnel, plusieurs structures sont possibles.

Bdd 1 table



Bdd 3 tables

Structure d'un ISBN :

EXEMPLE : ISBN 978 - 2 - 7177 - 2113 - 4				
978 Livre	2 Pays francophone	7177 BnF	2113 N° de monographie	4 Clé de contrôle

Exemple d'un code-barres d'un livre :

Récupérer l'ISBN via un système de reconnaissance :

Résumé du projet :

L'objectif est de pouvoir récupérer l'ISBN d'un ouvrage en photographiant/scannant le code-barres situé au dos du livre. Une fois récupéré, on interroge la base de données pour trouver l'ISBN correspondant. Si la recherche ne donne rien, l'application se tourne alors vers d'autres services pour obtenir le meilleur résultat possible. Une fois les références obtenues, l'utilisateur peut alors les exporter pour créer une bibliographie ou les ajouter à un document existant.

Pistes de travail à améliorer dans les prochains jours :

- Améliorer la technique de *Web scraping* pour obtenir une meilleure qualité, au détriment de la quantité. Le fait de ne pas avoir à retravailler chaque donnée est un gain de temps précieux.
- Continuer à construire une base de données de qualité.
- Commencer l'interface graphique de l'application.
- Se pencher sur les systèmes de reconnaissances et sur l'IA qui pourrait être utile à la bonne avancée du projet.
- Améliorer et mieux structurer le code python (notamment en POO)

Scripts développés en python :

- Programme pour récupérer des ISBN (*Web scraping*), interroger les ressources, formater les réponses, inscrire le résultat dans un csv.
- Petits scripts pour adapter l'importation des données à partir de csv d'origines différentes (ou d'autres formats).
- Script où l'utilisateur rentre manuellement un ISBN et reçoit la réponse sous format *txt* ou *csv*.
- Idem mais l'utilisateur rentre une liste d'ISBN à partir d'un fichier *txt/csv.*
- Nombreuses fonctions pouvant être par la suite réutilisées dans l'application. Il est cependant nécessaire de revoir la POO pour une meilleure efficacité.