Ecole **N**ationale **S**upérieure
d'**I**nformatique et d'**A**nalyse
Des **S**ystèmes

Mémoire de Projet de Fin d'Année

Option

e-Management and Business Intelligence

Sujet

Data application for extracting and visualizing research papers data from AfricArXiv

Soutenu par :

ELYOUSSFI Ismail

EL-FATIH Ziad

Sous l'encadrement de :

Mme. Lamia BENHIBA

Jury:

Mme. L. BENHIBA

M. Y. Alami

Année Universitaire 2018-2019

“Si tu n’échoues pas de temps à l’autre, c’est le signe que tu ne fais rien de très innovant.” - Woody Allen.

Remerciements

Au terme de ce travail, nous tenons à remercier tout particulièrement et à témoigner toute notre reconnaissance à notre encadrante Mme. Lamia BENHIBA pour sa collaboration, son écoute, ses conseils précieux et sa disponibilité totale.

Nous adressons aussi nos plus sincères remerciements à Mr. JANATI IDRISSI Mohamed Abdo qui veille à ce que la formation e-Management et Business Intelligence se déroule dans les conditions les plus favorables.

On profite aussi pour remercier tout le cadre professoral de l'ENSIAS qui nous assures une formation qualifiante.

Enfin je voudrais faire part de ma gratitude à tous ceux qui ont participé de près ou de loin au bon déroulement de ce travail, Merci à tous.

Résumé

Le présent document est la synthèse de notre travail dans le cadre du Projet de Fin de 1ère Année. Le projet a pour finalité de mettre en place une plateforme d'extraction des données sur les documents de recherche publiées sur AfricArXiv via l'API offert par le site hébergeant OSF.

Par la manipulation des données collectées, on arrive a créér des visualisations qui donne un aperçu global sur la communaute scietifique d'AfricArXiv, et des relations entre ses éléments.

En fait, ce site web va fournir à ses utilisateurs une pagination leur permettant d'avoir un accès fluide aux différentes visualisations.

Pour une bonne gestion de projet, il a fallu en premier temps passer par la phase de formation et de compréhension, suivie par la phase d'analyse et de conception pour enfin arriver à la phase de mise en œuvre.

Mots-clés : *AfricaArXiv, API, OSF, visualisation, extraction, pagination.*

Abstract

This document is a synthesis of our work within the framework of the 1st Year End Project. The project aims to set up a Data application for extracting research papers data published on AfricArXiv via the API offered by the hosting site OSF.

By manipulating the collected data, we create visualizations that give a global overview of the website, and the relationships between its elements.

In fact, this website will provide its users with a pagination allowing them to have fluid access to the different visualizations.

For a good project management, it was necessary to go through the training and understanding phase first, followed by the analysis and design phase, and finally to the implementation phase.

Keywords: *AfricaArXiv, API, OSF, Data application,* visualization, *extraction, pagination.*

Liste des abréviations

ABRÉVIATION	DÉSIGNATION
API	Application programming interface
BD	Base de Données
CSS	Cascading Style Sheets
HTML	Hypertext Markup Language
JSON	JavaScript Object Notation
OSF	open source framework
REST	Representational State Transfer
SQL	Structured Query Language

Liste des figures

Table Des matières

Introduction générale

AfricArxiv s'est donné pour mission d'améliorer et d'ouvrir la recherche et la collaboration entre les scientifiques africains, de co-concevoir l'avenir de la communication savante. [1]

Bien que ce portail fasse partie de l'Open Science Framework (OSF), un logiciel libre et gratuit qui permet aux chercheurs de se connecter et de partager leurs travaux. Ces données sont représentées de telle sorte d'être exploitées pour utilisation personnelle, or ceci ne nous donne pas une vue globale sur le site et sur leur progression. [2]

C'est dans ce cadre que s'inscrit notre projet afin de faciliter la lecture de ces données et intégrer des nouvelles solutions dans l'open source africain afin d'atteindre les objectifs de l'Open Data.

Notre projet consiste à créer un site web qui fournit l'extraction des données sur les documents de recherche publiées sur AfricArXiv via l'API offert par le site hébergeant OSF, En manipulant ces données collectées, on arrive à créer des visualisations qui donne un aperçu global sur le site, et des relations entre ses éléments.

Le but ultime de notre application est de contribuer au développement, au soutien et à l'avancement du l'open source africain, et réaliser une collaboration scientifique entre l'ENSIAS et ce réseau d'archive.

Ce rapport est la synthèse de travail réalisé durant notre projet de fin de première année et qui s'articule autour de trois chapitres :

- ✓ Le premier chapitre présente le contexte général du projet et ses objectifs, tout en explicitant les problématiques et la démarche adoptée pour sa conduite, et la collaboration scientifique faite.
- ✓ Le second est un chapitre d'analyse et de conception qui met en place les besoins fonctionnels du système futur, suivie d'une description de la macroarchitecture du projet.
- ✓ Le dernier chapitre de notre rapport est celui de la mise en œuvre où sera réellement explicité le travail des phases qui précèdent. Ce chapitre portera sur les différentes technologies utilisées dans ce projet ainsi qu'une illustration par quelques écrans de notre plateforme.

Chapitre 1

1. Contexte générale du projet

Ce chapitre débute par une explication de la problématique, de la motivation et des objectifs du projet. Il traite ensuite les généralités du projet, en détaillant la collaboration scientifique, ainsi que la démarche suivie et la planification du projet.

1.1 Présentation du Projet

La science ouverte est de plus en plus populaire dans le monde et offre des opportunités sans précédent aux scientifiques en Afrique, en Asie du Sud-Est et en Amérique latine. Les scientifiques africains font face à plusieurs difficultés lorsqu'ils tentent de faire publier leurs travaux dans des revues à comité de lecture - il existe un petit nombre de plateformes de publication, un manque de connaissances et des difficultés d'accès liées aux revues existantes (dont la visibilité sur le Web n'est pas très bonne). Les méthodes et techniques (y compris le processus d'évaluation par les pairs) qui sont mises au point pour sa diffusion ne sont pas nécessairement adaptées au contexte d'autres régions du monde, dont l'Afrique. En effet, de nombreuses revues savantes africaines évaluées par des pairs ne sont pas en mesure d'héberger leur contenu en ligne en raison des ressources limitées et de la fracture numérique. [3]

L'open data, que l'on traduit par « données ouvertes », est une pratique de publication sous licence ouverte qui garantit un accès libre aux données et autorise leur réutilisation sans conditions techniques, juridiques ou financières. Ces données offrent de nombreuses opportunités pour étendre le savoir et créer de nouveaux produits et services de qualité. [4]

Toutefois, la présentation des données offertes par le site web n'est pas toujours exploitable. De plus, les données ne sont pas bien présentées et ne permettent pas une bonne compréhension et connaissance du métadonnées.

La data extraction suivie de la data visualisation apparaît comme un excellent moyen d'analyser les données, de suivre le Progress, et de comprendre les différentes relations entre les éléments du site web.
L'extraction des données depuis AfricArXiv est faite par exploitation de l'API fournit par la plateforme mère OSF,

La visualisation de données permet de synthétiser les informations que contiennent ces publications pour mettre en évidence les informations clés qu'ils renferment. En fait, Chaque nouvelle visualisation est susceptible de nous apporter des informations sur nos données. Certaines sont peut-être déjà existantes, alors que d'autres peuvent être complétement nouvelles, voire surprenantes. Pour dire les choses en peu de mots, la visualisation de données c'est de les mieux faire parler. Ainsi fût née notre idée de « Data application for Extracting and Visualizing Research Papers Data From AfricArXiv. », qui a pour but de rassembler les données et de les transformer en graphes significatifs et plus compréhensible.

1. 2 Objectifs

Le projet vise à mettre en place un site web qui offre aux utilisateurs :

- Une vue générale sur les métadonnées
- Accès au données d'AfricArXiv de manière ordonnée par stockage sur notre propre API
- Des diagramme (charts) interactifs pour visualiser les données

1.3 La collaboration scientifique

Lors de notre initialisation sur ce projet, l'idée nous a venu de contacter les développeurs du site, pour en savoir plus sur ses objectifs à long terme qu'il n'est d'autre que l'amélioration et l'ouverture de la recherche et de la collaboration entre les scientifiques africains et à concevoir l'avenir de la communication scientifique.

Une proposition pour une collaboration scientifique entre l'ENSIAS et AfricArXiv surgit de leur part.

1.4 Conduite du projet

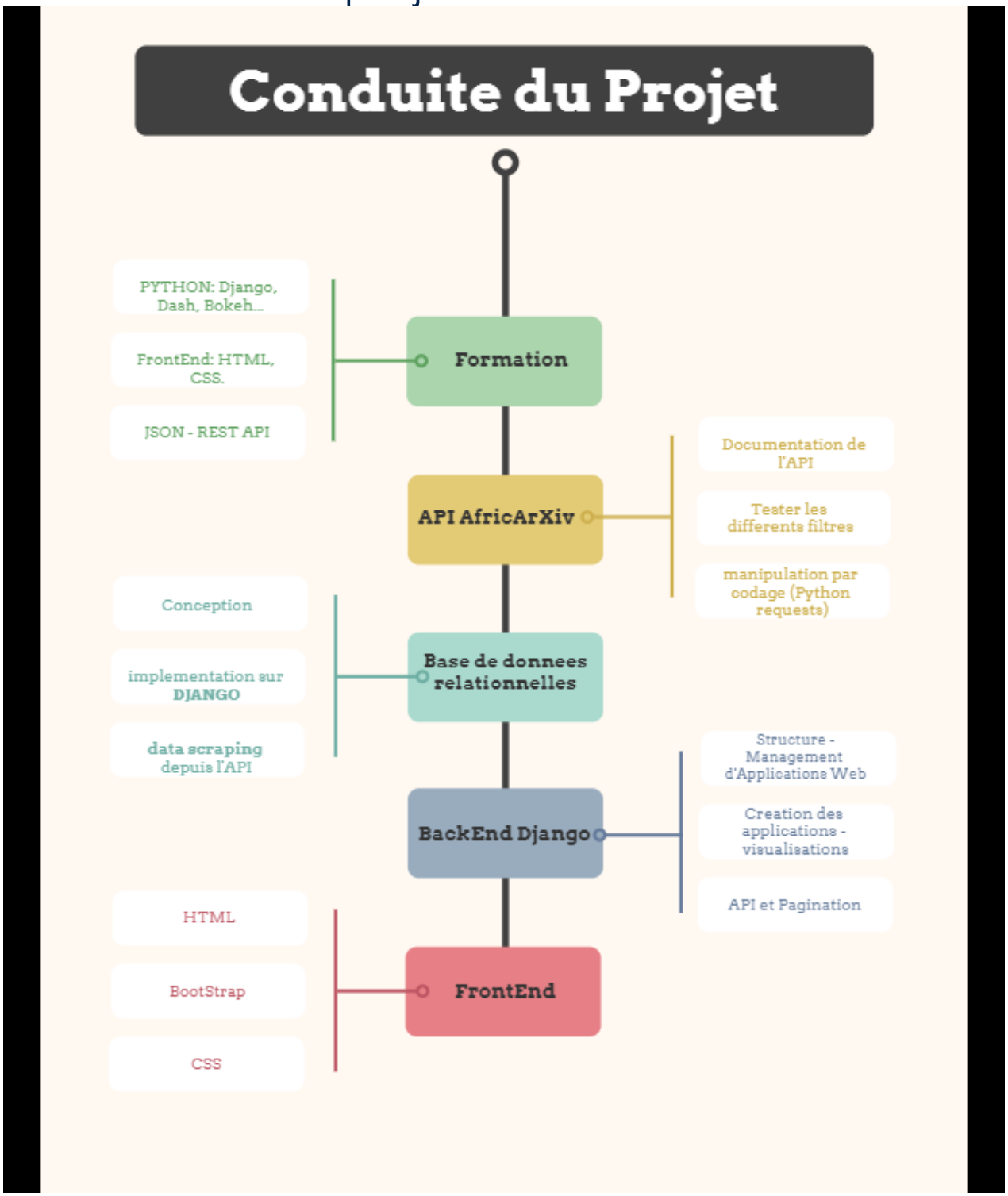


Figure 1 Conduite de projet

Chapitre 2

2. Analyse et Conception

Ce chapitre sera consacré à la spécification des besoins fonctionnels et opérationnels du système à réaliser et de présenter la macro architecture du projet.

2.1. Cartographie fonctionnelle

L'objectif de notre projet est de créer des différentes visualisations interactives qui représente un ensemble de données.

Ainsi, pour atteindre notre but nous serons amenées à suivre les étapes ci-dessous :

Figure 2 Cartographie fonctionnelle

2.2. Exploitation de l'API et acquisition des données

2.2.1. Exploitation de l'API

L'API d'OSF nous offre de diverses données qui ne s'intéresse pas nécessairement sur notre site objectif AfricArXiv. En manipulant la documentation de cet API (https://developer.osf.io/#section/https:api.osf.iov2), et en maitrisant les diffèrent filtres offerts, on a pu naviguer sur l'API (https://api.osf.io/v2/) et atteindre un « *sous-API*» (https://api.osf.io/v2/providers/preprints/africarxiv/?format=json) qui affiche les métadonnées du site.

2.2.2. Acquisition des données

Grace au métadonnées obtenues sous forme JSON, et en analysant ces données, on arrive à créer un modèle MCD, qu'on prévoit stocker sous forme d'objets sur notre base de données, en respectant les différentes contraintes d'intégrité et les associations entre ses entités.

2.3. Base de données relationnelles

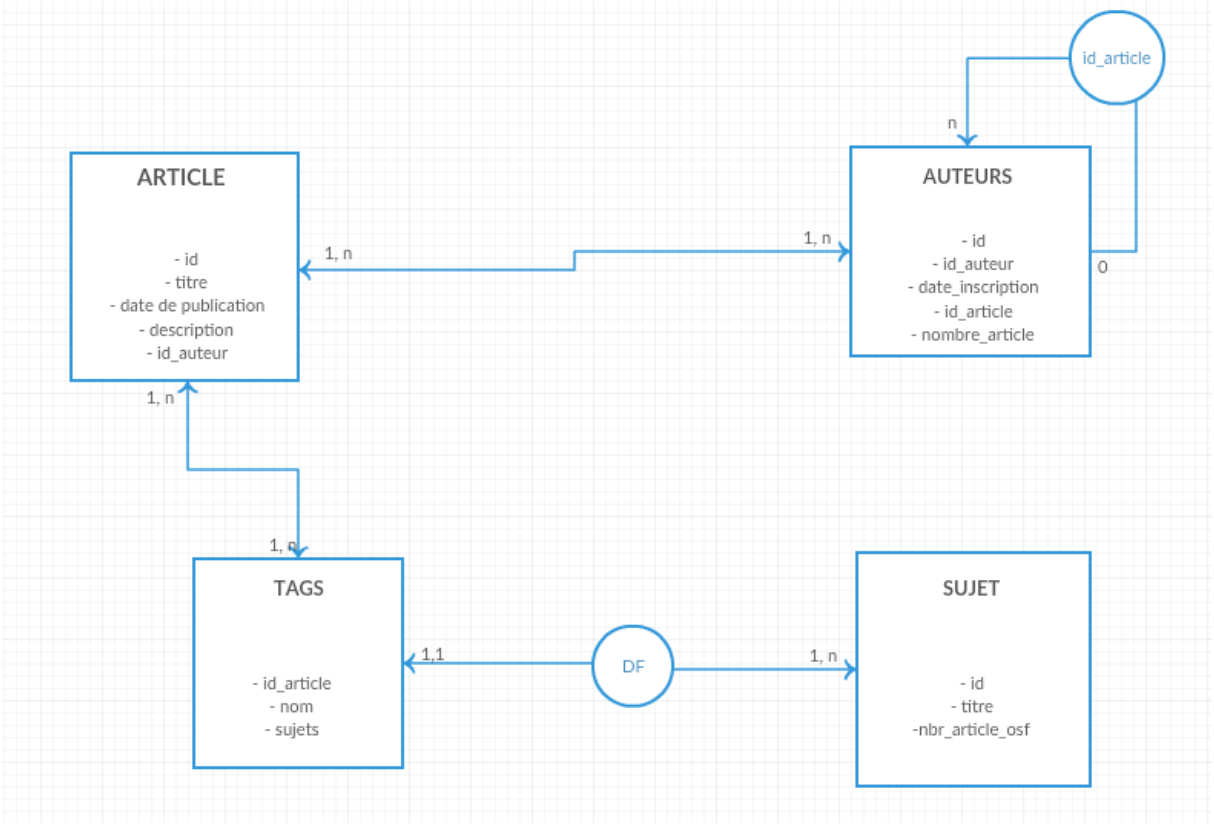


Figure 3: modélisation de la base de données relationnelles avec diagramme uml

2.4. Visualisation des données

Dans cette phase, on a des données bien structurées dans la BD et on peut les transformer en matière visuelle.

La visualisation des données se fait via une plateforme web. La plateforme se compose de plusieurs pages décrites comme suit :

Une page d'accueil qui re directe vers les différentes visualisations :

- Le nombre des tags par chaque sujet.
- Le nombre des articles contribués par mois.
- '' Top 5 auteurs '' montrant les cinq auteurs avec le plus grand nombre de contributions, et récemment actif.
- La domination des article d'AfricArXiv par rapport au site hébergeant OSF.

2.5. Conclusion

Ce chapitre avait pour objectif la description des étapes d'analyse et de conception, étapes cruciales pour le bon déroulement du projet. Une cartographie fonctionnelle a été présentée délinéant les étapes du projet : extraction des données en exploitant l'API fournit par l'OSF, suivie de la conception de la BD, ensuite la visualisation des données. Le chapitre suivant va concerner la phase de la mise en œuvre de notre site web.

Chapitre 3

3. Mise en œuvre

Ce dernier chapitre sera consacré à la description de l'architecture technique, des outils et technologies utilisés pour le développement de la plateforme et à la spécification des détails de l'implémentation de la solution conçue.

3.1 Architecture technique

Cette partie correspond à l'implémentation technique de la cartographie fonctionnelle du projet mentionnée dans le chapitre précèdent. (Figure 2)

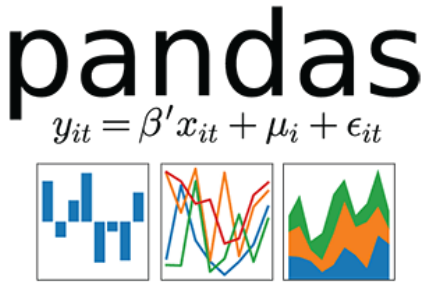
Figure 4: implémentation technique [5] [6] [7] [8] [9]

3.2 Description des outils et technologies

3.2.1 Extraction et Traitement des données

Python est un langage de programmation objet, multiparadigme et multiplateformes. Python est un flexible, puissant et facile à utiliser. Il possède de puissantes bibliothèques pour la manipulation et l'analyse des données. Son unicité provient du fait qu'il est assez fort pour attaquer les problèmes les plus difficiles dans pratiquement tous les domaines, tout en étant facile à utiliser pour le calcul analytique et quantitatif. C'est un langage

moderne, dont l'agilité et la productivité des solutions est légendaire.
Les bibliothèque REQUESTS et JSON sont utilisée très souvent dans notre projet [5].



Pandas est une bibliothèque écrite pour le langage de programmation Python permettant la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles. Les principales structures de données sont les séries (pour stocker des données selon une dimension - grandeur en fonction d'un index), les DataFrames (pour stocker des données selon 2 dimensions - lignes et colonnes), les Panels (pour représenter des données selon 3 dimensions, les Panels4D ou les DataFrames avec des indexes hiérarchiques aussi nommés MultiIndex (pour représenter des données selon plus de 3 dimensions - hypercube). Alignement intelligent des données, déformation/pivotement de données flexibles et haute performance de fusion des ensembles de données. Python muni de pandas est utilisé dans une grande variété de domaines, y compris la finance, les neurosciences, l'économie, les statistiques, la publicité, les Web Analytics, et encore plus [6].

3.2.2 Back End

Attribué le slogan suivant : « Le *Framework web pour les perfectionnistes sous pression* «. Django est donc clairement orienté pour les développeurs ayant comme besoin de produire un projet solide rapidement et sans surprise_... c'est à dire à tous les développeurs ! Comme il est toujours compliqué de partir de rien, Django propose une base de projet

solide. Django est donc une belle boîte à outils qui aide et oriente le développeur dans la construction de ses projets. [7]

SQLite est une bibliothèque écrite en langage C qui propose un moteur de base de données relationnelle accessible par le langage SQL. Contrairement aux serveurs de bases de données traditionnels, comme MySQL ou PostgreSQL, sa particularité est de ne pas reproduire le schéma habituel client-serveur mais d'être directement intégrée aux programmes. L'intégralité de la base de données (déclarations, tables, index et données) est stockée dans un fichier indépendant de la plateforme. [8]

3.2.3 Front End

HTML est l'abréviation de HyperText Markup Language, soit en français « langage de balisage hypertexte ». Grâce au HTML, on va par exemple pouvoir indiquer au navigateur que tel texte doit être considéré comme un simple paragraphe ou que tel autre est un titre. Le HTML va également nous permettre d'insérer différents types d'éléments dans nos pages web : du texte, des liens, des images, etc.

CSS est le diminutif de Cascading StyleSheets, ou feuilles de styles en cascade. Le CSS va nous permettre par exemple de définir la taille, la couleur ou l'alignement d'un texte. Nous allons donc utiliser le CSS sur notre code HTML, afin d'enjoliver le résultat visuel final. [9]

Chart.js est une bibliothèque open-source maintenue par la communauté (elle est disponible sur GitHub) qui vous aide à visualiser facilement les données en utilisant JavaScript. Il prend en charge 8 types de graphiques différents (y compris les barres, les lignes et les tartes), et

ils sont tous réactifs. En d'autres termes, vous configurez votre graphique une fois pour toutes, et Chart.js s'occupera de soulever les objets lourds pour vous et s'assurera qu'ils sont toujours lisibles (par exemple en enlevant quelques détails non critiques si le graphique devient plus petit). [10]

Dash est un Framework Python pour la création d'applications web. Il s'intègre dessus de Flask, Plotly.js, React et React Js. Il vous permet de construire des tableaux de bord en utilisant Python pur. Dash est open source, et ses applications s'exécutent sur le navigateur web. [11]

3.3 Mise en place de la '' Data Application ''

3.3.1 Extraction et conception du base de données

Apres obtention des données souhaitées, on les ajoute à notre propre base de données.

Ceci est fait par un algorithme de stockage qu'on a mis en œuvre, et qui effectue les tâches suivantes :

- Extraction de données depuis l'API – Data Munging : En utilisant la bibliothèque Requests en python, ainsi que la bibliothèque JSON, on arrive à extraire les différentes données depuis l'API d'OSF, sous forme de JSON.

```
#filling_the_Preprints & Authors & Tags
url='https://api.osf.io/v2/providers/preprints/africarxiv/preprints/?format=json'
dict = requests.get(url).json()
```

Figure 5: extraction depuis l'API d'OSF

- En cas d'erreurs rencontrées lors de la connexion, on refait l'opération de connexion au plus trois fois, une fois ce nombre est dépassé, on déclenche notre propre erreur.

- Respecter le model conceptuel énoncé auparavant (figure 3) : cette opération implique le traitement des erreurs d'intégrité et de redondance de données.

- Stockage sur notre propre base de données, ce stockage est réalisé par la BD connectée à Django (SQLite)
- Sérialisation et création de l'API : dans cette phase on a sérialisé notre Object afin de rendre capable d'être déchiffrées et lu par n'importe quel système. Ainsi les objets issus de notre base de données sont sauvegardés et affichées (GET) sur notre propre API.

Ici un exemple de l'un des résultats obtenus :

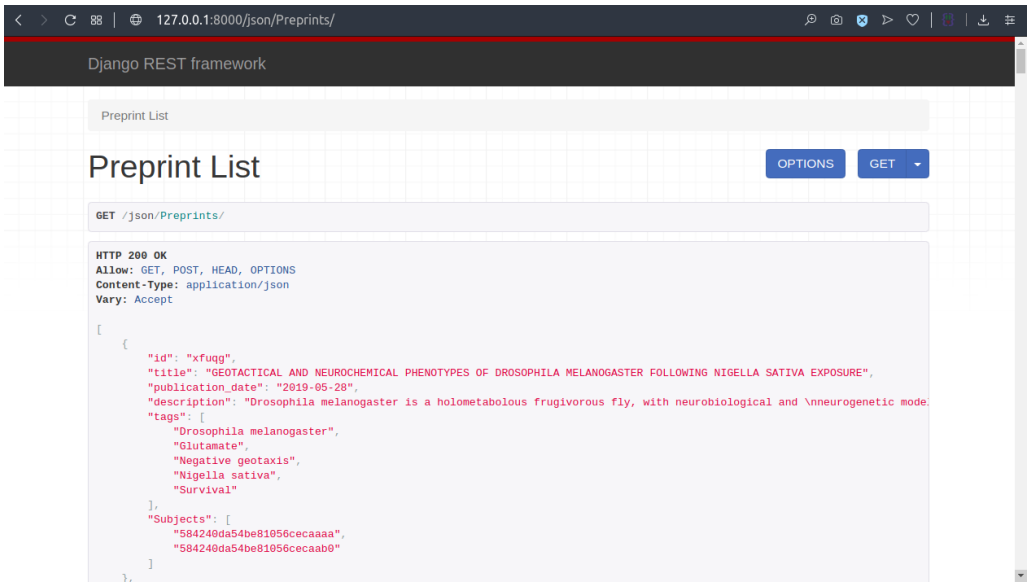


Figure 6 : liste des preprints

3.3.2 Visualisation

Après avoir préparé les données dans les étapes précédentes, dans cette phase on les transforme en graphes significatifs. Ensuite, on les met sur notre plateforme interactive.

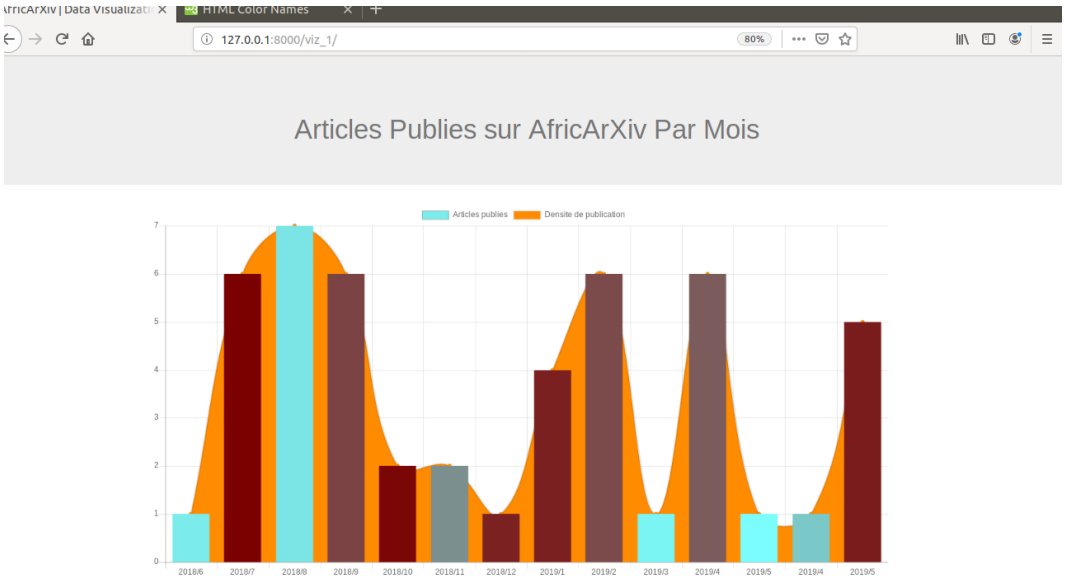
Quand l'utilisateur ouvre le site web, il est dirigé vers la page d'accueil qui illustre chaque visualisation traitée. Tout cela est bien présenté dans les captures d'écran ci-dessous :

Figure 7: page d'acceuil

Si l'utilisateur choisit la rubrique '' articles par Sujet '', il va se trouver face à ce Diagram (chart) qui illustre le nombre des articles par sujet.

En choisissant la rubrique '' articles publiés par domaine '', l'utilisateur va se trouver face à un Diagram interactif qui illustre le nombre des articles contribués au total sur AfricArXiv, par comparaison de ceux de l'OSF, puisque chaque colonne représente un domaine spécifique, par faire passer le curseur sur l'un des colonnes, le nombre d'article par ce domaine-là s'affiche à l'écran.

Figure 8: visualisation d'articles par mois



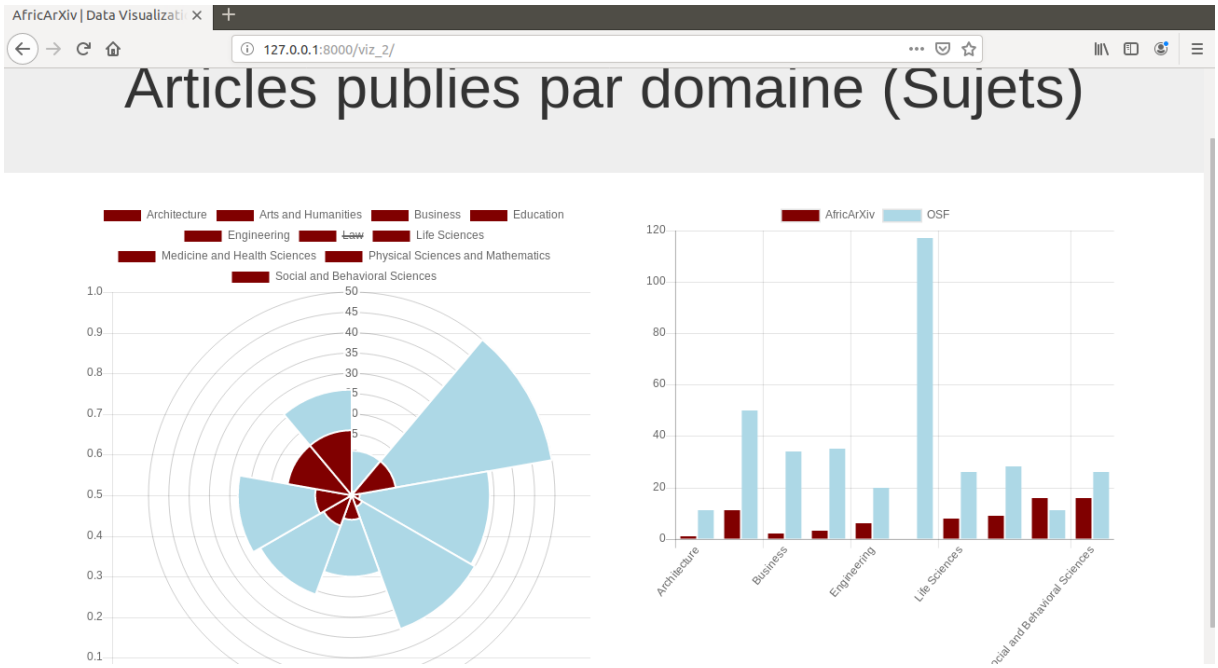


Figure 9: visualisations ; Articles publiés par domaine

La rubrique '' nombre de mots clés employées par domaine '', permet à l'utilisateur d'accéder à une visualisation interactive qui illustre le nombre des mots clés par domaine (sujet).

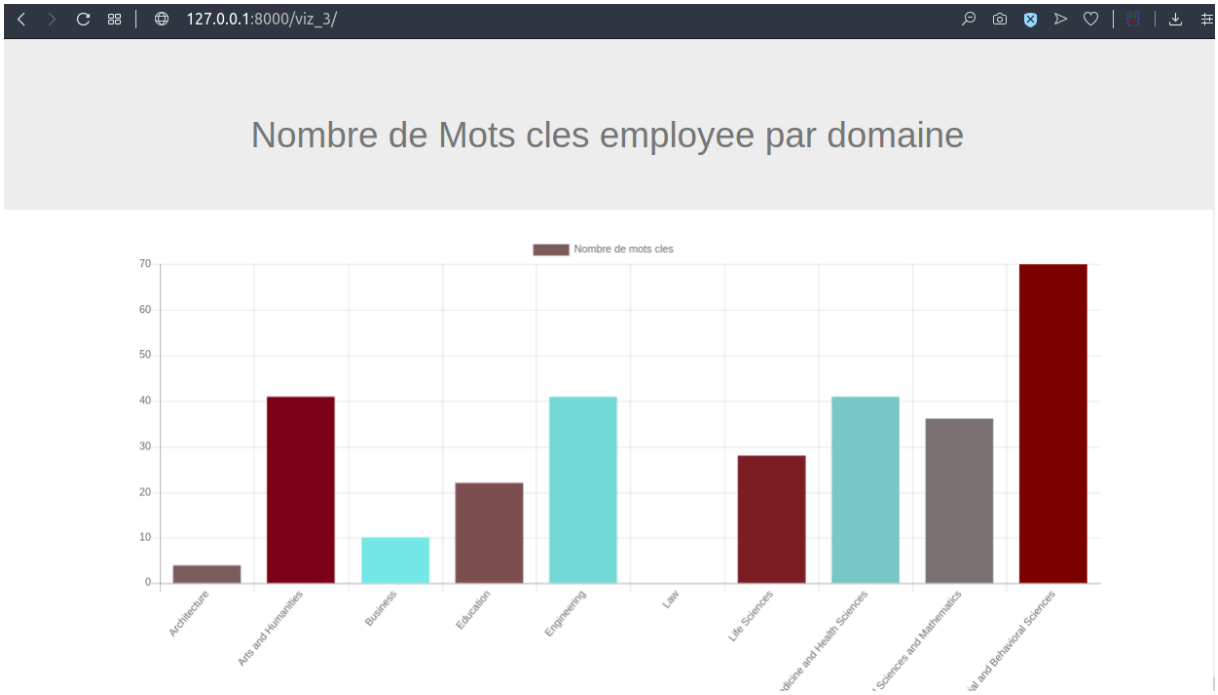


Figure 10: Nombre de mots clé par domaine

La rubrique « zone source de contribution » nous permet à l'utilisateur d'accéder à une visualisation interactive qui illustre l'effectif du contribution par pays.

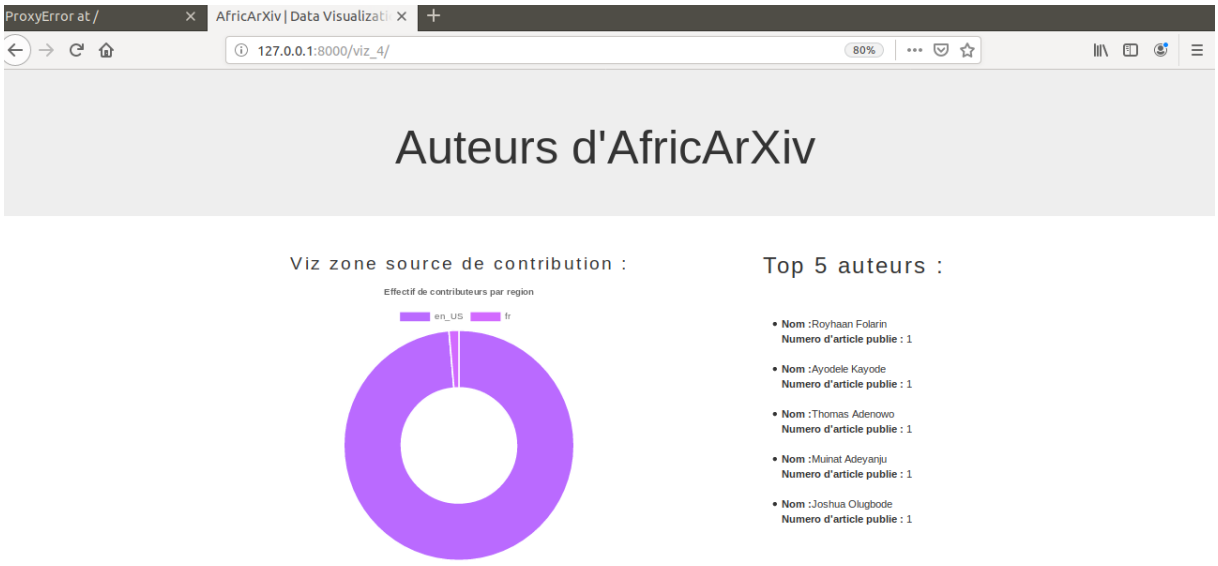


Figure 11: zone de contribution

La rubrique « Réseau d'auteurs» nous permet à l'utilisateur d'accéder à une visualisation interactive qui illustre les collaborations établis par les différents auteurs..

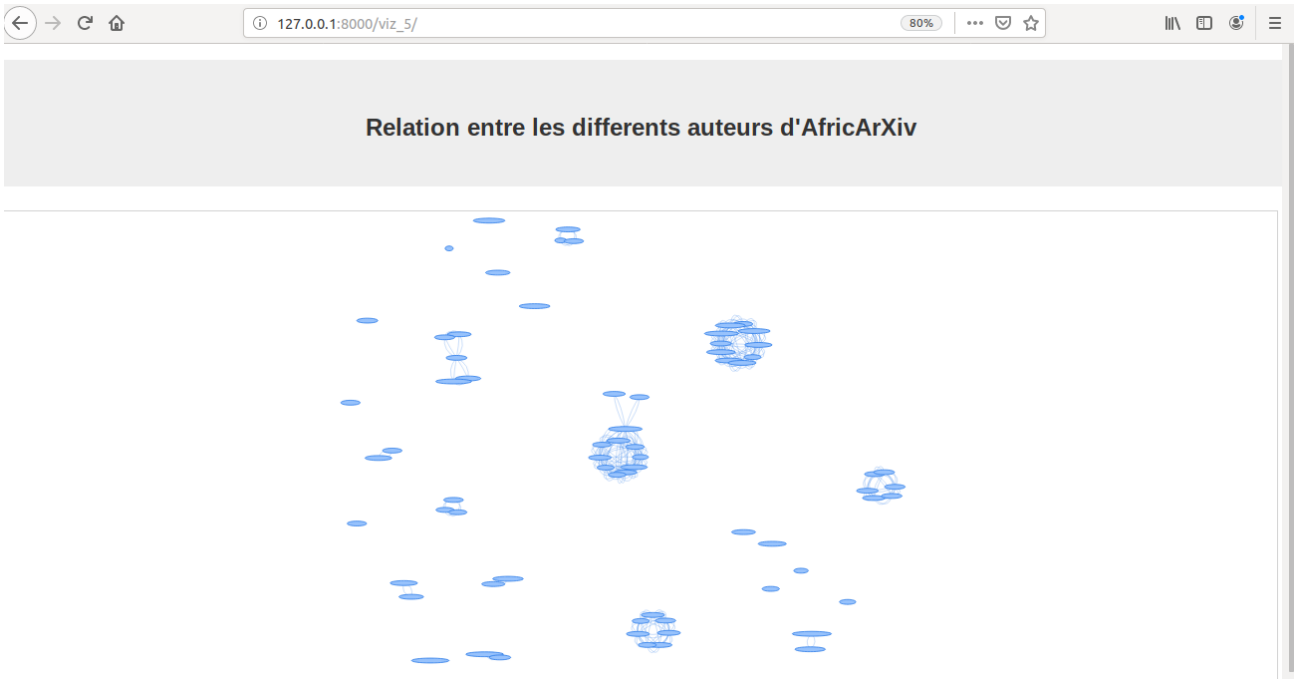


Figure 12 Réseau d'auteurs

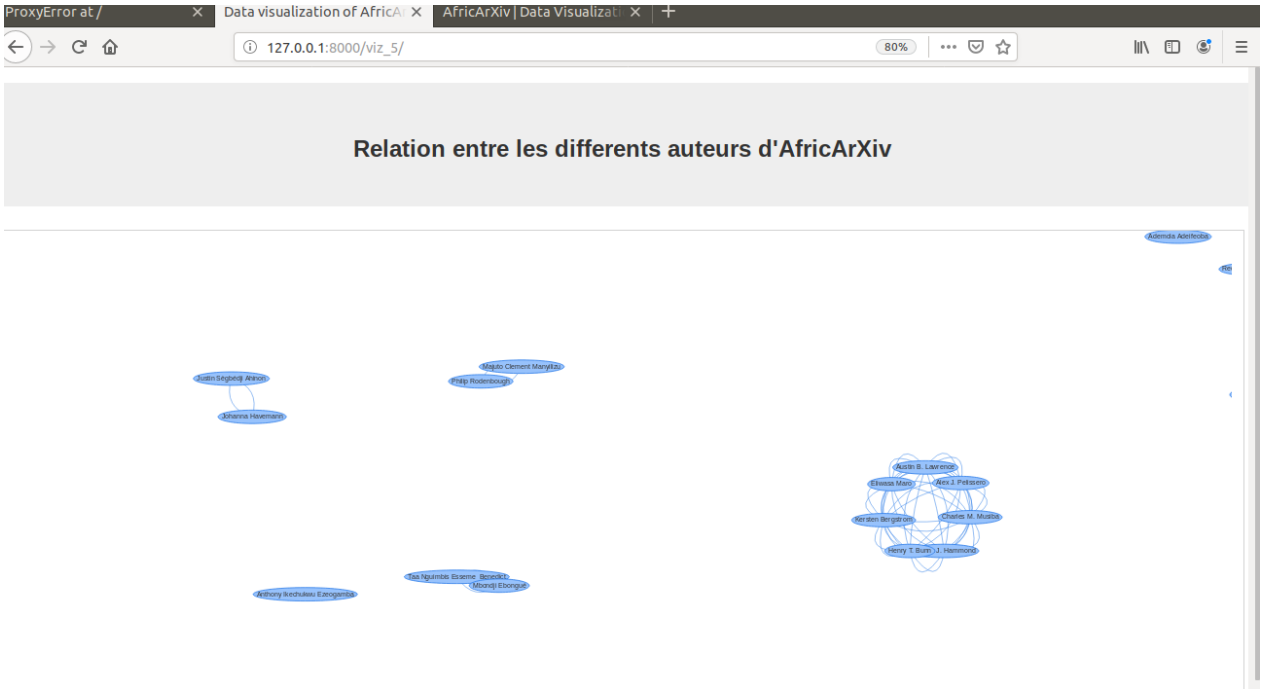


Figure 13: Réseau d'auteurs (zoom)

3.4 CONCLUSION

Dans ce chapitre on a décrit les outils utilisés pour la réalisation du site web. Ensuite, on a montré le résultat final attendu de notre projet.

Conclusion générale

L'objectif de ce projet de fin de première année était de créer une plateforme interactive pour la visualisation des données gouvernementales. Il s'agissait de mettre en place des visualisations concernant différents aspects du site web AfricArXiv et qui implémentent des graphes dynamiques illustrant les différents indicateurs de ce site.

Pour se faire, nous avons passé par une phase de formation pour bien se documenter, puis nous avons entamé la phase d'analyse et de conception qui consiste à identifier les besoins fonctionnels, ensuite nous avons passé à la phase de conception où nous avons présenté notre architecture technologique et nous l'avons décrit, finalement on a exposé notre produit final à l'aide de quelques captures d'écran.

Comme nous avons eu l'opportunité de mettre en œuvre les différentes connaissances acquises durant notre première année à l'Ecole Nationale Supérieure d'informatique et d'analyse des Systèmes. Nous avons également pu approfondir nos compétences en informatique décisionnelle.

Par ailleurs, nous avons pu raffiner nos capacités techniques et théoriques, et développer notre esprit de recherche et d'auto-formation. En outre, ce projet était l'occasion pour nous d'améliorer notre méthodologie de travail et développer notre esprit d'équipe.

En guise de perspective, le site est récemment créé, ce qui nous offre une pauvre base de données, ceci impacte sur le nombre de visualisations possible. Cette contrainte nous empêche à utiliser du Machine Learning - Apprentissage supervisé, pour prévoir le taux de contribution par domaine au futur.

Webographie

[1] AfricArXiv, le dépôt de pré-impression pour la recherche africaine. https://info.africarxiv.org

[2] AfricArXiv, Open Science in Africa, Published by SEGBEDJI on 21st November 2018, https://info.africarxiv.org/open-science-in-africa/

[3] AfricArXiv , African scientists launch their own preprint server, Published by SEGBEDJI on 25th June 2018, https://info.africarxiv.org/africarxiv-launch/

[4] Wikipedia: Open Data, https://fr.wikipedia.org/wiki/Open_data

[5] Wikipedia, Python https://fr.wikipedia.org/wiki/Python_(langage)

[6] Wikipedia, Pandas https://fr.wikipedia.org/wiki/Pandas

[7] fullstackpython, Djangohttps://www.fullstackpython.com/django.html

[8] SQLite, https://www.sqlite.org/about.html

[9] https://www.pierre-giraud.com/html-css/cours-complet/html-css-definition-role.php

[10] Medium, https://medium.com/javascript-in-plain-english/exploring-chart-js-e3ba70b07aa4

[11] Datacamp, https://www.datacamp.com/community/tutorials/learn-build-dash-python