



UNIVERSITE MOHAMMED V
FACULTE DES SCIENCES RABAT
DÉPARTEMENT D'INFORMATIQUE

Année universitaire 2024-2025



RAPPORT De Projet de Fin de module

Dans le cadre du module

Intégration de données et Business Intelligence
Master : Intelligent Processing Systems

Par :

- EL MAHRAOUI AMAL

Titre

Intégration de Données Bibliographiques dans un Système Documentaire
PMB

Pr. AMMARI MOHAMMED

Encadrant

Abstract

Ce projet vise la mise en œuvre d'un processus complet d'intégration de données multilingues (français et arabe) en suivant l'approche ETL (Extract, Transform, Load). À partir de plusieurs jeux de données hétérogènes, les informations sont extraites, nettoyées, transformées et intégrées dans une base de données relationnelle SQL, conçue pour assurer la cohérence, l'accessibilité et la performance des requêtes analytiques. Cette base de données alimente ensuite des tableaux de bord interactifs développés sous Power BI, permettant une visualisation intuitive et pertinente des indicateurs clés. Enfin, les données consolidées sont exportées et intégrées dans la plateforme phpMyBibli (PMB) à l'aide de scripts PHP, afin de faciliter leur exploitation dans un environnement documentaire. Ce projet met en évidence la maîtrise du cycle de vie des données, la gestion des langues à double encodage, et l'interopérabilité entre des technologies variées, dans une optique d'aide à la décision et de valorisation de l'information.

Mots-clés : Intégration de données, ETL, SQL, Power BI, PHP, phpMyBibli (PMB).

Table des matières

1 Fondements	9
1.1 Introduction à l'intégration de données	9
1.1.1 Définition et objectifs	9
1.1.2 Approche ETL	10
1.1.3 Défis de l'intégration des données	12
1.2 Introduction à la Business Intelligence (BI)	13
1.2.1 Définition	13
1.2.2 Les composants clés de la BI	14
1.2.3 Processus typique de la BI	16
1.3 Cadrage du projet	18
1.3.1 Contexte général et Problématique	18
1.3.2 Objectif du projet	18
1.3.3 Solution proposée	18
2 Méthodologie	20
2.1 Analyse et préparation des données	20
2.1.1 Description des datasets	20
2.1.2 Nettoyage des données	21
2.1.3 Transformation des données	23
2.1.4 Normalisation des données	23
2.2 Visualisation des données	25
2.2.1 Présentation de Power BI	25
2.2.2 Visualisation des données avec Power BI	25

2.3	Intégration dans la base de données PMB	25
2.3.1	Présentation de PMB (PhpMyBibli)	25
2.3.2	Intégration de données	26
2.3.3	Réindexation de la base de données	26
	Bibliography	28

Table des figures

1.1	Schéma Extract, Transform, Load (ETL)	10
1.2	Différents formats de sources de données	11
1.3	Processus OLAP	15
1.4	Rotate-cube	15
1.5	Enter Caption	17
1.6	Solution proposée	19
2.1	Extrait du fichier buf	20
2.2	Extrait du fichier bua	21
2.3	avant et après correction des erreurs typographiques	23
2.4	Extrait du tableau : Reformatage de l'inventaire avec ajout du préfixe FSR	23
2.5	Les données en français après la classification Dewey	24
2.6	Les données en arabe après la classification Dewey	24

Liste des tableaux

2.1	Nombre de doublons détectés avant et après nettoyage	21
2.2	Nombre de valeurs manquantes (NaN) par colonne avant et après nettoyage	22
2.3	Classification Décimale Dewey	24

Introduction

DANS un monde en constante évolution numérique, les organisations produisent une quantité exponentielle de données. Ces données, souvent multilingues et hétérogènes, nécessitent une intégration structurée et efficace pour être exploitées de manière optimale. Dans ce contexte, la mise en œuvre d'un processus d'intégration de données devient essentielle pour consolider l'information, la rendre accessible et en tirer des connaissances stratégiques (Kimball & Ross, 2011).

Le présent projet a pour objectif de traiter et d'intégrer un ensemble de jeux de données multilingues (en français et en arabe) en suivant une démarche d'ETL (Extract, Transform, Load). Ce processus vise à extraire les données de différentes sources, à les transformer pour garantir leur cohérence, puis à les charger dans une base de données relationnelle SQL (Microsoft, 2023). L'intégration de données multilingues comporte plusieurs défis techniques, notamment la gestion des encodages, des structures syntaxiques et des formats de données (Mohamed & Alhassan, 2020).

Une fois les données intégrées dans la base SQL, elles sont connectées à l'outil de visualisation Power BI pour produire des tableaux de bord interactifs permettant une meilleure interprétation des données (Microsoft, 2023). Ces visualisations offrent des indicateurs clés et facilitent la prise de décision dans un environnement opérationnel. Enfin, les données sont exportées vers le système documentaire phpMyBibli (PMB) via des scripts PHP, afin d'en permettre la consultation au sein d'une plateforme de gestion de bibliothèque (Services, 2021).

La méthodologie adoptée dans ce projet repose sur une succession d'étapes logiques, allant de la compréhension des sources de données jusqu'à leur exploitation finale. Cette approche permet non seulement d'automatiser une partie du traitement,

mais aussi d'assurer la qualité et la cohérence de l'information tout au long du cycle de vie des données.

Ce rapport présente les différentes étapes du projet, en commençant par un cadrage théorique et une revue des technologies utilisées. Il enchaîne avec une présentation des jeux de données traités, suivie de la conception du processus ETL et de la base de données. Ensuite, la visualisation avec Power BI et l'intégration via PMB sont décrites. Le document se termine par une synthèse des résultats, une discussion critique ainsi qu'une conclusion exposant les limites et les perspectives du projet.

Chapitre 1

Fondaments

1.1 Introduction à l'intégration de données

1.1.1 Définition et objectifs

Depuis que les systèmes de gestion ont commencé à collecter des données la combinaison des différentes sources de données pose un problème [Talend (2025)]. Ce n'est qu'au début des années 1980 que les informaticiens ont commencé à concevoir des systèmes prenant en charge l'interopérabilité des bases de données hétérogènes ou différentes.

L'intégration des données permet de combiner des données provenant de différentes sources dans une vue unifiée : de l'importation au nettoyage en passant par le mapping et la transformation dans un gisement cible, pour finalement rendre les données plus exploitables et plus utiles pour les utilisateurs qui les consultent [SAP (2025)]. Les entreprises sont en train de mettre en place des initiatives d'intégration de leurs données pour les analyser et les exploiter plus efficacement.

L'un des premiers systèmes d'intégration des données a été développé par l'Université du Minnesota en 1991 [Talend (2025)] : son objectif était d'assurer l'interopérabilité des milliers de bases de données sur la population. Le système utilisait une approche d'entreposage de données permettant d'extraire, de transformer et de charger des données de sources disparates dans un schéma de visualisation pour les rendre compatibles.

1.1.2 Approche ETL

1.1.2.1 Définition et Origine

ETL (Extract, Transform, Load) est un processus [Voir la figure 1.1] d'intégration de données qui combine des données provenant de diverses sources en un ensemble de données unifié et cohérent. Ce processus consiste à extraire des données brutes de différents systèmes, à les transformer pour répondre à des exigences spécifiques, puis à les charger dans une base de données cible, un entrepôt de données ou un data lake house [OVHcloud (2025)]. Il en résulte un référentiel centralisé de données structurées et de haute qualité, prêtes à être analysées.

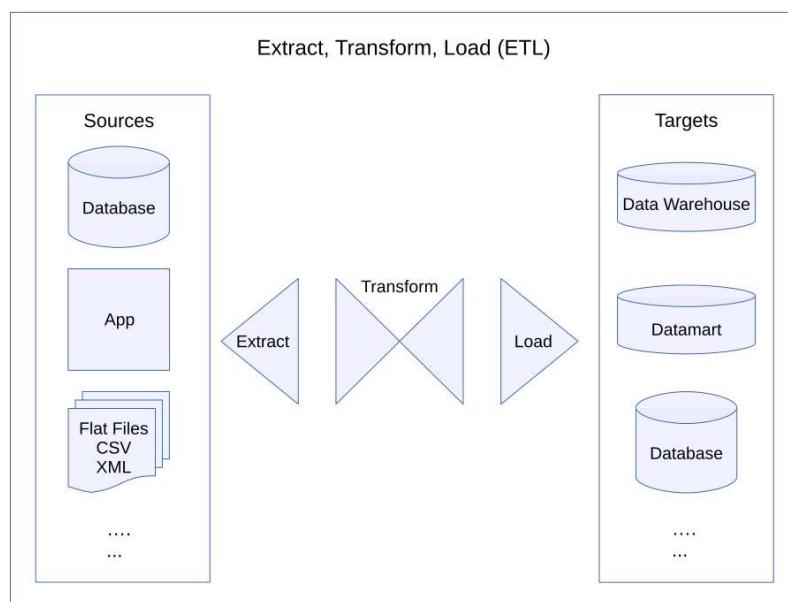


FIGURE 1.1 – Schéma Extract, Transform, Load (ETL)

Source : Wikipédia

L'ETL a gagné en popularité dans les années 1970 lorsque les organisations ont commencé à utiliser plusieurs référentiels de données, ou bases de données, pour stocker différents types d'informations commerciales. La nécessité d'intégrer les données réparties dans ces bases s'est rapidement imposée. L'ETL est devenu la méthode standard pour prendre des données de sources disparates et les transformer avant de les charger dans une source cible, ou destination.

À la fin des années 1980 et au début des années 1990 [SAS (2025)], les data

warehouses ont fait leur apparition. Type distinct de base de données, les data warehouses offrent un accès intégré aux données provenant de systèmes multiples – ordinateurs mainframe, mini-ordinateurs, ordinateurs personnels et feuilles de calcul. Mais souvent, différentes organisations choisissent des outils ETL différents utilisables avec des data warehouses différents. Avec les fusions et les acquisitions, de nombreuses entreprises se sont retrouvées avec plusieurs solutions ETL différentes qui n'étaient pas intégrées.

Au fil du temps, le nombre de formats, de sources et de systèmes de données [Voir la figure 1.2] s'est considérablement accru. Aujourd'hui, l'extraction, la transformation et le chargement n'est plus qu'une des nombreuses méthodes utilisées par les organisations pour collecter, importer et traiter les données. L'ETL et l'ELT sont tous les deux des composants importants de la stratégie élargie d'intégration des données d'une organisation.

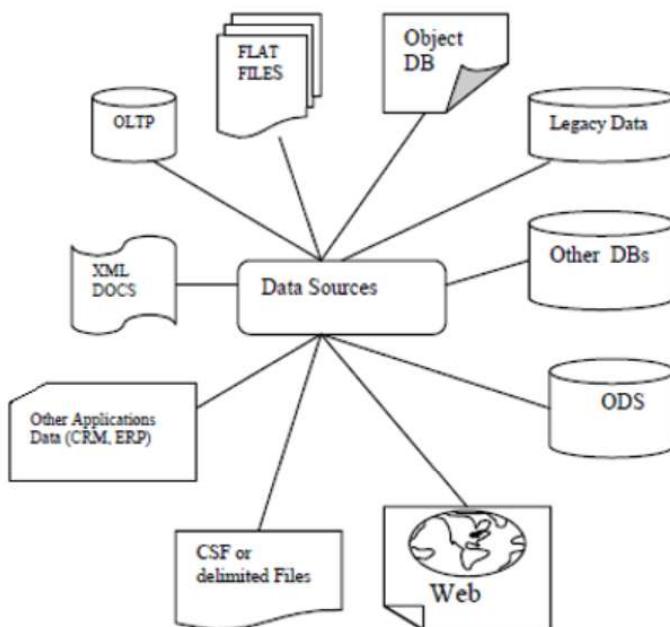


FIGURE 1.2 – Différents formats de sources de données

Source : lightsondata

1.1.2.2 Extraction (Extract)

L'extraction des données est la première étape du processus ETL [Voir la figure 1.1]. Ici, les données sont collectées à partir de différentes sources, qui peuvent inclure des bases de données, des tableurs, des API ou même des pages web. Ces données sont souvent non structurées et désordonnées, ce qui nécessite un nettoyage et une organisation avant de pouvoir être utilisées.

1.1.2.3 Transformation (Transform)

Lors de l'étape de transformation des données, les données extraites sont nettoyées, converties et restructurées [Voir la figure 1.1]. Ce processus peut impliquer le filtrage des données non pertinentes, le remplissage des valeurs manquantes, la standardisation des formats et l'agrégation des données. L'objectif est de transformer les données brutes en un format cohérent et utilisable.

1.1.2.4 Chargement (Load)

La dernière étape du processus ETL est le chargement des données, où les données transformées sont chargées dans un entrepôt de données ou une base de données cible [Voir la figure 1.1]. Ces données sont désormais organisées, propres et prêtes pour l'analyse, le reporting et, en fin de compte, la prise de décision.

1.1.3 Défis de l'intégration des données

L'un des principaux défis de l'intégration des données est la complexité et le coût élevé des méthodes traditionnelles d'intégration des systèmes. Toutefois, les progrès technologiques ont permis de mettre à disposition plusieurs solutions visant à simplifier le processus d'intégration [Actian (2024)]. Qu'il s'agisse de développement interne ou de solutions tierces, le choix de la bonne approche d'intégration est crucial pour réussir. Les responsables informatiques, les gestionnaires d'applications, les ingénieurs et les architectes de données jouent un rôle essentiel dans ce processus de planification, en veillant à ce que l'approche d'intégration choisie corresponde aux buts et aux objectifs de l'organisation.

1.2 Introduction à la Business Intelligence (BI)

Le terme Business Intelligence (BI) désigne les technologies, applications et pratiques de collecte, d'intégration, d'analyse et de présentation de l'information. L'objectif de la Business Intelligence est de soutenir une meilleure prise de décision des verticales métiers, commerciale, marketing, finance [Oracle (2025)]. Essentiellement, les systèmes de Business Intelligence sont des systèmes d'aide à la décision axés sur les données.

1.2.1 Définition

La Business Intelligence (BI) est un processus technologique d'analyse des données et de présentation d'informations pour aider les dirigeants, managers et autres utilisateurs finaux de l'entreprise à prendre des décisions business éclairées. La Business Intelligence englobe une grande variété d'outils, d'applications et de méthodologies qui permettent aux organisations de collecter des données à partir de systèmes internes et de sources externes. Ces données sont ensuite préparées pour l'analyse afin de créer des rapports, tableaux de bord et autres outils de Data Viz pour rendre les résultats analytiques disponibles aux décideurs et aux opérations.

1.2.1.1 Sources de données

Les sources de données utilisées pour enrichir son outil de BI comprennent les systèmes de gestion de la relation client (CRM), les informations sur la chaîne logistique, les tableaux de bord des performances commerciales, les analyses marketing, les données d'appel des call center. Les applications de Business Intelligence aident les entreprises à regrouper toutes ces sources disparates en une seule vue unifiée fournissant des rapports, des tableaux de bord et des analyses en temps réel.

Bien que la business intelligence ne donne pas de directives aux utilisateurs sur les décisions à prendre, ou ce qu'ils doivent faire ou ce qu'il se passera s'ils suivent un certain cours, la Business Intelligence ne se limite pas non plus au reporting. Au contraire, la BI offre aux employés un moyen d'examiner les données de l'entreprise et de comprendre les tendances du marché et d'en tirer des "learnings" pour faire

avancer toutes les verticales métiers de leur entreprise.

1.2.2 Les composants clés de la BI

Les composantes clés de la business intelligence sont : les logiciels ETL [Voir la page 15], la data warehouse, l'OLAP, la data mining, les analyses de données, les rapports et les tableaux de bord. Chaque composante joue un rôle précis.

1.2.2.1 Entrepôts de données (Data Warehouses)

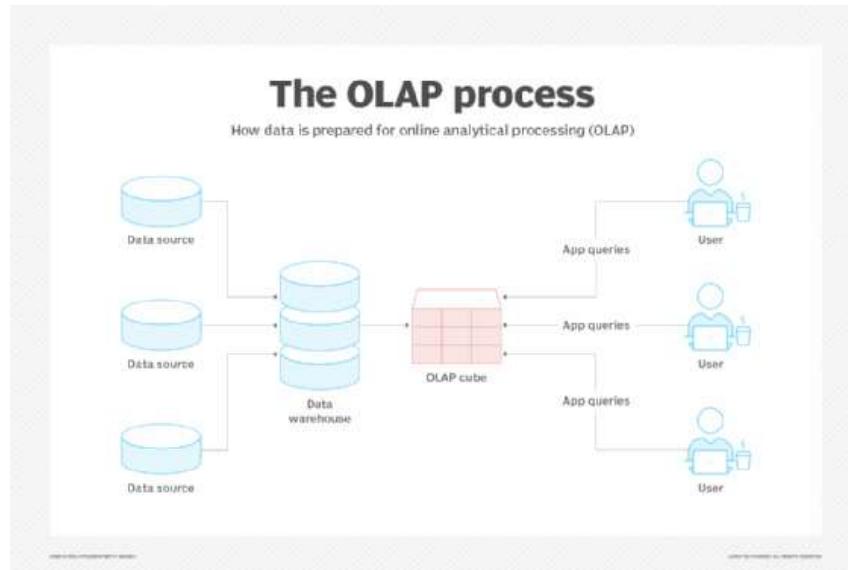
Le terme de 'Data Warehousing' désigne le processus consistant à collecter et à gérer des données en provenance de sources diverses [DataScientest (2021)], afin d'en dégager de précieuses informations exploitables par l'entreprise.

Un entrepôt de données est une plateforme utilisée pour collecter et analyser des données en provenance de multiples sources hétérogènes. Elle occupe une place centrale au sein d'un système de Business Intelligence.

1.2.2.2 L'analyse de données et l'OLAP

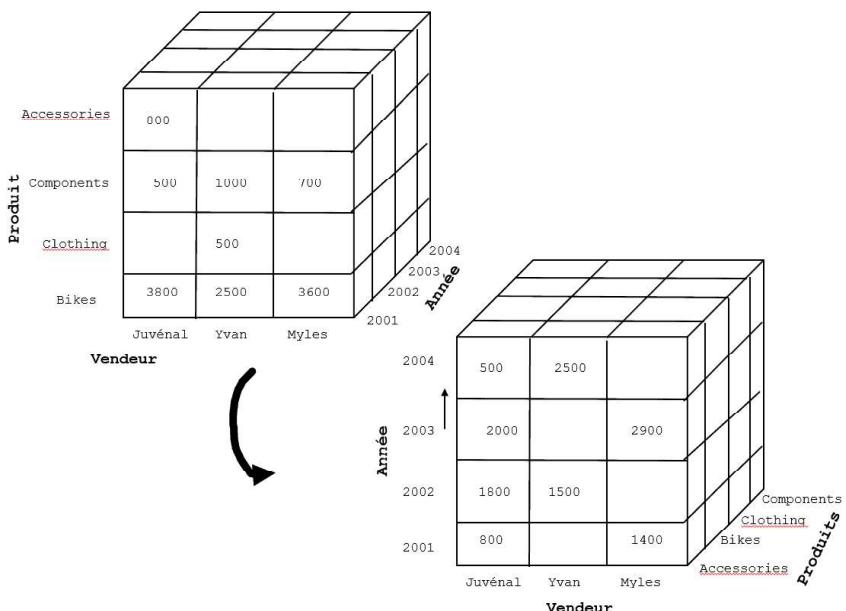
L'analyse de données est essentielle pour détecter les tendances et les modèles cachés dans les informations collectées. Elle joue un rôle clé dans l'identification des opportunités de croissance et la prise de décisions plus éclairées [Lemon Learning (2023)]. Parmi les méthodes généralement utilisées pour examiner les données, on distingue la modélisation prédictive, l'analyse statistique, etc.

L'OLAP (Online Analytical Processing) modifie les bases de données pour les rendre exploitables. À la fin du processus, il permet d'obtenir deux types de tables [Voir la figure 1.3], à savoir de dimensions et de faits. Les tables de dimensions fournissent une analyse des résultats produits par votre entreprise. Elles étudient vos ventes suivant trois paramètres : le client, la géographie et les produits.

**FIGURE 1.3 – Processus OLAP**

Source : techtarget

Les tables de faits recensent les résultats enregistrés par votre société : les livraisons, les ventes, les prestations, etc. Elles regroupent toutes les données que vous désirez analyser. La configuration de ces tables forme une structure d'information appelée cube OLAP [Voir la figure 1.4] et permet de présenter les résultats obtenus aux utilisateurs finaux.

**FIGURE 1.4 – Rotate-cube**

Source : Data Transition Numérique

1.2.2.3 Les rapports et tableaux de bord

Dans un contexte où les organisations collectent de grandes quantités de données provenant de diverses sources, la Business Intelligence (BI) joue un rôle essentiel pour transformer ces données brutes en informations exploitables. Deux des principaux outils de la BI sont les rapports et les tableaux de bord.

1. Les rapports BI

Les rapports BI sont des documents structurés, souvent générés automatiquement, qui présentent des données analysées sous forme de tableaux, de graphiques et de métriques. Ils permettent :

- De communiquer des résultats précis à un instant donné,
- D'effectuer des analyses détaillées sur des sujets spécifiques,
- De répondre à des besoins ponctuels ou périodiques (ex. : rapport hebdomadaire des ventes, rapport mensuel de performance).

Ils sont généralement conçus pour être lus et interprétés par des analystes, des managers ou des décideurs.

2. Les tableaux de bord BI

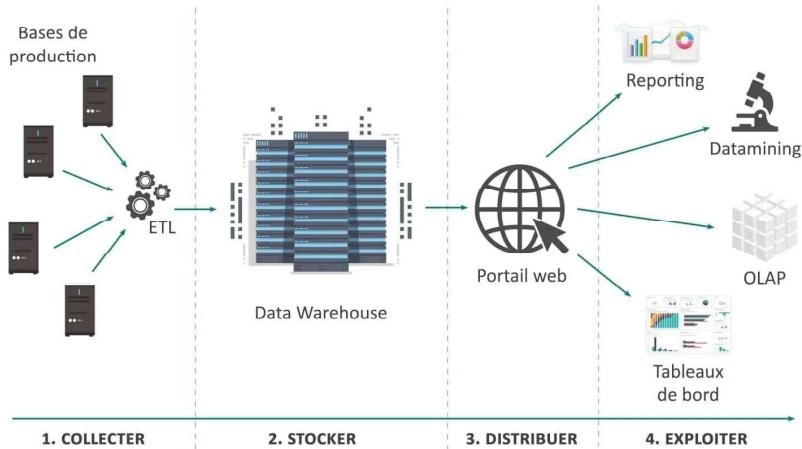
Les tableaux de bord (dashboards) sont des interfaces visuelles interactives qui affichent des indicateurs clés de performance (KPI) en temps réel ou quasi réel. Leur but est de :

- Donner une vue d'ensemble synthétique des données pertinentes,
- Permettre une prise de décision rapide grâce à une visualisation intuitive,
- Surveiller l'évolution des activités (ventes, trafic web, satisfaction client, etc.).

Un bon tableau de bord est clair, dynamique et personnalisable selon les besoins des utilisateurs. Des outils comme Power BI, Tableau, ou Google Data Studio sont largement utilisés pour leur création.

1.2.3 Processus typique de la BI

Les outils de Business Intelligence peuvent apporter de nombreux bénéfices à une entreprise. Ils permettent d'accélérer et d'améliorer la prise de décision, d'op-

**FIGURE 1.5 – Enter Caption**

Source : biworks

timiser des processus d'affaires internes, d'augmenter l'efficacité opérationnelle, la génération de nouveaux revenus, d'obtenir un avantage concurrentiel face à la concurrence.

Pour mettre en place un projet Business Intelligence, voici les 4 phases [Voir la figure 1.5].

1. La phase d'alimentation

Cette phase fait intervenir des processus ETL (Extract Transform Load) qui se chargeront de récupérer les données issues de différentes sources de stockage, de les formater, nettoyer et consolider.

2. La phase de modélisation et de stockage

Dans cette phase, le datawarehouse se charge de stocker et de centraliser les données en vue de la constitution du système d'information décisionnel.

3. La phase de restitution

Cette étape fait intervenir les outils de restitution des données afin de distribuer et de faciliter l'accessibilité des informations selon les fonctions et les types d'utilisation.

4. La phase d'analyse

Les utilisateurs finaux interviennent dans cette dernière étape pour exploiter et analyser les données qui leur sont fournies.

1.3 Cadrage du projet

1.3.1 Contexte général et Problématique

Dans un contexte marqué par une croissance exponentielle des données numériques, les organisations sont confrontées à un défi majeur : comment exploiter efficacement des données hétérogènes, souvent multilingues, issues de sources variées ? La coexistence de contenus en plusieurs langues (notamment en français et en arabe), de formats multiples et de structures non normalisées rend difficile leur traitement direct. Cette diversité empêche souvent une analyse cohérente et limite la prise de décision fondée sur des faits tangibles. De plus, l'absence de structuration et d'intégration adéquates peut entraîner une perte de valeur significative de l'information disponible.

1.3.2 Objectif du projet

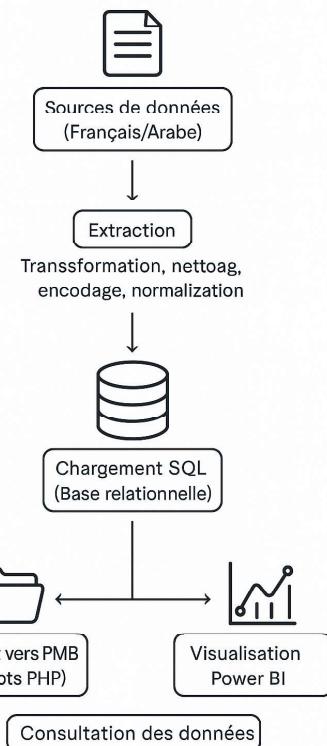
Ce projet vise à concevoir et mettre en œuvre un processus complet d'intégration de données, basé sur la méthodologie ETL (Extract, Transform, Load), pour répondre aux enjeux précédemment évoqués. Les objectifs principaux sont les suivants :

- Extraire des données issues de différentes sources, en français et en arabe ;
- Assurer leur transformation afin de garantir cohérence, uniformité et qualité ;
- Charger les données dans une base relationnelle SQL structurée ;
- Exploiter ces données via un outil de visualisation dynamique (Power BI) ;

Intégrer les données finales dans un système documentaire (PMB) pour une consultation centralisée.

1.3.3 Solution proposée

Pour atteindre ces objectifs, une solution modulaire et progressive a été mise en place. Elle s'articule autour des étapes suivantes [Voir la figure 2.3] :

**FIGURE 1.6 – Solution proposée**

- Extraction des données à partir de fichiers plats, bases existantes ou systèmes multilingues ;
- Transformation, impliquant la normalisation des formats, l’harmonisation des encodages (UTF-8 notamment pour l’arabe), et la gestion des correspondances sémantiques ;
- Chargement dans une base de données relationnelle SQL robuste permettant des jointures et requêtes complexes ;
- Visualisation à travers Power BI, afin de produire des rapports interactifs utiles aux parties prenantes ;
- Exportation vers le système documentaire PMB via des scripts PHP, pour une exploitation au sein d’une interface utilisateur finalisée.

Chapitre 2

Méthodologie

2.1 Analyse et préparation des données

2.1.1 Description des datasets

2.1.1.1 buf

Les données francophones représentent une sélection d'ouvrages littéraires, culturels et historiques rédigés en langue française. Chaque enregistrement correspond à un livre identifié par sa cote bibliothécaire et décrit selon plusieurs attributs : le titre de l'œuvre, le nom de l'auteur, le lieu de publication, la maison d'édition, ainsi que l'année de parution. Ces informations sont essentielles pour permettre le classement, la recherche, et l'exploitation de ces ouvrages dans un cadre bibliothécaire ou documentaire. Le tableau suivant présente un extrait de ces données.

Cote	Titre	Auteur	Lieu	Editeur	Année	Nb pages	Matière	Inventaire
0 001_091./01	La conquête du savoir = exploiting the earth a...	ASIMOV Isaac	NaN	Marabout	1982	436	Science et connaissance	13105
1 001_951./02	La science chinoise et l'occident	NEEDHAM Joseph	Paris	Seuil	1979	252	Science et connaissances en général	11225
2 001.18/01	Pour une prospective de la science . recherche...	PAPON Pierre	Paris	Seghers	1983	383	Avenir de la science	11828
3 001.18/02	Traité élémentaire de prévision et de prospective	DECOUFLE André-Clement	Paris	Puf	1978	432	Avenir de la science . Anticipation	14366
4 001.2/01	Les scientifiques parlent.	JACQUARD Albert	Paris	Hachette	1987	326	rapport des sciences les unes avec les	18176

FIGURE 2.1 – Extrait du fichier buf

2.1.1.2 bua

Ce jeu de données contient des informations relatives à des ouvrages rédigés en langue arabe, couvrant des thématiques variées telles que la religion, l'histoire, la

littérature et la pensée islamique. Chaque entrée fournit une description bibliographique complète : une cote unique, le titre de l’ouvrage, le nom de l’auteur, le lieu d’édition, le nom de la maison d’édition et l’année de publication. Ces métadonnées sont particulièrement utiles dans un contexte de gestion de fonds documentaires en langue arabe. Le tableau ci-dessous illustre quelques exemples issus de ce fichier.

Cote	Titre	Auteur	Lieu	Editeur	Année	Nb pages	Matière	Inventaire
0 "011/07	فهرس المخطوطات العربية التركية والقارئية ...والتو	فهيم علي، (إعداد) ناميتس	- لندن - سراغفو	مكتبة الغاربي منسرووك	1998.0	NaN	بليوغرافيا	27943
1 "02/02	مقدمة إلى نظم المكتبة المبنية على الحاسوب	أ. تيد. لوسي	عمان	المنظمة العربية للعلوم الإدارية	1985.0	465	علم المكتبات	17847
2 "3/43	عندما يختتم الصراع دليل عملي لاستخدام الوساطة	... أ. سلبيكو كارل	القاهرة	الدار الدولية للنشر والتوزيع	1999.0	348	العلوم الاجتماع	29267
3 "32/83	التخليل السياسي الحديث	أ.د. آل روبرت	القاهرة	مركز الأهرام للترجمة والنشر	1993.0	204	العلوم السياسية	28505
4 "330.114/01	جدول حول النقدم	أباظة، إبراهيم دسوقي	البيضاء	دار الكتاب	1975.0	30	الاقتصاد السياسي	11531

FIGURE 2.2 – Extrait du fichier bua

2.1.2 Nettoyage des données

Le nettoyage des données a également porté sur la gestion des valeurs manquantes (NaN) et des doublons. Les tableaux ci-dessous présentent les effectifs de valeurs manquantes dans chaque colonne avant et après traitement, ainsi que le nombre de doublons détectés et supprimés dans les deux jeux de données df_ar et df_fr. On constate une suppression complète des valeurs manquantes et doublons, assurant ainsi une qualité optimale pour les analyses ultérieures.

Jeu de données	Doublons avant nettoyage	Doublons après nettoyage
df_ar	0	0
df_fr	2	0

TABLE 2.1 – Nombre de doublons détectés avant et après nettoyage

Colonne	Nan avant nettoyage	Nan après nettoyage
<i>Jeu de données df_ar</i>		
Cote	0	0
Titre	26	0
Auteur	0	0
Lieu	474	0
Editeur	766	0
Annee	1286	0
Nb pages	82	0
Matière	13	0
Inventaire	0	0
Classe Dewey	0	0
<i>Jeu de données df_fr</i>		
Cote	0	0
Titre	0	0
Auteur	13	0
Lieu	313	0
Editeur	108	0
Annee	5	0
Nb pages	60	0
Matière	1	0
Inventaire	0	0
Classe Dewey	0	0

TABLE 2.2 – Nombre de valeurs manquantes (NaN) par colonne avant et après nettoyage

Lors de la préparation des données en arabe, une étape essentielle a concerné le nettoyage des cotes. Ce traitement a consisté à supprimer les guillemets et à effectuer une normalisation du format. En effet, la présence de guillemets dans certaines cotes provoquait des erreurs d’interprétation au sein du système PMB par exemple :

```
SELECT * FROM documents WHERE cote = "001_091_01";
```

compromettant ainsi l’exécution correcte des requêtes SQL. Ce nettoyage a donc été réalisé afin de garantir la cohérence syntaxique des champs de cote et d’éviter toute erreur d’indexation ou de recherche dans le catalogue. Un tableau illustratif avant/après ce traitement est présenté ci-après.

ID	Avant Nettoyage	Après Nettoyage	Problème Résolu
1	"001_091_01"	001/091/01	Guillemets parasites + séparateur incohérent
2	'Science/2023'	Science/2023	Guillemet simple mal placé
3	"Bio-Chimie-22"	Bio/Chimie/22	Espaces + tirets + guillemets
4	"Mécanique\\Fluides"	Mécanique/Fluides	Double backslash dangereux

FIGURE 2.3 – avant et après correction des erreurs typographiques

2.1.3 Transformation des données

Une étape de transformation des données a été réalisée afin d'améliorer la qualité et la cohérence des informations contenues dans l'inventaire. Cette transformation a notamment porté sur le reformatage de l'inventaire, qui a consisté à modifier la structure des codes d'inventaire afin d'harmoniser leur présentation et de faciliter leur traitement automatisé. Le nouveau format, débutant systématiquement par le préfixe FSR [Voir le tableau 2.4], permet d'identifier plus rapidement et clairement les documents appartenant à cette collection spécifique. Cette uniformisation garantit une meilleure cohérence des données, limite les risques d'erreurs lors des recherches, et simplifie considérablement la gestion informatique de l'inventaire.

Cote	Titre	Auteur	Lieu	Editeur	Année	Nb pages	Matière	Inventaire	Classe Dewey
0 001_091_01	La conquête du savoir = exploring the earth a...	ASIMOV Isaac	NaN	Marabout	1982	436	Science et connaissance	FSR_13105	Généralités
1 001_951_02	La science chinoise et l'occident	NEEDHAM Joseph	Paris	Seuil	1979	252	Science et connaissances en général	FSR_11225	Généralités
2 001.18/01	Pour une prospective de la science : recherche...	PAPON Pierre	Paris	Seghers	1983	383	Avenir de la science	FSR_11828	Généralités
3 001.18/02	Traité élémentaire de prévision et de prospective	DECOUFLE André-Crément	Paris	Puf	1978	432	Avenir de la science . Anticipation	FSR_14366	Généralités
4 001.2/01	Les scientifiques parlent.	JACQUARD Albert	Paris	Hachette	1987	326	rapport des sciences les unes avec les	FSR_18176	Généralités

FIGURE 2.4 – Extrait du tableau : Reformatage de l'inventaire avec ajout du préfixe FSR

2.1.4 Normalisation des données

Par la suite, une classification Dewey [Voir la table 2.3] a été appliquée aux cotes des fichiers en français et en arabe. La cote, en bibliothèque, constitue un code unique permettant d'identifier, de classer et de localiser un document. Elle remplit trois fonctions principales : elle sert d'adresse physique (emplacement sur les rayons), de clé de recherche dans le système d'information, et d'identifiant

thématique basé sur le domaine de connaissance. Nous avons adopté la Classification Décimale Dewey (CDD), reconnue à l'échelle internationale, notamment dans les bibliothèques publiques.

Code	Thématique	Exemples Courants
000	Généralités	Informatique, Encyclopédies, Bibliographie
100	Philosophie	Éthique, Psychologie, Logique
300	Sciences Sociales	Économie, Droit, Éducation, Politique
500	Sciences Exactes	Mathématiques, Physique, Chimie, Astronomie
800	Littérature	Romans, Poésie, Théâtre, Critique littéraire

TABLE 2.3 – Classification Décimale Dewey

Par exemple, une cote comme 330 .114/01 peut se décomposer ainsi : 330 correspond au domaine de l'économie, .114 désigne un sous-thème spécifique (par exemple le marché du travail), et /01 renvoie à un identifiant local [Voir les deux tableaux (2.5) et (2.6)]. Ce système assure une organisation logique et uniforme des documents, facilitant leur accès par les utilisateurs.

Cote	Titre	Auteur	Lieu	Editeur	Année	Nb pages	Matière	Inventaire	Classe Dewey
0 011/07	فهرس المخطوطات العربية التركية... والفارسية واليونانية على	فهيم علي، (إعداد) ناجيال سارفو	لندن - سارفو	مكتبة الفازى منسوبك	1998.0	NaN	بليوغرافيا	27943	Généralités
1 02/02	مقدمة إلى نظم المكتبة المبنية على الحاسوب	أ. تيد. لويس	عمان	المؤسسة العربية للعلوم الإدارية	1985.0	465	علم المكتبات	17847	Généralités
2 3/43	عندما يختبر الصراح دليل عمل لاستخدام الوساطة...	أ. سليماني كارل	القاهرة	الدار الولائية للنشر والتأريخ	1999.0	348	العلوم الاجتماعية	29267	Généralités
3 32/83	التحليل السياسي الحديث	أ.د. إل. روبرت	القاهرة	مركز الأهرام للترجمة والنشر	1993.0	204	العلوم السياسية	28505	Généralités
4 330.114/01	جدل حول التقدم	أيادى إبراهيم دسوقي	البيضاء	دار الكتاب	1975.0	30	الاقتصاد السياسي	11531	Sciences sociales

FIGURE 2.5 – Les données en français après la classification Dewey

Cote	Titre	Auteur	Lieu	Editeur	Année	Nb pages	Matière	Inventaire	Classe Dewey
0 001_091_/_01	La conquête du savoir = exploring the earth a...	ASIMOV Isaac	NaN	Marabout	1982	436	Science et connaissance	13105	Généralités
1 001_951_/_02	La science chinoise et l'occident	NEEDHAM Joseph	Paris	Seuil	1979	252	Science et connaissances en général	11225	Généralités
2 001.18/01	Pour une prospective de la science : recherche...	PAPON Pierre	Paris	Seghers	1983	383	Avenir de la science	11828	Généralités
3 001.18/02	Traité élémentaire de prévision et de prospective	DECOUFLE André-Clement	Paris	Puf	1978	432	Avenir de la science . Anticipation	14366	Généralités
4 001.2/01	Les scientifiques parlent.	JACQUARD Albert	Paris	Hachette	1987	326	rappart des sciences les unes avec les	18176	Généralités

FIGURE 2.6 – Les données en arabe après la classification Dewey

2.2 Visualisation des données

Dans le cadre de ce projet, plusieurs outils de visualisation des données ont été envisagés. Le choix s'est porté sur Power BI en raison de sa convivialité, de sa puissance analytique, ainsi que de sa capacité à gérer des jeux de données volumineux et diversifiés.

2.2.1 Présentation de Power BI

Power BI¹ est une plateforme de business intelligence développée par Microsoft qui permet de créer des rapports interactifs et des tableaux de bord dynamiques. Cet outil facilite l'analyse visuelle des données et la prise de décisions grâce à son interface intuitive et ses nombreuses fonctionnalités avancées.

2.2.2 Visualisation des données avec Power BI

Après avoir effectué les différentes étapes de nettoyage, transformation et classification des données, celles-ci ont été chargées dans Power BI pour permettre une exploration visuelle et une analyse interactive [Voir l'annexe]. Cette étape facilite la compréhension globale des données, la détection d'éventuelles anomalies, ainsi que la mise en évidence de tendances ou de regroupements pertinents.

2.3 Intégration dans la base de données PMB

2.3.1 Présentation de PMB (PhpMyBibli)

PMB (PhpMyBibli)² est un système intégré de gestion de bibliothèque (SIGB), un logiciel libre et open source développé pour gérer les bibliothèques. Il permet de gérer le catalogue, les emprunts, les auteurs.

PMB est un logiciel gratuit, utilisable par divers types de bibliothèques (universitaires, publiques, spécialisées), et est connu pour sa flexibilité et sa possibilité de personnalisation. Il est basé sur les technologies PHP et MySQL, ce qui facilite son adaptation aux besoins spécifiques de chaque bibliothèque.

1. <https://powerbi.microsoft.com>

2. <https://www.sigb.net>

2.3.2 Intégration de données

Les données nettoyées et transformées ont été importées dans la base de données pmb_db [Voir l'annexe], qui constitue le système de gestion documentaire utilisé par PMB. Cette intégration s'est faite via des scripts PHP, assurant la mise à jour automatique et cohérente des informations. Cette connexion entre la base de données et le logiciel PMB permet de maintenir l'inventaire à jour et de garantir l'accès rapide et fiable aux documents par les utilisateurs finaux.

2.3.3 Réindexation de la base de données

Suite à l'importation des nouvelles données, une étape essentielle de réindexation [Voir l'annexe] a été réalisée afin de mettre à jour les index de recherche dans PMB. La réindexation assure que toutes les modifications apportées aux données sont prises en compte dans le moteur de recherche, améliorant ainsi la rapidité et la pertinence des résultats pour les utilisateurs. Cette opération est cruciale pour maintenir la performance optimale du système documentaire.

Conclusion

Ce projet a permis de réaliser un traitement complet et structuré des données bibliographiques, allant du nettoyage à l'intégration dans un système documentaire. L'ensemble des étapes menées – nettoyage des cotes, gestion des doublons et des valeurs manquantes, normalisation par la classification Dewey, transformation des inventaires – a considérablement amélioré la qualité et la cohérence des données.

L'utilisation de Power BI pour la visualisation a facilité l'analyse et la prise de décision, tandis que l'importation dans la base de données pmb_db et la réindexation ont permis une exploitation directe dans le système PMB. Le travail accompli garantit une meilleure accessibilité aux ressources documentaires et optimise leur gestion au sein de l'infrastructure existante.

Dans une optique d'amélioration continue, plusieurs pistes peuvent être envisagées. L'automatisation des processus de traitement permettrait de gagner en efficacité, notamment pour la gestion de volumes de données plus importants. Il serait également pertinent d'enrichir les visualisations avec de nouveaux indicateurs, tels que les taux de consultation ou d'emprunt. Enfin, le développement de fonctionnalités avancées dans le système PMB, comme la recherche intelligente ou la recommandation documentaire, renforcerait l'utilité de la base au service des utilisateurs.

Annexe

- Pour une analyse visuelle complète et détaillée des données, toutes les visualisations réalisées sont présentées dans le fichier annexe intitulé **rappor t_visualizations_données_powerbi**, accompagné du fichier Power BI (.pbix) correspondent.
- Les étapes détaillées d'importation des données dans la base de données pmb_db, ainsi que la réindexation effectuée via les scripts PHP reliés à PMB, sont entièrement décrites dans le document **rappor t_deploiement_pmb**. Ce document contient également les captures d'écran associées à ces étapes.

Bibliographie

Actian. L'importance d'une intégration efficace des données, 2024. URL <https://www.actian.com/fr/blog/data-integration/data-integration-automation-data-management/>. Consulté le 18 mai 2025.

DataScientest. Data warehouse : Qu'est-ce que c'est ? qui s'en sert ?, 2021. URL <https://datascientest.com/data-warehouse>. Consulté le 18 mai 2025.

Ralph Kimball and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*. Wiley, 2011.

Lemon Learning. Business intelligence (bi), le guide complet, 2023. URL <https://lemonlearning.com/fr/blog/business-intelligence>. Consulté le 18 mai 2025.

Microsoft. Power bi documentation, 2023. URL <https://learn.microsoft.com/en-us/power-bi/>.

A. Mohamed and I. Alhassan. Challenges and approaches in multilingual data integration. *International Journal of Data Engineering*, 5(2) :45–56, 2020.

Oracle. Qu'est-ce que la business intelligence ?, 2025. URL <https://www.oracle.com/fr/database/business-intelligence-definition/>. Consulté le 18 mai 2025.

OVHcloud. Qu'est-ce que l'etl (extract, transform, load) ?, 2025. URL <https://www.ovhcloud.com/fr/business/intelligence-artificielle/etl-extract-transform-load/>.

//www.ovhcloud.com/fr-ma/learn/what-is-etl/. Consulté le 18 mai 2025.

SAP. Intégration des données : la clé pour connecter vos données, 2025. URL <https://www.sap.com/france/products/technology-platform/what-is-enterprise-integration/data-integration.html>.

SAS. Qu'est-ce que l'etl (extract, transform, load) ?, 2025. URL https://www.sas.com/fr_ch/insights/data-management/what-is-etl.html. Consulté le 18 mai 2025.

PMB Services. phpmybibli (pmb) : Sigb open source, 2021. URL <https://www.sigb.net>.

Talend. Qu'est-ce que l'intégration des données ?, 2025. URL <https://www.talend.com/fr/resources/what-is-data-integration/>.