Partie 3 : Plan de gestion des données (PGD)

1. Description des données et collecte ou réutilisation de données existantes

Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ? Les données textuelles utilisées pour la version initiale de la plateforme seront extraites des travaux réalisés par F. M. Rey (Université de Loraine) et E. Reymond (Yale University) dans leur édition critique des manuscrits hébreux de Ben Sira. Ces données, actuellement au format InDesign, seront à convertir au format XML TEI. La numérisation des manuscrits sera assurée par Cambridge (Taylor-Schechter Genizah Research Unit) et par F. M. Rey en collaboration avec la Bibliothèque de l'Alliance israélite universelle.
Les données initiales du projet sont donc pré-existantes et seront ré-utilisées pour la mise en œuvre de la plateforme.

Quelles données (types, formats et volumes par ex.) seront collectées ou produites ? Les données collectées au cours du projet seront de trois types distincts :

- Les données liées à la numérisation de manuscrits, où nous pouvons distinguer :
  - Les données liées à la numérisation de manuscrits, réalisées spécifiquement pour le projet. Ces images seront en format TIFF et JPG. En termes de volumétrie, nous nous attendons à une centaine d'images.
  - Les données liées à la numérisation de manuscrits, telles qu'importées par les utilisateurs de la plateforme. Ces images seront en format TIFF et JPG, et il est difficile de fixer une volumétrie, car celle-ci dépendra de la popularité du projet.
- Les données liées à la transcription des données, qui seront soient importées depuis des données transcrites au format XML TEI, soit réalisées directement sur la plateforme par les utilisateurs. Concernant les données de la plateforme pilote, nous nous attendons à la transcription d'environ 300 pages papier. Concernant les ajouts utilisateurs, la faible volumétrie des données XML converties en JSON, ainsi que la capacité des bases NoSQL à gérer un grand volume de données, nous assure d'une base solide pour la volumétrie envisagée.
- Les données liées à l'utilisation de la plateforme :
  - Les informations liées à l'inscription sur la plateforme pour accéder à l'éditeur collaboratif : nom, prénom, organisme de recherche si pertinent, mots de passe...
  - Les informations liées à l'utilisation de plateforme et de s'assurer de la bonne performance de la plateforme : cookies, référencement...
  - Les informations liées au travail de recherche sur la plateforme : métadonnées renseignées par l'utilisateur, travail en cours de transcription...

  L'ensemble des métadata représente un volume négligeable par rapport au reste des données, surtout vu l'efficacité de la base de données choisie.

Concernant les données de type image, celles-ci seront réalisées dans du stockage objet, de type S3 comme MinIO [1], de manière à les stocker de manière efficace et pérenne. Les autres données, textuelles et utilisateurs, seront stockées à l'aide d'une base NoSQL, orientée document, comme par exemple MongoDB, de manière à nous garantir une intégration simple face au standard Web HTTP/HTTPS. Toutefois, l'architecture du projet sera définie de manière à proposer des interfaces génériques pour permettre le développement des connecteurs aux données, et minimiser ainsi l'adhérence sur la technologie utilisée, pour permettre la connexion à la base Digital Qumrân par l'API REST mise à disposition. Cette modularité permet de garantir la pérennité du projet vis-à-vis des sources de données.

2. Documentation et qualité des données

Quelles métadonnées et quelle documentation (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ? Pour chaque manuscrit importé, il sera demandé à l'utilisateur de soumettre un ensemble de métadonnées accompagnant sa soumission : titre et auteur de l'œuvre, provenance du manuscrit, date de la numérisation…
Les schémas de la base de données, ainsi que l'API REST permettant d'accéder aux données, sera strictement documentée

1. https ://min.io/

en respectant les standards d'OpenAPI [2], de manière à faciliter la réutilisation des données, à la fois pour les contributeurs de la plateforme, mais aussi par des organismes externes souhaitant accéder aux données.
Le projet en lui-même, entièrement OpenSource, sera documenté à la fois en proposant une documentation de type ReadTheDocs [3], qui nous permettra d'héberger le guide utilisateur, le guide technique, ainsi qu'une description du schéma de la base de données.

Quelles mesures de contrôle de la qualité des données seront mises en œuvre ? La qualité des données sera validée par les pairs, de par la nature collaborative de la plateforme. Chaque utilisateur devra être enregistré pour contribuer, et sa contribution sera soumise à une validation par les pairs à chaque soumission.
La qualité des données et leur intégration seront de plus validées à l'aide d'outils automatiques, sous la forme de tests unitaires ou de tests d'intégration, comme une vérification systématique des termes dans un dictionnaire avant la publication de la donnée.
L'intégrité du format de la donnée sera, elle, vérifiée par des schémas de sérialisaion/désérialisation à chaque passage de la donnée par l'API REST.

3. Stockage et sauvegarde pendant le processus de recherche

Comment les données et les métadonnées seront-elles stockées et sauvegardées tout au long du processus de recherche ? Tout au long du projet, les données seront sauvegardées à la fois sur le serveur qui hébergera à terme les données, mis à disposition par l'Université de Lorraine, et que nous utiliserons pour réaliser les développements, et seront dupliquées sur un stockage S3 de type NAS qui nous permettra de garantir une sauvegarde automatique récurrente des données (par exemple, deux fois par jour).

Comment la sécurité des données et la protection des données sensibles seront-elles assurées tout au long du processus de recherche ? Au long du processus de recherche, les données seront hébergées sur un serveur, accessible seulement par les partenaires ayant part au développement du projet, en utilisant un mécanisme de connexion par protocole ssh [4], l'un des plus sécurisés. Le serveur sera hébergé par l'université de Lorraine, ce qui nous garantit de contrôler précisément la localité des données. L'application étant *containerisée* l'ensemble du transfert des données sera fait sur un réseau privé.
Une fois que le processus de développement aura suffisamment abouti, nous mettrons en place le plus rapidement possible un mécanisme d'authentification, par exemple grâce à Keycloak [5], qui nous garantira une protection efficace de l'API et des données exposées.

4. Exigences légales et éthiques, codes de conduite

Si des données à caractère personnel sont traitées, comment le respect des dispositions de la législation sur les données à caractère personnel et sur la sécurité des données sera-t-il assuré ? La plateforme reposant sur un mécanisme d'authentification, nous demanderons aux utilisateurs de fournir à minima un pseudonyme, un e-mail, et s'il/elle le souhaite, une affiliation, de manière à garantir la qualité des données exposées sur la plateforme. Ces données étant des données personnelles, selon la définition de la CNIL car permettant d'identifier la personne, nous nous engageons à respecter les principes de la CNIL : les données collectées ne viseront qu'à garantir le bon fonctionnement de la plateforme, elles seront récoltées en toute transparence et les utilisateurs pourront se rétracter à tout moment ce qui entraînera la suppression de l'ensemble des données collectées, nous ne réaliserons aucun traitement supplémentaire dessus, et nous garantirons leur sécurité grâce à un hébergement privé. Les données personnelles seront de plus encryptées, protégeant nos utilisateurs même en cas de vol de disque.

Comment les autres questions juridiques, comme la titularité ou les droits de propriété intellectuelle sur les données, seront-elles abordées ? Quelle est la législation applicable en la matière ? En termes de propriété intellectuelle, les données présentées sur la plateforme seront placées sous une licence *Creative Commons NC* : les données fournies seront donc librement accessibles, à la fois *via* l'application Web en elle-même, et *via* une API REST que nous mettrons à disposition des utilisateurs souhaitant réutiliser les données, même pour des services tiers. Toutefois, celles-ci seront *Non Commercial*, et nous-mêmes et les utilisateurs de la plateforme ne devrons donc dériver aucune réutilisation commerciale de nos données.

2. https ://www.openapis.org/
3. https ://readthedocs.org/
4. https ://www.openssh.com/
5. https ://www.keycloak.org/

Tout le code réalisé sera placé sous licence *Apache Licence 2.0* [6], et sera hébergé sur l'outil de versionning collaboratif Github [7], pour que chacun puisse contribuer librement au projet.

Comment les éventuelles questions éthiques seront-elles prises en compte, les codes déontologiques respectés ? A notre connaissance, il n'y aura pas de questions éthiques spécifiques liées aux partages des données.

5. Partage des données et conservation à long terme

Comment et quand les données seront-elles partagées ? Y-a-t-il des restrictions au partage des données ou des raisons de définir un embargo ? Les données seront partagées dès que les développements seront assez avancés pour assurer la mise à disposition de données fiables. La création de la base de données et de l'API étant indispensables avant de commencer le développement Web, elles seront le développement prioritaire dans la phase initiale du projet. Nous envisagerons de plus la publication des jeux de données en OpenSource sur des plateformes comme Zenodo [8] dès que les jeux de données ont été suffisamment consolidés.

Comment les données à conserver seront-elles sélectionnées et où seront-elles préservées sur le long terme (par ex. un entrepôt de données ou une archive) ? L'ensemble des données concernant les images des manuscrits et leur transcription seront conservées à la fois sur le serveur en ligne, hébergé par un serveur de l'Université de Lorraine, et des sauvegardes versionnées seront réalisées régulièrement sur un ou des NAS, eux aussi hébergés par nos soins.

Quelles méthodes ou quels outils logiciels seront nécessaires pour accéder et utiliser les données ? Les données seront accessibles soit en JSON à l'aide du protocole HTTP/HTTPS, ce qui nécessite un accès à internet et un outil permettant de réaliser des requêtes HTTPS (navigateur Web, *curl*, *wget*...), soit accessibles en format brut directement mis à la disposition des utilisateurs sous forme de fichiers plats (JSON, XML, CV...).

Comment l'attribution d'un identifiant unique et pérenne (comme le DOI) sera-t-elle assurée pour chaque jeu de données ? L'attribution du DOI sera gérée par Zenodo.

6. Responsabilités et ressources en matière de gestion des données

Qui (par exemple rôle, position et institution de rattachement) sera responsable de la gestion des données (c'est-à-dire le gestionnaire des données) ? Le gestionnaire des données sera le porteur du projet, Jean-Sébastien Rey. Celui-ci s'assurera de la qualité des données et du bon respect du PDG. Il s'assurera aussi de la bonne intégration avec les partenaires et leurs données.

Quelles seront les ressources (budget et temps alloués) dédiées à la gestion des données permettant de s'assurer que les données seront FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable) ? Un stockage adéquat et la mise à disposition des données étant central pour le projet, la majeure partie de la phase initiale d'architecture logicielle sera consacrée au *design* de la base de données et de l'API REST de mise à disposition des données. Nous assurerons le principe FAIR selon les modalités suivantes :

- **Facile à trouver** : le lien vers la plateforme sera référencé sur les moteurs de recherche, ainsi que sur le site de nos partenaires. De plus, tout le code sera mis à disposition sur le site de versionnement collaboratif Github, qui redirigera l'utilisateur vers la documentation exhaustive de la base de données. Il sera aussi possible d'accéder facilement *via* une URL spécifique au schéma OpenAPI de l'API REST.
- **Accessible** : l'utilisation de protocole HTTP/HTTPS, ainsi que la possibilité d'export des données sous forme de fichier plat assure l'accessibilité des données.
- **Interopérable** : l'architecture logicielle choisie sera la andsibleplus générique possible, grâce à l'utilisation d'interfaces agnostiques à la technologie de la base et ainsi nous assurer de la pérennité du projet. L'intégralité de la plateforme sera *containerisée* [9] et facilement déployable, par exemple par l'utilisation d'Ansible [10], ce qui assure un déploiement automatisé et idempotent sur n'importe quel serveur. La possibilité d'exporter les données au format XML TEI et le développement de l'API assureront l'interopérabilité des données et du système.

6. https ://www.apache.org/licenses/LICENSE-2.0
7. https ://github.com/
8. https ://zenodo.org/
9. https ://www.docker.com/
10. https ://www.ansible.com/

— **Réutilisable** : par l'ouverture de l'intégralité du code à la communauté OpenSource, ainsi que par sa facilité de déploiement grâce aux outils de containerisation, d'automatisation, et de déploiement notre travail sera très facilement réutilisé.

Tableau de synthèse PGD

Source	Nature du résultat	Format(s) des données	Standard(s) des données	DOI prévus pour vos données	Nombre d'objets	Date de début de mise à disposition pour Biblissima+	Licence d'utilisation souhaitée	Solution(s) de stockage, versioning, sauvegarde pendant le projet	Solution et responsable du stockage et de l'archivage d'après le projet
Données collectées pour le projet	**Corpus d'image**	Tiff, JPG, PDF	IIIF	NA	100	31/01/2023	CC NC ou en fonction des institutions de convervation	Serveur local et BDD objet / Minidump	F Rey
	Edition critique de Ben Sira	InDesign / XML TEI	XML TEI	oui	300 pages A4 de texte	31/01/2023	CC NC	Serveur local	F Rey
Données collectées par la plateforme	**Données textuelles créées par les utilisateurs (BDD)** (BDD)	BSON / SQL dump	Export en JSON, XML TEI, txt, csv	Oui (Zenodo)	NA	NA	NA	Serveur local et serveur distant (Université de Göttingen)	F Rey
	Métadonnées utilisateurs	BSON	NA	NA	NA	NA	CC NC	Serveur local	F Rey
Code	**Application Web**	.js	NA	NA	NA	31/12/2023	Apache 2.0	Github	F. M. Rey
	Moteur d'inférence TAL	.py	NA	NA	NA	31/12/2023	Apache 2.0	Github	F. M. Rey
	Code API REST	.py	NA	NA	NA	31/12/2023	Apache 2.0	Github	F. M. Rey