EMAN

Plan de gestion de données créé à l'aide de DMP OPIDoR

Créateur du PGD : Richard Walter

Affiliation du créateur principal : CNRS

Modèle du PGD : ANR - Modèle de PGD (français)

Dernière modification du PGD : 04/10/2021

Résumé du projet :

EMAN (Édition de Manuscrits et d'Archives Numériques) est une plate-forme de publication numérique pour la diffusion et l'exploitation de documents et de fonds d'archives. Elle repose sur le logiciel Omeka à partir duquel elle développe des extensions pour l'édition numérique savante. EMAN diffuse des corpus en respectant les standards de l'édition numérique et de l'interopérabilité. La plate-forme soutient et héberge plus d'une cinquantaine de projets scientifiques d'édition de corpus. Ces projets publient, explorent, analysent des objets et des documents, à différents stades d'élaboration, produits par des écrivains, des artistes et des scientifiques, de l'Antiquité au XXIe siècle.

Chercheur Principal : Richard Walter

Contact pour les Données : Richard Walter

EMAN

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

1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?

En amont d'EMAN : opérations de préparation des données (formats, structurations, plan de nommage des fichiers).
Méthodologie : format recommandé des données pour imports (manuels ou automatiques) : format structuré sous forme de tableur. Utilisations de référentiels pour recueillir les données ?
Logiciel pour recueillir les données préexistantes et produire les nouvelles données : Omeka + plugins & thème EMAN
Procédures normalisées d'imports : recueil des données avec métadonnées associées, import automatique sur le site projet
Provenance des données documentée à deux endroits : pages de présentation sur le site projet + métadonnées DC.Source (ou autre)
Restrictions sur les données existantes : selon le projet
Sources existantes écartées : ce qu'on n'a pas eu le droit de faire avec ses données => choix faits, périmètre des données

1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?

49 corpus composés de :
fichiers images (iconographie, texte, etc.), fichiers sonores, fichiers multimédia : formats ouverts préconisés mais non obligatoires
fichiers transcriptions XML/TEI
base de données avec métadonnées associées aux fichiers
pages web de présentations des projets
plugins d'exploitation (géolocalisation, indexation, relations, visualisation, graphe...)
archives documentées de la plate-forme (wiki, manuels, guides, carnet de recherche)

Formats ouverts, standards, documentés et interopérables. Base de données MySQL, utilisation de PHP et javascript en code ouvert.
pages HTML, fichiers JPG, PNG, PDF + format sonore & video. Les autres formats sont stockés et uniquement téléchargeables
+ métadonnées au format Dublin Core + métadonnées personnalisées complémentaires + possibilités d'établir des relations dans des langages documentaires standards (frbr, foaf, dc étendu)
Volume :
quantité Giga/Tera
quantité d'objets (fichiers, items, enregistrements dans la base de données),

2. Documentation et qualité des données

2a. Quelles métadonnées et quelle documentation (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ?

Métadonnées Dublin Core + métadonnées personnalisées complémentaires
Mode d'organisation des données : structure Omeka classic (collection, item, fichier)
Structuration des dossiers : 1 Omeka par projet. Sur un Omeka, une base de données, un répertoire de fichiers à différentes tailles, répertoire des outils (plugins)
Documentation (manuel d'utilisation, wiki, lettres d'infos)

2b. Quelles mesures de contrôle de la qualité des données seront mises en œuvre ?

Curation des données (exemple : présence des métadonnées droits & éditeurs) :
jeu de requêtes : vérification des liens morts, des doublons, des champs vides, etc.
=> vérification de la présence des données et de leurs formes

Hiérarchie des responsabilités : procédure de validation des contenus proposés avec des profils contributeurs/responsables (avec signatures différentes et pages personnelles "profils")
Utilisation de listes contrôlées et de l'autocomplétion pour la saisie des métadonnées

Insertion des projets dans une structure collective avec partages d'expériences réguliers et perfectionnement continu des méthodologies.

3. Stockage et sauvegarde pendant le processus de recherche

3a. Comment les données et les métadonnées seront-elles stockées et sauvegardées tout au long du processus de recherche ?

Stockage individuel par projet des données avant introduction sur la plate-forme EMAN (guide de recommandations). Stockage des documents en haute définition (la plate-forme n'a pas la vocation d'archiver des corpus haut définition) est du ressort des projets.
Les données sont stockées sur une machine virtuelle gérée par Huma-Num qui en assure l'accessibilité et la sauvegarde.
Les métadonnées sont stockées dans un serveur de bases de données MySQL et chaque projet est un de ces bases de données.
La base de données est ouverte à la création du projet.

3b. Comment la sécurité des données et la protection des données sensibles seront-elles assurées tout au long du processus de recherche ?

Sécurité et protection de la machine virtuelle sont assurées par les outils d'Huma-Num.
Les données sensibles ne sont pas présentes [a priori] dans les projets EMAN (données à caractère personnel, politiquement sensibles des informations ou secrets commerciaux). [Si c'est le cas, une déclaration CNIL devra être faite.]
Chaque site a un système d'affichage public et privé. L'accès aux données privées (non encore publiées) est fait par mot de passe donné par un administrateur.

4. Exigences légales et éthiques, codes de conduite

4a. 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 ?

Les données à caractère personnel ne sont pas traitées sur la plate-forme EMAN. [données à caractère personnel :]
[pages personnelles des collègues et des étudiants ? ces données sont mises par ceux-ci, ce ne sont pas des données collectées par autrui. Consultation à prévoir d'une juriste pour savoir si les CV, photos et présentation des profils sont des données personnelles, si on les mets soi-même. Il faut qu'elles sachent clairement que ces données sont publiées sur le site ("consentement éclairé")]
Les porteurs de projets sont responsables des contenus publiés sur le site. Ils s'engagent à retirer tout contenu contesté.

Les données présentant les participants aux projets sont stockées dans une base de données chiffrée sur une machine virtuelle gérée par Huma-Num.

4b. 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 ?

Législation française s'applique car serveur géré en France.
Les droits sur les données diffusées sont gérés par les porteurs de projet [ce qui est dans le tableau de bord vous concerne].
Les droits sur les documents publiés et les droits sur les contenus créés sur EMAN sont indiqués dans la balise DC.Droits.
Tout contenu sur le site est validé et signé par un(e) ou des responsable(s).

4c. Comment les éventuelles questions éthiques seront-elles prises en compte, les codes déontologiques respectés ?

Le projet se place dans le cadre des codes et des règles déontologiques des institutions partenaires du projet, pour le respect de la propriété intellectuelle et des mentions des sources utilisées.
La plate-forme respecte les obligations des hébergeurs de contenus numériques, pour l'intégrité des données. [pas d'intervention de l'admin dans les contenus] Les porteurs de projets sont signataires de conditions générales d'utilisation (CGU) pour accéder aux services de la plate-forme.

5. Partage des données et conservation à long terme

5a. 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 ?

Chaque projet est responsable de son calendrier de publication des données et des éventuels embargos. Des données sont publiques, d'autres au statut privé et l'acte de publication est du ressort du responsable.
La plate-forme s'engage à faire une curation des données juridiques (présence systématique des indications de responsabilité et de propriété intellectuelle).
Les données sont partagées via des sites web Omeka personnalisés et l'interopérabilité des données est assuré par un entrepôt OAI-PMH ouvert dès la création du site.
Par défaut, les données produites numériques sur la plate-forme sont placées sous la licence Licence Creative Commons Attribution – Partage à l'Identique 3.0 (CC BY-SA 3.0 FR). [quid des transcriptions qui doivent être modifiées ?]

5b. 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) ?

[EMAN est sur une machine d'Huma-Num : la sauvegarde est assurée mais pas l'archivage, qui est une opération spécifique. Aucune politique d'archivage à long terme est faite actuellement sur les données EMAN]
Chaque projet est responsable de la sélection de ses données et de leur stockage sur la plate-forme.
Chaque projet de la plate-forme a un entrepôt de données sur le modèle de l'OAI-PMH et qui sera accessible au-delà de la fin du projet.
Les responsables de projets peuvent récupérer à tout moment leurs données pour les réutiliser et les archiver ailleurs.

5c. Quelles méthodes ou quels outils logiciels seront nécessaires pour accéder et utiliser les données ?

Le choix de la plate-forme, via l'utilisation du logiciel Omeka, est d'avoir une accessibilité totale en ligne des données. Un navigateur suffit pour consulter les données publiques comme privées.
Les données peuvent être publiques, accessibles sans restriction, ou être privées, accessibles par mot de passe selon la décision du porteur de projet.
La constitution et l'utilisation des données se font via des formulaires et des plugins du logiciel Omeka adaptés par la plate-forme.
Le logiciel Omeka, développé par le Center for History and New Media (CHNM), repose sur un langage informatique non propriétaire, maintenu par

une large communauté de développeurs ; les données peuvent être récupérées sous forme de tableurs de données et de répertoires de fichiers, pour des exploitations autres.

5d. Comment l'attribution d'un identifiant unique et pérenne (comme le DOI) sera-t-elle assurée pour chaque jeu de données ?

Chaque notice de collections, d'items et de fichiers a un identifiant unique dans la base de données Omeka.
A terme, il est envisagé d'obtenir des identifiants pérennes comme des DOI par le dépôt des métadonnées auprès d'organismes *ad hoc*.
Une politique d'obtention de numéros ISSN pour le site du projet ou les éditions numériques présentes sur ce site est envisagée par un groupe de travail spécifique.

6. Responsabilités et ressources en matière de gestion des données

6a. 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 de la plate-forme et ses institutions de rattachement sont responsables du stockage, de la sauvegarde, de l'archivage et du partage des données.
2021 : Richard Walter, laboratoire Thalim (CNRS-ENS-Sorbonne nouvelle).
Les responsables de projets sont les gestionnaires de leurs données : saisie des données, production des métadonnées, qualité des données [nommer les personnes et leurs institutions].
La mise en oeuvre du PGD et ses mises à jour sont assurées respectivement par le gestionnaire de la plate-forme et par les responsables de projets.

6b. 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) ?

Structurellement, EMAN respecte les principes FAIR. La plate-forme repose sur des standards et des logiciels conçus dans le respect de ces principes. La FAIRISATION des données est directement intégrée dans le fonctionnement de la plate-forme et est pris en charge par le gestionnaire de la plate-forme et ses institutions de rattachement. [2021 : salaire chargé d'un IR + prestations de service développement]
Les projets sont responsables de la documentation de leur démarche et de la description de leurs données. La mise à disposition des méthodologies et des outils de traitement propres à EMAN repose sur une culture commune de respect des principes FAIR et de la science ouverte.