GTnum LS2N "L'impact de l'Intelligence Artificielle à travers l'Éducation Ouverte" #IA_EO – Plan de gestion de données - Groupes thématiques numériques de la Direction du numérique pour l'éducation (Ministère de l'Éducation nationale et de la Jeunesse) 2020-2022

Colin de La Higuera, Mélanie Pauly Harquevaux

▶ **To cite this version:**

Colin de La Higuera, Mélanie Pauly Harquevaux. GTnum LS2N "L'impact de l'Intelligence Artificielle à travers l'Éducation Ouverte" #IA_EO – Plan de gestion de données - Groupes thématiques numériques de la Direction du numérique pour l'éducation (Ministère de l'Éducation nationale et de la Jeunesse) 2020-2022. 2023. hal-04000486

HAL Id: hal-04000486
https://hal.science/hal-04000486

Submitted on 22 Feb 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

DMP du projet "GTnumérique L2SN / LAB STICC #IA_EO"

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "ANR - Modèle de PGD (français)" fourni par Agence nationale de la recherche (ANR).

Renseignements sur le plan

Titre du plan	DMP du projet "GTnumérique L2SN / LAB STICC #IA_EO"
Domaines de recherche (selon classification de l'OCDE)	
Langue	fra
Date de création	2021-03-30
Date de dernière modification	2023-02-13
Identifiant	- Identifiant : GTnum 2020-2022 LS2N LAB STICC #IA_EO

Renseignements sur le projet

Titre du projet

GTnumérique L2SN / LAB STICC #IA_EO

Résumé

L’objectif de ce groupe de travail est d’étudier l’impact de l’IA sur l'Éducation, en particulier sur l’Open Education, au travers de multiples opportunités que l’IA permet d’ouvrir, de manière intelligible, facilement exploitable et reproductible. Les questions abordées s’articulent d’abord autour des 3 axes suivants :

- 1) Exploitation intelligible des Learning Analytics par l’utilisateur (apprenant, enseignant, institutions, parents, etc.) et leur combinaison avec les techniques de l’IA.
- 2) Expérimentation de nouvelles manières d’enseigner et d’apprendre avec l’IA, à travers la recommandation et la personnalisation des apprentissages, l’aide à l’orientation, la prédiction des élèves en décrochage, ou encore le développement des compétences méta-cognitives et sociales chez l’apprenant.
- 3) La création et la co-conception de ressources éducatives ouvertes (en situation d’urgence ou pas), par l’identification de ressources pertinentes, et l’aide au respect de questions éthiques liées à l’exploitation de ces ressources. Parmi les aspects à développer/étudier, la question des métadonnées, la possibilité de s’en affranchir avec l’IA.

Sources de financement

- DNE / MEN :

Produits de recherche :

1. Collection de RELs (Jeu de données)

Contributeurs

Nom	Affiliation	Rôles
Mélanie Pauly Harquevaux		• Coordinateur du projet • Personne contact pour les données (Collection de RELs) • Responsable du plan de gestion de données

DMP du projet "GTnumérique L2SN / LAB STICC #IA_EO"

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

Collection de RELs

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

La démarche suivie pour recueillir, puis produire les données relatives au projet est la suivante :

Étape 1 : Établir, puis enrichir une liste d'annuaires, de catalogues de collections de ressources éducatives libres potentielles. Cela se fait sur la base de nos propres analyses, de notre connaissance du domaine, mais aussi et surtout par l'interaction avec les acteurs de l'éducation.

Étape 2 : Exploration des annuaires, vérification sur la base de différents critères. Les annuaires suggérés peuvent contenir des ressources. Mais celles-ci peuvent être difficilement accessibles si l'organisation du site n'est pas correcte. Et de plus, se pose le problème des licences. Si les ressources ne sont pas correctement licenciées, il est inutile, contre-productif, interdit ou dangereux de récupérer les ressources. Par "correctement licencié", nous souhaitons dire : (1) utilisant une licence ouverte reconnue comme telle, typiquement de la famille Creative Commons, et (2) cette licence doit être accessible et clairement liée à la ressource. Une partie de cette étape 2 consiste à évaluer l'opportunité de récupérer ces ressources.

Étape 3 : Établissement de listes de ressources dûment licenciées, avec les métadonnées existantes. Si un site a été identifié comme utile au sens défini ci-dessus, une liste d'adresses de ressources est établie. Et pour chaque ressource, les métadonnées disponibles existantes sont récupérées.

Étape 4 : Enrichissement via des algorithmes d'intelligence artificielle des ressources. La liste est traitée par la chaîne d'analyse X5-GON qui va utiliser un certain nombre d'algorithmes pour extraire le texte (transcription), traduire ce texte en plusieurs langues, effectuer l'analyse sémantique et construire un certain nombre de métadonnées supplémentaires. Une liste de méthodes peut être trouvée sur le site du projet X5-GON (https://www.x5gon.org/science/deliverables/).

Étape 5 : Stockage sur les bases de données de Posta. Les données originales ne sont pas conservées. Les métadonnées construites par les méthodes décrites ci-dessus sont stockées dans la base de données de X5-GON (les serveurs sont ceux de Posta, le service public de Poste en Slovénie). Ces données sont accessibles par tous avec des APIs ouvertes. En particulier, un dump de la base de données est récupéré régulièrement à Nantes.

La stratégie décrite ci-dessus vise à ne collecter que des données dont les licences permettent clairement un traitement ultérieur.

De plus, à aucun moment, des données personnelles ou des données utilisateur ne seront collectées.

Le premier jeu de données en langue française que nous voulons traiter ainsi est celui d'Édubase, géré par le ministère de l'Éducation nationale (https://edubase.eduscol.education.fr/).

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

Nous distinguons ici les données collectées des données produites.

Les données collectées sont les ressources éducatives libres dûment licenciées provenant de collections en langue française.

En premier lieu, il s'agit des données de ressources partiellement identifiées et organisées autour de la plateforme Édubase.

À partir de la page http://edubase.eduscol.education.fr/, il est possible d'accéder à toutes les ressources recensées dans Édubase. À date du 17 mai 2021, 12 777 sont disponibles et le tableau ci-dessous détaille le nombre de ressources publiées par chaque académie.

Académie	Nombre de ressources
Nantes	1263
Versailles	1215
Aix Marseille	839
Poitiers	750
Normandie	663
Créteil	601
Grenoble	564
Besançon	510
Amiens	502
Nancy-Metz	461
Lyon	454
Dijon	449
Strasbourg	434
Rennes	410
Orléans-Tours	386
Bordeaux	363
Toulouse	357
Reims	327
Nice	300
Paris	284
Martinique	284
Clermont-Ferrand	249
Montpellier	240
Limoges	194
Lille	193
Guyane	181
La Réunion	143
Guadeloupe	64
Corse	25
Nouvelle-Calédonie	24
Autres	22
Ministère de l'Éducation Nationale	21
Polynésie Française	5
TOTAL	**12 777**

Au fur et à mesure de l'avancée du projet, de nouvelles données seront ajoutées. Les sites identifiés à ce jour sont ceux dépendant du Ministère de l'Enseignement Supérieur et de la Recherche : les Universités Numériques Thématiques.

Les données produites sont des modèles construits à partir des données collectées. Il s'agit donc du résultat d'algorithmes d'intelligence artificielle : des modèles, des données intégrées dans la base de données X5-GON (https://www.x5gon.org/science/deliverables/).

1. Produites par Nantes : une liste d'URLs + métadonnées de RELs françaises correctement licenciées et donc susceptibles d'être ingérées dans X5-GON.
2. Production par X5-GON : des métadonnées enrichies par IA.

2. Documentation et qualité des données

Collection de RELs

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

Pour la collecte, les métadonnées correspondent à certains éléments descriptifs des formats Dublin Core et ScoLOMFR.

Le format Dublin Core comprend un niveau simple, qui permet de fournir une quinzaine d'éléments descriptifs relatifs au contenu, à l'instanciation et à la propriété intellectuelle. Le niveau qualifié propose des éléments supplémentaires, des qualificatifs. (cf. https://www.dublincore.org/specifications/dublin-core/usageguide/#basicprinciples).

Le ScoLOMFR est un format de description des ressources pédagogiques pour l'enseignement scolaire. Dans la version la plus récente (actuellement 6.0), le profil d'application ScoLOMFR contient 46 vocabulaires spécifiques à l'enseignement scolaire. (cf. https://www.reseau-canope.fr/scolomfr/data/fr/).

Pour la production, ce sont les métadonnées du projet X5-GON accessibles par API. Nous trouverons ici une liste (en langue anglaise) de ces métadonnées : https://www.x5gon.org/wp-content/uploads/2020/04/D3.3_final.pdf.

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

Étant donnée la stratégie choisie, un filtrage sur la qualité a lieu avant la collecte.

Ainsi, dans l'étape 2 décrite ci-dessus, le travail d'ingénierie consiste à évaluer la qualité d'une collection : liens absents, versions différentes de celles identifiées dans les méta-données.

Un second enjeu de qualité est lié au caractère pédagogique de ces données : pourquoi un cours est-il de qualité (suffisante) ? Le choix qui est fait est de juger la qualité d'une collection plutôt que la qualité d'une ressource. Ainsi, toujours à l'étape 2 (exploration), nous devons trouver ces éléments de qualité avant de proposer l'ingestion de la collection.

La qualité pédagogique des ressources est donc du ressort des curateurs de la collection.

Pour donner un exemple, le ministère de l'Éducation nationale a publié une liste de Youtubers reconnus et recommandés. Les vidéos individuelles de ces youtubers ne seront pas analysées. Si ces vidéos ont la licence correcte, elles seront ajoutées à la collection.

La qualité des données produites est liée à l'évaluation très positive du projet X5-GON. Ce projet H2020 a été évalué en avril 2021. Le rapport d'évaluation sera rendu public prochainement. Notons en particulier :

- la qualité du partenariat,
- la qualité des méthodes de transcription et de traduction automatiques,
- la qualité des algorithmes d'analyse sémantique, qui est aussi évaluée par la qualité des articles scientifiques publiés.

Les données sur le site sont bien documentées, bien organisées. Les données sur les sites académiques sont de natures très disparates. Une cartographie plus précise est en cours de construction.

3. Stockage et sauvegarde pendant le processus de recherche

Collection de RELs

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

1. Pendant le processus de recherche, les données seront stockées sur les serveurs de recherche du laboratoire LS2N (solutions cloud). Ces serveurs sont sécurisés car utilisés pour d'autres applications particulièrement sensibles (données médicales, industrielles).
2. Le serveur Posta contiendra les données "IA" produites. Posta Slovenicke est l'organisme postal public Slovène, partenaire du projet X5-GON et agréé pour ce stockage.

Rappelons que les données dont nous parlons sont des ressources éducatives sous licence libre et donc non sensibles.

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

Sécurisation des données de bout en bout.
Les données étant ouvertes, une fois accédées, la question ne se pose plus.

Les moyens affectés au projet ne permettent pas d'avoir accès à un accompagnement spécifique sur ces questions de sécurité. Nous suivrons strictement les consignes établies par le Laboratoire des Sciences du Numérique de Nantes et de l'Université de Nantes.

4. Exigences légales et éthiques, codes de conduite

Collection de RELs

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 ?

Nous n'aurons pas accès à des données personnelles, à l'exception du nom des auteurs des cours. Mais cette information est volontairement publique, puisque c'est l'intention de l'auteur de se revendiquer en tant que tel. De plus, les licences Creative Commons considérées usuellement contiennent toutes l'attribut "BY", indiquant justement que les droits de propriété intellectuelle sont respectés.

Pour les questions de sécurité, nous suivrons strictement les consignes établies par le Laboratoire des Sciences du Numérique de Nantes et l'Université de Nantes.

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 ?

Cette question est essentielle. C'est bien ce qui justifie le choix de ne collecter que des ressources éducatives correctement licenciées. La famille des licences Creative Commons est connue et les traitements algorithmiques envisagés sont compatibles avec les licences suivantes : CC-BY, CC-BY-SA, CC-BY-NC, CC-BY-NC-SA. D'autres licences sont également compatibles et nous bénéficions de l'aide de l'équipe de Mme Patricia Serrano, Maître de Conférences à l'Université de Nantes, pour résoudre ces questions de compatibilité.
Avec les choix proposés, nous sommes dans un cadre juridique rassurant.
Ajoutons que nous suivrons également strictement les consignes établies par le Laboratoire des Sciences du Numérique de Nantes et l'Université de Nantes.

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

En adossant ce travail à la fois à la Chaire Unesco RELIA et à un laboratoire public de Recherche, nous assurons la capacité de discuter, d'effectuer de la veille sur ces questions d'éthique, en contact avec les meilleurs spécialistes internationaux.

Nous suivrons également dans ce cas les consignes établies par le Laboratoire des Sciences du Numérique de Nantes et l'Université de Nantes. Nous avons également accès à la commission éthique du LS2N et, en relation avec nos partenaires internationaux, à des forums internationaux sur les enjeux d'éthique en Sciences de Données.

5. Partage des données et conservation à long terme

Collection de RELs

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 ?

Les données collectées sont partagées par nature, puisque la licence le précise. Les données produites sont partagées par décision du consortium X5-GON. Elles sont partagées de différentes manières :

- par des interfaces adaptés, comme X5-Learn, X5-Blind ou X5-Moodle ;
- par des APIs permettant à des développeurs de proposer de nouveaux produits bâtis autour de ces ressources.

Les conditions de partage des données sont donc celles des licences sous lesquelles les ressources/données ont été déclarées.

Et dans ce cas, aucun embargo n'est nécessaire.

Une question se pose : celle de la maintenance dans le cas où un cours disparaît, par décision du site d'origine. Les systèmes de X5-GON permettent de détecter cette situation et de mettre à jour la base de données.

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) ?

Il y a 3 niveaux :

- Les données d'origine restent sur les serveurs d'origine, et aucune conservation par l'Université de Nantes ou par X5-GON n'est nécessaire. La copie effectuée pour effectuer l'extraction du texte est ensuite éliminée.
- Les annuaires enrichis produits après l'identification des ressources utiles (adresses des ressources et métadonnées) sont basés sur des données publiques (les métadonnées sont également licenciées). Ces données seront stockées sur le cloud de l'Université de Nantes pendant 4 ans. Puis, avant la fin 2025, il sera décidé de prolonger ou non ce stockage.
- Les métadonnées enrichies par intelligence artificielle qui sont stockées sur Posta jusqu'en 2023 en attendant un accord avec l'Unesco pour une extension pérenne : rappelons que l'objectif est de rendre ces données disponibles pour tous, donc l'objectif est d'assurer la pérennité des données.

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

Dans la mesure où aucune donnée personnelle n'est collectée, les seules données manipulées sont celles pré-existantes et hébergées par les académies qui nourrissent ensuite les chaînes de traitement X5-GON. Les résultats sous forme de graphes RDF sont stockés dans un logiciel de gouvernance des données permettant leur import et export dans les formats définis par les normes de partage du W3C, typiquement une application Web comme TopBraid EDG ou WebProtégé.

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

Les données en question, étant toutes sous licence Creative Commons, sont destinées par nature à être partagées par et avec le plus grand nombre.

Aucune donnée personnelle, d'usage, de tracking, ne sera ni collectée ni stockée.

Les jeux de données se distinguent en fonction de leur position dans le cycle de vie. Concernant les données collectées, à savoir les ressources éducatives libres hébergées par différentes académies et listées dans des catalogues et annuaires, elles sont publiquement exposées via le protocole OAI-PMH qui garantit, entre autres, leur identification unique et pérenne au sein des entrepôts. En accord avec les spécifications d'OAI-PMH, ces identifiants sont formatés selon les principes de l'URI définis par l'IETF. Les données collectées doivent donc être indexées en réutilisant ces identifiants.

Concernant les données produites, celles-ci sont instanciées pendant leur traitement par les chaînes X5-GON, et à l'issue de ce traitement des données collectées elles peuplent des graphes de connaissances au format RDF, en assignant là encore des URIs aux jeux de résultats obtenus.

6. Responsabilités et ressources en matière de gestion des données

Collection de RELs

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 responsable scientifique du projet est Colin de la Higuera.

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) ?

Les moyens alloués sont ceux du projet, puisqu'in fine le projet consiste justement à rendre FAIR les données.

Précisons :

(F). Le but est de permettre à un moteur de recherche multilingue de trouver les données. De plus, ce moteur a vocation à être multi-plateforme plutôt qu'interne. Par conséquent, l'objectif est bien de rendre les données faciles à trouver.

(A). L'accessibilité a deux sens. La question de l'accessibilité à tous est primordiale et au cœur de nos préoccupations. Ainsi, dans le contexte du projet X5-GON, nos partenaires slovènes ont développé l'outil X5-Blind dont l'objectif est de permettre aux malvoyants d'avoir accès aux mêmes outils. De plus, nous cherchons systématiquement des solutions pour permettre l'accès au plus grand nombre, et, dans le cadre de nos partenariats, nous travaillons avec de nombreux pays, en particulier du *Global South*.

(I). L'interopérabilité est assurée par des solutions dynamiques basées sur des APIs ouvertes.

(R). La réutilisabilité est au cœur de la définition des ressources éducatives libres. Un enseignant doit pouvoir reprendre un cours et le transformer. C'est là aussi une priorité du projet