Qu'est-ce que le Data Monitor ?

Novembre 2023, GTSO Données de Couperin

Auteurs : Meriç Akdogan, Sorbonne Université et Laetitia Bracco, Université de Lorraine

Table des matières

I. Contexte

Cette note a été commandée au GTSO Données par le bureau professionnel de Couperin, suite à la stratégie de communication opérée par Elsevier auprès des ateliers de la donnée et des établissements pour vendre la solution commerciale Data Monitor.

II. Qu'est-ce que le Data Monitor (DM) ?

DM se fixe pour objectif de constituer, à l'échelle d'une institution, un catalogue de jeux de données déposés dans 2000 entrepôts différents. D'après le site internet de l'outil, DM :

- « Identifie et récolte automatiquement les métadonnées de jeux de données ;
- Nettoie et enrichit les enregistrements ;
- S'intègre avec les CRIS et les SI de l'établissement ;
- Permet de contrôler la conformité avec les politiques nationales, des financeurs, des éditeurs ou des institutions ;
- Fournit des informations sur la politique institutionnelle en matière de science ouverte et de données et améliore la prise en charge du cycle de vie des données pour les chercheurs".

Deux établissements en Europe ont à notre connaissance mis en place publiquement cet outil :

→ L'Université de Groningen

→ L'Université de Milano-Bicocca : Open Archive Research Data

En France, l'Université Paris-Saclay utilise également cet outil.

L'Université de Groningen (Pays-Bas) a participé à la phase pilote de l'outil, lancée en 2019, aux côtés de la Vrije Universiteit d'Amsterdam, de l'Université du Sud Danemark et des universités belges Vrije Universiteit de Bruxelles et de Namur.

Groningen utilise Pure, un système d'information pour la recherche d'Elsevier, auquel peut être intégré DM. Une analyse commanditée par Elsevier sur la dissémination des jeux de données d'une institution dans différents entrepôts (Analysis of research data for 11 Institutions - Data Monitor) suggère que 90% des jeux de données d'une université sont en général déposés en-dehors de l'entrepôt institutionnel. Il est intéressant de noter que c'est l'API de DM qui a été utilisée dans cette analyse réalisée pour justifier l'utilité de DM : "*For each institution, we counted the number of datasets published in their Institutional Data Repository (IDR) and tracked the number of public research datasets hosted in external data repositories via the Data Monitor API*". Ce chiffre n'est cependant pas étonnant puisque la bonne pratique est en effet de déposer en priorité dans un entrepôt disciplinaire adapté puis seulement dans un entrepôt institutionnel si rien d'autre n'est disponible.

III. Retour de l'Université de Milano-Bicocca

L'Université de Milano-Bicocca a contacté Elsevier en 2020 après avoir réalisé une analyse comparative de différentes solutions de gestion des données de recherche (*Research data management*, RDM). Dans le cadre du processus d'évaluation, ils ont défini des paramètres pour évaluer les principaux logiciels de RDM et ont demandé des informations auprès de différents référents. Un paramètre essentiel était la capacité de moissonner les métadonnées des jeux de données à partir d'autres entrepôts de données, et le module Data Monitor d'Elsevier a fourni une solution appropriée à cette exigence.

La motivation qui a conduit leur choix vers Digital Commons Data (Mendeley Data, un entrepôt de données alimenté par Digital Commons Data) et par conséquent vers Data Monitor, était principalement due à la possibilité de moissonner les métadonnées des jeux de données déposés (et liés aux auteurs affiliés à Bicocca) sur d'autres entrepôts de données, et de les publier sur leur plateforme. Parmi les options disponibles en 2020, seul le produit d'Elsevier offrait cette fonctionnalité à un coût raisonnable. Il n'est pas possible d'avoir cette fonctionnalité de moissonnage sans avoir Data Monitor.

Concernant les résultats, ils sont partiellement satisfaits des services fournis. D'une part, Data Monitor a réussi à moissonner et publier avec succès des milliers de ressources affiliées à Bicocca, répondant à leurs attentes. Cependant, ils ont rencontré des difficultés avec le support et l'assistance fournis par Elsevier, car plusieurs problèmes signalés, tels que des erreurs dans les affiliations, y compris ceux liés à Data Monitor et Digital Commons Data, restent sans solution depuis plusieurs mois.

En ce qui concerne la correction des erreurs, la version précédente de Data Monitor leur permettait d'exclure manuellement les jeux de données affiliés incorrectement à l'Université de Milano-Bicocca. Cependant, la version actuelle ne dispose pas de cette fonctionnalité, et ils ne peuvent pas enrichir les ressources moissonnées en ajoutant des métadonnées ou en apportant des corrections.

Pendant le processus d'implémentation de l'outil, ils ont initialement fourni à Elsevier une liste des informations sur les utilisateurs de l'Université de Milano-Bicocca, y compris leurs noms, prénoms, adresses e-mail et identifiants ORCID. Le module Data Monitor était déjà intégré à la solution Mendeley Data lors du lancement de l'entrepôt.

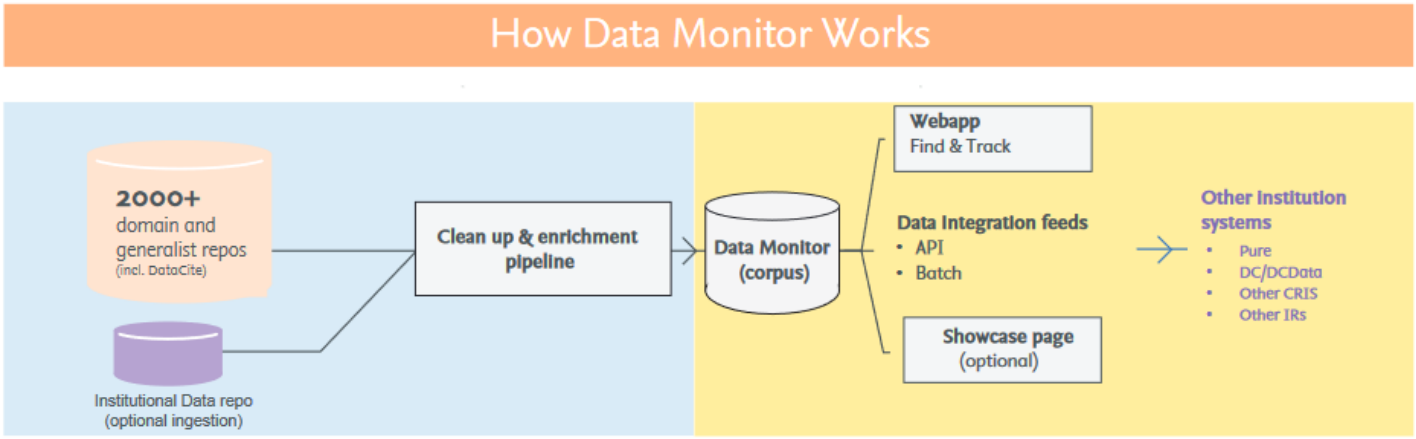
L'Université de Milano-Bicocca utilise les résultats de Data Monitor pour organiser les jeux de données moissonnés en collections départementales, permettant d'avoir un aperçu du nombre de ressources affiliées à chaque département. Cependant, l'ajout de ressources à une collection spécifique ne peut être effectué que manuellement. Ils génèrent également des rapports internes, tels que des statistiques sur les vues et les téléchargements, en utilisant des APIs et des processus automatisés avec les informations fournies par Data Monitor.

Bien que Data Monitor ne soit pas actuellement utilisé à des fins de pilotage au sein de leur organisation en raison de sa relative immaturité, ils reconnaissent son potentiel et espèrent qu'il deviendra plus précieux dans leur travail quotidien à l'avenir. En ce qui concerne le coût, ils ont récemment renouvelé le contrat avec Elsevier pour une année supplémentaire. Le coût annuel du module Data Monitor s'élève à environ 10 000 €.

Sur le catalogue public de l'Université Bicocca, on peut constater quelques problèmes, tels que des doublons, des erreurs 404, des erreurs d'attribution et des erreurs sur les types de données identifiés.

IV. Comment fonctionne l'outil ?

Peu de détails techniques sont disponibles sur le site. On peut néanmoins trouver ce schéma :



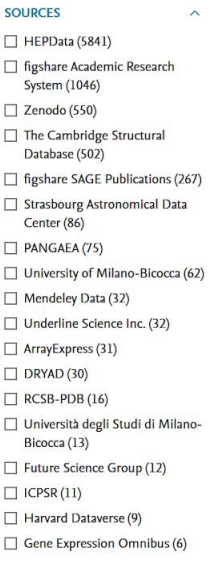
How we track the Research Data

- Harvest metadata from 2000+ repos
  - Optional: ingest the Institutional Data Repository
- Normalize metadata (OpenAIRE schema)
- Clean-up i.e. remove duplicates, non-research data e.g. articles
  - How: ML techniques, integration w/ Scopus, experts
- Enrich metadata with publications/ author / institution
  - How: NLP techniques; integration w/ Scopus, Scholix, DataCite

What users see and get:

- Webapp
- Data Integration feeds
  - API
  - Bulk
- Optional: Showcase page

L'exemple du portail de l'Université Milano-Bicocca nous permet de comprendre certains éléments : DM a été intégré à leur entrepôt de données institutionnel, ce dernier fonctionnant avec la solution Digital Commons Data, dont dépend également Mendeley Data, un autre produit d'Elsevier.



L'interface permet de chercher par identifiant Scopus, par financement, par DOI…

DM affirme être en mesure d'extraire des métadonnées de plus de 2000 "entrepôts" différents, puis de normaliser et dédoublonner les références. Elles sont ensuite enrichies avec des affiliations issues de Scopus. Enfin, les données peuvent être exposées sur l'application web DM ou bien intégrées à un autre système via l'API, comme le fait Bicocca dans Digital Commons Data. On peut voir ici un exemple de sources agrégées dans leur portail. **DM est ainsi capable de rapatrier des jeux de données n'ayant pas de DOI, mais un numéro d'accession**, par exemple celui-ci : FUS KO mRNA sequencing and anti-FUS RNA immunoprecipitation sequencing.

Une fonctionnalité de preview pour certaines sources dans l'entrepôt est disponible, même pour les jeux de données déposés ailleurs.

Dans sa Content and Data Policy, Elsevier indique que DM indexe les entrepôts généralistes, disciplinaires et institutionnels directement ou en moissonnant les métadonnées que ces entrepôts mettent à disposition par l'intermédiaire de DataCite.

V. Cet outil est-il performant ?

Le DM est performant pour repérer des données déposées dans des entrepôts variés, parfois peu connus. Il indexe Recherche Data Gouv au niveau du fichier et pas du jeu de données ; de ce fait, il affiche un total de près de 44 000 jeux de données alors qu'il n'y en n'a en réalité qu'environ 3 500. Il gère mal les fusions de laboratoires. Il s'avère

qu'une recherche sur un nom d'établissement donne de mauvais résultats ; seule la recherche par ID Scopus est pertinente pour établir la liste des jeux de données d'une unité de recherche. Ceci renforce la dépendance aux produits Elsevier, puisqu'un établissement ne disposant pas de Scopus aura des données de moins bonne qualité. En outre, les jeux de données ne sont pas validés par les établissements et les délais pour corriger des erreurs sont très longs.

VI. Quels acteurs français sont concernés ?

Le Data Monitor se place dans un contexte comprenant deux acteurs majeurs :

- Le Baromètre français de la Science Ouverte (BSO) dédié aux données de la recherche et aux logiciels, dont la vocation est de proposer en 2024 aux établissements des listes de jeux de données de leur institution avec des indicateurs d'ouverture.
- Le module Catalogue de Recherche Data Gouv, dont la mission est de moissonner les métadonnées de jeux de données déposés ailleurs que dans l'entrepôt national.

VII. L'avis du GTSO Données

L'utilisation de Data Monitor, couplée à la galaxie de produits Elsevier tels que Pure et Digital Commons, mène inévitablement à une dépendance forte à des produits commerciaux sur lesquels nous n'avons ni visibilité, ni souveraineté, qui nous amènera à réaliser des dépenses mais aussi à investir du temps en efforts de curation qui sont ensuite propriétarisés comme nous le faisons pour des bases bibliographiques. **A ce stade pour le GTSO Données, l'intérêt d'une souscription à cet outil semble très limité, d'autant plus qu'elle amènerait les établissements à payer pour un service propriétaire alors qu'une solution gratuite, ouverte, institutionnelle et intégrée à l'écosystème français est déjà en développement et sera prochainement disponible**.