Data paper - Une incitation à la qualification et à la réutilisation des jeux de données

Synthèse webinaire du 5 novembre
Quels enseignements - Quelles perspectives ?

Le GT Inter réseau « Atelier Données » a organisé le 5 juin 2020 un webinaire intitulé : Data paper - Une incitation à la qualification et à la réutilisation des jeux de données.

Ce webinaire avait pour objectif d'aborder diverses questions autour du data paper pour nous permettre de mieux comprendre les pratiques de recherche qui s'y rattachent et la place qu'occupe aujourd'hui (et peut-être encore demain) ce nouveau modèle de publication dans l'écosystème de la donnée.

La matinée, introduite par Stéphane Renault[1] s'est organisée en deux parties. La première session a débuté avec une intervention de Sophie Pamerlon[2] qui a présenté le workflow et les outils intégrés de publication de données du GBIF. Sa présentation a été illustrée d'un retour d'expérience d'Annegret Nicolai[3], utilisatrice des outils GBIF pour la publication de jeux de données collectés à l'occasion d'un bioblitz et la rédaction de son datapaper.

En seconde partie, Denise Pumain[4] et Clémentine Cottineau[5] ont apporté un éclairage différent en abordant le sujet du point de vue de l'éditeur. Elles ont présenté le processus d'évaluation des data papers mis en place au sein de la revue Cybergeo. Le dernier intervenant, Victor Gay[6] a apporté le témoignage d'un producteur de data paper. Il a présenté en tant que chercheur sa démarche vers la valorisation des données issues de sa recherche sur l'écosystème TRF-GIS et les raisons qui l'ont motivée.

Pour finir, Joachim Schöpfel[7], invité en tant que Grand Témoin de ce webinaire a partagé ses observations et formulé quelques perspectives de progrès.

Ce document synthétise autour de 5 grandes parties les principaux enseignements et questionnements soulevés lors des interventions de cette

1 Stéphane Renault, Responsable éditorial, CNRS, LAMPEA, Réseau Medici
2 Sophie Pamerlon, Gestionnaire de données Biodiversité, GBIF France – USM Patrimoine naturel
3 Annegret Nicolai, Chercheuse, Univ. Rennes 1 – UMR ECOBIO, Station Biologique de Paimpont
4 Denise Pumain, Professeure de géographie et co-fondatrice de la revue Cybergeo, Univ. Paris 1 – UMR Géographie-Cités
5 Clémentine Cottineau, Géographe et responsable de la rubrique datapaper de la revue Cybergeo, CNRS – Centre Maurice Halbwachs
6 Victor Gay, Chercheur, Univ. Toulouse 1 – École d'Économie de Toulouse
7 Joachim Schöpfel, Chercheur, Université Lille 3 – GERiiCO

matinée et tente de rendre compte de la richesse et de la teneur des échanges entre les participants.

FONCTION, USAGE DU DATA PAPER ET RÉUTILISATION DES DONNÉES

Tout au long de ce webinaire, la fonction et l'usage du data paper ont été évoqués, débattus et questionnés.

Sophie Pamerlon a présenté le data paper comme une publication scientifique dont le but est de décrire un jeu de données ou un ensemble de jeux de données plutôt que de rendre compte d'analyses et de résultats de recherche. L'observation des pratiques actuelles montre toutefois que la distinction entre data paper et article scientifique ne va pas toujours de soi.

Les débats et questionnements ont montré que le data paper trouve sa place dans un large **processus de publication scientifique** et d'ouverture des données et peut se situer en conflit ou en complément avec la rédaction d'un article scientifique classique, l'ajout de « «complementary data » pour l'analyse des données ou même être perçu comme redondant avec la rédaction d'outils de planification comme le **Data Management Plan** (DMP).

Le data paper a été défini également comme un moyen de mettre en valeur les jeux de données, mais qu'en est-il de la définition du jeu de données ? Quelles données faut-il déposer (froides, tièdes, brutes, retravaillées etc.) ? Comment sélectionne-t-on un jeu de données et à quel niveau de granularité ? Tous les jeux de données ont-ils vocation à figurer dans un data paper ? Quel est le véritable objet du data paper in fine ? Faut-il y décrire le contenant (une base de données, une collection), le contenu (le jeu de données), les deux ?

Si des recommandations et spécifications strictes et précises sont énoncées par certaines revues comme Cybergeo, il ne semble pas, selon Joachim Schopfel, y avoir véritablement de consensus en la matière. On ne peut que constater une grande diversité en fonction des communautés, des pratiques, des thématiques, des équipements, infrastructures et des méthodes.

Des questions peuvent se poser également quant à l'orientation possiblement différente des objectifs du data paper en fonction de qui en assure la rédaction : un *producteur* de données ou un *consommateur* de données. Le principe de réutilisation des données n'implique-t-il pas la création d'un nouveau data paper ?

En général, lorsqu'elles sont réutilisées, les données font l'objet d'un article scientifique et pas d'un data paper. C'est logiquement au *producteur* d'une donnée qu'il revient d'en assurer la description et non au *consommateur*. Si toutefois les données sont réintégrées à une autre recherche, complétées et enrichies par d'autres jeux, la transformation des données initiale peut probablement engendrer une autre pratique méthodologique qui justifierait la rédaction d'une nouveau data paper.

Les retours d'expériences présentés par Annegret Nicolai et Victor Gay ont montré que le principal avantage du data paper réside surtout dans la mise à disposition et l'accessibilité pour les chercheurs de jeux de **données structurées, compréhensibles et reexploitables**. Ces données constituent un véritable réservoir de matériaux disponibles pour la recherche sur un territoire pour une thématique singulière.

Annegret Nicolai a, en particulier, montré la variété des données principales (dans le domaine de la biodiversité) et annexes (exploitables par des géographes, sociologues, physiciens, chimistes etc.) recueillies lors de son inventaire éclair de la biodiversité (bioblitz). Ces données une fois documentées et partagées offrent aux futurs utilisateurs des éléments riches redéployables dans différents domaines scientifiques pour différents usages.

La visibilité des données est au cœur de la problématique. La citation du jeu de données via l'attribution de DOI de même que celle de l'article de recherche et du data paper (et le renvoi des uns vers les autres) prend ici tout son sens. Toutefois, au-delà de l'accessibilité et de la réutilisation des données, se pose aussi la question de pouvoir identifier plus précisément les usages et profils d'utilisateurs (cercle restreint de collaborateurs ou public plus large et inconnu).

Si l'on en croit le retour de Victor Gay, le manque de données disponibles et de **nomenclatures interopérables** dans certains domaines de la connaissance (*comme c'est le cas pour la recherche empirique sur la France de la 3ème république par exemple*) constitue une incitation forte pour les chercheurs à s'orienter vers la rédaction d'un data paper et un véritable **gain de temps** dès lors qu'ils peuvent en disposer.

Il semble que le data paper soit aussi une réponse à la **crise de reproductibilité** et au manque de valorisation du travail effectué sur les données elles-mêmes.

DES OUTILS ET MÉTHODES POUR ACCOMPAGNER LA RÉDACTION DE DATA PAPERS

Il faut savoir que des outils existent pour accompagner la publication des données et appuyer la rédaction et la soumission d'un data paper.

Dans certains domaines disciplinaires, des politiques se sont constituées pour mettre à disposition des chercheurs des **protocoles et outils pour la gestion**, le partage et la valorisation des données. Sophie Pamerlon a ainsi présenté les outils (IPT[8] et Arpha Writing Tool) et le processus déployé au GBIF pour publier les données et métadonnées de la biodiversité selon les principes FAIR.

Le travail de remplissage des métadonnées et la production automatisée d'un manuscrit de data paper est facilité par l'usage d'un outil comme IPT (Integrated publishing Tool).
Comment cela fonctionne ? Un grand nombre de catégories sont proposées (non obligatoires) pour accompagner la **curation** et même si cette phase peut

8 Integrated Publishing Tool

paraître fastidieuse, le temps passé à la tâche n'est plus à reproduire au moment de la rédaction du data paper.

Le processus nécessite d'appliquer des **standards d'échange** (pour le formatage des données) qui dépendent fortement de la discipline et de la nature des données : Darwin Core, ABC par exemple pour les données d'occurrences, les données taxonomiques ou d'échantillonnage, EML (Ecological Metadata Langage) pour la description des données et métadonnées.

Une fois générées et publiées dans l'IPT, les données (et métadonnées) sont moissonnées, elles apparaissent sur le GBIF et sont réutilisables par tous. Un fichier EML est accessible et pourra être soumis à une revue d'édition pour évaluation. Il va de soi que plusieurs revues proposent un appui technique et des outils automatisés pour aider à la rédaction.

L'outil Arpha, quant à lui, facilite la mise en page, la soumission, le processus de relecture, la publication, l'hébergement et l'archivage d'articles scientifiques.
Deux voies de publications sont envisageables – il est possible de choisir la voie « *par document texte* » en début de rédaction ou choisir de « *récupérer le fichier EML.xml* » qui, une fois intégré, va proposer un brouillon de data paper.
Dans ce workflow, il faut souligner que les outils sont interopérables entre eux, mais pas seulement, des liens sont aussi possibles avec Zenodo ou d'autres entrepôts et le **formatage des citations** est facilité.

Annegret Nicolai a utilisé les outils du GBIF pour publier les données collectées pour sa recherche (inventaire de la biodiversité : 660 espèces, 1819 occurrences) à la station biologique de Paimpont. En retour d'expérience, elle témoigne de la souplesse de l'outil qui offre la possibilité de pointer certains protocoles utilisés via l'intégration de schémas et de cartographies, ce qui représente un réel avantage (et permet par exemple un focus sur une série de temps avec 77 espèces d'oiseaux apparu et disparus ces 60 dernières années).

Elle signale en revanche la difficulté parfois à appliquer **les référentiels** proposés qui peuvent s'avérer inappropriés (certaines espèces se sont avérées difficiles à identifier car non reconnues dans la taxonomie) ou la difficulté à gérer des problèmes d'accentuation qui ont généré du travail supplémentaire et la **perte de certaines occurrences.**

On constate ici l'importance que revêt l'aspect normatif des données dans la publication d'un data paper et l'enjeu que représente la FAIRisation des données. On peut toutefois s'interroger, à l'instar des participants au webinaire, sur la cible d'une « normalisation » ou « FAIRisation » efficace ? Est-il préférable de FAIRiser les données, l'entrepôt de données, ou les articles de données ?

Quoi qu'il en soit la question de la **confiance** dans la réutilisation d'un jeu de données est essentielle. Il s'agit véritablement de rendre les données fiables, crédibles et dignes d'intérêt. Les principes FAIR sont là pour guider le processus. Les entrepôts doivent également être soigneusement étudiés et les démarches vers leur **certification** (bien qu'encore balbutiantes) en facilitent le choix et l'acceptation.

Plus que pour un article classique, on observe, avec la présentation de Sophie Pamerlon, à quel point le data paper s'intègre dans un **écosystème** plus large

qui va des infrastructures de recherche vers les entrepôts de données des plateformes de revues. La question du workflow et de l'interopérabilité des outils, propriétaires (pensoft) ou ouvert (dataverse) se pose. Il est fort à parier que des **modèles communautaires** vont se développer autour d'une thématique, d'un réseau, de certains outils connus et utilisés.

QUELQUES CONSEILS À RETENIR AVANT DE PUBLIER UN DATA PAPER

Le projet de recherche de Victor Gay a porté sur la création d'un système d'information géographique (SIG) pour la France de la 3ème république. Il a constitué pour cela une grande base de données qui propose des nomenclatures et shapefiles[9] annuels de 1870 à 1940 pour un ensemble de circonscriptions administratives (Il s'agit au total de la constitution de 16 bases de données, 901 nomenclatures et 830 shapefiles). A ce travail s'ajoute la mise à disposition de l'ensemble du matériel de reproduction (code source, données source, archives consultées), le tout accessible sur dataverse.

De manière synthétique, sa démarche vers la publication d'un data paper a montré l'importance :

- **D'accorder un soin particulier à la méthodologie de construction et au choix de catégorisation**,

Il faut selon lui pouvoir montrer les limites et la valeur des données par rapport à l'état de l'art (certaines bases de données sont créées dans un but défini et cela doit être précisé). Les éléments techniques doivent être décrits et le potentiel de réutilisation exprimé.

- **De travailler de manière simultanée à la construction de la base de donnée *et* à la rédaction du data paper,**

Il faut réfléchir au moment de la conception de la base de donnée à la manière dont elle sera présentée dans le data paper et réfléchir à la constitution de la base de donnée au moment de la rédaction du data paper. C'est un travail conséquent.

- **De se préparer très en amont du projet à la FAIRisation des données** …

Avant même la constitution d'une base de données, Victor Gay recommande de déterminer le choix des métadonnées (DDI, Dublin Core, DataCite …), du vocabulaire contrôlé (Library of Congrès …), des identifiants persistants, des formats ouverts, standards répandus (COB, INSEE, GEOFLA IGN…), licence appropriée … sachant que certains de ces éléments sont gérés différemment selon les entrepôts.

- **… Et à la dissémination et la valorisation des données**

Il convient aussi de se poser la question du format et du lectorat approprié. Le choix du dépôt de donnée est particulièrement important (il faut pouvoir

9 Le Shapefile est un format de fichier pour les systèmes d'information géographique. Il contient des informations liées à la géométrie des objets décrits (points, lignes, polygones)

examiner les critères, besoins, usages, respect des principes FAIRs) mais il faut étudier également les capacités de stockage offertes. La structure de la base (*13 000 fichiers en tout*) peut potentiellement poser problème si elle est complexe et que l'entrepôt ne dispose pas d'une ergonomie efficace pour le producteur ou l'utilisateur.

- **Sans oublier le rôle crucial de la formation !**

La rédaction d'un data paper suppose l'acquisition d'un large spectre de connaissances bien souvent inconnu et peu diffusé dans les écoles doctorales il y a encore deux ans. Il n'existe que peu d'appui local ou de structures d'incitation. La tache peut s'avérer complexe malgré une démarche proactive des personnels d'accompagnement et le suivi de formations continues (URFIST, MSHS-T DoRANum).

Le Rôle des Ingénieurs et Techniciens (IT) dans la rédaction du data paper

Conformément aux souhaits du comité d'organisation, les interventions et nombreux échanges au cours de ce webinaire ont permis de clarifier le rôle des ingénieurs et techniciens dans la rédaction des data papers.

Au-delà de la mise à disposition des données, le data paper est un gage de reconnaissance pour tous les auteurs et acteurs de la recherche menée (le partenariat Gbif et Pensof a d'ailleurs été initialement motivé par la volonté de rendre compte du temps et de l'investissement des personnels dans l'informatisation des données et la construction des bases de données).

Mais le data paper peut-il avoir comme fonction de rendre compte de l'investissement des ingénieurs et techniciens qui œuvrent au côté des chercheurs ? Un travail conjoint de rédaction est-il envisageable ?

Selon Victor Gay, certaines problématiques disciplinaires en terme d'usage et de diffusion peuvent rendre la tâche difficile, les structures d'incitation et les responsabilités peuvent ne pas s'avérer compatibles. Selon Denise Pumain en revanche, les IT en sciences sociales ont toujours été étroitement associés à la recherche et au travail sur les données qui suppose une implication forte de leur part pour apporter du sens et construire les données. Les IT signent et co-signent dans la revue Cybergeo et rien n'interdit à un ingénieur de publier des analyses ou articles techniques.

Si tel est le cas, peut-on alors considérer légitime la participation d'un IT dans un comité de lecture ? Clémentine Cottineau assure que le statut de l'évaluateur au sein de la revue Cybergeo n'est pas un critère déterminant dans le choix d'un comité d'évaluation, il s'avère qu'au moins 20% des évaluateurs du dernier comité étaient des ingénieurs et des techniciens**.**

Pour Joachim Schöpfel, le besoin de travailler ensemble et avec les IT n'est plus à démontrer. Des interrogations toutefois persistent autour de la définition des métiers, fonctions, activités et compétences qui par ailleurs vont de pair avec la mise en place d'incitations ou de gages de reconnaissance.

QU'EN EST-IL DE L'ÉVALUATION DU DATA PAPER ?

Comment se déroule le processus d'évaluation d'un data paper ?

La revue Cybergeo est une revue de géographie en ligne crée en 1996, qui s'inscrit dans la mouvance de l'Open Science et qui a initié en 2017 une rubrique dédiée à la publication de data paper.

Dans cette revue, libre, gratuite et bilingue, le data paper est évalué par un comité scientifique comprenant au moins un évaluateur sur le thème et un sur les données. Il est composé de géographes, statisticiens et ingénieurs mais en fonction des besoins et des thèmes il peut être décidé de recourir à des relecteurs extérieurs.

Le processus de peer reviewing ne consiste pas à évaluer une simple description du jeu de données. L'évaluation porte principalement sur l'argumentation autour de la construction du jeu et comment il peut être réutilisable par les auteurs dans le futur.

Les consignes aux auteurs sont rédigées dans le souci du partage, de la transparence des données et de la reproductibilité des analyses. Aussi, il est demandé aux auteurs un travail rigoureux pour décrire et documenter les données, préciser les spécifications liées aux formats particuliers des données géographiques (les échelles, les composantes spatiales et scalaires, géolocalisation …), la compatibilité avec les logiciels etc.

L'ensemble des procédures et processus de validation doivent être justifiés, de même que la chaine de traitement et les choix opérés. Les auteurs doivent expliquer comment ils ont utilisé, transformé et agrégé leurs sources pour produire la base de données. Ainsi, il semble clair que l**a revue évalue le processus de validation des données plutôt qu'elle n'évalue des données elles-mêmes.**

Au sein de Cybergeo, la publication des données précède la soumission des articles. Le dépôt est à l'initiative de l'auteur mais il doit être pérenne, ouvert et de préférence institutionnel.

Le processus d'évaluation de Cybergo pose aussi un certain nombre de difficultés :

L'évaluation « *dubble blind* » n'est pas vraiment compatible avec le dépôt des données car il est difficile dans ces conditions de préserver l'anonymat. La revue autorise par ailleurs la double soumission d'un data paper et d'un article détaillant une analyse issue des jeux de données. Certains auteurs sont alors tentés de détailler la construction des données dans l'article classique et de soumettre pour le data paper une liste de métadonnées comme dans d'autres disciplines ou d'autres revues.

Pour Joachim Schöpfel, la question de l'évaluation est un enjeu majeur et le fonctionnement de Cybergeo apparaît ici relativement plus ouvert que celui des revues scientifiques classiques. Le principe d'évaluation pose néanmoins la question cruciale de la « matière » évaluée : data paper ? données ? dépôt des

données ? métadonnées ? A déposer dans un environnement ouvert ? fermé ?

Il pose aussi la question de la meilleure façon d'évaluer ? En double aveugle ? Dans l'anonymat ? Par qui ? Par les pairs ? Mais qui sont les pairs ? Des ingénieurs et techniciens ou des chercheurs ? Les deux ? Qui doit proposer un data paper ? La communauté ou les usagers ? Pourquoi ne pas laisser le soin d'évaluer les données et data paper en aval par d'autres, les usagers ?

Mais peut-on se permettre de publier sans se soucier du **prestige de la revue** ? Publier un data paper dans une revue de haut rang n'est-il pas important pour l'évaluation du chercheur ?
En réponse à cette question, Denise Pumain précise qu'en SHS le rang d'une revue a moins de signification que dans d'autres domaines, car il est assez difficile et souvent peu pertinent d'établir des hiérarchies entre les revues dont les thèmes sont plus ou moins spécialisés avec une grande diversité culturelle et locale. Même si de « grands éditeurs » avec facteurs d'impacts existent, les indicateurs ne représentent pas suffisamment bien la qualité des revues et les membres des comités nationaux d'évaluation sont très défiants vis-à-vis d'une bibliométrie purement quantitative.

Ce qui compte pour les auteurs, c'est véritablement la procédure d'évaluation, la qualité des avis reçus et la visibilité conférée à une publication (visibilité non démentie par l'audience et le taux de téléchargement des articles de Cybergeo qui s'avère aussi important que n'importe lequel des ranking internationaux)

A ce sujet, Joachim Schöpfel nous interpelle sur l'importance du débat autour des **mesures d'impacts** dont on connaît finalement assez mal la réalité (nombre de citations, de téléchargement etc.). Il semble que pour Elsevier, le taux de consultation ou de téléchargement des données publiées dans *Mendeley Data* n'est pas corrélé par exemple avec la richesse des métadonnées des jeux présentés.

Faut-il accorder autant d'importance à l'impact du data paper dans l'évaluation d'une recherche ? Ne doit-on pas plutôt chercher à valoriser les données que pointe l'article, plutôt que l'article lui-même qui n'est qu'un simple outil de valorisation ?

En présentant la rubrique data paper de la revue Cybergeo, Denise Pumain a présenté un modèle économique basé sur le mode « Freemium », un modèle selon Joachim Schöpfel assez inhabituel dans le paysage de l'économie des données. Les données ont une valeur économique indéniable et les revues et data journals fonctionnent généralement selon un modèle GOLD avec APC. Ce modèle économique génère auprès des éditeurs un chiffre d'affaire déjà assez conséquent et la question peut se poser de savoir si un risque de « *predatory publishing* » pour les data papers est à prévoir dans un avenir plus ou moins lointain.

ET POUR CONCLURE

La conclusion de notre Grand Témoin, en fin de webinaire, nous oriente vers des perspectives diverses mais encore difficilement prédictibles. Une augmentation

du nombre de data papers (potentiellement absorbés par des journaux réguliers ou des data journals) est envisageable tout comme un rapprochement des plateformes de publications et des données.

On peut aussi s'attendre à voir se développer la génération automatique d'articles depuis les entrepôts.
Est-ce que cela concernera tous les domaines scientifiques ou uniquement certains ? Les années à venir nous le diront. Il n'en demeure pas moins que l'on constate aujourd'hui un intérêt (politique, économique, professionnel et scientifique) croissant pour la gestion des données et pour le data paper qui s'inscrit dans une réalité dynamique, en pleine évolution.