

# Data paper

*Une incitation à la qualification et à la réutilisation des jeux de données*

Synthèse, questions

Joachim Schöpfel

Joachim Schöpfel, Dominic Farace, Hélène Prost, Antonella Zane. Data papers as a new form of knowledge organization in the field of research data. *12ème Colloque international d'ISKO-France : Données et mégadonnées ouvertes en SHS : de nouveaux enjeux pour l'état et l'organisation des connaissances ?*, ISKO France, Oct 2019, Montpellier, France. <https://halshs.archives-ouvertes.fr/halshs-02284548>

# Un intérêt croissant

*Partenaires du GT inter-réseaux - Atelier Données*

Renatis, Médici, QeR, Resinfo, BDD, Calcul, Devlog, SIST, DIST-CNRS



Le Réseau Calcul



Le Réseau DEVlog



Le Réseau Medici



Le Réseau Resinfo



Le Réseau SIST de l'INSU



Le Réseau Renatis



La Direction de l'Information  
Scientifique et Technique du  
CNRS



Le Réseau QeR

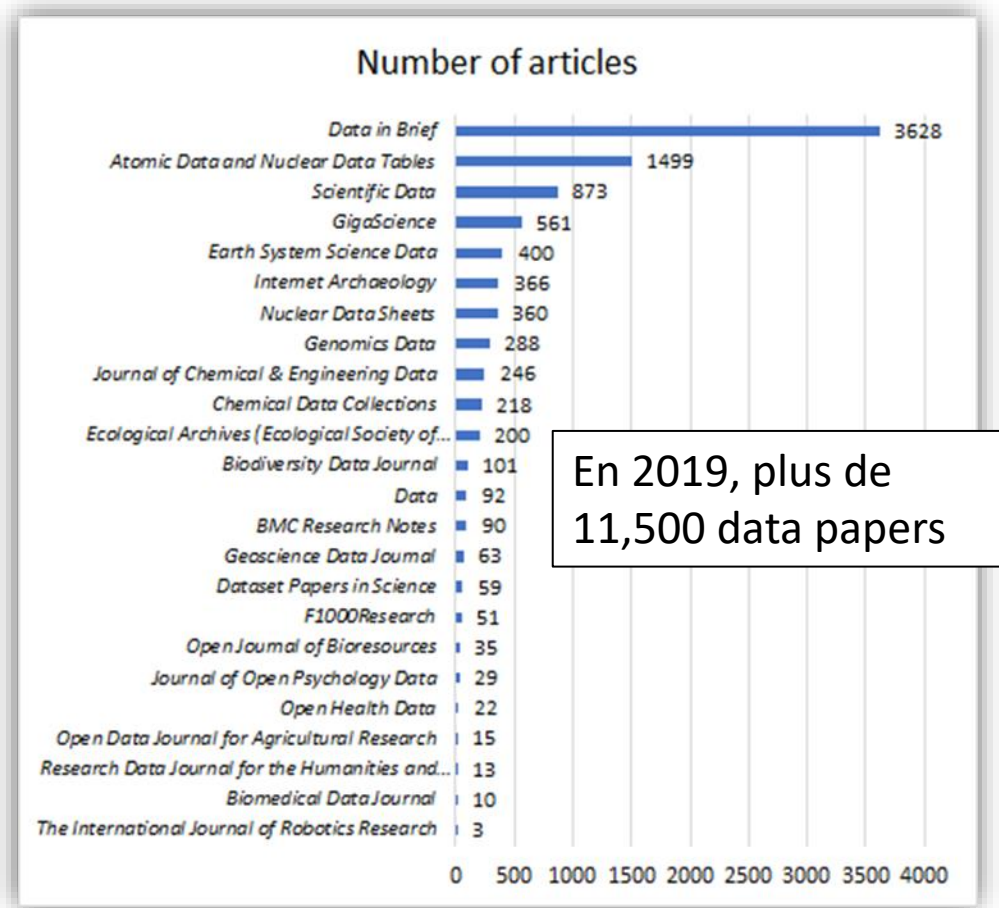


Le Réseau Rbdd

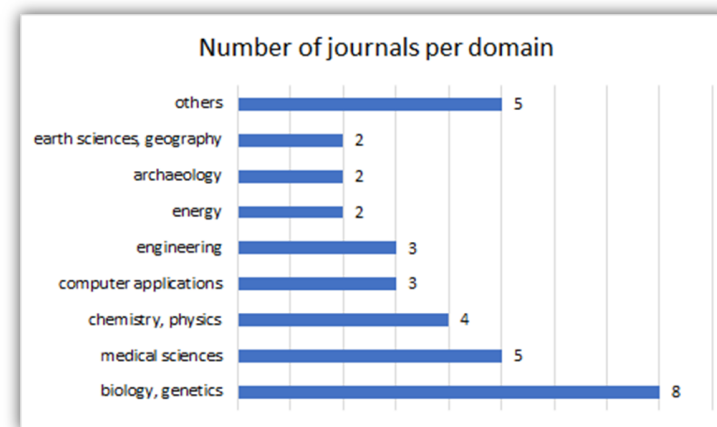
- Pour la gestion des données
- Pour les articles de données
- Politique, économique, professionnel, scientifique

→ Dans le cadre du soutien public aux revues, recommander l'adoption d'une politique de données ouvertes associées aux articles, le développement des articles de données et des revues de données.

# Une réalité en évolution



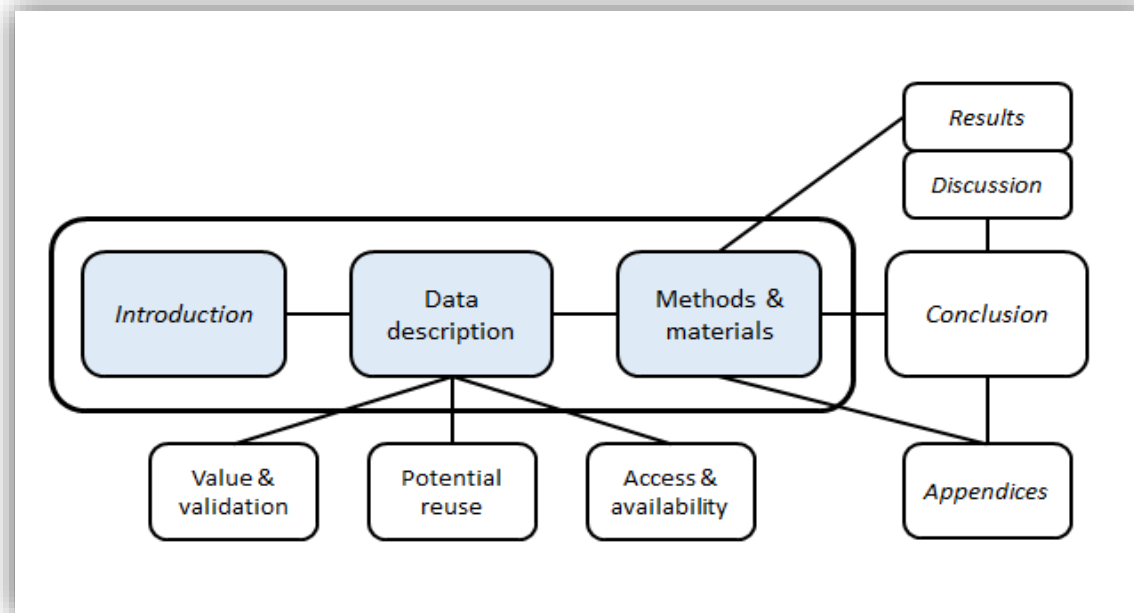
- Un type d'article relativement récent
- Un nombre d'articles en augmentation constante
- Mais peu d'articles (<0.1%)
- Surtout STM



# Une grande diversité

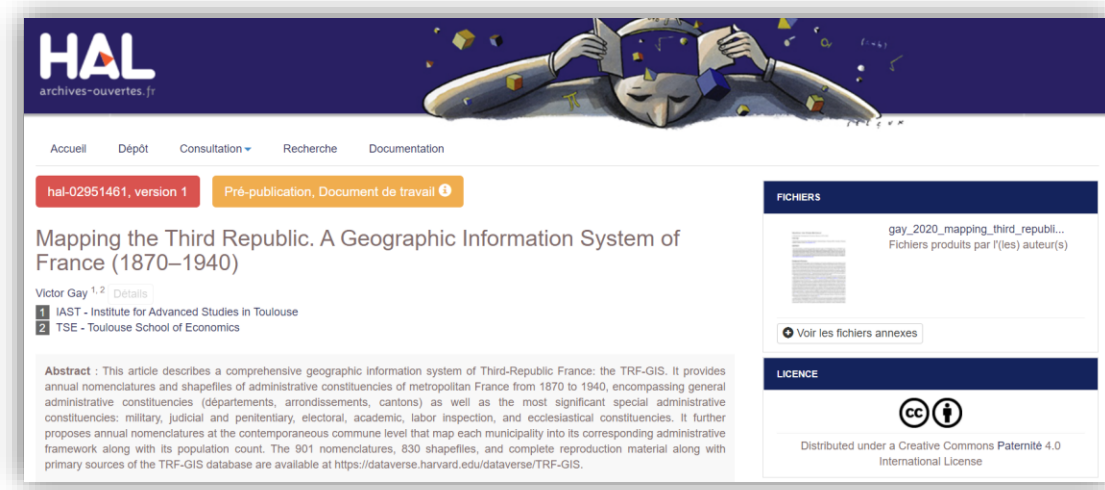
La définition la plus simple :

- *“information on the what, where, why, how and who of the data”*
- Des “contenus coeur”



| Les normes de présentation de votre article                   |  |
|---------------------------------------------------------------|--|
| Les métadonnées                                               |  |
| Le titre                                                      |  |
| Le résumé                                                     |  |
| Les mots clés                                                 |  |
| Le texte                                                      |  |
| Les images et les fichiers supplémentaires                    |  |
| Les images                                                    |  |
| Rappels (carto)graphiques                                     |  |
| Le titre de vos illustrations                                 |  |
| La source de vos illustrations                                |  |
| Le dépôt des fichiers supplémentaires                         |  |
| Le format des images                                          |  |
| Quelques recommandations essentielles                         |  |
| Dans quels cas utiliser ces formats d'image                   |  |
| Pour obtenir une qualité optimale                             |  |
| La bibliographie                                              |  |
| La Feuille de style (Zotero)                                  |  |
| Les articles                                                  |  |
| Les extraits d'ouvrages                                       |  |
| Les livres                                                    |  |
| Les articles électroniques                                    |  |
| <b>Guide pour les Data Papers et modèle des métadonnées</b>   |  |
| <b>Cybergeog Data Papers Guidelines and Metadata template</b> |  |
| Specific guidelines for GeoData Papers                        |  |
| Metadata template                                             |  |
| Format for metadata                                           |  |

# Décrire quoi : le contenant ou le contenu ? Ou les deux ?



- Quel est l'objet d'un data paper :
  - Un jeu de données ?
  - Une base de données ?
  - Un dispositif d'acquisition ou de traitement de données ?
  - Un entrepôt de données ?
- Quelles données ? Sélection ?
- Un large éventail de « documents de données »

[https://www.isko.org/cyclo/data\\_documents](https://www.isko.org/cyclo/data_documents)

# L'évaluation

Et par qui ? Data papers par  
« la communauté » ?  
Science ouverte ? Open review ?

Évaluer un data paper  
retour d'expérience de la revue

cybergeog

european journal of geography  
revue européenne de géographie

Clémentine Cottineau, CNRS  
Christine Kosmopoulos, CNRS  
Denise Pumain, Paris 1

Webinaire du groupe de travail inter-  
reseaux  
Atelier Données de la MITI (CNRS)  
Jeudi 5 novembre 2020

- Evaluer quoi :
  - le data paper ?
  - les données ?
  - le dépôt des données ?
- Evaluer comment :
  - en double aveugle ?
  - par les pairs (mais qui) ?
- Evaluer pourquoi :
  - pour la qualité de la revue ?
  - pour l'impact des données ?

# La normalisation

**Consignes spécifiques: metadata**

Format for metadata

|                          |                  |            |      |                |
|--------------------------|------------------|------------|------|----------------|
| Title                    |                  |            | text | 250 characters |
| Abstract                 |                  |            | text | 100 words      |
| Temporal reference       |                  |            |      |                |
|                          | Time lapse       |            |      |                |
|                          |                  | time_begin | date | dd-mm-year     |
|                          |                  | time_end   | date | dd-mm-year     |
|                          | Publication date |            | date | dd-mm-year     |
|                          | Latest update    |            | date | dd-mm-year     |
|                          | Creation date    |            | date | dd-mm-year     |
| Responsible organization |                  |            | text | 250 characters |
| Responsible role         |                  |            | text | 250 characters |
| Keywords                 |                  |            | text | 250 characters |
| Use                      |                  |            | text | 50 words       |

> Normes

|                                                         |       |  |                     |                                                                  |
|---------------------------------------------------------|-------|--|---------------------|------------------------------------------------------------------|
| Type of spatial representation                          |       |  | ISO controlled list | vector - grid - table - tin - stereoscopic model - photo - video |
| Spatial resolution (scale or minimum cartographic unit) |       |  | text                | 50 characters                                                    |
| Language                                                |       |  | text                | 50 characters                                                    |
| Themes                                                  |       |  | text                | 100 characters                                                   |
| Geographic extension                                    |       |  | text                |                                                                  |
|                                                         | min y |  | number - float      |                                                                  |
|                                                         | min x |  | number - float      |                                                                  |
|                                                         | max y |  | number - float      |                                                                  |
|                                                         | max x |  | number - float      |                                                                  |
| Reference system                                        |       |  | text                | 50 words                                                         |
| Source                                                  |       |  | text                | 100 words                                                        |

- Lien avec principes FAIR
- Mais : FAIRisation des données, des entrepôts de données ou des articles de données ?
- Métadonnées disciplinaires ou/et génériques ?
- Quels formats ?



# La confiance

## Consignes spécifiques: dépôt

"Make sure to deposit your data in an appropriate and sustainable repository and provide the appropriate authorization certificates if the data includes elements under proprietary license."

- **Dépôt extérieur** (figshare, zenodo, nakala etc.) de préférence institutionnel
- **Recommandé** : Attribution d'un DOI par l'entrepôt institutionnel
- **Attention aux licences d'utilisation**
- **Responsabilité des auteurs**

- Choix du dépôt de données: critères, besoins, usages?
  - Respect (et gestion) des principes FAIR.
  - Espace de stockage (13K fichiers, 6GB).
  - Ergonomie producteur et utilisateur (structure complexe).
  - Visibilité.
  - Générique, disciplinaire, interne?
  - Possibilités: Harvard Dataverse, Nakala, Figshare,...

- Besoin de rendre les entrepôts (et les données) fiables, crédibles, « dignes de confiance » (*trustworthy*)
- Lien avec certification (CoreTrustSeal) ?
- Lien avec principes TRUST ?
- Lien avec principes FAIR ?
  - Surtout F & R ?

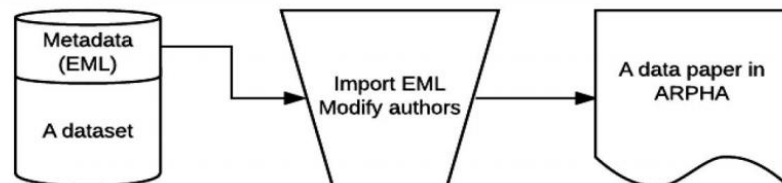


# L'intégration dans un écosystème

## Outils et fonctionnalités interopérables

Exemple : métadonnées GBIF (IPT) → Data paper (ARPHA)

Generate and import an **entire manuscript**



| Name                 | Title                    | Include                             | is Submitting            |
|----------------------|--------------------------|-------------------------------------|--------------------------|
| Viktor Senderov      | dataone@pensoft.net      | <input checked="" type="checkbox"/> | <input type="checkbox"/> |
| Nikolene Plass-Chang | nikolene@plass-change.fr | <input type="checkbox"/>            | <input type="checkbox"/> |
| Oscar Schofield      | oscar@marine.rutgers.edu | <input type="checkbox"/>            | <input type="checkbox"/> |

Photosynthetic pigments of water column samples analyzed using High Performance Liquid Chromatography (HPLC), sampled during Palmer LTER field season at Palmer Station Antarctica, 1991 - 2009.

- Un écosystème d'infrastructures de recherche, d'entrepôts de données, de plateformes de revues
- Le besoin de l'interopérabilité
- Workflow « propriétaire » ou « ouvert » ?
  - Ex.: Dataverse, Pensoft
  - Outils libres ?

# Des écosystèmes communautes

## Démarche de publication

- Nouvelle méthodologie sur la reconstitution des réseaux de transport antiques
- Explication de la démarche et mise à disposition des outils pour la reproduire

**Article principal**  
(*Journal of Computer Applications in Archaeology*)

Démarche de modélisation des itinéraires entre les sites archéologiques et résultats

**Data paper**  
(*Journal of open archaeology data*)  
Présente le jeu de données, sa construction et sa qualité

**Dépôt**  
(Zenodo)  
Données d'entrées, résultats, métadonnées et outils de traitement



## Data Papers provide an Innovative Tool for Information and Data Management

This study seeks to demonstrate how the data paper provides an innovative tool for information and data management, as part of an "ecosystem" of conference proceedings, journal articles, research data and open repositories. It relies upon GreyNet's current collection of 46 published datasets and 16 data papers. The study

highlights the importance of the human contribution for the writing of data papers and the enrichment of their metadata. To this end, key shared components of GreyNet's collection of data papers are discussed, namely the stakeholders, linked metadata, open data archiving, preservation, and issues of quality and information rights. The study concludes from a user perspective by addressing the value of data papers drawn from available statistics. The results are expected to move beyond a simple case study to a use case in which the key components of data papers can be implemented in other communities of practice dealing with grey literature.

## Data Papers Extend GreyNet's Document Trail

| Document Types | Proposal / Abstract | Conference Poster | Conference Slides | Video Presentation | Conference Paper | Published Dataset | Data Paper Preprint | Data Paper Article |
|----------------|---------------------|-------------------|-------------------|--------------------|------------------|-------------------|---------------------|--------------------|
|                |                     |                   |                   |                    |                  |                   |                     |                    |



← Data Papers are a multifaceted Tool

↓ Data Papers increase file Downloads

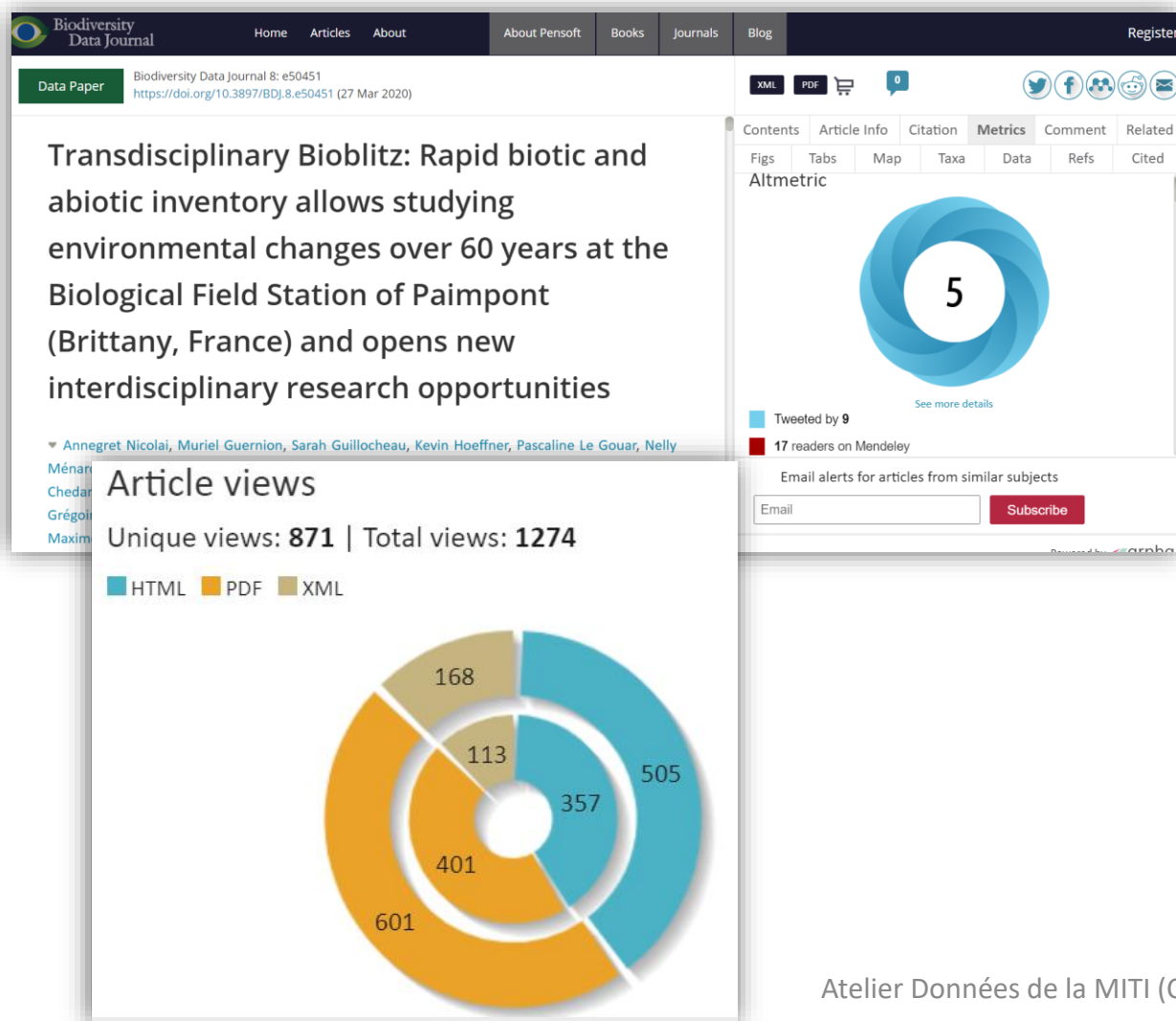
| As of August 31, 2020         | With Data Papers | Without Data Papers |
|-------------------------------|------------------|---------------------|
| GreyNet's Datasets in DANS    |                  |                     |
| 46 Datasets                   | 16 (34.8%)       | 30 (65.2%)          |
| 899 Downloads                 | 423(47.1%)       | 476 (52.9%)         |
| Average Downloads per Dataset | 26.4             | 15.9                |

Dominic Farace, GreyNet International

Joachim Schöpfl, France

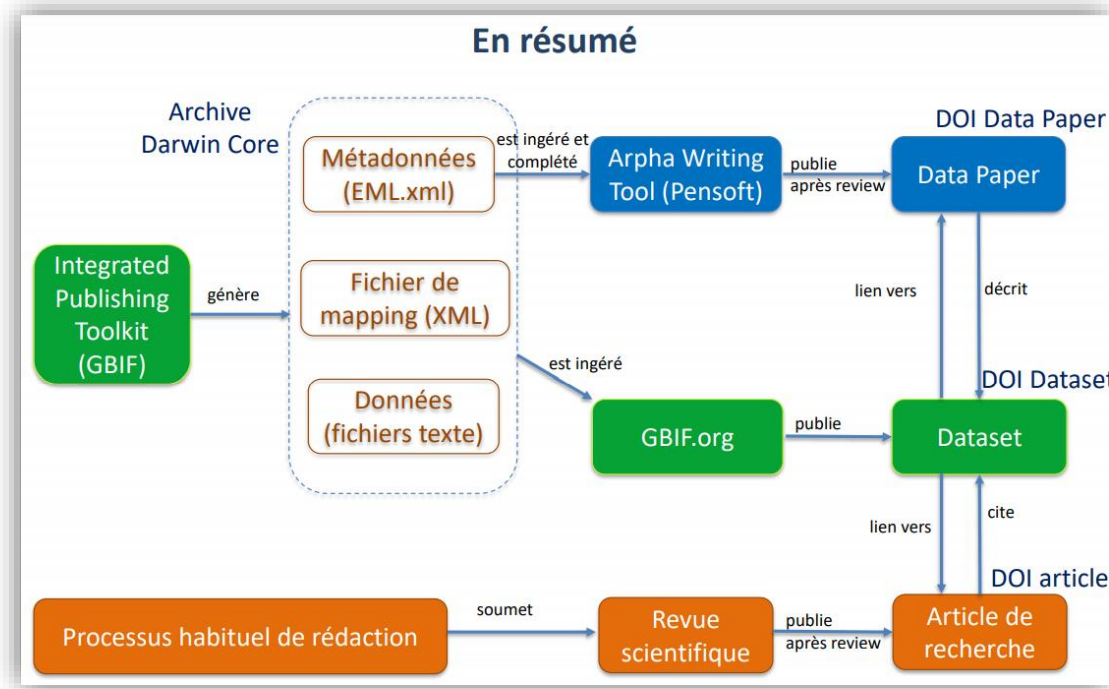


# L'impact d'un data paper



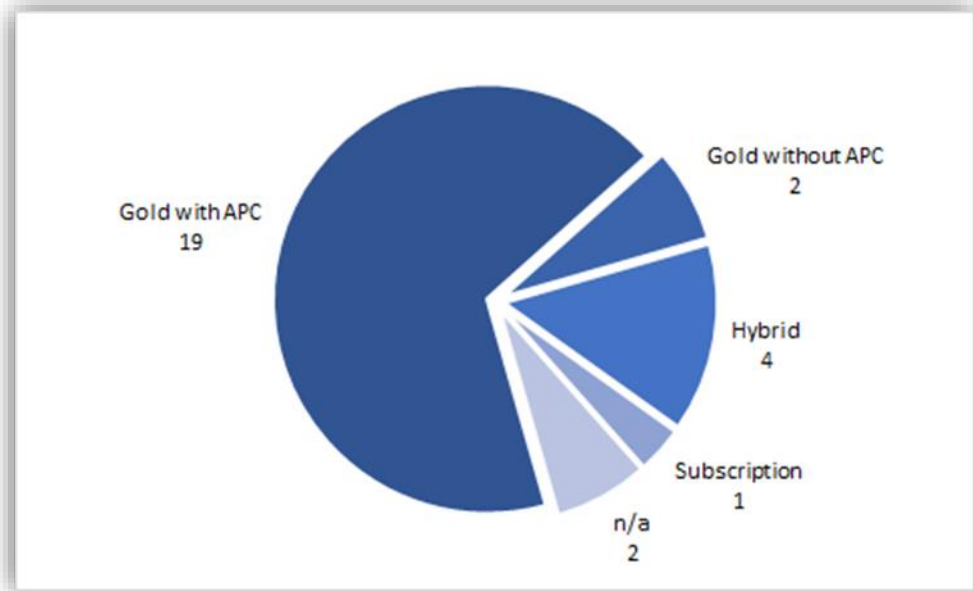
- Consultations
- Téléchargements
  - Mais réutilisation ?
- Citations
- *Dataset ou data paper ?*
- Lien avec richesse des métadonnées ?

# L'automatisation



- Génération automatique à partir d'un entrepôt ?
  - Quelle valeur ajoutée humaine ?
- Lecture du data paper par une machine ?
  - Formalisation
  - Normalisation

# L'économie de l'écosystème



- Data papers comme un moyen pour développer l'activité des éditeurs et augmenter leur chiffre d'affaires ?
- Quel risque d'un « predatory publishing » ?

- Except for Taylor & Francis, all big five academic publishers (Elsevier, Springer-Nature, Wiley-Blackwell and SAGE) have their own data journals.
- Other data journals are published or hosted by newcomers, especially by OA publishers such as Ubiquity Press, BioMed Central, Hindawi, MDPI, Copernicus Publications, Pensoft or Faculty of 1000.

# Lien avec processus (cycle de vie)

## La rédaction du data paper

- Le data paper n'est pas qu'un codebook descriptif.
- Description détaillée sur plusieurs plans
  - Méthodologie de construction
  - Données source
  - Choix de catégorisation
  - Limites et valeur par rapport à l'état de l'art
  - Éléments techniques
  - Réutilisation
- Plusieurs modèles possibles, format évolutif (ici: *Scientific Data*).
- La conception des données et du data paper doit être simultanée.
- Un travail conséquent!

- Quelles données :
  - données « froides » ?
  - données « chaudes » ou « tièdes » ?
- Lien avec plan de gestion ?
- Lien avec publication ?
  - Avant, après, à la place ?

# Le besoin de travailler ensemble

## Quels rôles pour les fonctions support?

- Large spectre de connaissances nécessaires.
- Rôle primordial de la formation
  - Pas d'enseignement à l'école doctorale.
  - Suivi de formations continues (URFIST, MSHS-T, DoRANum).
- Pas d'appui local (UT1: économie et droit très en retard).  
⇒ Besoin démarche proactive des deux parties.
- Un travail conjoint chercheurs / fonctions support?
  - Spécificités scientifiques disciplinaires: usages, diffusion,...
  - Des structures d'incitation et responsabilité incompatibles?

- Quels métiers ?
- Quelles fonctions ?
- Quelles activités ?
- Quelles compétences ?
  
- Reconnaissance ? Incitation ?
  - Indexation ? Monitoring ?



# Quelles perspectives ?



**HAL**  
archives-ouvertes.fr

Accueil Dépôt Consultation Recherche Documentation

hal-00481614, version 1 Ouvrage (y compris édition critique et traduction)

Data Paper – High Resolution Vegetation Cover Data for the Southern Western Ghats of India. (IFP\_ECODATA\_VEGETATION)

Quentin Renard<sup>1</sup>, B. R. Ramesh<sup>1</sup>, G. Muthusankar<sup>1</sup>, Raphaël Pelissier<sup>1,2</sup> [Détails](#)

**1** IFP - Institut Français de Pondichéry  
**2** UMR AMAP - Botanique et Modélisation de l'Architecture des Plantes et des Végétations

**Abstract :** The Western Ghats form a 1,600 km long escarpment that runs parallel to the southwestern coast of Peninsular India. This relief barrier, which orographically exacerbates the summer monsoon rains, is responsible for steep bioclimatic gradients that have long been recognized as one of the major ecological determinants for the forest vegetation of the region. We report here girded vegetation data at 30' lat/long (ca. 1 km) resolution that cover an area of about 70,000 km<sup>2</sup> of the southern Western Ghats, between 74 to 78° E and 8 to 16° N. These data have been extracted from: the 1:250,000 scale forest maps of South India published by the French Institute of Pondicherry (FIP), which have been digitized and simplified; the 2004 MODIS (Moderate Resolution Imaging Spectroradiometer) database, for the IGBP (International Biosphere Geosphere Programme) global vegetation Land Cover Type and Normalized Difference Vegetation Index (NDVI) of March 2004.

- Augmentation du nombre de *data papers* (?)
- Substitution aux *regular papers* ?
  - Comment faire le *monitoring* ?
- Développement des *data journals* ?
  - Multidisciplinaires ?
  - Disciplinaires / communautaires ?
- Des « *mega data journals* » ?
- Ou absorption par les *regular journals* ?
  - Ou preprints ? Ou HAL ?
- Rapprochement des plateformes publications et données ?
  - Intégration verticale ?
- Tous les domaines scientifiques ?
  - SHS ?