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

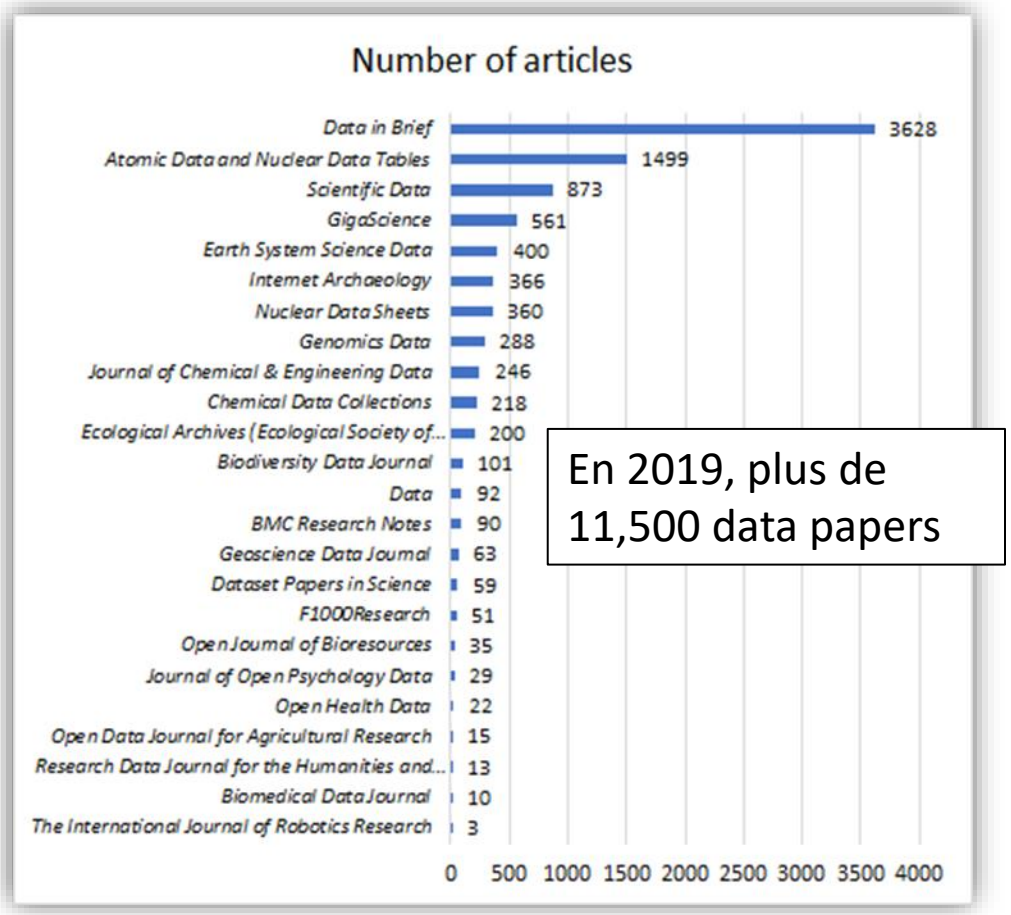
Un intérêt croissant

Partenaires du GT inter-réseaux - Atelier Données
Renatis, Médici, QeR, Resinfo, BDD, Calcul, Devlog, SIST, DIST-CNRS

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

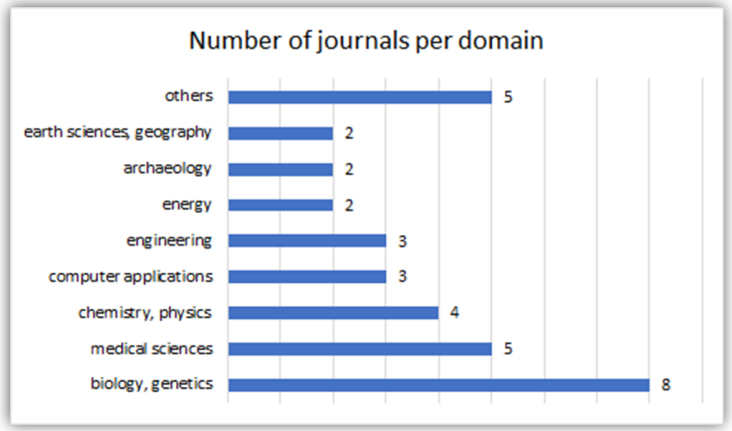
➔ Dans le cadre du soutien public aux revues, recommander l'adoption d'une politique de données ouvertes associées aux articles, le développement des articles de données et des revues de données.

Une réalité en évolution



En 2019, plus de 11,500 data papers

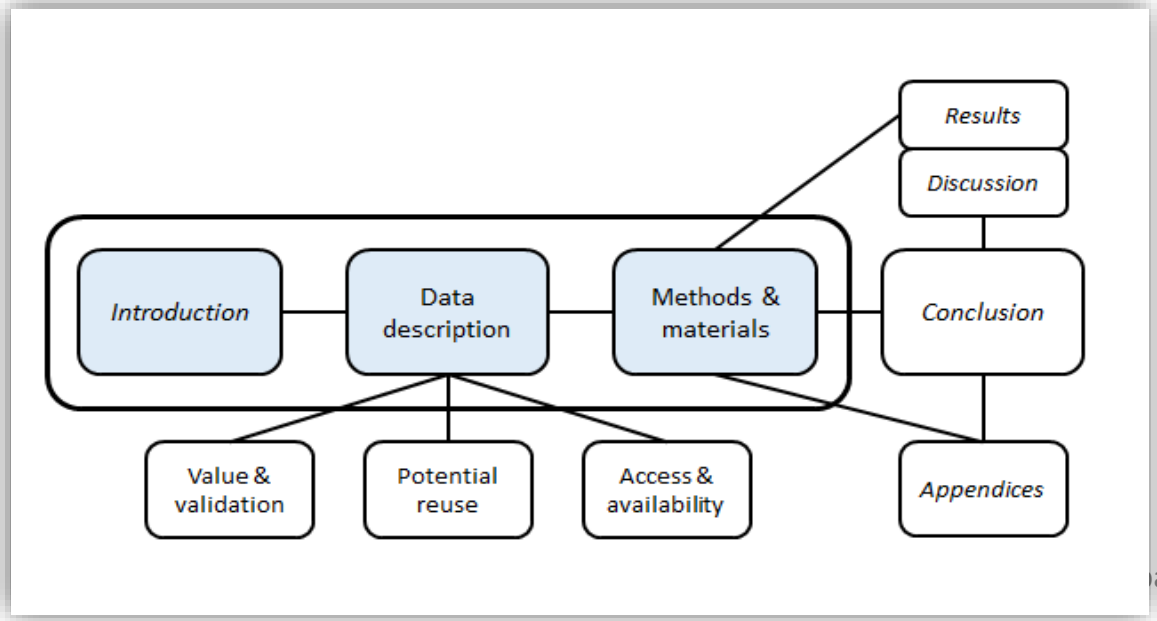
- Un type d'article relativement récent
- Un nombre d'articles en augmentation constante
- Mais peu d'articles (<0.1%)
- Surtout STM

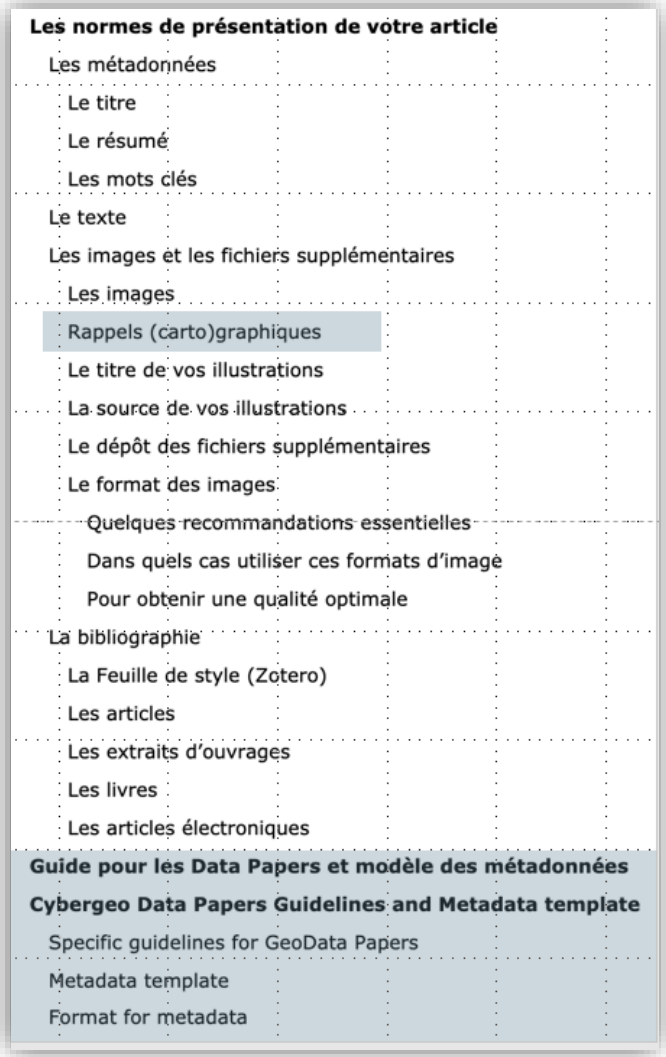


Une grande diversité

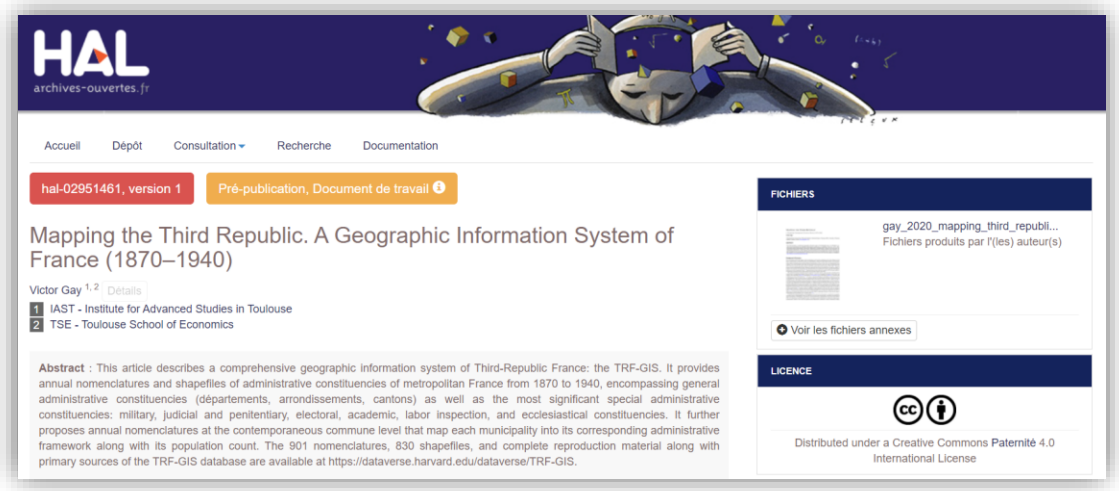
La définition la plus simple :

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





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 »

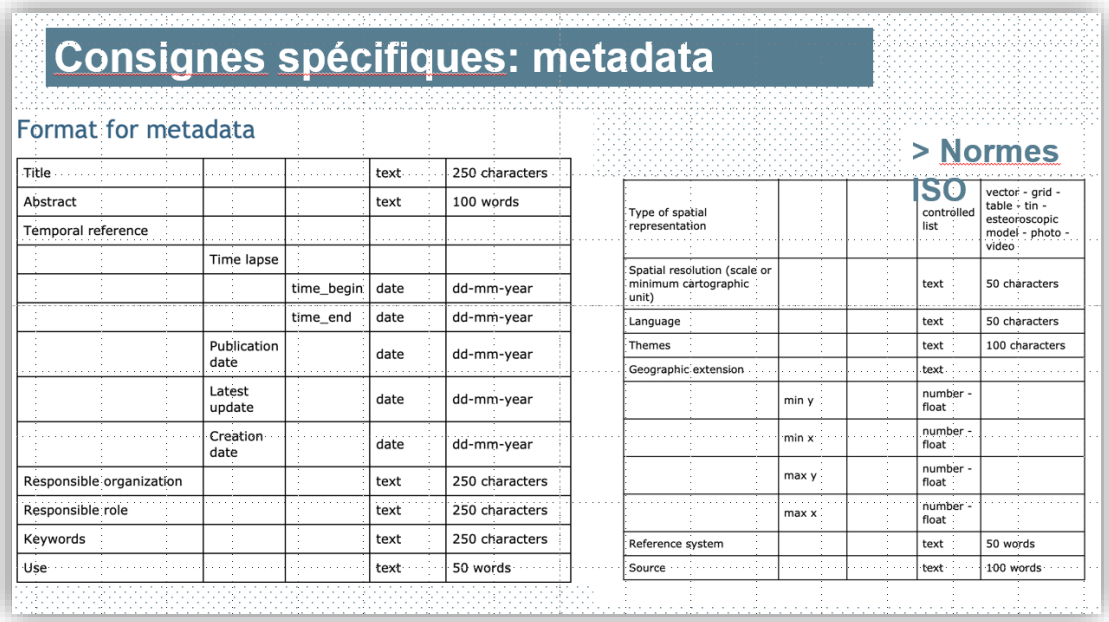
https://www.isko.org/cyclo/data_documents

L'évaluation

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

- 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

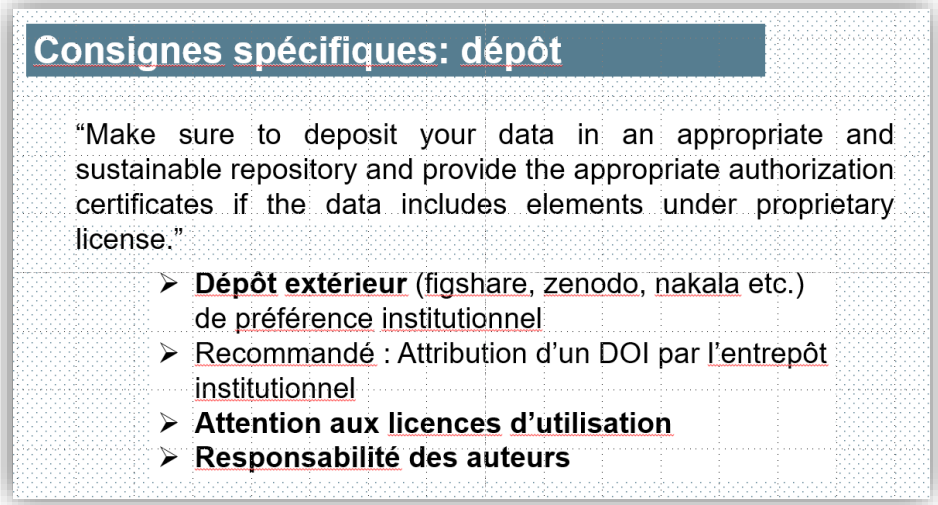
Title			text	250 characters
Abstract			text	100 words
Temporal reference				
	Time lapse			
		time_begin	date	dd-mm-year
		time_end	date	dd-mm-year
	Publication date		date	dd-mm-year
	Latest update		date	dd-mm-year
	Creation date		date	dd-mm-year
Responsible organization			text	250 characters
Responsible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

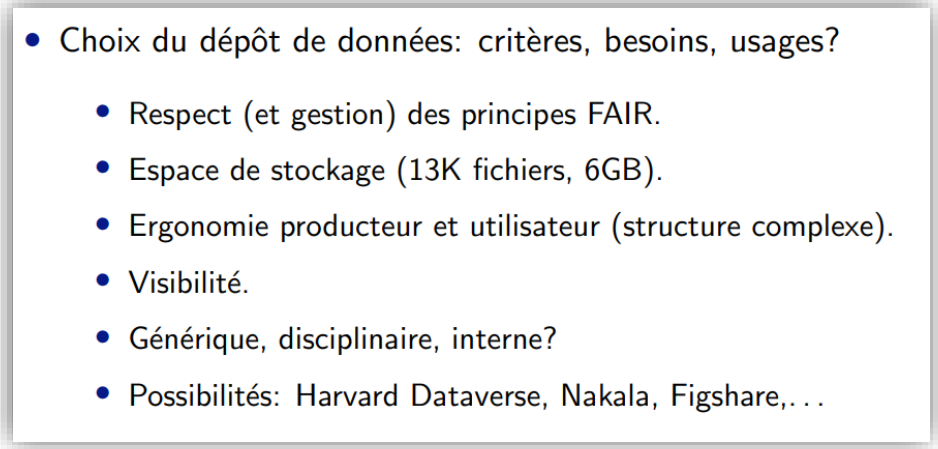
> Normes ISO

Type of spatial representation			controlled list	vector - grid - table - tin - esteoroscopic 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 ?

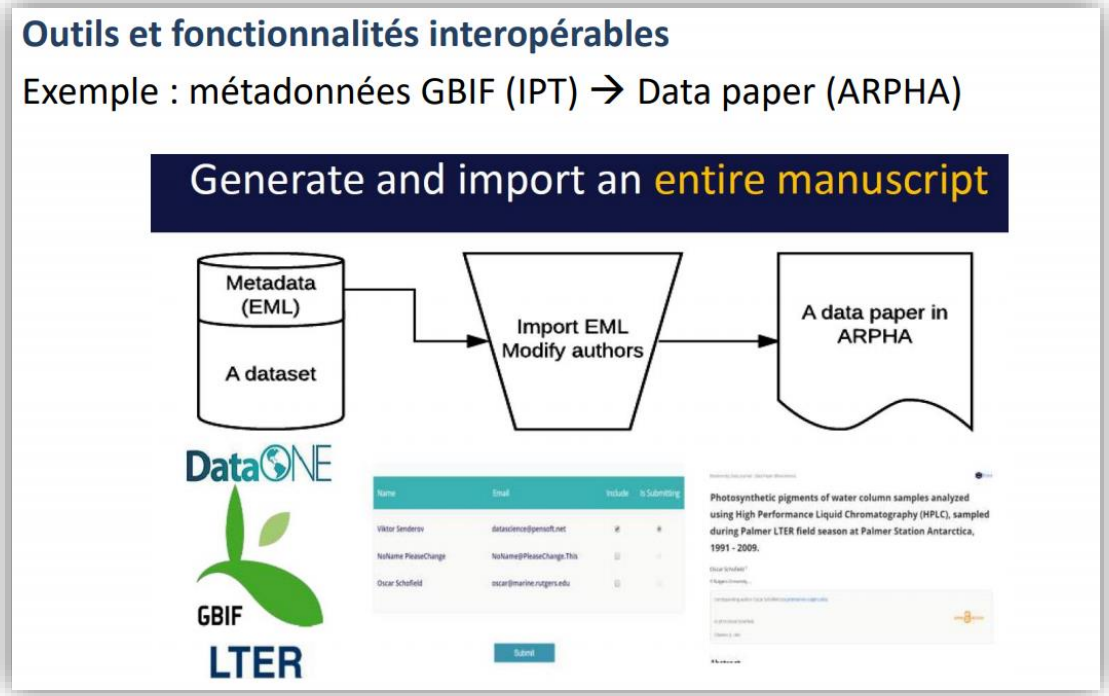
La confiance





- 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



- 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 communautaires

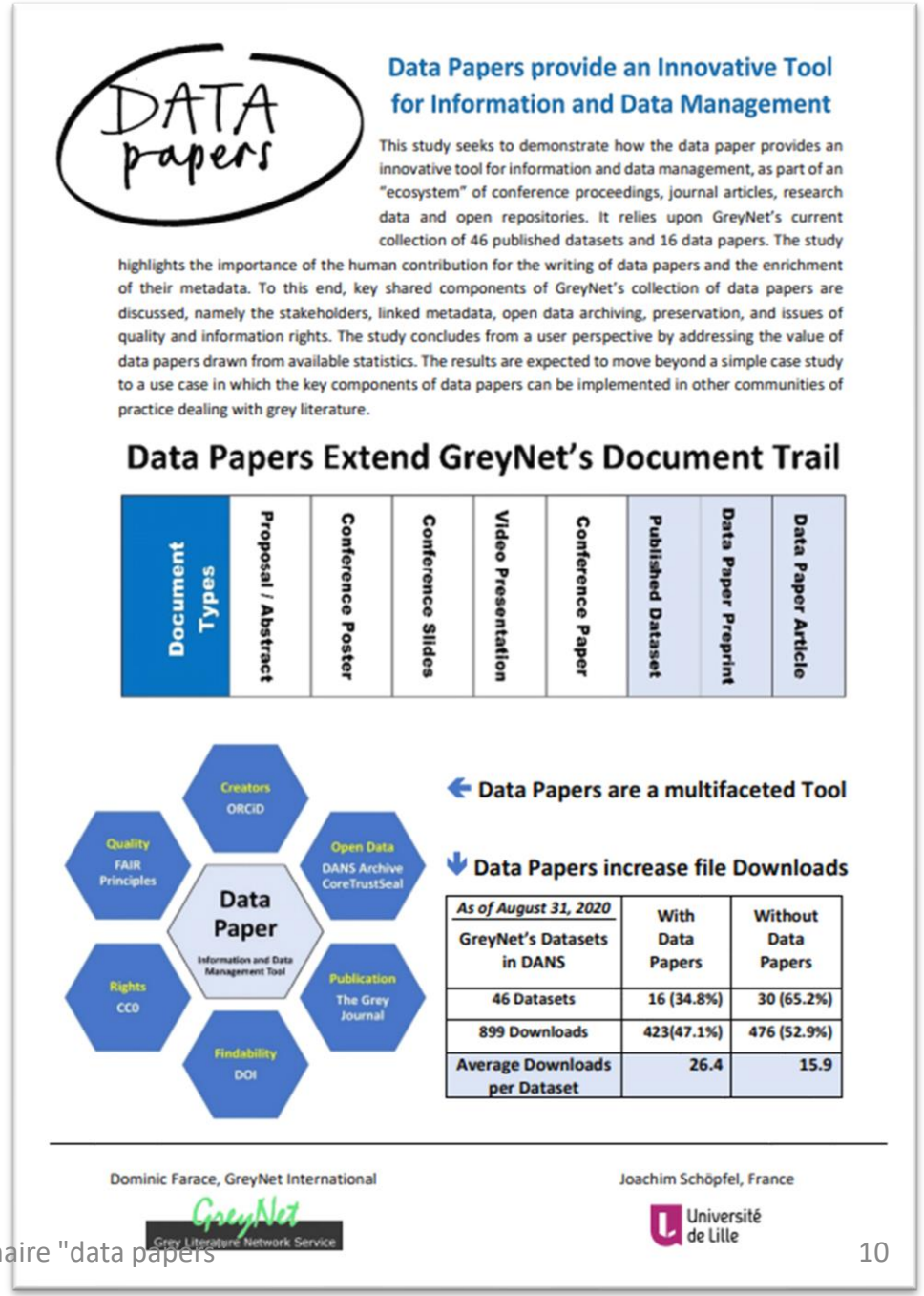
Démarche de publication

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

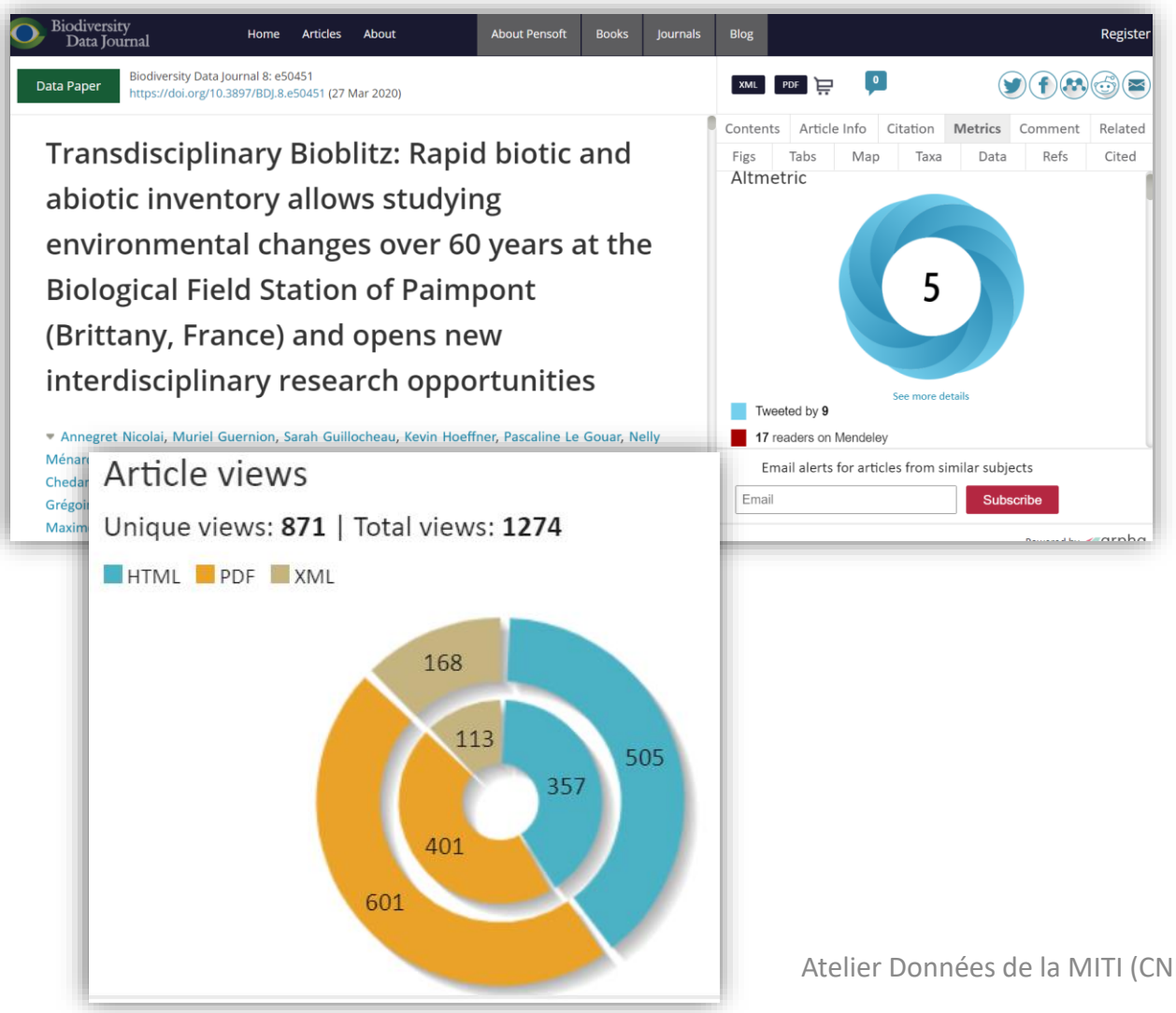
Article principal
(Journal of Computer Applications in Archaeology)
Démarche de modélisation des itinéraires entre les sites archéologiques et résultats

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



L’impact d’un data paper

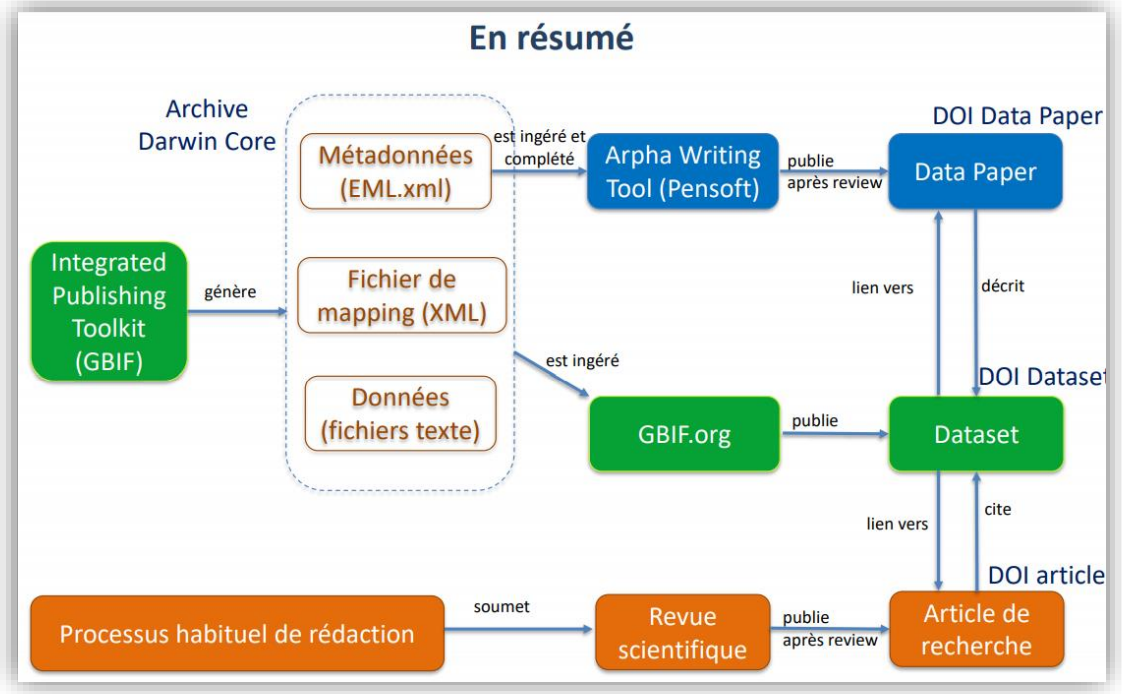


- Consultations
- Téléchargements
  - Mais réutilisation ?
- Citations

- *Dataset* ou *data paper* ?

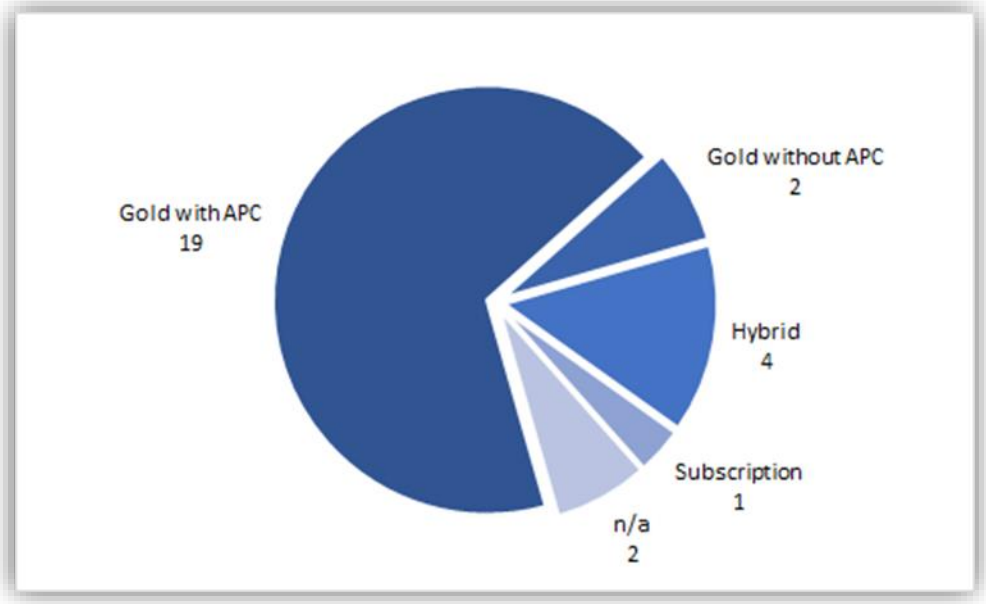
- Lien avec richesse des métadonnées ?

L’automatisation



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

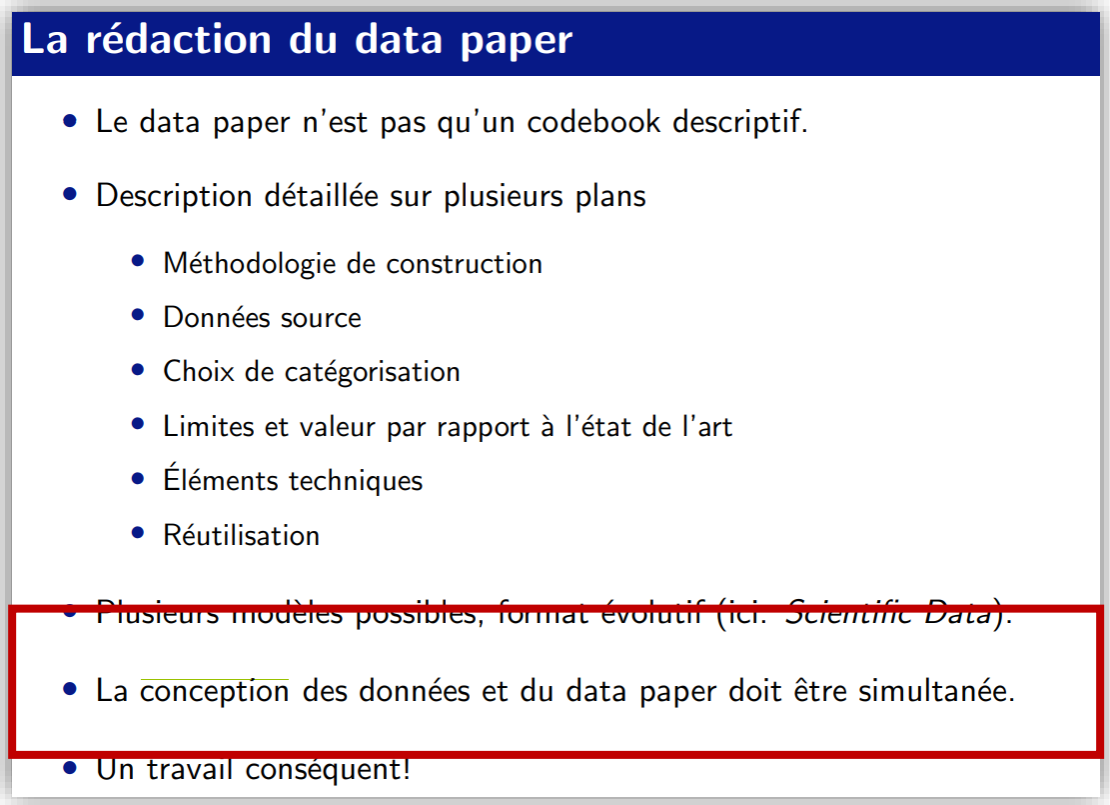
L’économie de l’écosystème



- 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)

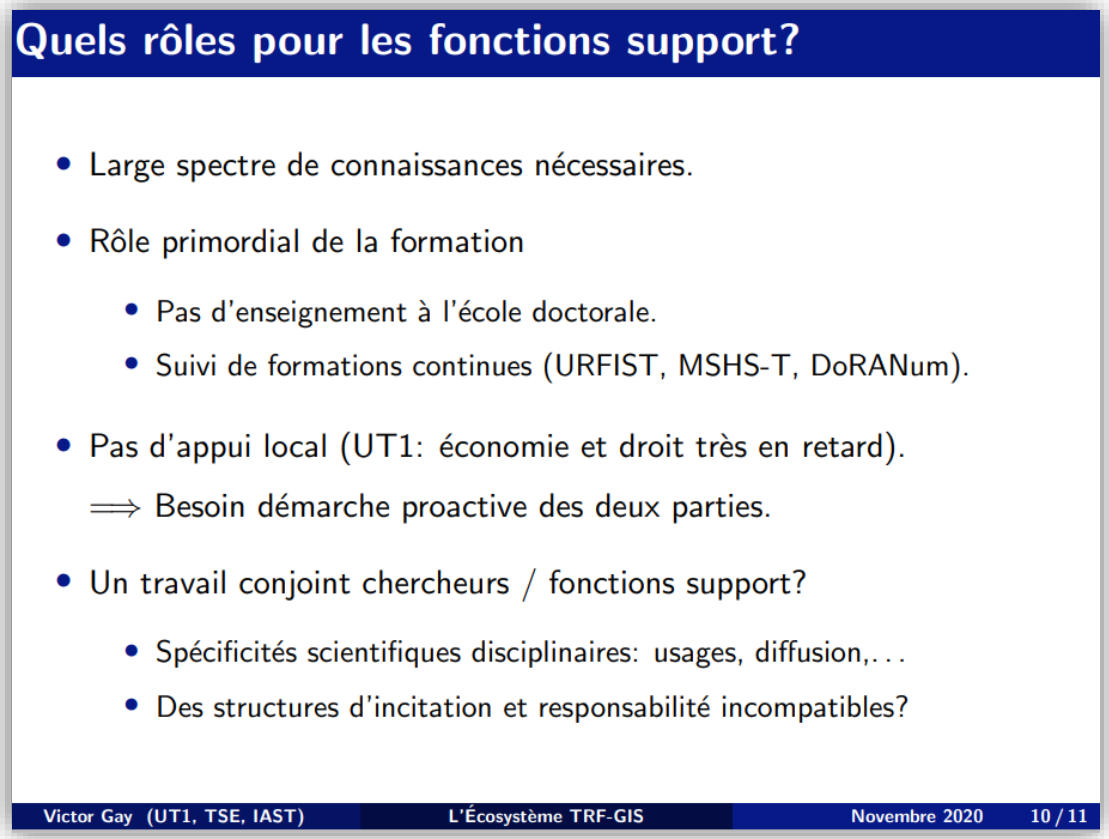


La rédaction du data paper

- 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 métiers ?
- Quelles fonctions ?
- Quelles activités ?
- Quelles compétences ?

- Reconnaissance ? Incitation ?
  - Indexation ? Monitoring ?

Quelles perspectives ?

- 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 ?