Analyse de l'impact des data papers de l'UMR TETIS

Rémy Decoupes

Analyse de l'impact des data papers

Rémy Decoupes

2024-01-31

L'objectif de cette étude est d'analyser l'impact des data papers sur la réutilisation des données mise à disposition par l'UMR TETIS. Pour cela, nous proposons d'analyser les indicateurs de réutilisation des jeux de données (consultation, téléchargement et citation) dépôsés sur Dataverse.Cirad.fr (CIRAD) ou Entrepôt.Recherche.Data.Gouv.Fr (INRAE) en comparant ceux accompagnés d'un data paper à ceux sans data paper.

Pour se faire, plusieurs API (*Application Programming Interface*) sont utilisées. Hal (pour INRAE) et Agritrop (pour le CIRAD) constituent la source des data papers, alors que les indicateurs pour les jeux de données sont obtenus via Dataverse.Cirad.fr (CIRAD) et Entrepôt.Recherche.Data.Gouv.Fr (INRAE).

En résumé des expérimentations menées par cette étude, il apparaît que les data papers TETIS (*14*) utilisent majoritairement le Dataverse CIRAD. L'utilisation de Entrepôt.Recherche.Data.Gouv.Fr est plus récente (*2021*), avec un nombre moyen de téléchargements par jeux de données de *10* pour INRAE (*30* jeux) et *55* pour le CIRAD (*120* jeux). Par ailleurs, les jeux de données associés à un data paper sont beaucoup plus téléchargés, en médiane, on observe plus de *100* téléchargements contre *5* pour les jeux sans data papers.

	INRAE	CIRAD
Data Papers		
Entrepôts de Données		

Sommaire

1. Analyse bibliométrique des articles de type `data paper`
   1.1 Collection TETIS/INRAE via Hal
   1.2 Collection TETIS/CIRAD via Agritrop
   1.3 Fusion des sources de données
2. Analyse bibliométrique des entrepôts de données (dit dataverse)
   2.1 Collection TETIS/INRAE via Entrepôt de Recherche Data Gouv
   2.2 Collection TETIS/CIRAD via Dataverse CIRAD
3. Analyse des jeux de données décrits par les data papers
   3.1 Répartition Dataverse CIRAD / RDG INRAE dans les data papers
   3.2 Comparaison du nombre de téléchargements avec ou sans data paper
4. Conclusion
5. Annexe
   5.1 Champs HAL utilisés
   5.2 Liste des data papers TETIS
   5.3 Liste des jeux de données TETIS sour Entrepôt Recherche Data Gouv

Version	Date	Description de la Modification	Auteur
1.0.	2024-01-25	Version initiale du document	Rémy Decoupes
1.1.	2024-01-31	Ajout de data papers non Scientific data ou Data in Brief.	

1. Analyse bibliométrique des articles de type `data paper`

1.1 Collection TETIS/INRAE via Hal

En interrogeant l'API, les data papers publiés dans la collection de TETIS peuvent être exportés à travers le champs Hal `docSubType_s: "DATAPAPER"`. 12 data papers ont été publiés depuis 2019.

La Figure 1 montre la série temporelle de publication de data papers depuis 2019. Il est à noter qu'il y a une accélération du nombre de datapaers depuis fin 2022. Les 2 revues, **Scientific Data** et **Data in Brief** sont les plus utilisées.

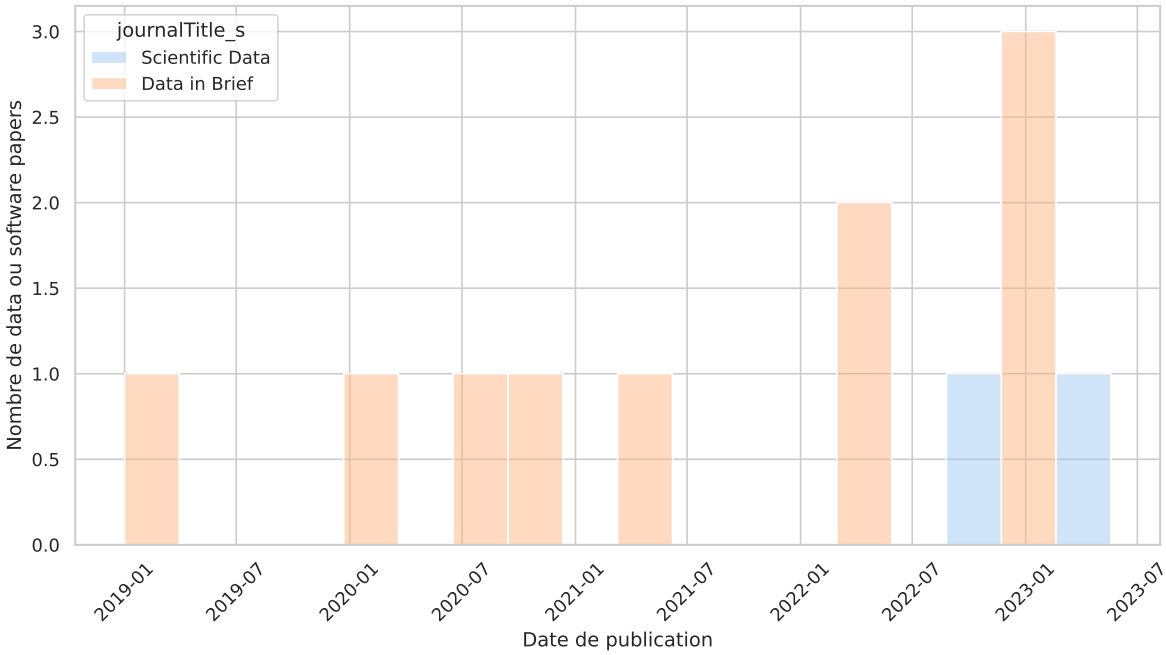

Figure 1: Histogramme des publications de data papers de l'UMR TETIS sous Hal depuis 2019

1.2 Collection TETIS/CIRAD via Agritrop

Agritrop ne semble pas proposer d'API. Il est donc nécessaire de filtrer manuellement, à travers le site web, les résultats d'Agritrop via `Affiliation : TETIS` et titre de revues de data papers : `Scientific data` et `Data in brief`. En effet, aucun champs ne permet de filtrer sur les articles de type data papers. 15 data papers ont été téléversés sur Agritrop depuis 2019. Seuls ces deux journaux ont été sélectionnés pour cette étude car ce sont les seuls présent dans la collection Hal de TETIS. Dans la version 1.1 du document, le data paper de Jolivot et al. (2021) publié dans *Earth System Science Data* a été ajouté manuellement à la collection.

Pour plus d'information, le document CIRAD de Laurence Dedieu (2017) propose une liste exhaustive des journaux acceptant les data papers.

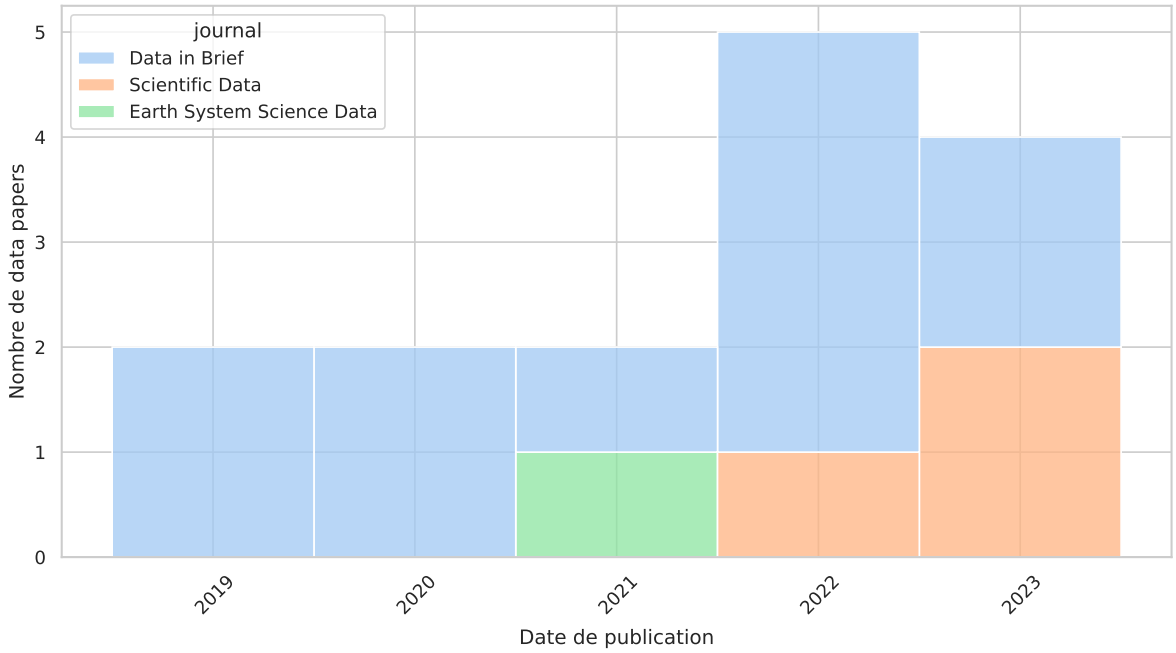

Figure 2: Histogramme des publications de data papers Agritrop depuis 2019

La Figure 2 propose l'historique de dépôts de data papers sur Agritrop.

1.3 Fusion des sources de données

Table 3: Liste des articles présent uniquement dans l'une des collections

	source	titre
0	agritrop	A manually annotated corpus in French for the ...
1	agritrop	Land use / land cover map of Vavatenina region...
2	agritrop	Harmonized in situ datasets for agricultural l...

Après fusion des deux sources puis suppression des articles présent dans les deux plateformes, le nombre de data papers publiés depuis 2019 est 15. Les articles seulement présents dans Hal ou Agritrop sont détaillé dans la Table 3.

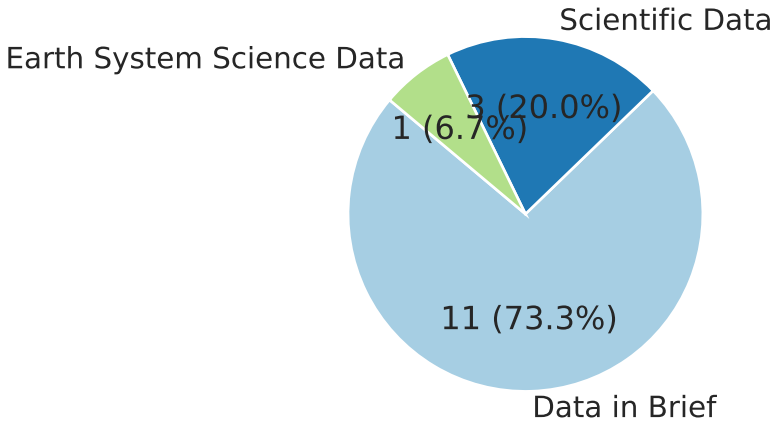

Figure 3: Répartition de l'ensemble des publications par revue

La liste complète des data papers est proposée en Annexe (Table 6).

Table 4: Liste des data papers TETIS ayant reçu au moins une citation

titre	citation_count	year
Harmonized in situ datasets for agricultural land use mapping and monitoring in tropical countries	10	2021
Land use / land cover map of Vavatenina region (Madagascar) produced by object-based analysis of very high spatial resolution satellite images and geospatial reference data	1	2022
PADI-web corpus: Labeled textual data in animal health domain	7	2019
Mapping land cover on Reunion Island in 2017 using satellite imagery and geospatial ground data	15	2020
Land cover maps of Antananarivo (capital of Madagascar) produced by processing multisource satellite imagery and geospatial reference data	1	2020
COVID-19 and Media datasets: Period- and location-specific textual data mining	6	2020
Food packaging permeability and composition dataset dedicated to text-mining	3	2021
Partial n-Ary relation instances on food packaging composition and permeability extracted from scientific publication tables	2	2022
VegAnn, Vegetation Annotation of multi-crop RGB images acquired under diverse conditions for segmentation	5	2023

En ce qui concerne les citations des data papers, nous les obtenons avec l'API de CrossRef, agence d'attribution de DOI. La Table 4 montre les data papers TETIS ayant au moins une citation. 60.0% des data papers TETIS ont obtenu au moins 1 citation.

2. Analyse bibliométrique des entrepôts de données (dit dataverse)

2.1 Collection TETIS/INRAE via Entrepôt de Recherche Data Gouv

Le nombre de jeux de données décrit dans la collection TETIS de l'entrepôt de Recherche.Data.Gouv est 33, pour un volume total de 7,851,073,543 bytes. La liste complète est disponible en annexe (Table 8).

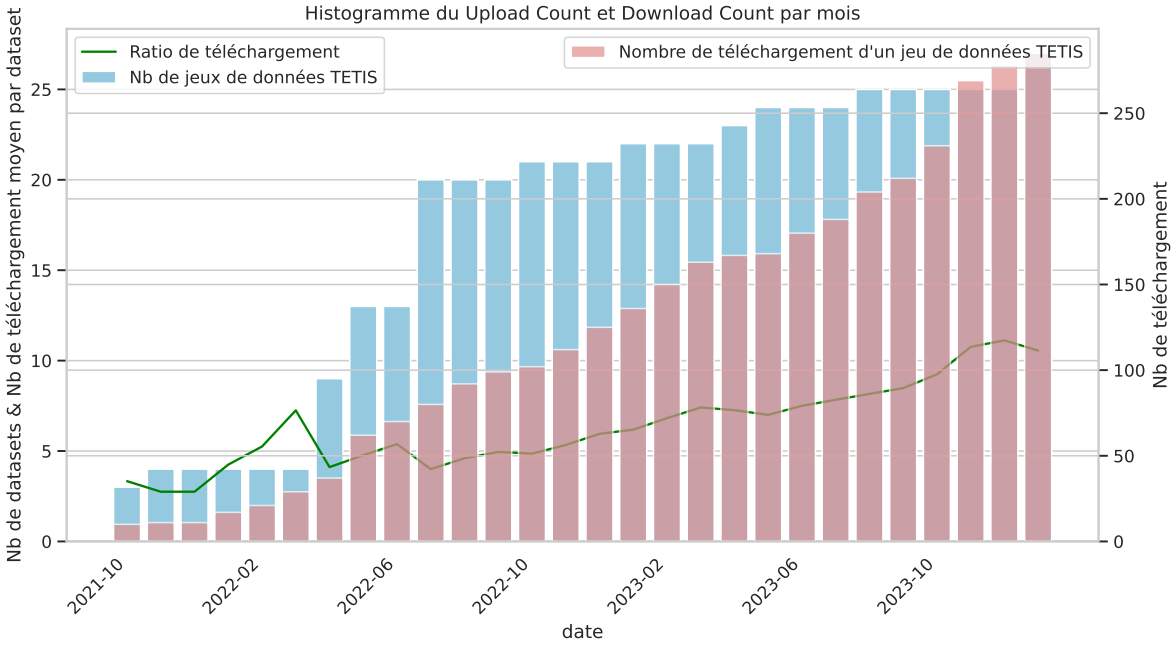

Figure 4: Analyse globale de la dynamique des dépôts de jeux de données ainsi que leurs téléchargement

La Figure 4 propose 3 variables d'intérêt pour suivre la dynamique globale de la collection TETIS. La première (en bleu) est le nombre de dépôts cumulés. La deuxième (en rouge) représente, en cumulé également, le nombre total des téléchargements de jeux de données TETIS. Enfin, en vert, nous affichons la moyenne du nombre de téléchargements par jeu de données. On observe une augmententation plutôt lente de cette moyenne. En début de 2024, la moyenne est de ~10 téléchargements par jeu de données.

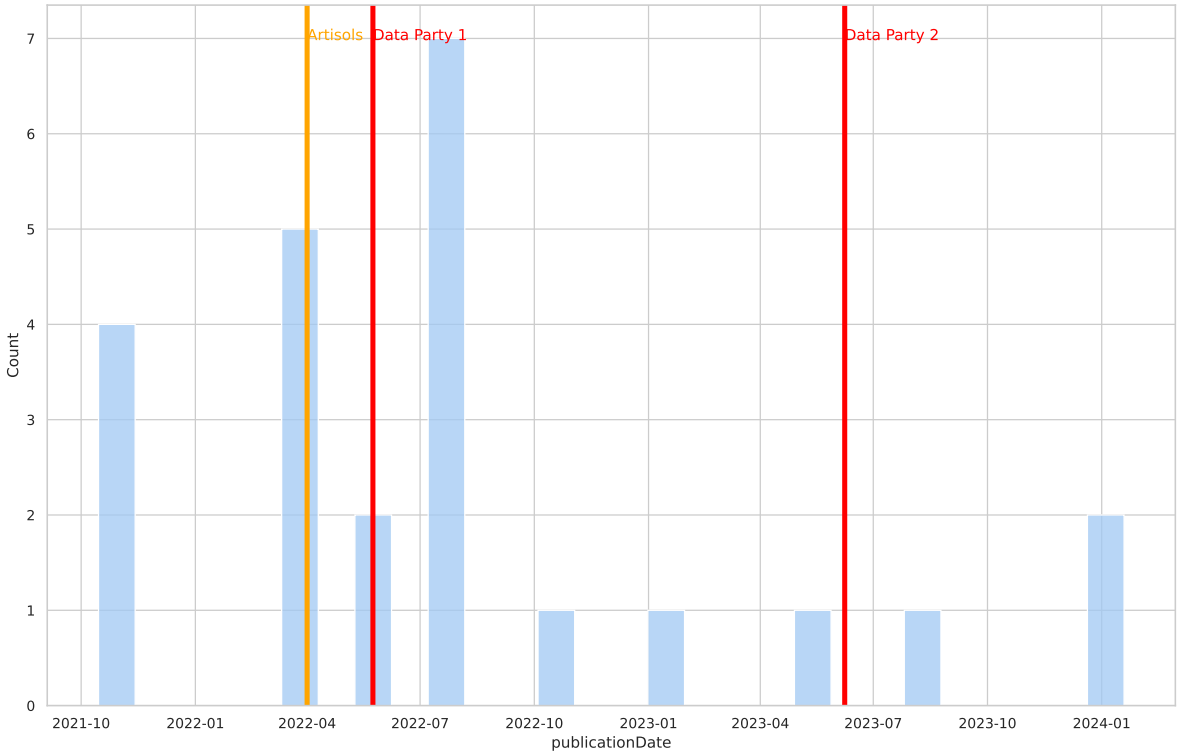

Figure 5: Histogramme des soumissions de jeux de données sous RDG

Plusieurs évènements autour de la gestion de la donnée à TETIS ont marqué l'évolution des soumissions de jeux de données dans la collection TETIS. La Figure 5 propose un histogramme de ces soumissions avec trois marqueurs temporels. Le premier est la date de mise à disposition des données produites par le projet Artisols. Le second et troisième sont les deux data parties organisées à TETIS. On observe que beaucoup de dépôts ont été réalisés après la première data party. La seconde, axée sur les Plan de Gestion des Données et la gestion des codes, n'a visiblement pas entraîné de nouveaux dépôts.

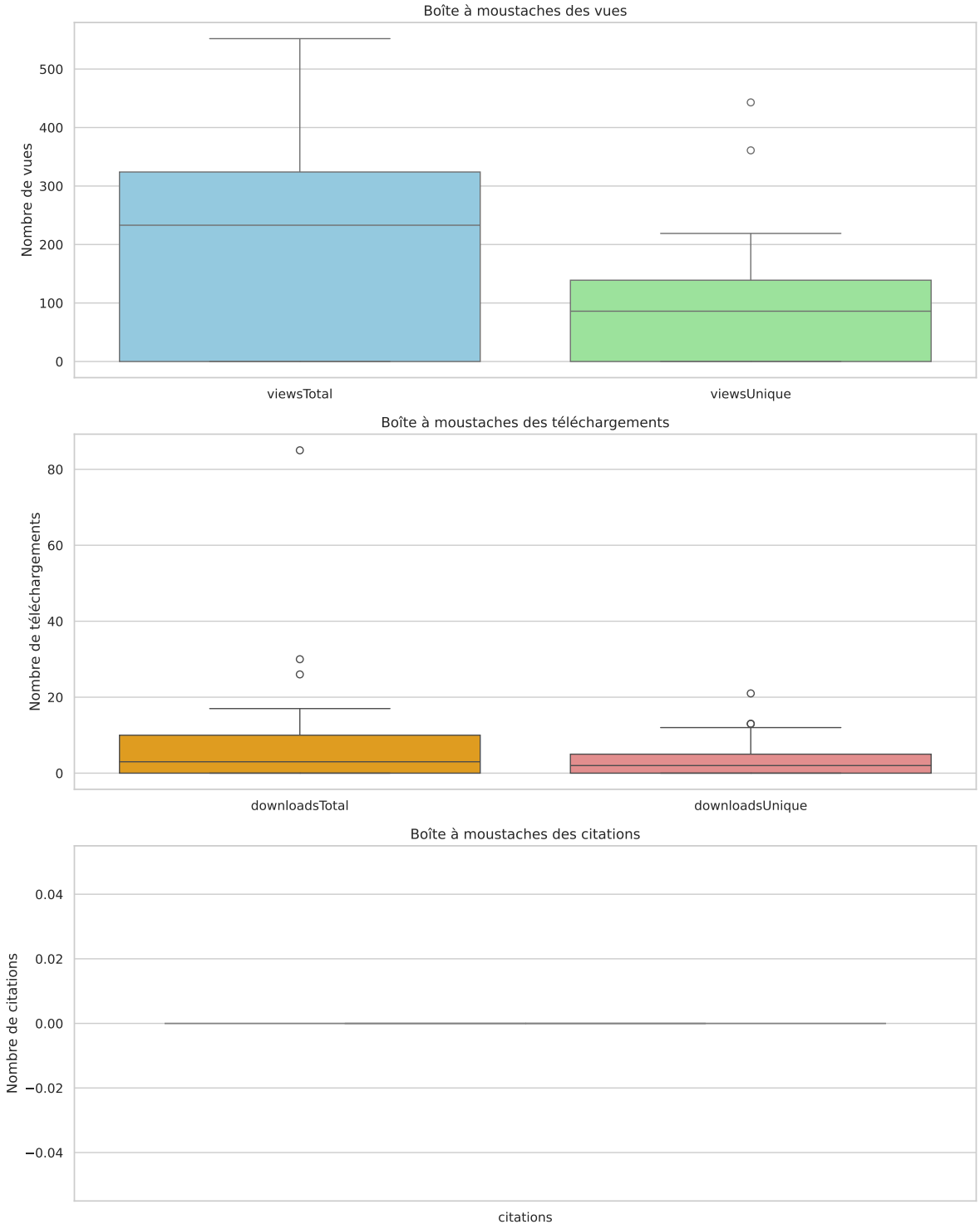

Figure 6: Distribution statistique des consultations (total et unique), des téléchargements (total et unique) ainsi que du nombre de citation

La Figure 6 montre les distributions statistiques des nombres de consultations, téléchargements et citation des jeux de données de la collection TETIS. Nous pouvons observer une grande variabilité de ces métriques pour les jeux de données. Certains jeux de données semblent très consultés & téléchargés comparativement à la moyenne des autres. Cette observation est confirmée lorsque l'on compare la moyenne du nombre de téléchargement par jeu de données de la Figure 4 (aux alentours de 10) et la médianne de la Figure 6 (inférieur à 5). Quelques jeux de données sont donc très téléchargés.

2.2 Collection TETIS/CIRAD via Dataverse CIRAD

La version de l'API actuelle du dataverse CIRAD (5.13) ne permet pas, malheureusement, d'obtenir les informations à l'échelle du jeu de données. En effet, seule une analyse globale (au niveau de la collection TETIS) est possible.

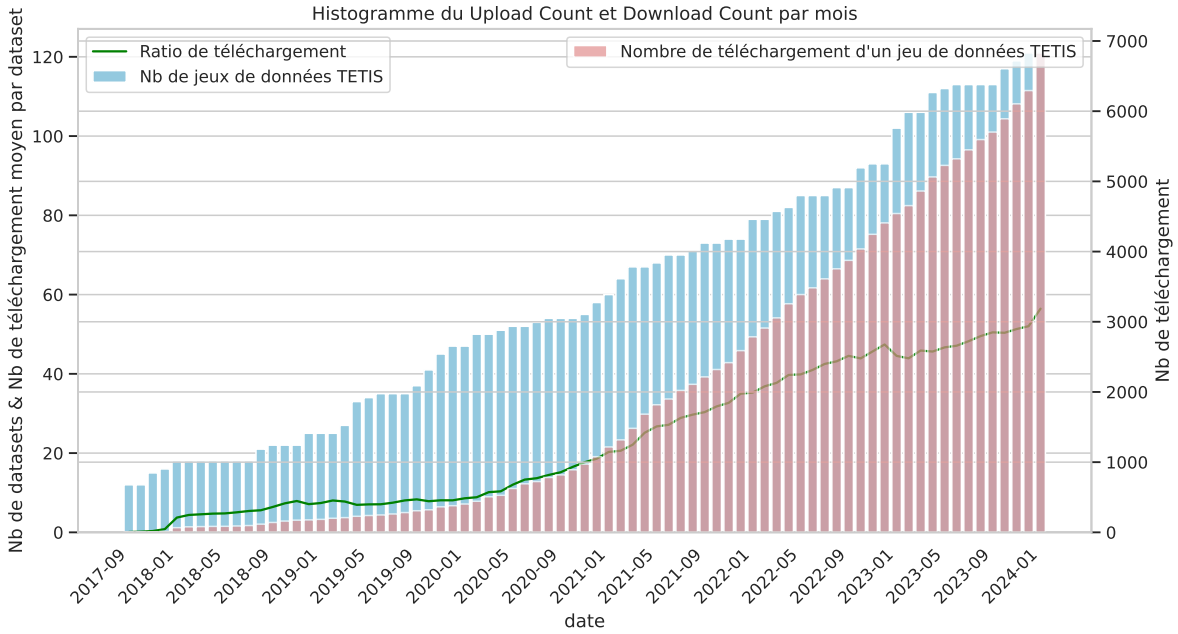

Figure 7: Analyse globale de la dynamique des dépôts de jeux de données ainsi que leurs téléchargement

Comme illustré par la Figure 7, la pratique de dépôt de jeux de données par le CIRAD est plus ancienne (2017 au lieu de 2021 pour INRAE). Le nombre de jeux de données est aussi bien plus important (120 Cirad pour 30 INRAE), ainsi que le nombre de téléchargement (6000 Cirad pour 250 INRAE). La moyenne du nombre de téléchargement par jeux de données est largement supérieur (55 Cirad pour 10 INRAE).

Comme indiqué en introduction de cette section, il est impossible d'accèder aux indicateurs à l'échelle des jeux de données. Il est donc impossible de tracer la distribution statistique du nombre de téléchargements afin de comparer la médianne et la moyenne.

3. Analyse des jeux de données décrits par les data papers

La méthode utilisée pour cette section est de télécharger les data papers référencés par le paragraphe 1.3 puis d'en extraire les liens vers les entrepôts dataverse CIRAD et Recherche.Data.Gouv. Avec cette liste, il est alors possible d'obtenir (soit par API soit par Web scrapping), le nombre de téléchargements.

La liste des publications, fusionnées de Agritrop et de Hal, possède les liens DOI vers les data papers. Cependant, les redirections (DOI landing page vers le site de l'éditeur) ainsi que les pages Web dynamiques des éditeurs rendent le scrapping difficile (le contenu de ces pages web sont générés dynamiquement en fonction de la navigation de l'utilisateur). Aussi, nous avons été contraint de ne travailler que sur les documents Hal, soit 80.0% des data papers. En effet, à partir de leur HalID, il est alors facile de télécharger le fichier PDF et de l'analyser.

Ainsi, après extraction du texte des fichiers PDF, deux motifs d'URL ont été cherchés, l'un correspond au prefix DOI de Recherche.Data.Gouv (10.15454) et l'autre au prefix CIRAD (10.18167). Ainsi pour chaque article Hal, nous avons la liste des liens DOI vers les entrepôts CIRAD ou Recherche.Data.Gouv.

3.1 Répartition Dataverse CIRAD / RDG INRAE dans les data papers

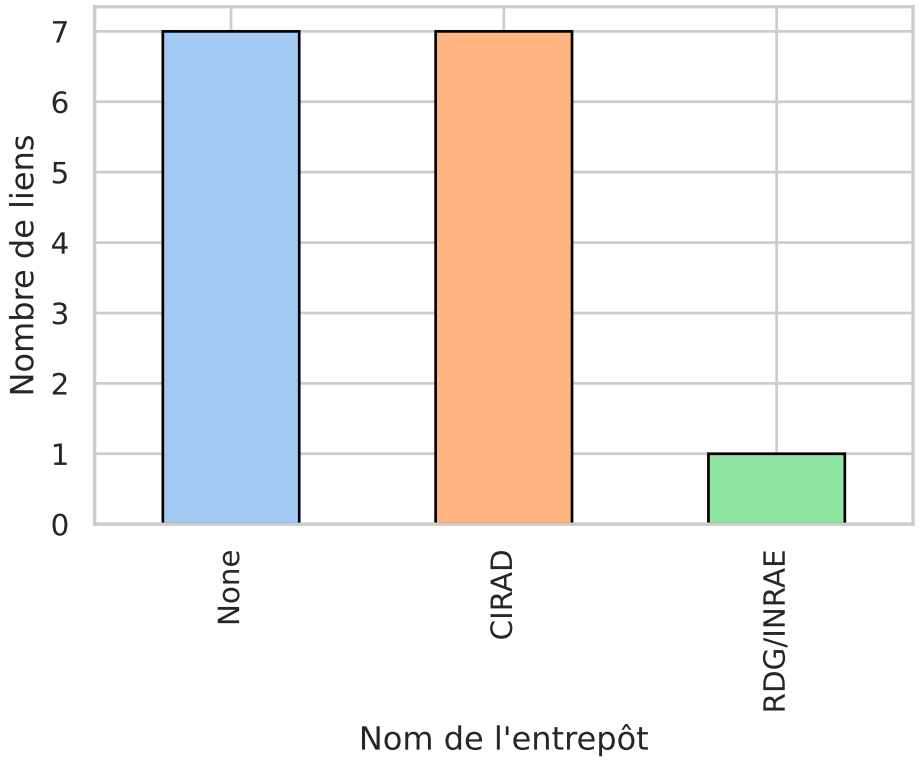

Figure 8: Nombre de liens vers les entrepôts RDG/INRAE (Recherche.Data.Gouv), CIRAD (Dataverse.Cirad) et None (lorsqu'aucun lien n'a été trouvé)

Parmi les data papers sur Hal (e.g. 15), 7 n'ont pas de liens vers les entrepôts INRAE ou CIRAD. Par ailleurs, l'entrepôt CIRAD est le plus utilisé comme illustré par Figure 8.

3.2 Comparaison du nombre de téléchargements avec ou sans data paper

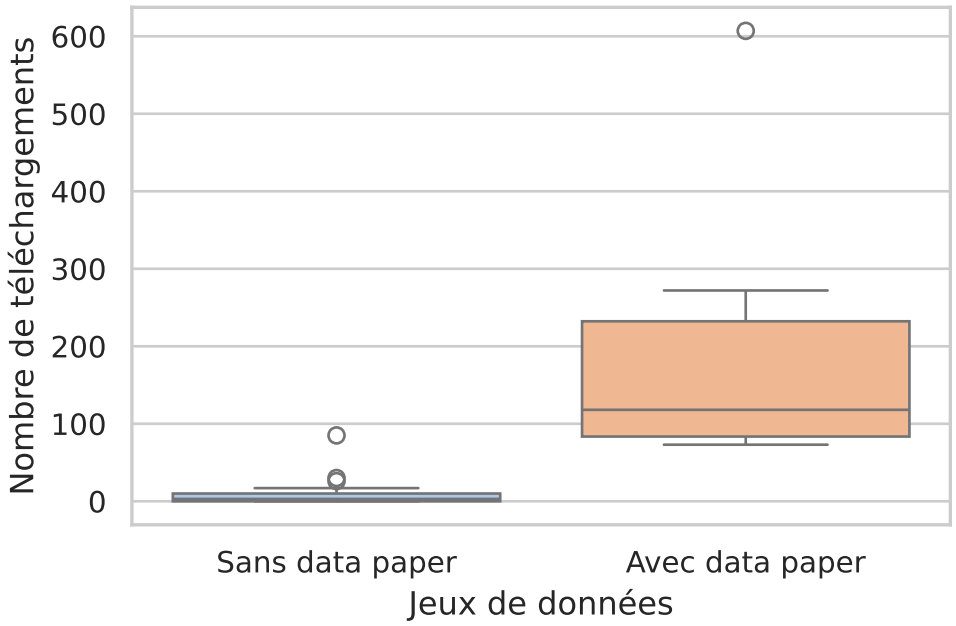

Figure 9: Boîtes moustache du nombre de téléchargements avec et sans Data Paper

Afin de voir l'impact sur la réutilisation des jeux de données accompagnés ou non par un data paper, nous proposons de comparer l'indicateur *Nombre de téléchargement* entre deux listes de jeux de données. La première contient l'ensemble des jeux de données de la collection TETIS sur Recherche.Data.Gouv (dont 1 provenant de l'unique Data paper TETIS (Schaeffer et al. (2022)) dont les données sont déposées dans cet entrepôt). Le second est la liste des jeux de données cités par les data papers de TETIS. La Figure 9 illustre cette comparaison en affichant la boîte à moustache de ces deux distributions.

Nous pouvons observer que la médianne des téléchargement des jeux de données accompagnés par un data paper est nettement supérieur (aux alentours de 110 pour 5). Même les jeux de données parmi les 5% les moins téléchargés de la distribution des data papers, restent nettement supérieur aux 5% les plus téléchargés des sans data papers.

La limite à cette comparaison est que les jeux de données sans data paper proviennent de la collection INRAE, nettement plus récente que celle de CIRAD, et dont les jeux de données sont moins téléchargées (cf Comparaison collections CIRAD / INRAE)

Conclusion

Les data papers sont utiles à l'Open Science en aidant à améliorer la reproductibilité de la recherche. En effet, ils permettent de fournir une documentation complète pour décrire l'origine, la pertinence du jeu de données (pour sa communauté de recherche) et proposent des potentiels cas de réutilisation.

De plus, ils permettent de donner une visibilité importante à la production de données ou logiciels de TETIS. En effet, cette étude montre que les jeux de données accompagnés par des data papers sont beaucoup plus téléchargés (*110* contre *5* téléchargements en mediane). Les data papers constituent donc un moyen efficace pour porter à connaissance notre production de données à nos communautés de recherche.

Le CIRAD a entrepris une démarche d'ouverture de ses données depuis une plusieurs années. Cette politique porte clairement ses fruits, elle a permis de mettre à disposition plus de *120* jeux de données (contre *30* pour INRAE) avec un total de *6000* téléchargements (contre *250* INRAE). Depuis 2021, des initiatives similaires sont en cours à INRAE (soutenu notamment par les Référents Données Opérationnels (RDO) de TETIS), il est nécessaire de les poursuivre.

Plusieurs pistes de reflexion peuvent être menées pour accompagner davantage la réutilisation de notre production de données. Tout d'abord, d'autres indicateurs que le nombre de téléchargements doivent être pris en compte pour évaluer le taux de réutilisation (Est-ce que les jeux de données des data papers ne sont pas automatiquement moissonnés par des plateformes ce qui a pour effet d'augmenter le nombre de téléchargements ? Si oui, comment le mesurer ?). En complément des data papers, quel type de promotion pouvons-nous mettre en place ? Nous pouvons envisager le dépôt des jeux de données dans des entrepôts communautaires (comme HuggingFace pour les modèles d'Intelligence Articficielle à travers le groupe TETIS par exemple). Nous pouvons également organiser des Hackathons comme cela a été fait pour le projet MOOD.

Afin de répondre à ces questions, il serait pertinent de conduire une enquête (sondage) auprès du personnel TETIS afin d'ajouter d'autres indicateurs d'impact. Nous pourrions également recenser les méthodes utilisées à TETIS pour promouvoir la réutilisation de nos jeux de données (avec ou sans data paper).

Annexe

5.1 Champs HAL utilisés

Champ	Description
docid	Identifiant du document dans la base de données HAL.
halId_s	Identifiant HAL unique associé au document.
title_s	Titre du document.
authFullName_s	Noms complets des auteurs.
submittedDate_s	Date de soumission du document.
abstract_s	Résumé du document.
journalDate_s	Date de publication dans le journal.
publicationDate_s	Date de publication du document.
producedDate_s	Date de production du document.
docType_s	Type de document (ART pour article, COUV pour couverture, etc.).
doiId_s	Identifiant DOI associé au document.
journalPublisher_s	Éditeur du journal dans lequel le document a été publié.
journalTitle_s	Titre du journal dans lequel le document a été publié.
journalIssn_s	ISSN du journal dans lequel le document a été publié.
researchData_s	Identifiants (DOI) des données de recherche associées.
docSubType_s	Sous doc type: data paper, preprint, …

5.2 Liste des data papers TETIS

5.2.1 Agritrop & HAL

Table 6: Liste des data papers TETIS (Agritrop & Hal)

titre	authors
VegAnn, Vegetation Annotation of multi-crop RGB images acquired under diverse conditions for segmentation	[‘Simon Madec’, ‘Kamran Irfan’, ‘Kaaviya Velumani’, ‘Frederic Baret’, ‘Etienne David’, ‘Gaetan Daubige’, ‘Lucas Bernigaud Samatan’, ‘Mario Serouart’, ‘Daniel Smith’, ‘Chrisbin James’, ‘Fernando Camacho’, ‘Wei Guo’, ‘Benoit de Solan’, ‘Scott Chapman’, ‘Marie Weiss’]

Table 6: Liste des data papers TETIS (Agritrop & Hal)

titre	authors
Experimental variables in sugarcane intercropping in Reunion Island for data matching	[‘Sandrine Auzoux’, ‘Billy Ngaba’, ‘Mathias Christina’, ‘Benjamin Heuclin’, ‘Mathieu Roche’]
An annotated dataset for event-based surveillance of antimicrobial resistance	[‘Nejat Arınık’, ‘Wim van Bortel’, ‘Bahdja Boudoua’, ‘Luca Busani’, ‘Rémy Decoupes’, ‘Roberto Interdonato’, ‘Rodrique Kafando’, ‘Esther van Kleef’, ‘Mathieu Roche’, ‘Mehtab Alam Syed’, ‘Maguelonne Teisseire’]
LEAP4FNSSA lexicon: Towards a new dataset of keywords dealing with food security	[‘Mathieu Roche’, ‘Agneta Lindsten’, ‘Tomas Lundén’, ‘Thierry Helmer’]
Elaboration of a new framework for fine-grained epidemiological annotation	[‘Sarah Valentin’, ‘Elena Arsevska’, ‘Aline Vilain’, ‘Valérie de Waele’, ‘Renaud Lancelot’, ‘Mathieu Roche’]
Labeled entities from social media data related to avian influenza disease	[‘Camille Schaeffer’, ‘Roberto Interdonato’, ‘Renaud Lancelot’, ‘Mathieu Roche’, ‘Maguelonne Teisseire’]
Partial n-Ary relation instances on food packaging composition and permeability extracted from scientific publication tables	[‘Martin Lentschat’, ‘Patrice Buche’, ‘Luc Menut’, ‘Romane Guari’, ‘Mathieu Roche’]
Food packaging permeability and composition dataset dedicated to text-mining	[‘Martin Lentschat’, ‘Patrice Buche’, ‘Juliette Dibie-Barthelemy’, ‘Luc Menut’, ‘Mathieu Roche’]
Mapping land cover on Reunion Island in 2017 using satellite imagery and geospatial ground data	[‘Stéphane Dupuy’, ‘Raffaele Gaetano’, ‘Lionel Le Mézo’]
Land cover maps of Antananarivo (capital of Madagascar) produced by processing multisource satellite imagery and geospatial reference data	[‘Dupuy Stéphane’, ‘Defrise Laurence’, ‘Gaetano Raffaele’, ‘Andriamanga Valérie’, ‘Rasoamalala Eloise’]
COVID-19 and Media datasets: Period- and location-specific textual data mining	[‘Mathieu Roche’]
PADI-web corpus: Labeled textual data in animal health domain	[‘Julien Rabatel’, ‘Elena Arsevska’, ‘Mathieu Roche’]
A manually annotated corpus in French for the study of urbanization and the natural risk prevention	Koptelov Maksim, Holveck Margaux, Crémilleux Bruno, Reynaud Justine, Roche Mathieu, Teisseire Maguelonne

Table 6: Liste des data papers TETIS (Agritrop & Hal)

titre	authors
Land use / land cover map of Vavatenina region (Madagascar) produced by object-based analysis of very high spatial resolution satellite images and geospatial reference data	Lelong Camille, Herimandimby Hasina
Harmonized in situ datasets for agricultural land use mapping and monitoring in tropical countries	Jolivot Audrey, Lebourgeois Valentine, Leroux Louise, Ameline Maël, Andriamanga Valérie, Bellon Béatriz, Castets Mathieu, Crespin-Boucaud Arthur, Defourny Pierre, Diaz Santiana, Dieye Mohamadou, Dupuy Stéphane, Ferraz Rodrigo, Gaetano Raffaele, Gély Marie, Jahel Camille, Kabore Bertin, Lelong Camille, Le Maire Guerric, Lo Seen Danny, Muthoni Martha, Ndao Babacar, Newby Terrence, de Oliveira Santos Cecília Lira Melo, Rasoamalala Eloise, Simoes Margareth, Thiaw Ibrahima, Timmermans Alice, Tran Annelise, Bégué Agnès

5.2.2 Seulement Hal

Année de publi	Référence
2019	Rabatel, Arsevska, and Roche (2019)
2020	Roche (2020)
2020	Dupuy et al. (2020)
2021	Lentschat et al. (2021)
2022	Valentin et al. (2022)
2022	Schaeffer et al. (2022)
2022	Roche et al. (2022)
2022	Lentschat et al. (2022)
2023	Auzoux et al. (2023)
2023	Madec et al. (2023)
2023	Arınık et al. (2023)

5.3 Liste des jeux de données TETIS sour Entrepôt Recherche Data Gouv

Table 8: Liste des jeux de données TETIS / INRAE

	title
0	Données et produits lidar drone, réserve fores...
1	Package Fordead
2	Labeled Entities from Social Media Data Relate...
3	Données lidar acquises par drone UE Citrus, Sa...
4	Projet Artisols - Coefficient de dispersion pa...
5	Projet Artisols - NoData SPOT 6/7 - Occitanie ...
6	Projet Artisols - Atlas cartographique - Occit...
7	Projet Artisols - Indice de compacité des tach...
8	Projet Artisols - Indice de fragmentation des ...
9	Coupes rases - méthode de détection ”bi-date” ...
10	Panoramas des paysages de Madagascar acquis pa...
11	BioDispersal
12	Carte d'occupation du sol des corridors rivula...
13	Mise au point d’un capteur optique pour le sui...
14	Valorisation des données OCSGE Nouvelle Généra...
15	Valorisation des données OCSGE Nouvelle Généra...
16	Projet GESSICA
17	Valorisation des données OCSGE Nouvelle Généra...
18	Valorisation des données OCSGE Nouvelle Généra...
19	Valorisation des données OCSGE Nouvelle Généra...
20	Valorisation des données OCSGE Nouvelle Généra...
21	Valorisation des données OCSGE Nouvelle Généra...
23	Carte d'occupation du sol du bassin de l'Or 2014
24	Etude exploratoire_sensibilité des cultures à ...
25	ULM_Ciron
26	Avian Influenza events from different digital ...
27	Métabolisme territoire Chaource
28	An annotated Avian Influenza dataset from two ...
29	cartographie des projets de l'UMR TETIS
30	Modèle numérique de terrain du campus d'AgroPa...
31	Annotated datasets from PADI-web for event-bas...
32	Enhanced Spatial Disambiguation in the GeoViru...
33	Données et analyse SmartSens - Data and analys...

5.4 Carte des terrains d'étude des data papers

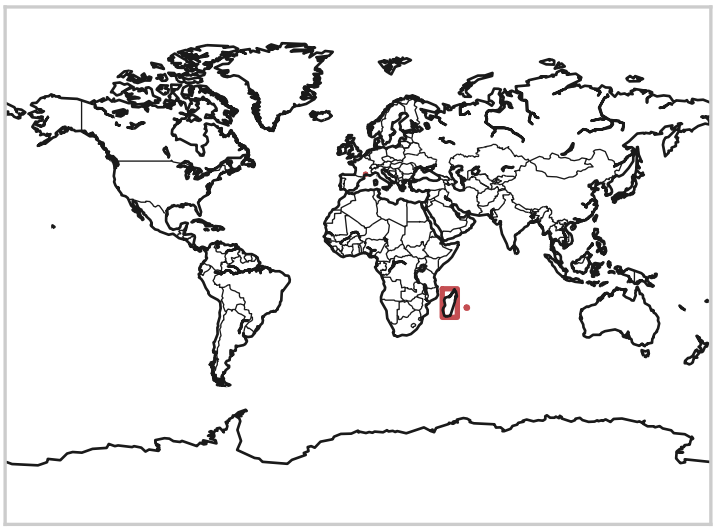

Figure 10: Lieux présents dans les abstract des data papers

Pour finir ce rapport, travaillant à TETIS, il est impossible de ne pas ajouter une carte. La Figure 10 propose de visualiser les étendus géographiques des lieux présent dans les *Abstracts* des data papers.

Pour obtenir la liste des étendues spatiales, nous appliquons un modèle extraction d'entités nommées basé sur le modèle de langue pré-entrainé *BERT*. Ce modèle extrait une liste des lieux que nous géocodons avec l'API de *Photon* utilisant les données d'*OpenStreetMap*.

Bibliography

Arınık, Nejat, Wim Van Bortel, Bahdja Boudoua, Luca Busani, Rémy Decoupes, Roberto Interdonato, Rodrique Kafando, et al. 2023. "An Annotated Dataset for Event-Based Surveillance of Antimicrobial Resistance." *Data in Brief* 46 (February): 108870. https://doi.org/10.1016/j.dib.2022.108870.

Auzoux, Sandrine, Billy Ngaba, Mathias Christina, Benjamin Heuclin, and Mathieu Roche. 2023. "Experimental Variables in Sugarcane Intercropping in Reunion Island for Data Matching." *Data in Brief* 46 (February): 108869. https://doi.org/10.1016/j.dib.2022.108869.

Dedieu, Laurence. 2017. "Revues Publiant Des Data Papers."

Dupuy, Stéphane, Defrise Laurence, Gaetano Raffaele, Andriamanga Valérie, and Rasoamalala Eloise. 2020. "Land Cover Maps of Antananarivo (Capital of Madagascar) Produced by Processing Multisource Satellite Imagery and Geospatial Reference Data." *Data in Brief* 31 (August): 105952. https://doi.org/10.1016/j.dib.2020.105952.

Jolivot, Audrey, Valentine Lebourgeois, Louise Leroux, Mael Ameline, Valérie Andriamanga, Beatriz Bellón, Mathieu Castets, et al. 2021. "Harmonized in Situ Datasets for Agricultural Land Use Mapping and Monitoring in Tropical Countries." *Earth System Science Data* 13 (12): 5951–67. https://doi.org/10.5194/essd-13-5951-2021.

Lentschat, Martin, Patrice Buche, Juliette Dibie-Barthelemy, Luc Menut, and Mathieu Roche. 2021. "Food Packaging Permeability and Composition Dataset Dedicated to Text-mining." *Data in Brief* 36 (June): 107135. https://doi.org/10.1016/j.dib.2021.107135.

Lentschat, Martin, Patrice Buche, Luc Menut, Romane Guari, and Mathieu Roche. 2022. "Partial n-Ary Relation Instances on Food Packaging Composition and Permeability Extracted from Scientific Publication Tables." *Data in Brief* 41 (April): 108000. https://doi.org/10.1016/j.dib.2022.108000.

Madec, Simon, Kamran Irfan, Kaaviya Velumani, Frederic Baret, Etienne David, Gaetan Daubige, Lucas Bernigaud Samatan, et al. 2023. "VegAnn, Vegetation Annotation of Multi-Crop RGB Images Acquired Under Diverse Conditions for Segmentation." *Scientific Data* 10 (1): 302. https://doi.org/10.1038/s41597-023-02098-y.

Rabatel, Julien, Elena Arsevska, and Mathieu Roche. 2019. "PADI-Web Corpus: Labeled Textual Data in Animal Health Domain." *Data in Brief* 22 (February): 643–46. https://doi.org/10.1016/j.dib.2018.12.063.

Roche, Mathieu. 2020. "COVID-19 and Media Datasets: Period- and Location-Specific Textual Data Mining." *Data in Brief* 33 (December): 106356. https://doi.org/10.1016/j.dib.2020.106356.

Roche, Mathieu, Agneta Lindsten, Tomas Lundén, and Thierry Helmer. 2022. "LEAP4FNSSA Lexicon: Towards a New Dataset of Keywords Dealing with Food Security." *Data in Brief* 45 (December): 108680. https://doi.org/10.1016/j.dib.2022.108680.

Schaeffer, Camille, Roberto Interdonato, Renaud Lancelot, Mathieu Roche, and Maguelonne Teisseire. 2022. "Labeled Entities from Social Media Data Related to Avian Influenza Disease." *Data in Brief* 43 (August): 108317. https://doi.org/10.1016/j.dib.2022.108317.

Valentin, Sarah, Elena Arsevska, Aline Vilain, Valérie De Waele, Renaud Lancelot, and Mathieu Roche. 2022. "Elaboration of a New Framework for Fine-Grained Epidemiological Annotation." *Scientific Data* 9 (1): 655. https://doi.org/10.1038/s41597-022-01743-2.