***In* Partage et valorisation des données de la recherche. Développements, tendances et modèles, Chapitre 5, pp. 87-106**
May 2023
Eds Schöpfel Joachim, Rebouillat Violette
ISBN 9781789480733
https://archimer.ifremer.fr/doc/00836/94838/

Archimer
https://archimer.ifremer.fr

SEANOE - un entrepôt thématique

Merceur Frédéric [1], Petit De La Villeon Loic [1], Van Iseghem Sylvie [1]

[1] Ifremer, France

Please note that this is an author-produced PDF of an article accepted for publication following peer review. The definitive publisher-authenticated version is available on the publisher Web site.

Comment décrire la mission générale de l'entrepôt SEANOE ?

SEANOE[1] permet de publier facilement, rapidement et gratuitement des jeux de données en sciences marines. Il est ouvert à l'ensemble de la communauté scientifique internationale dans le domaine de la recherche marine. SEANOE est uniquement dédiées aux données des différents domaines des sciences marines, comme l'océanographie physique, la géologie marine, la biologie marine, etc.

Chaque jeu de données publié par SEANOE bénéficie d'un DOI (identifiant numérique d'objet utilisé comme mécanisme d'identification des ressources) dans lequel l'auteur (le producteur de la donnée) est clairement identifié. Cela permet une citation précise, fiable et pérenne.

SEANOE offre ainsi une solution pour répondre aux revues qui demandent que les données utilisées dans un article soient accessibles en ligne (ex. : PLoS ONE).

Les données publiées par SEANOE peuvent être utilisées en respectant les conditions de la licence Creative Commons sélectionnée par l'auteur des données. Un embargo limité à 2 ans sur un jeu de données est possible (pour les publications scientifiques en cours).

SEANOE accepte des jeux de données dont la taille est inférieure à 100 Go.

Quel est son contexte administratif ?

SEANOE est géré par l'équipe par le centre de données Sismer[2].

En 1971, le gouvernement français a confié au CNEXO (Centre National pour l'Exploitation des Océans), la mission de représenter la France auprès de IOC (Intergovernmental Oceanographic Commission) de UNESCO. A cette époque, seule la gestion des données issues des campagnes océanographiques était assurée. Lors de la création de l'Ifremer en 1984, résultat de la fusion des organismes CNEXO et ISPTM, cette mission lui a été transférée.

L'Ifremer a alors mis en place le centre de données Sismer (Systèmes d'Information Scientifique pour la Mer) pour gérer la collecte, le traitement et la diffusion de ces données marines. Le périmètre des données prises en charges a été étendu progressivement depuis 1984. Le personnel du Sismer est aujourd'hui une équipe de 28 personnes (21 CDI, 6 CDD). Le Sismer conduit et/ou participe à un ensemble de banques de données thématiques nationales et internationales :

- La base internationale d'océanographie physique Coriolis[3] ;
- La base nationale Harmonie du SIH[4] (Système d'information halieutique) de données de la pêche ;
- La base nationale Quadrige pour de données environnementales ;
- Le catalogue des campagnes océanographiques françaises[5] ;
- Les bases CATDS et CERSAT de données océanographiques spatiales.

Le centre de données Sismer bénéficie de l'aide de deux autres équipes dans ses missions :

- L'équipe ISI (14 CDI, 7 CDD) gère, dans un contexte de sous-traitance, le développement de ces banques données ;

[1] https://www.seanoe.org
[2] https://data.ifremer.fr/SISMER
[3] http://www.coriolis.eu.org/
[4] http://sih.ifremer.fr/
[5] https://campagnes.flotteoceanographique.fr/

- L'équipe RIC (20 CDI, 4 CDD) maintient les infrastructures techniques (ex : serveurs) de l'Ifremer. Environ 40% de l'activité de l'équipe RIC est dédiée à la gestion des infrastructures de ces banques de données.

En 2019, le budget total de ces trois équipes était de 5,3 M€. Le budget est financé à 100% par l'Ifremer.

Au niveau national, Sismer est désormais une composante du pôle de données pour l'océan Odatis[6], qui fédère au niveau national des activités de gestion de données et d'expertise scientifique en océanographie.

Comment fonctionne l'équipe de SEANOE ?

SEANOE est hébergé sur les serveurs centraux mutualisés de l'Ifremer qui sont administrés par l'équipe RIC (Oracle, ElasticSearch, Tomcat, système de fichiers, système d'archivage). Le coût spécifique pour l'Ifremer du projet SEANOE en termes d'infrastructure et de gestion de cette infrastructure est donc marginal dans la mesure où il est basé sur des services existants et mutualisé avec un grand nombre de services.

Le projet SEANOE est piloté par un chef de projet de l'équipe Sismer. La maintenance évolutive du code de SEANOE est aujourd'hui sous-traitée à une société de service (Altran). En 2015, l'adaptation d'Archimer pour mettre en place SEANOE a couté 4 mois de travail. Depuis 2015, environ un mois de travail annuel est nécessaire pour assurer la maintenance évolutive. La gestion de la sous-traitance est assurée par une personne de l'équipe ISI qui assure également une partie du support aux usagers.

La validation des dépôts est actuellement assurée à tour de rôle par deux personnes de l'équipe Sismer (dont la responsable du projet). En cas d'absence, la validation est assurée par le service d'assistance de l'équipe Sismer. Un service continu est donc assuré

Nous contactons systématiquement les déposants, pour leur proposer notre aide, qui ont initié un dépôt mais qui ne l'ont pas finalisé immédiatement. C'est souvent l'occasion de répondre à des questions sur la gestion des versions par exemple.

Nous devons également, de temps en temps, apporter notre aide pour uploader les fichiers de données quand la taille de ceux-ci dépasse les 20-30 Go. En cas de difficulté, nous proposons aux usagers de mettre les données à disposition sur un site ftp de leur organisme pour que nous puissions les télécharger et les uploader nous-mêmes dans SEANOE à partir du réseau Ifremer.

Au total, entre une demie et une journée par semaine sont nécessaire pour :

- Le pilotage du projet ;
- L'assistance aux usagers (par exemple, répondre aux questions sur la gestion des versions, etc.) ;
- La validation et la mise à jour des dépôts ;
- La veille sur les citations ;
- La sollicitation de nouveaux dépôts auprès des auteurs d'articles corédigés par l'Ifremer ;
- La gestion de la sous-traitance du code informatique.

Quand a été lancé SEANOE, et quels ont été les besoins ?

SEANOE a été lancé en novembre 2015.

[6] https://www.odatis-ocean.fr/

Le contexte réglementaire Européen puis Français avec la loi sur la république numérique promulguée en 2016 recommande une diffusion en libre accès des données de la recherche publique. De plus des éditeurs comme PLoS imposent que les données exploitées dans un article soient accessibles librement en ligne et citées à l'aide d'un DOI ; la publication simultanée des données et d'un article peut renforcer la crédibilité de l'étude.

Or, en 2015, le centre de données Sismer n'est pas organisé pour répondre à ce besoin :

- Le Sismer gère un ensemble de bases de données qui couvrent l'essentiel des thématiques des sciences marines. Les données qui alimentent ces bases de données sont issus de dispositifs automatisés (ex. : flotteurs autonomes Argo), de réseaux de surveillance nationaux (ex. : Rephy), de données de campagnes océanographiques…
- L'ingestion d'un jeu de données dans une de ces bases se fait souvent manuellement avec la nécessité de transformer le format des données et les de qualifier avant de pouvoir les charger dans une des bases de données.
- Certaines de ces bases ne sont pas en libre accès (ex. : données de pêche).
- Les données chargées dans ces bases sont anonymisées et ne permettent pas de créditer un auteur en particulier.

Quel était le cahier des charges initial ?

Dans le contexte cité ci-dessus, Sismer a alors décidé de développer un système de publication de jeux de données qui offre une interface web de dépôt simple, un système de validation et la possibilité d'attribuer un DOI à ces jeux de données. Le modèle que nous avions alors était la base Pangaea développé et maintenu en Allemagne[7].

Les scientifiques qui publient un jeu de données pour le citer dans un article, le faisaient souvent en urgence, parfois après avoir subi un refus de la part de l'éditeur car les données exploitées n'étaient pas accessibles en ligne (c'est un peu moins vrai aujourd'hui où une partie des auteurs a intégré cette exigence de la part des éditeurs et commencent à anticiper la publication de leurs données avant de soumettre un article). Le système devait donc permettre l'attribution d'un DOI rapidement, idéalement en moins de 24 heures.

Les données devaient être accessibles librement, après un éventuel embargo limité dans le temps.

Pour le Sismer, l'objectif de ce nouveau service était également d'élargir le périmètre de collecte des données et de capter de nouveaux jeux de données susceptibles d'alimenter ses bases thématiques.

Lors de la validation, le Sismer doit uniquement s'assurer que le jeu de données déposés est bien un jeu de données, que la thématique est bien respectée, et qu'il est suffisamment décrit. De fait, seuls deux dépôts ont pour l'instant été refusés : un article déposé par erreur dans SEANOE et un jeu de données insuffisamment décrit et dont l'auteur semblait incapable de décrire correctement.

Lors de la validation, le Sismer peut réclamer des métadonnées supplémentaires (ex. : l'emprise géographique, l'affiliation des auteurs, …), recommander, quand c'est possible, de convertir les données dans un autre format (ex. : Excel vs csv), ou de mieux décrire les données (ex. : ajouter des unités manquantes dans des colonnes de données).

La taille des jeux de données acceptées ne doit pas excéder 100 Go. Cette limite a été fixée car, au-delà, il commence à devenir difficile pour les usagers de télécharger les données. Il est alors préférable

[7] Cf. https://www.pangaea.de/

de sélectionner, des systèmes adossés à des moyens de calculs où les données sont directement exploitables sans avoir à les télécharger.

Ce cahier des charges n'a pas été modifié depuis le lancement.

Quels moyens techniques ont été mobilisés pour SEANOE ?

Pour mettre en place ce service, deux pistes ont été envisagées : utiliser l'infrastructure Sextant[8], un Système d'information Géographique (SIG) basé sur le système GeoNetWork ; ou adapter Archimer[9], l'archive Institutionnelle d'Ifremer. La solution Archimer présentait alors l'avantage de proposer un système de dépôt simple ainsi qu'un système de validation et d'être très rapidement opérationnelle. C'est la solution qui a été retenue.

SEANOE est donc un sous-ensemble d'Archimer, l'Archive Institutionnelle de l'Ifremer. Archimer est une réalisation interne du service de documentation adossée aux serveurs centraux de l'Ifremer (Oracle, ElasticSearch, Tomcat). Archimer a été adaptée pour créer SEANOE. Ce sont deux déclinaisons d'un même système. Le code de SEANOE est commun a plus de 95% avec celui d'Archimer. La très grande majorité des métadonnées de description d'un rapport dans Archimer et d'un jeu de données dans SEANOE sont communes. Seules quelques métadonnées spécifiques ont été ajoutées pour décrire un jeu de données (ex. : étendue temporelle, niveau de traitement des données, zone géographique…). Le formulaire de chargement des fichiers a également dû être adapté. En effet, SEANOE accepte des jeux de données dont la taille peut aller jusqu'à 100 Go. Pour charger en ligne de tels volumes de fichiers, il faut implémenter des mécanismes particuliers qui permettent de gérer des chargements qui peuvent durer plusieurs heures sans risque d'erreur http de type « TimeOut ».

Les données sont publiées sous le nouveau nom de domaine seanoe.org. Nous avons en effet pris le soin d'anonymiser SEANOE par rapport à l'Ifremer pour ne pas donner l'impression que l'Ifremer souhaitait s'approprier les données publiées.

L'interface et les métadonnées sont uniquement disponibles en anglais. Même pour attirer les jeux de données d'équipes françaises, il faut se présenter comme un service international.

Comment assurez-vous la promotion de SEANOE auprès des chercheurs ? Comment communiquez-vous avec eux ?

Nous avons présenté SEANOE dans l'ensemble des unités de recherche de l'Ifremer. Nous l'avons également présenté à plusieurs réseaux de collecte de données nationaux.

Nous avons contacté les éditeurs PLOS, Elsevier, Nature pour qu'ils listent SEANOE dans leurs listes d'entrepôts de données recommandés.

Enfin, nous adressons, depuis 2016, un message aux auteurs de l'ensemble des articles corédigés par des auteurs des UMR auxquelles l'Ifremer est associée pour leur proposer de publier les données associées à leurs articles dans SEANOE. Les retours positifs immédiats à ces sollicitations sont inférieurs à 1%. Par contre, quand ces auteurs sont confrontés à une demande d'un éditeur qui impose la publication des données pour accepter un article, ils se souviennent alors parfois du service proposé par SEANOE.

Comme nous nous y attendions, et c'est confirmé par les statistiques d'utilisation, Google est la principale source d'accès aux jeux de données dans SEANOE ; en 2019, 50% des consultations

[8] https://sextant.ifremer.fr/
[9] https://archimer.ifremer.fr/

proviennent de Google. Améliorer la visibilité d'un jeu de données, c'est donc avant tout travailler sur son référencement et c'est quelque chose que nous travaillons particulièrement.

Un bon référencement ne peut être obtenu sans métadonnées de qualité et parmi les métadonnées, le titre est particulièrement important. Quand un auteur soumet un jeu de données, nous refusons donc par exemple, les titres qui ne sont pas explicites. Un exemple : un jeu de données initialement publié avec le titre « *GULF_IND_PLOS_One* » sera finalement publié sous le titre « *Output from a 1/12-degree Global experiment with the Hybrid Coordinate Ocean Model (HYCOM), forced with with NCEP Reanalysis products - Data for the Persian Gulf and Strait of Hormuz* ».

Techniquement, le code HTML des landing pages des jeux de données est ensuite structuré pour optimiser sa visibilité. Il contient par exemple les métadonnées structurées au format JSON-LD schema.org.

Enfin, nous cherchons à enrichir le réseau de liens croisés automatiques entre les jeux de données publiés dans SEANOE et des ressources extérieures. Google comptabilise le nombre de liens vers une ressource dans son calcul de popularité.

A ce stade, un jeu de données publié dans SEANOE peut proposer des liens croisés vers :

- La page ORCID de ses auteurs ;
- Le CV de ses auteurs Ifremer ;
- Des documents dans Archimer ;
- Des lots d'images dans l'Océanothèque de l'Ifremer ;
- Des campagnes dans le catalogue des campagnes océanographiques françaises ;
- Des échantillons dans le catalogue des campagnes océanographiques françaises ;
- Les articles Elsevier et Taylor & Francis via Scholix.

Enfin, le projet SeaDataNet et le pôle de données Odatis, dans lesquels l'Ifremer est un élément moteur, ont sélectionné SEANOE comme leur service d'attribution de DOI.

Nous effectuons également une veille sur les citations des jeux de données dans les articles internationaux à l'aide d'alertes sur le mot SEANOE et le préfix des DOI dans Google Scholar et sur le site de plusieurs éditeurs (ex. : Elsevier). Quand nous repérons un article qui cite un jeu de données, nous ajoutons sa référence à la notice SEANOE. Sa notice est alors mise à jour automatiquement auprès de DataCite avec le couple DOI de données / DOI d'article. Ces couples sont ensuite également poussés par DataCite vers le projet Scholix qui est utilisé par plusieurs éditeurs.

Quand les auteurs de l'article ne sont pas les auteurs du jeu de données, nous signalons la citation aux auteurs de la donnée.

Parfois le jeu de données est mal cité. L'erreur la plus courante est la citation de l'URL de la landing page à la place du DOI. Nous signalons alors l'erreur aux auteurs de l'article pour qu'ils corrigent la citation si ce n'est pas trop tard.

Enfin, pour quelques jeux de données majeurs qui bénéficient d'un nom qui n'est pas trop concurrentiel sur Internet (ex. : Argo, Rephy, …) nous avons également mis en place une veille sur le nom du jeu de données. Si nous repérons un article qui cite un jeu de données sans citer le DOI correspondant, nous interpellons les auteurs de l'article pour qu'ils ajoutent la citation du DOI SEANOE. A défaut, quand c'est trop tard, les auteurs promettent généralement de respecter la consigne de citation dans leurs prochains articles.

Quels ont été les changements depuis le lancement ?

SEANOE a assez peu évolué depuis son lancement. Dès le début, nous nous sommes aperçus qu'il manquait quelques champs de métadonnées. Par exemple, lors des premiers dépôts, plusieurs scientifiques enregistraient leurs remerciements à la fin du champ description. Nous avons donc ajouté un champ « Acknowledgement » spécifique.

Pour les besoins spécifiques du projet Argo[10], nous avons développé en avril 2016, la possibilité de gérer plusieurs versions d'un jeu de données au sein du même DOI et d'y accéder de manière différenciée à l'aide de fragment (#). Cette possibilité est depuis ouverte à l'ensemble des jeux de données publiés dans SEANOE.

Pour les données Argo, une image (snapshot) de l'ensemble des données est figée et conservée tous les mois. Dans une première version, un DOI principal avait été attribué au jeu de données Argo et des DOI spécifiques avaient été attribués à chaque snapshot mensuel.

A la demande de la direction scientifique du projet Argo, un DOI unique a été attribué par SEANOE aux données Argo. Ce DOI unique permet de citer le jeu de données global ou un snapshot spécifique à l'aide du même DOI. Dans cette perspective, chaque snapshot est uploadé dans SEANOE qui lui attribue une URL et une clé. La clé 42350 a par exemple été attribuée au snapshot du 2016-02-08.

La citation du jeu de données global s'effectue en citant le nouveau DOI sans paramètre. Un exemple :

- *Argo (2000). Argo float data and metadata from Global Data Assembly Centre (Argo GDAC). SEANOE. http://doi.org/10.17882/42182*

La citation d'un snapshot spécifique se fait en ajoutant sa clé précédée du caractère # au DOI :

- *Argo (2016). Argo float data and metadata from Global Data Assembly Centre (Argo GDAC) – Snapshot of Argo GDAC of February, 8th 2016. Seanoe. http://doi.org/10.17882/42182#42350*

Cette possibilité d'attribuer un DOI unique à un jeu de données évolutif présente de nombreux avantages, en particulier :

- Les consignes de citation sont plus simples à donner aux utilisateurs du jeu de données. Ils n'ont qu'un seul DOI à appréhender. C'est un point important car mêmes les consignes simples ont du mal à être appliquées : nous nous en apercevons par exemple en constatant que, régulièrement, des scientifiques demandent un DOI pour citer leurs données dans un article, et, au final, ils citent l'URL de la landing page à la place du DOI.
- Pour améliorer la visibilité du DOI dans les moteurs de recherches (ex. : Google), il est préférable de diffuser une seule landing page qui concentre un maximum de citations et donc de liens (backlink) plutôt que de multiples landing page de contenu quasiment identiques.
- Si un DOI spécifique est attribué à chaque version, les moteurs de recherches ne présenteront par forcément la version la plus récente en avant dans leurs listes de résultats. Un internaute peut donc découvrir une version obsolète du jeu de données et ne pas se rendre compte qu'il existe une version plus récente.

Récemment, nous avons mis en place la possibilité de dupliquer les données publiées SEANOE dans EMODnet Ingestion. EMODnet (European Marine Observation and Data Network) est un réseau

[10] http://www.argo.ucsd.edu/

d’organisations soutenu par la politique maritime intégrée de l'UE. Le projet EMODnet gère un ensemble de bases données thématiques, dont :

- EMODnet Biology (https://www.emodnet-biology.eu)
- EMODnet Bathymetry (https://www.emodnet-bathymetry.eu)
- EMODnet Physics (https://www.emodnet-physics.eu)

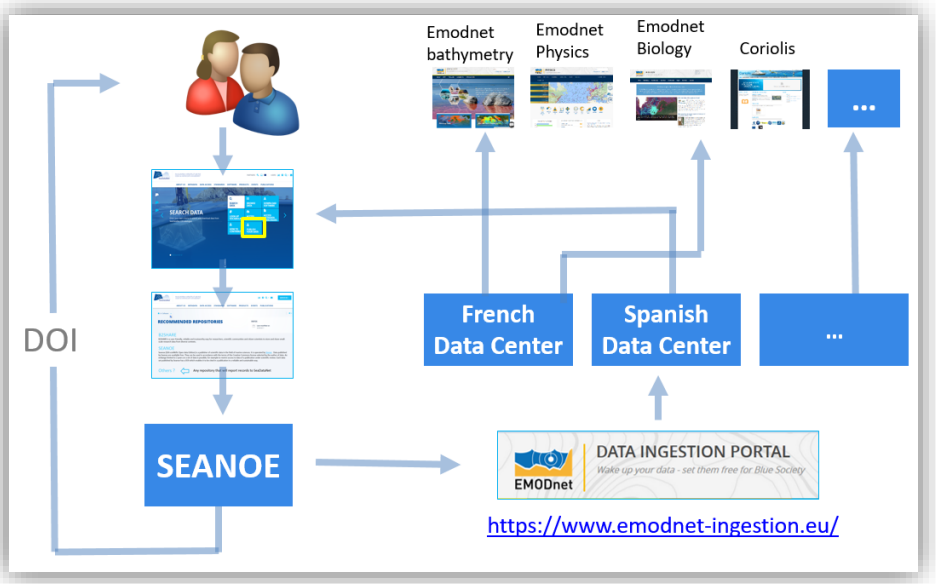

Figure 1 : La duplication des données SEANOE dans EMODnet

Les jeux de données publiés dans SEANOE peuvent être poussés automatiquement dans EMODnet Ingestion (cf. figure 1). EMODnet Ingestion les attribue alors au centre de données national correspondant à la nationalité du « corresponding auteur ». Ce centre de données national doit alors vérifier si les données peuvent être ingérées dans l'un ou l'autre de ces portails et si c'est le cas, les mettre en forme et les ingérer dans la base thématique la plus adéquate. L’équipe Sismer effectue une sélection des jeux de données à dupliquer de SEANOE à EMODnet Ingestion : les jeux de données sont dupliqués quand les données sont qualifiées, que la licence de diffusion sélectionnée le permet et que le jeu de données n’est pas déjà issu de l’export d’une base internationale.

Par ailleurs, le lien avec EMODnet est à ce jour l’uniquement interconnexion actuelle de SEANOE avec une autre infrastructure.

Quel est à ce jour le nombre de dépôts ?

Le 15 mai 2020, 594 jeux de données étaient publiés dans SEANOE. Depuis fin 2018, deux dépôts spontanés sont enregistrés toutes les semaines en moyenne. Plus en détail : en 2019, 118 jeux de données ont été publiés par SEANOE dont 108 déposés spontanément par les auteurs et 10 dépôts liés au projet EGO Gliders. Les dépôts spontanés sont très majoritairement liés à des publications d’articles. Plusieurs éditeurs (ex. : PLoS ONE, Elsevier, …) imposent désormais que les données exploitées dans un article soient accessibles librement en ligne et citées à l’aide d’un DOI. Figure 2 montre l’évolution des dépôts.

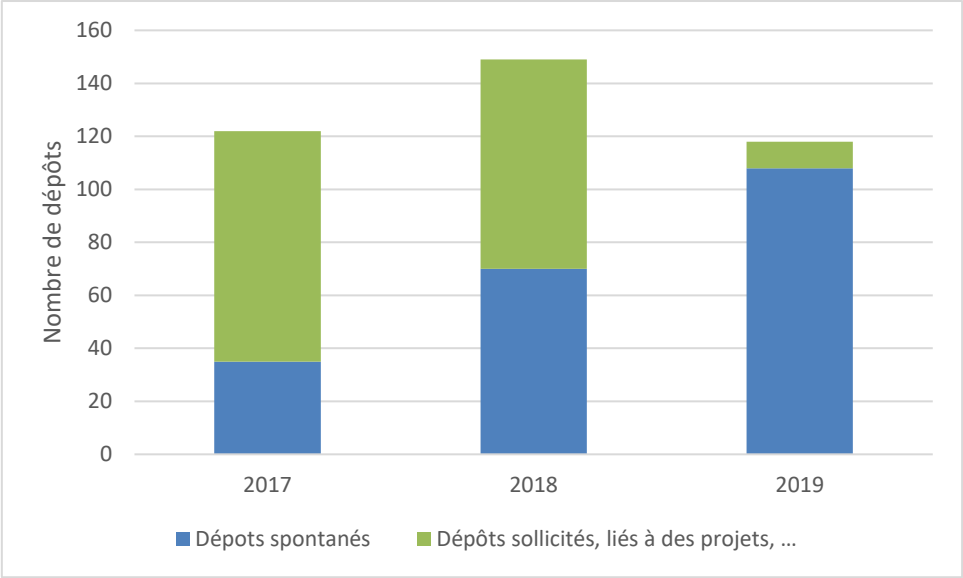

Figure 2 : Nombre de dépôts dans SEANOE 2017-2019

Figure 3 présente la thématique des jeux de données déposés en 2019. Un dépôt peut être classé dans plusieurs thématiques. Dans ce graphique, seuls les jeux de données déposés spontanément par leurs auteurs ont été pris en compte. Les domaines les plus représentatifs sont l'océanographie physique, biologique et chimique.

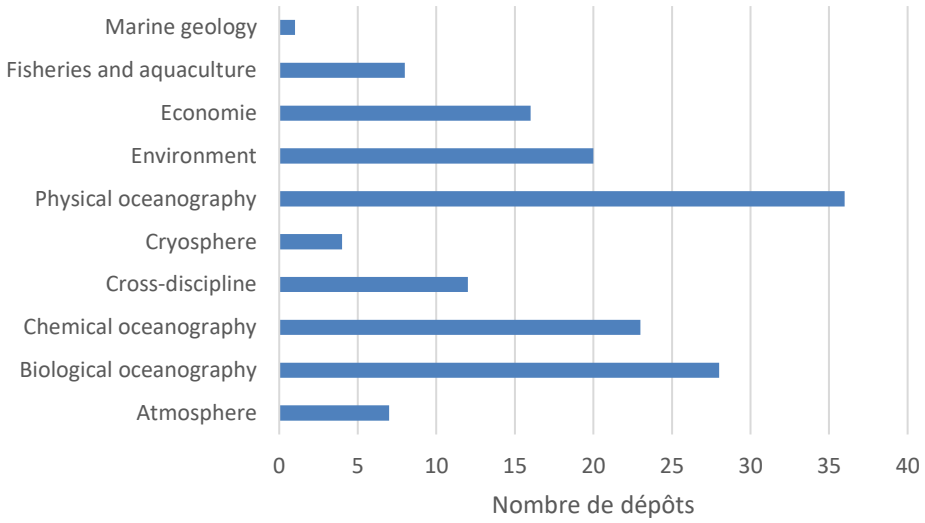

Figure 3 : Les thématiques des dépôts spontanés en 2019 (108)

La somme des fichiers de données déposés en 2019, si on inclut les mises à jour de versions (ex. : Argo, Cora, …) est de 0,56 To. La taille médiane des dépôts est de 10 à 100 Mo. Figure 4 montre la répartition du volume des données, DOI par DOI. Si un jeu de données est publié sous la forme de plusieurs fichiers, la taille des fichiers est cumulée.

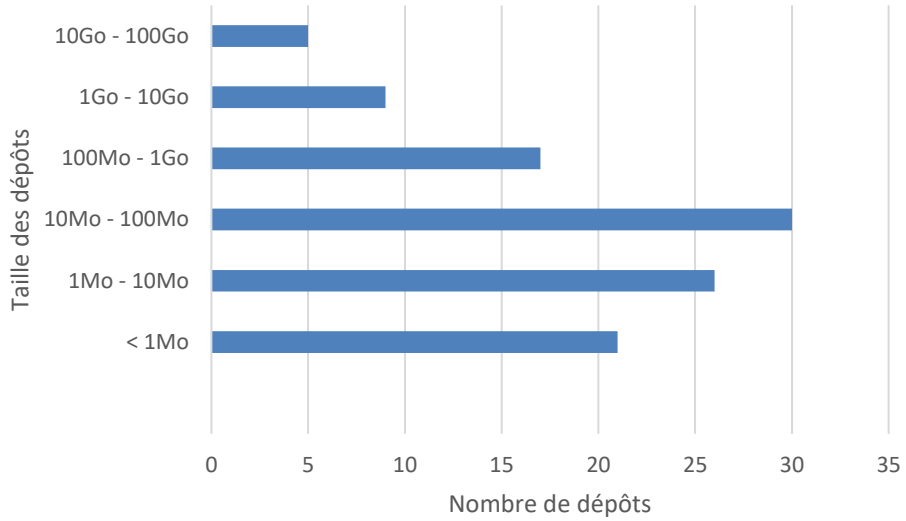

Figure 4 : Volumes des fichiers de données par DOI pour les dépôts 2019

Quant au format des dépôts, NetCDF et CSV sont les formats les plus utilisés. Les fichiers qui sont fournis au format Excel sont généralement transformés en CSV avant la publication. Mais ce n'est pas toujours possible quand, par exemple, le fichier Excel contient des éléments de présentation nécessaires à la compréhension des données. Les fichiers PDF proposent généralement des informations de description des données (format, acquisition). Certains jeux de données sont composés d'un grand nombre de fichiers, parfois de formats différents, qui sont alors zippés.

Quels sont les usages en termes de connexions, de sessions utilisateurs, de téléchargements… ?

Voici quelques éléments d'information à partir de l'analyse des logs du serveur Web Apache de l'Ifremer. En 2019, les landing pages ont été consultées 24 000 fois depuis l'extérieur de l'Ifremer (hors robots). Les fichiers de données ont été téléchargés plus de 9 700 fois. La progression depuis 2016 a été significative ; entre 2017 et 2019, consultations et téléchargements ont plus que doublés (Figure 5).

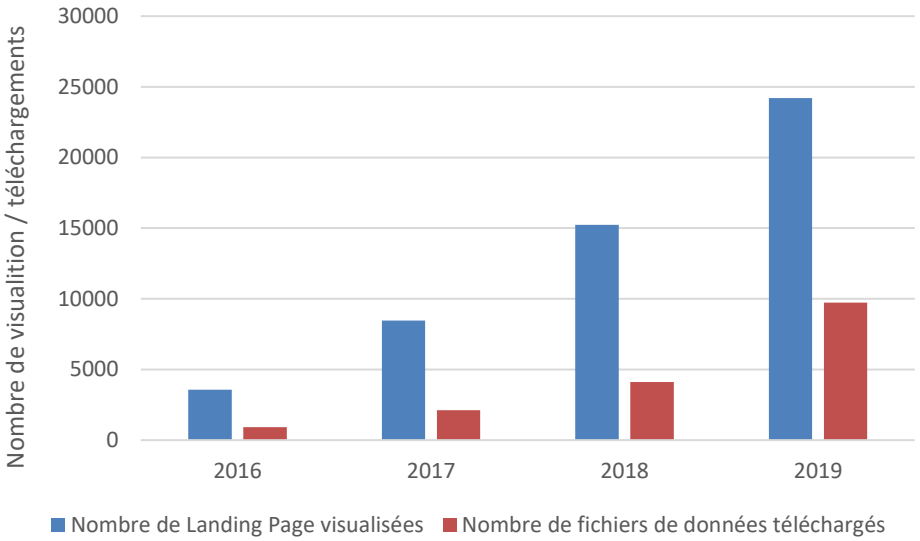

Figure 5 : Consultations et téléchargements depuis l'extérieur de l'Ifremer (2016-2019)

Le dépôt le plus demandé a été téléchargé 1002 fois en 2019 (« Labeled SAR imagery dataset of ten geophysical phenomena from Sentinel-1 wave mode [TenGeoP-SARwv] ») ; les dix jeux de données le plus téléchargés représentent 46% des téléchargements en 2019.

Les consultations et téléchargements proviennent de tous les continents (cf. Figure 6). Les premiers pays, en termes d'usages, sont la Chine, la France, les Etats Unis, l'Inde et l'Italie, suivis par le Royaume Uni et l'Allemagne.

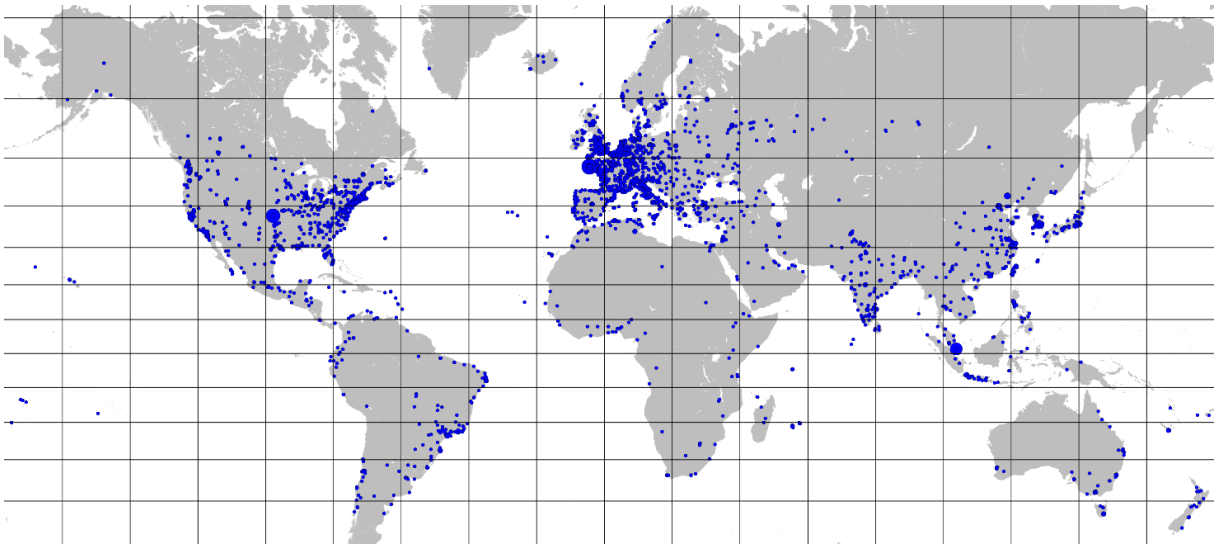

Figure 6 : Origine géographique des visualisations des landings pages en 2019 (sans Ifremer)

Que sait-on des déposants ? Quel est le retour de la part des utilisateurs ?

Entre janvier et mai 2019, 47 nouveaux jeux de données ont été publiés dans SEANOE. La quasi-totalité de ces jeux de données ont été publiés par des scientifiques ou des ingénieurs dans la perspective d'une citation du jeu de données dans un projet d'article. 31 de ces 47 jeux de données ont été publiés par des scientifiques français. Les autres nationalités sont diversifiées (Australie, Brésil, Canada, Chine, Allemagne, Italie, Japon, Corée, Mexique, Qatar, Etats-Unis).

Les retours des scientifiques qui publient leurs données sont globalement très positifs. De fait, nous revoyons régulièrement les mêmes scientifiques publier de nouveaux jeux de données. Parmi les points positifs, l'aide apportée lors du dépôt et la rapidité du service reviennent le plus souvent.

Pour vous, quels sont les points forts de SEANOE ?

SEANOE a bénéficié de l'expérience du projet Archimer en matière de publication. Les mécanismes de référencement dans les moteurs de recherche comme Google n'avaient par exemple jamais été mis en œuvre jusqu'à présent dans les systèmes gérés par le Sismer. Les liens avec ORCID, la gestion des affiliations des auteurs, les liens avec les DOI d'articles, la fourniture de statistiques de téléchargement aux usagers, sont d'autres exemples de l'apport du monde documentaire à ce projet de publications de données.

Inversement, la gestion de fichiers de plusieurs dizaines de Go, les conseils sur les formats de données, les liens avec le projet EMODnet Ingestion, sont des apports de l'équipe Sismer en termes de gestion de données.

Parmi les points forts de SEANOE, nous citerions :

- La disponibilité et l'implication du support lors du dépôt et après la publication du jeu de données ;
- La pérennité du système adossé à un centre de données qui existe depuis 1984 ;
- La veille sur les citations des articles ;
- Le système de gestion de versions à l'aide fragment au sein du même DOI ;
- Les mécanismes de référencement dans Google ;
- La richesse du réseau de liens croisés automatiques ;
- La lisibilité et la richesse des landing pages des DOI.

Est-ce que SAENOE prend en compte les principes FAIR ?

D'une manière générale, oui, nous nous efforçons de respecter les recommandations FAIR dans SEANOE. A noter que le centre de données Sismer, ce qui inclut SEANOE, dispose de la certification CoreTrustSeal[11]. Le Sismer a aussi été accrédité comme centre national de données océanographique par le programme IODE[12].

Voici en détail la conformité de SEANOE avec les différents principes FAIR[13]

Findability (faciles à trouver)

F1. (Meta)data are assigned a globally unique and persistent identifier

C'est bien le cas, chaque jeu de données publiés dans SEANOE dispose d'un DOI.

F2. Data are described with rich metadata (defined by R1 below)

SEANOE impose une liste de métadonnées obligatoires (titre, auteurs, description, …) et propose un ensemble de métadonnées optionnelles. Nous refusons de valider les dépôts que nous jugeons insuffisamment décrits (ex. : titre non explicite, …). Mais le détail de la description d'un jeu de données diffère selon les auteurs.

F3. Metadata clearly and explicitly include the identifier of the data they describe

La landing page du jeu de données présente de manière claire le DOI à la fois dans les métadonnées et dans la suggestion de citation. De plus, certains auteurs nous demandent de leur réserver un DOI avant le dépôt pour qu'ils puissent l'ajouter dans les fichiers de données. C'est une pratique que nous devrons encourager à l'avenir.

F4. (Meta) data are registered or indexed in a searchable resource

Les jeux de données sont accessibles depuis le front-office de SEANOE. Les métadonnées sont également stockées de manière structurée au format schema.org dans les landing pages ce qui permet à Google de les indexer dans son outil Dataset Search. Les métadonnées sont également accessibles via le moteur OAI-PMH de SEANOE. L'API REST de SEANOE devrait être rendue accessible sur Internet dans une prochaine version (même si nous n'avons pas eu de demande dans ce sens pour l'instant).

Accessibility (accessibles)

[11] https://www.coretrustseal.org/wp-content/uploads/2019/11/IFREMER-SISMER.pdf
[12] https://archimer.ifremer.fr/doc/00389/50015/50604.pdf
[13] Cf. https://www.go-fair.org/fair-principles/

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

Les métadonnées et les données sont accessibles librement sans authentification à l'aide de liens HTTPS (à l'exception des données sous embargo qui sont accessibles à la demande).

A2. Metadata are accessible, even when the data are no longer available

Dans les conditions d'utilisations de SEANOE il est spécifié qu'il est interdit de demander le retrait d'un jeu de données déjà publié. Il est possible de mettre à jour un jeu de données mais pas de le supprimer. Dans les faits, il nous arrive de replacer un jeu de données sous embargo à la demande d'un auteur le temps qu'il diffuse une correction quand une erreur est détectée dans son jeu de données.

Interoperability (intéroperables)

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

Les métadonnées sont accessibles au format JSON LD depuis les landing pages et au format Dublin Core depuis le moteur OAI-PMH.

I2. (Meta)data use vocabularies that follow FAIR principles

C'est un point que nous devons améliorer. Par exemple en implémentant la possibilité de lister les paramètres à l'aide de listes de vocabulaires définis dans le projet SeaDatanet.

I3. (Meta)data include qualified references to other (meta)data

Il est possible de lier le jeu de données à d'autres ressources, à l'aide :

- Des ORCID des auteurs ;
- Des IGSN des échantillons géologiques ;
- Des DOI des articles associés ;
- Des DOI des jeux de données associés.

L'ensemble de ces couples DOI SEANOE / PID ressources sont mis à jour dans la notice DataCite.

Par contre, nous ne permettons pas de détailler la nature du lien avec les articles et les jeux de données. Nous n'utilisons que le rôle IsAssociatedTo. Permettre de préciser la nature du lien pourrait faire l'objet d'une évolution dans une prochaine version.

Reusability (réutilisables)

R1.1. (Meta)data are released with a clear and accessible data usage license

Les jeux de données dans SEANOE sont obligatoirement publiés sous une licence CC.

R1.2. (Meta)data are associated with detailed provenance

Cette information est souvent disponible mais disséminée dans la description et dans le champ "Sensor metadata ». Mais nous ne proposons pas de champ de métadonnées spécifique pour enregistrer cette information.

R1.3. (Meta)data meet domain-relevant community standards

Les métadonnées collectées par SEANOE sont compatibles avec le schéma de DataCite et du Dublin Core. L'ajout de champs (ex. : paramètres) basés sur les listes de vocabulaire de SeadataNet est un projet à moyen terme.

SEANOE est-il lié à l'EOSC ou prévoit-il de l'être ?

SEANOE est un service de gestion de données de l'EOSC[14] ; SeaDataNet utilise SEANOE pour aider les chercheurs dans la publication de leur jeu de données.

Quels sont les challenges actuels et à venir ?

En termes d'outil, l'interface du back-office est vieillissante et nécessitera une refonte à moyen terme. Mais c'est surtout la faiblesse relative du nombre de dépôts qui pourrait être problématique à long terme. Le principal challenge pour SEANOE est d'attirer davantage de dépôts. L'atteinte d'une masse critique de dépôts, qu'il est d'ailleurs difficile d'estimer, assurerait à SEANOE la visibilité suffisante au niveau international pour que cette visibilité entraine elle-même de nouveaux dépôts.

Comment allez-vous développer SEANOE ?

Les formulaires de dépôts dans SEANOE sont vieillissants. Ils sont basés sur un système qui a été développé en 2009. La refonte du back-office de SEANOE est envisagée à partir de 2021. La première impression donnée par un back-office d'un entrepôt de données est importante. En effet, quand un auteur doit publier les données associées à un article, il a à sa disposition plusieurs centaines d'entrepôts généralistes ou thématiques. Nous remarquons parfois le passage de scientifiques qui remplissent rapidement quelques champs avant d'abandonner leur dépôt. Nous supposons qu'ils passent donc quelques dizaines de secondes sur une première sélection d'entrepôts avant d'en retenir un pour publier leurs données. Et quand un auteur a fait l'expérience d'un système de dépôt qui lui convient, il lui reste le plus souvent fidèle pour ses prochains dépôts. Pour SEANOE, le manque d'attrait immédiat du système de dépôt actuel est pénalisant.

Le développement d'un système d'indexation à l'aide de liste de vocabulaires standardisés du projet SeaDataNet (ex. : liste des paramètres mesurés, pays et organismes du point de contact, …) permettrait de finaliser l'automatisation des transferts de données entre SEANOE et EMODnet Ingestion. La description des jeux de données par des vocabulaires standardisés est également une recommandation FAIR.

A quoi ressemblera SEANOE dans cinq ou dix ans ?

A ce stade, SEANOE remplit correctement le rôle qui lui a été attribué : offrir une interface de publication simple et rapide pour publier un jeu de données et obtenir un DOI pour le citer dans un article. Dans dix ans, ce besoin devrait continuer d'exister et sera sans doute encore plus fort du fait de la demande sociétale de la publication en libre accès des données de la recherche.

Dans dix ans, SEANOE devrait donc être sensiblement identique à la version actuelle. Nous espérons qu'il proposera un accès libre à un nombre important de nouveaux jeux de données.

Et dans dix ans, les jeux de données publiés par SEANOE devraient s'inscrire dans un écosystème élargi de ressources internationales interconnectés.

Questions posées par Joachim Schöpfel

[14] https://marketplace.eosc-portal.eu/services/seadatanet-doi-minting-service