Les données et métadonnées dans DYNALIT

Synthèse des discussions 2015-2016

28 Juin 2016

Liste des contributeurs principaux :
Aurélie Batany, Guillaume Brissebrat, Arnaud Caillo, Lucie Coquempot, Robert Davidson, Rémy Decoupes, Christophe Delacourt, Yuji Kato, Olivier Lobry, Myriam Mahabot, Denis Marin, Christophe Martin, Mohand Medjkane, Cécilia Pignon-Mussaud, Mathias Rouan, Loïc Salaün.
Adresse de messagerie : dynalit-md@listes.univ-brest.fr
Site Web : https://www-iuem.univ-brest.fr/pops/projects/donnees-observatoire

GT Métadonnées

SOMMAIRE

Introduction - La donnée dans DYNALIT

Présentation DYNALIT

DYNALIT regroupe le SOERE “Trait de Côté, Aménagement Littoraux” (labellisation AllEnvi février 2011) et le SNO “Dynamiques du Littoral et du Trait de Côte” (labellisation INSU-SIC avril 2014).
Il s’agit d’un réseau de 29 sites-ateliers qui adressent des questions scientifiques concernant:

- L’hydrodynamique du littoral
- Les transports sédimentaires
- La morphodynamique

Types de données produites dans DYNALIT

Les mesures au sein de DYNALIT peuvent porter sur :

- **La bathymétrie :** SMF, SONAR, vidéo, LIDAR, etc.
- **La topographie (MNT):** TLS, ALS, LIDAR, DGPS, drone, etc.
- **La topographie (profils) :** DGPS, photos aériennes, images sat, LIDAR, etc.
- **La ligne de rivage** : photos aériennes, images sat, LIDAR
- **Les courants/houles :** ADCP, ADV, Limnimètre, courantomètre, Datawell, capteurs de pressions, etc.
- **Les niveaux d’eaux :** capteurs de pression, marégraphes, vidéo, etc.
- **La sédimentologie :** échantillons, benne van veen, carotte, etc.
- **La turbidité :** sondes YSI/ADCP, bouée MAREL sonde SMATCH, etc.
- etc.

GT Métadonnées

Cycle de la donnée (cf. INIST)

Le Cycle de vie des données de recherche (Research Data Lifecycle) décrit le processus d'utilisation des données de leur création à la publication à leur réutilisation ultérieur.
Il existe plusieurs représentations de ce cycle, nous proposons d'appliquer le modèle du Centre d'Archives des donnés de la recherche, UK Data Archive, aux données DYNALIT :

Création de données :
Conception de campagnes de mesures / Développements de protocoles d'acquisition /Collecte de données.

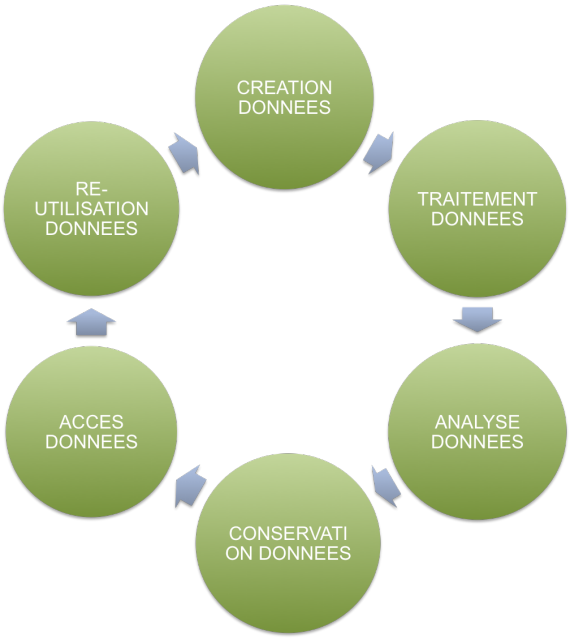
Traitement de données :
Développement et mutualisation de routines / Partage de logiciels dédiés/ Nettoyage, validation, vérification.

Analyse des données :
Interprétation/ Création d'indicateurs dérivés/ Production de résultats de recherche.

Conservation des données :
Format et support de stockage/Procédures de sauvegarde/Renseignement des métadonnées

Accès aux données :
Développement d'une interface dédiée/ Interopérabilité des systèmes existants/ DOI

Réutilisation des données :
Charte d'utilisation de la données/ Licences



Questions clés 2016 :

Analyse des données

Production de résultat de recherche - délais de publication

Au sein du SNO DYNALIT, un constat sur la production de recherche permet de d'identifier :
Plusieurs niveaux de distribution, qui sous-entendent des délais de publication différents :

- Donnée brute ;
- Donnée traitée ;
- Donnée interprétée.

Une prise en compte de contextes particuliers :

- exemple des thèses ;
- données produites dans le cadre de projets;
- rattachement à une structure internationale (→ délais de diffusion courts)

Proposition du groupe :
Mise en place de délais systématiques de publication des données DYNALIT ?
Inscrire une date de publication pour toute donnée bancarisée au moment de la bancarisation : tout de suite ; 6 mois, 1 an; durée thèse, 5 ans, etc.

Conservation des données

Fiche de métadonnées

Eléments de contexte :

- Standards liés à la directive INSPIRE
  - Métadonnées de données : cf. Guide de saisie des éléments de métadonnées INSPIRE
    Ce guide décrit pour chaque champs/élément de métadonnée, les exigences INSPIRE ainsi que les recommandations nationales, le tout agrémenté d'exemples et de contre exemples.
  - Métadonnées de services : cf. Guide de saisie des métadonnées de service INSPIRE
    INSPIRE définit une architecture de services, composée de services de plusieurs types. Les services concernés par la Directive INSPIRE sont les « services de données géographiques » (spatial data services), c'est-à-dire les opérations qui peuvent être exécutées à l'aide d'une application informatique sur les données géographiques contenues dans des séries de données géographiques ou sur les métadonnées qui s'y rattachent (art. 3 de la directive). Parmi ces services de données géographiques, on distingue un sous-ensemble particulier de cinq types de services, connus sous la dénomination « services en réseau » (network services) pour lesquels INSPIRE définit des règlements et des guides techniques spécifiques.
  - Identifiant de ressource unique : cf. Guide sur les identifiants de ressource uniques
  - Outils de saisie et de validation :
    - http://www.geocatalogue.fr/#!ServicesValidationMD
    - http://inspire-geoportal.ec.europa.eu/editor/
  - Afin de faciliter la saisie et la validation de fiches de métadonnées il est recommandé d'établir des modèles (template)

Accès aux données - attribution de DOI

Objectif des DOI

L'objectif d'un Digital Object Identifier (DOI) est de permettre l'identification et donc la citation de ressources numériques. A ce titre, les DOI déjà utilisés pour référencer des publications, sont de plus en plus utilisés pour identifier de manière non-ambiguë des jeux de données. Les DOI vont typiquement être utilisés dans les publications scientifiques pour désigner les jeux de données sur lesquels s'appuie l'étude réalisée. Le lecteur de l'article peut alors utiliser ce DOI pour retrouver sur le Web les métadonnées du jeu de données qui lui permettent ensuite de retrouver les données elles-mêmes. Il pourra ainsi rejouer les calculs qui ont été effectués à partir de ces jeux de données afin de vérifier les conclusions de l'étude.

Pour le producteur d'un jeu de données, l'intérêt est de permettre que son jeu de données soit cité correctement et de manière précise et favoriser ainsi la valorisation et la dissémination de son travail. Cela lui permet également de retrouver plus facilement les citations vers son jeu de données puisque l'identifiant apparait directement dans les publications.

Contraintes :

L'attribution de DOI à des jeux de données doit cependant répondre à deux contraintes :

- Il doit être facilement utilisable dans le sens où il doit être facile de citer les données que l'on utilise.
- Il doit permettre de retrouver le jeu de données dans l'état dans lequel il était au moment de la citation.

Ces deux contraintes posent des problèmes de mise en œuvre que nous abordons par la suite.

Fonctionnement et Organisation

Les DOI permettent une identification au niveau international et son gérés au plus haut niveau par DataCite, une organisation à but non lucratif basée à Londres qui vise à faciliter l'accès aux données de la recherche. L'organisation est représenté au niveau local (national) par des organisations bien identifiées. Au niveau français cette représentation est assuré par l'Institut de l'information scientifique et technique (Inist) du CNRS.

Le DOI est composé d'un préfixe et d'un suffixe (préfixe/suffixe). Le préfixe est relatif à une organisation chargée de l'attribution de DOI. Ce préfixe est obtenu auprès des représentant nationaux. En France, pour mettre en place des DOI, il convient donc de contacter l'Inist qui délivrera un préfixe en échange d'une redevance annuelle (180 € en 2016). Le suffixe est quant à lui composé d'une suite libre de caractères alphanumériques. Chaque organisation peut donc décider de construire et structurer les noms de ses DOI comme elle l'entend. Nous reviendrons par la suite sur cet aspect.

A chaque DOI est associé un ensemble de métadonnées qui décrivent le jeu de données. Ces informations que l'on retrouve dans la norme ISO 19115 permettent notamment de connaitre la nature et la qualité des données, leur propriétaires, financeurs, les conditions et restrictions d'utilisation, ...

Le fonctionnement d'un DOI est assez simple pour un utilisateur :

- A un DOI correspond une URL qui permet de retrouver un document HTML contenant les métadonnées correspondantes ;
- Dans ces métadonnées figure une URL qui conduit vers une page ou un service web qui donne accès au jeu de données lui-même.

Mise en œuvre et obligations

La délivrance de DOI consiste en un ensemble d'actions et d'obligations :

- Il convient tout d'abord d'obtenir un préfixe auprès de l'Inis. Notons qu'il est possible d'obtenir gratuitement un préfixe de test qui servira tout au long de la phase de mise en œuvre des mécanismes d'attribution ;
- Il est nécessaire de réfléchir préalablement à une stratégie d'attribution et de nommage des DOI (voir plus loin) ;
- Pour chaque DOI, il convient de remplir, manuellement ou automatiquement, une fiche de métadonnées
- L'organisme d'attribution doit être en mesure de renvoyer, pour chaque DOI, une page web appelée "Landing Page" qui permettra de récupérer le jeu de données correspondant ;
- Il devra mettre en outre un service de téléchargement des jeux de données correspondants. L'attribution d'un DOI suppose que les données soient librement téléchargeable sur Internet ;
- Enfin, il doit s'assurer de la non-modification d'un jeu de données référencé (voir plus loin).

Considérations et choix de mise en œuvre

Une organisation qui met en place des DOI va être amenée à se poser un certain nombre de questions et faire des choix de mise en œuvre.

Critères de découpage et d'attribution

Lorsqu'une organisation gère de grosses bases de données provenant de diverses sources, mesurant différents paramètres, en différents lieux, … elle doit se demander comment attribuer des DOI à l'ensemble de ces données, car les critères de découpage peuvent être multiples et il n'y en a pas forcément qui sortent du lot. Cela est d'autant plus vrai pour les observatoires qui visent généralement la pérennisation des sites d'observation. Les jeux de données ne correspondent plus à une expérimentation, une étude ponctuelle de quelques paramètres ou d'un chercheur particulier mais à l'observation systématique et sur le long terme d'un ensemble de variables observées par un ensemble d'acteurs.
Il convient alors de trouver la bonne granularité pour l'attribution des DOI, comment regrouper les données dans des jeux de données et pour cela trouver des critères de regroupement. Un DOI doit arriver à cibler des données, les discriminer des autres suivant une logique pertinente. Il est par exemple pertinent qu'un article puisse citer les données qui ont précisément servi pour une étude, une analyse, plutôt que de citer la base entière. Néanmoins, multiplier les DOI peut conduire à une multitude de citations dans le cas où une étude ou un rapport concernerait une grosse partie de la base.
Pour pallier cela, une approche peut consister à hiérarchiser l'information et à attribuer des DOI à plusieurs niveaux : au niveau de la base, d'un ensemble de paramètres, d'un ensemble de stations, …
D'autres critères peuvent également entrer en jeu dans le découpage : la qualité des données, les conditions d'utilisation, les droits d'accès, l'évolution des données, etc. Le découpage doit se faire en respectant une certaine homogénéité de ces critères-là.
En définitive, il convient de trouver un compromis entre la facilité d'utilisation (granularité, cohérence des données) des DOI qui va permettre de citer correctement les données et la facilité de réalisation (capacité à tenir les obligations).

Mise à jour d'un jeu de données

La mise à jour d'un jeu de données pose une difficulté particulière dans la mesure où un DOI doit en théorie référencer un jeu de données qui n'évolue pas. En effet, lorsqu'un jeu de données est cité pour appuyer une étude scientifique, il doit être possible de vérifier les conclusions de l'étude, par exemple en refaisant tourner un modèle scientifique, en se basant sur les mêmes entrées que lorsque l'étude a été menée. Si le jeu de données a évolué entre-temps, il est impossible de refaire les mêmes analyses.
Le cas des jeux de données continus où l'on se contente de rajouter des données ne pose pas de problème, car il est possible de retrouver les données qui ont servies initialement en se basant sur la date de publication. L'attribution d'un DOI n'impose d'ailleurs pas qu'un jeu de données ait une date de fin définitive.

En revanche, modifier les données, pour des raisons de corrections, de validation ou de suppression de valeurs aberrantes, est problématique. Pour palier ce problème, deux solutions sont envisageables. La première consiste à versionner un jeu de données et d'attribuer un DOI différent à chaque version. Chaque DOI devra permettre de retrouver la version du jeu correspondant, indépendamment des évolutions postérieures au jeu. Cette solution est relativement simple à mettre en œuvre dans le cas de jeux de données peu volumineux, mais peut en revanche poser un problème de stockage dans le cas contraire.

Une autre solution consiste à n'attribuer qu'un seul DOI mais à conserver l'historique de modifications. Cet historique devra pouvoir être récupéré en même temps que le jeu de données dans sa valeur actuelle, de sorte qu'il sera possible de retrouver l'état du jeu à la date de publication. Cette fois, seules les modifications apportées ont besoin d'être conservées en plus du jeu de données, limitant ainsi la place totale occupée. Mais il nécessite de mettre en place des mécanismes permettant de détecter et historiser chaque modification.

Choix du suffixe

Le suffixe d'un DOI étant libre. Il doit être unique et doit permettre de retrouver le jeu de données correspondant, mais sa structuration est complètement libre. Il peut notamment possible d'y mettre des informations sémantiques telles que la provenance, le propriétaire ou le gestionnaire du jeu. Nous n'encourageons cependant pas cette stratégie, car alors le choix du suffixe peut alors être relativement complexe lorsque la production d'un jeu de données entre dans le cadre de plusieurs contextes : projets financeurs, structures de rattachement (observatoires, laboratoires, pôles de données, instituts, …) Trouver les termes et l'ordre de ces termes, qui conviendrait à tous les acteurs sans que cela n'entraîne de susceptibilités, peut vite devenir un casse-tête. Nous encourageons plutôt de ne mettre aucune information sémantique dans le suffixe (par exemple une suite de chiffre) sachant que toutes les informations concernant la propriété intellectuelle peuvent être renseignées dans les métadonnées et la citation du jeu.

Alias

Les DOI doivent être uniques et ne doivent référencer qu'un seul jeu de données. En revanche, rien n'interdit qu'un même jeu de données soit référencer par plusieurs DOI.

Donner plusieurs DOI à un même jeu peut notamment s'avérer nécessaire lorsque ce jeu rentre dans le cadre de plusieurs structures qui ont chacune mise en place un système d'attribution de DOI (par exemple au niveau d'un observatoire, d'un Service National d'Observation, d'un pôle de données). Ce qui est en revanche important est que ces différents DOI permettent de retrouver le même jeu et les mêmes métadonnées sur ce jeu, modulo les métadonnées propres à chaque structure.

Notons cependant que si cela est possible, attribuer plusieurs DOI à un même jeu n'a pas d'intérêt particulier sauf dans le cas où des mécanismes automatiques d'attribution ont été mis en place. Dans le cas contraire, simplement remonter les DOI vers les structures englobantes peut s'avérer un meilleur choix.

DOI et catalogue de données

La mise en œuvre de DOI peut aller de pair avec la mise en place d'un catalogue de données. Dans ce cas, il pourra être utile de penser les deux systèmes de manière conjointe et cohérente afin d'éviter un travail superflu. Typiquement il convient dans les deux cas de choisir un découpage, de renseigner des métadonnées et d'offrir un moyen de retrouver les données. Même si le catalogage n'impose pas de pouvoir retrouver les données directement sur internet (une adresse mail de contact suffit), il serait dommage de ne pas profiter, dans le cadre du catalogue, de la nécessaire mise en place d'un tel mécanisme dans le cadre des DOI. Les métadonnées devant être renseignées dans le cadre des DOI étant très proches de celles de la norme ISO 19115 et de la directive INSPIRE il peut être intéressant de les gérer de manière commune (par exemple en générant celles des DOI à partir des métadonnées ISO 19115).

Ouverture des données

L'attribution d'un DOI présuppose l'ouverture des données qu'il référence : les données doivent pouvoir être téléchargeables en ligne gratuitement. Cela n'empêche cependant pas de tracer les données via par exemple un mécanisme d'authentification permettant d'identifier la personne qui télécharge les données. En outre, des conditions et restrictions d'utilisation peuvent être spécifiées dans les métadonnées et tout utilisateur s'engage à citer les données à l'aide de la citation correspondante.

Réutilisation de la donnée - Mise en place d'une data policy / charte d'utilisation des données ?

Une bonne charte de données est nécessaire à un échange efficace et satisfaisant au sein du projet et au-delà. Elle permet de définir les droits et devoirs mutuels des fournisseurs et des utilisateurs de données, et doit répondre à leurs besoins.

Besoins des utilisateurs :

- accéder à des données de "bonne" qualité et disponibles "rapidement"
- bénéficier de l'expertise des producteurs de données

Besoins des producteurs :

- participer aux (ou au moins être au courant des) activités des utilisateurs
- cosigner (ou au moins être remercié dans) les publications

Les institutions ou les bailleurs de fonds ont besoin d'être remerciés dans les publications / contributions.

La charte doit définir : les règles globales d'accès et d'utilisation des données, et, si les données ne sont pas en accès libre, la procédure d'enregistrement des utilisateurs, les différents groupes d'utilisateurs et les règles spécifiques à chacun. Elle précise aussi le périmètre des données concernées et l'évolution dans le temps des règles (durée de vie d'un compte utilisateur, périodes d'embargo, ouverture progressive des accès aux données). Chaque jeu de données peut définir des règles d'utilisation plus spécifiques (phrase de remerciement particulière, par exemple).

Dans la charte, sont aussi précisés les devoirs des fournisseurs de données : fournir la meilleure version des données, informer les utilisateurs des corrections, respecter un délai de fourniture des données.

Plusieurs niveaux d'accès sont possibles :

- accès libre
- enregistrement des utilisateur
- enregistrement modéré (plusieurs niveaux d'accès, possibilité de refuser l'accès à un utilisateur)

Dans tout les cas, l'établissement de règles générales de "bon comportement" est nécessaire. Par exemple :

- pas de redistribution des données
- pas d'utilisation commerciale
- s'engager à contacter le pi (surtout en cas de publication)
- utiliser la phrase de remerciement du projet (ou du jeu) dans toute publication utilisant les données

Si laisser libre l'accès aux données est plus simple, imposer aux utilisateurs de s'enregistrer offre certains avantages. Savoir qui accède à quel jeu de données permet de :

- prévenir le pi à chaque téléchargement
- favoriser la collaboration entre utilisateur et producteur de données

- informer les utilisateurs en cas de mise à jour ou de correction d'un jeu de données
- rassurer les fournisseurs de données et obtenir plus facilement les données
- fournir aux financeurs des statistiques sur l'utilisation des données

Licence pour la diffusion des données ?

- Définition et intérêt
- Tour d'horizon des types de licence possibles pour les données scientifiques (http://coop-ist.cirad.fr/gestion-de-l-information/gestion-des-donnees-de-la-recherche/rendre-publics-ses-jeux-de-donnees/6-les-principales-licences-de-diffusion-des-jeux-de-donnees)

Liens avec HAL

- Contexte : HAL, Hyper Articles en Ligne, est une archive ouverte. Le nombre d'articles uploadés sur cette plateforme constitue un critère d'évaluation HCERES. Il est possible de regrouper les publications de HAL dans des collections comme celle de Dynalit (disponible sur https://hal.archives-ouvertes.fr/DYNALIT). On parle alors de "tamponnage" d'articles. Le tamponnage peut être manuel ou automatique.
-
- Problématique : Comment enrichir la collection DYNALIT automatiquement via des critères de tamponnage ?
- L'attribution de DOI sur les jeux de données Dynalit peut répondre à cette problématique. Par exemple, on peut envisager de créer un critère de tamponnage qui remonte les préfix DOI de Dynalit.
- Cependant, est-ce que la solution DOI va être retenue ? Et si oui, quand serons-nous capable de l'implémenter ?
- Est-ce qu'il ne faudrait pas réfléchir à une solution intermédiaire ? C'est à dire, demander aux chercheurs de "tagguer" leurs publications reposant sur Dynalit? Si oui quels tag?
- Voici le résultat d'une recherche du mot "Dynalit" en texte libre sur l'ensemble des publications HAL : http://api.archives-ouvertes.fr/search/?q=DYNALIT&wt=xml. Ce type de requête (SOLR) peut être un critère de tamponnage automatique. Le problème est que pour l'instant cette requête ne remonte qu'une publication (c'est à dire qu'il n'y qu'une publication ou le mot Dynalit apparaît). Pour info, voici un site d'aide à la construction de ce type de requête : http://api.archives-ouvertes.fr/docs/ref/resource/structure.
-
- Autre question : est-ce qu'il y a du sens à lier nos fiches de métadonnées "geonetwork" aux publications présentes dans Hal? Si oui, les implémentations de serveurs CSW sont elles compatibles avec le moissonnage OAI (Open Archive Initiative) qu'offre HAL? Ceci permettrait d'enrichir nos fiches de métadonnées avec les publications, on montrerait ainsi comment nos jeux de données sont utilisés.
- Exemple de requête de moissonnage OAI (critère : collection Dynalit) :
- http://api.archives-ouvertes.fr/oai/hal/?verb=ListRecords&metadataPrefix=oai_dc&set=collection:DYNALIT

Glossaire

Métadonnées	Informations descriptives des données ou des services sur les données, et rendant possibles leur recherche, leur inventaire et leur utilisation (INSPIRE)
Données	Dans ce document, le terme « données » est un terme générique désignant à la fois les séries et les ensembles de séries
Thésaurus	ou dictionnaire de mots-clés Ensemble de mots clés organisés de façon synonymique et hiérarchique permettant de classer ou relier des ressources
DOI	Un DOI (Digital Object Identifier) est un identifiant pérenne qui permet l'identification unique d'un objet physique ou numérique et sa citation. Il fournit un lien stable à des ressources en ligne, comme les données de la recherche (INIST)
IDS	Une IDS (*Infrastructure de Données Spatiales*; ou IDG, Infrastructure de Données Géographiques, en anglais SDI, pour Spatial Data Infrastructure) est une organisation qui repose sur des accords de partage, une coordination entre ses membres et des systèmes informatiques qui intègrent un ensemble de services (catalogues, serveurs, logiciels, données, applications, pages web, ...) utilisés pour la gestion de l'information géographique (cartes, orthophotoplans, images satellitaires…). Une partie des services informatiques des IDS est déployée sur le web et respecte un ensemble de conditions d'interopérabilité (normes, spécifications, protocoles, interfaces, ...). Cela permet à l'utilisateur de pouvoir utiliser les services à travers un simple navigateur web ainsi que de combiner les services proposés par différentes IDS selon ses besoins. (wikipédia)