Cookbook pour dictionnaire de données pour l'interopération des données et logiciels de recherche à l'échelle globale

Auteurs: R David, L Bouveret, L Coché, P Corrêa, R Edmunds, A Heredia, JL Jung, Y Kondo, I Le Berre, Y Le Bras, E Lerigoleur, L Mabile, J Machicao, B Madon, Y Murayama, M O'Brien, T Osawa, H Raoul, A Richard, S Santos, A Specht, D Stepanyan, D Vellenich, L Wyborn Contact: romain.david@erinha.eu Website: www.erinha.eu twitter: @ERINHA_RI

Notre société est aujourd'hui confrontée à de profonds changements (biodiversité, climat, pandémie, etc.). Les impacts humains et leur atténuation dépendent de notre capacité à mobiliser la recherche au niveau mondial. Le développement durable de la société dépendra en grande partie du **développement sur le long terme d'une science globale et d'outils de recherche scientifique, de leurs résultats et des écosystèmes de recherche partagés**. Cette globalisation de la recherche nécessite d'inter-opérer nos systèmes d'observation et d'expérimentation pour mieux comprendre ces changements, pour mieux simuler leurs effets.

La pandémie de Covid-19 fait désormais rage dans le monde. La *reproductibilité de la recherche et des résultats à travers les régions dans des contextes différents devrait accélérer les réponses sociétales*. Le partage des données et le développement de la recherche de synthèse avec agrégation de données à large échelle sont essentiels pour permettre de tels processus.

L'utilisation partagée de connaissances formalisées, de vocabulaires, de normes et de procédures communes à grande échelle est nécessaire.

Objectifs:

Ce poster présente une méthodologie commune, un "livre de cuisine" de dictionnaire de données, qui propose une feuille de route pour construire des dictionnaires de données à grande échelle.

L'objectif est de présenter les défis relevés lors de la construction de dictionnaires de données dans trois projets à l'échelle globale liés à la recherche sur la biodiversité et/ou les maladies infectieuses:

PARSEC, Kakila, ERINHA-Advance.

cookbook pour Dictionnaire de Données

Générique

Définir le périmètre de votre communauté
[Questions scientifiques]

Expliquer et convaincre
[Principes du dictionnaire de données]

Identifier vos objets scientifiques
[Listes d'Entités]

Identifier les aspects des objets scientifiques que vous devez évaluer
[Liste de QUALITES]
pour chaque entité

Nommer et définir les variables nécessaires et leurs dimensions : [Noms des variables : Entité_Qualité_Dimensions] pour chaque Qualité de chaque Entité

⚠ Réutiliser les standards, vocabulaires, concepts et définitions existants

Faire valider par l'ensemble de la communauté
[Liste de DEFINITIONS]

Parsec

Périmètre de la communauté défini par:
[Questions scientifiques]
Certaines images satellites peuvent-elles être utilisées comme proxy des indicateurs socio-économiques ?

[Principes du dictionnaire de données]
- Choisir des définitions communes pour permettre des comparaisons entre les pays
- Valider les définitions réutilisables pour assurer la reproductibilité des résultats du projet

A propos de la question scientifique, des exemples d' **[ENTITES]**
- Municipalité
- Images satellites
- Secteurs
- ...

Pour chaque entité, des exemples de
[QUALITES]
- Indicateurs de développement humain (HDI) de [municipalité ou images satellites ou secteurs]
- Population de [municipalité ou images satellites ou secteurs]
- ...

Pour chaque qualité, un exemple de
[Nom de variable: Entité_Qualité_Dimensions]
- Municipality_HDI_value
- Municipality_population_value
- ...

Réutiliser des **vocabulaires et des ontologies existants**, avec un défi : lorsqu'il y en a plusieurs, choisir les meilleurs termes.
*Tous les termes sont **approuvés par la communauté***

Kakila

Périmètre de la communauté défini par:
[Questions scientifiques]
Peut-on caractériser la présence des cétacés dans les eaux de l'archipel guadeloupéen à l'aide de plusieurs bases de données de science citoyenne ?

[Principes du dictionnaire de données]
- Choisir un vocabulaire commun pour décrire les champs de bases de données hétérogènes
- Adopter une description et des valeurs de variables communes entre les bases de données

A propos de la question scientifique, des exemples d **[ENTITES]**
- Observation de cétacés
- Observateurs
- ...

Pour chaque entité, des exemples de
[QUALITES]
- Individu (observation de cétacés)
- Age individu (observation de cétacés)
- Expérience (d'observateur)

…

Pour chaque qualité, un exemple de
[Nom de variable: Entité_Qualité_Dimensions(&unit)]
- Cetacean_Obs_Age
- Observer_Exp_[category]
- ...

Les deux défis sont :
- D'obtenir un consensus sur les valeurs possibles des variables entre bases de données
- **D'aligner les données** existantes sur le vocabulaire Darwin Core

*Tous les termes sont **approuvés par la communauté***

Erinha Advance

Périmètre de la communauté défini par:
[Questions Scientifiques]
Les laboratoires de niveau de biosécurité 4 peuvent-ils comparer leurs résultats d'études virologiques *in vivo* ?

[Principes du dictionnaire de données]
- Répertorier tous les modèles in vivo, les virus, les protocoles et la sensibilité des données
- Lister et homogénéiser la description des variables pour d'éventuelles méta-analyses

A propos de la question scientifique, des exemples d **[ENTITES]**
- Virus
- Hamster
- ...

Pour chaque entité, des exemples de
[QUALITES]
- Taux de mortalité (après infections par le virus)
- Âge (du hamster)
- Race (du hamster)
- ...

Pour chaque qualité, un exemple de
[Nom de variable: Entité_Qualité_Dimensions]
- Virus_Mortalite_Percentage
- Hamster_Age_Nbweeks
- Hamster_Breed_[BreedName]
- ...

Les deux défis sont :
- De réutiliser les vocabulaires et les ontologies existants (s'il y en a plusieurs, de choisir le meilleur).
- D'obtenir un consensus pour les noms et les définitions des variables.

*Tous les termes sont **approuvés par la communauté***

Le dictionnaire de données avec **[Liste des DEFINITIONS]** approuvé par l'ensemble de la communauté contient TOUTES les **[Questions scientifiques], [Liste des ENTITES], [Liste des QUALITÉS], [Noms des variables : Entity_Quality_Dimentions_Units]** et les définitions des variables (voir https://www.qudt.org/)

3 projets, le même "livre de cuisine"

Le **projet PARSEC** construit de nouveaux outils pour le partage et la réutilisation des données grâce à une étude transnationale sur l'impact **S**ocio-**EC**onomique des **A**i**R**es **P**rotégées.

La **base de données Kakila** centralise et harmonise les données d'observations des mammifères marins du sanctuaire Agoa autour de l'archipel français de la Guadeloupe, Antilles françaises.

Le **projet ERINHA-Advance** vise à soutenir les opérations de l'infrastructure de recherche ERINHA qui est conçue pour générer des données à partir d'accès transnational aux moyens de recherche sur les agents hautement pathogènes.

Dans ces 3 projets à large échelle, des défis similaires se sont posés : **agréger et inter-opérer des données hétérogènes préexistantes** à l'échelle globale, et **partager des outils** pour surveiller, maintenir la qualité, analyser l'échelle et faire face à l'incertitude.

Faire face à la complexité

- Développer la littératie des dictionnaires de données est un travail essentiel pour impliquer les scientifiques dans les définitions de variables,
- Permettre l'adaptabilité, la portabilité, la réplicabilité et la reproductibilité implique que les logiciels et les workflows doivent être définis comme toutes les données du dictionnaire de données,
- Aborder les problèmes de dimension dans chaque contexte est nécessaire pour toutes les variables.

Aussi simple que possible!!

L'expérience commune de nos trois projets a montré que **nous devons procéder étape par étape le plus simplement possible** et veiller à ce que chaque étape soit *compréhensible pour l'ensemble de la communauté*. Il est nécessaire pour cela d'*améliorer l'accès et la réutilisation de tous les éléments sémantiques existants* et ne pas essayer de construire une cathédrale avec une petite cuillère.

*** References:**
- ★ David, R., Mabile, L., Specht, A., Stryeck, S., Thomsen, M., Yahia, M., Jonquet, C., Dollé, L., Jacob, D., Bailo, D., Bravo, E., Gachet, S., Gunderman, H., Hollebecq, J.-E., Ioannidis, V., Le Bras, Y., Lerigoleur, E., Cambon-Thomsen, A. and Research Data Alliance – SHAring Reward and Credit (SHARC) Interest Group, T.R.D., 2020. FAIRness Literacy: The Achilles' Heel of Applying FAIR Principles. Data Science Journal, 19(1), p.32. DOI: http://doi.org/10.5334/dsj-2020-032
- ★ Coché L, Arnaud E, Bouveret L, David R, Foulquier E, Gandilhon N, Jeannesson E, Le Bras Y, Lerigoleur E, Lopez PJ, Madon B, Sananikone J, Sèbe M, Le Berre I, Jung J-L (2021) Kakila database: Towards a FAIR community approved database of cetacean presence in the waters of the Guadeloupe Archipelago, based on citizen science. Biodiversity Data Journal 9: e69022. https://doi.org/10.3897/BDJ.9.e69022

Remerciements:
PARSEC est financé par le Belmont Forum avec le support de la National Science Foundation (NSF), São Paulo Research Foundation (FAPESP), French National Research Agency (ANR), et Japan Science and Technology Agency (JST). ERINHA Advance est financé par le programme Européen ERINHA-Advance Nº824061. Kakila database est financé par le LabEx DRIIHM "Investissements d'Avenir" (ANR-11-LABX-0010) et soutenu par le SO-DRIIHM project (ANR-19-DATA-0022). Ce travail est partiellement financé par le programme Européen EOSC-Life (No. 824087).

Affiliations : Romain David (ERINHA, European Research Infrastructure on Highly Pathogenic Agents) AISBL, FR), https://orcid.org/0000-0003-4073-7456; Laurent Bouveret (OMMAG), https://orcid.org/0000-0003-3621-1005; Lorraine Coché, (LETG, Université de Bretagne Occidentale) https://orcid.org/0000-0003-4909-4848 ; Pedro Pizzigatti Corrêa (University of São Paulo, BR), https://orcid.org/0000-0002-8743-4244, Rorie Edmunds (World Data System); Ana Heredia (ORCID), https://orcid.org/0000-0001-7862-8955; Jean-Luc Jung (Univ Brest, ISYEB-CNRS, Sorbonne Université), https://orcid.org/0000-0002-8795-8056 ; Iwan Le Berre, (LETG, Université de Bretagne Occidentale), https://orcid.org/0000-0002-8933-743X; Yvan Le Bras, (PNDB, UMS 2006 PatriNat), https://orcid.org/0000-0002-8504-068X ; Emilie Lerigoleur, (CNRS, UMR GEODE, Université Toulouse 2), https://orcid.org/0000-0002-0864-659X ; Laurence Mabile (University of Toulouse - INSERM, FR) ,https://orcid.org/0000-0002-7724-1721 ; Jeaneth Machicao (University of São Paulo, BR), https://orcid.org/0000-0002-1202-0194 ; Bénédicte Madon (LETG, Université de Bretagne Occidentale), https://orcid.org/0000-0001-8608-3895 ; Yasuhiro Murayama (National Institute of Information and Communications Technology, JA), https://orcid.org/0000-0003-1129-334X; Margaret O'Brien (University of California Santa Barbara, USA), https://orcid.org/0000-0002-1693-8322; Takeshi Osawa (Tokyo Metropolitan University), https://orcid.org/0000-0002-2098-0902 ; Hervé Raoul (ERINHA); Audrey Richard (ERINHA), https://orcid.org/0000-0002-0207-0139; Solange Santos (Scientific Electronic Library Online), https://orcid.org/0000-0002-2713-6202, Alison Specht (The University of Queensland, AU), https://orcid.org/0000-0002-2623-0854; Shelley Stall, American Geophysical Union, USA), https://orcid.org/0000-0003-2926-8353; Diana Stepanyan (ERINHA; Danton Ferreira Vellenich (University of São Paulo, BR), https://orcid.org/0000-0002-3223-6996 ; Yasuhisa Kondo (Research Institute for Humanity and Nature), https://orcid.org/0000-0001-7670-4475 ; Lesley Wyborn (Australian National University, AU), https://orcid.org/0000-0001-5976-4943.

17ème plénière RDA - *à distance*, Edinburg, 20-22 Avril 2021