Data dictionary cookbook for research data and software interoperability at global scale

Romain David, Laurent Bouveret, Lorraine Coché, Pedro Pizzigatti Corrêa, Rorie Edmunds, Ana Heredia, Jean-Luc Jung, Yasuhisa Kondo, Iwan Le Berre, Yvan Le Bras, et al.

Data dictionary cookbook for research data and software interoperability at global scale

Authors: 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

We are now facing profound changes (biodiversity, climate, pandemic, etc.). Human impacts and their mitigation will depend on our ability to mobilize research at the global level. The sustainable development of the society will largely depend on the **sustainable development of global science and scientific research tools, outputs, and research ecosystems**. This globalization of research requires inter-operating our observation and experimentation systems in order to better understand these changes, to better simulate their effects.

The Covid-19 pandemic is now raging around the world. The *reproducibility of research and results across regions in different contexts should accelerate human responses*. Data sharing and the development of Synthesis Research with data aggregation at large scale is critical to enable such processes.

The use of common knowledge, vocabularies, standards and procedures at a large scale is necessary.

Objectives:

This poster proposes a draft common methodology, a data dictionary cookbook, which will provide a roadmap towards the building of large scale data dictionaries.

The objective is to report on the challenges met while building data dictionaries in three global projects related to biodiversity and/or disease research:

PARSEC, Kakila, ERINHA-Advance.

Data dictionary cookbook

Generic

Define your community perimeter
[Scientific Questions]

Explain and convince
[Data dictionary principles]

Identify your scientific objects
[List of ENTITIES]

Identify the aspects of your scientific objects you need to assess
[List of QUALITY]
for each Entity

Name and define necessary variables and their dimensions:
[Variables names: Entity_Quality_Dimentions]
for each Quality of each Entity

⚠ Reuse existing standards, vocabularies, concepts and definitions

Validate by the whole community
[List of DEFINITIONS]

Parsec

Community perimeter defined by:
[Scientific Questions]
Can some satellite images be used as proxys for socio-economic indicators?

[Data dictionary principles]
- Choose common definitions to enable comparisons between countries
- Validate reusable definitions to ensure reproducibility of project results

About the scientific question, examples of
[ENTITIES]
- Municipality
- Satellite images
- Sectors
- ...

For each entities, examples of
[QUALITIES]
- Human Development Indicators of [Municipality or Satellite images or Sectors]
- Population of [Municipality or Satellite images or Sectors]
- ...

Due for each qualities, examples of
[Variables names: Entity_Quality_Dimentions]
- Municipality_HDV_value
- Municipality_population_value
- ...

Reuse **existing vocabularies and ontologies**, with a challenge: when there are several, choosing the better terms.
*All terms are **community approved***

Kakila

Community perimeter defined by:
[Scientific Questions]
Can we characterize cetacean presence in the waters of the Guadeloupe archipelago using several citizen science databases?

[Data dictionary principles]
- Choose common vocabulary to describe fields from heterogeneous databases
- Adopt common variables description and values between databases

About the scientific question, examples of
[ENTITIES]
- Cetacean observation
- Observers
- ...

For each entities, examples of
[QUALITIES]
- Individual Species (Cetacean observation)
- Individual Length (Cetacean observation)
- Experience (of Observer)

…

For each qualities, examples of
[Variables names: Entity_Quality_Dimentions(&unit)]
- Cetacean_Obs_Length
- Observer_Exp_[category]
- ...

The two challenges are:
- to obtain a **consensus for variables values** between **databases**
- to **align with the existing data** to Darwin Core vocabulary

*All terms are **community approved***

Erinha Advance

Community perimeter defined by:
[Scientific Questions]
Can Biosafety level 4 laboratories compare their virologic *in vivo* studies results?

[Data dictionary principles]
- List all *in vivo* models, viruses, protocols and their data sensitivity
- List and homogenize variables description for possible meta-analyses

About the scientific question, examples of
[ENTITIES]
- Virus
- Hamster
- ...

For each entities, examples of
[QUALITIES]
- Mortality rate (of the virus)
- Age (of the hamster)
- Breed (of the hamster)
- ...

For each qualities, examples of
[Variables names: Entity_Quality_Dimentions]
- Virus_Mortality_Percentage
- Hamster_Age_Nbweeks
- Hamster_Breed_[BreedName]
- ...

The two challenge are:
- to Reuse **existing vocabularies and ontologies** (if several, choosing the better one).
- to obtain a consensus between variables names and definitions.

*All terms must be **community approved***

Data dictionary with **[List of DEFINITIONS]** approved by the whole community contain ALL **[Scientific Questions], [List of ENTITIES], [List of QUALITIES], [Variables names: Entity_Quality_Dimentions_Units]** and **variables definitions**

3 projects, the same cookbook

The **PARSEC Project** is building new tools for data sharing and reuse through a transnational investigation of the socioeconomic impact of protected areas.

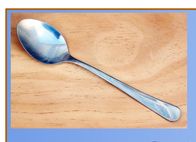
The **Kakila database** centralizes and harmonizes marine mammal observation data for the AGOA sanctuary around the French archipelago of Guadeloupe, French Antilles.

The **ERINHA-Advance project** aims to support the operations of the ERINHA research infrastructure which is designed to generate data from transnational access research activities on highly pathogenic agents.

In these 3 global case-studies, similar challenges have arisen: **to aggregate and interoperate pre-existing heterogeneous data** at the global scale, and **to share common tools** to monitor, maintain quality, scan scale and cope with uncertainty.

Dealing with complexity

- Developing data dictionary literacy is an essential work to imply scientists in variables definitions,
- Enabling adaptability, portability, replicability and reproducibility implicates that software and workflows must be defined as all data in the data dictionary,
- Addressing dimension issues in each context is necessary for all variables.



As simple as possible!!

The common experience of our three projects showed that we **need to proceed step by step as simply as possible** and to ensure that each step is **understandable for the whole community**. It is necessary to *improve access and re-use of all existing semantic materials* and not trying to build a cathedral with a little spoon.

*** 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 P., 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: Data paper. *submitted* Dataset: https://doi.org/10.48502/cg6n-1103

Acknowledgements:
PARSEC is funded by the Belmont Forum through the National Science Foundation (NSF), The São Paulo Research Foundation (FAPESP), the French National Research Agency (ANR), and the Japan Science and Technology Agency (JST). ERINHA Advance is funded by ERINHA-Advance european program under grant agreement Nº824061.. Kakila database is funded by the LabEx DRIIHM French program "Investissements d'Avenir" (ANR-11-LABX-0010) and supported by the SO-DRIIHM project (ANR-19-DATA-0022). This work is partially funded by the EOSC-Life European program (grant agreement No. 824087).

Author 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.