From Raw Data to Data Standards through Quality Assessment and Semantic Annotation

Julien Sananikone[‡], Elie Arnaud[‡], Olivier Norvez[§], Sophie Pamerlon[|], Anne-Sophie Archambeau[¶], Yvan Le Bras[‡]

‡ MNHN, Concarneau, France, Metropolitan
§ FRB, Paris, France
| OFB - Office Français de la Biodiversité (French Biodiversity Agency), Paris, France
¶ IRD, Paris, France, Metropolitan

Corresponding author: Yvan Le Bras (yvan.le-bras@mnhn.fr)

Abstract

Data quality and documentation are at the core of the FAIR (Findable, Accessible, Interoperable, Reusable) principles (Wilkinson et al. 2016). Regarding biodiversity and more broadly ecology domains, complementary solutions of the well-known data standard (notably through Darwin Core (Wieczorek et al. 2012)) orientation are emerging from the intensive use of EML (Ecological Metadata Language (Michener et al. 1997)) metadata standard. These notably capitalize on using:

1. semantic annotation from EML metadata documents that describe data attributes, and
2. FAIR quality assessment as proposed by DataOne (Data Observation Network for Earth) network.

Here we propose to present this point of view by orchestrating the production of rich (with attributes description and links with terminological resources terms) EML metadata from raw datafiles and, through the generation of FAIR metrics for direct assessment of FAIRness and creation of data standards like Darwin Core. Using EML, we can describe each data attribute (e.g., name, type, unit) and associate each attribute one to several terms coming from terminological resources. Using the Darwin Core vocabulary as a terminological resource, we can thus associate, on the metadata file, original attributes terms to corresponding Darwin Core ones. Then, the data and their metadata files can be processed in order to automatically create the necessary files for a Darwin Core Archive. By acting at the metadata level, associated with accessible raw data files, we can associate raw attribute names to standardized ones, and then, potentially create data standards.

Keywords

Ecological Metadata Language, EML, FAIR, FAIR assessment, terminological resources, ontologies, thesaurus

Presenting author

Yvan Le Bras

Presented at

TDWG 2022

Conflicts of interest

References

- Michener W, Brunt J, Helly J, Kirchner T, Stafford S (1997) Nongeospatial metadata for the ecological sciences. Ecological Applications 7 (1): 330-342. https://doi.org/10.1890/1051-0761(1997)007[0330:nmftes]2.0.co;2
- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7 (1). https://doi.org/10.1371/journal.pone.0029715
- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, 't Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18