Geophysical Research Abstracts
Vol. 20, EGU2018-5313, 2018
EGU General Assembly 2018

Data Integrity test to measure Data "FAIRness"

Fernando Aguilar Gómez and Jesús Marco de Lucas
IFCA-CSIC, Santander, Spain (aguilarf@ifca.unican.es)

Data Life Cycle management is essential to ensure a good quality as well as address the four FAIR principles to make data Findable, Accessible, Interoperable and Reusable. Especially in Environmental Sciences, where the open data production is being increased, the use of a proper standard is needed to include different types of metadata: administrative, descriptive and structural.

A rich enough metadata schema can detail different characteristics at diverse levels: physical and file level, data included in the set, location, etc. The Ecological Metadata Language (EML), is the metadata standard used to describe the datasets related to different environmental sciences initiatives. This standard includes a set of modules capable of describing different aspects of a dataset, including the parameters involved, physical details of the dataset file, associated software etc. Since EML is based on XML, metadata can not only be understood by humans, but also it can be processed automatically and parsed by algorithms. Therefore, different tests can be automatized in order to check the integrity of the datasets.

Metadata describing a dataset can be analyzed in order to check if both address the four FAIR principles along the Data Life Cycle. The "Data Integrity Test", based on a Python script, proposes an approach to validate the data "FAIRness" in the different stages:

Plan: Checks the existence of a Data Management Plan associated.
Collect: Checks if the dataset or digital resource is available and integrate (checksum).
Curate: Validate the Qa/Qc methods applied.
Analyze: Checks the parameter definition in the metadata and validate the automatic processing.
Ingest: Validate the use of Persistent identifiers and the catalog/repository, if it explicit any open protocol like OAI-PMH.
Preserve: Checks the License definition and the preservation details.

The EML is complete enough to describe every single detail that makes a dataset FAIR, so it is the base for the data integrity test. The presentation will explain the different implementation details that allow checking automatically the FAIR data production and how they can be applied to different cases.