Geophysical Research Abstracts
Vol. 21, EGU2019-16513, 2019
EGU General Assembly 2019

Long-term Archive Challenges: Enhancing Data Discovery via Multilevel Metadata Aggregations At Scale

Graham Parton, Ag Stephens, Richard Smith, and Joe Singleton
STFC Ruthford Appleton Laboratory, RAL Space, Harwell Oxford, Didcot, United Kingdom (graham.parton@stfc.ac.uk)

Data archives require accurate, content-rich data catalogues that are fit-for-purpose to support meaningful data discovery. However, sourcing suitable high-quality metadata to populate catalogues at scale can be problematic when manual workflows are no longer able to cope. One solution is automated metadata harvesting directly from archive contents, drawing on technical solutions to Big Data challenges faced by rapidly evolving, petabyte-scale, heterogeneous archives. Yet other issues quickly arise, including: changes in, or lack of, metadata standards over time; missing or incorrect metadata; diversity of, and lack of interoperability between, formats and metadata conventions; and, changes in data availability over time. These are further compounded when dealing with historical archives and legacy systems stretching back decades before comprehensive end-to-end metadata harvesting workflows were envisioned.

The Centre for Environmental Data Analysis (CEDA) Archive has long-term archiving responsibility for the UK atmospheric, climate change and Earth observation communities. With highly heterogeneous deposits of both historic and rapidly growing fresh data holdings, spanning over 5,500 datasets, 200 million files and over 5 Pb of online storage, the CEDA Archive is no newcomer to the 4Vs of the Big Data challenge: Volume, Velocity, Variety and Veracity. Recently CEDA's combined use of parallel archive and processing systems with noSQL Elasticsearch indexes has addressed many of these challenges. This has permitted file-level metadata such as parameter, spatial and temporal information to be indexed and then aggregated to populate most dataset records in CEDA's ISO-driven data catalogue. However, there remain significant shortcomings in the quality and coverage of the metadata harvested via this pipeline (less than 50% of files return parameter information for example) that require additional, complementary approaches.

The key remaining challenges are how to address the following issues: where file-based metadata is incomplete or missing (e.g. unscannable file-formats; incomplete metadata); coping with incorrect file-based metadata (e.g. incorrect geo-temporal information; mapping between coordinate systems); dealing with removed data files; and, coverage for offline/external content. To address this a complementary YAML-driven 'Manual Metadata Store' has been set up operating at the dataset records level in the CEDA data catalogue. This enables a manually maintainable, traceable, versioned alternate metadata source and splicing rules to determine how the information should be used in conjunction with automatically harvested metadata such as CEDA's file-level index. Splicing options include: partial or complete replacement; appending of information; and, setting default information.

Whilst this has allowed CEDA to begin addressing issues with automated metadata harvesting, it has also raised additional questions, especially as CEDA seeks to allow file-level faceted searches in the future, where file-derived metadata is known to be incorrect or missing. At what level should amendments to metadata be applied? The dataset record, file-level index or the original metadata themselves? And how to manage the conflicting pressures of limited resources, preserving original data integrity and the desire to deliver quality information at all levels and at scale. Essentially, the next issue to face can be summed up with the questions: 'What is Truth? And at what level to convey it?'