EGU2020-21258
https://doi.org/10.5194/egusphere-egu2020-21258
EGU General Assembly 2020

Scaling metadata catalogues with web-based software version control and integration systems

Tara Keena, **Adam Leadbetter**, Andrew Conway, and Will Meaney
Marine Institute, Ocean Science and Information Services, Oranmore, Galway, Ireland (tarakeena12@gmail.com)

The ability to access and search metadata for marine science data is a key requirement for answering fundamental principles of data management (making data Findable, Accessible, Interoperable and Reusable) and also in meeting domain-specific, community defined standards and legislative requirements placed on data publishers. One of the foundations of effective data management is appropriate metadata cataloguing; the storing and publishing of descriptive metadata for end users to query online. However, with ocean observing systems constantly evolving and the number of autonomous platforms and sensors growing, the volume and variety of data is constantly increasing, therefore metadata catalogue volumes are also expanding. The ability for data catalogue infrastructures to scale with data growth is a necessity, without causing significant additional overhead, in terms of technical infrastructure and financial costs.

To address some of these challenges, GitHub and Travis CI offers a potential solution for maintaining scalable data catalogues and hosting a variety of file types, all with minimal overhead costs.

GitHub is a repository hosting platform for version control and collaboration, and can be used with documents, computer code, or many file formats

GitHub Pages is a static website hosting service designed to host web pages directly from a GitHub repository

Travis CI is a hosted, distributed continuous integration service used to build and test projects hosted at GitHub

GitHub supports the implementation of a data catalogue as it stores metadata records of different formats in an online repository which is openly accessible and version controlled. The base metadata of the data catalogue in the Marine Institute is ISO 19115/19139 based XML which is in compliance with the INSPIRE implementing rules for metadata. However, using Travis CI, hooks can be provided to build additional metadata records and formats from this base XML, which can also be hosted in the repository. These formats include:

DataCite metadata schema - allowing a completed data description entry to be exported in support of the minting of Digital Object Identifiers (DOI) for published data

Resource Description Framework (RDF) - as part of the semantic web and linked data

Ecological Metadata Language (EML) - for Global Biodiversity Information Facility (GBIF) – which is used to share information about where and when species have been recorded

Schema.org XML – which creates a structured data mark-up schema to increase search engine optimisation (SEO)

HTML - the standard mark-up language for web pages which can be used to represent the XML as a web pages for end users to view the catalogue online

As well as hosting the various file types, GitHub Pages can also render the generated HTML pages as static web pages. This allows users to view and search the catalogue online via a generated static website.

The functionality GitHub has to host and version control metadata files, and render them as web pages, allows for an easier and more transparent generation of an online data catalogue while catering for scalability, hosting and security.