Plan Overview

A Data Management Plan created using DMPonline

Title: Applying Linked Data and Semantic Web Standards to Improving Interoperability of GBIF Datasets

Creator: Marcos Zárate

Principal Investigator: Marcos Zárate

Data Manager: Marcos Zárate

Affiliation: Other

Template: DCC Template

ORCID iD: 0000-0001-8851-8602

Project abstract:

Data quality is one of the highest priorities for Global Biodiversity Information Facility (GBIF), the national nodes and other providers, depends on both automatic methods and community experts to detect and correct data issues. Not all issues can however be automatically detected or corrected, so community assistance is needed to help improve the quality of exposed biological data. Semantic Web is a new approach where conventional web documents can be extended with additional data that add meaning to them rather than structure alone. On the Semantic Web, data can be retrieved from seemingly unrelated fields automatically in order to combine them, find relations, and make discoveries. The research proposal aims to use Semantic Web technologies to convert datasets published in GBIF to Linked Open Data (LOD) in order to perform quality controls in an automated way and promote the integration of information with other datasets relevant for Biodiversity.

ID: 39871

Last modified: 27-03-2021

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customise it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

Applying Linked Data and Semantic Web Standards to Improving Interoperability of GBIF Datasets

Data Collection

What data will you collect or create?

The primary data that will be used come from Darwin Core files (DwC-a) published through the institutional IPT belonging to the National Patagonian Center (CENPAT-CONICET) http://ipt.cenpat-conicet.gob.ar:8081/ .

Notes:

Darwin Core Archive (DwC-A) is a biodiversity informatics data standard that makes use of the Darwin Core terms to produce a single, self contained dataset for sharing species-level (taxonomic), species-occurrence data, and sampling-event data. An archive is a set of text files, in standard comma- or tab-delimited format, with a simple descriptor file (called ***meta.xml***) to inform others how your files are organised. The format is defined in the Darwin Core Text Guidelines. ***It is the preferred format for publishing data in the GBIF network.***

How will the data be collected or created?

The data is collected by our samples or observations of marine and terrestrial species, and then digitized using the tools provided by GBIF to create the DwC-A files

Documentation and Metadata

What documentation and metadata will accompany the data?

The preferred format for publishing data to the GBIF network is the Darwin Core Archive, and its Integrated Publishing Toolkit uses EML as its metadata standard.

Ethics and Legal Compliance

How will you manage any ethical issues?

How will you manage copyright and Intellectual Property Rights (IPR) issues?

Storage and Backup

How will the data be stored and backed up during the research?

Qualitative data will be backed up and secured by the lead country researcher on a regular basis and metadata will include clear labelling of versions and dates. There are some potential sensitivities around some of the data being collected, so the project will establish a system for protecting data while it is being processed, including use of passwords and safe back-up hardware.

The primary data after the conversion, is stored in GraphDB triplestore, accessible publicly from http://web.cenpat-conicet.gob.ar:7200/ . This server makes copies daily.

How will you manage access and security?

GraphBD manages the users and the privileges that each user has to access the repository. In case of possible unauthorized users, graphDB controls and reports this situation to the administrator.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

In this research, because it is primary biodiversity data that can be used in various analyzes, it is important that the primary data and metadata are persistent for several years.

What is the long-term preservation plan for the dataset?

Data will be kept for at least 10 years. After this time the data may be subject to deletion it has not been reused, accessed, or cited.

Data Sharing

How will you share the data?

The most appropriate means of sharing the data generated through the project will be online, through institutional websites. The project will have a dedicated space on the CENPAT nstitutional website to facilitate this, and all other involved institutions will also be encouraged to host the data on their websites.
Linked Open Data is available for access through GraphDB http://web.cenpat-conicet.gob.ar:7200/ , it can be downloaded as data dump or accessed through the SPARQL end point.

Are any restrictions on data sharing required?

Researchers should respect the following rights statement:
The publisher and rights holder of this work is CCT CONICET-CENPAT Centro Científico Tecnológico. This work is licensed under a Creative Commons Attribution Non Commercial (CC-BY-NC) 4.0 License.

Responsibilities and Resources

Who will be responsible for data management?

Renato Mazzanti Centralized Computer Service Manager CENPAT-CONICET Bvd. Brown 2915U9120ACF Puerto Madryn, CHUBUT, Argentina renato@cenpat.edu.ar 54-0280 - 4451024 - 445040

What resources will you require to deliver your plan?