Linking Taxonomy to CF

Roy Lowry
NOC Emeritus Fellow
Storing Biological Data in CF

- CF labels measurements using Standard Names controlled vocabulary
  - Content governance through community discussion on GitHub
  - Technical governance through versioned files (text and XML) and vocabulary servers
- Expanding the controlled vocabulary requires resource
- Massive expansion has performance or even viability implications for technical governance
- Much biological data has one measurement per taxon
- Tens of measurements for hundreds of taxa threatened massive expansion of Standard Names vocabulary
Storing Biological Data in CF

• Vocabulary expansion avoided by considering the taxon as a data co-ordinate
• A single timestep species distribution map has abundance values stored in a 3D array with the co-ordinates
  • Latitude
  • Longitude
  • Taxon
• Each co-ordinate has one or more 1D array auxiliary variables that need to be populated
• The abundance values and each auxiliary need labelling with a standard name
Population of the Taxon Auxiliary Variables

• Each element in each of the taxon co-ordinate auxiliary variables stores a taxon label for one of the map layers in the 3D array
• Obvious label to use is the plaintext taxon name
  • Advantages
    • Human-readable
    • No management overhead
  • Disadvantages
    • Subject to spelling errors
    • Subject to naming variations e.g. Eristalis abusivus and Eristalis abusive
    • Potential to be populated by total garbage e.g. taxonomist’s pet name
• Included as readability makes data file self-contained
• Given the standard name biological_taxon_name
Population of the Taxon Auxiliary Variables

• A machine-readable, standard-conformant taxon label is highly desirable
• Having such a label connected to a machine-readable taxonomy is even more desirable
• Well-managed internet resources that provide taxa with permanent identifiers
• Identifiers may be incorporated into resolvable URIs
• Machine-readable label(s) may be provided by populating one or more taxon auxiliary variables with either PIDs or URIs
Population of the Taxon Auxiliary Variables

• PID/URI sources need to be approved by the CF community
• Initially two were selected
  • World Register of Marine Species (WoRMS) – good coverage of marine flora and fauna
  • International Taxonomic Information System (ITIS) – good coverage of terrestrial flora and fauna
• Others may be proposed for consideration through GitHub
• WoRMS give each taxon a PID known as an AphiaID
• ITIS give each taxon a PID known as a Taxonomic Serial Number (TSN)
• Both PIDs may be incorporated into URIs that resolve through service APIs
Population of the Taxon Auxiliary Variables

- Initially it was proposed to have two partially-filled auxiliary variables, one for AphialID-based URIs and the other for TSN-based URIs.
- Following feedback it was decided to use the Life Science ID (LSID) syntax as a single vehicle capable of carrying either ID.
- LSID syntax:
  - `urn:lsid:<Authority>:<Namespace>:<ObjectID>[::<Version>]
  - `urn:lsid:marinespecies.org:taxname:104464" for AphialID 104464
  - `urn:lsid:itis.gov:itis_tsn:180543 for TSN 180453
  - URN to URL by prefixing with http://www.lsid.info/
- The auxiliary variable has been given the Standard Name: `biological_taxon_lsid`
- The abundance data array has been given the Standard Name: `number_concentration_of_organisms_in_taxon_in_sea_water`
Population of the Taxon Auxiliary Variables

• LSIDs are not universally loved but they provided a practical way to combine AphialIDs and TSNs into a single auxiliary variable in a machine-understandable manner

• Other identifiers – either PIDs or URIs - could be incorporated
  • Each identifier added needs a Standard Name to be proposed and agreed by the community through the GitHub procedure
  • The first time this is done the Conventions Document will also need minor amendments

• Not aware of the mechanism being used in anger

• Oceanographic biogeochemical community being steered towards CF so its day will come