Workshop Wiki: CMNH & ANS Workshop: Empowering Database Decision-Making - iDigBio
Draft Agenda: DRAFT Workshop Agenda - Google Docs

Agenda

Best Practices: Data Standards & Controlled Vocabularies

- Importance of standardization
- Implementing controlled vocabularies
- Tools & resources

Data Quality: Common Issues

- Common data quality issues

Best Practices: Data Cleaning & Preparation

- Data cleaning techniques
- Data preparation steps

OUTLINE

1. Introduction
    a. Why should I listen?
    b. Works
    c. Workflow Overview
        i. Guide for sessions
2. Terminology
    a. Basic Terms
    b. KOS Types https://docs.google.com/spreadsheets/d/1Clusc65jid6j6Jj-FBveN5tchmBPavTda8sGcDyU5r4/edit?gid=1762766037#gid=1762766037
    c. Human and machine-readable
    d. Preferred and alternate labels
    e. Verbatim values and Preservation strategy
3. Data Standards
    a. A standard that defines a set of data according to a set of data structuring rules so that the set of data can be interchanged between one computer system and another [ISO/IEC TR 10032]
    b. Interoperability: the ability of two or more systems or applications to exchange information and to mutually use the information that has been exchanged

c. TDWG
    i. What is TDWG
    ii. Standards ratification process https://www.tdwg.org/about/process/
    iii. Participation
d. Darwin Core: https://dwc.tdwg.org/terms/
    i. Flagship specimen-level metadata standard
e. Latimer Core: https://ltc.tdwg.org
    i. Collections-level Metadata
    ii. Stand-alone standard
f. Minimum Information about a Digital Specimen https://github.com/tdwg/mids
    i. Digitization Metric Framework
    ii. https://dev.mids.dissco.tech/
g. Mineralogy Extension
    i. In Process: https://docs.google.com/spreadsheets/d/1NjoS_JVuh-N26e3tAVNoO7nnnDZerUEl03WPHmmHi8E/edit?gid=1828619460#gid=1828619460
    ii. Vocabulary Enhancement
h. International Image Interoperability Framework https://iiif.io/
    i. A set of open standards for delivering high-quality, attributed digital objects online at scale
    ii. API Response Specifications
i. Dublin Core
    i. High-level, general terminology
    ii. Frequently utilized by domain standards such as LTC and DWC
    iii. DCMI Metadata https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
j. Schema.org
    i. Another general-use standard
    ii. https://schema.org
k. SKOS
    i. Mapping

4. Controlled Vocabularies
    a. Contain values assigned to terms (Data standards contain the terms)
    b. Authoritative Sources
    c. Terms with Definitions

   d. DWC/TDWG Controlled Vocabularies
   e. The availability of simple yet rigorous domain-specific controlled vocabularies that can be easily integrated into existing systems is a pending opportunity and something I spend time working on
   f. See ANSI/NISO Z39.19-2005
5. Data Quality
   a. Define Data Quality Dimensions - Framework
      i. Formats (Lat/Lon)
      ii. Valid Values (ISO 3166)
      iii. Annotations (? to indicate uncertainty)
      iv. Data types (numeric, string, date time)
   b. Names
      i. Understandability - A name should describe the concept it represents
      ii. Conciseness - A name should use only the words necessary to communicate the concept it represents
      iii. Consistency - Names should be used for formatted consistently
      iv. Distinguishability - A name should be visually and phonetically distinguishable from other names
   c. Identifiers
      i. Human vs Machine Consumption
      ii. Dimensionless, Arbitrary and Agnostic
   d. Document and Define Everything
   e. Examples
      i. Geologic Timescale
      ii. Type Status

GLOSSARY

Data Standard

A standard that defines a set of data according to a set of data structuring rules so that the set of data can be interchanged between one computer system and another [ISO/IEC TR 10032:2003]

Interoperability
the ability of two or more systems or applications to exchange information and to mutually use the information that has been exchanged [ISO/IEC TR 15944-14:2020]

Data Interoperability
interoperability concerning the creation, meaning, computation, use, transfer, and exchange of data [ISO/IEC 20944-1:2013]

Controlled Vocabulary
Finite set of values that represent the only allowed values for a data item [ISO 11179]

Machine Readable
Pertaining to data in a form that can be automatically generated by and input to a computer. [ISO/IEC/IEEE 32675:2022]

Term
Written or verbal designation of a general concept in a specific domain [ISO 5127]
A machine-friendly written or verbal designation of a general concept in a specific domain that follows a specific naming convention suitable for machine use (*Data Standards)*

Label
A human-friendly (readable) representation of a concept that follows a common naming convention of proper or lowercasing

Column Name
An alteration representation of a concept that follows a naming convention of a database

Field Name
Alternate representation of a concept specifically for use in spatial data forms

RESOURCES

Datasets

https://www.gbif.org/occurrence/download/0014609-241007104925546
GBIF Distinct Values: https://github.com/tdwg/dwc-qa/tree/master/data/GBIFDistinctValues

Examples

Type Status: https://docs.google.com/spreadsheets/d/1psE6bs0z-rM6oC7F2cTyQcM9D9nidj5CdLIpB2BggsQ/edit?usp=sharing
Geologic Timescale
https://docs.google.com/spreadsheets/d/1CmrpRDT_-klZ23zJbpqhb0qxgUN6HQCk876O39hu9HQ/edit?usp=sharing

Publications

Chapman AD (2005) Principles of Data Quality. Global Biodiversity Information Facility. https://doi.org/10.15468/doc.jrgg-a190

GBIF Data Quality Checklist
https://ipt.gbif.org/manual/en/ipt/latest/data-quality-checklist

ANSI/NISO Z39.19-2005 (R2010). Guidelines for the construction, format, and management of monolingual controlled vocabularies. ISBN: 1-880124-65-3. Available at: https://www.niso.org/publications/ansiniso-z3919-2005-r2010

KOS Types Vocabulary https://nkos.dublincore.org/nkos-type.html

TDWG Standards Specification:
https://github.com/tdwg/vocab/blob/master/sds/documentation-specification.md

Collections Data Workflow Document
https://docs.google.com/document/d/13EsbLwJy8JPcdLvJcrqBbF3-IMn1CwVoO92wnKhmwRg/edit?usp=sharing