Road map for GBIF vocabularies

Introduction

GBIF's vocabulary server (Apache-2.0) is a dynamic, web-based platform that empowers the community to easily access and utilize controlled terms that standardize the content shared across the GBIF network. For users, GBIF's vocabulary server functions as an intuitive access point where anyone can browse, explore, and share URLs to term definitions within any vocabulary. These URLs are stable and always reflect the most up-to-date content.

For editors, GBIF's vocabulary server serves as a collaborative environment for creating, defining, and managing new vocabularies and concepts.

The vocabularies available on GBIF's vocabulary server, as well as those under consideration, are typically based on already established external vocabularies or ontologies. A GBIF-relevant vocabulary may be constructed from multiple external sources. However, GBIF aims to avoid redundancy by not hosting vocabularies that are already well-defined, complete, stable, and accessible elsewhere. Instead, these vocabularies are integrated directly from their sources, such as the Getty Thesaurus of Geographic Names, GADM, and certain vocabularies curated by the Catalogue of Life.

Most vocabularies hosted on the GBIF vocabulary server focus on standardizing verbatim values from data sources. This includes mapping different language variations, abbreviations, phrasings, and grammatical differences to the controlled terms within a given vocabulary.

The development and maintenance of GBIF's Vocabulary Server, along with the creation of new vocabularies, are managed through an active GitHub repository.

Below is the roadmap for GBIF's vocabulary server for 2024-2027. Please note that this plan is subject to updates as new work programs are initiated. The first item, 'Occurrence vocabularies and implementation' is part of the 2024/25 Work Program:

- Revising the documentation on record-quality requirements, including occurrence identifier stability, and use of standardised values, including vocabularies
- Encourage growth of the community of vocabulary editors to accelerate the collection of community-curated vocabularies

Item #2-3 are expected to be handled within the Strategic Framework 2023-2027.

1. Occurrence vocabularies and implementation

Vocabulary prioritisation

Future vocabulary development will be prioritized within GBIF's annual work programs. The initial phase of this development is centred on enhancing GBIF's internal processes, specifically, those related to the interpretation of source data, to refine terms critical to the Darwin Core (DwC) occurrence core. However, certain vocabularies may be deferred to a backlog, pending the availability of external resources or further community engagement.

Vocabulary construction, mapping and implementation

The following vocabularies are planned to be finalised and implemented in interpretation pipelines:

- TypeStatus
- GeoTime (One vocabulary that covers 10 terms for chronostratigraphy/geological time defined by Darwin Core)
- Sex
- Month
- GeodeticDatum

This includes mapping concepts to verbatim values and uploading vocabularies to the test (UAT) and the production environment (PROD).

Updates to implemented vocabularies

Some of the vocabularies already in production will be revisited and checked to see if further concepts should be added from existing, external vocabularies and if hidden values should be mapped accordingly:

- EstablishmentMeans
- LifeStage

Vocabulary definitions and implementation during the strategic framework 2023-2027

The following vocabularies are expected to be migrated or implemented during the strategic framework but require scoping, community input or consideration:

- DatasetCategory (currently on UAT, implementation requires consideration)
- Preparations (may be split into multiple vocabularies and dependent on TDWG community efforts)
- CountryName
- Realm
- Biome
- EnvironmentalMaterial
- EventType

Realm, Biome and EnvironmentalMaterial will be defined and requested as new Darwin Core terms in a TDWG task group (WP25). The vocabularies may be hosted externally, depending on the output of the task group.

2. Community engagement & updates to the Registry

Explore the option for internal (GBIFS) and external management of vocabularies by creating roles and scopes similar to the permissions in GRSciColl (GitHub issue). As vocabularies come into production, maintenance will be required to update items based on data standards and research community development. Updates will be handled by expert communities and the GBIF Secretariat.

Engage in Biodiversity Information Standards (TDWG) development to align GBIFS vocabulary efforts with the ongoing work in TDWG.

3. Additional technical development

Vocabulary maintenance and management

During the production phase of a vocabulary, its maintenance should be supported by various implementations to ensure ongoing relevance and accuracy. As more vocabularies or terms become available, versioning will be essential to maintain a well-structured system. The database schema (Fig. 1) will be periodically evaluated to ensure its effectiveness. GBIFS will also explore the potential benefits of integrating A Simple Standard for Sharing Ontology Mappings (SSSOM) and enhancing compliance with the Simple Knowledge Organization System (SKOS) structure to determine if these measures can optimize or reduce the technical demands on interpretation pipelines.

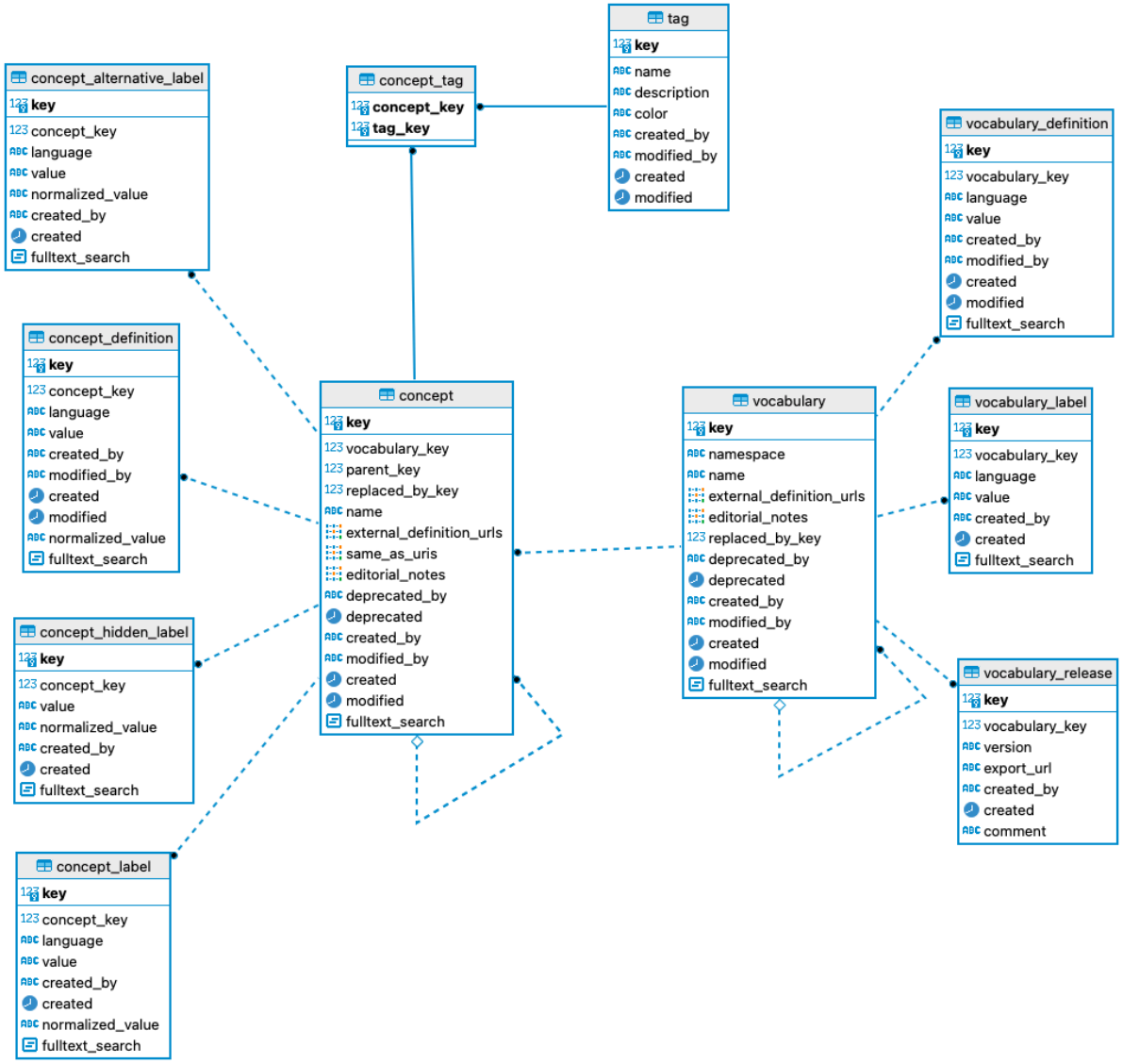

Figure 1) current database schema of GBIFs vocabulary server.

Integration of vocabularies on GBIF web pages

GBIFS will investigate the most effective ways to integrate the server and its associated vocabularies across GBIF's services and products. This exploration will cover the presentation of vocabulary definitions, the display of values in multiple languages, and the implementation of search functionality for hierarchical controlled terms. A well-designed user interface (UI) is crucial to ensure that these vocabularies are accessible and easy to navigate, enhancing the overall user experience. By thoughtfully displaying vocabularies and their associated data, GBIF can better support users in finding and applying the correct terms, ultimately improving data quality and consistency across the network.