7

HOW TO USE A TAXONOMIC RESOLUTION ENGINE WIKIBASE

Target group

Researchers interested in the linked data concepts that **underpin digital taxonomy**, such as digitally available type specimens, name registries and taxonomic treatments.

Goal

The challenge of managing such links was addressed by TETTRIs using a custom Wikibase application, taking an explicit taxonomic focus. The data model was conceived with links between these different concepts in mind, as well as their provenance. The prime focus of the **Taxonomic Resolution Engine (TRE)** was to link **taxonomic names** to their reference type specimens, but the tool can be used to manage other key links as well, such as **material citations** in published taxonomic treatments, taxonomic authors or other parts of the Biodiversity Knowledge Graph (Page 2016).

Cetonia aurata | Photo by Amelie Höcherl

Summary/Description

A Wikibase is a generic implementation of a linked open data platform, consisting mainly of items that are annotated with properties, which themselves can link to other items, string values or external data sources. A Wikibase comes with an intuitive user interface, several **Application Programming Interfaces (APIs)** for bulk modifications and exports, a strong versioning system for each item and various services **supporting communal contributions**, such as discussion pages and templates for certain types of items. It is designed as a secondary data source, always linking back to the original sources, while making relations between concepts explicit and easy to query for.

The **TRE** is a Wikibase set up with a primary focus of managing explicit links between taxonomic names and the specimens used as types to register these names. It includes tools to populate the Wikibase, validate links as data in the source systems changes and provenance mechanisms, including some fallback measures where the stability of source identifiers is unclear. The Wikibase was set up as a service hosted by the Wikibase.cloud platform, maintained by **Wikimedia Germany**. For more detailed background information, read TETTRIs milestone 3.

You Will Need

Human resources:

- Data scientists with some familiarity with data wrangling and batch API requests in Python.
- Taxonomic experts who understand the intricacies of taxonomy and can navigate the landscape of different data sources (often with their own intricacies).

Steps to implement the Course

1 **Acquire an account on the Wikibase.**
This can be done by signing up to the TRE on this page. Please note that you may see a strange anime cat girl character first: this is normal, the figure is the mascot of the Anubis website protection software, which is used by the maintainers of the Wikibase.cloud platform to block malicious and excessive use of some of their websites.

Accounts need to be approved by administrators of the TRE, to ensure the security and usability of the site.

2 **Define the initial scope**
Draw out what you want to contribute. This can be a certain taxonomic group, or a more general dataset that you have worked on. Identify the different data elements that you want to publish and compare them to the data models in Milestone 3. The TRE currently has a strong focus on documenting type specimens, but this can be extended to include material citations in published taxonomic treatments, taxonomic authors or other parts of the Biodiversity Knowledge Graph.

3 **Discuss any required extensions to the existing data model with input from diverse stakeholders**
If new properties are needed, it requires community discussion and the involvement of the administrators. A community discussion can be opened on the Wikibase by creating a new page and opening a discussion. Only do this if a significant userbase is active on the TRE. If not, other discussion venues should be used, such as a CETAF working group like the CETAF Information Science & Technology Commission (ISTC).

4 **Acquire data from different sources**

The TRE is designed to identify its key items through persistent identifiers. If these are not part of your dataset, or other key metadata is missing, you will need to acquire and harmonize it from different sources. Example R scripts that do this are available on GitHub.

5 **Link the different data concepts**

The TRE is designed with explicit links in mind. There are items for the different concepts (specimens, names, datasets) and there are items for the links between them (e.g. typification assertion of a specimen for a name). This allows the links to be easily updated if new information arises or the link was inaccurately claimed.

6 **Develop or modify the scripts and other modifications needed for automated import**

Python scripts using the wikibaseintegrator package are available on GitHub. These make use of the Wikimedia API and require an OAuth 1.0a consumer to be generated for the user's account.

Users can also make use of the Quickstatements module to load batch additions or modifications into the Wikibase from a CSV file. A user needs to have the status of "autoconfirmed" in order for them to be able to use the Quickstatements service. This can be accomplished by making 50 manual contributions or contacting an administrator.

7 **Execute the import and/or modification operations**

It is recommended to set up the Wikimedia API script of the import with a limit of no more than 10 requests per second, and to work in batches of 10 to 40,000 changes. This makes a timeout less likely.

Quickstatements are directly set up as a job on the server itself and do not require a local script to run. It is recommended to work in smaller batches, though, closer to 10,000 at a time, as the current version of Quickstatements does not support request limiting and hence a large batch may cause the user to exceed an API limit.

8 **Set up a bot or cron job to periodically perform automatic updates**

A fully automated pipeline can be set up, which periodically acquires data from different sources, links the concepts and publishes it to the TRE. The pipeline can and should also request data from the TRE through the SPARQL interface to look for duplicated or out-of-date information, such as broken links, changed metadata or drifting identifiers. An example of such a pipeline can be found (soon) on https://github.com/matdillen/ttrypes

Timeframe

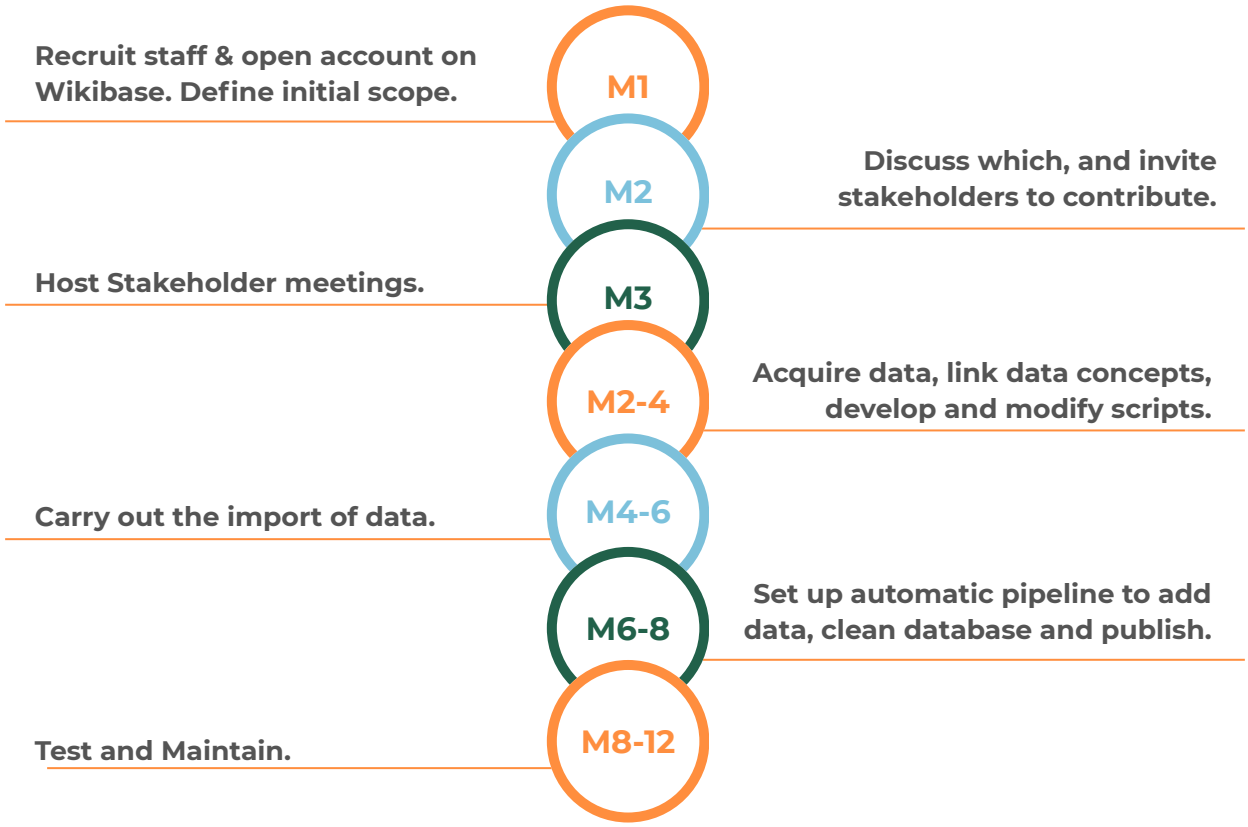
Estimating the time frame is very difficult. In theory, all of these steps can proceed very efficiently and the contribution to the TRE can be up and running with automated updating scripts after a few weeks. If the scripts are set up properly, 100,000 additions or updates can be reliably finalised **in the course of a day**.

In practice, there are many potential delays in every step and the ultimate setup workflow is more likely to take several months. Organising workshops with stakeholders is also likely to take weeks of **planning and finding dates to bring enough contributors together**. Data acquisition and linking can run very smoothly, but bridging different data models can cost extra time. In particular data stability and versioning across infrastructures may differ widely and introduce additional complications. A full year is a more realistic and safe time frame.

From the above: here is a rough timeframe to include in the cookbook:

The timeframe is 3 - 12 months.

In theory, all of these steps can proceed very efficiently and the contribution to the TRE can be up and running with automated updating scripts after a few weeks. If the scripts are set up properly, 100,000 additions or updates can be reliably finalised in the course of a day. In practice there are potential delays in each step, thus we have given a 12 month time frame.



Estimated Budget (Indicative)

Provide a simple cost estimate for implementing this recipe. Break it down by major categories such as:

PERSONNEL

- Data scientist (for 1 year)
- Taxonomic expert (for 1 year)
- Site administrator (for 1 year)

TRAVEL & LOGISTICS

- 3 stakeholder workshops. These could be virtual or fit into a larger event, reducing costs.
- Server hardware, cloud hosting or service hosting fees.
- Dedicated PC/server to perform the automated updates.

Meloe violaceus Triangulos larvae | Photo by Amelie Höcherl

What went well / Even better if

- **Wikibase is a very powerful tool for this purpose.** It has the data structure, the APIs, the query engine, the user interface, customization features and more for what someone might want for a novel platform.

- **Using wikibase.cloud removed many potential complications, but also introduced a few others.** Occasional bugs slowed the development of the platform, sometimes taking weeks before the development team could address them. On the other hand, these issues could have crept up on a self-managed installation as well and then we might have needed to consult external support to address them - or spend extra time ourselves.

- **There are many intricacies with different data sources that complicate this process.** A lot but not all data is open and can thus be used freely. Other data sources however have overly restrictive licenses making them unavailable for ingestion in the TRE. Versioning and stability of records is still problematic. Identifiers for concepts are not always consistent, do not persist or multiply.

- **Wikibase is not too fast in ingesting data updates without direct backend access.** So a dedicated machine to perform the bulk updates and log retries is very important. Updates could be performed much more rapidly with a self-hosted Wikibase, through direct database access, but this is accompanied by a very significant risk of data corruption. The software stack behind a Wikibase is not a simple one and requires significant experience to manage properly.

- **Operate on an invite-only basis** (i.e. new accounts need admin approval to work, anonymous contributions are not allowed, which took up more staffing time). This was a necessary restriction as spammers and malicious users are everywhere and instances of the wikibase.cloud platform are trivial to locate.

Optional: Related Deliverables or Resources

- TRE test Wikibase.

Meloe violaceus | Photo by Amelie Höcherl

Panorpa communis | Photo by Amelie Höcherl

Funded by
the European Union