TNLS: Taxonomic activity notification service

1 Introduction

Work package 2 of the TETTRIs project and the TETTRIs satellite project TNLS focussed on mechanisms to name match a list of scientific names against an aggregator's taxonomic dataset. Apart from checking the data quality and other properties of the scientific name, users get the possibility to link their name to a name-identifier (ID) the database provides. This would enable a researcher to follow scientific names and get information of taxonomic or nomenclatural changes affecting them. The idea is to enable users to be informed (via push or pull mechanisms) about changes that affect their name. Changes (as described below) would refer to defined "versions" of the dataset, e.g. daily or weekly snapshots.

Here, we are looking into some aspects of such a system, classifying potential audiences, exploring possible changes affecting names in a taxonomic dataset, and identifying technologies that may be used to implement such a service. We are focussing on changes in the core taxonomic and nomenclatural data (classification, acceptance, nomenclature) but this could extend to the addition of new occurrence records or specimens.

2 Analysis of possible services

2.1 Audiences

1. **Taxonomists** responsible for maintaining a coherent classification of organisms.
2. **Biologists** who study aspects of organisms separate from classification of those organisms, such as ecologists and morphologists.
3. **Policy makers** who take decisions and create legislation based on the work of taxonomists and other biologists. This could range all the way from politicians to those running border posts.
4. **Data scientists** analyze large datasets to identify patterns, create predictive models, and help organizations solve complex problems by combining skills in statistics, computer science, and domain knowledge.

2.2 Changes affecting names in taxonomic datasets

Changes may occur within the scope of a single taxon or within a list of taxa. Taxonomists are most likely to be interested in activity within a taxon and its descendants (i.e. global checklist) but biologists and policy makers are more likely to want notifications based on a list of taxa, such as those in a certain geography, a theme or governed by certain regulations.

Confusion can arise in object graphs and relational models if we don't have clear, shared definitions of what we mean by change. This is because the boundary of an object may vary for different audiences in different contexts. Has the object changed or was the change just in something linked to that object? Here we set out some pragmatic categories of changes with comments on which audience may be interested in such changes. It assumes that names are identified by immutable, opaque identifiers. The things that change are properties of the thing (name) signified by the identifier.

1) **Nomenclatural changes:** Changes concerning the data that is controlled by the code of nomenclature applicable to the organism concerned (ICZN, ICN, ICNAFP), having no effect on taxonomy.

    a) **Rendering:** These are changes that will affect the way the full name (including authors and year) appears when written. It might be a spelling correction or change to preferred author abbreviation. This is of a wide interest to all audiences, taxonomists, biologists and policy makers.

European Innovation Council
Co-funded by the European Union

b) **Technical:** These are changes to the data associated with the publication of the name. It might include changes in the citation of the place of publication or a lectotypification event. The changes won't affect the way the name is written but would be of interest to taxonomists who study this group.

2) **Taxonomic:**

a) **Taxonomic status:** This is a change in the taxonomic status of a name. It is relevant to all audiences.

i) **Accepting:** Fired when a name becomes the accepted name of a taxon, when it wasn't the accepted name of a taxon before.

ii) **Sinking:** Fired when a name becomes a synonym within a taxon when it was previously an accepted name or not in the classification at all. This may include nomenclaturally invalid designations and names at superfluous ranks.

b) **Placement**

i) **Removing:** When a name has been removed from the classifications (when it was previously an accepted name or synonym). This might happen to names at superfluous ranks or doubtful nomenclatural status (nomen nudum). If possible, link the removed name to a new equivalent (sinking). Relevant to all audiences.

ii) **Moving:** When the direct 'parent' of the name changes. This may be when a genus moves to a different family or when a synonym moves between accepted names. This change is likely to be of interest to taxonomists and biologists but less likely to be relevant to other audiences.

iii) **Ancestral:** When there is change in the path to the root of the classification above the level of the immediate parent. E.g., if a species is being followed and the genus of the species is placed in a different family (implies change in membership, too). This is of most interest to taxonomists and of less interest to other audiences although may be significant where legislation is bound to higher level taxa, such as CITES.

c) **Membership:** When the members of a taxon change. Unless restricted by a sub filter this would include any name (accepted or synonym) added, removed or moved in the followed taxon or any of its descendants (including direct taxa). Tell me any changes below this level! Membership changes are likely to be mainly of interest to taxonomists but may be significant for others when policy is bound to higher level taxa.

i) **Synonyms:** A synonym is added or removed from the followed taxon directly.

ii) **Children:** A taxon is added or removed from the followed taxon.

iii) **Descendants:** A taxon is added, removed or moved below the level of the immediate children of the followed taxa. E.g. A subspecies is added to a species in a genus that is being followed.

iv) **Potential**: A name without clearly resolved placement has been found that may be a descendant of the taxon followed. This is of particular interest to taxonomists responsible for the names within a group.

d) **Siblings:** When a taxon is added or removed from the parent of the followed taxon. E.g. A species is added to the genus of a species that is being followed. This is probably only of interest to taxonomists as an indication of a changing classification.

3) **Identifiers:** Sometimes identifiers are created in error. Systems should not break when this happens but simply resolve the IDs to the correct place. Consumers should be notified when changes occur:

   a) **Same-as:** The identifier has a new same-as relationship with another identifier.

   b) **Different-from:** The identifier no longer has a same-as relationship.

   c) **Unknown**: The identifier has become of doubtful application.

2.3 Implementation technologies/methods

Below, we list a number of existing technical solutions for notifications based on taxonomic databases. The order of appearance is random and does not indicate any ranking.

- **Linked Data Event Streams** (LDES) is an append-only collection of members described using the Resource Description Framework (RDF). The specification says how a client must replicate the history of an event stream, and how it can then remain synchronized as new members are published (https://w3id.org/ldes/specification)
    - Example https://www.marineregions.org/feed (applied on Aphia soon)
- The **ActivityPub** protocol is a decentralized social networking protocol based upon the ActivityStreams 2.0 (https://www.w3.org/TR/activitypub/#bib-ActivityStreams) data format. It provides a client to server API for creating, updating and deleting content, as well as a federated server to server API for delivering notifications and content.
    - Example: The **Fediverse** is a collection of social networking services that can communicate with each other (formally known as federation) using common protocols, mainly ActivityPub. (https://jointhefediverse.net/).
    - A possible approach would be for a taxonomic database to become its own Mastodon server with each accepted taxon (or subset of ranks) appearing as their own accounts. Users of other ActivityPub implementations such as BlueSky could then subscribe to these users.
- **Pub/Sub** is an asynchronous and scalable messaging service that decouples services producing messages from services processing those messages (https://cloud.google.com/pubsub/docs/overview)
- **RSS** (RDF Site Summary or Really Simple Syndication) is a widely used news federation technology that is implemented in a number of related standards
    - Example: Aphia RSS Feeds (https://www.compositae.org/gcd/aphia.php?p=stats, right top)
    - Example: WFO RSS Feeds for changes within the Rhakhis editor at the Order and Family level (https://rhakhis.rbge.info/rhakhis/api/downloads/rss/). These are updated daily and linked to from within the interface but are only really relevant to data editors, not a wider audience
- **OAI-PMH** (Open Archives Initiative Protocol for Metadata Harvesting) is a widely used protocol for synchronising data between archives first launched in 2001. Its primary use is between databases rather than as a notification service for users.
- **Email**
    - Example: Aphia weekly digest. Taxonomists receive weekly updates tailored to the groups to which they are subscribed. These updates may include changes in taxonomic names, species distributions, specimen records, images, and other relevant data. Within their Aphia profile, users can (un)subscribe to any of the groups for which they are responsible.
    - Example: WFO has a manual process of notifying TENs (Taxonomic expert networks). After a synchronisation with IPNI a python script runs over to extract changed nomenclatural data and chunks the data into records belonging to each TEN. These are pushed to GitHub. An email is sent to the TENs mailing lists with a list of the TENs with an overview list

detailing the number of records per TEN.

- **Draxlr** can connect it to any database and define SQL queries that monitor for data changes, supporting "Send Alerts Only on Change", so if you have a table tracking species records, you could write a query detecting e.g. “new species added” or “status changed”

3 Activity in literature of interest to curators of taxonomic databases

We also identified that in addition to the aforementioned services, some service alerting taxonomists about events occurring in the taxonomic literature is necessary. The content of the taxonomic databases we manage originates in the scientific literature. Each nomenclatural and taxonomic change should always be evidenced by a written publication. There should therefore exist a continuous process of scrutinizing the literature for new evidence to incorporate. There is potential for this to become more automated. We recognise 3 broad classes of event:

1. Individual events such as **new species names or taxonomic changes** being published as parts of larger works.
2. **Complete taxonomic works** (floras, faunas and monographs) that need to be assessed as a whole by taxonomic experts.
3. Implicit/**implied changes**. When a work makes a taxonomic change that may affect other taxa not mentioned directly in the work.

Each of these events could lead to any of the changes listed in the databases section above. A notification service could be as simple as to notify of an event in the literature that falls within the scope of a taxon being followed or it could categorise that event (using the vocabulary above) or it could go as far as initiating the necessary database changes to incorporate that new evidence.

To be able to achieve proper literature notifications, we need a number of tools or components

Table 1. Components

Component	Description
Trigger	Finding out that a new (and relevant) publication was published
Extraction	Ingesting articles (unstructured PDF or structured formats), identifying and extracting species treatments. Manual, semi-manual (i.e. via templates) and automatically (i.e. for structured formats)
Standardisation & enhancement	Semantic enrichment, format conversion, linking to identifiers
Quality control	Manual review, training, metadata health, contributor workflows
Storage, indexing & search	Repository storage, indexing, search portal, export formats
Dissemination	API access, exports (i.e. DwCA, RDF) for reuse

Notifications based on literature should be compared against an existing database. That way updates and additions can be confirmed “new” before sending the notification to the user.

The major drawback of such a notification is **getting access to the scientific literature**. Only 20% of the biodiversity publications is published as open and available[1].

The problem of data not being available in a structured format can partly be solved by tools that recognize scientific names and relations in full texts.

A number of journals provide open access to a machine readable version of its papers. (i.e. ZooKeys, PhytoKeys).

Combining this with the other components above puts this combination very close to a fully automated literature notification system. But it is limited by only one publisher (Pensoft).

Some real-world examples of (combinations of) the above components:

- The WFO Plant List uses a semi automated approach that **pushes relevant publication citations into a shared Zotero library**. Within this library each Taxonomic Expert Network has their own folders that contain publications to be assessed
- A **semi/automatic import of new taxa from Pensoft** is available in Aphia. A demo video at https://www.marinespecies.org/photogallery.php?album=5456&pic=149305
- **Plazi TreatmentBank** (TB) is a service to liberate data from scholarly publications, and convert, enhance, link, store, and disseminate it as FAIR data (https://plazi.org/treatmentbank/)
- **BioRSS** feed of newly published taxa (https://biorss.herokuapp.com)
- **GROBID** is a machine learning software for extracting information from scholarly documents (https://github.com/kermitt2/grobid)
- **TaxonGrab** extracts taxonomic names from text (https://journals.ku.edu/jbi/article/view/17)
- **TaxoNERD** is a collection of deep neural models for the recognition of taxonomic entities in the ecological and evolutionary literature. (https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.13778)
- **LINNAEUS** is a species name recognition and normalization software (https://linnaeus.sourceforge.net/)
- **SPECIES** is a standalone command line application capable of identifying taxonomic mentions in documents and mapping them to corresponding NCBI Taxonomy database entries (https://species.hcmr.gr/)
- **Global Names Finder** finds scientific names in web pages, PDFs, Microsoft Office documents, images, or freeform text (https://finder.globalnames.org/)
- **Google Scholar** allows you to retrieve emails based on search keywords, i.e. groups, species names (https://scholar.google.com)
- And many more…

4 Conclusions

The preparation of this document has been a valuable exercise in collaboratively exploring the taxonomic notification service problem domain. Although apparently complex it is relatively straightforward to formalise the kinds of changes users may be interested in. There are a host of mature technologies available that could be used to implement notification services that could be used to communicate these changes to users. We do not need to invent anything new. Different audiences are likely to be interested in notification about different changes and they are also likely to require the use of different technologies. The next step is to explore potential services with real world users to establish which would be of genuine use to them. For

[1] Mandeville et al. 2021, 10.1093/biosci/biab072

example, implementing a service based on Mastodon and BlueSky would be of no use if the target audience were reluctant to adopt the use of these platforms. RSS is known to be of use to paranoid database curators to monitor changes being made by other users but would taxonomic experts, biologists or policy makers be willing to install RSS feed readers to be notified just about their taxonomic groups? These questions are probably only answerable by building pilot systems and letting users try them out.