TNLS: Common name matching specification

1 Background

The Taxonomic Name Linking Services enables TETTRIs (Transforming European Taxonomy through Training, Research and Innovations) participants to link their services with those of the major, global taxonomic datasets thus aligning multiple datasets to a common, global taxonomic backbone system. The first work area in the project is to develop a common name matching design pattern, that is applicable across multiple projects. This document gives the findings of task 1 of the project, developing a common name matching specification.

2 The Name Matching Problem

The process of combining biodiversity data from multiple sources currently starts with matching of the Latin name strings for the organisms used in each dataset (thus, we exclude any other types of names). Datasets often contain names that cannot be unambiguously matched or miss out some names entirely. The reasons for this are complex and outlined in Appendix 1. From experience comparing datasets between 10% and 20% of names will typically fail to match perfectly and may need some human interaction or acceptance of error. A formal study to quantify the failure rate hasn't been carried out partially because of the lack of standardisation around matching. With datasets of many thousands of names this is a major hurdle that has to be crossed every time data is combined in an analysis. The problem is exasperated when more than two datasets are used. The number of cross comparison tasks increase exponentially with the formula $x = \frac{n^{(n-1)}}{2}$. A study combining data from ten sources would require 45 pairwise comparisons, each with a potential 10% issue rate. Large scale, long term projects adopt a "taxonomic backbone" approach to avoid making multiple comparisons. In this approach a single, authoritative classification, sometimes curated by experts, is used as a standard against which others are compared. The number of matching tasks drops to n x 1.

Once matching is achieved the data can be tagged with unambiguous name IDs avoiding the need for name matching in the future. The resolution to a unique identifier of some kind is what differentiates name matching from search. This is expanded on in Appendix 2.

There is no single, global taxonomic backbone. Major global projects have developed their own bespoke lists and name matching services. This is the case even within Europe. The heterogeneity leads to three main issues:

1. It is complex for users to interact with more than one service because they have to understand the different conceptual and technical approaches taken.
2. It is not easy to build tools and infrastructures that work across suppliers.
3. It is complex to integrate data between systems.

These issues were highlighted by Islam, S. *et al* (2024) "Navigating taxonomic complexity: A use-case report on FAIR scientific name-matching service usage in ENVRI Research Infrastructures".

> *The paper underscores the importance of **standardised tools**, enhanced communication, training, collaboration and shared resources. Addressing these needs can facilitate more effective FAIR [Findability, Accessibility, Interoperability, and Reuse] implementation within the ENVRI [Environmental and Earth System Research Infrastructures] community and biodiversity research.*

> *This, in turn, will empower RIs [Research Infrastructures] to seamlessly integrate and leverage scientific names, unlocking the full potential of their data for research and policy implementation.*
>
> Islam, S. et al (2024) - our emphasis

This task within TNLS takes the first steps in developing a standardised mechanism across matching services that, we hope, will lead to a formally defined common API in the future. We took a pragmatic, two-pronged approach. We surveyed the existing name matching services looking for commonalities and distilled this down to a short list of common input and output fields available. This led to the proposed common terminology outlined below. This work was informed by a conceptual review of the name matching process as outlined in the appendices.

3 Matching Services Surveyed

Twenty different name matching services were surveyed. Further details of supported fields are given in an accompanying spreadsheet.

- **Algaebase** (https://www.algaebase.org/) A database of information on algae that includes terrestrial, marine and freshwater organisms.
- **Atlas of Living Australia** (https://www.ala.org.au/)
  The Atlas of Living Australia (ALA) is a collaborative, digital, open infrastructure that pulls together Australian biodiversity data from multiple sources, making it accessible and reusable.
- **Australian Plant Name Index** (https://www.anbg.gov.au/apni/)
  A tool for the botanical community that deals with plant names and their usage in the scientific literature, whether as a current name or synonym. APNI is maintained by the Australian National Botanic Gardens as part of its larger IBIS database
- **Catalogue of Life** (https://www.catalogueoflife.org/)
  Possibly the most complete authoritative list of the world's species - maintained by teams of global taxonomists. Currently containing 2.1M species and 5.3M names across the whole of biology.
- **Checklist Bank (**https://www.checklistbank.org/tools/name-match**)**
  A tool to support the publication and curation of checklists and to provide a platform for their consistent discovery, use and citation. Checklistbank is a joining project between GBIF & Catalogue of Life.
- **Euro+Med PlantBase** (https://europlusmed.org/)
  Euro+Med Plantbase covers all native and introduced vascular plants from Europe, the Mediterranean and the Caucasus.
- **GBIF Taxonomic Backbone** (https://www.gbif.org/tools/species-lookup)
  The Global Biodiversity Information Facility is an international network and data infrastructure funded by the world's governments and aimed at providing anyone, anywhere, open access to data about all types of life on Earth. Within GBIF, an uncurated working list of taxonomic names is being used to manage the taxonomy. The aim of GBIF is to replace this with a subset of CheckListBank.
- **Global Names Verifier** (https://verifier.globalnames.org/)
  GNV is a tool provided by the Global Names Architecture (GNA) is a system of web-services which helps people to register, find, index, check and organize biological scientific names and interconnect on-line information about species, mostly based on existing name lists
- **Index Fungorum** (https://www.indexfungorum.org/)
  A global fungal nomenclator coordinated and supported by the Index Fungorum Partnership, containing names of fungi (including yeasts, lichens, chromistan fungal analogues, protozoan fungal analogues and fossil forms) at all ranks.
- **Integrated Taxonomic Information System** (https://www.itis.gov/)
  ITIS is an authoritative taxonomic information on plants, animals, fungi, and microbes of North

America and the world. Run as a partnership of U.S., Canadian, and Mexican agencies (ITIS-North America); other organizations; and taxonomic specialists.

- **International Plant Name Index** (http://namematch.science.kew.org/)
  IPNI provides nomenclatural data (spelling, author, types and first place/date of publication) for the scientific names of vascular plants from family to infraspecific ranks. The nomenclatural data and name registration provided by IPNI is produced by a collaboration between The Royal Botanic Gardens, Kew; The Harvard University Herbaria and The Australian National Herbarium hosted by the Royal Botanic Gardens, Kew.
- **LifeWatch e-lab** (https://lifewatch.be/e-lab)
  LifeWatch Belgium is one of the eight national nodes composing LifeWatch ERIC, the e-Science European Infrastructure consortium providing data resources, web services and Virtual Research Environments (VRE) to biodiversity and ecosystem research. Within LifeWatch Belgium a user-friendly GUI or "e-lab" was developed to interrogate web services and enrich data files without the need for any technical/programming skills.
- **National Center for Biotechnology Information** (https://www.ncbi.nlm.nih.gov/)
  The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information. The taxonomic backbone contains the names and phylogenetic lineages of more than 160,000 organisms that have molecular data in the NCBI databases.
- **The Paleobiology Database** (http://paleobiodb.org/data1.1/taxa/single.json?name=)
  A public database of paleontological data maintained by an international non-governmental group of paleontologists.
- **PESI / eu-nomen** (https://www.eu-nomen.eu/portal/taxamatch.php)
  PESI provides the first all-taxa inventory for European species, consisting of: the European Register of Marine Species (ERMS), Fauna Europaea, Euro+Med PlantBase and Index Fungorum
- **ROpenSci taxize** (https://docs.ropensci.org/taxize/articles/datasources.html)
  Taxize is an R package that allows users to search over many taxonomic data sources for species names (scientific and common) and download up and downstream taxonomic hierarchical information.
- **Taxonomic Name Resolution Service** (https://tnrs.biendata.org/)
  The Taxonomic Name Resolution Service (TNRS) is a tool for standardization of plant scientific names. The TNRS corrects spelling errors and alternative spellings to a standard list of names, and converts out of date names (synonyms) to the current accepted name.
- **Tropicos** (https://legacy.tropicos.org/NameMatching.aspx)
  The Tropicos database links over 1.39M scientific names with over 7.4M specimens and over 2.4M digital images. The data include over 171K references from over 56.1K publications offered as a free service to the world's scientific community. Tropicos is hosted by Missouri Botanic Garden.
- **World Flora Online Plant List** (https://list.worldfloraonline.org/matching.php)
  WFO is the international initiative to achieve Target 1 of the Global Strategy for Plant Conservation and provides a global overview of the diversity of plant species.
- **World Register of Marine Species** (https://www.marinespecies.org/aphia.php?p=match)
  The aim of a World Register of Marine Species (WoRMS) is to provide an authoritative and comprehensive list of names of marine organisms, including information on synonymy. You can use the WoRMS Taxon Match Tool to automatically match your species list or taxon list with WoRMS.

Conclusion: we define a "standard name matching specification" as the availability of the following common input and output parameters.

4 Common Input Parameters

The accompanying spreadsheet gives a list of the 40 parameters and variations of parameters taken by the name matching services listed above. The table below gives our selection of just six parameters between these services along with suggested standard names and definitions.

Table 1 Suggested Input Parameters

Parameter name	Type	Definition	Notes
name	string	The full scientific name of the organism as defined by an appropriate nomenclatural code.	This **may** include the rank part of a name, the authority string and the year of publication where appropriate.
authorship (optional)	string	A string representation of the authors of the name.	If the client is able to pass this as a separate field.
year (optional)	integer	The year of publication of the name as an integer. The match must be published in this year if it is provided.	More relevant in zoology than in botany
rank (optional)	string (from enumeration)	The matching name must be at this rank or ranks if provided.	It is recommended the service provides a controlled vocabulary for this field.
Kingdom (optional)	string (from enumeration)	The matching name must be within this kingdom (e.g. Animalia, Plantae, Fungi, Protista, Archaea, and Bacteria).	Useful for data aggregators covering all forms of life. Limits matching to a single nomenclatural code and prevents cross code homonyms.
include_accepted (default or parameter)	boolean	If the name matched is considered a synonym by the data source then also return the accepted name.	Differentiates between pure nomenclators that are only providing name matching services and taxonomic services.

5 Common Output Fields

The accompanying spreadsheet gives a list of 60 different fields returned by the name matching services considered above. The table below gives our suggestions of a minimum set of common fields. Services are likely to provide many more than these core fields depending on their user requirements.

Table 2 Suggested Output Fields

Field name	Type	Definition	Notes
ID	string	A persistent, unique (within the scope of the database or globally) identifier for this name object and or taxon.	Services may provide standard ways to convert local IDs to global (e.g. prepend URL). Should not change when name spelling changes.
name	string	The full scientific name of the organism as defined by an appropriate nomenclatural code.	The correct way to cite the name according to the service provider.
name_html (optional)	string	The same as name but with the code mandated italics parts of the name delimited with <i> tags.	This is to allow client software to render the names correctly without needing to re-parse them.

authorship (optional)	string	A string representation of the authors of the name.	The correct author string for the name.
accepted (optional)	boolean	Whether the name is an accepted name of a taxon or not.	Can be ignored by pure nomenclator services.
placement (optional)	array	Where the name is placed in the classification.	How this is presented will depend on the service. It could be a single string path or a more complex structure.

6 Appendix 1: Why names are often ambiguous

Here we summarise a few of the mechanisms whereby scientific names can be ambiguous either because different strings of characters refer to the same published name or the same string of characters refers to a different published name.

6.1 Homonyms

Homonyms are names that are spelt the same but refer to different things. Under the codes of nomenclature one of the names will always have precedence over another but because there have not been universal name registries it has not been possible to prevent creation of duplicate names. In the strict sense homonym means the full name, including the author(s) names, are identical. Homonym is often used in a looser sense of just applying to the words that make up the name, excluding the author string. This is because author strings are often not standardised or omitted entirely. Homonyms may occur within or between codes, that is the same name string may be used for two plants or for a plant and an animal. Furthermore, there are two types of homonyms:

6.1.1 Isonyms

Isonyms occur when a name is based on the same type specimen but published in multiple places. The majority of isonyms are created by the author publishing the name again (perhaps in a paper and in a flora, fauna or catalogue) and so have the same author(s). There is no scope for taxonomic confusion in botany and the only scope for nomenclatural confusion caused by isonyms is citing the wrong reference as a place of original publication. In zoology the name string may have different dates thus causing matching failures even though the intent of the author(s) was to name the same taxon.

6.1.2 True Homonyms

True homonyms are names based on different type specimens and, usually, published by different authors. If they are published by different authors (homonym in the loose sense) and the author(s) names are included in the full name then they should not be ambiguous during matching however author(s) names may be omitted, causing false matches, or use nonstandard forms, causing false mismatches. When a species is placed in a different genus from where it was originally published the authors of this new combination are cited by convention as part of the name in botany but not in zoology. This leads to a slightly greater theoretical chance of causing homonyms in zoology than botany.

6.1.3 Author(s) String variation

Most ambiguity caused by identically spelt scientific names can be resolved if the author(s) of the name are included in the full name. Unfortunately, this is fraught with difficulties in real data.

1. The author(s) are frequently omitted. If material is created for a general audience then inclusion of author(s) names can be considered confusing especially if the scientific (Latin) form of the name is being used in addition to a well-known vernacular name. When data is being shared within a specialist scientific community who only work on a few species then the author(s) are omitted

because there is no chance of confusion in that particular research context. Omission has also been influenced by legacy systems having restrictions on the length of data fields and restricted character encoding.

2. The zoological code of nomenclature does not consider the author(s) to be part of the name and inclusion is only customary although usually advisable. It is recommended on first use in a publication. Zoologists usually do not include the names of authors of new combinations (species placed in different genera). Botanists are more consistent in use of authors but do not include the year of publication of a name which is customary in zoology.
3. Standard author abbreviations are not mandated in either botany or zoology although in botany there is a more established convention to use the author abbreviations as maintained by IPNI and also community curated in Wikidata property P428. Publication editors will sometimes mandate changes to, or further abbreviation of, author strings, for example the use of ‘et. al.’ when there are more than two authors and the addition of spaces after periods.
4. Use of ‘ex’ is inconsistently applied and error prone. The nomenclatural code allows the author(s) who validly publish a name to acknowledge previous author(s) who published the name incorrectly by including the original authors' names in the citation. When they do this the two sets of author(s) names are separated by 'ex'. In botany the original authors come before the ‘ex’ but in zoology the two sets of authors are presented the other way around. Citing the original authors on subsequent use of the name is considered optional in both codes. Often there is confusion as to which set of authors to leave out resulting in multiple combinations of author strings for a single name being in circulation, some legal and some illegal. Some of those versions will, by chance, match unrelated names.
5. Encoding issues. Prior to widespread adoption of UTF-8. Author names may include accented characters but taxon names should only include common characters from the original ASCII code page.

In summary: Even within the scope of vascular plants covered by IPNI, where it should be possible to follow a standard, there are always errors in trying to match full name strings that include the author(s) and therefore differentiating potential homonyms.

6.2 Orthographic variants

The way the names are spelled in the original publication is not fixed. Nomenclatural rules have been introduced to mandate particular spelling changes. In the botanical code, article 60.7. specifies that diacritical signs (such as ö) and ligatures (such as œ) cannot be used in names and that where they occur in existing names they should be replaced by a combination of characters. Some of these can be handled automatically through the matching process because they are specified in the code. The botanical congress in 2024 voted to change the spelling of names containing the offensive "caffra" to "afra" and initiate a process to consider changes to other names that cause offence. There are various rules in the botanical code specifying how the ending of a name should agree with the gender of the genus and how the gender of the genus should be established. These can result in changes of spelling through time. Because the rules are retrospective and all versions of the spelling of the name will continue to be in circulation from historical data even if there is a prescribed version that should be used at the moment, matching algorithms can never rely on stable spelling of all names.

6.3 Errors

Only at the end do we get to stochastic or unforced errors. These may occur from the following main sources:

1. **Human error:** People simply make a spelling mistake when writing a name, perhaps by misremembering it.
2. **Transcription errors:**

1. **Human:** Handwriting, particularly on specimens can be difficult to read and is often transcribed wrongly.
2. **OCR:** Although Optical Character Recognition (OCR) is now very accurate on typed text, it is still error prone on hand writing and much of the data in circulation is the result of OCR using early software versions which weren't as accurate as those available today.

3. **Generative AI Hallucination:** An increasing issue is that generative machine learning algorithms will create new names that look like scientific names. This has been observed with prompts requesting the scientific name for a vernacular name as well as in translations between languages where an algorithm, that is not ostensibly a generative AI, will create a new scientific name from a vernacular name in the source language.

All errors have the potential to create zombie names but generative AI is perhaps the most concerning source going forward.

6.4 Zombie Names

Zombie names are name strings that may have occurred in the literature or a database via the errors outlined above or may have been legitimately published for a long-discredited classification. They have subsequently been propagated from one dataset to the next without ever being deleted. They soak up time and resources because each time datasets are combined the process has to resolve zombie names to an original place of publication starts again. Zombie names are particularly problematic in the age of big data. If we delete them they will keep coming back again from different data sources and each time they are rediscovered they will use up more resources.

The only way to deal with zombie names is to have a central register that includes them and flags them as not being resolvable. This is the approach World Flora Online has taken with the notion of deprecated names. Because the names are tracked and have IDs similar to other names they can be resurrected in the unlikely event they are discovered to have been correctly published.

7 Appendix 2: Matching vs Searching

The processes of matching and searching are often confused. It is worth defining what is meant by "name matching" here.

In general parlance matching means one person or thing resembles or corresponds to another. In computing an exact match is usually implied, for example, all the lines in a file that contain a certain string match the query. Both senses imply a comparison.

In taxonomic name matching one of the two things compared is a string of characters representing the name of an organism and the other is a list of known names of organisms. The intent is to go from a string of characters on the page to a protolog and the type specimen that anchors the name into a classification. To achieve this the matching process must only return a single result. If it returns multiple results it is searching the list for potential matches but not actually matching the name string.

Table 3 Comparison of matching and searching

Matching	Searching	Explanation
Normative	Informative	A normative response is prescriptive. It gives an answer on how to comply with a standard. An informative response provides useful or interesting

		information to help to understand what the name means.
Definitive	Indefinite	A matching process will return the value from a controlled vocabulary that has been specified *a priori*. Searching returns a result (or results) and the domain can be vaguely defined.
One or zero returned	zero to many returned	A match is only expected to return a single result (usually an ID) or to fail. A search may result in many candidates answers that are of interest to a user.
Correct or Incorrect	Relevant or not	If a matching process succeeds the result is either correct or incorrect. Searching produces results with a degree of relevance (depending on the algorithm used).
Targeted at machines	Targeted at humans	The result of a matching process is typically applicable to machine interpretation, e.g. linking resources. Search results are usually consumed by humans.

7.1 Examples of Matching

1. Looking up a telephone number for a name in an old-fashioned telephone book is an example of matching. The domain is the city the book covers. The answer will be the number prescribed by the telephone company for that name. It will be found or not found. It could be the wrong number but it couldn't be only partially correct. The number is for consumption by a machine, the telephone, not a human.
2. Finding the DOI (Digital Object Identifier) for a publication based on a written citation in APA (American Psychological Association) format. The publication will only have a single DOI which is authoritative and controlled by the DOI Foundation. The DOI is used to find the official metadata and full text of the publication and of no interest to a human in itself.

7.2 Examples of Searching

1. A Google search of the internet. Many results are typically returned from the provided search terms ranked by relevance.
2. Booking a flight route from New York to Paris via a flight aggregator.
3. Using BLAST to find regions of similarity between nucleotide or protein sequence and those in a database.

7.3 Areas of Confusion

It is possible to look up a DOI for a publication by searching for the authors and title on Google. It is very likely that the DOI of the paper will be in the first few results returned. This feels like matching but requires

a human to pick the most appropriate result. Google is helping a human do the matching.
Most matching algorithms will fall back to searching if an unambiguous match is not found. When presented with a list of names an algorithm might automatically match the majority of them, perhaps using a degree of "fuzzy matching" as Taxamatch does, but if confidence falls below a predefined threshold a list is presented for a human to pick a preferred result if possible. At this point the matching algorithm has done a search for candidates and the human is actually doing the matching, not the machine.