Annotations and identifiers in biodiversity publishing

This is the combined ms with BiCIKL WP T6.3 identifiers task
Original: https://docs.google.com/document/d/1Zq_1RoTn2oYbO-cWhgRAzNpOWxz3cyHi/edit
BiCIKL sources: https://docs.google.com/document/d/1VR54ecDppm8LcVr-eaER3nFdfsTE-uGNYPSW3Vkt_6s/edit#

New table that will serve as appendix: https://docs.google.com/spreadsheets/d/1n_-v8Ty9VO6Cg4mUL794K37I1QC0PJxs/edit#gid=1287527644

Dissco
https://docs.google.com/document/d/1qZ8D5T7TQjzscw0YTUn-h-kI-mG4MFykEVwLMbwcd9k/edit

Also check the DOI use for content elements in DiSSCo (see attached) and EOSC. (https://ec.europa.eu/info/sites/default/files/research_and_innovation/ki0420576enn-updt.pdf)

To read:
https://dissco.tech/2020/04/11/identifiers-for-our-institutes-grid-and-ror/
https://cetafidentifiers.biowikifarm.net/wiki/Main_Page

Target audience:
Publishers, authors, researcher

introduce each element as:

- Definition (what is e.g. a serial publication)
- What are its identifiers
- How to discover them
- How to mint an identifier
- How to annotate and cite them
- Examples
- Recommendation

Table: priorities (list only the elements that we recommend?

Commented overview table

MS starts here

Recommendations for use of annotations and persistent identifiers in taxonomy and biodiversity publishing

Authors

Donat Agosti (1), Laurence Bénichou (2), Wouter Addink (16), Christos Arvanitidis (17), Terence Catapano (13), Guy Cochrane (18), Mathias Dillen (3), Markus Döring (14), Teodor Georgiev (4), Isabelle Gérard (5), Quentin Groom (15), Puneet Kishor (19), Andreas Kroh (7), Jiří Kvaček (6), Patricia Mergen (8), Daniel Mietchen (9), Joana Pauperio (10), Guido Sautter (12), Lyubomir Penev (11)

1. Plazi, Bern, Switzerland; ORCID: 0000-0001-9286-1200
2. Muséum national d'histoire naturelle, Paris, France; ORCID: 0000-0002-0713-0751
3. Meise Botanic Garden, Meise, Belgium; ORCID: 0000-0002-3973-1252
4. Pensoft Publishers, Sofia, Bulgaria; ORCID: 0000-0001-8558-6845
5. Royal Museum for Central Africa, Tervuren, Belgium; ORCID: 0000-0003-4375-3750
6. National Museum, Prague, Czechia; ORCID: 0000-0003-2001-121X
7. Natural History Museum, Vienna, Austria; ORCID: 0000-0002-8566-8848
8. Royal Museum for Central Africa, Tervuren, Belgium and Meise Botanic Garden, Meise, Belgium; ORCID: 0000-0003-2848-8231
9. Leibniz Institute of Freshwater Ecology and inland Fisheries, Berlin, Germany; ORCID: 0000-0001-9488-1870
10. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, CB10 1SD, United Kingdom; ORCID: 0000-0003-2569-0768
11. Pensoft Publishers and Institute for Biodiversity and Ecosystem Research, Bulgarian Academy of Sciences, Sofia, Bulgaria; ORCID: 0000-0002-2186-5033
12. Plazi, Bern Switzerland; ORCID: 0000-0002-6073-3658
13. Plazi, Bern, Switzerland; ORCID: 0000-0002-6857-0021
14. Global Biodiversity Information Facility, Copenhagen, Denmark; ORCID: 0000-0001-7757-1889
15. Meise Botanic Garden, Meise, Belgium. ORCID: 0000-0002-0596-5376
16. Naturalis Biodiversity Centre, Netherlands. ORCID: 0000-0002-3090-1761
17. LifeWatch ERIC, Seville, Spain; ORCID: 0000-0002-6924-5255
18. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, CB10 1SD, United Kingdom; ORCID: 0000-0001-7954-7057
19 Plazi, Bern, Switzerland; ORCID: 0000-0002-1746-6195

Keywords: semantic publishing, taxonomy publishing, semantic annotation, biodiversity, persistent identifiers, taxa, specimens, sequences, treatments, XML, JATS, TaxPub, tagging

Abstract

The paper summarises many years of discussions and experience of biodiversity publishers, organisations, research projects and individual researchers, and proposes recommendations for implementation of persistent identifiers for article metadata, structural elements (sections, subsections, figures, tables, references, supplementary materials and others) and data specific to biodiversity (taxonomic treatments, treatment citations, taxon names, material citations, gene sequences, natural history collections) in taxonomy and biodiversity publishing. The paper proposes best practices on how identifiers should be used in the different cases and on how they can be minted, cited, and expressed in the backend article XML to facilitate conversion to and further re-use of the article content as FAIR data. The paper also discusses several specific routes for post-publication re-use of semantically enhanced content through large biodiversity data aggregators such as the Global Biodiversity Information Facility (GBIF), the International Nucleotide Sequence Database Collaboration (INSDC) and others, and proposes specifications of both identifiers and XML tags to be used for that purpose. A summary table provides an account and overview of the recommendations. The guidelines are supported with examples from the existing publishing practices.

Introduction

Specifics of taxonomic publications

Deans et al (2012) very elegantly stated that “Taxonomists are arguably the most active annotators of the natural world, collecting and publishing millions of phenotype data annually through descriptions of new taxa. By formalising these data, preferably as they are collected, taxonomists stand to contribute a data set with research potential that rivals or even surpasses genomics”.

Taxonomic publications communicate the discovery of new biological taxa or new data on already known taxa in the form of taxonomic treatments, well delimited sections of text for each taxon (Fig. 1; Catapano 2010, Penev et al. 2011, Agosti & Egloff 2021). New research results are added to the already existing treatments by citing previous treatments using a “treatment citation”. Altogether, the treatments and data related to them represent the basis for the knowledge graph on the Earth’s biological diversity. Treatments have been used from the beginning of modern taxonomy by Linnaeus in 1753 for plants and in 1758 for animals. Treatments begin with a nomenclature section including a unique identifier for the taxonomic name, the Latin Binomen for species or Latin Name for a supraspecific taxon such as genus, family or order. This is followed by one or more sections covering the citation of previous treatments of the same taxon, description, diagnosis, etymology, distribution, material citations or conservation. New taxa are based on type and other specimens in natural history collections and data on these specimens are included in the treatment in the form of dedicated “material citations”. This new style of presenting information on biological taxa required a certain degree of comprehension and adoption but was widely accepted by the taxonomists in the second half of 18th century.

Apis mellifera Linnaeus, 1758

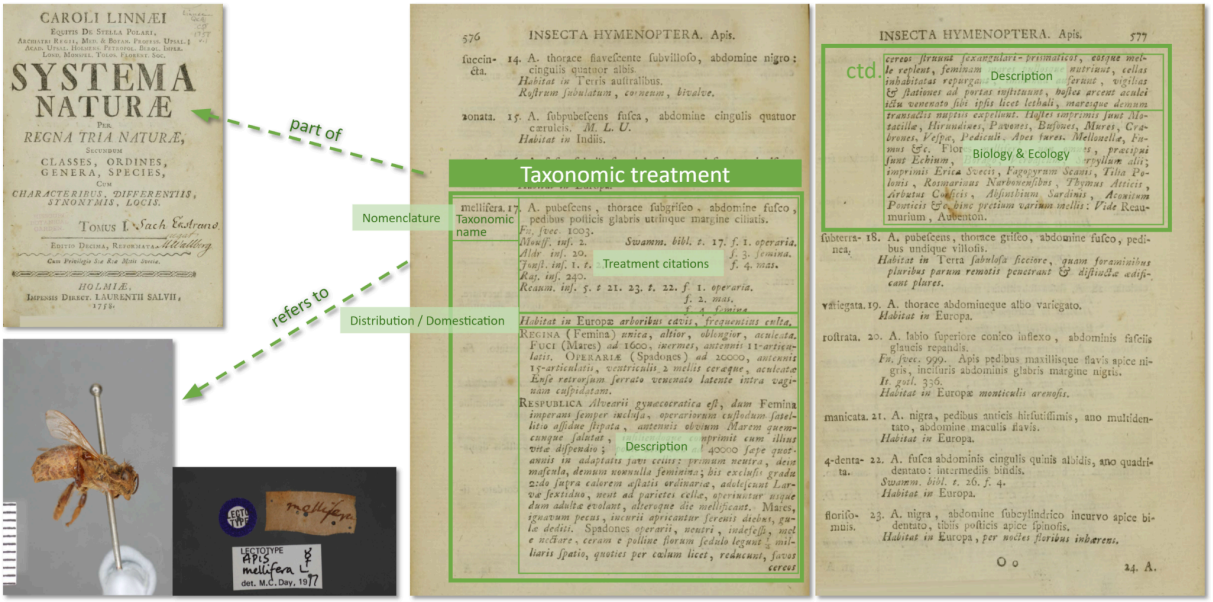


Fig. 1 Schematic representation of the taxonomic treatment of *Apis mellifera* Linnaeus, 1758. Sources: text: https://doi.org/10.5962/bhl.title.542; figures: https://linnean-online.org/16019/: composite: https://doi.org/10.5281/zenodo.5168465

Translated into today's digital world, this simple framework of presenting biological taxa in both human readable and machine actionable format is sufficient, given that it is present as digital accessible knowledge (DAK, Fawcett *et al.* 2022), to build a knowledge graph of the Earth's biological diversity. By "machine actionable" we mean that the data are structured systematically so that computers can be programmed to process and interpret the data. This requires that the elements *taxonomic treatment*, *taxonomic name*, *treatment citation*, *material citation* and other important terms of relevance are annotated in publications following a community accepted standard, and are made citable through inclusion of the respective identifiers of the cited elements (e.g., treatment in *treatment citations*, *taxonomic name*, specimens or digital specimen for the *material citations*). Thus to explore known biodiversity, this is the minimal degree of digital accessible knowledge needed to allow us to ask questions such as "What do I know about taxon X?", "What are the synonyms of a taxonomic name", and "What are the facts used to make the changes?".

Research results presented in the biodiversity literature are one of the best curated data (Deans *et al.* 2012) providing expert linking of taxonomic names, molecular, including omics data, phenomics data, specimens, geographical, environmental and climatic data, taxonomies and phylogenies, previously published data, publications and people via accession codes, material citations, treatment citations, bibliographic references or personal identifiers, respectively. The semantic annotation or semantic role labelling (e.g. Walls *et al.* 2014) of texts, provides an additional feature for identifying the role of people and taxonomic names. For example, a person mentioned in a material citation can be inferred to be a collector, whereas a person's name in a taxonomic name indicates the role of authority of the taxon's name, and an author of the publication in which the taxonomic name has been published. A taxonomic name in the nomenclature section functions as a label for the treatment, while a taxonomic name in the treatment body outside the nomenclature section indicates some sort of connection between the two taxa.

In today's digital arena these structured texts are an ideal prerequisite to enhance the publications by making them machine actionable (Chester et al. 2019). This includes making, for example, treatments

and figures open, findable, accessible, interoperable and reusable (FAIR) digital objects[1], then adding identifiers to the cited materials, gene sequences, and authors, and annotating them to add a semantic meaning to those tokens. The use of persistent identifiers is intended for many purposes, including building a knowledge graph, understanding the use of specimens and their collections in research, to give credit to individual scientists and institutions, and more broadly to allow reuse by aggregators, such as the Global Biodiversity Information Facility (GBIF) or ChecklistBank. Persistent identifiers also contribute to mitigating the taxonomic impediment recognized by conservation policy (Abrahamse *et al.* 2021), create new knowledge management systems, and bridge gaps between different domains such as taxonomy, ecology and molecular biology in the life sciences. The first working examples of knowledge graphs in the biodiversity realm are OpenBiodiv (Senderov et al. 2018, Penev et al. 2019, Dimitrova et al. 2021), Ozymandias (Page, 2019) and Synospecies (Gmür & Agosti 2021),

Use of identifiers

An identifier (ID) is a label for any subject, conceptual, physical or digital (Dillen *et al.* 2021). An ID can be called persistent (PID) (European Commission 2020a) if it can be maintained as a label in the longer term, in spite of any changes to the subject itself. For example, IDs for people can be persistent even if their name(s) change or they move to another location or change jobs. Hence**, an ID aims to disambiguate the entity it relates to.** To be a PID, it also needs to be Globally Unique, Persistent and Resolvable (GUPRI, European Commission 2020a). It should thus be unique at the context in which it is used and come with a system that maintains the link between the ID and its subject. For example, in the case of resolvable Uniform Resource Identifiers (URIs), this system is the internet's Domain Name System (Hyam *et al.* 2012). However, it is still up to the organisation that mints the URI to ensure it remains persistent, as Domain Names are not. Digital Object Identifiers (DOIs) are another example, making use of the Handle system to maintain the link between ID and subject. Nevertheless, the onus still remains on the organisation holding the digital object to ensure that the DOI resolves to the right object.

As an identifier serves to unequivocally label an entity, it may also be employed to track the use of it, particularly when that entity is digital. Performance indicators are an important tool for the efficient management and development of organisations and infrastructures. Such indicators are used to channel appropriate funding internally, and also to request funding externally. The more we are able to show the impact and reach of our field, the easier it is to gather financial support to develop natural history collections, maintain services, digitise objects and conduct research.

Parallel to the rise of e-publishing, IDs minted and used in biodiversity informatics have diversified and become increasingly important to link to objects, specimens, and their digital representation, as well as the component parts of literature (Guralnick et al. 2015; McMurry et al. 2017; Page 2016, 2019; Madden & Woodburn 2021). They form a scaffold on which to form a biodiversity knowledge graph.

Usage of identifiers can be broad and complex. PIDs are used to identify and link digital and physical objects or concepts. One of the very first uses of DOIs was identifying individually published articles as well as references in the bibliographies, which enhanced the visibility and citability of these articles. DOIs are also used for data and figures, and are proposed for the digital objects in DiSSCo (Hardisty *et al*., 2022). For physical specimens in natural history collections there are the persistent HTTP URIs proposed and implemented by CETAF (Güntsch *et al.* 2017) and in Zenodo as the DataCite "physical object" (Boschert & Dikow, 2022). Likewise, Life Science Identifiers (LSID) were once used in

[1] https://www.go-fair.org/fair-principles/

biodiversity informatics *inter alia* for concepts (taxonomic names), while ORCID or Wikidata identifiers serve as identifiers for people. A PID is needed for any digital object posted on the Web so it may be easily found, cited, linked, annotated, and reused. Furthermore, in publications, PIDs and their respective metadata can be provided for many types of research-related content such as journals, chapters, grants or funders, datasets, data, text, and images (Guralnik *et al.* 2015). An emerging consensus for PIDs is the current development in DiSSCo infrastructure and the BiCIKL project to use DOIs as community-agreed, unified identifiers for curation of a digital specimen. Digital specimens will be treated as a Federated Digital Objects (FDO), that is, as an aggregator of several existing identifiers of data related to a specimen, such as the identifier for the physical specimens itself, IDs of material citations of the specimen published in the literature, IDs of gene sequences from the specimen (INSDC accession codes) and others (Hardisty *et al.* 2021).

The Internet's world wide coverage clearly makes it evident that globally unique identifiers are a prerequisite to locate the cited resources, and, consequently, through conversion and transformation of data, to build a knowledge graph, where all these resources can be identified and linked to each other through their PIDs. Since digitization of objects (e.g. an article) can occur in parallel, this can lead to collision between identifiers for physical objects, or across domains, between articles or specimens. Identifiers of different kinds have a long tradition in biodiversity research — they served specific purposes such as to label specimens from an expedition or a natural history collection, and have been understandable and resolvable within their respective context. At the same time, most of these "internal" identifiers are in formats that are not easily recognisable or interpreted by non-specialists or machines, and thus assigning a unique PID to each digital object is necessary. However, this will require look-up tables linking historic identifiers with the respective PIDs or extending non-unique IDs with a prefix to make it unique. Ideally the connection between the legacy ID and the unique PID is made either at the metadata level of each object, or within the specimen record (material citation) in publications.

Because of the current transitional period of digitising biodiversity data, new and different kinds of PIDs might be minted for the same object. To connect different PIDs for the same object we will need a discovery mechanism to build look-up tables. The different data accessible via the resolution of the PIDs will then provide complementary, sometimes conflicting data about the same objects (such as is discovered by GBIF's clustering mechanism[2] for seemingly similar occurrences) and thus increase the knowledge about an object.

To minimise the costs of the significant and non-trivial effort of disambiguation of entities and building and maintaining look-up tables, the recommendations in this paper strongly encourage the use of harmonised PIDs that are compliant with a community accepted standard across different journals and publishers and serve, therefore, multiple scientific disciplines or domains. A good basis for harmonisation, for example, are the recommendations of the European Open Science Cloud (EOSC) for the use of PIDs that should be taken into account (European Commission 2020a; European Commission 2020b).

On the need of harmonisation

The recommendations in this paper are produced collaboratively by several organisations, research projects and biodiversity scientists. They are based on nearly 15 years of experience on annotating unstructured legacy publications by Plazi (Agosti & Egloff 2009), and on TaxPub XML-based structured

[2] https://www.gbif.org/news/4U1dz8LygQvqIywiRIRpAU/new-data-clustering-feature-aims-to-improve-data-quality-and-reveal-cross-dataset-connections

publishing by Pensoft, including 38 journals since 2010 (Penev *et al.* 2010 [3]). Furthermore, during several EU-funded projects such as pro-iBiosphere, EU BON, and COST Mobilise, the focus of discussions was on building an infrastructure to provide FAIR data, for example, the Biodiversity Literature Repository (BLR) as well as on the implementation of persistent identifiers in article XMLs of Plazi and Pensoft (Catapano 2010, Penev *et al.* 2010, 2011). Finally, part of this discussion was carried out in the CETAF e-publishing group's ongoing work on unique identifiers.

The paper has been largely elaborated and finalised in a collaboration between several partners in the Biodiversity Community Integrated Knowledge Library project (BiCIKL) (Penev *et al. 2022*). In a similar fashion to the harmonisation of PIDs that the Research Organisation Registry (ROR), Datacite, Crossref and ORCID have agreed (Demeranville *et al.* 2021). This has reinforced the use of their PIDs in the scientific community, and has been the foundation for disambiguation and interlinking of institutional and biographical data, article metadata and datasets.

Taxonomy is ruled by nomenclatural codes which state the requirements for a nomenclatural act to be validly published, whether in print or online. These rules have evolved with the emergence of online journals, and mandate the use of certain identifiers within the publication and especially in the full-text XML of articles, for example the LSID of the publication in which a new nomenclature act is published, or the mention of the ISSN for the journal (see Penev *et al.*, 2016 and Bénichou *et al.* 2018). Hence, as a consequence of this main mandate, we outline the use of structured data and their identifiers to allow machines to assess whether a new taxonomic name is available according to the Codes.

The objective of this paper is to list the main structural elements and data types present in taxonomic publications, the existing identifiers currently in use, and make proposals for use of additional PIDs where these do not exist yet. The paper aims at providing both recommendations, best practices and practical advice to technical editors and publishers in taxonomy on how to implement identifiers in their work and how they can be leveraged. For each element, the use of an identifier is discussed from the perspective of taxonomic publishing, its pros and cons are given, and short explanations of how and where to implement these PIDs. We recommend that authors and publishers provide as many identifiers and links as possible, facilitating in this way the conversion of the published content into a digitally accessible knowledge. This would be not only a starting point for the reuse of this important data at scale, but also spur new research based on this incredibly rich resource. It will also allow linking data in taxonomy with other scientific disciplines to build the future practice of evidence-based knowledge, that is to bridge the gap from a taxonomic name to machine actionable data about it.

Publications, publication sections, sub-article data elements and their identifiers

Modern taxonomic articles follow a rather strict structure that facilitates their representation in a structured XML format following the widely used TaxPub[4] schema and enabling efficient data exchange (Catapano 2010; Penev et al. 2012). Based on the Journal Article Tag Set (JATS) standard[5], a journal article is composed of up to three optional parts, which should appear in the following order: Front matter is required while body and back are optional all in that order.

[3] https://tb.plazi.org/GgServer/dioStats/stats?outputFields=doc.uploadUser+bib.source&groupingFields=doc.uploadUser+bib.source&FP-doc.uploadUser=%25ensoft%25&format=HTML

[4] https://github.com/plazi/TaxPub/

[5] https://jats.nlm.nih.gov/publishing/

In its broader sense, publishing is the act of making content available to the public. In this paper, we refer specifically to peer-reviewed publications, either in the form of monographs (books) or periodicals (journals), print or electronic. While not all taxonomic publications are peer-reviewed, most of the comments and recommendations made here would apply to them too. Publishing taxonomic content, and specifically publishing nomenclatural acts, has a specific meaning and requires compliance to the rules, which are defined in various codes of nomenclature.

For Zoology, the *International Code for Zoological Nomenclature* (ICZN)[6] defines a publication in Article 8:

publication, n.
(1) Any published work. (2) The issuing of a work conforming to Articles 8 and 9.
electronic publication
A publication issued and distributed by means of electronic signals.
publish, v.
(1) To issue any publication. (2) To issue a work that conforms to Article 8 and is not excluded by the provisions of Article 9. (3) To make public in a work, conforming to (2) above, any names or nomenclatural acts or information affecting nomenclature.

In botany and mycology, the International Code of Nomenclature for algae, fungi, and plants (ICN)[7] defines a publication in its Article 29:

Publication is effected, under this Code, by distribution of printed matter (through sale, exchange, or gift) to the general public or at least to scientific institutions with generally accessible libraries. Publication is also effected by distribution on or after 1 January 2012 of electronic material in Portable Document Format (PDF; see also Art. 29.3 and Rec. 29A.1) in an online publication with an International Standard Serial Number (ISSN) or an International Standard Book Number (ISBN).

Front matter of the publication

Definition

The article's front matter contains metadata for the article and its host journal: title, authors' list with their affiliation, the date of publication, abstracts, keywords, a copyright statement, etc (see front matter structures in JATS XML here[8]). These frontmatter components should be encoded with JATS XML elements, the following with PIDs included: the article itself, the journal in which it is published, and the authors (see the section on Person names below).

Persistent unique identifiers for the front matter

International Standard Serial Number (ISSN)

Definition

The ISSN is an 8-digit number used to uniquely identify a serial publication. The system was designed in 1971, then published as a standard in 1975, and can be used for a journal as well as for book series,

[6] https://www.iczn.org/the-code/the-code-online/
[7] https://www.iapt-taxon.org/nomen/main.php
[8] https://jats.nlm.nih.gov/publishing/tag-library/1.1d1/n-me42.html

and even for some websites in the scholarly domain. It is unique and designates the publication medium, for instance if a journal is published in both print and digitally it must have a different ISSN for each media: a Print-ISSN and an E-ISSN (a different ISSN should also be given for any mobile version or CD-Rom version). One also needs an ISSN in case of a different language version of the same journal. When the publication is provided in different media, it is recommended to display all ISSN numbers on each version of the publication, if the latter is published, e.g. in different languages in different journals. The ISSN does not offer any resolution mechanism and is only a media-oriented identification.

Why does a journal need an ISSN?

The ISSN is mandatory for any journal or serial publication. In taxonomy, to be compliant with most nomenclatural codes, the nomenclatural acts should be published in a journal or series identified with an ISSN or a book with an ISBN.

Indeed, the International Code of Nomenclature for algae, fungi, and plants (ICN) stipulates that a nomenclatural novelty to be considered effectively published, it should be present in a publication distributed either in print (through sale, exchange, or gift) to the general public or at least to scientific institutions with generally accessible libraries, or (after 1 January 2012) in an online publication with an International Standard Serial Number (ISSN) or an International Standard Book Number (ISBN) and in Portable Document Format (PDF) (see also Art. 29.3 and Rec. 29A.1).

The *International Code of Zoological Nomenclature* (ICZN) stipulates in Art. 8.5 (ICZN 2012) that for an e-publication to be considered published in the terms of the Code, a work issued and distributed electronically must:

- have been issued after 2011
- have the date of publication stated in the work itself
- be registered in Zoobank (see below)
- contain the evidence of such registration (LSID of the publication or of the new name must be indicated in the work itself).

In Zoobank, the entry must have the name of an organisation other than the publisher that intends to permanently archive the work in a manner that preserves the content and layout, and is capable of doing so. The ISSN or ISBN of the publication must be registered in the Zoobank entry.

How to discover an existing ISSN

To find the ISSN of a series or journals, one may consult the ISSN portal, which provides a comprehensive list of all ISSNs and some associated metadata.

How to obtain an ISSN

To get an ISSN for a journal or series, all the necessary information, is available at https://portal.issn.org/requesting-issn. In some countries the ISSN might not be free and may require a registration fee between 25 € and 50 €, depending on the country assigning the ISSN.
It seems possible to obtain an ISSN before the first publication of a print serial, however, it is very common to be asked to wait until number 2 of the series to be printed. Online publications are usually assigned an ISSN after the first or second issue is published (with at least 5 publications published), or in some countries, after the website of the new periodical has gone live and is fully functional.

How to annotate and display an ISSN

Display: ISSN: 1313-2989 (for the print version of a serial, if exists)
e-ISSN: 1313-2970 (for the online version of a serial, if exists)

For linked data purposes, it is even better to use the resolvable version of the ISSN: https://portal.issn.org/resource/ISSN/1313-2989

Annotate in JATS:

Tag the ISSN number using the <issn> element, using the *publication-format* attribute to specify the format or medium of the publication (e..g, "print", "electronic", "video", "audio", "ebook", and "online-only")[9]

```
<issn publication-format="ppub">[ISSN number]</issn>
<issn publication-format="epub">[ISSN number]</issn>
```

Example of an ISSN

2118-9773 is the ISSN of the *European Journal of Taxonomy* (EJT). As the journal is an e-only journal, it has only one online or e-ISSN.
ZooKeys has two ISSN depending on its version: 1313-2989 (for the print version) 1313-2970 (for the online version).
An ISSN can also identify a series of books, e.g. The *Mémoires du Muséum national d'histoire naturelle* have one for the print version (ISSN: 1243-4442) and one for the online version (e-ISSN: 1768-305X).

Sample annotation in JATS (for ZooKeys):

```
<journal-meta>
      <issn publication-format="ppub">1313-2989</issn>
      <issn publication-format="epub">1313-2970</issn>
</journal-meta>
```

Recommendation

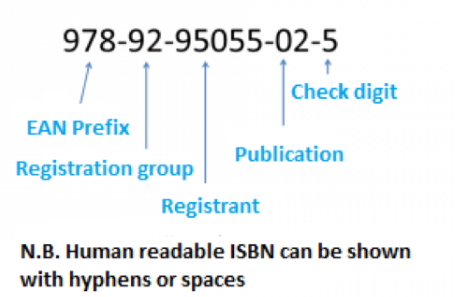
Considering that ISSN (or ISBN) are mandatory for online publication in taxonomy to be compliant to both ICN and ICZN codes, and that an ISSN makes your journals or series more easily identifiable and findable, attributions of an ISSN or ISBN to taxonomic publications must be considered mandatory. A unique ISSN should be assigned to each version of the journal, print and electronic. Each linguistic version of the journal should also have its own ISSN.

International Standard Book Number (ISBN)

Definition

The ISBN was internationally approved as an ISO standard in 1970, and published in 1972, and is a unique international identifier for monographic publications. Correct use of the ISBN allows different product forms and editions of a book, whether printed or digital, to be clearly differentiated, ensuring that it identifies the specific version it relates to. Similarly to the ISSN, each version of the book, print, e-book, pdf etc., must have a different ISBN. A book included in a book series, or published as a monograph in a journal, can be provided with both an ISBN and the ISSN of the series in which it is published.
ISBN is a 13-digit number that identifies a book. As it is typically used in a barcode format, it is prefixed by an European Article Number (EAN). It is constructed as follows:

[9] https://jats.nlm.nih.gov/publishing/tag-library/1.3/element/issn.html



Why does a book need an ISBN?

ISBN is important for cataloguing a book and for its findability, discovery, and dissemination. Its display is obligatory in the first pages of the book, along with the book title, author(s) name(s) and the publisher. ISBN is the main international record of your publication and is important for indexing and dissemination. It aims at facilitating the compilation of book trade directories and bibliographic databases, which in turn facilitate their dissemination as book dealers can use them to order books efficiently and unambiguously.

In taxonomy, it is crucial to have ISBN assigned to any taxonomic monograph with nomenclatural acts. For instance, as explained above, Zoobank requires an ISBN to register a nomenclatural act published within a book (ICZN 2012, art. 8.5). It is also mentioned in the ICN art. 29.3 as an alternative to ISSN when the nomenclatural novelty is published in an electronic book.

How to discover an existing ISBN

As a unique identifier, ISBN is part of the metadata associated with any book. To find the ISBN of any published book, whatever version of the book, PDF, e-book or print version, a simple query on the internet with the title followed by the mention of the ISBN will bring the answer. WorldCat is a good place to retrieve all the ISBNs of a book. Beware that a book may have as many ISBNs as format versions: one ISBN for the print version, another one for the ebook, or for second edition and so on.

How to obtain an ISBN

All the information needed to get an ISBN for a publication is available at **https://www.isbn-international.org/content/how-get-isbn.**

When an ISBN has been assigned to a publication, it should always be displayed to facilitate its identification. The ISBN is also crucial for dissemination as it is displayed in a barcode format, so libraries and bookshops can process incoming stock and outgoing sales quickly and accurately. On a printed book, an ISBN should be included on the copyright page, also called the title verso page, or at the foot of the title page if there is no room on the copyright page. If there is no barcode, then the ISBN should also be on the back cover or jacket preferably on the lower right. Each version of the book needs to be provided with its own ISBN. More details on when to assign an ISBN are available here.

The publisher will then fill in the ISBN in the legal deposit form with all the additional metadata of the book for cataloguing purposes at their respective national ISBN agencies.

How to annotate and display an ISBN

Display: ISBN: 0-23-8675-309

Annotate in JATS:

Tag the ISBN number using the <isbn> element, using the *publication-format* attribute to specify the format or medium of the publication (e..g, “print”, “electronic”, “video”, “audio”, “ebook”, and “online-only”)[10]

```
<isbn publication-format="[format type]">[ISBN number]</isbn>
```

Example of an ISBN

The *Flora of New Caledonia* published by the Muséum national d’Histoire naturelle in 2020 on *Apocynaceae, Phellinaceae* and *Capparaceae* has an ISBN for its print version (978-2-85653-939-2), one for the PDF version on sale (978-285653-954-5) and one for its bundle (print + e-book: 978-2-38036-955-2).

Sample annotation in JATS with multiple ISBN numbers:

```
<book-meta>
   <isbn publication-format="print">978-2-85653-939-2</isbn>
   <isbn publication-format="PDF">978-285653-954-5</isbn>
   <isbn publication-format="bundle">978-2-38036-955-2</isbn>
<book-meta>
```

Recommendation

An ISBN is mandatory to properly identify a published book. Each version of the book (PDF, print, ebook, each linguistic version, second edition) should have its own ISBN. Considering that ISSN and ISBN are mandatory for nomenclature purposes, we must consider the use of ISBN *mandatory for taxonomic publications*.

Digital Object Identifier (DOI)

Definition

The two most commonly used agencies that register DOIs in the scholarly domain are Crossref and DataCite. Both are membership organisations providing DOIs to research outputs but for different purposes. The main difference lies in the type of digital objects they identify, the scale of numbers of DOIs needed and the metadata associated with the DOI.

Crossref is a non-profit membership organisation specifically serving scholarly publications. Its members are publishers, research institutions, university presses, societies and funders. Membership in Crossref is open to organisations that produce professional and scholarly materials and content. In addition, applicants should be able to meet the terms and conditions of membership.

DataCite is a global non-profit organisation that provides persistent identifiers (DOIs specifically) for research data and other research outputs and resources. DataCite’s members work with data centres, stewards, libraries, archives, universities, publishers and research institutes that host repositories and who have responsibility for managing, holding, curating, and archiving data and other research outputs.

[10] https://jats.nlm.nih.gov/publishing/tag-library/1.3/element/isbn.html

In their respective websites, a schema (Fig. 2) explains the rationale behind each of these two agencies (e.g. https://www.crossref.org/community/datacite/).

The DOI includes three parts:

To create the DOI, the DOI prefix given to an organisation is combined with a suffix of choice. The DOI becomes active once registered with Crossref. CrossRef provides a complete documentation on best practices to construct the suffixes.

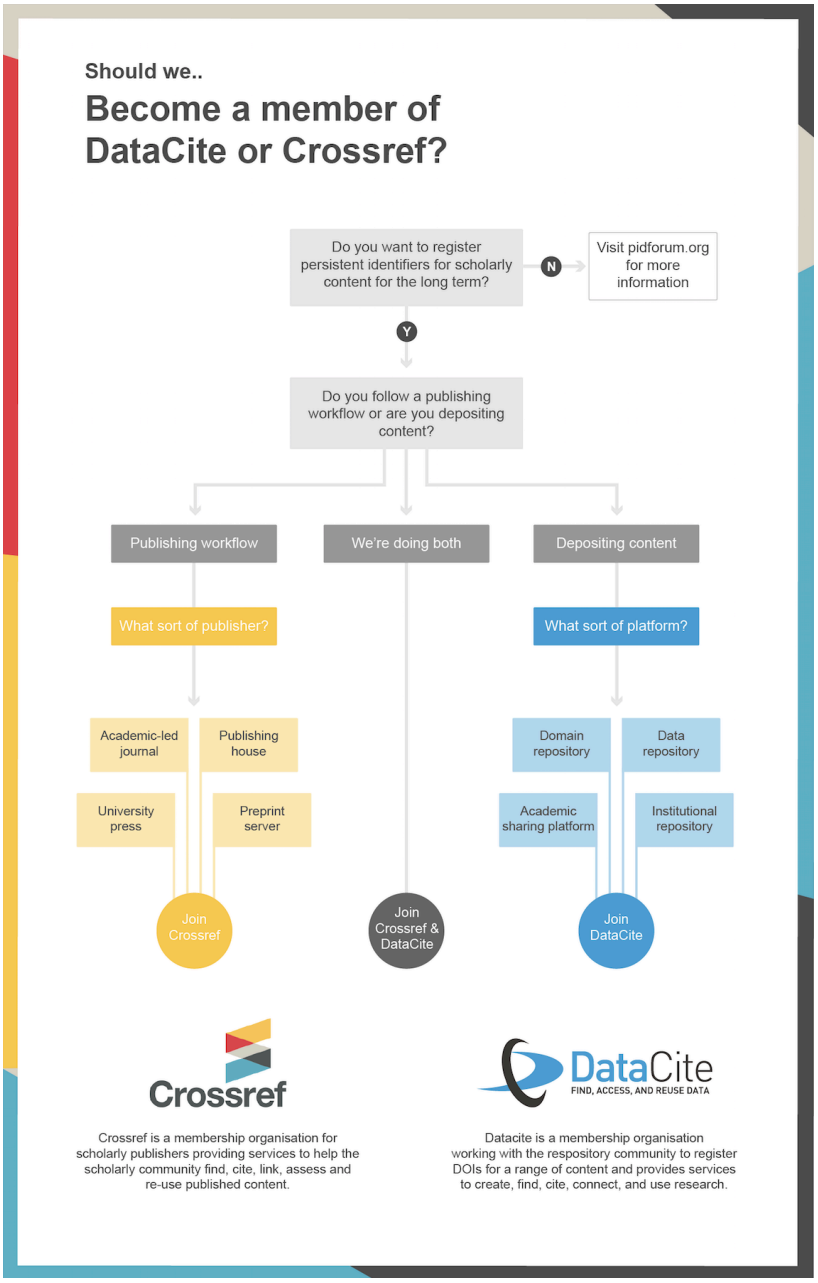


Fig. 2. Dichotomic decision tree for the registration of CrossRef and DataCite DOIs, from the DataCite - Crossref release.

How to discover an existing Digital Object Identifier (DOI)?

To find the corresponding DOI registered in Crossref, enter the title, the author or any metadata in Crossref or DataCite search engines or use alternatively the ReFindit tool.

How to mint a Digital Object Identifier (DOI)

All agencies providing DOIs are listed here: **https://www.doi.org/registration_agencies.html.** Each of them may have different rules and apply different fees. Alternative repositories to mint DOI for legacy publications are the Biodiversity Heritage Library, the Biodiversity Literature Repository and institutional libraries retro-digitising legacy publications, such as E-Periodica at the Federal Institute of Technology, Zurich.

To deposit a DOI to Crossref, one has to be a member. Membership fees begin at 275 USD and depend on the revenue of the applicant. Once a member, a DOI prefix is assigned to the joining organisation and will form the stem of links to all its metadata records. Fees vary per record type, books, research grants, preprints, etc., from 0.15 USD for a legacy article to 1 USD for a newly published article. Each DOI has to be registered by direct deposit of XML, using Open Journal System Plugin for instance or, alternatively, through an online web deposit form.

Component DOIs are often registered for figures, tables, and supplemental materials associated with a journal article. They have their own metadata distinct from that of the parent article DOI.
The registration of the DOI includes all the metadata, i.e. basic information such as dates of publication , publication outlet, including the ISSN or ISBN, article title and authors. There is a Crossref membership obligation: accurate metadata should be deposited for all DOI registered, and the metadata should be maintained for the long term, including updating any URLs that change. It is also an obligation to include DOIs in the reference lists for existing works which have DOIs. A free public API is available to retrieve all existing Crossref DOIs.

To register a DOI with Datacite, one has to be a member. Membership is open to all organisations whose missions include research output sharing. A membership fee of 2,000 euros applies to member organisations. Once a member, non-for-profit members will have to pay another 500 € annual fee to make use of DOI registration services. Each DOI, up to 1,999, will cost 0,80 €. There are two ways to register a DOI: using an API or a Web Interface. All information is provided here.

How to cite and annotate a DOI

Cite:
https://doi.org/10.3897/zookeys.1083.72939 (preferred), or DOI: 10.3897/zookeys.1083.72939

Annotate in JATS:
Tag the cited DOI with the <ext-link> element, using the *xlink-href* to provide the DOIs https version and the *ext-link-type* attribute with value "doi".[11]

```
<ext-link xlink:href="[[https-version of DOI]]"
ext-link-type="doi">[https-version of DOI]</ext-link>
```

Example of a DOI

https://doi.org/10.3897/zookeys.1083.72939 refers to an article published in *ZooKeys*.

Sample annotation in JATS:

```
<ext-link xlink:href="10.3897/zookeys.1083.72939"
ext-link-type="doi">https://doi.org/10.3897/zookeys.1083.72939
</ext-link>
```

–Recommendation

Use Crossref DOI for articles and bibliographic references. For supplementary material, figures or tables, create Crossref component DOI or DataCite DOI. Generate a DataCite DOI for data. If none is available, try to find a way to create a DOI using an alternative repository such as BHL, BLR or E-Periodica, or DOIs issued for datasets deposited at large international repositories, such as GBIF, DataONE, Dryad, Zenodo and others.

[11] https://jats.nlm.nih.gov/publishing/tag-library/1.3/element/ext-link.html

Display all the identifiers, ISSN, ISBN, DOI, on the corresponding publication page and register all the corresponding metadata associated with the DOIs with CrossRef or DataCite. Always include the DOI in the metadata for other publication-related registration purposes, for example at ZooBank, IPNI, MycoBank, Zenodo, Dryad and others.

Body of the article

Definition

Most academic journals require the authors to write their articles following the IMRaD format. IMRaD stands for Introduction, Method, Result and Discussion which are the four main sections that constitute the structure of most scientific papers in the Science, Technical and Medical (STM) fields. The body of the article is the main textual and graphic content of the article and is situated between the front and the back matters. This usually consists of sections, subsections, and paragraphs, which may themselves contain figures, tables, etc.

In a taxonomic article, the body of the article includes specific items, such as taxonomic treatments, material citations, descriptions, differential diagnoses, details of collecting permits, etc.

Sections

Definition

Most journal articles are divided into sections, each with a title that describes the content of the section, such as “Introduction”, “Materials and Methods”, or “Conclusions”[12]. A special section in taxonomic publications is the taxonomic treatment as described below. The different sections include different kinds of data and information that are important to reproduce the research. For example, the section “Materials and Methods” lists the collections studied, software used to analyse the data, or instruments used to make measurements.

What are their identifiers?

Sections are normally tagged with internal UUIDs in the article XML. In addition, the names of the sections, which are used more or less consistently in various science domains, e.g., “Introduction”, “Material and Methods”, “Results”, “Conclusions” etc. can be used for inferring a semantic meaning of their content, an approach that is currently used for the conversion to RDF and export to the OpenBioDiv knowledge graph.

How to annotate sections

In JATS, the sections are annotated using the following elements and attributes:

```
<sec sec-type="[section type]" id="[internal identifier]">
   <sec-meta>
      <mixed-citation>
         <object-id object-id-type="uuid">[UUID]</object-id>
      </mixed-citation>
...
</sec>
```

According to the JATS tag library, the *sec-type* attribute “[n]ames the main semantic type of section content”. Following the recommendation that *sec-type* “is most useful when a list of values is maintained, and articles are tagged accordingly”, for JATS the values: "cases", "conclusions",

[12] https://jats.nlm.nih.gov/publishing/tag-library/1.3/chapter/nfd-body-and-sec.html

"discussion", "intro", "materials", "methods", "results", "subjects", "supplementary-material", are recommended.[13]

The “id” attribute is a unique internal identifier of an element; it allows the element to be cross-referenced [and linked to]. The value must be unique across a document…[*id*] holds an internal document identifier that can be used by software to perform a simple link. An *id* should not be confused with elements…that are used to hold externally defined identifiers such as a DOI”[14] For an externally defined identifier assigned to the section, a *<sec-meta>*[15] element may be used to provide metadata for a section, which includes *<mixed-citation>*[16] containing an *<object-id>*[17] element to record an identifier, for example, a UUID.

Though not recommended, a lighter-weight solution for associating an external identifier with a section is to “overload” the *id* attribute of <sec> by using an external identifier such as a UUID as the value. However, the “id” attribute “must start with a letter of the alphabet”[18], so UUIDs (which may start with a digit) should be prefixed with a string starting with an alphabetic character, e.g., “uuid-”, to validate.

Example

Annotation of a section “Methods” including an object identifier taken from the article of Bueno-Soria *et al.* (2022).

```
<sec sec-type="methods" id="SECID0E1H">
      <sec-meta>
            <mixed-citation>
                  <object-id
            object-id-type="uuid">07ED460C-7F70-42EC-A7F8-59CC2F512131</o
            bject-id>
            </mixed-citation>
      </sec-meta>
      <title>Methods</title>
      <p>
The specimens of the genus Xiphocentron studied here were borrowed
from the collections of the National Museum of Natural History,
Smithsonian Institution in Washington, DC, and from the Colección
Nacional de Insectos, Instituto de Biología de la Universidad
Nacional Autónoma de México.
      </p>
      <p>
The type materials are deposited as indicated in each species
description, in the collections: National Museum of Natural
History, Smithsonian
      </p>
</sec>
```

[13] https://jats.nlm.nih.gov/publishing/tag-library/1.3/attribute/sec-type.html
[14] https://jats.nlm.nih.gov/publishing/tag-library/1.3/attribute/id.html
[15] https://jats.nlm.nih.gov/publishing/tag-library/1.3/element/sec-meta.html
[16] https://jats.nlm.nih.gov/publishing/tag-library/1.3/element/mixed-citation.html
[17] https://jats.nlm.nih.gov/publishing/tag-library/1.3/element/object-id.html
[18] https://jats.nlm.nih.gov/publishing/tag-library/1.3/attribute/id.html

Recommendation
Section and subsection titles should be tagged as such and Internal UUIDs should be assigned to them in the article XMLs.

Figures, figure captions and citations

Definition
A figure is either a photo or a scientific drawing illustrating biological species or part(s) of them, landscapes, habitats or equipment, or visualisation of data or results from statistical analyses. Figures and their captions convey an essential part of the information contained in a scientific paper and are of particular interest for the community.

The ICN states the importance of illustrations in its Art. 43.2:
"A name of a new fossil-genus or lower-ranked fossil-taxon published on or after 1 January 1912 is not validly published unless it is accompanied by an illustration or figure showing the essential characters or by a reference to a previously and effectively published such illustration or figure."

According to article 40.3, illustrations can also be a type specimen prior to 1 January 2007[19].

The figures related to a taxonomic treatment (see definition below) are usually cited at the beginning of the treatment and are part of it.

What are their identifiers
DOIs being either Crossref component DOIs or DataCite DOIs are usually used when the figures are deposited in a repository.

How to mint an identifier
For minting DOIs, see section "Digital Object Identifiers" above. If no DOIs are minted for figures, these can be identified with internal UUIDs minted by software during the compilation of the full-text article XML, and a hash of the figure allows to uniquely identify the respective figure.

How to annotate and cite a figure, figure caption and figure citation

Cite: Figures are cited within the text following the long-established practice in scholarly publishing (e.g., "according to Fig. X" or "see detail (Fig. Y)"). Citation style should follow the journal's or publisher's instructions for the authors.

Annotate in JATS:

Figure:
```
<fig id="[Internal identifier]">
     <object-id object-id-type="doi">[DOI]</object-id>
     <caption>[free_text]</caption>
</fig>
```

In-text figure citation
```
<xref ref-type="fig" rid="[Internal identifier]">[Figure
reference]</xref>
```

[19] https://www.iapt-taxon.org/nomen/pages/main/art_40.html

Note that the citation details of the figure as delivered by the Biodiversity Literature Repository (BLR) at Zenodo should contain both the DOI of the article and the component DOI of the figure. In the case where no CrossRef component DOI exists, a DataCite DOI is minted for the figure at BLR. The “rid” (reference to an identifier) attribute is needed to perform the linking with the <fig> element via the embedded “id” element.

Examples

Annotation of a figure and in-text figure citation (from Blahnik and Andersen 2022):

Figure

```
<fig id="F4">
     <object-id
object-id-type="doi">10.3897/zookeys.1111.77586.figure1</object-id>
     <caption>
          <p>Malaise trap across Kaputu Stream at 1535 m altitude
     in the Mazumbai Forest Reserve in the West Usambara
     Mountains, northeastern Tanzania (Photo: Trond Andersen).</p>
     </caption>
</fig>
```

In-text figure citation

```
<xref ref-type="fig" rid="F4">Fig. 4</xref>
```

https://doi.org/10.3897/zookeys.1088.78139.figure1 is a CrossRef component DOI assigned to Figure 1 of the article https://doi.org/10.3897/zookeys.1088.78139.

Recommendation

Use Crossref component DOI to identify each figure within an article. The component DOI has the important feature of a link from the figure DOI to its parent article DOI. If no Crossref DOI is available, use alternatives from DataCite.

In all cases, and especially If no DOIs are minted for figures, it is recommended to assign internal UUIDs minted by software during the compilation of the full-text article XML as well as a hash for unique identification.

When compiling the full-text XML, it is highly recommended to cross-reference (anchor) the in-text figure citations to their respective figures in the article body.

Tables, table citations

Definition

A table is a concise and effective way of presenting large amounts of data usually displayed in rows and columns for reference.

Tables are increasingly important because they contain, in many cases, a compilation of the specimens used, their sequence accession codes, specimen codes that allow linking to the cited specimens, as well as traits, such as measurements or qualitative descriptions or even the results of an analysis performed on the raw data taken from the specimens or from their environment. Each

row can be envisioned to represent a structured material citation, and if used to list species used in a study, together with a taxonomic name, an entire taxonomic treatment.

Table identifiers

In TreatmentBank, tables are identified by a UUID and a persistent http URI ID. In the Pensoft article XMLs, tables are identified by internal UUIDs.

How to mint a table identifier

DOIs or Crossref component DOIs relating to the article, should be minted and submitted for registration to Crossref by the publisher. If no DOIs are minted for tables, these can be identified with internal UUIDs minted by a software during the compilation of the full-text article XML.

Annotating and citing tables

Tables are cited within the text following the long-established practice in scholarly publishing (e.g., "according to Tab. X" or "see data (Tab. Y)"). Citation style should follow the journal's or publisher's instructions for the authors.

Annotate in JATS:

Table

```
<table-wrap id="[Internal identifier]">
     <object-id object-id-type="doi">[Digital Object Identifier]</object-id>
      <caption> Text</caption>
      <table>
      ...
      </table>
</table-wrap>
```

In-text table citation

```
<xref ref-type="table" rid="[Internal identifier]">[Table reference]</xref>
```

The "rid" attribute is needed to perform the linking to the <table> element via the "id" attribute of the target <table-wrap> element, which itself has optional, repeatable <object-id> elements recording identifiers for the table it contains. The JATS tag library defines the <object-id> element as a "Unique identifier (such as a DOI or URI) for a component within an article (for example, for a figure or a table).", further stating that, "The <object-id> element holds an external identifier, typically assigned to an object such as a table by a publisher. The contents of this element should not be confused with the @id attribute, which holds an internal document identifier that can be used by software to perform a simple link inside the document."[20]

Examples

Annotation of a table and in-text table citation (from Blahnik and Anderson 2022)

Table

```
<table-wrap id="TID0ER4DK">
      <object-id content-type="table"
object-id-type="doi">10.3897/zookeys.1111.77586.table1
```

[20] https://jats.nlm.nih.gov/publishing/tag-library/1.3/element/object-id.html

```
      </object-id>
      <caption> Table 1. African genus Chimarra
      </caption>
      <table>
      ...
      </table>
</table-wrap>
```

In-text citations

```
<xref ref-type="table" rid="T1TID0ER4DK">Tab. 1</xref>
```

Recommendation

Ideally, a table should be provided with a Crossref component DOI related to the article. In all cases, and especially If no DOIs are minted for tables, it is recommended to assign internal UUIDs minted by software during the compilation of the full-text article XML.

When compiling the full-text XML, it is highly recommended to cross-reference (anchor) the in-text table citations to their respective tables in the article body.

Taxonomic treatments

Definition

Taxonomic treatments are sections of publications documenting the features or distribution of a related group of organisms (taxon) (Catapano 2010). Each taxonomic name relates to at least one taxonomic treatment: a publication, or more frequently a section of a publication documenting the features of a taxon in ways adhering to highly formalised conventions. Some of these descriptions are over two centuries old and are maintained by taxonomic community ethical and professional norms regulated by the Nomenclatural Codes. The modelling of taxonomic treatments in TaxPub XML is designed to follow the FAIR principles and provide clarity and repeatability of the research, which both are integral parts of the modern evidence-based science.

The features and structure of treatments have changed over time, and vary between and within publications. Often an indication follows the name of whether the taxon is new to science, e.g. “species nova, “sp. nov.” or “genus novum”, “gen. nov.” and the name or names of the persons who attribute the naming. A listing of taxa that are already known to science, citations of earlier treatments (treatment citations), often follows in a section In cases when taxonomic names change as a result of a taxonomic revision, for example because of a raise in its rank, or because a taxon is synonymized, this is followed by a label stating the change, such as for example “syn. nov.” or “nov. stat.”. Other information, such as persistent identifiers and references to physical specimens, may also be found.

A number of other sections may follow the nomenclature section. One of the most significant sections, frequently titled “Materials Examined”, includes citations to specimens used as the basis of the treatment and data about their properties (e.g., DNA sequences). This section often includes the circumstances of collection and/or deposition at a museum or other institution. Historically, these details have allowed scientists to visit the holding institution, or to seek a loan, for further scientific investigation of the same material that was described by the treatment. Also common is a “Description” section providing information — often in highly structured language, and sometimes in tabular form — on the distinctive features of the collected organisms, with an aim toward characterising the entire taxonomic class such material represents.

Similar to a “Description” section, there is a “Diagnosis” section, which contains descriptions of only those features or unique combinations of features “*that distinguish that species from others, in the same way that the disease identification you receive when you visit the doctor is called the diagnosis because the doctor has distinguished your illness from all other possibilities based on the basis of your symptoms and tests*” (Winston 1999). Most treatments describing new taxa include an “Etymology” section explaining the origin of the assigned Latin name, a “Distribution” section summarising the spatial and temporal distribution of the taxon, or an “Ecology” section discussing behaviour and relationships to habitat or details on the environmental variables measured during the collecting events of the specimens. For higher level taxa (such as genera and families) a “Key” presenting a set of instructions, in the form of a decision tree or even workflow, for distinguishing lower level taxa from one another is also common (Catapano 2010).

Similar to publications and following the FAIR principles, the treatments can be extracted from the publications, preserved separately and made freely accessible to the public (Fig. 3; Agosti & Egloff 2009; Paterson *et al.* 2014).

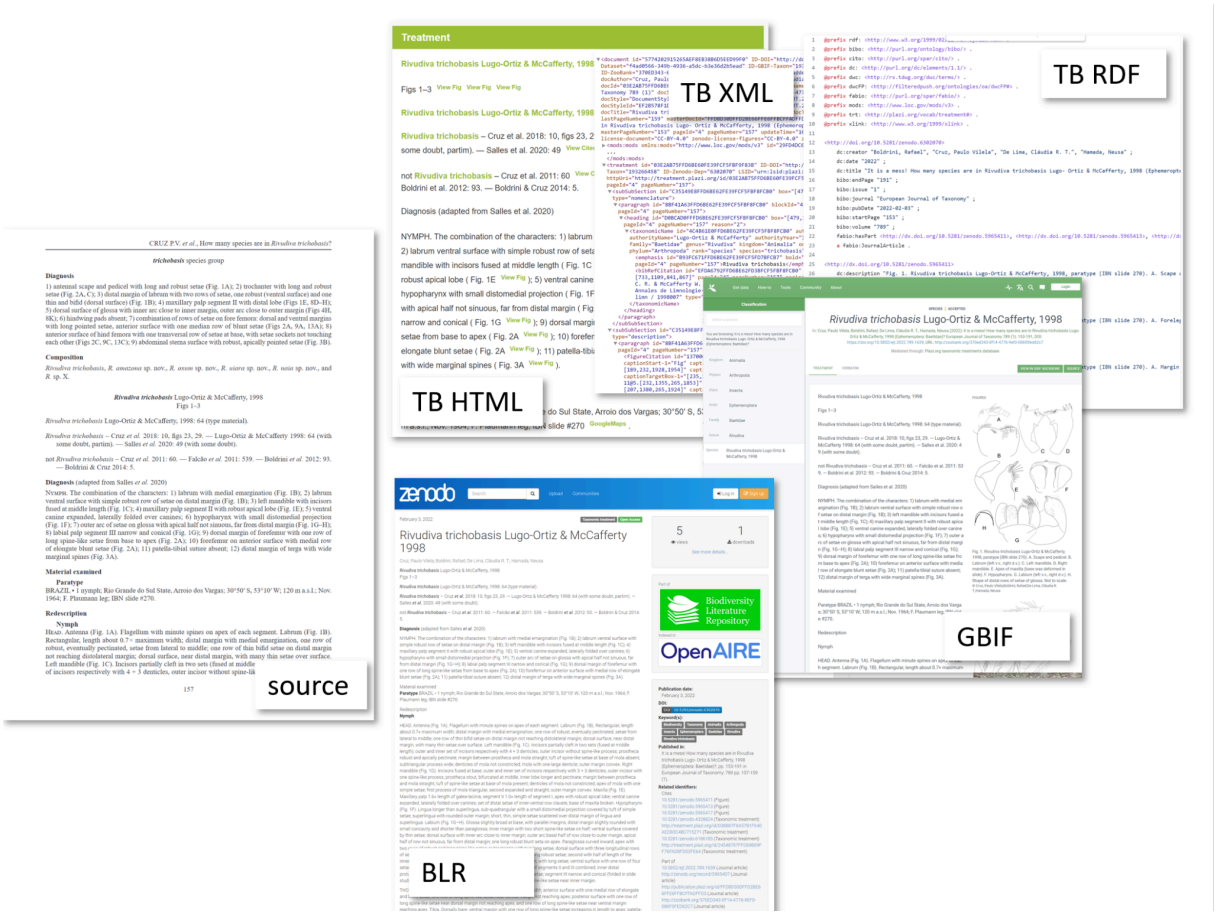


Fig 3. An example of a treatment accessible in various formats. Source: https://europeanjournaloftaxonomy.eu/index.php/ejt/article/view/1639/5873
Alternative format:.
TreatmentBank (TB) HTML:
https://tb.plazi.org/GgServer/html/03E2AB75FFD6BE60FE39FCF5FBF9F83B
TB XML: https://tb.plazi.org/GgServer/xml/03E2AB75FFD6BE60FE39FCF5FBF9F83B

TB RDF:
https://github.com/plazi/treatments-rdf/blob/main/data/03/E2/AB/03E2AB75FFD6BE60FE39FCF5FBF9F83B.ttl
BLR: http://doi.org/10.5281/zenodo.6302070
GBIF: https://www.gbif.org/species/193266458

An XML tagset for Taxonomic treatments has been formalised as an extension of the Journal Article Tag Suite (JATS) (Catapano 2010), and adopted in 2010 by Pensoft Publishers in their journal production process (Penev et al. 2010), now including 38 journals[21]. The export of treatments from published PDFs has been adopted by CETAF's *European Journal of Taxonomy* and Muséum national d'Histoire naturelle (5 journals[22]). Legacy publications are annotated and treatments are made accessible by TreatmentBank (780,000 treatments as of August 2022) and the Biodiversity Literature Repository (390,000 treatments), including current content from 52,000 articles. Together with the treatments exported by Pensoft, the total number of processed articles exceeds 70,000.

Treatments are reused by GBIF upon extraction, where they are imported as part of a dataset in a Darwin Core Archive format compiled from taxonomic treatments and cited figures. Currently these article-based datasets represent almost 60% of all the datasets published in GBIF.

In Wikidata, taxonomic treatments can be annotated with the property *taxonomic treatment* (P10594)[23], with protologue as a subclass referring to the treatment used to describe a new taxon, that is to create an available name sensu the ICN.

The Barcode of Life Data Systems (BOLD) Barcode Identification Numbers (BINs) (Ratnasingham & Hebert 2013) are functionally similar to treatments, though they are not sections of taxonomic publications and provide less information. BINs are dynamically generated by the Barcode of Life Data System (BOLD), through an online framework that clusters barcode sequences and generates an identifier and web page for each cluster. This framework uses a clustering algorithm based on graph theoretic methods to assign BINs (Ratnasingham & Herbert 2013). Each BIN is assigned two identifiers, a resolvable URI generated by BOLD, that consists of an alphanumeric identifier composed by the prefix BOLD followed by 3 letters and a 4-digit number (e.g. BOLD:AAA0111), and a DOI. When the submission of new information leads to the merge of two BINs, the most recently registered BIN is synonymized. But, when the analysis splits a BIN into two, new BINs are established and a disambiguation option is suggested. In any case DOI amendments are made to ensure that original identifiers are not lost.

The UNITE Species Hypotheses (SHs) are functionally similar to treatments including all the clustered public fungal ITS sequences to which a unique DOI is assigned by UNITE. UNITE is a database and sequence management environment for the molecular identification primarily of fungi but now also of other taxa. It focuses on nuclear ribosomal internal transcribed spacer (ITS) region sequences that are considered the fungal barcode. All species hypotheses have a unique URL where the associated public sequences are displayed (Nilsson *et al.* 2019). These sequences are referenced through their accession numbers and linked to their original records at International Nucleotide Sequence Database Collaboration (INSDC, Arita et al. 2021).

[21] https://tinyurl.com/4e6w6utj
[22] https://tinyurl.com/38hp7tms
[23] https://www.wikidata.org/wiki/Property:P10594

Identifiers for taxonomic treatments

DOI

A subtype "taxonomictreatment" has been added in Zenodo as a DataCite digital object identifier (DOI) to the "publication" type. The metadata for taxonomic treatments in Zenodo are enhanced with added custom keywords based on existing domain specific vocabularies (e.g. Darwin Core), links to the source publication, cited figures or related identifiers such as the http URIs minted by TreatmentBank (see below). In case of treatments deposited in BLR *via* TreatmentBank, the respective HttpURI are included in the metadata.

For BINs[24] and Species Hypotheses[25], DataCite DOIs are minted.

HTTP URI

The "HttpURIs' were created by Plazi for treatments in 2009 parallel to the development of the persistent HTTP URIs for specimens now widely accepted in CETAF. The HTTP URIs are used by GBIF when reusing TreatmentBank treatments. The HTTP URIs are kept persistent and are built based on a unique UUID and the prefix "http://treatment.plazi.org/id/UUID" (e.g. http://treatment.plazi.org/id/0000C505-BB5D-484C-76BE-9AB6999DEB23). The original intention was to share the UUID with Zoobank whereby the Zoobank UUID would resolve to the taxonomic name and to the respective taxonomic treatment in TreatmentBank. Unfortunately this synchronisation has been discontinued.

UUID

During the publication of a taxonomic article, Pensoft journals assign UUIDs to each taxon treatment. Those UUIDs are further used by Plazi to mint the HTTP URIs of the treatments at TreatmentBank.

How to discover identifiers

The DOI of a treatment can be found by searching ReFindit or for those minted by Biodiversity Literature Repository, through the search engines of Zenodo or TreatmentBank. The HTTP URIs can be found through GBIF or the Biodiversity Literature Repository (BLR) or TreatmentBank.

Via ReFindit API (search by author, year, and taxon name, the latter as title):
https://refindit.org/find?search=advanced&author=Kronestedt&year=2011&title=Pardosa%20zyuzini (requires some further matching to pick correct result)
Via Zenodo UI (full text search):
https://zenodo.org/search?page=1&size=20&q=Pardosa+zyuzini+Kronestedt+2011
Via TreatmentBank statistics API (exact match search on author, year, and taxon name):
https://tb.plazi.org/GgServer/srsStats/stats?outputFields=doc.uuid+bib.author+bib.title+bib.pubDate+bib.origin+bib.firstPage+bib.lastPage+bib.articleFirstPage+bib.articleLastPage+tax.name+tax.genusEpithet+tax.speciesEpithet+tax.authName+tax.authYear+tax.status&groupingFields=doc.uuid+bib.author+bib.title+bib.pubDate+bib.origin+bib.firstPage+bib.lastPage+bib.articleFirstPage+bib.articleLastPage+tax.name+tax.genusEpithet+tax.speciesEpithet+tax.authName+tax.authYear+tax.status&format=JSON&FP-tax.name=%22Alloplasta%25japonica%22&FP-tax.authName=Watanabe&FP-tax.authYear=2022 (simplified API under development)
Via TreatmentBank UI search (fuzzy search for taxon name):

[24] http://www.boldsystems.org/index.php/Public_BarcodeIndexNumber_Home
[25] https://unite.ut.ee/

https://tb.plazi.org/GgServer/search?taxonomicName.isNomenclature=true&taxonomicName.exactMatch=true&taxonomicName.taxonomicName=Pardosa+zyuzini
Via GBIF UI (dataset search):
https://www.gbif.org/search?q=Prosymna+lisima
Via GBIF API (species search):
http://api.gbif.org/v1/species?name=Prosymna+lisima

How to mint an identifier

Currently, Zenodo is the only place to mint a treatment DOI. UUIDs are generated by some publishers during the article processing before publication (all Pensoft journals, for example). HTTPURIs are minted by TreatmentBank. This is not an exclusive solution, however, since a treatment is a subtype of the DataCite publication type at Zenodo.

How to annotate and cite treatments

Cite: A citation of a treatment can be provided either by its DOI or its HTTP URI generated by Plazi's TreatmentBank. The citation of other treatments normally happens within a given treatment's Nomenclature section (in the so-called "nomenclature-citation-list" of the JATS/TaxPub XML representation), where they can also introduce a nomenclatural change, indicated with a label (e.g. syn. nov., comb. nov., nom. nov, etc.).

Annotate in JATS (*treatment and subsections in JATS/TaxPub):*

```
<tp:taxon-treatment>
      <tp:treatment-meta>
            <mixed-citation>
                  <object-id object-id-type="doi">[digital object
            identifier]</object-id>
            </mixed-citation>
      </tp:treatment-meta>
</tp:taxon-treatment>
```

For the nomenclature subsection

```
<tp:nomenclature>
      <tp:taxon-name>[Taxon name]</tp:taxon-name>
</tp:nomenclature>
```

For all other subsections

```
<tp:treatment-sec sec-type="[section type name]">
```

Section types should, if possible, make use of the following vocabulary terms: description, diagnosis, discussion, distribution, ecology_behavior, conservation, etymology, materials_examined, reference_group, and vernacular_names which will add a semantic meaning to (sub-)section titles and facilitate the extraction and reuse of the data.

Examples

Annotation of treatments, nomenclature section and subsections:

Treatment

```
<tp:treatment-meta>
     <mixed-citation>
        <object-id
content-type="doi">10.5281/zenodo.6964440</object-id>
     </mixed-citation>
```

Nomenclature section

```
<tp:nomenclature>
      <tp:taxon-name>[Tribasodites yatung]</tp:taxon-name>
</tp:nomenclature>
```

Subsection

```
<tp:treatment-sec sec-type=”description">
```

Recommendation

Tag each taxonomic treatment in the article full-text XML and then assign a CrossRef Component DOI or Datacite DOI or internal UUID for it. Register all the metadata associated with the DOI.

Treatment citations

Definition

A treatment citation is a reference to a previous treatment, in many cases the original description of the taxon, or protologue (Fig. 4). Treatment citations reflect the history of the taxon and its nomenclatural relationships with other taxon concepts, either by indicating a change proposed in the treatment, e.g. a new synonymy or a new combination, or by reconfirming previous changes. They also refer to treatments that contributed new research results to an existing taxon. Thus, treatment citations can be grouped in several categories, e.g. by type of a nomenclatural change (“syn. n.”, “comb. n.”, etc.,) or by confirmation of previous taxon name status, and those categories allow formal annotation during the text mining process and further-re-use.

Treatment citations are the source and basis for creating synonymic lists and taxonomic catalogues. Treatment citations are analogous to bibliographic references in a publication citing previous works.

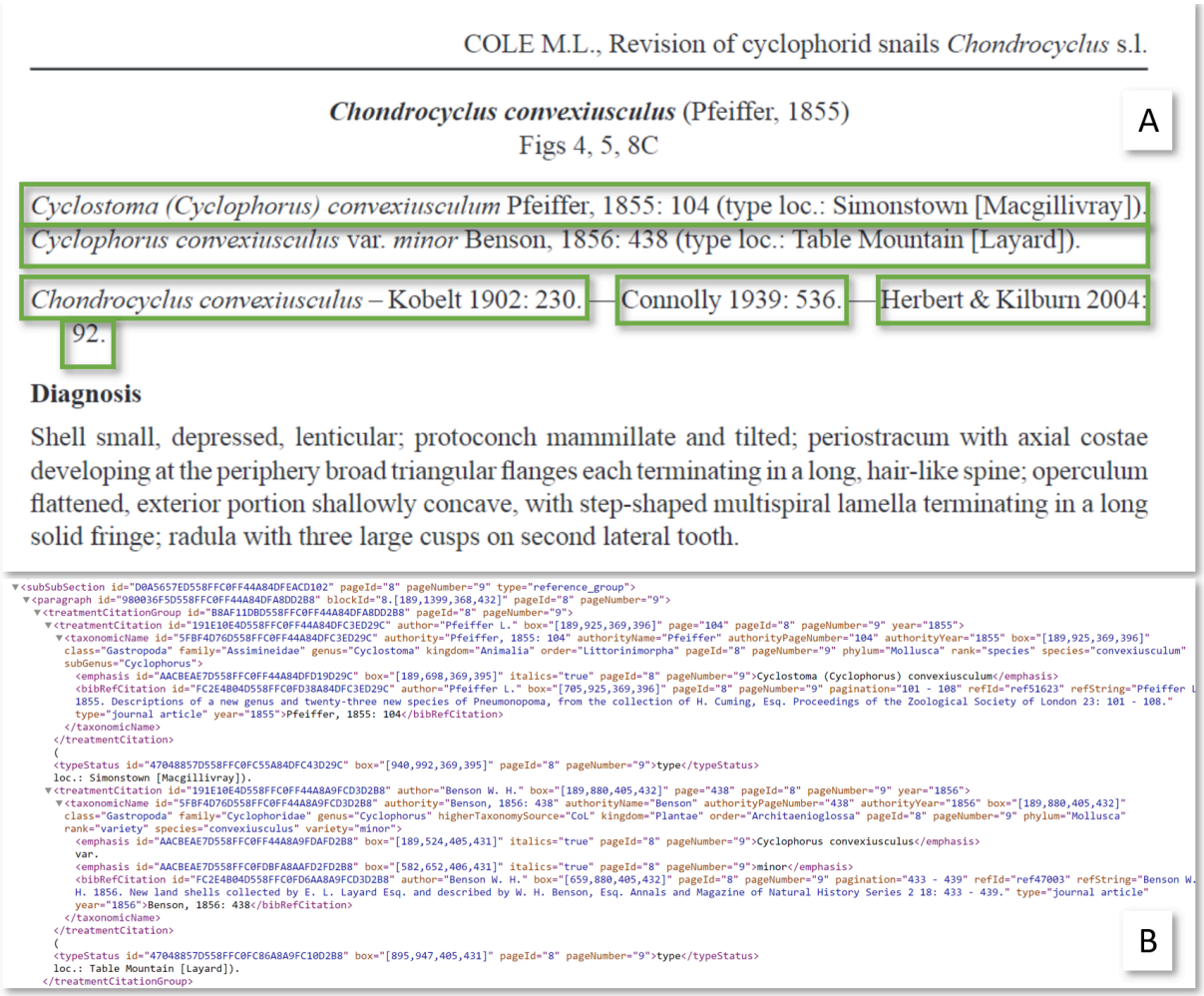


Fig. 4. Treatment citation. Green boxes: treatment citations. A. published example. B: annotated treatment citations in TreatmentBank using TB internal XML. Source: https://europeanjournaloftaxonomy.eu/index.php/ejt/article/view/787/1829
TreatmentBank XML:
https://treatment.plazi.org/GgServer/xml/101687E3D558FFC3FDF9A837FECED008
.

Identifiers for treatment citations

No identifiers are known, however citations can and should be tagged in the backend XML of the article to be made discoverable and processed for further use.

How to discover treatment citations

Treatment citations are listed subsequent to the nomenclatural sections of a taxonomic treatment. They usually consist of a taxonomic name, the authority and year, and a page number, especially in zoology. In combination, the authority and year are also a bibliographic citation of the original publication of the respective treatment, albeit often implicit, because traditionally, taxonomists do not include this kind of bibliographic references in the article reference list (Bénichou *et al.* 2018), a procedure now suggested by CETAF, BHL and SPNHC (Bénichou *et al.* 2022) and strongly encouraged by Pensoft's journals. In case of multiple citations for the same taxonomic name, a further element (treatment citation list) is included that allows that the taxonomic name is not to be repeated in each case.

How to mint an identifier

There is no established procedure for minting treatment citations, except for possible assignment of internal UUIDs to them.

How to annotate and cite a treatment citation

The treatment citation annotations are attributed with persistent HTTPURIs of the respective treatment(s) in TreatmentBank. The treatment citation element is currently being remodelled and thus the recommendations might change in the next version of TaxPub.

Annotate in JATS/TaxPub:

```
<tp:nomenclature>
      <tp:taxon-name>
            <tp:taxon-name-part
            taxon-name-part-type="genus">Cus</tp:taxon-name-part>
            <tp:taxon-name-part
            taxon-name-part-type="species">dus</tp:taxon-name-part>
      </tp:taxon-name>
      <tp:nomenclature-citation-list>
            <tp:nomenclature-citation>
                  <tp:taxon-name>Aus bus</tp:taxon-name>
                              <mixed-citation><object-id
                        content-type="taxonomic_treatment"
                        object-id-type="doi">[DOI]</object-id></mixed-citati
                        on>
                  </tp:nomenclature-citation>
        </tp:nomenclature-citation-list>
    </tp:nomenclature>
```

Example

Annotation of treatment citation in the treatment of *Chondrocyclus convexiusculus* (Pfeiffer, 1855) in Cole (2019):

```
          <tp:nomenclature-citation-list>
        <tp:nomenclature-citation>
            <tp:taxon-name>Cyclostoma (Cyclophorus)
convexiusculum</tp:taxon-name>  Pfeiffer, 1855: 104
            (Type loc.: Simonstown [Macgillivray]).
        </tp:nomenclature-citation>
        <tp:nomenclature-citation>
            <tp:taxon-name>Cyclophorus convexiusculus var.
minor</tp:taxon-name>  Benson, 1856: 438
            (type loc.: Table Mountain [Layard]).
        </tp:nomenclature-citation>
    </tp:nomenclature-citation-list>
    <tp:nomenclature-citation-list>
        <tp:taxon-name>Chondrocyclus convexiusculus</tp:taxon-name>
        - <tp:nomenclature-citation>Kobelt 1902:
230</tp:nomenclature-citation>
        <tp:nomenclature-citation>Connolly 1939:536
        </tp:nomenclature-citation>
        <tp:nomenclature-citation>Herbert &amp;amp; Kilburn 2004: 92
        </tp:nomenclature-citation>
    </tp:nomenclature-citation-list>
```

Recommendation

Treatments should be cited with their PIDs either through their inclusion in the Nomenclature section of a treatment or as a standalone in-text citation in any part of the article as follows: “Based on Treatment: [hyperlinked treatment PID, where a treatment PID can be either the DOI of the treatment provided by BLR or Plazi’s HTTP identifier available from TreatmentBank] I conclude that ....”.

Treatment citations should be tagged in the article XML as separate entities and, if available, should contain the existing PIDs of the cited treatments.

Recently, a joint statement of CETAF, SPNHC and BHL has been published (Benichou *et al.* 2022) recommending extended citation details of taxon names by adding richer bibliographic citation detail to each taxon concept. We provide here a shortened version of these recommendations:

1. Provide each scientific name of a taxon, at least on its first mention in the paper, with authorship, date, and corresponding entries to the publication’s “Bibliographic references” section.
2. If the publisher’s guidelines do not allow you to list it as a reference, cite it properly as a bibliographic reference by adding the page number after the date for instance. For example, for a species described in *EJT* http://dx.doi.org/10.5852/ejt.2022.828.1851 p. 48, it is preferable to use the notation *Infrantenna fissilis* Liu & Sittichaya, 2022: 48 instead of *Infrantenna fissilis* Liu & Sittichaya, 2022.
3. Provide the corresponding persistent identifier (PID) to each of these references where they exist, i.e. a Crossref DOI minted by the publisher or minted by the Biodiversity Heritage Library (BHL) when the legacy publication has been digitised retrospectively and provided with a DOI, or a DataCite DOI minted by organisations digitising legacy literature (e.g. e-Periodica at the Federal Institute of Technology Zurich) or the Biodiversity Literature Repository (BLR) at Zenodo.
4. Provide the PID of the taxonomic treatment where they exist, using for instance, the DOI of the treatment deposited in BLR, or for articles with primary taxonomic descriptions minted by BHL (for example: https://www.biodiversitylibrary.org/part/304567).

Material citations

Definition

A material citation is a reference to, or citation of, one or multiple specimens in scholarly publications (https://dwc.tdwg.org/terms/#materialcitation). Material citations can be situated within the respective treatments, in tables, or as supplementary material, and refer to the specimen data used in the study. They provide the best, expert-curated identification of specimens in collections including, in many cases, explicit links to the institution, specimen, gene sequences and geographic data. Often they are the only evidence of the existence of a specimen in the digital world, for example, if published through the GBIF infrastructure.
The GBIF occurrences can create a rich linking network for specimens because a GBIF specimen record can be linked to a material citation published in a scholarly article, or at least to the treatment or publication containing that record.

Identifiers of material citations

TreatmentBank and the Biodiversity Data Journal issue internal UUIDs for material citations. They are reused in conjunction with the treatment UUID in GBIF in the form of "treatment UUID.mc.material citations UUID". GBIF is minting an identifier for each material citation present as an occurrence record in their infrastructure. TB maintains the links and identifiers of the occurrences in GBIF with their respective material citations in TB.

How to discover material citations

These identifiers are currently minted post-publication by TreatmentBank, or before publication by the Biodiversity Data Journal, and can be found using TreatmentBank data access interface (https://tb.plazi.org/GgServer/srsStats) which can also provide access to the related GBIF occurrence ID.

Via TreatmentBank statistics UI with a wide variety of search fields and output fields to choose from. The link (visit TreatmentBank statistics UI) shows taxon name, author, year, and type status.

Via TreatmentBank statistics API with a wide variety of search fields and output fields to choose from in the UI. The link (visit TreatmentBank statistics API) shows taxon name, author, year, and type status, retrieved as JSON. A more simplified API is under development.

Via GBIF API (occurrence search for taxon name, restricted to materials citations)
http://api.gbif.org/v1/occurrence/search?basisOfRecord=MATERIAL_CITATION&scientificName=Lebertia+insignis Other search fields are also available, e.g. country, might require further matching efforts to find additional matches from specific source publications in GBIF.

How to mint an identifier

Follow your standard procedure for minting UUIDs.

How to annotate and cite material citation

Annotate in JATS/TaxPub:

Use "object-id" to provide an identifier for a material citation in the article which allows it to be cited unambiguously.

```
<tp:material-citation>
      <object-id content-type="[content type]">[Identifier]</object-id>
      [material citation string]
 </tp:material-citation>
```

To provide an external identifier for a component of a material citation (e.g., a catalog number or occurrence id), use <named-content>, specifying the type of identifier in the *content-type* attribute.

```
<named-content content-type="[content
type]">[Identifier]</named-content>
```

The <uri> element may be used to tag an identifier that is a URI and provide a live link to the representation of the identified resource:

```
<named-content content-type="[content type]">
<uri xlink:href="[URI]">[URI]</uri>
</named-content>
```

Examples

A material citation from *Monomorium dryhimi* that can be unambiguously cited (Aldawood & Sharaf, 2011) .

```
<tp:material-citation>
      <object-id
content-type="arpha">B5596AA1-CDF9-DDA3-D5CD-D922E1723751</object-id>
       Holotype worker. SAUDI ARABIA, Al Bahah province, Amadan
forest, Al Mandaq governorate, 20°12'N, 41°13'E, 1881 m.a.s.l.
19.V.2010 (M. R. Sharaf &amp;amp; A. S. Aldawood Leg. KSMA
</tp:material-citation>
```

A material citation citing a specimen from the MNHN Paris (Paton *et al.*, 2016))

```
<tp:material-citation>
      <named-content content-type="dwc:occurrenceID">
      <uri
xlink:href="http://coldb.mnhn.fr/catalognumber/mnhn/p/p04158076">ht
tp://coldb.mnhn.fr/catalognumber/mnhn/p/p04158076</uri>
      </named-content>
…
</tp:material-citation>
```

Finer grained markup

```
<tp:material-citation>
      <object-id
content-type="arpha">B5596AA1-CDF9-DDA3-D5CD-D922E1723751</object-i
d>
      <tp:type-status>Holotype</tp:type-status> worker.
      <tp:material-location>King Saud Museum of Arthropods (KSMA),
College of Food and Agriculture Sciences, King Saud University,
Riyadh, Kingdom of Saudi Arabia.</tp:material-location>
      <tp:collecting-event>
            <tp:collecting-location>
                  <tp:location>SAUDI ARABIA, Al Bahah province,
            Amadan forest, Al Mandaq governorate</tp:location>
            </tp:collecting-location>,
                  <named-content
            content-type="dwc:verbatimCoordinates">20°12'N,
            41°13'E</named-content>
, 1881 m.a.s.l. 19.V.2010 (
                  <named-content content-type="dwc:recordedBy">M. R.
            Sharaf</named-content>
&amp;
```

```
                  <named-content content-type="dwc:recordedBy">A. S.
            Aldawood</named-content>
Leg.);
      </tp:collecting-event>
</tp:material-citation>
```

Besides GBIF issuing an occurrence ID for the material citations, and Pensoft's Biodiversity Data Journal, no other publishers are using IDs for material citation so far. For EJT and the journals of the MNHN Paris, Plazi is adding the material citations attribute after extracting the data from the published papers.

In legacy publication annotations, material citations are attributed with a unique UUID in TreatmentBank. These UUIDs are resolvable via Plazi SRS[26], and are included in the Darwin Core Archive submitted by TreatmentBank to GBIF where they are reused in combination with the parent taxonomic treatment UUID as identifiers for the published material citation.

The TreatmentBank UUID for the material citation is reused in GBIF as a couple of treatment UUID * material citation UUID:

```
Identifier =
03A10B47FFDFFFAFFDE0FA60FB18F865.mc.3B60B00CFFD1FFAFFF38FED7FD4AFE22
```

In the Biodiversity Data Journal, the material citations are exported to Darwin Core Archive and indexed by GBIF automatically on the date of publication. The internal material citation UUID is minted and entered in the "occurrenceID" of Darwin Core. If the "occurrenceID" is already occupied by the original ID supplied by the author, it should be moved to the "associatedOccurrences" field of Darwin Core, while the "occurrenceID" field should be used again for the internal material citation ID provided by the journal.

Example from BDJ:

```
<tp:treatment-sec sec-type="materials">
      <title>Materials</title>
      <list list-type="alpha-lower" list-content="occurrences">
            <list-item>
                  <p>
                        <bold>Type status:</bold>
                              <named-content
                         content-type="dwc:typeStatus">Holotype</named-conten
                         t>
.
                        <bold>Occurrence:</bold>
catalogNumber:
                              <named-content
                         content-type="dwc:catalogNumber"><ext-link
                         xlink:href="[url]">IFRD9449</ext-link></named-conten
                         t>
; recordedBy:
                              <named-content
                         content-type="dwc:recordedBy">Liu
                         Yu-Wei</named-content>
; occurrenceID:
```

[26] https://tb.plazi.org/GgServer/srsStats

```
                              <named-content
                          content-type="dwc:occurrenceID">UUID</named-content>
; associatedOccurrences:
                              <named-content
                           content-type="dwc:associatedOccurrences">living
                           culture IFRDCC3104</named-content>
                    </p>
               </list-item>
          </list>
</tp:treatment-sec>
```

Recommendation

Publishers should use unambiguous separators, such as a Unicode character U+2022 "•", for the material citations within an article and identify these with UUIDs in the backend article JATS XML. When material citations represent a holotype or other type specimens, this specific status, the collecting event and the collection should be tagged unambiguously in the backend XML to facilitate harvesting and reuse.

Taxonomic names

Definition

A taxonomic name, or more generally scientific name, is the formal name, that is the scientific identity, given to a species or, more generally, a taxon, following the rules of nomenclature and used widely beyond taxonomy to link data to a particular taxon. Although the concept of scientific names, along with rules on the interrelationships of taxa, was introduced in the ancient times by Aristotle (c. 350 BC), and subsequently by Voultsiadou *et al.* (2017), binomial names were introduced by Linnaeus in 1753 and since then, have served as a precursor to today's persistent identifiers. Taxonomic names play different roles inferred by their position in a publication. In other words, the context of their use defines their role. A taxonomic name in the treatment's nomenclature section is the nominate taxonomic name of that treatment. A taxonomic name used in a treatment citation of an existing treatment relates that earlier treatment to the nominate treatment, and represents its taxonomic history; it can also be accompanied with a label indicating nomenclatural changes such as a synonymy or a new combination. These can be nomenclatural acts or subjective synonyms. Any mention of a taxon name in any other section of the article is regarded as a Taxon Name Usage (TNU).

Identifiers for new taxa descriptions and other types of nomenclatural acts, and their on-line registration, are used increasingly, and the process is regulated by zoological (ICZN) and botanical (ICN) codes (Ride *et al.* 2012, Turland *et al.* 2017). Currently registration of nomenclatural acts, other than new taxa descriptions, as a part of a valid publication, whether electronic or print, is mandatory only in mycology including palaeomycology. Registration of identifiers in other disciplines is mandatory only for new taxa descriptions but not yet for other nomenclatural acts. It is, however, planned for implementation (Barkworth *et al.* 2016a, b).

The Catalogue of Life (COL) consortium, in a collaboration with the Global Biodiversity Information Facility (GBIF), aims to provide a global list of accepted names (Garnett et al. 2020; Hobern et al. 2021) by using a combination of automated and manual integration of existing checklists including large scale checklists such as WoRMs, as well as checklists originating from individual taxonomic publications submitted to GBIF. At the moment, COL provides persistent identifiers for taxonomic names but not for taxon concepts. However, WoRMS provides persistent identifiers for each available name, including higher taxa, in its infrastructure, Aphia. Aphia uses Life Science Identifiers (LSIDs) as unique and stable identifiers. TreatmentBank provides a persistent identifier for each available name annotated in the nomenclature section of a taxonomic treatment in legacy literature, both for new

taxa or re-descriptions. Taxon concept identifiers are planned as part of ChecklistBank, a repository and index for taxonomic data. The taxonomic name for a taxon, which can include a large number of taxonomic name usages (e.g. synonyms), are separated from their role in nomenclature (Hobern *et al*. 2021) and in a subsequent section in the treatment after the nomenclature section.

The National Centre for Biotechnology Information (NCBI) taxonomy database holds unique identifiers (taxIDs) for taxonomic names for which sequence data is available at the INSDC (Schoch *et al.* 2020). All records at INSDC have their taxonomic information linked to the NCBI taxIDs. This database, however, does not comprise a complete list of taxonomic names. BOLD taxonomy browser also contains entries for taxonomic names, with associated identifiers. The ChecklistBank allows mapping of these identifiers to the entries in COL.

Identifiers for new names or nomenclatural acts

Fungi

Pre-publication registration of identifiers for names, typifications and other nomenclatural acts is mandatory for fungi since 1st January 2013. The identifiers must be published in the protologue or in nomenclatural changes.

Living vascular plants: IPNI (International Plant Names Index)

In botany the registration of nomenclatural acts was accepted at the XIX International Botanical Congress in Shenzhen 2017 (Turland *et al.* 2017a, b).

Post-publication indexing is a well-established practice of the IPNI which covers seed plants, ferns and lycophytes, but not bryophytes or algae. IPNI is produced collaboratively by The Royal Botanic Gardens, Kew, The Harvard University Herbaria, and The Australian National Herbarium and is hosted by the Royal Botanic Gardens, Kew. Pre-publication indexing and inclusion of IPNI record identifiers in the publication was first implemented by the Pensoft journal PhytoKeys (Penev *et al.* 2016), and later on by *EJT*. IPNI provides nomenclatural information (spelling, author, types and first place and date of publication) for the scientific names of non-fossil vascular plants from family down to infraspecific ranks, including an index of authors for all the groups under the International Code of Nomenclature for algae, fungi, and plants (ICNafp).

Algae

PhycoBank is the registration system for nomenclatural acts (new names, new combinations and types) of algae (Kusber *et al.* 2019). However, the registered identifiers are not required to be listed in the original publication.

Fossil plants (except for fossil fungi and diatoms)

Pre-publication indexing is established in the Fossil Plant Names Registry (FPNR) and the International Fossil Plant Names Index (IFPNI). Registration of taxa is not mandatory.

Bryophytes

IDs for new bryophyte names can be obtained from the Index of Mosses Database (W³MOST).

Animals

ZooBank provides registration of new nomenclatural acts, published works, and authors. It is an authoritative online, open-access, community-generated registry for zoological nomenclature provided as a service to taxonomists, biologists, and the global biodiversity informatics community. It is also the *official register* of the International Commission on Zoological Nomenclature (ICZN).

The registration of Type Specimens is allowed in Zoobank but yet not fully implemented. Registration is mandatory for electronic publications publishing new nomenclatural acts since 1st January 2012. Each electronic publication receives an identifier (LSID) minted by ZooBank.

Identifiers for taxa in Catalogue of Life, NCBI taxonomy, and TreatmentBank.

The Catalogue of Life and the NCBI taxonomy are two widely used reference taxonomies. Both issue taxon name IDs. For references in articles, authors can use hyperlinked taxon IDs of either COL or NCBI just as they use sequence accession numbers.
TreatmentBank mints persistent identifiers for taxonomic names as part of the annotation and FAIRizing of treatments in legacy literature. They are a combination of the treatment UUID extended with ".taxon".

How to discover identifiers of names

The following web sites provide the search facility for discovering the identifiers of names.

Fungi
https://www.mycobank.org/

Living vascular plants
International Plant Names Index (IPNI)

Algae
https://www.phycobank.org /

Fossil plants (except fossil fungi and diatoms)

IFPNI (International fossil plant names index)

PFNR (Plant Fossil Names Registry)

Animals
Zoobank.org
Identifiers of nomenclatural acts can also be found through other services, for example the World Register of Marine Species www.marinespecies.org or uBio http://ubio.org/

Catalogue of Life
https://www.checklistbank.org/tools/name-match

Via Catalogue of Life UI (advanced name search):
https://www.catalogueoflife.org/data/search?q=Pardosa+zyuzini

Via Catalogue of Life API (name usage search):
https://api.catalogueoflife.org/dataset/3LR/nameusage/search?q=Pardosa+zyuzini

Via GBIF UI (species search):
https://www.gbif.org/species/search?q=Pardosa+zyuzini (results include list of other identifiers)
Via GBIF API (species search):
https://api.gbif.org/v1/species?name=Pardosa+zyuzini (requires further matching to pick desired identifiers, e.g. ZooBank UUID of name)

NCBI taxonomy
https://www.ncbi.nlm.nih.gov/taxonomy

TreatmentBank
Identifiers for taxonomic names can be found using TreatmentBanks stats https://tb.plazi.org/GgServer/srsStats

How to mint an identifier

Fungi
Mycobank is an on-line database aimed as a service for the mycological and scientific community by documenting mycological nomenclatural novelties, that is, new names and combinations, and associated data such as descriptions and illustrations.

Index Fungorum, the global fungal nomenclator coordinated and supported by the Index Fungorum Partnership, contains names of fungi including yeasts, lichens, chromistan fungal analogues, protozoan fungal analogues and fossil forms, at all ranks. As a result of changes to the ICN relating to registration of names, Index Fungorum provides a mechanism to register names of new taxa, new names, new combinations and new typifications.

The third index that provides registration of nomenclatural acts for fungi is the Fungal Names maintained in China.

Authors of novel fungal taxa must register the new names in only one registry, e.g. either in MycoBank or Index Fungorum or Fungal Names. These registries regularly coordinate sharing of data and have arranged an informal agreement to only accept the first listed name in case it appears in more than one registry. Registration of the same new name in multiple registries is considered an inappropriate practice that creates a considerable amount of confusion and extra work for the registries and necessitates the deprecation of the duplicated registrations at a later stage.

Living vascular plants
IPNI (International Plant Names Index) uses LSIDs as unique identifiers for plant names and provides a mechanism to register those LSIDs. IPNI records LSIDs for names of new taxa, new combinations and replacement names for living and vascular plants. LSIDs are not mandatory for valid publication of a plant name. However, if an IPNI LSID is needed, it can be pre-registered on the IPNI website. For new taxa, the holotype data can also be provided. The new plant name will be provided with a LSID that will be activated once the article is published. It is important to note that IPNI can only provide LSIDs for “vascular plants”, i.e., extant ferns, lycophytes and seed-bearing plants. Thus, IPNI will not give LSIDs for fungi, bryophytes (mosses), macroalgae (Rhodophyceae etc.), diatoms, or any fossil vascular plant.

Algae

PhycoBank is the registration system for nomenclatural acts such as new names, new combinations and types of algae (Kusber *et al.* 2019). However, it is not required as a part of valid publication. PhycoBank provides a user interface for curatorial and voluntary data entry. Each nomenclatural act according to the provisions of ICN Art. 7 is identified by a stable http identifier that links directly into the PhycoBank portal. The identifier is generated automatically when a reference is linked to a scientific name. Preparation of a record can be done while the manuscript is in the review process. If the preparation is not public, a registration identifier in a manuscript will return the status 'in preparation'. Curation can be done once the publication is finalised and reference details like page numbers and volume are available. The registration can be published on PhycoBank once the scientific paper is published.

Fossil plants (except fossil fungi and diatoms)

PFNR (Plant Fossil Names Registry) is a database of preferably new names, but also previously published names of plant fossils and associated nomenclatural acts excluding fossil diatoms and fossil fungi . It is run by the National Museum Prague for the International Organisation of Palaeobotany. An LSID links the name to its original publication. The registration of a new nomenclatural act results in a registration number that is added to the manuscript. This part is not public and, if necessary, all data can be changed during manuscript processing. These data are available only to the account owner who registered the manuscript, and to the editors of the database. When the paper is published, the missing data should be added and completed. A more detailed guide for name and typification registration is available.

IFPNI (International fossil plant names index) is a comprehensive literature-based record of the scientific names of all fossil plants, algae, fungi, allied prokaryotic forms, protists (ambiregnal taxa) and microproblematica. IFPNI provides an authoritative online, open-access, community-sourced registry of fossil plant nomenclature as a service to the global scientific community. A dynamic database documents all nomenclatural novelties including new scientific names of extinct organisms and associated data, including registration of the scientific publications containing nomenclatural acts and author-generated taxonomic literature in palaeobotany and palaeontology. IFPNI issues LSIDs for each kind of data object to locate biologically significant data over a network. LSIDs are designed to be automatically machine resolvable. Read more about IFPNI coverage.

Animals

To obtain an LSID for a new publication or a new name, the article has to be pre-registered in Zoobank by filling in a form with all the metadata: type of publication, article or monographs in a series, date of publication, authors, full title, ISSN of the journal, DOI of the article, volume, number, pages, online archive (Penev *et al.*. 2016). Tutorials are available online on the Zoobank website to register a publication, a new name, an existing record, etc.

How to annotate and cite a taxon name

In JATS/TaxPub

```
<object-id  object-id-type="Taxon name service">[taxonomic name
identifier]</object-id>
```

Examples of annotations

Fungi

The new fungal species *Acrocalymma chuxiongense* Y. W. Liu & X. Y. Zeng, sp. nov., published in BDJ (Liu Y-W, Zeng X-Y 2022, https://doi.org/10.3897/BDJ.10.e89635) has the identifier "MycoBank 844399" which resolves to the MycoBank record for this new name, available after logging in MycoBank.

In the article JATS XML, this record is annotated as:

```
<tp:taxon-treatment>
      <tp:nomenclature>
            <tp:taxon-name>
                  <object-id
            content-type="arpha">C9A75F0A-E0D9-56CA-94F3-B5F6D1C6A8B0</object
            -id>
            <tp:taxon-name-part
      taxon-name-part-type="genus">Acrocalymma</tp:taxon-name-part>
            <tp:taxon-name-part
      taxon-name-part-type="species">chuxiongense</tp:taxon-name-part>
            <object-id  object-id-type="MycoBank">844399</object-id>
      </tp:taxon-name>
      <tp:taxon-authority>Y. W. Liu &amp;amp; X. Y. Zeng</tp:taxon-authority>
      <tp:taxon-status>sp. nov.</tp:taxon-status>
  </tp:nomenclature>
</tp:taxon-treatment>
```

Living vascular plants

The new plant species *Ardisia whitmorei* Julius & Utteridge, sp. nov. Is published in PhytoKeys (Julius & Utteridge 2022, https://doi.org/10.3897/phytokeys.204.86647) and bears the IPNI ID in the protologue: urn:lsid:ipni.org:names:77302868-1. The IPNI ID is directly linked to the IPNI record.

In the article JATS XML, this record is annotated as:

```
<tp:taxon-treatment>
      <tp:nomenclature>
            <tp:taxon-name>
            <object-id
      content-type="arpha">7A38F1AF-338B-5EAE-BE17-7FE4DD689BD9</ob
      ject-id>
            <object-id
      content-type="ipni">urn:lsid:ipni.org:names:77302868-1</objec
      t-id>
           <tp:taxon-name-part taxon-name-part-type="genus"
reg="Ardisia">Ardisia</tp:taxon-name-part>
           <tp:taxon-name-part taxon-name-part-type="species"
reg="whitmorei">whitmorei</tp:taxon-name-part>
            </tp:taxon-name>
      </tp:nomenclature>
</tp:taxon-treatment>
```

Algae

The new algae *Pseudostaurosira occulta* E.Morales, C.E.Wetzel & Ector, published in PhytoKeys (Morales, Wetzel and Ector 2021, https://doi.org/10.3897/phytokeys.187.73338) has the "PhycoBank 100352103289" identifier which resolves to the PhycoBank record for this name.

In the JATS XML of the article, this record is annotated as:

```
<tp:taxon-treatment>
      <tp:nomenclature>
            <tp:taxon-name>
                  <object-id
            content-type="arpha">C499B059-A1A2-5442-872F-AA446F6CBEE
            </object-id>
                  <named-content content-type="phycobank"
            xlink:href="http://phycobank.org/103328">Phycobank
            103328</named-content>
                  <tp:taxon-name-part taxon-name-part-type="genus"
            reg="Pseudostaurosira">Pseudostaurosira</tp:taxon-name-p
            art>
                  <tp:taxon-name-part taxon-name-part-type="species"
            reg="occulta">occulta</tp:taxon-name-part>
            </tp:taxon-name>
            <tp:taxon-authority>E. Morales, C.E. Wetzel &amp;
      Ector</tp:taxon-authority>
            <tp:taxon-status>sp. nov.</tp:taxon-status>
      </tp:nomenclature>
</tp:taxon-treatment>
```

Animals

A new species of giant *Eunice*, *Eunice dharastii* published in Zanol and Hutchings (2022, https://doi.org/10.3897/zookeys.1118.86448). The taxonomic name resolves at Zoobank.

In the article, the JATS XML is annotated as follows:

```
<tp:nomenclature>
      <tp:taxon-name>
            <object-id content-type="arpha"
      object-id-type="UUDI">BFC45050-4831-5B2D-9E7C-F99DC97D5DF1</o
      bject-id>
            <object-id content-type="zoobank"  object-id-type="UUID"
      xlink:href="https://zoobank.org/63BC2367-9654-45DA-8021-FD175
      84DFFD" >63BC2367-9654-45DA-8021-FD17584DFFDC</object-id>
            <tp:taxon-name-part taxon-name-part-type="genus"
      reg="Eunice">Eunice</tp:taxon-name-part>
            <tp:taxon-name-part taxon-name-part-type="species"
      reg="dharastii">dharastii</tp:taxon-name-part>
      </tp:taxon-name>
      <tp:taxon-status>sp. nov.</tp:taxon-status>
 </tp:nomenclature>
```

Expression of links to taxon names in JATS from the Catalogue of Life and NCBI Taxonomy

The link to *Formica rufa* in the Catalogue of Life is as follows:

```
<object-id
content-type="COL">https://www.catalogueoflife.org/data/taxon/6JGM9
</object-id>
```

The link to *Formica rufa* taken from the NCBI taxonomy

```
<object-id
content-type="NCBI">https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/w
wwtax.cgi?id=258706</object-id>
```

Links to other taxon specific catalogues can be added by entering a respective `content-type`.

Linking a taxonomic name anywhere in the text to the taxonomic name in the respective treatment, an `id` and `<xref rid>` element can be included.

```
<tp:taxon-treatment>
      <tp:nomenclature>
            <tp:taxon-name id="tn1">Formic rufa <object-id
      object-id-type="UUID">C41B87B7-C473-EE75-A7E8-8565FCE1A945.ta
      xon</object-id>
            </tp:taxon-name>
      </tp:nomenclature>
      ...
</tp:taxon-treatment>
```

In a following section of the same article

```
<sec>
      <title>Conclusions</title>
      <p>
In this study,
      <xref rid="tn1">
            <tp:taxon-name>Formica rufa<object-id
      content-type="UUID">C41B87B7-C473-EE75-A7E8-8565FCE1A945.taxo
      n</object-id></tp:taxon-name>
      </xref>
is different from Formica rufa sensu Collingwood, 1966.
      </p>
</sec>
```

Recommendation

Provide a pre-publication registration and include identifiers of new taxa or nomenclatural acts in the original article whenever possible, even when this is not required by a Code. Where and how to register new taxa and get identifiers for different groups of organisms such as algae, fungi, plants, or animals is explained in the sections above.

Specimens

Definition

Physical specimens held in collections may be cited directly, for example, material citations as part of taxonomic treatments, or in other sections of the article. In other cases, data derived from the specimens such as genetic sequences may include a reference to the specimen source. To keep track of the use of these specimens, collections should assign them with at least locally, but better with globally unique IDs (catalogue numbers)[27]. For this reason, the Darwin Core (DwC) triplet comprising the catalogue number, collection code and institution code is often used to assign an *ad hoc* PID to a specimen. However, while the DwC triplets are used commonly, they are far from perfect as these codes are poorly standardised and can change over time (Guralnick *et al.* 2014).

Specimens are often cited by combinations of metadata other than the DwC triplet, such as a who-what-when-where combination, e.g. a specimen "X", collected in locality "Y" by collector "Z" on date XX-YY-ZZZZ, belonging to Taxon A, identified by Person "B". This may include names of the person(s) who collected it, where and when this happened and/or a taxonomic identification. A field number may also be used, which acts as a unique identifier for the collection event as minted by the collectors. These numbers are not unique beyond this narrow context and may not have a systematic syntax. The combination of these properties may allow a specimen to be uniquely identified, but this is not a trivial task and natural language processing as well as disambiguation efforts are required.

Increasingly, the aim is to keep track of physical specimens through digital twins, called Digital or Extended Specimens (Hardisty *et al.* 2019, Lannom *et al.* 2020, Addink and Hardisty 2020). These twins will be minted with a PID, such as a DOI, which can be used to reference the specimen in publications. The Digital twin itself still needs to maintain a relationship to its physical source, such as a CETAF identifier used for a physical specimen, but this is done at the level of the Digital Specimen, rather than through citations in a publication. Once Digital Specimens and their PIDs become available, authors should use these to cite specimens whenever they mention these in the text.

What are the identifiers for specimens

Darwin Core "triplet" of Institution Code, Collection Code and Catalogue Number

The Darwin Core triplet is a concatenation of three Darwin Core properties associated with the physical specimen:

- institutionCode: a code that is commonly associated with the institution where the specimen is held. Often, this is the acronym or one of the acronyms of this institution's name, either in English or in the native language. In botany, codes from the Index Herbariorum[28] are commonly used.
- collectionCode: a code that describes a collection held at the institution indicated above. Institutions may curate multiple collections at different locations and/or with different underlying themes, such as higher taxonomy or geography.
- catalogNumber: a (mostly) alphanumeric code, often a barcode, that is used by the curator of the specimen to uniquely identify it.

By combining institutional provenance and locally used identifiers, the triplet is a simple solution to turn specimen metadata into a globally unique identifier. While it is opaque to the specimen's data, such as the who-what-when-where properties for the event when it was collected, it is constructed of human-readable metadata elements that do not necessarily require a resolver. Triplets have been

[27] CETAF Stable Identifiers https://cetaf.org/resources/bestpractices/
[28] Thiers, B. M. (updated continuously). Index Herbariorum. http://sweetgum.nybg.org/science/ih/

adopted for this reason to different extent by various infrastructures and their data providers. At GBIF, for example, providers of specimen data still regularly publish data using triplets as occurrence IDs. The infrastructure itself uses the triplet as a fallback measure to keep track of updates to occurrence records for specimens, if the combination of the ID for the data provider and the occurrence ID is not sufficient. At INSDC, guidelines[29] recommend data providers use triplets as unique identifiers for voucher specimens.

Triplets are concatenations that function unambiguously as machine-readable identifiers only if they are concatenated consistently using the same methodology, which is not always the case. In practice, the three elements may be separated by different characters, including ":", "/", "<" and "-". The institution and collection elements may change, as institutional branding or internal organisation evolve. The use of simple acronyms and other codes carries the risk of introducing homonymous triplets.

Catalogue or Specimen Number

CETAF stable HTTPIdentifiers

The CETAF specimen identifier concept was established to create a general PID for physical specimens in CETAF collections, rooted in the concept of Semantic Web[30] (Güntsch *et al.* 2017). These identifiers resolve to http URIs which can redirect to human or machine-readable resources such as RDF or JSON, with data on the specimen. To date, more than 29 institutions have implemented the CETAF identifier specification to various extents[31]. They are used on institutional portals and published to GBIF, but are only rarely used elsewhere, such as in a pilot by Güntsch et al. (2021).

CETAF IDs are globally unique by virtue of the Domain Name System, but require institutional investment and policy to ensure their persistence, as domain names may change and they need to be opaque to any potential technical modifications to the infrastructure hosting them. Failure to accommodate these requirements may lead to link rot. Additionally, CETAF IDs have the disadvantage of not being easily discoverable.

Digital Specimen PIDs

The Digital Specimen concept was coined in 2019 during the Biodiversity Next meeting in Leiden and will represent "*a digitised physical specimen, containing information about a single specimen with links to related supplementary information*" (Hardisty *et al.* 2019). Currently, the vision of Digital Specimen accepted as a ground stone of the DiSSCo Research Infrastructure, leads to a global level implementation through becoming a new TDWG standard to be aligned with the vision for Extended Specimens, developed in the USA by the iDigBio project and others (Addink & Hardisty 2020). The practical employment of the Digital Specimens is going through establishing a global Registration Agency using DOIs and the Handle system, developed within the BiCIKL and DiSSCo Prepare projects.

Physical object ID

The PhysicalObject is a DataCite type referring to an inanimate, three-dimensional object or substance used for artefacts or specimens[32]. It has been implemented by the BLR project in Zenodo[33] and first used in 2022 (Boschert & Dikow 2022), including custom metadata using DwC and Audubon core vocabulary terms. It can be cited using a DOI.

[29] https://www.insdc.org/submitting-standards/feature-table/
[30] https://cetaf.org/resources/best-practices/cetaf-stable-identifiers-csi-2/
[31] https://cetaf.org/resources__trashed/best-practices/cetaf-stable-identifiers-csi-2/
[32] https://support.datacite.org/docs/schema-mandatory-properties-v41
[33] https://zenodo.org/communities/biosyslit/search?page=1&size=20&type=physicalobject

How to discover specimen identifier

The DOIs assigned to Digital Specimens will be found through the registration agency's website and other supporting tools, such as ReFindit. The DOIs for the "Physical Object" can be found through ReFindit.

How to mint a specimen identifier

Catalogue numbers should be linked to the physical specimens, often through bar- or QR-codes. They need to be physically minted by the curating institutions.

CETAF identifiers are also minted by the institutions, but require an IT infrastructure that implements the specification, and an institutional pledge to keep the URIs resolving even if the underlying infrastructure is changed.

Digital Specimen PIDs (DOIs) are minted by the authorised natural history institutions, including national focal points and submitted for registration to a centralised registration agency, maintained by an international consortium of stakeholders.

Physical Object PIDs (DOI) can be minted using the Zenodo repository.

How to annotate and cite them

Specimen IDS should be cited either through their inclusion in the specimen record, as with material citations, or as a standalone in-text citation as follows: "Based on Specimen: [hyperlinked physical specimen PID or digital specimen DOI], where a physical specimen PID can be any resolvable PID including e.g. CETAF ID, ARK, DOI., I conclude that….", in tabular format, or alternatively, in supplementary material. The latter, however, is not preferred because of technical limitations to find and extract the data subsequently.

Annotate in JATS:

```
<named-content content-type="issuing institution specimen
identifier" xlink:href="specimen http URI">institution
ID</named-content>
```

Examples

In the article of Patterson et al. (2020) in ZooKeys, all studied specimens are listed in Appendix 1 with their local specimen catalogue numbers. Whenever the collection data is available also in GBIF, then the catalogue number is hyperlinked to the corresponding GBIF record. For example the specimen of the bat *Hipposideros ater* is preserved at the University of Kansas Museum under catalogue number KU 164242. This catalogue number resolves to the GBIF record of this specimen: https://www.gbif.org/occurrence/686491354, which, in turn, contains the original catalogue number of the specimen. Although such a practice is "better-than-nothing", ideally, the specimenID should be a globally unique, persistent, resolvable identifier (GUPRI) which would resolve to the digital specimen serving as a Fair Data Object (FDO) for the physical specimen (see above).

In the article XML, this annotation is expressed as follows:

```
<named-content content-type="cetaf_specimen_id"
xlink:href="https://uri_prefix/CETAF_ID">CETAF_ID</named-content>
```

Recommendation

Use specimen identifiers whenever possible, especially when they are persistent; introduce the practice of citing a specimen, analogously to INSDC accession numbers; keep the IDs separate from HTTPs in the backend XML; use Digital Specimens DOIs when they become available; authors should be encouraged to cite specimens through their IDs.

Whenever CETAF identifiers are available, authors should use them to cite specimens rather than combinations of variable specimen properties.

In case CETAF identifiers are not available, we recommend using the local catalogue number of a specimen with explicit mention of the Collection and/or Institutional Codes where the specimen is preserved.

Sequence data

Definition

Nucleotide sequence data has become fundamental in both basic and applied areas of research related to biology and living organisms. This includes DNA and RNA sequences, genomes and transcriptomes with optional annotations, metagenomes and raw sequence data among other data types.

Sequence data is usually submitted by researchers to public sequence repositories, such as the International Nucleotide Sequence Database Collaboration (INSDC) and cited in publications through their accession numbers in the INSDC databases such as GenBank, ENA, and DDBJ (see below for details). The sequence data is synchronised between the databases using the same accession number but with different prefixes. In ENA, the human readable access is by using the prefix https://www.ebi.ac.uk/ena/browser/view/ and the machine operable version by https://www.ebi.ac.uk/ena/browser/api/embl/. In NCBI it is https://www.ncbi.nlm.nih.gov/nuccore/, and in DDBJ it is http://getentry.ddbj.nig.ac.jp/getentry/na/".

What are the identifiers of gene sequences

INSDC accession numbers

The International Nucleotide Sequence Database Collaboration (INSDC) is a global initiative committed to sharing sequence data and its associated metadata. This collaboration includes three nodes, the DNA Data Bank of Japan (DDBJ), the European Nucleotide Archive (ENA) at EBI and GenBank at NCBI, that comprise the largest repositories of nucleotide sequence data.

INSDC accession numbers are unique identifiers assigned to data submitted to INSDC. These are unique and stable alphanumeric codes that identify each sequence, sample or project, and that also provide information about the type of data and the INSDC partner to which it was submitted[34]. Accession numbers resolve to the data for a particular sequence in the database it has been submitted to.

BOLD Process IDs

The Barcode of Life Database (BOLD) is a platform developed at the Centre for Biodiversity Genomics in Canada for the storage and analysis of barcode sequence data. BOLD Process IDs are unique

[34] https://ena-docs.readthedocs.io/en/latest/submit/general-guide/accessions.html

identifiers in the BOLD system, created to connect specimen metadata such as taxonomy, collection information and images, to the DNA barcode sequence. These have a standard format that includes the project code and a numeric code, followed by the year the record was submitted. The sequence ID corresponds to the Process ID followed by the genetic marker code sequenced.

All public sequences of the BOLD database are periodically mirrored to GenBank, so the public BOLD sequence IDs are associated with INSDC accession numbers.

How to discover identifiers

Sequence data identifiers can be searched for in the respective sequence databases, but these are also linked to other specific molecular databases or portals and even associated with Operational Taxonomic Units (OTUs) such as BOLD Barcode Identification Numbers (BINs), or UNITE Species Hypothesis (SHs), or specimen and distribution data such as occurrences in GBIF.

Via ENA UI (search by taxon name):
https://www.ebi.ac.uk/ena/browser/text-search?query=Formica+selysi

How to mint an identifier

Accession numbers and other sequence identifiers are automatically generated when the data is submitted to the public databases.

How to annotate and cite identifiers

Cite: Sequences, BINs or SHs should be cited either through their hyperlinked persistent identifiers included in the specimen record, that is, the material citation, or in other parts of the articles such as figure legends, tables, appendices or free text, as a standalone in-text citation in the following way, e.g.: “Based on Sequence [hyperlinked accession number], I conclude that….”

Annotate in JATS XML:

```
<named-content content-type="Institution Name"
xlink:href="httpURI">accession name</named-content>
```

Examples

In the main text:
“The following sequences [hyperlinked accession numbers] were used in the analysis… ”

In the Data accessibility section:
“DNA sequences: GenBank accessions xxx to yyy.”
“The data for this study have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEBxxxx (https://www.ebi.ac.uk/ena/browser/view/PRJEBxxxx).”[35]

Sample annotation in JATS for the sequence COI: AM 932788:

NCBI (National Centre for Biotechnology Information)

```
<named-content content-type="NCBI"
xlink:href="https://www.ncbi.nlm.nih.gov/nuccore/AM932788">AM932788
</named-content>
```

[35] https://ena-docs.readthedocs.io/en/latest/submit/general-guide/accessions.html

ENA (European Nucleotide Archive)

```
<named-content content-type="ENA"
xlink:href="https://www.ebi.ac.uk/ena/browser/view/AM932788">AM9327
88</named-content>
```

Recommendation

Publishers should take care not only to include accession numbers in the content but also hyperlink them to the source database and tag them in the backend article XML.

Annotation to only one institution, either ENA or NCBI, should be provided for each accession number.

Persons

Definition

People have different roles in publications, which can often be inferred. The role is identified by the context in which the person's name occurs, which itself is indicated by annotating the section of text within which it occurs. A person appearing in the author section is an author of the publication. A person appearing after a taxonomic name has two roles: as authority of the taxonomic name and also possibly an author of the publication in which the taxon is described. A person whose name appears in a material citation is most likely either a collector or identifier of the cited specimen. A person's name appearing in a short or long bibliographic reference is an author or editor of the cited publication. A person's name in the etymology section is probably honorific, in which case at least part of the name will be Latinized in the taxon name (Article 60.8, Turland *et al.* 2018; Article 31.1, Ride 1999). Hence, a person's unambiguous identity is the key piece of information required, as a person's role can typically be inferred from context. Nevertheless, people's names can suffer from considerable ambiguity and do not resolve on their own. Therefore, it is generally safer to identify people by a persistent identifier that establishes their identity unambiguously, in addition to the context in which their particular role is specified.

There are local and thematic biographical databases of scientists and collectors that provide lists of names, affiliations, birth and death dates and the period a scientist has been active (*floruit)*. In many cases, locally unique identifiers are provided. Such repositories either have only the names of nationals or internationals affiliated to national institutions, while others also refer to non-locals, for example if they are co-authors or collaborate within projects.

Zoobank and IPNI provide LSIDs respectively for zoologists and botanists. For botany the International Plant Name Index (IPNI) also provides a standard form for taxonomic name authorities abbreviations widely used in publishing.

Person names identifiers

ORCID

The Open Research and Contributor ID (ORCID) is an identifier for researchers, with the principal aim of uniquely identifying and connecting them to their publications. It is maintained by the not-for-profit ORCID organisation, which operates through fees paid by member organisations. These organisations are mainly research institutions such as universities and commercial publishers, all of which benefit from widespread ORCID adoption and can make use of the ORCID APIs. ORCID identifiers are a subset of International Standard Name IDs (ISNI), which extend beyond research to other media content creators.

ORCID identifiers have widespread adoption and support, and are easy to register and manage by the researchers themselves. They do suffer from a few downsides: ORCID profiles with scarce metadata and limited or no linked publications are difficult to disambiguate or track back to the person. This may happen when research institutions mandate ORCIDs for their staff, but these new records are only poorly maintained if at all and duplicates may even be created. ORCIDs are also not suitable for deceased researchers, as they are intended to be self-maintained. Finally, ORCID identifiers are intended for use by individuals, not groups, teams or organisations. Also, an individual can register several ORCIDs which adds considerable ambiguity. A particular downside of ORCID is the condition for confirming that only part of the personal data can be made public; despite that this is a GDPR-required condition, it makes the use of the data difficult in some cases. Still, ORCID identifiers are currently the most commonly used means to identify scientific researchers in publications.

ISNI (International Standard Name Identifier)

ISNI is an ISO standard (International Standard Name Identifier, ISO standard 27729) established in 2010, and is widely used to identify people as well as organisations involved in creative activities, and public personas of both such as pseudonyms, stage names, record labels or publishing imprints. The original ISNI database has been populated and is regularly updated from the Virtual International Authority File (VIAF) database. Thus, it is not used exclusively for people. An analysis of a random sample of 10,000 Wikidata items with an ISNI number reveales that about 90% are individuals, and thus ISNI identifiers seem to be widely used[36]. Further analysis is needed to confirm this finding. ISNI identifiers have been used to disambiguate taxonomists (Dillen *et al.* 2021), but are often only sparsely linked to their taxonomic publications.

ISNI is an open standard and its database is populated by harvesting the information from other resources using matching algorithms. The ISNI community would like to promote its usage worldwide, but had to meet the challenges linked to the requirements of the Global Data Protection Regulations (GDPR). A revised version of their data policy was published in March 2021[37]: https://isni.org/resources/pdfs/isni-data-policy.pdf. They have about 37 registered Agencies, including many large and smaller libraries, but also global players like Youtube. The twenty-nine ISNI Members have full access to the ISNI database and the tools or facilities that surround it, including batch and API options for search and ISNI assignment. Members may make ISNI assignment requests for their own needs but are not permitted to act on behalf of other customers or clients outside their organisation. They are thus accessible and re-usable under certain conditions, only to the members and access does not seem open to all.

VIAF

The Virtual International Authority File is a service maintained by the cooperative Online Computer Library Centre (OCLC), a global organisation of libraries. In VIAF, multiple national authority files are compiled into a single authority file where authors are disambiguated. VIAF identifiers may be more appealing for use in taxonomic publications than ISNI identifiers as they can be used to identify authors through their work available in library catalogues, and because of their deliberate overlap (see ISNI).

Wikidata ID

Wikidata is an open graph database hosted by the Wikimedia Foundation. Similar to how Wikipedia was conceived as a community-curated encyclopaedia of all knowledge that is sufficiently notable,

[36] https://w.wiki/5XnQ
[37] https://isni.org/resources/pdfs/isni-data-policy.pdf

Wikidata is the same for linked data. Any Wikidata-notable[38] concept, object, person or organisation can be added to Wikidata, and linked to various characteristics, claims and other associations particular to it. Like Wikipedia, Wikidata can be added to and edited by anyone. This makes it an intriguing registry to reference persons or organisations that fall outside the scope of the databases mentioned earlier, such as deceased researchers or subdivisions of major research organisations. Wikidata is also useful as a broker facilitating interoperability between different databases (Dillen *et al.* 2021).

Taxonomic researchers databases

Numerous databases exist to keep track of taxonomic names, treatments and literature. Many of these have data on people involved in taxonomic research and in observing/collecting specimens, for which person identifiers may be minted. These databases are often under closed curation, but the identifiers may be used to identify people who do not appear in any other system. For instance, LSIDs for authors are used by Zoobank for zoology and the International Plant Names Index (IPNI) for botany. Examples of such databases are also the List of the entomologists of the world on Wikipedia or the Harvard index of botanists.

ResearcherID

The ResearcherID is an identifier for authors, reviewers and editors of scientific publications. It is hosted and maintained by Clarivate, a commercial company that is also responsible for the Web of Science publication index, Endnote bibliography management software, and the Publons review tracking database. As such, this identifier easily connects authors to their work tracked by the Clarivate infrastructures, but paid access to these services is required[39,40].

Scopus Author ID

Scopus is the database of research publications maintained by the Dutch commercial company Elsevier. Author IDs are automatically minted as content enters this database, and may be merged or split as needed or prompted by author feedback forms. The system encourages authors to connect their Scopus Author ID to their ORCID profile, as they can manage the latter themselves[41].

How to discover persons' identifiers

ORCID

ORCID IDs can be found through: https://orcid.org/orcid-search/search?searchQuery=name

ISNI (International Standard Name Identifier)

There is a database search engine to search ISNI identifiers: https://isni.org/page/search-database/

Wikidata IDs

Wikidata provides several ways - ranging from generic to very specific - to find identifiers for people. There is a generic search box available on every Wikidata page, and it can be used for searching with name strings (example). This often yields large numbers of results, including many irrelevant ones, so the query can be refined, e.g., to yield only humans (example) or humans meeting some additional criteria, be it an additional string like "botanist" (example) or an additional identifier like the ZooBank author ID (example) or any additional statement, e.g., a specific place of birth (example).

[38] https://www.wikidata.org/wiki/Wikidata:Notability
[39] https://www.researcherid.com/#rid-for-researchers
[40] https://utas.libguides.com/ManageID/ResearcherID
[41] https://utas.libguides.com/ManageID/Scopus

ResearcherID

Clarivate explains how to search an author identifier in the specific page in the Web of science core collection. However, one has to be registered to access the database.

Scopus Author ID

Use Scopus search engine at:
https://www.scopus.com/freelookup/form/author.uri?zone=TopNavBar&origin=NO%20ORIGIN%20DEFINED

How to mint an identifier

ORCID

In the ORCID model, researchers are invited to register an ID for themselves, and include this ID whenever they publish new work. Many publishers are already ORCID members and technically support easy inclusion of ORCID with author metadata. The combination of this linked body of work and a few pieces of metadata, such as name and (past) affiliation(s), allows unique identification of researchers and facilitates keeping track of their interests, performance and collaborations.

ISNI

The website does not indicate the membership fees upfront. If you want to get an ISNI for yourself, you need to contact the registration agency that provides this service.

Wikidata

Wikidata is an open database, so it is very straightforward to add new records or amend existing ones. Any volunteer can contribute anonymously, in which case IP address will be logged, or through a free registered account. New content needs to comply with community guidelines or may be removed by other volunteers, and moderators can enforce stronger restrictions.

ResearcherID

Researchers can mint their own identifier by registering at https://www.researcherid.com/#rid-for-researchers. ResearcherIDs may also be minted automatically for researchers in the Clarivate system that appear in multiple records but have no ID yet.

Scopus Author ID

Scopus Authors profile's are automatically generated by metadata extracted from documents indexed in Scopus. The profile cannot be edited by the researcher. If correction is required, a request has to be sent to Scopus. Scopus Authors IDs are aligned with ORCID.

How to annotate and cite them

Cite: A person ID is usually cited only in the authors section.

Annotate in JATS:

```
<contrib contrib-type="author" corresp="no">
                                                    <contrib-id
contrib-id-type="orcid">https://orcid.org/0000-0002-0241-1036</cont
rib-id>
                    <name name-style="western">
                        <surname>Liu</surname>
```

```
            <given-names>Yuxin</given-names>
        </name>
        <xref ref-type="aff" rid="A1">1</xref>
    </contrib>
```

Examples

ORCID

0000-0001-5864-8676 is Philippe Bouchet's ORCID

ISNI (International Standard Name Identifier)

0000 0001 2127 4957 is Carl Linnaeus' ISNI

Wikidata

Q1043 is Linnaeus' Wikidata ID.

Recommendation

For persons, ORCIDs are the recommended identifiers if available. If not, ISNI, VIAF. IPNI or Zoobank identifiers could be used if these do exist, which may be possible for people not involved in scientific research or who died before they could create their own ORCID. For any other case, Wikidata is the recommended resource. IPNI or Zoobank identifiers should be added for nomenclature purposes.

Institutions and collections

Definition

The institution is an organisation or infrastructure having custody of the objects included in its holding. The collection or dataset can include specimens of a shared origin, history or collecting campaign, normally part of the activities of an institution. Collection is sometimes also used in the sense of institution, or one institution may have several thematic collections (e.g., Vertebrates, Insects, Non-insect invertebrates), thus often causing confusions between the two terms.

What are the institution identifiers

ROR

ROR is in many ways an equivalent of ORCID for research organisations. It is built on the data of the Global Research Identifier Database (GRID) system, which has a similar scope, but is no longer updated and is currently curated by a commercial company. Like ORCID, ROR aims for an open approach to creation and maintenance of data, operating under community oversight and establishing close links to other infrastructures such as DataCite and CrossRef (Demeranville *et al.* 2021).

To identify research organisations, ROR is currently the most recommended option. However, many research institutions have hierarchical structures, with many different faculties, departments, labs, groups, libraries and archives branching off of a single organisation tree. ROR is intended only for the top-level institution, not any subunit. For legacy reasons, such subunits may still be present in ROR (and GRID) but this is under continuous discussion. For now, adding new 'subunits' to ROR is not recommended. Other solutions should be found for these cases, as some top-level organisations may be vast, e.g., government agencies and big universities, and identification of a lower-level department may be preferable.

GRSciColl

The Global Registry of Scientific Collections (GRSciColl) is a community-curated clearing house of scientific collections[42], hosted by GBIF. The data model[43] underpinning GRSciColl covers basic metadata for Institutions, the collections they hold and the staff who manage them. Content is either 1) curated directly in GRSciColl by a wide pool of editors including the global GBIF Nodes community and projects like iDigBio, or 2) can originate from an external system and further annotated in GRSciColl to associate the entity with additional identifiers for supporting linkages. GRSciColl currently synchronises weekly with Index Herbariorum. Other possible sources of information are dataset metadata and organisations registered in GBIF. GRSciColl entries linked to those sources are updated in real time.

As a clearing house, GRSciColl does not provide a PID on its own, but is able to associate the following identifiers with entities, allowing GRSciColl to act as a collaborative space to link records.

Institutions	GRSciColl ID, institutionCode[44], GRID, ROR, DOI, LSID, CITES and arbitrary IDs such as URL, UUID etc.
Collections	GRSciColl ID, collectionCode[45], Index Herbariorum IRN, DOI, LSID and arbitrary IDs such as URL, UUID etc.
Staff	ORCID ID, Wikidata[46], ResearcherID, HUH, ISNI, VIAF, Index Herbariorum IRN

GRSciColl provides an open API[47] and lookup services allowing for systems integration.

NCBI Biocollections

The NCBI Biocollections database holds curated metadata for institutions and collections, e.g., natural history museums, culture collections or herbaria, associated with sequence records available at the INSDC. It is maintained by the NCBI taxonomy group and used to support the construction of the DwC triplet voucher annotations added to sequence and sample records at the INSDC (Sharma *et al.* 2018).

How to discover an institution's identifier

ROR

Query after name of the institution at: https://ror.org/search?query=name

GrSciColl

Use the search interface ar: https://www.gbif.org/fr/grscicoll/institution/search

NCBI

Search the list of biocollection at: https://www.ncbi.nlm.nih.gov/biocollections/advanced

[42] https://doi.org/10.3897/biss.5.74354
[43] https://github.com/gbif/registry/blob/master/registry-persistence/docs/GrSciColl-db-model.png
[44] http://rs.tdwg.org/dwc/terms/institutionCode
[45] http://rs.tdwg.org/dwc/terms/collectionCode
[46] Besides people, Wikidata covers many institutions and collections too, and identifiers for them can be minted and used in essentially the same way as described for persons above.
[47] https://www.gbif.org/developer/registry#collections

How to mint an identifier for an institution

ROR

ROR works on a community-based curation model and creating a new ROR record or updating an existing one cannot be done directly. Any change can be proposed through a feedback form, after which it will be subjected to community discussion on a Github repository.

GrSciColl

As a community-based curated database, anyone can suggest updates to a GrSciColl record that will be submitted for approval to reviewers that may include institution editors, country mediators, or administrators. GBIF has restored community-curation functionality, enabling those working within each of the institutions and collections, to help maintain up-to-date information in the registry.

NCBI

To register a new collection, send an email to the NCBI contact.

How to annotate and cite identifiers

Cite: An institution can be cited by its abbreviated name, provided it is fully spelled during the first use in the article and linked to a PID to prevent ambiguity.

Annotate in JATS:

```
<named-content  content-type="institution"
xlink:href="https://www.gbif.org/grscicoll/institution/09d260e0-a509-40c7-bef
5-b0f5444be84a" id="NCID0E3CAE">Western Australian Museum</named-content>

<aff id="A1">
      <label>1</label>
      <addr-line content-type="verbatim">Pensoft Publishers, Sofia,
Bulgaria</addr-line>
      <institution>Pensoft Publishers</institution>
      <addr-line content-type="city">Sofia</addr-line>
      <country>Bulgaria</country>
      <uri content-type="ROR">https://ror.org/01znaqx63</uri>
</aff>
```

Examples

The French National Museum of Natural History (MNHN, Paris) has 27 collections identified in GRSciColl. The same institution has a ROR Id (https://ror.org/03wkt5x30) and several other IDs such as ISNI (0000000121749334), GRID grid.410350.3, Crossref Funder ID 501100007522 and Wikidata Q838691.

The institution code for the Western Australian Museum

```
<named-content  content-type="institution"
xlink:href="https://www.gbif.org/grscicoll/institution/09d260e0-a509-40c7-bef
5-b0f5444be84a" id="NCID0E3CAE">Western Australian Museum</named-content>

The institution code for Pensoft Publishers
<aff id="A1">
      <label>1</label>
      <addr-line content-type="verbatim">Pensoft Publishers, Sofia,
Bulgaria</addr-line>
```

```
      <institution>Pensoft Publishers</institution>
      <addr-line content-type="city">Sofia</addr-line>
      <country>Bulgaria</country>
      <uri content-type="ROR">https://ror.org/01znaqx63</uri>
</aff>
```

The institution code for Naturalis, Leiden

```
<named-content content-type="dwc:institutionCode"
xlink:title="National Museum of Natural History, Naturalis"
xlink:href="http://grbio.org/institution/national-museum-natural-history-naturalis">RMNH</named-content>
```

Recommendations

Use ROR for author's affiliations and GRSciColl for specimens.

Encourage institutions to ensure their metadata is up to date in GRSciColl. Ensure that the collectionCode and institutionCode used on specimen records match the one registered in GRSciColl. Make use of the Darwin Core terms collectionID and institutionID on specimen records, populated with the GRSciColl identifier.

Promote the use of GRSciColl identifiers as a means to link entities.

Ensure that all identifiers related to an institution are linked one to another, e.g., make sure all identifiers of the institution are mentioned in its ROR ID.

Back matter

Definition

The article back matter contains information that is ancillary to the main text, such as cited references, acknowledgements, declaration on authors' contributions, declarations on conflicts of interest, declarations on funding, footnotes, supplementary materials, appendices, glossary, etc. The JATS structure for back matter is listed here.

This kind of information can be included either in the article text, or in the article back matter metadata alone, or in both places which may often cause confusion.

Acknowledgements

Definition

Section at the end of the manuscript where the authors thank colleagues or institutions who helped them in their work, e.g., contributed to it in producing some research, provided information, assisted the research, reviewed the manuscript, etc. It also mentions and acknowledges the projects and grants that funded the research. Linking funding agencies, grant numbers and persons to their PIDs or their homepages is potentially a very important source of information to build alternative metrics for individual contributions or the impact of a funding agency.

What are the identifiers

Persons

For persons contributing to the publication, use the contrib tag.

Funding agencies

Funding agencies are the institutions funding research including science foundations, other funding agencies, private charity trusts, or others. Funding agencies normally provide a grant award number that needs to be cited. There are, however, two large international sources that list projects and funders with their identifiers, the EU infrastructure OpenAIRE and the Funder Registry at CrossRef.

How to discover them

Both OpenAIRE and Crossref Funder Registry provide a search interface and API which helps publishers integrate the data about projects and funders with their editorial systems. Wikidata has different properties to acknowledge funding or for general acknowledgements.[48]

How to annotate and cite funding agencies

Funding agencies

A funding agency should be cited by its name, tagged and linked with its identifier to avoid ambiguity and promote findability. The following example includes the agency name, the grant award number as well as a DOI of the agency name.

Person names

See the respective section above.

Examples

(1) Funder: European Commission. Funded project: Building the European Biodiversity Observation Network (EU BON). Grant number: 501100000780.

(2) Funding program: Institute of Museum and Library Services, award # LG-246400-OLS-20, "Extending data curation to interdisciplinary and highly collaborative research" (from Kouper and Cook KJ, 2021, https://doi.org/10.3897/biss.5.79084).

Annotate in JATS:

```
<funding-source>
      <named-content content-type="funder_name">XXX</named-content>
      <named-content
content-type="funder_identifier">XXX</named-content>
      <named-content content-type="funder_doi">XXX</named-content>
</funding-source>
```

Sample annotations in JATS:

```
<funding-source>
      <named-content content-type="project_title">EU BON: Building the
European Biodiversity Observation Network</named-content>
      <named-content content-type="project_identifier">308454</named-content>
      <named-content
content-type="project_funder_id">501100000780</named-content>
      <named-content content-type="project_funder_name">European
Commission</named-content>
```

[48] Sample queries for funding acknowledgments and general acknowledgments: https://scholia.toolforge.org/sponsor/Q304878 and https://w.wiki/5XnJ .

```
      <named-content
content-type="project_funder_doi">http://doi.org/10.13039/501100000780</named
-content>
</funding-source>
<funding-source>
      <named-content content-type="funder_name">Institute of Museum
and Library Services</named-content>
      <named-content
content-type="funder_identifier">100000208</named-content>
      <named-content
content-type="funder_doi">http://doi.org/10.13039/100000208</named-
content>
</funding-source>
```

Recommendation

The Acknowledgment and List of references sections should be included in the manuscript text.

Authors' contributions, supporting funders/projects and declarations of conflict of interest should be part of the machine-readable article back matter metadata. A common and highly recommended way to facilitate the conversion to machine-readable format is to request this information during the submission process through APIs or dropdown list of items for: (1) CrossRef Funder Registry and OpenAIRE databases, from which the authors can select the acknowledged funder and research project; (2) CRedIT taxonomy[49] to formalise the authors contributions, and (3) standardised answers to formalise the question on conflict of interest.

Funder and project names in the backend article XML should be assigned external identifiers, such as Funders' IDs from OpenAIRE or CrossRef Funder Registry and grant numbers, when acknowledging grants.

We recommend authors to fill in supporting grants and funders also in the article metadata during the submission process. The integration of OpenAIRE and CrossRef's Funder Registry during the submission system would disambiguate the grant and funder titles/names which should be included in the article-meta tag of JATS.

Bibliographic references and citations

Definition

The references cited within a paper are listed at the end of the manuscript and usually organised by alphabetical order based on the surname of the first author, and often include the DOI of the publication. They refer to the sources used and quoted in the text. In the humanities, the bibliography often refers to all the materials consulted and sources the authors went through to conceive their research, and are not all necessarily cited in the text. On the contrary, in STM, all references relate to in-text citations. All in-text citations must refer to a reference listed at the end of the article.

Bibliographic citations have a specific use in taxonomy. Indeed, traditionally a taxonomic name is followed by the mention of the author who first described it. A bibliographic citation is implicit in the taxonomic name, and both together, the taxon name and the authority, form the "taxon concept". For example, "L. 1758" in *Formica rufa* L. 1758 is a bibliographic reference to Linnaeus, 1758. Bénichou *et al.* 2018 recommend adding a proper bibliographic reference, such as *Formica rufa*, Linnaeus, 1758:

[49] https://credit.niso.org/

580, and then add the full bibliographic reference in the list of references. This will help facilitate the reader to find the referenced article. From a machine actionable point of view, it will enable building a citation network of the article, and give credit to the authors of the referenced publications. Other publishers such as Pensoft and Magnolia Press have been recommending this for a long time, however it has not been mandatory because of the additional burden on the authors for compiling excessively long reference lists for all taxon names mentioned in a paper.

Another specific use of a citation that has yet to gain wider traction is to indicate the citation context, for which CiTO - the Citation Typing Ontology[50] - provides a well-defined set of options, including that the citing work "agrees with", "extends"or "refutes" the cited work. The citation network can be visualised using Wikidata's Scholia[51].

What are the identifiers for bibliographic references

Digital Object Identifiers

DOIs issued by CrossRef by or on behalf of the publisher are most commonly accepted as a norm by most publishers. BHL provides Crossref DOIs for digitised articles from legacy literature. BLR at Zenodo provides DataCite DOIs for legacy articles which do not have DOIs or for several sub-article elements such as treatments, figures and others.

Handle

Handles are sometimes used by institutional repositories such as the Digital Repository at the American Museum of Natural History (AMNH), however, their use in the publishing world is very limited.

How to discover identifiers

The DOIs can be found using the search form provided by CrossRef or DataCite for their minted DOIs, or Refindit, which includes all the providers of DOIs.

How to mint DOIs

DOIs are minted by the publishers, or on behalf of publishers, and subsequently submitted to CrossRef for registration. DataCite DOIs are minted by the DataCIte partnering organisations and can be taken from there by the data aggregators and publishers.

Use CrossRef to mint DOIs for legacy publications that have no DOI, following their policies for retrospective DOI assignment. One may also use repositories such as BLR at Zenodo that provide free DataCite DOIs. To avoid duplication, it's strongly recommended to carefully check and ensure that the article has not been assigned a DOI before.

How to annotate and cite them

The DOI follows the bibliographic reference and is hyperlinked (see example below)

Annotate in JATS:

```
<ext-link  ext-link-type="doi"
xlink:href="10.5962/bhl.title.542">https://doi.org/10.5962/bhl.title.542</ext-link>
```

Examples

Linnæus C. 1758. *Systema naturae per regna tria naturae, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis*, locis Laurentii Salvii, Holmiae. Vol. Tomus I, Editio decima, reformata Edition: i-ii, 1-824. https://doi.org/10.5962/bhl.title.542

[50] https://sparontologies.github.io/cito/current/cito.html
[51] https://scholia.toolforge.org/cito/

Recommendation

Each bibliographic citation must refer to the full bibliographic reference included in the List of references at the end of the article.

Each bibliographic reference should be complemented with a DOI. If no DOI exists, a DOI should be minted and added following rules and policies for retrospective DOI assignment at CrossRef and DataCite. When minting DOIs for legacy content, always check if a DOI hasn't already been assigned to the same content by someone else.

All DOIs minted by a publisher, or anyone else, should be registered at the corresponding registration agency, CrossRef or DataCite, respectively.

Bibliographic citations in the text (e.g. Linnaeus 1758) should be cross-referenced to their bibliographic reference, mandatorily including DOIs for recently published articles, and hopefully for historical ones, in the backend article XML. This would allow easy harvesting and tracking of citations and discovery of the original literature source behind the reference.

Supplementary material

Definition

Supplementary material is used to add detail, background, or context to an article by providing backend data and information which are not formally part of the manuscript text, for example, multimedia objects such as audio clips and applets, raw data in a spreadsheet, additional XML-tagged sections, tables, or figures, or a source code of a software application in a repository.[52]

What are its identifiers

CrossRef component DOIs, or DataCite DOIs, minted and submitted for registration by the publisher to CrossRef or DataCite, respectively.

How to discover identifiers

DOIs can be located using the search form provided by CrossRef or DataCite for their minted DOIs, or Refindit, which includes all the providers of DOIs.

How to mint them

CrossRef DOIs are minted by the publishers, or on behalf of publishers. DataCite DOIs are minted by the DataCIte partnering organisations and can be taken from there by the data aggregators and publishers.

How to annotate and cite them

Cite: The supplementary material's component DOI is linked to the parent DOI of the article (see example below) but can also be hyperlinked and made accessible independently from the article DOIs. It can be cited using the community accepted citing convention for data which should include its DOI, e.g., https://doi.org/10.3897/bdj.5.e14650.suppl1.

Annotate in JATS:

```
<ext-link  ext-link-type="doi"
xlink:href="https://10.3897/bdj.5.e14650.suppl1">10.3897/bdj.5.e146
50.suppl1</ext-link>
```

[52] https://jats.nlm.nih.gov/archiving/tag-library/1.3/element/inline-supplementary-material.html .

Examples

Van Achterberg K, Taeger A, Blank S, Zwakhals K, Viitasaari M, Yu D, de Jong Y (2017) Supplementary material 1 from: van Achterberg K, Taeger A, Blank S, Zwakhals K, Viitasaari M, Yu D, de Jong Y (2017) Fauna Europaea: Hymenoptera – Symphyta & Ichneumonoidea. *Biodiversity Data Journal* 5: e14650. https://doi.org/10.3897/BDJ.5.e14650. https://doi.org/10.3897/bdj.5.e14650.suppl1

In JATS XML, this example is annotated as:

```
<supplementary-material id="S3638143" orientation="portrait"
position="float">
      <object-id
object-id-type="arpha">52DE64E0-3C7A-52B0-B88B-804A668FA4D4</object-id>
      <object-id object-id-type="doi">10.3897/BDJ.5.e14650.suppl1</object-id>
      <object-id object-id-type="zenodo_dep_id">904725</object-id>
      <label>Supplementary material 1</label>
      <caption>
            <p>
                  <tp:taxon-name>
                        <tp:taxon-name-part
                        taxon-name-part-type="superfamily">Ichneumonoidea</t
                        p:taxon-name-part>
                  </tp:taxon-name>
— Fauna Europaea mapping
            </p>
      </caption>
      <p>Data type: xlsx</p>
      <p>
Brief description: Cross-validation of Fauna Europaea (version 2.6.2) and
            <tp:taxon-name>
                  <tp:taxon-name-part            taxon-name-part-type="genus"
                  reg="Taxapad">Taxapad</tp:taxon-name-part>
            </tp:taxon-name>
(version 2016). Discrepancies are annotated. For details on data ownership
and correct citation please check the Fauna Europaea and
            <tp:taxon-name>
                  <tp:taxon-name-part            taxon-name-part-type="genus"
                  reg="Taxapad">Taxapad</tp:taxon-name-part>
            </tp:taxon-name>
Websites.
      </p>
      <p>File: oo_145108.xlsx</p>
            <media  xlink:href="bdj-05-e14650-s001.xlsx"  mimetype="Microsoft
            Excel     Document"     mime-subtype="xlsx"     position="float"
            orientation="portrait">
            <uri
      content-type="original_file">https://binary.pensoft.net/file/145108</ur
      i>
            </media>
                  <attrib specific-use="authors">Kees van Achterberg, Dicky
            Sick Ki Yu and Yde de Jong</attrib>

</supplementary-material>
```

Recommendation

All metadata of supplementary material should be available in a standard, machine-readable format in the article backend XML and in human-readable citation formats suggested by the publisher at the article webpage.

Use CrossRef component DOI to identify each supplementary material files related to an article. The component DOI has the important feature to link the supplementary files DOI to its parent article DOI. If no CrossRef DOI are available, use DOIs from DataCite.

Conclusions

Over the last decades, the communities and the relevant scientific networks have realised the benefits of using PIDs, and they're developing initiatives to expand their use in other components of the publications. Adoption of identifiers in the scientific literature is a stepwise process and it may happen in two different ways: (1) data and sub-article level content is liberated from publications and identifiers are assigned to it retrospectively, or (2) they are prospectively published where the datum identifiers are provided upfront, at the moment of publication, and are available from the articles themselves, mostly from their published backend XML versions. These identifiers should be aligned through standards and community norms, for which the present paper provides guidance, and then re-used by data aggregators, for example, GBIF, Biodiversity Literature Repository, TreatmentBank, ARPHA-XML, OpenBiodiv, SIBiLS and others. This is leading to a unique and large corpus of FAIR literature data which can be linked to their original data sources or re-used for generation of new knowledge.

Though the development of semantically enhanced publications in taxonomy is much advanced, especially when compared to some much better funded science branches, such as molecular biology or ecology, its wider adoption in the publishing world is still to come due to a combination of technical and sociological issues. The composition and granularity of annotated sub-article elements discussed here are the first steps to providing machine access to the rich data in publications that would facilitate further exploration.

Future steps could include more structured geographic data or species traits which are still to be explored for their suitability for text mining and annotation to allow re-use in ongoing research (e.g. Upham *et al.* 2021). Examples of such initiatives are: (a) the functional trait data, examples of which are the trait data structures, such as the TraitBank[53] (Parr et al. 2016) and the Ecological Trait-data Standard Vocabulary[54] (Schneider et al. 2019); (b) the geographic data, a good example of which is the Marine Regions Gazetteer Ontology and the Marine Regions Geographic Identifiers (MRGID) [55]). These two types of data are also important components of the taxonomic treatments and biodiversity literature overall and our further task is to automate the process of their markup, extraction, annotation and re-use.

To summarise, our recommendations presented in detail in the article text and in a concise form in Appendix 1, are outlined here as best practices to be followed when implementing persistent identifiers in the conversion of legacy literature, and even more importantly, in prospectively published scholarly articles:

[53] https://eol.org/traitbank
[54] https://terminologies.gfbio.org/terms/ets/pages/
[55] https://www.marineregions.org/about.php

1. Persistent identifiers should be used as widely as possible for article metadata and sub-article structural elements and data.
2. Global unique persistent resolvable identifiers (GUPRI) provided by established international organisations such as CrossRef, DataCite, ORCID, INSDC, GBIF, etc., should always be preferred over locally assigned identifiers (UUIDs).
3. Persistent identifiers should be incorporated in the backend article XML to the maximum extent possible. The best practice in doing this is GUPRIs to be assigned as two different properties to an element: (1) as a "plain" UUID, and (2) as a resolvable UUID, that is, including its HTTP prefix. However more than one property for a PID per element should be allowed by the XML schema used. In case only one property per PID is allowed, then it is preferable to use the GUPRI or some other kind of resolvable PID instead of a UUID alone.
4. Assigning a persistent identifier to a named entity adds a semantic layer to it, however semantics and identifications do not always overlap. There are many cases when there is no need to add a PID to a sub-article element, however this element should still be appropriately tagged in the backend article XML. This is because much of the structure of the text, for example that of taxonomic treatments, implies the semantics needed for machine actionability and processing of the content.
5. In the world of biodiversity, assigning PIDs to data or other information entities should be aligned whenever possible with the current Darwin Core vocabulary and terms and other standards accepted by the biodiversity community in the Biodiversity Information Standards (TDWG) process. In Darwin Core coded data, PIDs should be placed consistently in the appropriate data fields intended for their use and management.

The data published in scholarly literature is normally considered as high-quality data, at least because it passes a review process and editorial evaluation and also because scientists associate their names and authority with the quality of data and content they publish. Hence, the data published in the literature is of special value to researchers, and to science as a whole, therefore it needs much more attention and exploration in the Internet-era compared to when text annotation and extraction tools did not exist.

The next important step is to convince more publishers, research infrastructures and biodiversity researchers to follow some or all of these guidelines, to achieve a complete integration of the published literature in the research lifecycle, not only in its usual human-readable form, but even more importantly, as data liberated from the narrative and re-imported back into the data lifecycle. For that goal, it is important that all actors in the research and publishing domains contribute to this process. For example:

1. Publishers need to implement semantic technologies in the publishing process that will facilitate text conversion to structured data. By doing this, they will benefit from higher visibility, citability and re-use of the content they publish.
2. Biodiversity research infrastructures participating in the BiCIKL and, more generally, in the *alliance for biodiversity knowledge* process, should commit to reusing the data extracted from literature by providing linkages between their source and related published data. A good example for that would be if INSDC automatically linked any mention of a particular sequence in the literature, or GBIF linked any mention of a specimen in the literature to its respective specimen record on their infrastructure. By doing this, research infrastructures will enrich their content, link it to other valuable sources of information and benefit their users by providing additional incentives to publish structured data.
3. Researchers, when writing their manuscripts, should use PIDs for data or sub-article elements whenever possible, and insist that publishers keep these PIDs in the published articles. By doing this, they will benefit from far richer sources of data from the published literature

provided to them in a structured form and with less effort, gain higher visibility of their work and hopefully, increased opportunities for collaboration with other researchers.

4. All involved stakeholders, by implementing all or some of the recommendations in their daily work, should anticipate others to follow by demonstrating the impact of the identifiers in place, for example, by providing cases of generation of new knowledge based on data reuse, or by using PIDs to build alternative metrics to measure scientific output.

Acknowledgements

Financial support has been provided by the European Union's Horizon 2020 Biodiversity Integrated Knowledge Library (BiCIKL) project, under grant agreement No 101007492. The following colleagues provided valuable input: Anton Güntsch, Alex Ioannidis, Connie Rinaldo, and Clément Oury.

References

Abrahamse T, Andrade-Correa MG, Arida C, Galsim R, Häuser C, Price M, Sommerwerk N 2021. The Global Taxonomy Initiative in Support of the Post-2020 Global Biodiversity Framework. CBD Technical Series No. 96. Secretariat of the Convention on Biological Diversity, Montreal, 103 pages. (Plazi: participant) doi: https://doi.org/10.5281/zenodo.5728812

Aldawood AS, Sharaf MR 2011. Monomorium dryhimi sp. n., a new ant species (Hymenoptera, Formicidae) of the M. monomorium group from Saudi Arabia, with a key to the Arabian Monomorium monomorium-group. ZooKeys 106: 47-54. https://doi.org/10.3897/zookeys.106.1390

Aristotle (c. 350 BC). *Historia Animalium*. (English translation): http://classics.mit.edu/Aristotle/history_anim.5.v.html

Arita M, Karsch-Mizrachi I, Cochrane G (2021) The international nucleotide sequence database collaboration. Nucleic Acids Research, 49, D121-D124. doi: 10.1093/nar/gkaa967

Agosti D, Egloff W 2009. Taxonomic information exchange and copyright: the Plazi approach. BMC Research Notes 2009, [2:53]. doi: https://doi.org/10.1186/1756-0500-2-53

Barkworth M.E., Watson M., Barrie F.R., Belyaeva I.V., Chung R.C.K., Dašková J., Davidse G., Dönmez A.A., Doweld A.B., Dressler S., Flann C., Gandhi K., Geltman D., Glen H.F., Greuter W., Head M.J., Jahn R., Janarthanam M.K., Katinas L., Kirk P.M., Klazenga N., Kusber W.-H., Kvaček J., Malécot V., Mann D., G. Marhold K., Nagamasu H., Nicolson N., Paton A., Patterson D.J., Price M.J., Prud'homme van Reine W.F., Schneider C.W., Sennikov A., Smith G.F., Stevens P.F., Yang Z.-L., Zhang X.-C. & Zuccarello G.C. 2016b. – Proposals to provide for registration of new names and nomenclatural acts. *Taxon* 65 (3): 656–658. doi: https://doi.org/10.12705/653.37

Barkworth M.E., Watson M., Barrie F.R., Belyaeva I.V., Chung R.C.K., Dašková J., Davidse G., Dönmez A.A., Doweld A.B., Dressler S., Flann C., Gandhi K., Geltman D., Glen H.F., Greuter W., Head M.J., Jahn R., Janarthanam M.K., Katinas L., Kirk P.M., Klazenga N., Kusber W.-H., Kvaček J., Malécot V., Mann D., G. Marhold K., Nagamasu H., Nicolson N., Paton A., Patterson D.J., Price M.J., Prud'homme van Reine W.F., Schneider C.W., Sennikov A., Smith G.F., Stevens P.F., Yang Z.-L., Zhang X.-C. & Zuccarello G.C. 2016a. – Report of the Special Committee on registration of algal and plant names (including fossils). *Taxon* 65 (3): 670 –672. doi: https://doi.org/10.12705/653.43

Bénichou, L., Gérard, I., Laureys, Éric, & Price, M. (2018). Consortium of European Taxonomic Facilities (CETAF) best practices in electronic publishing in taxonomy. *European Journal of Taxonomy*, (475). https://doi.org/10.5852/ejt.2018.475

Bénichou L, Buschbom J, Campbell M, Hermann E, Kvacek J, Mergen P, Mitchell L, Rinaldo C, Agosti D (2022). Joint statement on best practices for the citation of authorities of scientific names in taxonomy by CETAF, SPNHC and BHL. *Research Ideas and Outcomes* 8: e94338: 1-7. https://doi.org/10.3897/rio.8.e94338

Blahnik R, Andersen T (2022) New species of the genus *Chimarra* Stephens from Africa (Trichoptera, Philopotamidae) and characterization of the African groups and subgroups of the genus. In: Pauls SU, Thomson R, Rázuri-Gonzales E (Eds) Special Issue in Honor of Ralph W. Holzenthal for a Lifelong Contribution to Trichoptera Systematics. ZooKeys 1111: 43-198. https://doi.org/10.3897/zookeys.1111.77586

Boschert C, Dikow T 2022. Taxonomic revision of the mydas-fly genera Eremohaplomydas Bequaert, 1959, Haplomydas Bezzi, 1924, and Lachnocorynus Hesse, 1969 (Insecta, Diptera, Mydidae). African Invertebrates 63(1): 19-75. doi: https://doi.org/10.3897/afrinvertebr.63.76309

Bueno-Soria J, Vilarino A, Barba-Alvarez R, Ballesteros-Barrera C (2022) Three new species of Xiphocentron Brauer, 1870 (Trichoptera, Xiphocentronidae) from Mexico. In: Pauls SU, Thomson R, Rázuri-Gonzales E (Eds) Special Issue in Honor of Ralph W. Holzenthal for a Lifelong Contribution to Trichoptera Systematics. ZooKeys 1111: 199-213. https://doi.org/10.3897/zookeys.1111.73371

Catapano T 2010. TaxPub: An Extension of the NLM/NCBI Journal Publishing DTD for Taxonomic Descriptions. Proceedings of the Journal Article Tag Suite Conference 2010 doi: 10.5281/zenodo.3484285

Chester C., Agosti D., Sautter G., Catapano T., Martens K., Gérard I. & Bénichou L. 2019. *EJT* editorial standard for the semantic enhancement of specimen data in taxonomy literature. *European Journal of Taxonomy* 586: 1–22. doi: https://doi.org/10.5852/ejt.2019.586

Cole, Mary L. 2019. Revision of Chondrocyclus s. l. (Mollusca: Cyclophoridae), with description of a new genus and twelve new species, European Journal of Taxonomy 569, pp. 1-92 https://doi.org/10.5852/ejt.2019.569

Deans A.R., Yoder M.J. & Balhoff J.P. 2012. Time to change how we describe biodiversity. *Trends in Ecology & Evolution* 27(2): 78-84. doi: https://doi.org/10.1016/j.tree.2011.11.007 .

Demeranville T., Gould M., Krznarich L., Lammey R., Petro J. & Vierkant P. 2021 (posted on September 2022). Making the world a PIDder place: It's up to all of us!. DataCite Member Meeting 2021, virtual. Zenodo. doi: https://doi.org/10.5281/zenodo.5532631

Dillen M., Groom Q., Cubey R., von Mering S., Hardisty A., Humphries J. Butcher G., Robertson T., Ernst M. 2021. – A best practice guide for semantic enhancement and improvement of semantic interoperability. *DiSSCo Prepare deliverable 5.4 report.* doi: 10.34960/ajxs-zr25

Dimitrova M, Senderov VE, Georgiev T, Zhelezov G, Penev L (2021) Infrastructure and Population of the OpenBiodiv Biodiversity Knowledge Graph. Biodiversity Data Journal 9: e67671. https://doi.org/10.3897/BDJ.9.e67671

Directorate-General for Research and Innovation (European Commission), EOSC Executive Board, Valle M., Heughebaert A., Kotarski R., Weigel T., Ritz R., Matthews B., Manghi P., Sparre C.A., Hellström M., Wittenburg P. 2020a. – A Persistent Identifier (PID) policy for the European Open Science Cloud (EOSC). *Report from the European Open Science Cloud FAIR and Architecture Working Groups.* 20 pages. doi: https://doi.org/10.2777/926037

Directorate-General for Research and Innovation (European Commission), 2020b. – *Solutions for a sustainable EOSC A FAIR Lady (olim Iron Lady) report from the EOSC Sustainability Working Group*. 48 pages. doi: https://doi.org/10.2777/870770

Fawcett S, Agosti D, Cole SR, Wright CF 2022. Digital accessible knowledge: Mobilizing legacy data and the future of taxonomic publishing. Bulletin of the Society of Systematic Biologists 1(1): 1-12. doi: https://doi.org/10.18061/bssb.v1i1.8296

Garnett S.T., Christidis L., Conix S., Costello M.J., Zachos F.E., Bánki O.S., et al. (2020) Principles for creating a single authoritative list of the world's species. *PLoS Bioloby* 18(7): e3000736. doi: https://doi.org/10.1371/journal.pbio.3000736

Gmür R., Agosti D. 2021. Synospecies, an application to reflect changes in taxonomic names based on a triple store based on taxonomic data liberated from publication. Biodiversity Information Science and Standards 5: e75641. doi: https://doi.org/10.3897/biss.5.75641

Groom Q., Güntsch A., Huybrechts P. *et al.* 2020. – People are essential to linking biodiversity data. *Database* (2020). doi: https://10.1093/database/baaa072

Groom Q.J., Dillen M., Huybrechts P., Johaadien R., Kyriakopoulou N., Fernandez F.J.Q., Trekels M., Wong W.Y. 2021. – Connecting molecular sequences to their voucher specimens. *BioHackrXiv*. doi: https://doi.org/10.37044/osf.io/93qf4

Güntsch A, Groom Q, Ernst M, Holetschek J, Plank A, Röpert D, et al. (2021) A botanical demonstration of the potential of linking data using unique identifiers for people. PLoS ONE 16(12): e0261130. doi: https://doi.org/10.1371/journal.pone.0261130

Güntsch A., Hyam R., Hagedorn G., Chagnoux S., Röpert D., Casino A., Droege G., Glöckler F., Gödderz K., Groom Q., Hoffmann J., Holleman A., Kempa M., Koivula H., Marhold K., Nicolson N., Smith V.S., Triebel D.., 2017. Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects, *Database* 2017, bax003. doi: https://doi.org/10.1093/database/bax003

Guralnick R.P., Cellinese N., Deck J., Pyle R.L., Kunze J., Penev L., Walls R., Hagedorn G., Agosti D., Wieczorek J., Catapano T., Page R. - 2015 Community next steps for making globally unique identifiers work for biocollections data. ZooKeys 494: 133-154 doi: https://zenodo.org/record/6945039#.YuksZ3ZBz8A

Hardisty A.R., Addink W., Glöckler F., Güntsch A., Islam S., Weiland C. 2021. A choice of persistent identifier schemes for the Distributed System of Scientific Collections (DiSSCo). *Research Ideas and Outcomes* 7: e67379. doi: https://doi.org/10.3897/rio.7.e67379

Hardisty A.R, Ellwood E.R., Nelson G., Zimkus B., Buschbom J., Addink W., Rabeler R.K., Bates J., Bentley A., Fortes J.A,B,, Hansen S., Macklin J.A., Mast A.R., Miller J.T., Monfils A.K., Paul D.L., Wallis E., Webster M., 2022. Digital Extended Specimens: Enabling an Extensible Network of Biodiversity Data Records as Integrated Digital Objects on the Internet, BioScience, biac060, https://doi.org/10.1093/biosci/biac060

Hobern, D., Barik, S.K., Christidis, L. et al. Towards a global list of accepted species VI: The Catalogue of Life checklist. *Organism, Diversity and Evolution* 21, 677–690 (2021). doi: https://doi.org/10.1007/s13127-021-00516-w

Hyam, R., Drinkwater, R. E., & Harris, D. J. (2012). Stable citations for herbarium specimens on the internet: an illustration from a taxonomic revision of Duboscia (Malvaceae). *Phytotaxa*, 73(1), 17-30. doi: https://doi.org/10.11646/phytotaxa.73.1.4

Kouper I, Cook KJ (2021) Challenges in Curating Interdisciplinary Data in the Biodiversity Research Community. Biodiversity Information Science and Standards 5: e79084. https://doi.org/10.3897/biss.5.79084

Kusber W.-H., Kohlbecker A., Mohamad H., Güntsch A., Berendsohn W.G. & Jahn R. 2019. – Registration of Algal Novelties in Phycobank: Serving the scientific community and filling gaps in the global names backbone. *Biodiversity Information Science* 3, e37285. doi: https://doi.org/10.3897/biss.3.37285

Linnæus, C. 1753. Species plantarum : exhibentes plantas rite cognitas ad genera relatas, cum diferentiis specificis, nominibus trivialibus, synonymis selectis, locis natalibus, secundum systema sexuale digestas. Holmiæ, impensis Laurentii Salvii, 1753. https://doi.org/10.5962/bhl.title.37656

Linnæus, C. 1758. Systema naturae per regna tria naturae, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis Laurentii Salvii, Holmiae. Vol. Tomus I, Editio decima, reformata Edition: i-ii, 1-824. doi: https://doi.org/10.5962/bhl.title.542

Liu Y.-W., Zeng X.-Y. 2022. Acrocalymma chuxiongense sp. nov., a new species of Acrocalymmaceae (Pleosporales) on leaves of Quercus. Biodiversity Data Journal 10: e89635. https://doi.org/10.3897/BDJ.10.e89635

Madden F. & Woodburn M. 2021. Towards a National Collection: Persistent Identifiers as IRO Infrastructure. British Library, 19pp. doi: https://doi.org/10.22020/k99s-we61

Mathis W, Zatwarnicki T (2013) A revision of the shore-fly genus Hydrochasma Hendel (Diptera, Ephydridae). ZooKeys 363: 1-161. doi: https://doi.org/10.3897/zookeys.363.6482

McGill, B.J., Enquist, B.J., Weiher, E., & Westoby, M. 2006. Rebuilding community ecology from functional traits. Trends in Ecology & Evolution, 21: 178–185. doi: https://doi.org/10.1016/j.tree.2006.02.002

McMurry JA, Juty N, Blomberg N, Burdett T, Conlin T, Conte N, Courtot M, Deck J, Dumontier M, Fellows DK, Gonzalez-Beltran A, Gormanns P, Grethe J, Hastings J, Hériché JK, Hermjakob H, Ison JC, Jimenez RC, Jupp S, Kunze J, Laibe C, Le Novère N, Malone J, Martin MJ, McEntyre JR, Morris C, Muilu J, Müller W, Rocca-Serra P, Sansone SA, Sariyar M, Snoep JL, Soiland-Reyes S, Stanford NJ, Swainston N, Washington N, Williams AR, Wimalaratne SM, Winfree LM, Wolstencroft K, Goble C, Mungall CJ, Haendel MA, Parkinson H. (2017). Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. PLoS Biol. 15(6):e2001414. doi: https://doi.org/10.1371/journal.pbio.2001414 .

Morales EA, Wetzel CE, Ector L (2021) New and poorly known araphid diatom species (Bacillariophyta) from regions near Lake Titicaca, South America and a discussion on the continued

use of morphological characters in araphid diatom taxonomy. PhytoKeys 187: 23-70. https://doi.org/10.3897/phytokeys.187.73338

Nilsson R.H., Larsson K.-H., Taylor A.F.S., Bengtsson-Palme J., Jeppesen T.S., Schigel D., Kennedy P., Picard K., Glöckner F.O., Tedersoo L., Saar I., Kõljalg U. & Abarenkov K. 2019. The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Research* 47: D259-D264. doi: https://doi.org/10.1093/nar/gky1022

Page RDM (2016) Surfacing the deep data of taxonomy. In: Michel E (Ed.) Anchoring Biodiversity Information: From Sherborn to the 21st century and beyond. ZooKeys 550: 247–260. https://doi.org/10.3897/zookeys.550.9293

Page RDM. 2019. Ozymandias: a biodiversity knowledge graph. *PeerJ* 7:e6739 https://doi.org/10.7717/peerj.6739

Parr, C.S., Schulz, K.S., Hammock, J., Wilson, N., Leary, P., Rice, J., & Corrigan, R.J. 2016. TraitBank: Practical semantics for organism attribute data. Semantic Web, 7: 577–588. doi: https://doi.org/10.3233/SW-150190

Paton A, Phillipson P, Suddee S (2016) Records of Wenchengia (Lamiaceae) from Vietnam. Biodiversity Data Journal 4: e9596. https://doi.org/10.3897/BDJ.4.e9596

Patterson B.D., Webala P.W., Lavery T.H., Agwanda B.R., Goodman S.M., Kerbis Peterhans J.C., Demos T.C. 2020. Evolutionary relationships and population genetics of the Afrotropical leaf-nosed bats (Chiroptera, Hipposideridae). ZooKeys 929: 117-161. https://doi.org/10.3897/zookeys.929.50240

Patterson DJ, Egloff W, Agosti D, Eades D, Franz N, Hagedorn G, Rees J, Remsen DP 2014. Scientific names of organisms: attribution, rights, and licensing . BMC Research Notes 2014, 7:79 doi: https://doi.org/10.1186/1756-0500-7-79

Penev L, Agosti D, Georgiev T, Catapano T, Miller J, Blagoderov V, Roberts D, Smith V, Brake I, RyrcroftS, Scott B, Johnson N, Sautter G, Chavan V, Robertson T, Remsen D, Stoev P, Parr C, Knapp S, Kress W, Thompson F, Erwin T (2010) Semantic tagging of and semantic enhancements to systematics papers: ZooKeys working examples. ZooKeys 50: 1-16. doi: https://doi.org/10.3897/zookeys.50.538

Penev L., Catapano T., Agosti D., Georgiev T., Sautter G, Stoev P. 2012 - Implementation of TaxPub, an NLM DTD extension for domain-specific markup in taxonomy, from the experience of a biodiversity publisher. In: Journal Article Tag Suite Conference (JATS-Con) Proceedings 2012 [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2012. doi: https://doi.org/10.5281/zenodo.804247

Penev L, Koureas D, Groom Q, Lanfear J, Agosti D, Casino A, Miller J, Arvanitidis C, Cochrane G, Hobern D, Banki O, Addink W, Kõljalg U, Copas K, Mergen P, Güntsch A, Benichou L, Benito Gonzalez Lopez J, Ruch P, Martin CS, Barov B, Demirova I, Hristova K (2022) Biodiversity Community Integrated Knowledge Library (BiCIKL). Research Ideas and Outcomes 8: e81136. doi: https://doi.org/10.3897/rio.8.e81136

Penev L, Lyal C, Weitzman A, Morse D, King D, Sautter G, Georgiev T, Catapano T, Agosti D (2011) XML schemas and mark-up practices of taxonomic literature. ZooKeys 150: 89-116. https://doi.org/10.3897/zookeys.150.2213

Penev L, Paton A, Nicolson N, Kirk P, Pyle RL, Whitton R, Georgiev T, Barker C, Hopkins C, Robert V, Biserkov J, Stoev P (2016) A common registration-to-publication automated pipeline for

nomenclatural acts for higher plants (International Plant Names Index, IPNI), fungi (Index Fungorum, MycoBank) and animals (ZooBank). In: Michel E (Ed.) Anchoring Biodiversity Information: From Sherborn to the 21st century and beyond. ZooKeys 550: 233–246. https://doi.org/10.3897/zookeys.550.9551

Penev L, Dimitrova M, Senderov V, Zhelezov G, Georgiev T, Stoev P, Simov K. OpenBiodiv: A Knowledge Graph for Literature-Extracted Linked Open Data in Biodiversity Science. *Publications*. 2019; 7(2):38. https://doi.org/10.3390/publications7020038

Ratnasingham S. & Hebert P.D. 2013. A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. *PLoS ONE* 8(8): e66213. doi: 10.1371/journal.pone.0066213

Ride W.D.L., Cogger H.G., Dupuis C., Kraus O., Minelli A., Thompson F. C. & Tubbs P.K. 2012. – *International code of zoological nomenclature*. Fourth edition. On-line version.

Schneider, F.D., Güntsch, A., Fichtmueller, D., Jochum, M., Le Provost, G., Penone, C., Gossner, M.M., König-Ries, B., Manning, P., Ostrowski, A. & Simons, N.K. 2019. Towards an ecological trait-data standard. Methods in Ecology and Evolution 10: 2006 - 2019. doi: 10.1111/2041-210X.13288

Sharma S., Ciufo S., Starchenko E., Darji D., Chlumsky R., Karsch-Mizrachi I., Schoch C.L. 2018. The NCBI Biocollections Database. Database Vol 2018: article ID bay006. Doi: https://doi.org/10.1093/database/bay006

Senderov V, Simov K, Franz N, Stoev P, Catapano T, Agosti D, Sautter G, Morris RA, Penev L 2018. OpenBiodiv-O: ontology of the OpenBiodiv knowledge management system. Journal of Biomedical Semantics 9: 5. doi: https://doi.org/10.1186/s13326-017-0174-5

Turland N.J., Wiersema J.H., Barrie F.R., Greuter W., Hawksworth D.L., Herendeen P.S., Knapp S., Kusber W.-H., Li D.-Z., Marhold K., May T.W., McNeill J., Monro A.M., Prado J., Price M.J. & Smith G.F. (eds) 2018. – *International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code) adopted by the Nineteenth International Botanical Congress Shenzhen, China, July 2017.* Regnum Vegetabile 159. Glashütten, Koeltz Botanical Books.# doi: http://dx.doi.org/https://doi.org/10.12705/Code.2018

Upham N.S., Poelen J.H., Paul D., Groom Q.J., Simmons N.B., Vanhove M.P.M., Bertolino S., Reeder D.M., Bastos-Silveira C., Sen A., Sterner B., Franz N.M., Guidoti M., Penev L., Agosti D. 2021. Liberating host–virus knowledge from biological dark data. The Lancet Planetary Health. doi: https://doi.org/10.1016/S2542-5196(21)00196-0

Violle, C., Navas, M.-L., Vile, D., Kazakou, E., Fortunel, C., Hummel, I., & Garnier, E. 2007. Let the concept of trait be functional! Oikos 116: 882–892. doi: https://doi.org/10.1111/j.0030-1299.2007.15559.x

Voutsiadou E., Gerovasileiou V., Vandepitte L., Ganias K. & Arvanitidis C. 2017. Aristotle's scientific contributions to the classification, nomenclature and distribution of marine organisms. *Mediterranean Marine Science* 18: 468-478. doi: http://dx.doi.org/10.12681/mms.13874

Walls, R. L., Deck, J., Guralnick, R., Baskauf, S., Beaman, R., Blum, S., ... & Wooley, J. (2014). Semantics in support of biodiversity knowledge discovery: an introduction to the biological collections ontology and related ontologies. PloS one, 9(3), e89606. doi: https://doi.org/10.1371/journal.pone.0089606

Winston J. 1999. Describing Species. New York: Columbia University Press. doi: https://doi.org/10.5281/zenodo.7006542

Zanol J, Hutchings P (2022) A new species of giant Eunice (Eunicidae, Polychaeta, Annelida) from the east coast of Australia. ZooKeys 1118: 97-109. https://doi.org/10.3897/zookeys.1118.86448

Supplementary material

Overview table of recommendations

As the most recent version of TaxPub (V1rc2) extends [56] the JATS Journal Publishing Tag Set Version 1.1[57] that version is referred to in the table.
Table

[56] https://github.com/plazi/TaxPub/
[57] https://jats.nlm.nih.gov/publishing/1.1/

Comments

By Wouter in the Arpha writing tool

replace with: readable or interpretable, see: https://docs.google.com/document/d/1PPGW83siPDMLG5QRHsKM2cOUPT-O5SqS_jZvs_AO1Qw for definitions about actionable, readable and interpretable

readable (machina actionable means that a machine cannot only interpret the data but also knows what actions are possible by

replace with: readable or even machine actionable. Machine actionable means that data is not only interpretable by machines, but they know also how to act upon it through FAIR metadata, predictable PID resolution and knowing possible operations for the type of data. According to FDO Forum machine actionable ”are those elements in bit-sequences that are machine interpretable and belong to a type for which operations have been specified in symbolic grammar." (cite: https://docs.google.com/document/d/1PPGW83siPDMLG5QRHsKM2cOUPT-O5SqS_jZvs_AO1Qw)

Andreas Kroh

The third index that provides registration of nomenclatural acts for fungi is the Fungal Names maintained in China.
Has been deleted