Opinion paper

EJT editorial standard for the semantic enhancement of specimen data in taxonomy literature

Chloë CHESTER [1,*], Donat AGOSTI [2], Guido SAUTTER [3], Terry CATAPANO [4],
Koen MARTENS [5], Isabelle GÉRARD [6] & Laurence BÉNICHOU [7]

[1,7] Muséum national d'histoire naturelle, 57 rue Cuvier, 75005 Paris, France.
[2,3,4] Plazi, Zinggstrasse 16, 3007 Bern, Switzerland.
[5] Royal Belgian Institute for Natural Sciences, Vautierstraat 29, B-1000 Brussels, Belgium.
[6] Royal Museum for Central Africa, 13 Leuvensesteenweg, B-3080 Tervuren, Belgium.

[*] Corresponding author: chloe.chester@mnhn.fr
[2] Email: agosti@amnh.org
[3] Email: gsautter@gmail.com
[4] Email: catapano@plazi.org
[5] Email: koen.martens@naturalsciences.be
[6] Email: isabelle.gerard@africamuseum.be
[7] Email: laurence.benichou@mnhn.fr

Abstract. This paper describes a set of guidelines for the citation of zoological and botanical specimens in the *European Journal of Taxonomy*. The guidelines stipulate controlled vocabularies and precise formats for presenting the specimens examined within a taxonomic publication, which allow for the rich data associated with the primary research material to be harvested, distributed and interlinked online via international biodiversity data aggregators. Herein we explain how the *EJT* editorial standard was defined and how this initiative fits into the journal's project to semantically enhance its publications using the Plazi TaxPub DTD extension. By establishing a standardised format for the citation of taxonomic specimens, the journal intends to widen the distribution of and improve accessibility to the data it publishes. Authors who conform to these guidelines will benefit from higher visibility and new ways of visualising their work. In a wider context, we hope that other taxonomy journals will adopt this approach to their publications, adapting their working methods to enable domain-specific text mining to take place. If specimen data can be efficiently cited, harvested and linked to wider resources, we propose that there is also the potential to develop alternative metrics for assessing impact and productivity within the natural sciences.

Keywords. XML, publishing standard, taxonomy, FAIR data, Open Science.

Table of contents

Introduction

Taxonomy and the urgent need for integrated bioinformatics

Publications containing descriptive taxonomy and nomenclatural acts constitute a pillar for developing robust hypotheses on identity and relationships in the natural world. Species names and the treatments associated with them are a prerequisite for reliable research across the natural sciences (Wägele *et al.* 2011), which in turn plays a pivotal role in effective conservation management and sustainable development (e.g., Groombridge 1992; Heywood *et al.* 1995; McCook *et al.* 2010). At a time when there is growing concern across the planet for the cascading effects of climate change and biodiversity loss on agriculture, land use and human welfare (IPCC 2018), it is now urgent for taxon concepts to move beyond the description of isolated organisms and towards an integrative systems approach that is able to address species interactions, both with their environment and with other species (Hardisty *et al.* 2013).

Despite its fundamental role in subsequent fields of research, alpha taxonomy is suffering from impediments that effect end-user accessibility, as well as the valorisation of its authors and their output (Ebach *et al.* 2011). The establishment of best practices in the citation of collection specimens, taxon concepts and bibliographic works would contribute to the interlinked resource infrastructure that is called for by the community (Hobern et al. 2019) and help to resolve some of the issues. However, the challenge therein is twofold, as we must overcome both sociological (e.g., habits and 'traditional' ways of working) and technical barriers in order to engineer change. This notion was a guiding vector throughout the *European Journal of Taxonomy*'s project for the semantic enhancement of its

publications. Achieving dynamic data exchange was not the unique goal; as well as the chosen workflow (the technical framework), the applicability and relevance of the method (the sociological factor) were equally paramount to ensuring that authors embrace the increasingly technological research paradigm.

The *European Journal of Taxonomy*: spearheading innovative publishing workflows

The *European Journal of Taxonomy* (*EJT*) is a peer-reviewed international journal in descriptive taxonomy of eukaryotic organisms (zoology, entomology, botany, palaeontology). The journal was founded by a consortium of European natural history institutions to take advantage of the shift from paper to online publications, as supported by the recent modifications in the governing Codes of nomenclature in zoology (ICZN 2009, 2012) and botany (McNeill *et al.* 2012, Turland *et al.* 2018).

From the journal's onset in 2011, *EJT*'s articles have been published directly online as individual PDFs in Diamond Open Access (Bénichou *et al.* 2011). However, in accordance with the journal's founding principle to spearhead innovative publishing techniques, it had always been envisaged to provide publications in a machine-readable format, namely Extensible Mark-up Language (XML).

Beyond the advantages of XML's stability as an archiving format (Morrissey *et al.* 2010; cOAlition S 2019; Library of Congress 2019), which would serve to guarantee the longevity of its publications, *EJT* understood that the way to increase the visibility of taxonomic research was to provide machine-readable and semantically enhanced text (Agosti *et al.* 2007; Agosti & Egloff 2009; Penev *et al.* 2010; Penev *et al.* 2011a; Penev *et al.* 2012; Miller *et al.* 2012; Miller *et al.* 2015). Moreover, in light of recent academic movements such as the FAIR Data Principles (Findable, Accessible, Interoperable, Reusable) for Open Science (Wilkinson *et al.* 2016) and international initiatives to mobilise biodiversity data – e.g., GBIF (https://gbif.org), the Catalogue of Life (https://www.catalogueoflife.org), DiSSCo (https://www.dissco.eu) – it became an increasing priority to spread the alpha taxonomy data contained within *EJT* articles throughout the ecosystem of dynamic, stable, free-to-use and interconnected platforms available on the Web.

A community-wide solution

The promotion of taxonomy, systematics and collection-based research via scientific publishing is a vision that *EJT* shares with the Consortium of European Taxonomic Facilities (CETAF), who officially endorsed the journal as its flagship title in 2016. *EJT* therefore aims to provide the taxonomic community with all of the modern interactive web-based facilities expected of a high-level, high-impact journal.

An additional motivation behind the creation of an international, cross-institutional journal was to enable its members to collectively tackle the challenges of the digital transition. Indeed, the unprecedented technological advances associated with 21st-century scientific publishing have given rise to complex strategic and technical issues related to the visibility, access, format and financial structure of academic journals, especially publicly funded titles (Bénichou *et al.* 2011, 2012; Côtez *et al.* 2018).

XML has proven itself as an indispensable format. The availability of a "machine readable format (for example XML)" now features in the mandatory quality criteria for Plan S-compliant journals (cOAlition S 2019: Art. 9.2). However, this is a veritable technical barrier for many smaller Open Access publishers, especially those who do not impose author processing charges (APCs). Given the fact that *EJT* is representative of typical natural history journals and their technical set-up, publishing on behalf of public-sector authors, the *EJT*-XML project reflects the operational and sociological constraints experienced by many other independent journals.

Thus, in line with *EJT*'s role as incubator on behalf of its members, the solution had to be transposable to other institutional journals managed within the consortium's network, and even further afield into the wider community of taxonomic publishing. We here report on the process of defining a new post-publishing XML workflow and the associated formatting that enables optimised encoding of the published content.

Approach, tools and partners

Project analysis

The *EJT* steering committee, consisting of the general directors of the members of the *EJT* consortium, asked for an assessment of the potential ways to achieve a semantically enhanced XML version of *EJT*'s publications. In 2015, the Naturalis Biodiversity Centre (an *EJT* consortium member) was tasked with investigating the different methods of obtaining and utilizing an XML version of *EJT* content.

The report submitted by Naturalis in March 2017 concluded that, while the field of taxonomy provided exciting opportunities for interlinking data to the wider biodiversity community, elaborating an in-house workflow operated by the *EJT* production team was not advisable. Given the journal's independent position (i.e., entirely managed, produced and published directly by the consortium members), the development of a tailor-made XML production platform was considered too costly, and performing manual encoding would be too time-consuming and technical for the desk editors to undertake. The only existing platform that offered integrated services for XML encoding of taxonomy literature was the Pensoft ARPHA Journal Publishing System (see Penev *et al.* 2019). *EJT* felt that it was important to investigate an alternative to this unique solution which, although highly performant, represented issues of workflow rigidity, cost and commercial monopoly.

Plazi and GoldenGATE

Based on its findings, the Naturalis team carried out a proof-of-concept using GoldenGATE to apply mark-up to *EJT* articles published in PDF (Fig. 1). GoldenGATE is a semi-automatic retro-conversion tool for encoding taxonomic literature (Sautter *et al.* 2007). The open-source program was developed by the Swiss NGO Plazi (http://plazi.org) to support and promote the interoperability of taxonomic treatments with other relevant cyber infrastructure components (name servers, biodiversity resources, etc.). It was specifically conceived to process "born-digital" PDFs (as opposed to digitised paper publications) by recovering the text from the rendering instructions embedded in the file and discovering structural elements such as figures, tables and textual sub-sections e.g., taxonomic treatments. Once identified, these elements are parsed using the tags defined in the TaxPub extension to the JATS DTD (Catapano 2010), which in turn allows the text to be annotated with scalable links to external sources. The process can be highly automated by developing a journal-specific template, chaining the various steps in the conversion process and using batch processing.

The *EJT*-Plazi workflow

The solution of a collaboration between *EJT* and Plazi was chosen owing to the latter's valuable expertise in domain-specific mark-up for taxonomic literature (Agosti *et al.* 2019a), as well as for their established distribution network for the harvested content; both essential points for *EJT*. Furthermore, Plazi's data warehouse Treatment Bank (https://treatmentbank.org), combined with the group's active status as a trusted data provider for the Global Biodiversity Infrastructure Facility (GBIF) and the Biodiversity Literature Repository (BLR; https://zenodo.org/communities/biosyslit/; Agosti 2019b) offered further scope for the distribution of the taxonomic treatments, images, specimens and other data published by *EJT*.

Using GoldenGATE and TaxPub, the taxonomic treatments and specimen data featured within *EJT* articles are converted into Darwin Core (DwC) archives (Wieczorek *et al.* 2012), the biodiversity informatics data standard developed by the Biodiversity Information Standards (TDWG), which is the preferred format for publishing data to the GBIF network. Once converted, data relating to the treatments and specimens published in *EJT* is accessible via Treatment Bank and GBIF within a few days of publication, at no extra cost or effort for the author. The figures and captions, along with a full copy of the article, are also available via the Biodiversity Literature Repository. All data harvested from

an article is explicitly linked to the original publication using full original citations and a DOI. This means that the sub-article elements are integrated into a much larger network of biodiversity data, which helps to ensure their preservation but also maximises the reach of the article (Fig. 2).

Semantic enhancement of specimen citations

A goldmine of data

The details given by authors about the specimens studied in their research – especially for the type material used in nomenclatural acts – are extremely rich and highly structured. Within the realm of biodiversity informatics, the physical specimens studied in a systematic account can be used as a valuable anchor to unambiguously identify taxa, analyse the associated collecting data and track any research based on these vectors.

Fig. 1. The GoldenGATE interface. Using pattern recognition and natural language processing techniques, the article structure is encoded allowing the extraction of the taxon treatment sections. Furthermore, the TaxPub extension allows for fine-grain mark-up of sub-article elements such as the scientific names (A) and specimen citations (B).
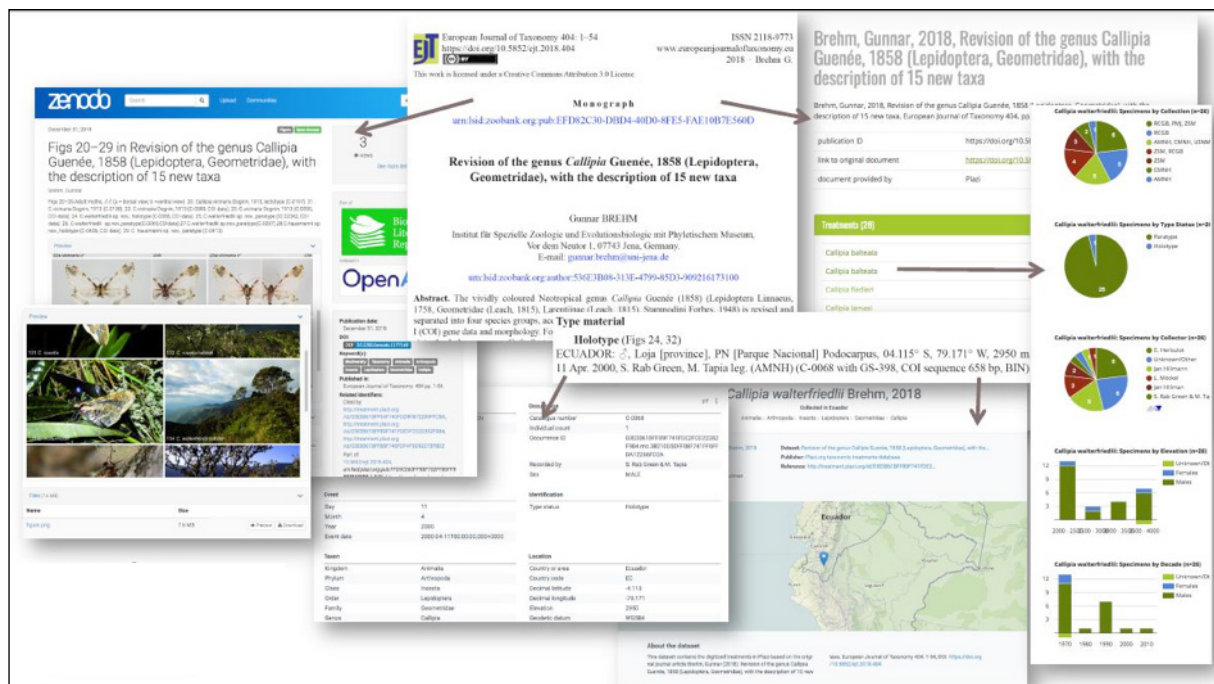
Fig. 2. The *EJT*-Plazi workflow. Plazi processes the PDF of an *EJT* publication to extract the sub-article elements and distribute them to biodiversity aggregators. Taxonomic treatments, images, tables, scientific names and even the fine-grain specimen data are semantically enhanced and available on a variety of platforms.
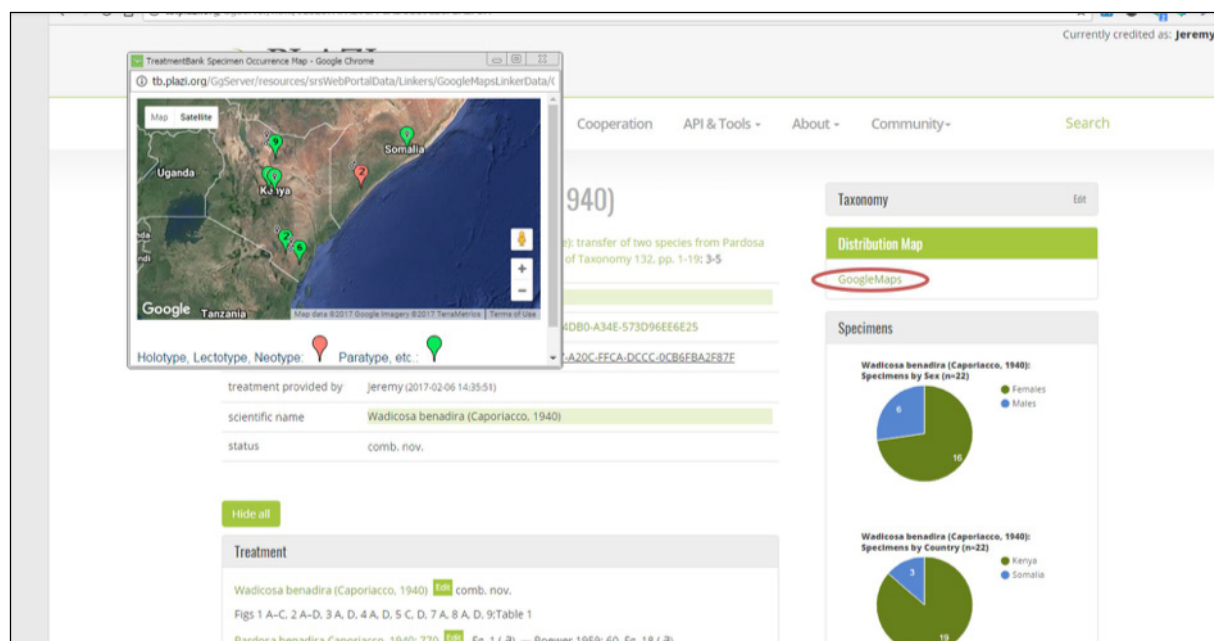
Fig. 3. A taxonomic treatment extracted from the *European Journal of Taxonomy* and displayed on the Plazi Treatment Bank. In the right-hand sidebar, the parsing performed on the specimen citations has been used to generate graphic charts of the material studied. In the inset, a map has been generated via GoogleMaps by plotting the extracted geocoordinates.

Plazi demonstrated how the information relating to the collection and preservation of physical specimens can be parsed to a fine degree and used to generate Darwin Core archives of occurrence records (Wieczorek *et al.* 2012). Once available in a machine-readable format, these records can be used to search, cite and track specimens through a multitude of criteria (by locality, date, collector, repository etc.) (Fig. 3), providing researchers and collection managers with sophisticated ways to query data sets and carry out their work (e.g., Nicolson & Tucker 2017; Miller 2019).

Structuring text for automatic parsing

During the first test phase, *EJT* submitted to Plazi a sample of 30 articles representing the diversity of disciplines (entomology, zoology, palaeontology, botany) and article types (species descriptions, revisions, check-lists...) published by the journal. These publications were processed within GoldenGATE using algorithms to automatically mark-up the article structure and the material sections.

Certain highly structured elements such as dates and geo-coordinates were correctly identified and encoded. However, the wide range of methods, punctuation and vocabularies used to describe the specimens examined resulted in a considerable amount of errors that were time-consuming to correct manually. Difficulties mainly stemmed from matching occurrence data to the correct specimen and delimiting unstructured data, e.g., distinguishing a habitat description from a locality within a text string.

Thus it became clear that a standardised format for the presentation of specimen data was required in order to facilitate fine-grain harvesting of details related to the physical material (specimens) studied. If correctly formatted, these data could be harvested, converted into occurrence records, then integrated into the wider infrastructure of biodiversity informatics (Dikow 2019). Moreover, by establishing standardised formatting that different authors and journals could follow, we also hoped to enable text mining for other taxonomy journals who publish in PDF.

Defining the *EJT* format

The specimen citations contained within the test set of articles, as well as those found in further articles published by *EJT* and other taxonomic journals (*Zoosystema*, *Adansonia*, *Geodiversitas*, *ZooKeys*, *Zootaxa*, *the Zoological Journal of the Linnean Society, the Botanical Journal of the Linnean Society*, *PhytoKeys*, *Phytotaxa*), were analysed to establish the types of data most commonly presented.

Once the specimen citations had been broken down into a list of highly recurrent fields, these fields were mapped to DWC terms (http://rs.tdwg.org/dwc/terms/). This permitted the creation of a flexible template for specimen citations, designed to fit a broad array of specimen data, which was tested on a wide sample of citations and refined according to feedback from Plazi, the *EJT* production staff and scientific editors, as well as several active *EJT* authors. The resulting template and formatting guidelines are described in detail hereafter as the Material Citations Formatting Guidelines.

Material citations formatting guidelines

In accordance with the *European Journal of Taxonomy*'s FAIR Data & Open Science policy (available from https://europeanjournaloftaxonomy.eu), the formatting guide for zoological and botanical specimen citations is presented below. Authors are encouraged to prepare their manuscripts according to this model prior to submission, but they will also be given an opportunity to comply upon acceptance of the article.

While *EJT* strongly recommends that authors adhere to the guidelines given below, the fine-grain formatting of the material citations is not compulsory; if an author decides not to comply or that the material is not appropriate, *EJT* will perform reduced formatting during production. In this case, the majority of the specimen data will not be tagged and converted into a machine-readable format; this means that the specimen-related information from the paper will not be included in major databases.

Only specimen data presented in the 'Materials examined' section will be tagged and converted for distribution. At this time, any specimen data presented in a separate table or section of the paper cannot be linked back to the specimen citation to form a full occurrence record.

General presentation

1. Order

Each material citation is composed of diverse data fields (number of specimens, locality, date collected, etc.) that *EJT* identifies using Darwin Core (DWC) terms. To efficiently perform this, it is important to ensure that the different fields of a material citation are consistently presented in the same order throughout the article or, at the very least, within a taxon treatment.

The preferred order for data fields differs for zoology and botany, and are as follows:

Zoology

COUNTRY • specimen(s) (e.g., "1 ♂"); **geographic / locality data** (from largest to smallest); **geographic coordinates**; **altitude / elevation / depth** (using alt. / elev. / m a.s.l. etc.); **date** (format: dd Mmm. YYYY, e.g., "16 Jan. 1998"); **collector** (followed by "leg."); **other collecting data** (e.g., micro habitat / host / method of collecting / "DNA voucher specimen"/ "vial with detached elements", etc.); **barcodes/identifiers** (e.g., "GenBank: MG779236"); **institution code and specimen code** (e.g., "CBF 06023").

Botany

COUNTRY • locality data (from least to most specific); **geographic coordinates**; **altitude / elevation / depth** (using alt. / elev. / m a.s.l. etc.); **other collecting data** (e.g., habitat / substrate / method of collecting / "herbarium specimen" / "DNA voucher specimen"); **date** (format: dd Mmm. YYYY, e.g., "16 Jan. 1998"); **phenology** (e.g., fl / fr); *collection number* (*collector's name + number in italics*); **determinator** (followed by "det."); **repository[identifier in the repository]** (format: herbarium acronym[identifier]; e.g., "NY[133679], NY[3774223], NY[3774224], L[L.1264510]!, PLP[PLP-0012346], BR[AWH10018408]); **genetic identifiers** (e.g., "GenBank no.: MG779236").

Details on how to format each data field are provided in the '**Data fields***' section.

2. Punctuation

A bullet point "•" (unicode: hex 2022, decimal 8226) is used to signify the beginning of a material citation. In Microsoft Word, the following keyboard shortcuts can be used to obtain a bullet point:

• for **Mac**: Alt + 8 (QWERTY keyboard) / Alt + shift + full stop (AZERTY)
• for **Windows**: Alt + 0149 on the numeric keypad

Within a citation, a **semicolon** "**;**" delimits each different field. Semicolons should not be used elsewhere in a material citation.

A single field can be composed of several details, which are separated by commas (e.g., the details region, area and town for the 'locality' field). In the following example, the 'locality' field is composed of two details: Province ("Eastern Cape Province") and town ("Cradock"):

> SOUTH AFRICA • ♂; Eastern Cape Province, Cradock; 30°19′ S, 25°39′ E; Aug.–Oct. 1985; museum staff leg.; pitfall trap; NMBA 1170.

3. Type material

Zoology

Type material should be presented in separate paragraphs with relevant subheadings (Holotype, Paratypes, etc.).

Botany

Basionyms & synonyms

In botanical articles, the type material of basionyms and homotypic synonyms is presented in the same paragraph as the relative scientific name and bibliographic reference (just under the treatment heading), preceded by the mention "**Type:**" in bold. All heterotypic synonyms under the recognised name are cited accordingly with their basionyms.

> ***Ohelopapa flexilis*** (Setch.) F.Rousseau, Martin-Lescanne, Payri & L.Le Gall comb. nov.
> Fig. 3, Table 2
>
> **Basionym**
>
> *Laurencia flexilis* Setch., *University of California Publications in Botany* 12: 101, pl. 19, figs 1–6 (Setchell 1926). **Type**: FRANCE • French Polynesia, Tahiti; reef at Tahara Mountain; *Setchell W.A., Setchell C.B. & Parks H.E. 5246*; holotype: JEPS[UC261333]. Original material is also preserved in SAP *fide* Masuda et al. (1999, 2006).

> ***Afroriccardia comosa*** (Steph.) Reeb & Gradst. comb. nov.
> Figs 3–4
>
> *Aneura comosa* Steph., *Botanical Gazette* 15 (11): 281 (Stephani 1890). – *Riccardia comosa* (Steph.) E.W Jones, *Transactions of the British Bryological Society* 3: 74 (Jones 1956, nom. inval.). – **Type**: FRANCE • La Réunion; 1889; *Rodriguez s.n.*; holotype: G[G00045027]!; isotype: PC[PC0103522]!

This presentation should be used regardless of whether the specimen has been examined (indicated by an exclamation mark in this context) or not.

Example of material designated as lectotype in a previous publication vs material designated within the paper:

Begonia acetosella Craib
Figs 6–7

Bulletin of Miscellaneous Information, Kew 1912: 153 (Craib 1912). – **Type**: THAILAND • Chiengmai, Doi Sootep; 21 Mar. 1909; 18°50′ N, 98°54′ E; *A.F.G. Kerr 557*; lectotype: K[K000761199], designated by Tebbit 2003a, sheet 2 K[K000761201]; isolectotype: B.

Begonia tetragona Irmsch., *Mitteilungen aus dem Institut für allgemeine Botanik in Hamburg* 10: 515 (Irmscher 1939). – **Type**: CHINA • Yunnan, Mengtze; *Henry 10737A*; lectotype: B[B100238046], **here designated**; isolectotype: E[E00315022].

Nᴇᴡ ᴛᴀxᴀ
Type material for new taxa, which do not have an existing bibliographic reference (e.g., it is proposed for the first time in the publication), should be presented as follows, under the heading "Type" or "Type material":

Type
PERU • Region Pasco, Prov. Oxapampa, Dist. Palcazu, Parque Nacional Yanachaga-Chemillén, sector Paujil; 150 m from entrance to Las Cavernas on trail from Paujil; 10°20′40″ S, 75°15′1″ W; alt. 432 m; 25 Feb. 2016; *Moonlight & Daza 318*; holotype: MOL; isotypes: E, MO, USM.

---

4. Repetitive data

Repetitive data can be indicated with terms such as "same data as for holotype", "same data as for preceding", "same locality", "ibid.", etc., as long as the method used is consistent throughout the paper.

Paratypes
SOUTH AFRICA • 1 ♂; same data as for holotype • 1 ♀; Limpopo Province, "Pietersburg" [now Polokwane]; 23°54′ S, 27°23′ E; 4 Nov. 1916; C.A. Thompson leg.; TM 13375 (formerly 2217).

If a material citation is identical to another with only minor differences, the exceptions should be listed after the mention of repeated data e.g.:

Zoology

SOUTH AFRICA • 1 ♂; Eastern Cape Province, Middelburg; 31°49′ S, 25°00′ E; 8 Sep. 1995; M. de Jager leg.; pit traps; NCA 95/394 • 1 ♂; same collection data as for preceding; 3 Sep. 1995; NCA 95/243 • 2 ♂♂; same collection data as for preceding; 8 May 1999; associated with termites; NCA 91/1051.

Botany

INDIA – **Arunachal-Pradesh** • Dirang Dzong; 8 Aug. 1938; *Ward 14055*; BM • Senge Dzong; 18 Aug. 1938; *Ward 14091*; BM. – **Mizoram** • Hmuifang; Jul. 1926; *Parry 45 • ibid.*; Jul. 1926; *Parry 46*; K • ibid.; Jul. 1926; *Parry 47*; K • Lungleh; 1 Sep. 1931; *Wenger 320*; K.

---

5. 'Missing' elements

Zoology

It is not necessary to include information such as "no date" or "no locality data"; list only the elements that are available.

Botany

All commonly used abbreviations can be used where appropriate ("s.n."/ "s.loc."/ "s.d.", "s.coll" etc.).

Type

IVORY COAST • Régio du Bas-Sassandra, Km 41 Sassandra-San Pedro road; 16 Nov. 1968; fl. bud, fr.; *Breteler 6052*; holotype: WAG; isotype: BR, K, MO n.v., PRE n.v., W n.v.

Basionym

Laurencia crustiformans McDermaid, *Phycologia* 28: 352, figs 2–8 (McDermaid 1989). **Type**: USA • Hawaii, Oahu, Lualualei Beach Park; 3 Jul. 1988; *K.J. McDermaid s.n.*; holotype: BISH[KM 2050] (Abbott 1999: 384).

---

6. Label citations

Double quotation marks (" ") are used to represent label citations that cannot be reliably interpreted and formatted as recommended in these guidelines. This data will simply be parsed as a verbatim citation. *EJT* recommends including photos of labels as figures if they contain data that cannot be standardised.

Only quotation marks should be used to present verbatim label data and they should not appear elsewhere in a material citation.

7. Author interpretation

Use **square brackets [ ]** to distinguish data that has been interpreted from a label e.g., coordinates interpreted from a locality, or translations of foreign text:

Lectotype (here designated)
MONTENEGRO • 1 specimen; "Popovo Höhle bei Njegus" [Popovo Cave near Njegus]; [43.5291° N, 19.2074° E]; 30 May 1903; Sturany leg.; NHMW 38260a.

Data fields

The different data fields of a material citation that *EJT* identifies for conversion and diffusion are explained below, along with the format required to achieve maximum output and accuracy.

1. Country/ Water body

The citations must be listed by either country or water body (e.g., ocean/sea), using a separate paragraph for each new zone. The country or water body is presented in capital letters.

If the material is organised by region, use the following format:

> MADAGASCAR – **Mahajanga Province** • Betsiboka Region, forêt de Kasijy, Kelifely; Nov. 1974; fl; *Morat 4700*; P, TAN • Boeny Region, Majunga; 30 Mar. 1912; fr; *Afzelius 259*; P • Soalala, Réserve Naturelle Intégrale de Namoroka, ca 40 km S of Soalala; 3 Feb. 2000; fr; *Davis, Rakotonasolo & Wilkin 2529*; BR, K, TAN • Majunga; dunes; Feb. 1915; fl, fr; *Perrier de la Bâthie 3504*; P • Ambongo; 17 Feb. 1841; fr; *Pervillé 680*; P.

> CHINA – **Guangdong Province** • 1 ♀; Tianma, Xinhui; 24 Apr. 1956; Z.B. Zhou leg.; SCAU • 4 ♀♀; Xinhui; Nov. 1956; SCAU • 1 ♀; Nanhai; 14 Oct. 1955; L.B. Huo leg.; SCAU. – **Guangxi Province** • 1 ♂; Longsheng; 6 Apr. 1974; Y.L. Luo leg.; SCAU • 1 ♂; Huaan; 26 Apr. 1982; Y.Q. Tang leg.; SCAU. – **Hainan Province** • 1 ♂, 3 ♀♀; Diaoluoshan; Jul. 1995; Z.Q. Peng leg.; SCAU • 1 ♂; Diaoluoshan; 8 May 2005; X.M. Wang leg.; SCAU • 1 ♂; Dongfang; 27 Nov. 1997; Z.Q. Peng leg.; SCAU.

2. Specimen count (zoology)

This field can contain several indications about the specimen(s) cited: **number**, **nature** (e.g., specimen, juv., shell, exuviae), **sex** and **type status**. All subsequent data in the same citation will be applied to the specimen(s) presented.

> SOUTH AFRICA • 1 ♂; Eastern Cape Province, Middelburg; 31°49′ S, 25°00′ E; 8 Sep. 1995; M. de Jager leg.; pit traps; NCA 95/394 • 2 ♀♀; same collection data as for preceding; 3 Sep. 1995; NCA 95/243 • 3 ♂♂; same collection data as for preceding; 8 May 1991; associated with termites; NCA 91/1051.

> THAILAND • 3 shells, same data as for preceding; HNHM 97479 • 16 specimens (preserved in ethanol); same data as for preceding; UF 76457.

> MYANMAR • 1 shell, holotype of *P. ponsonbyi* (D = 17.8 mm); "Burmah", Hlindet; NHMUK 1913.3.14.9.

3. Locality

The locality data is listed from least to most specific, using commas to divide each detail.

It is recommended to employ the English name in current usage where possible. If a different system is used, e.g., variant spellings or archaic names from label transcriptions, these should preferably be identified using quotes, with their current names given in square brackets.

> MONTENEGRO • "Popovo Höhle bei Njegus" [Popovo Cave near Njegus]; 43.5291° N, 19.2074° E; 30 May 1903; Sturany leg.; NHMW 38260a.

4. Geographic coordinates

Various formats are accepted but it is important to include the degree symbol (°) as well as the direction (N/E/S/W), which distinguishes the data as a geographic coordinate:

- degrees minutes seconds: 40°26′46″ N, 79°58′56″ W
- degrees decimal minutes: 40°26.767′ N, 79°58.933′ W
- decimal degrees: 40.446° N, 79.982° W

Geographic coordinates should be presented to a maximum of 5 decimal places. Latitude and longitude are separated with a comma. Latitude is cited first, then longitude.

5. Altitude/elevation/depth

This type of measurement should be explicit in the material citations, e.g.:

- Altitude: alt. 489 m or 547 m a.s.l.
- Depth: depth 20 m

6. Collection date

Format: d(d) Mmm. YYYY

Date ranges should be shown with an n-dash, e.g., Jan.–May 2018 / 5 Feb.–6 Apr. 2016 / 14 Dec. 2008–3 Feb. 2009 / 1950–1953.

7. Collector (zoology)

The name(s) of the collector(s) should always be followed by "leg."; for expeditions, "exped." can be used, e.g., "MNHN exped."

8. Collector and collection number (botany)

The collector's name and field number are cited together in italics.

For botanical disciplines that do not catalogue specimens on sheets (e.g., algae, diatoms), we ask that authors use "collected by: X", because the term "leg." does not have the same signification across all botanical fields.

9. Additional data

Ideally, the data fields identified above should be listed before any other collection data. If a different order is used, it is important to be as consistent as possible throughout the paper, or at least within a single treatment. Semicolons may be used to separate any additional data into appropriate fields, e.g.:

> SWEDEN • ♂; Halland, Halmstad; Gårdshult, Buskastycket; 56.41° N, 13.91° E; 3–25 May 2005; Swedish Malaise Trap Project leg.; trap 35; collecting event: 1786; Malaise trap; hay meadow; NHRS CEC1405 • ♂; Öland, Mörbylånga, llevi; 56.61° N, 16.60° E; 8 Apr.–9 May 2016; M. & C. Jaschhof leg.; Malaise trap; herb-rich meadow near forest; SDEI CEC1429.

Additional data can also be given in the appropriate field between brackets, e.g.:

> **Holotype**
> MYANMAR • 1 shell, holotype of *P. dextrorse* (D = 16 mm); Tenasserium Valley; NHMUK 1906.2.2.144.

> **Other material**
> THAILAND • 2 ♂♂, 1 ♂ (gonopods lost); Prachuap Khiri Khan Province, Kui Buri District, Hat Kham Subdistrict, Ban Yan Sue; 12°03′12″ N, 99°37′52″ E; ca 147 m a.s.l.; 31 Aug. 2007; ASRU members leg.; CUMZ.

10. Associated sequences

Specimen accession numbers, barcodes and DNA accessions should be identified as such, e.g., "GenBank: U34853.1", "accession no.: 5587453".

> BRAZIL • ♂; Pernambuco, off Recife, approximately 20 nautical miles from the coast; REC I, dredge 4; 8°08′51.5″ S, 34°34′08.0″ W; 65 m depth; 7 Feb. 2018; agglomerations of sponges, rhodoliths and calcareous algae; GenBank 16S gene: MK918616; MOUFPE 19470.

11. Repository data

The repository data field should be composed of an institution acronym followed by a specimen code/catalogue number/barcode (where available).

Zoology

INSTITUTION ACRONYM

Acronyms for repositories must feature in a distinct list in the **Materials and methods** section, under a heading called "**Repositories**", "**Institutional acronyms**" or "**Institutional abbreviations**". Institution codes must follow GRSciColl (https://gbif.org/grscicoll) where possible.

SPECIMEN CODE

Where a specimen code is available, it should be explicit which specimen it refers to. This guarantees unambiguous interpretation, both by human readers and upon encoding. For example, in the citation below, we cannot distinguish which specimens are catalogued under which code:

> NAMIBIA • 2 ♂♂, 4 imm.; Grootfontein, Nosib Cave; 8 Feb. 1995; SEGL leg.; SAMC B7732, B8870.

This citation should be presented as follows:

> NAMIBIA • 2 ♂♂; Grootfontein, Nosib Cave; 8 Feb. 1995; SEGL leg.; SAMC B7732 • 4 imm.; same data as for preceding; SAMC B8870.

Use the word "to" instead of a hyphen or an n-dash in order to show a range of specimen numbers. E.g., "NHMUK 213584 to 213595".

Botany

Acronyms of herbaria must follow **Index Herbariorum** (http://sweetgum.nybg.org/science/ih) and a phrase to this effect will be included in the 'Materials and methods' section under the heading '**Repositories**'. Any acronyms used for repositories that do not feature in the Index Herbariorum must also be given here, e.g.:

Material and methods

Repositories

Acronyms of herbaria follow Index Herbariorum (http://sweetgum.nybg.org/science/ih/), except for the following repository:

UANT =   University of Antwerp, Belgium

Identifier in the repository

Each identifier (barcode and/or other catalogue numbers) should be cited exactly as it is registered in the repository. Each individual code is presented within square brackets immediately after the herbarium acronym. E.g.:

USA • Hawaii, Hawaii Island, Isaac Hale Beach Park; alt. 250 m; Jan. 2008; *A. Kurihara s.n.* SISH[ARS03327][766726].

FRANCE – La Réunion • "sur les mousses, source pétrifiante de Hell-Bourge"; *G. de l'Isle 220*; PC[PC0716023] • "sous Piton de la Fournaise, le long GR2, Réserve de Mare Longue"; 21°20′30″ S, 55°44′30″ E; 175–300 m a.s.l.; *Vojko 9435B*; EGR • *G de l'Isle 570bis*; PC[PC0716024], G[G-00264057] • *Rodriques s.n.*; G[G-00264058] • "plaine des palmistes"; *s.coll. 56*; PC[PC0716026].

In case of type material, use the same convention for all fields, except for **repository[identifier in the repository]:** introduce the field with the nature of the type: holotype, lectotype...

For example:

"lectotype: BM[BM000043986], **here designated**; isolectotypes: P, K, WAG."
"holotype: BR[BR0000013468675]!"
"holotype: NY[133679]!; isotypes: NY[3774223][3774224], L[L.1264510]!"
"holotype: BM[slide BM001166601] illustrated in Figs 3–62, the valve representing the holotype is illustrated in Fig. 9; isotype: BRM[slide ZU10/14], Friedrich Hustedt Diatom Collection."

Discussion

First results

By respecting the standardised format explained in this paper, the specimens cited in *EJT* publications are quickly and efficiently converted into DwC archives by Plazi using a semi-automatic mark-up workflow developed within GoldenGATE. Since the project was launched in early 2018, *EJT* and Plazi have contributed 2155 treatments and 7888 materials citations to major aggregators of biodiversity data, as well as 2288 figures and 7324 bibliographic references: all directly linked to the original publication.

Spot checks carried out on the occurrence records obtained from the encoded specimen citations reveal that the use of the standard formatting greatly improves the precision and quantity of harvested data. Eventual discrepancies in the distributed data compared to the original data (stemming from misinterpretation by the parsing algorithms) can be easily reported ("contact" details available directly of the derived web pages) and corrected (updates performed by Plazi and automatically propagated throughout the network).

Workflow evaluation

The most evident drawback of the GoldenGATE workflow is the approach of retro-conversion, whereby pattern recognition and natural language processing techniques are used to decipher the PDF post-publication. This method is seldom chosen by publishers setting up digital workflows owing to the difficulties connected to a computer interpreting the text correctly. However, for the *EJT* "born-digital" PDF publications, the text is recovered from the rendering instructions embedded in the PDF, which facilitates the task. In addition, a template has been built to automate the conversion process, right up to the fine granularity of material citations. Using Plazi's expertise and the *EJT* standardised formatting as much as possible, we have not only improved the quality and quantity of the data harvested from the articles, but also the speed with which Plazi can perform encoding.

Finding a consensus for best practices

The process of establishing a standard for citing specimens within taxonomic publications exposed the variety of methods currently used and accepted by publishers. While it is important for researchers to benefit from flexibility in journals to present their work in a form that befits the nuances of their subject matter, it is clear that the taxonomic community needs to invest time and effort into reaching a consensus for standardising data and sharing best practices.

Reaching a consensus on working methods that call on community-wide standards, such as those suggested in the Material Citations Formatting Guidelines, could have a considerable impact on issues that contribute to the "taxonomic impediment" but the solutions – even those easily attainable by all – will only be effective through adequate buy-in and through our willingness to change. For example, the appropriate citation of taxonomic works, acts and authorities has long been evoked as a subject that could improve the recognition of taxonomy researchers (Ebach *et al.* 2011), yet there is still much confusion about which course of action should be followed, resulting in anarchical practices across different journals and even from one article to another (Bénichou *et al.* 2018). This is a purely sociological problem, which was directly addressed in the recent paper "Consortium of European Taxonomic Facilities (CETAF) best practices in electronic publishing in taxonomy" (Bénichou *et al.* 2018). An authoritative response is therefore available; the question now remains whether the community will assimilate these best practices.

Best practices and the interoperability of biodiversity data are recurrent themes in CETAF meetings, workgroups and publications (e.g., Keklikoglou et al. 2019), and other leading publishers of taxonomy are equally concerned with these issues, e.g., Pensoft (Penev *et al.* 2019), who also recommends that authors submitting to the journal *ZooKeys* apply the format developed by *EJT* and described in this paper (see: https://zookeys.pensoft.net/about#MaterialsExaminedFormattingGuidelines).

Another pertinent example of the need for controlled vocabularies and well-maintained registers is the citation of digital bio-collections in taxonomical work. Over the past decade, a huge effort has been made by natural history institutions to digitize their collections. This initiative demands serious resources, but also consensus on how to identify, retrieve and cite specimens once they are digitised. Once again, the CETAF has published directives on recommended best practices (Güntsch *et al.* 2017) for the implementation of globally unique identifiers (GUIDs), but bio-collection managers, informatics

communities, researchers and publishers still need to agree on which type of identifier to use, when to assign it and how to cite it, in order for real progress to be made (Guralnick *et al.* 2015). The citation of specimen GUIDs has not been covered within the present guidelines for this very reason but will be addressed in a subsequent paper.

Producing well-formatted occurrence records and handling associated GUIDs may take extra effort from researchers and from editors during manuscript preparation, but they constitute a solid foundation for subsequent study and will greatly contribute to the creation of an interlinked network of robust data sources. The standardised documentation of physical specimens and collection events would facilitate bio-collection management and, if stricter practices for the citation of material were consistently respected by researchers and publishers, improve interpretation by both humans and machines.

Using digital identifiers to assess impact

The traceable nature of linked data could potentially offer alternative metrics for evaluating research impact (McDade *et al.* 2011). However, in order to achieve metrics that are sufficiently representative in terms of quality and quantity to make reliable statements, it is urgent to reach a consensus in the use of controlled vocabularies and identifiers, starting with unique IDs for physical specimens, collections and institutions. This would allow specimens to be tracked from the moment they are collected to their deposition in a repository and throughout any subsequent movements, permitting the identification and retrieval of a given physical specimen despite eventual changes in local collection management. With such a tagging system in place, it would also be possible to infer the number of times a specimen has been used in research papers by tracking its identifier through every publication that cites it, or to imagine using GUIDs to gather statistics about the representation of museum collections across different fields of research (Guralnick *et al.* 2015; Nicolson *et al.* 2019). However, for as long as there is no standardised approach to the citation of biological specimens, any data obtained on a large scale will have a limited power.

As an institutional publisher, a specific area of interest for *EJT* was the identification of institution codes within research articles, indicating where specimens are held. However, the lack of an agreed method for identifying a given institution, in fields other than botany, means that these results cannot currently be exploited. For example, within the test batch of converted *EJT* articles, the Natural History Museum of London was identified by several acronyms: NHM, NHMUK and BMNH. Without a recognised register of official acronyms for zoological collections, as is the case with botanical collections and the institutional register Index Herbariorum, it would be difficult to make any statements about the collections represented by the journal's publications.

If we hope to align our efforts and build a truly seamless global infrastructure for biodiversity knowledge, the institutions dedicated to the curation, documentation and preservation of the primary research matter must take a central role in any proposed scheme. The recommendation in this paper is to follow Index Herbariorum (for botany) and for all other fields the Global Register of Scientific Collections (GRSciColl – a compendium of earlier directories recently merged together and re-launched by GBIF), but this resource will only be as strong as the will of its contributors, i.e., the community must make the effort to contribute, develop and utilise the registry. We therefore recommend that institutions take an active role in enriching and controlling this resource, which offers institutions and collections a foothold in the developing landscape of interlinked data, specimens, publications and institutions. While awaiting the community-curation functionality to be activated by GBIF (intended for late 2019), management should consult the website (https://www.gbif.org/grscicoll) and reflect upon how their institution and collections should be presented. This includes consensus on and communication about a unique and unambiguous acronym to represent their institution (and collections) that should be consistently used as an identifier in all scholarly output.

Conclusion

By contributing data to aggregators such as GBIF and BLR, the *EJT*-Plazi workflow not only improves access to and citation of taxonomic literature, but also means that the specimens studied and their linked taxonomic information are rapidly made available to the wider scientific community. Another huge benefit of this workflow is that the original data source is consistently cited and linked throughout all subsequent representations.

The *EJT*-XML project was presented to the international community of biodiversity scientists at the Biodiversity Next meeting in Leiden, The Netherlands in October 2019 (Bénichou *et al.* 2019), where it was received with enthusiasm. We hope that this paper will now serve as a reference for progress in the field of biodiversity informatics, as well as for *EJT* authors to prepare their manuscripts.

The standard formatting for material citations presented in this paper will be adapted and implemented by four other titles produced by the Muséum national d'histoire naturelle (Paris), and additional journals produced by *EJT* member institutions are expected to follow suit. We hope that any taxonomic journal or author who would like to contribute to the dynamic exchange of biodiversity data might consider adopting a similar approach to formatting and inform Plazi of their efforts.

Within the scope of this project, *EJT* has concentrated on standardising the presentation of specimen data to facilitate semantic enhancement and text mining. The next step will be to work on integrating persistent identifiers for sub-article elements (e.g., treatments), specimens and institutions, with the aim of further improving the accessibility of its publications and establishing alternative metrics for taxonomy. *EJT* intends to take an active role in addressing these issues (CETAF E-publishing workgroup; Biodiversity Next 2019; TDWG working groups) and will continue to collaborate closely with the CETAF to define best practices across European natural science institutions.

Acknowledgements

References

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

Agosti D., Klingenberg C., Sautter G., Johnson N., Stephenson C. & Catapano T. 2007. Why not let the computer save you time by reading the taxonomic papers for you? *Biológico* 69 (suplemento 2): 545–548.

Agosti D., Catapano T., Sautter G. & Egloff W. 2019a. The Plazi Workflow: The PDF prison break for biodiversity data. *Biodiversity Information Science and Standards* 3: e37046. https://doi.org/10.3897/biss.3.37046

Agosti D., Catapano T., Sautter G., Kishor P., Nielsen L., Ioannidis-Pantopikos A., Bigarella C., Georgiev T., Penev L. & Egloff W. 2019b. Biodiversity Literature Repository (BLR), a repository for FAIR data and publications. *Biodiversity Information Science and Standards* 3: e37197. https://doi.org/10.3897/biss.3.37197

Bénichou L., Dessein S., Duin D., Gerard I., Higley G. & Martens K. 2011. Towards a *European Journal of Taxonomy* (*EJT*). *Libreas*. Library Ideas 18.
Available from https://libreas.eu/ausgabe18/texte/09benichou_etal.htm [acessed 8 Nov. 2019].

Bénichou L., Martens K., Higley G., Gérard I., Dessein S., Duin D. & Costello M. J. 2012. *European Journal of Taxonomy*: A public collaborative project in Open Access scholarly communication. *Scholarly and Research Communication* 4 (1): 010134, 16 p. http://src-online.ca/index.php/src/article/view/37/114

Bénichou L., Gérard I., Laureys E. & Price M.J. 2018. Consortium of European Taxonomic Facilities (CETAF) best practices in electronic publishing in taxonomy. *European Journal of Taxonomy* 475: 1–37. https://doi.org/10.5852/ejt.2018.475

Bénichou L., Gerard I., Chester C., Agosti D. 2019. The European Journal of Taxonomy: Enhancing taxonomic publications for dynamic data exchange and navigation. *Biodiversity Information Science and Standards* 3: e37199. https://doi.org/10.3897/biss.3.37199

Catapano T. 2010. TaxPub: An extension of the NLM/NCBI journal publishing DTD for taxonomic descriptions. *Journal Article Tag Suite Conference (JATS-Con) Proceedings 2010*. National Center for Biotechnology Information, Bethesda (MD).
Available from https://www.ncbi.nlm.nih.gov/books/NBK47081/ [accessed 22 Jul. 2019]

Côtez E., Mabille A., Chester C., Rocklin E., Deroin T., Desutter-Grandcolas L., Lesur J., Merle D., Robillard T. & Bénichou L. 2018. 1802–2018: 220 ans d'histoire des périodiques au Muséum. *Adansonia* 40 (1): 1–41. https://doi.org/10.5252/adansonia2018v40a1

cOAlition S. 2019. *Making Full and Immediate Open Access a Reality*. Science Europe, Brussels. Available from https://www.coalition-s.org/wp-content/uploads/271118_cOAlitionS_Guidance.pdf [accessed 6 Aug. 2019].

Dikow T. 2019. Shaping our Taxonomic Legacy through Openly Sharing Primary Biodiversity Data in Taxonomic Revisions. *Biodiversity Information Science and Standards*. 3: e37062.
 https://doi.org/10.3897/biss.3.37062

Ebach M., Valdecasas A.G & Wheeler Q. 2011. Impediments to taxonomy and users of taxonomy: accessibility and impact evaluation. *Cladistics*. 27: 550–557.
https://doi.org/10.1111/j.1096-0031.2011.00348.x

Groombridge B. 1992. *Global Biodiversity: Status of the Earth's Living Resources. A Report Compiled by the World Conservation Monitoring Centre*. Chapman & Hall, London/Glasgow/New York/Tokyo/Melbourne/Madras. https://doi.org/10.1017/S0016756800011511

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. 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 E.D.M. 2015. Community next steps for making globally unique identifiers work for biocollections data. *ZooKeys* 494: 133–154. https://doi.org/10.3897/zookeys.494.9352

Hardisty A., Roberts D. & Community TBI. 2013. A decadal view of biodiversity informatics: challenges and priorities. *BMC Ecology* 13 (1): 16. https://doi.org/10.1186/1472-6785-13-16

Heywood V.H. 1995. *The Global Biodiversity Assessment.* United Nations Environment Programme Cambridge, Cambridge University Press, Cambridge.

Hobern D., Baptiste B., Copas K., Guralnick R., Hahn A., van Huis E., Kim E., McGeoch M., Naicker I., Navarro L., Noesgaard D., Price M., Rodrigues A., Schigel D., Sheffield C., Wieczorek J. 2019. Connecting data and expertise: a new alliance for biodiversity knowledge. *Biodiversity Data Journal* 7: e33679. https://doi.org/10.3897/BDJ.7.e33679

ICZN 1999. *International Code of Zoological Nomenclature, 4th Edition*. International Trust for Zoological Nomenclature, London. Available from http://iczn.org/iczn/index.jsp [accessed 5 Nov. 2019].

ICZN 2012. Amendment of Articles 8, 9, 10, 21 and 78 of the International Code of Zoological Nomenclature to expand and refine methods of publication. *Bulletin of Zoological Nomenclature* 69 (3): 161–169. Available from http://iczn.org/content/iczn-amendment-electronic-publication [accessed 5 Nov. 2018].

IPCC 2018. Summary for Policymakers. *In*: Masson-Delmotte V., Zhai P., Pörtner H.-O., Roberts D., Skea J., Shukla P.R., Pirani A., Moufouma-Okia W., Péan C., Pidcock R., Connors S., Matthews J.B.R., Chen Y., Zhou X., Gomis M.I., Lonnoy E., Maycock T., Tignor M. & Waterfield T. (eds) *Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty*. World Meteorological Organization, Geneva. Available from https://www.ipcc.ch/sr15/ [accessed 10 Sep. 2019].

Library of Congress. 2019. Recommended formats statement. I. Textual works and musical compositions. Available from https://www.loc.gov/preservation/resources/rfs/textmus.html#digital [accessed 24 September 2019].

McCook L.J., Ayling T., Cappo M., Choate J.H., Evans R.D., De Freitas D.M., Heupel M., Hughes T.P., Jones G.P., Mapstone B., Marsh H., Mills M., Molloy F.J., Pitcher C.R., Pressey R.L., Russ G.R., Sutton S., Sweatman H., Tobin R., Wachenfeld D.R. & Williamson D.H. 2010. Adaptive management of the Great Barrier Reef: A globally significant demonstration of the benefits of networks of marine reserves. *Proceedings of the National Academy of Sciences* 107 (43): 18278–18285. https://doi.org/10.1073/pnas.0909335107

McDade L.A., Maddison D.R., Guralnick R., Piwowar H.A., Jameson M.L., Helgen K.M., Herendeen P.S., Hill A. & Vis M.L. 2011. Biology needs a modern assessment system for professional productivity. *BioScience* 61 (8): 619–625. https://doi.org/10.1525/bio.2011.61.8.8

McNeill J., Barrie F.R., Buck W.R., Demoulin V., Greuter W., Hawksworth D.L., Herendeen P.S., Knapp S., Marhold K., Prado J., Prud'homme van Reine W.F., Smith G.F., Wiersema J.H. & Turland N.J. 2012. International Code of Nomenclature for algae, fungi, and plants (Melbourne Code) adopted by the Eighteenth International Botanical Congress Melbourne, Australia, July 2011. *Regnum Vegetabile* 154: 1–140. Available from https://www.iapt-taxon.org/melbourne/main.php [accessed 15 Nov. 2019].

Miller J., Braumuller Y., Kishor P., Shorthouse D., Dimitrova M., Sautter G. & Agosti D. 2019. Mobilizing data from taxonomic literature for an iconic species (Dinosauria, Theropoda, *Tyrannosaurus rex*). *Biodiversity Information Science and Standards* 3: e37078. https://doi.org/10.3897/biss.3.37078

Miller J., Dikow T., Agosti D., Sautter G., Catapano T., Penev L., Zhang Z.-Q., Pentcheff D., Pyle R., Blum S., Parr C., Freeland C., Garnett T., Ford L.S., Muller B., Smith L., Strader G., Georgiev T. & Benichou L. 2012. From taxonomic literature to cybertaxonomic content. *BMC Biology* 10: 87. https://doi.org/10.1186/1741-7007-10-87

Morrissey S.M., Meyer J., Bhattarai S., Kurdikar S., Ling J., Stoeffler M. & Thanneeru U. 2010. Portico: A Case Study in the Use of the Journal Archiving and Interchange Tag Set for the Long Term Preservation of Scholarly Journals. *In*: *Journal Article Tag Suite Conference (JATS-Con) Proceedings 2010*. Bethesda (MD) National Center for Biotechnology Information. Available from https://www.ncbi.nlm.nih.gov/books/NBK47087/ [accessed 24 Sep. 2019].

Nicolson N. & Tucker A. 2017. Identifying Novel Features from Specimen Data for the Prediction of Valuable Collection Trips. In: Adams N., Tucker A., Weston D. (eds) Advances in Intelligent Data Analysis XVI. IDA 2017. Lecture Notes in Computer Science, vol 10584. Springer, Cham. https://doi.org/10.1007/978-3-319-68765-0_20

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

Penev L., Hagedorn G., Mietchen D., Georgiev T., Stoev P., Sautter G., Agosti D., Plank A., Balke M., Hendrich L. & Erwin T. 2011. Interlinking journal and wiki publications through joint citation: Working examples from *ZooKeys* and Plazi on Species-ID. *ZooKeys* 150: 1–12. https://doi.org/10.3897/zookeys.90.1369

Penev L., Georgiev T., Senderov V, Dimitrova M. & Stoev P. 2019. The Pensoft data publishing workflow: The FAIRway from articles to Linked Open Data. *Biodiversity Information Science and Standards* 3: e35902. https://doi.org/10.3897/biss.3.35902

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. *Journal Article Tag Suite Conference (JATS-Con) Proceedings 2012*. National Center for Biotechnology Information, Bethesda (MD).
Available from https://www.ncbi.nlm.nih.gov/books/NBK100351/ [accessed 15 Nov. 2019].

Sautter G., Böhm K. & Agosti D. 2007. Semi-automated XML markup of biosystematic legacy literature with the GoldenGATE editor. *Pacific Symposium on Biocomputing* 12: 391–402. https://10.5281/zenodo.55665

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. Koeltz Botanical Books, Glashütten. https://doi.org/10.12705/Code.2018

Wägele H., Klussmann-Kolb A., Kuhlmann M., Haszprunar G., Lindberg D., Koch A. & Wägele J.W. 2017. The taxonomist – an endangered race. A practical proposal for its survival. *Frontiers in Zoology 2011* (8): 1–7. https://doi.org/10.1186/1742-9994-8-25

Wieczorek J., Bloom D., Guralnick R., Blum S., Döring M., Giovanni R., Robertson T. & Vieglais D. 2012 Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE* 7 (1): e29715. https://doi.org/10.1371/journal.pone.0029715

Wilkinson M., Dumontier M., Aalbersberg I.J., Appleton G., Axton M., Baak A., Blomberg N., Boiten J.-W., Bonino da Silva Santos L.O., Bourne P., Bouwman J., Brookes A. J. , Clark T., Crosas M., Dillo I., Dumon O., Edmunds S., Evelo C., Finkers R. & Mons B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3: 160018. https://doi.org/10.1038/sdata.2016.18