European Journal of Taxonomy: a deeper look into a decade of data

Laurence Bénichou, Marcus Guidoti, Isabelle Gérard, Donat Agosti, Tony Robillard, Fabio Cianferoni

European Journal of Taxonomy 782: 173–196
https://doi.org/10.5852/ejt.2021.782.1597

ISSN 2118-9773
www.europeanjournaloftaxonomy.eu
2021 · Bénichou L. *et al.*

Opinion paper

European Journal of Taxonomy: a deeper look into a decade of data

Laurence BÉNICHOU [1,*] Marcus GUIDOTI [2], Isabelle GÉRARD [3], Donat AGOSTI [4], Tony ROBILLARD [5] & Fabio CIANFERONI [6]

[1] Service des Publications scientifiques, Muséum national d'histoire naturelle, 57 rue Cuvier, CP 41, 75231 Paris Cedex 05, France.
[2] Plazi, Porto Alegre, Brazil.
[3] Publications Department, Royal Museum for Central Africa, 13, Leuvensesteenweg, 3080 Tervuren, Belgium.
[4] Plazi, Zinggstrasse 16, 3007 Bern, Switzerland.
[5] Institut de Systématique, Evolution et Biodiversité (ISYEB), Muséum national d'histoire naturelle, CNRS, SU, EPHE, UA, 57 rue Cuvier, CP 50, 75231 Paris Cedex 05, France.
[6] Istituto di Ricerca sugli Ecosistemi Terrestri (IRET), Consiglio Nazionale delle Ricerche (CNR), Via Madonna del Piano 10, 50019 Sesto Fiorentino (Florence), Italy.
[6] Zoologia, "La Specola", Museo di Storia Naturale, Università degli Studi di Firenze, Via Romana 17, 50125 Florence, Italy.

[*] Corresponding author: laurence.benichou@mnhn.fr
[2] Email: guidoti@plazi.org
[3] Email: isabelle.gerard@africamuseum.be
[4] Email: agosti@plazi.org
[5] Email: tony.robillard@mnhn.fr
[6] Email: fabio.cianferoni@cnr.it

Abstract. The *European Journal of Taxonomy (EJT)* is a decade-old journal dedicated to the taxonomy of living and fossil eukaryotes. Launched in 2011, the *EJT* published exactly 900 articles (31 778 pages) from 2011 to 2021. The journal has been processed in its entirety by Plazi, liberating the data therein, depositing it into TreatmentBank, Biodiversity Literature Repository and disseminating it to partners, including the Global Biodiversity Information Facility (GBIF) using a combination of a highly automated workflow, quality control tools, and human curation. The dissemination of original research along with the ability to use and reuse data as freely as possible is the key to innovation, opening the corpus of known published biodiversity knowledge, and furthering advances in science. This paper aims to discuss the advantages and limitations of retro-conversion and to showcase the potential analyses of the data published in *EJT* and made findable, accessible, interoperable and reusable (FAIR) by Plazi. Among others, taxonomic and geographic coverage, geographical distribution of authors, citation of previous works and treatments, timespan between the publication and treatments with their cited works are discussed. Manually counted data were compared with the automated process, the latter being analysed and discussed. Creating FAIR data from a publication results in an average multiplication factor of 166 for additional access through the taxonomic treatments, figures and material citations citing the original publication in TreatmentBank, the Biodiversity Literature Repository and the Global Biodiversity Information Facility. Despite the advances

in processing, liberating data remains cumbersome and has its limitations which lead us to conclude that the future of scientific publishing involves semantically enhanced publications.

Keywords. Diamond open access, taxonomic data, FAIR principles, interoperability, data liberation.

Bénichou L., Guidoti M., Gérard I., Agosti D., Robillard T. & Cianferoni F. 2021. *European Journal of Taxonomy*: a deeper look into a decade of data. *European Journal of Taxonomy* 782: 173–196.
https://doi.org/10.5852/ejt.2021.782.1597

Introduction

Natural History Institutions (NHIs), including natural history museums, herbaria, botanical gardens, and other research institutions, have traditionally been founded to contribute to the understanding of the natural world and to disseminate this knowledge. Their core mission can be divided into three main objectives: 1) to establish and maintain biological collections (carried out by herbaria, zoological archives, etc.); 2) to conduct scientific research associated with the collections; and 3) to disseminate scientific knowledge within the scientific community and to the general public. This is one of the reasons NHIs have been scientific publishers since their creation, some of them since the end of the 18th century (Bénichou *et al.* 2013).

The *European Journal of Taxonomy* (*EJT*), a journal jointly published by several NHIs in Europe, was created in 2011 by its founding institutions (National Museum of Natural History, Paris; Natural History Museum, London; Royal Belgian Institute of Natural Sciences, Brussels; Royal Museum for Central Africa, Tervuren; Meise Botanic Garden) with the intent to enable its members to collectively tackle the strategic and technical challenges related to the visibility, access, format and financial structure of academic journals, especially publicly-funded titles.

With the digital age new challenges arose. The creation of digital copies and the Internet allows instantaneous access to publications. The nature of the Internet, however, not only allows linking one page to another or a cited publication to its digital copy, but also linking a cited specimen to its digital copy, a cited gene sequence to the actual DNA sequence, a trait to the definition of the trait, or a taxonomic name to the taxonomic treatment. This also implies that data from within a publication need to be findable, accessible, interoperable, and reusable (FAIR: Wilkinson *et al.* 2016). Perfect fully-automated data extraction from unstructured text will hardly ever be possible, because of the degree of freedom to represent data, such as material citations or a bibliographic reference, with its constraints to be comprehensible by a human, uses of abbreviations and removal of repetitive texts and back references. Furthermore, machine processing depends on high quality text recognition and decoding of the portable document format (PDF) of the articles, text flows over pages, font sizes, etc. to recognize blocks of texts such as taxonomic treatments.

Openness

Openness is now a prerequisite set by most science foundations and funders in Europe that scholarly publications should be made accessible to all, free of charge for readers (and ideally at no cost to the author, through Diamond open access) and the movement towards open access has definitely won some important battles. It began with the Budapest open access Initiative (2002) and the Berlin Declaration (2003) when about 650 signatories (including funders, policy makers, governments and scientific institutions) committed to open access for the dissemination of research output. The San Francisco Declaration on Research Assessment (DORA 2013) sought to include scientific merit in their assessment, which could be the discovery of new species or subsequent taxonomic treatments in publications. The 96 research organisations and 210 individuals who signed the Bouchout Declaration for Open Biodiversity Knowledge Management in 2014 upheld the principles to provide free and open access to their digital

resources. The Amsterdam Call for Action on Open Sciences (2016) recommended full open access for taxpayer-financed research results by 2020 and to manage a fast transition to Gold open access (with Author Processing Charges, APC). In 2018, the European Commission Recommendation on Access to and Preservation of Scientific Information encouraged creation of infrastructures and synergies, development of policies, and pleaded for publications and data availability and preservation. Currently most European countries are involved in Plan S. "Plan S requires that, from 2021, scientific publications that result from research funded by public grants must be published in compliant Open Access journals or platforms" (https://www.coalition-s.org/). It constitutes a major step towards open access. Honouring such a commitment in such a short time frame requires – for the sake of simplicity – having recourse to two existing and non-exclusive roads:

1. The first road consists of publishing journals and books in open access ("Gold open access"), which implies either institutional funding of platforms or journals operated by institutions or scientific communities ("Diamond open access") on the one hand, or the payment of publication fees by authors to open access business models on the other hand; of course national recommendations to authors are in favor of journal with usual fees (even though some companies ask too high APC).

2. The second road consists of depositing and making the publication available in a reliable open access repository ("Green open access") without embargo or with the shortest possible embargo, taking respective national legislations into account.

Plan S still tends to focus on the APC option. Meanwhile, in line with the journal's rationale and philosophy, the *EJT* board has naturally opted for the Diamond open access route. Not only can the reader access the articles freely and reuse them according to the CC-BY (Creative Commons – Attribution) license, but no fees (APCs) are required of the authors. All the costs are borne and paid for by the member institutions.

While open access may be a prerequisite for any modern European academic journal, it is no longer enough. In light of recent academic movements such as the FAIR Data Principles (Findable, Accessible, Interoperable, Reusable) for Open Science (Wilkinson *et al.* 2016), the DORA declarations, and international initiatives to mobilise biodiversity data – e.g., Global Biodiversity Information Facility (GBIF: (https://gbif.org), the Catalogue of Life (https://www.catalogueoflife.org), DiSSCo (https://www.dissco.eu) – it becomes an increasing priority to ensure that all data contained within *EJT* are fed to all relevant databases of the biodiversity research fields. This is impossible to accomplish unless journals are openly available.

EJT rationale

The *European Journal of Taxonomy* was launched in 2011, under the 6th Framework Programme of the European Distributed Institute of Taxonomy (EDIT) Research Network of Excellence. The rationale behind the creation of the journal was to 1) prevent the further fragmentation of taxonomic information in small, sometimes obscure, journals; 2) bring together all actors involved in scientific publishing within natural history institutions in Europe as a means to avoid their isolation and empower them so they could face the technological revolution; and 3) enable the institutions to control their own editorial policy and establish efficient publishing skills within the institutions (for a more in-depth history of the creation of *EJT*, see Bénichou *et al.* 2010). The key aspects of the journal have always been to pool resources and expertise in order to set up a cross-institutional strategy at a European level for taxonomy. It is essential to document descriptive taxonomic research as a basis for safeguarding our planet's biodiversity. Providing access to taxonomic research results – including taxonomic treatments, well-demarcated sections of text about one taxon (Catapano 2010), material citations referencing the underlying specimens (TDWG 2021), digital representations of specimens, images, sound tracks and legacy publications – remains a core mission of natural history institutions and their staff.

The dissemination of original research along with the ability to use and reuse data as freely as possible is the key to opening the corpus of known published biodiversity knowledge, furthering advances in science and ultimately innovation. A further step is to add as many identifiers as possible to link to cited elements such as gene sequences using accession codes, specimens using a persistent specimen identifier issued by the Consortium of European Taxonomic Facilities (CETAF) institutions, or articles using digital object identifiers (DOI), and supported by the recently EU Horizon 2020 funded Biodiversity Community Integrated Knowledge Library (BiCIKL) project.

The *EJT* team strongly supports the idea that publishing is part of the research process (Fig. 1), and that natural history institutions should be able to preserve, expand, or recover their in-house publishing expertise. Even though natural history institutions and botanical gardens share a long tradition of scientific publishing (most have been scientific publishers since their creation), many chose to outsource their publication needs and consequently lost their publishing expertise in the 1990s (Bénichou *et al.* 2013). In-house publishing skills are now crucial to complete the transition to online publishing; these skills are even more necessary to embed the institution in an open science approach by making the publications, and especially the data therein, machine-readable and immediately reusable. Producing in-house ensures the institutions control the way they disseminate the scientific information they produce.

Currently, *EJT* belongs to 10 institutions and is run by a team of 11 desk editors hired by the institutions and scattered throughout Europe. Most of the desk editors run other journals for their institution and benefit from all the expertise and technology developed by *EJT* so they can reuse *EJT* tools and best practices to better serve the taxonomic community. Ten years after its creation, *EJT* fully plays its role as an incubator of innovation and as a pioneer in data integration on behalf of its members.

The promotion of taxonomy, systematics and collection-based research via scientific publishing is a vision that *EJT* shares with the CETAF, which 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. Moreover, *EJT* aims to liberate the data contained within its articles to the ecosystem of dynamic, stable, free-to-use and interconnected platforms available on the Web such as GBIF.

To achieve its goals in terms of data liberation, *EJT* signed a contract with Plazi in 2017 for the retro-conversion of the articles published in *EJT*, i.e., decode and recreate data imprisoned in a PDF or available as print only by adding semantic meaning using extensible markup language (XML). The workflow, based on the Plazi workflow (Agosti & Egloff 2009) and now integrated in TreatmentBank, has been described in depth by Côtez *et al.* 2018.

This process begins with the decoding of the original portable document format (PDF) into text and text streams, followed by the semantic enhancement at word to section level, for example by adding a taxonomic name XML tag to a taxonomic name, or a treatment XML tag around the section including a treatment. The creation of FAIR data for treatments and figures uses the Biodiversity Literature Repository including a DataCite digital object identifier (DOI), rich customized metadata, a licence, and allows output in both human- and machine-readable format (e.g., JSON). The Zenodo repository at the European Organization for Nuclear Research (CERN) has been chosen with the hindsight of its sustainability. With its focus on liberating data from publications, the Biodiversity Literature Repository complements the Biodiversity Heritage Library with its main focus on providing access to publications (BHL & Plazi 2021). Material citations are made FAIR by TreatmentBank and GBIF with a persistent identifier and metadata including the respective Plazi identifier, accessible through the GBIF application programming interface (API). When possible, collection, specimen and accession codes are attributed with

their respective identifiers. For later data search and import into GBIF, taxonomic names are attributed with their taxonomic hierarchy from Catalogue of Life.

The last step is to disseminate the converted articles to GBIF as a new treatment article data set packaged in a Darwin Core Archive. This process includes a push notification to GBIF which then collects and integrates the respective Darwin Core Archive (DWCA). It normally takes 5 minutes from the end of conversion in TreatmentBank to the notification and near-instantaneous integration in GBIF. Any changes in the annotation on the TreatmentBank side are similarly handled, ensuring that data sets in GBIF are synchronized and updated. This allows immediate responses to feedback regarding the conversion results (Plazi 2021). In addition to the original workflow, an automated quality control tool with subsequent manual curation has been integrated in the workflow to provide output defined by the needs of *EJT*. The desired output defines the settings of the gatekeeper – a tool that checks whether data meet a predefined standard in data quality and granularity – in this case whether the export to specific users is permitted (Simoes *et. al.* 2021).

Fig. 1. Publishing as part of the research life cycle. Once liberated, the data imprisoned in a single PDF immediately becomes an integral part of research and is reused and cited. Green figures illustrate the TreatmentBank workflow to liberate, FAIRize and disseminate data, and the Biodiversity Literature Repository, the long-term repository for the data. GBIF is one of the main reusers of the data, providing specific access to the material citations and making them FAIR. Finally, scientists reuse data from within GBIF for their research, for example new revisions, which then become incorporated in the liberation process, thus closing the research life cycle.

There are limitations to the output of the retro-conversion due to available financial resources, progress in automation and quality control tools. Whilst humans adapt at understanding texts and structures, a machine cannot handle a simple omission of symbols like a "." in "sp. nov.", and will be unable to find this new species. Finding treatments depends on structural clues such as the presence of a taxonomic name in a title at a given font size, which, in case of editorial changes, may not work anymore. Finding omitted treatments poses an interesting challenge that can partially be solved by using statistics, for instance by defining the expected maximal size of a treatment that cannot be exceeded without generating an error message. Without reverting to complete human control, which defeats the purpose of automation, deviation from the real content is to be expected.

Plazi's approach is to liberate data and then use a series of quality control steps and human curation. It is deemed better to release data, including possible errors, so as to provide a link to the source of any specimen cited of a given species, and to provide access to the treatment online to check immediately online.

This solution is based on Plazi's ability to react almost instantaneously to feedback which then affects all downstream products, including those uploaded to GBIF.

FAIR data for figures and treatments include a DataCite digital object identifier in Zenodo. For taxonomic treatments, "taxonomictreatment" is a specific publication subtype and inserted custom metadata linking terms to external – in the case of taxonomy, specific – vocabularies such as Darwin Core have been added in Zenodo. For material citation, i.e., the citation in the literature of a specimen, and to separate them from other occurrences in GBIF, the new term "MaterialCitation" has been accepted in the Darwin Core vocabulary.

The initial experiences with conversion of a publication to XML led to the publication of author guidelines (Chester *et al.* 2019), that have also been adopted by Pensoft journals, and implemented in *EJT* in coordination with processing of the material citations into its elements (such as collecting country, collector, collecting date or specimen and collection codes). In a first phase, before the enforcement of the above guidelines, material citations were not processed in detail, resulting in less data for analyses than after the implementation in 2019. As a result of the conversion, digital tools like dashboards can provide a novel automated look at the data in a journal.

Here we review the taxonomic knowledge generated in this first decade of *EJT* and explore the new opportunities offered by access to the data liberated from the journal.

Method

Data extraction from a PDF-based publication using Plazi's TreatmentBank data liberation research infrastructure follows the Plazi workflow outlined in Agosti & Egloff (2009), Côtez *et al.* (2018), Chester *et al.* (2019) and Fawcett *et al.* (in press, fig. 2). This highly-automated process includes the creation of the figures, treatments and material citations as FAIR data, in the first two cases via the Biodiversity Literature Repository as its repository, and in the third case via GBIF. Unique persistent identifiers are also provided for each treatment as well as for material citations in TreatmentBank, both used in GBIF to link back to the respective sources in TreatmentBank. The automation is based on the GoldenGate Imagine software program (Sautter *et al.* 2007*)*, creation of a template that describes the layout necessary for machine processing, quality control including error reporting and a user interface for error removal, a gatekeeper that prevents erroneous documents from being uploaded, tools to upload and interlink the figures, treatments and original publications to Biodiversity Literature Repository (including adding metadata and cross-links between the three objects), and finally the repackaging of the processed article as a Darwin Core Archive that is retrieved from TreatmentBank by GBIF. Within TreatmentBank, the

GBIF use case defines the criteria set to create fit-for-use data to export to GBIF by a gatekeeper that checks whether the data quality statistics in each processed article are ready for export. GBIF itself then reuses the included data to create a view of the article (see this example) as a treatment article database for each treatment, including the metadata, the text and figures, and material citation as occurrence. Each

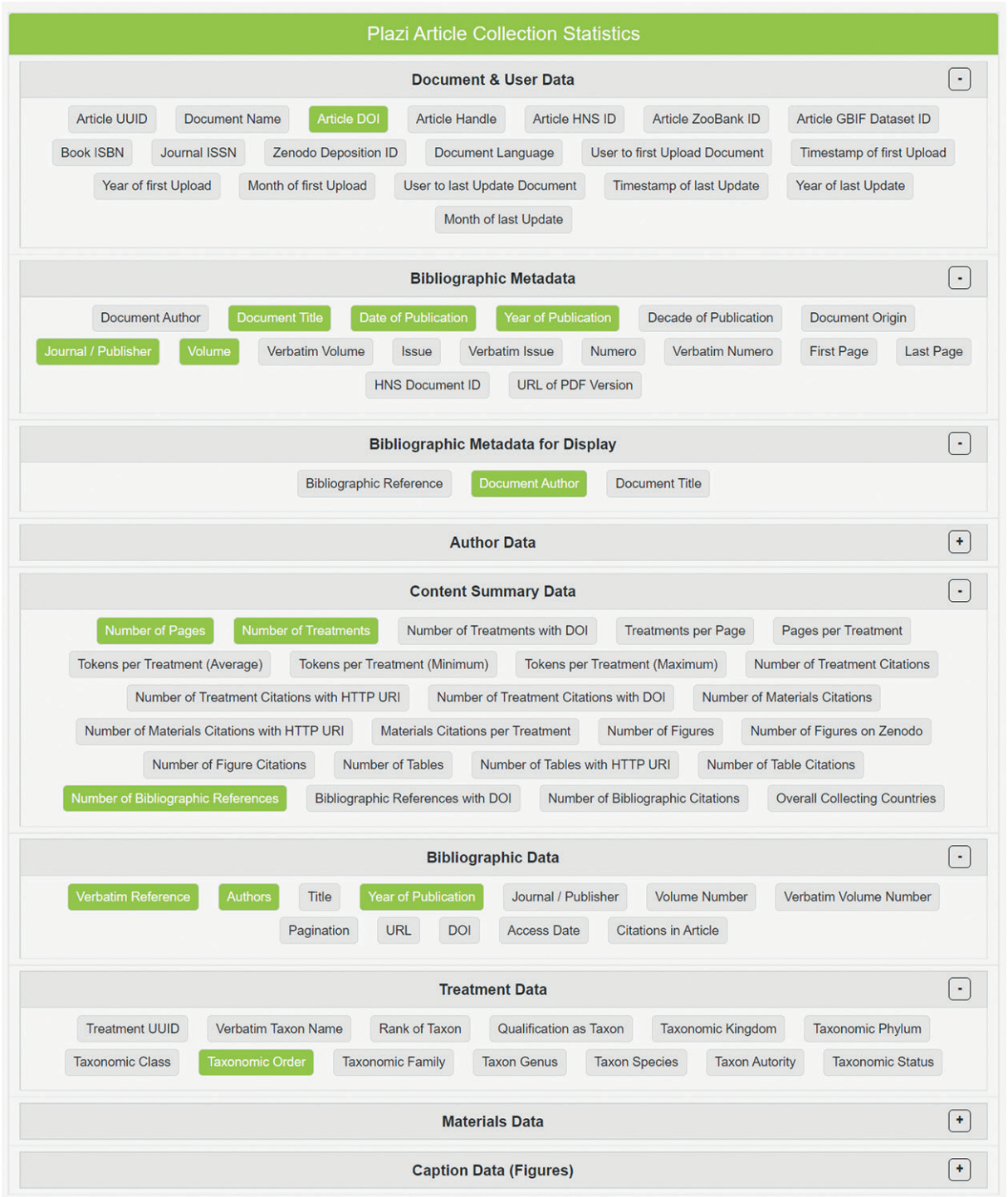

Fig. 2. Research within TreatmentBank's Article Collection Statistics to extract the bibliographic references (source: https://tb.plazi.org/GgServer/dioStats)

taxonomic name is attributed with the taxonomic hierarchy obtained from Catalogue of Life, or from GBIF's taxonomy backbone if the former is unavailable. Treatments and figures as FAIR data include in Biodiversity Literature Repository a DataCite DOI, rich customized metadata, a license, and output in both human- and machine-readable format (e.g., JSON).

In the present paper, we analyse the data extracted from the journals provided from two main sources, (1) TreatmentBank/Plazi and (2) an internal FileMaker database used by *EJT*'s editorial office to monitor progress on the journal. All analyses are based on the data up to 10th September 2021.

The data liberated by Plazi and included in TreatmentBank and Biodiversity Literature Repository and partners from 2011–2021 are available through TreatmentBank and articles data access (Fig. 2). All 900 publications have been processed. The granularity of the data has been augmented in 2019 to include all material citations thanks to the fact that the articles published since this date have been structured according to the precise guidelines set up by *EJT* (Chester *et al.* 2019) to allow greater accuracy of extracted data. The data for the analyses have been retrieved from TreatmentBank, then cleaned, wrangled, and analysed as needed using Python. The scripts and datasets (*EJT* analysis data 2021) are published on Zenodo. Some of the data included in the analyses, especially concerning treatment citations, have to be considered provisional since data clean-up is ongoing. However, all data is freely available on Plazi stats for further processing. Similarly, the extraction of author affiliation and annotation of collections or institutions is in its infancy, but its use provides an indication of its usefulness and potential. A typical search to retrieve data from TreatmentBank is illustrated in Fig. 2.

The following link provides an example of a query on TreatmentBank: http://tb.plazi.org/GgServer/dioStats/stats?outputFields=bib.year+bib.source+cont.pageCount+cont.treatCount+cont.treatCitCount+cont.matCitCount+cont.figCount+cont.tabCount+cont.bibRefCount+cont.countries&groupingFields=bib.year+bib.source&FP-bib.source=%22European%20Journal%20of%20Taxonomy%22&format=HTML

For the analyses and visualization, the data is available via BLR (*EJT* Analyses Data 2021).

The FileMaker database was compiled using data from TreatmentBank article statistics and from the *EJT* Editorial management system, and serves as the "gold standard" for comparing the output of Plazi's conversions. This includes the number of articles published, pages, new families, genera and species. Another table was created with the bibliographic references cited in *EJT* publications to calculate the timeline of production for each article and to calculate all figures regarding the age of bibliographic references cited within the articles published in *EJT*.

The internal FileMaker database has been compiled since the beginning of the journal to monitor certain figures such as the number of pages, number of new taxa, number of articles published, timeline during peer-review process, timeline during production process, and geographical origins of first authors based on their affiliation. These data have been manually inserted in the databases and make it possible to compare and analyse the accuracy of the data extracted in TreatmentBank.

A way to measure the increased access to a publication in comparison to its original single citable DOI is to calculate the journal multiplication factor as the sum of the liberated FAIR data. In the case of *EJT*, this includes three additional access points to the article via Biodiversity Literature Repository, TreatmentBank and GBIF; the access via treatments to the original article from Biodiversity Literature Repository, TreatmentBank and GBIF; the figures from Biodiversity Literature Repository, TreatmentBank, GBIF and Ocellus; and from material citations via GBIF. For example for a publication with 10 figures, 10 treatments, 30 material citations this is calculated as # Publication * 4 [in original, Biodiversity Literature Repository, TreatmentBank, GBIF] + # figures * 2 [in Biodiversity Literature Repository, GBIF] + # treatments

* 3 [in Biodiversity Literature Repository, TreatmentBank, GBIF] + # material citations * 1 [GBIF] = (1 * 4) + (10 * 2) + (10 * 3) + (30 * 1) = 84.

Results and Discussion

Bibliographic data

Most of the data discussed below are derived from the *EJT* internal database. However, all of them can be also extracted from TreatmentBank and are consistent with the monitoring done by the *EJT* team over the last decade.

Number of articles & published pages

From its first volume (9th September 2011) to volume 767 (6th September 2021), *EJT* has published 900 articles covering 31 778 pages. The number of pages available on TreatmentBank have increased since then and can be viewed here.

The journal's production capacity depends on the involvement of the institutions and thus the people hired to structure, copy-edit, proofread and publish the papers, ensure a fair and proper peer-review process, run the journal on a daily basis, and maintain the website.

EJT's publishing team has evolved over the past ten years according to the number of institutions involved in the journal and the in-kind time allocated to it by each institution. When founding *EJT*, we estimated based on our experience (Bénichou *et al.* 2010) that a desk editor working full-time for an academic journal with the high quality expected for *EJT* should be able to publish around 1000 pages per year. Altogether 11 desk editors scattered across the institutions participating in *EJT* work part-time, which represents 4.64 full-time equivalents. *EJT* thus has a projected production capacity of 4640 pages a year. Fig. 3

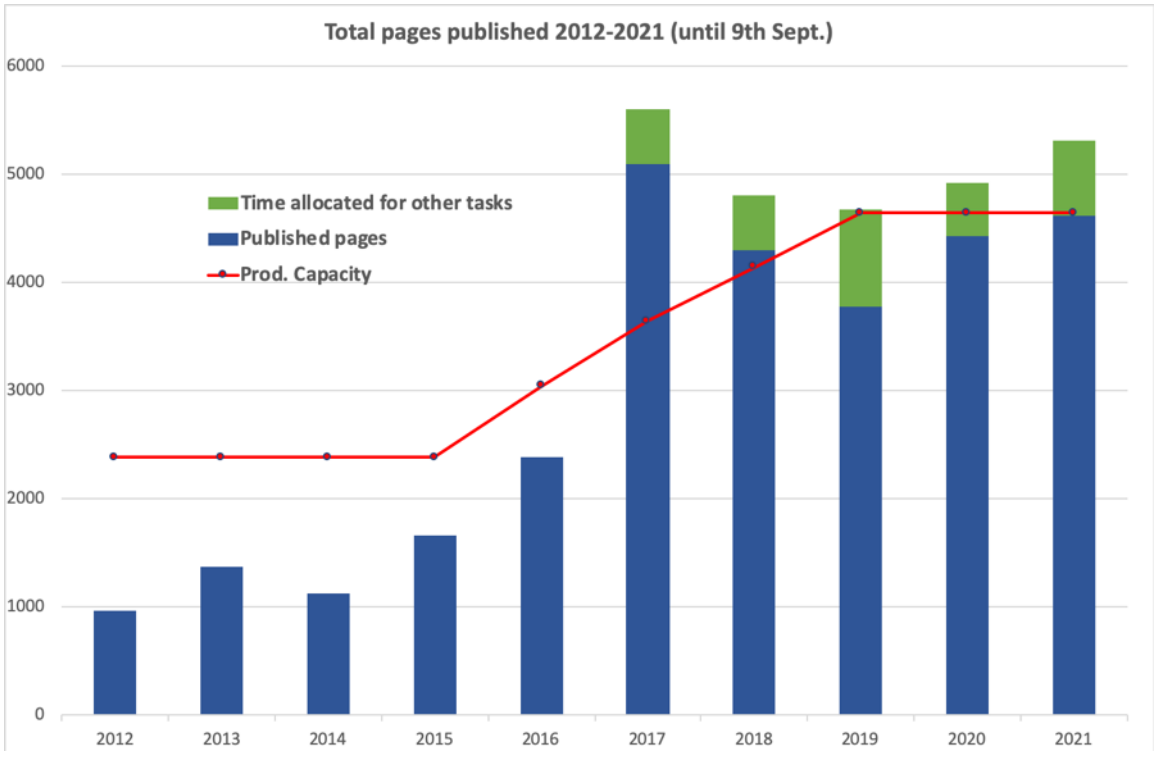

Fig. 3. Capacity production of *European Journal of Taxonomy.*

Table 1. FileMaker dataset showing the average number of references cited within *EJT* articles.

Total	Average per article
31 523 pages published, 897 articles	35 pages per article
42 278 references cited	47 references per article

shows that the current production often exceeds the team's projected capacity, which demonstrates its commitment. The time allocated by the team to other strictly production-related tasks (e.g., administrative work, addressing authors, referees and technical issues with the submission system) is represented in green in the graph, the time allocated to administrative tasks correlates with the number of submissions.

The average number of pages published per paper is 35 pages. Data is corroborated by TreatmentBank and can be viewed here.

Monographs (defined as an article of at least 50 pages) represent 15% of the total articles published, and 43% of all pages published in the journal so far. The publication of monographs tends to increase, with an increasing number of pages per monograph, yet their publication remains welcome in *EJT*.

Bibliographic references

To analyse the average number of bibliographic references cited within *EJT* articles, we have extracted the data from TreatmentBank as shown in Fig. 4. The data was then imported into a relational database created in FileMaker to generate the figures provided below. The dataset comprises all articles published from September 2011 to 6th September 2021 (i.e., 897 articles found in TreatmentBank) (Table 1).

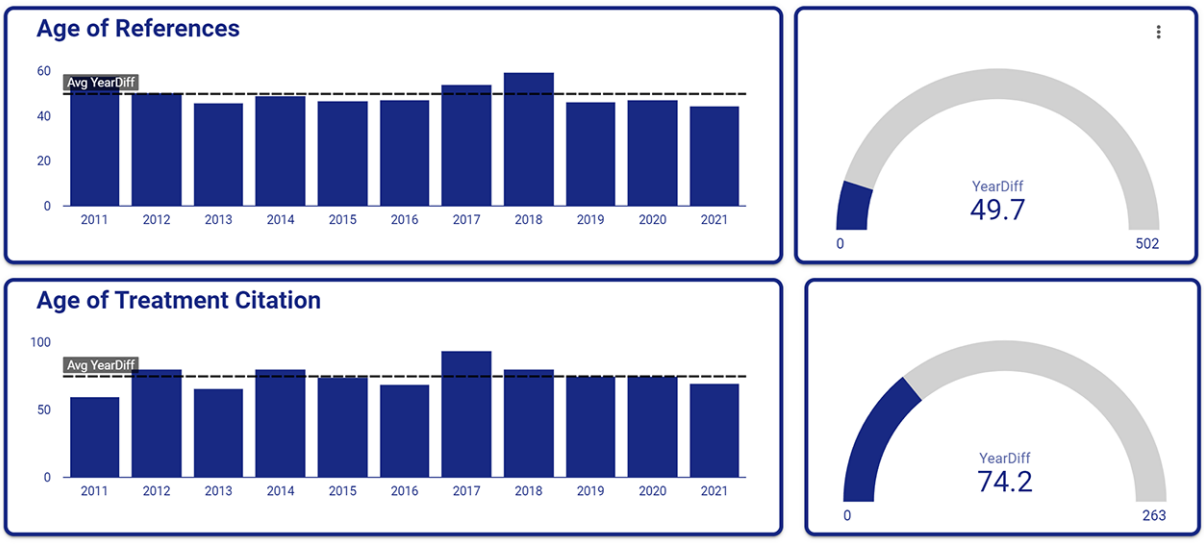

Fig. 4. Age of Reference: The average difference of the publication year of an article with those of the publications included in the bibliographic references. Age of Treatment Citations: The average difference between the year of publication of a taxonomic treatment and the cited treatments. New species generally do not include treatment citations. The biggest difference is when a treatment includes the citation of the original treatments by Linnaeus, 1753 for plants or 1758 for animals.

The relational database created in FileMaker enables us to calculate the averages. With the year of publication of the articles cited and the year of publication of the citing article (the article published in *EJT* that cites the reference), we can deduce the timespan between a publication and its citation, and more precisely calculate the average age of citation in the bibliography cited. On average in this dataset, the references cited are around 39 years old.

However, the Google Studio dashboard created with the dataset from TreatmentBank gives different figures, which is due to the fact that the data needs to be cleaned before the calculation which will be done in due course with additional resources available. Indeed, some references insert a year followed by a letter (1998a, 1998b) and can distort the calculation. Nonetheless, the figures in the dashboard reinforce the fact that the references cited are older citations (in this case, the average is 49 years old). The dashboard also allows calculation of the number of years between the treatment citation and the publication in which the treatment is cited: the average is 74 years!

These values have barely changed over 10 years. This underscores the fact that the calculation of the Clarivate impact factor or Scopus CiteScore has too short a horizon to be relevant to the field of taxonomy. Indeed, Clarivate calculates the impact factor of a journal by dividing the number of current year citations to the source items published in the said journal in the previous two or five years. CiteScore (Scopus calculation) is the number of citations received by a journal in one year to documents published in the three previous years, divided by the number of documents indexed in Scopus published in those same three years. In both cases, it does not reflect the age of references cited in taxonomic articles.

Number and geographical origins of authors

The *EJT* team also monitored the geographic locations of all authors of submitted papers for several years (Table 2), as well as those of published papers (Table 3) before it became too time-consuming to collate it manually, authors are regrouped in both tables by continent. The geographic origins of the authors are based on the institutional addresses indicated in their manuscript. Authors publishing in or submitting a paper to *EJT* come from a wide range of geographic locations. While the journal itself is published by a consortium of European institutions and describes species conserved in collections in Europe, submissions are open to authors from all over the world.

The data extracted in TreatmentBank show a similar pattern but with an incomplete analysis, as evidenced by fields with 'NULL' entries. For 2016 for instance, TreatmentBank recorded 182 authors (only first authors have been taken into account) of which 35% have a null entry. When Plazi and *EJT* together analysed the structure of the papers for retro-conversion it was clear that numerous inconsistencies in the author affiliations and in the name used for the institutions led to incomplete results and generated confusion. Both datasets show how global the authorship of *EJT*'s articles have become during this past decade.

In 2016, the number of African authors, particularly important to monitor for the Royal Museum for Central Africa, represented 2% of all authors in published papers, while 3% of authors submitting a paper that year were affiliated in an African institution (Table 4). This proportion compares to the 2% of *Zootaxa*'s authors from Africa (Zhang 2021).

An analysis of the first authorship in the articles published from 2016 to 2020 shows that from 2016 to 2020, 5% of the articles published in *EJT* were co-authored by at least one author affiliated with an African institution (Table 4). The data come from the internal databases manually maintained by the *EJT*'s team. The same analysis could be detailed country by country using the TreatmentBank statistics.

Table 2. Affiliation countries of authors of submitted papers. Figures manually counted for 2015 and 2016. In 2015, 164 papers were submitted, in 2016 285 papers were submitted.

Continents	2015	2016	2017
Africa	6	29	22
Asia	122	185	96
Europe	209	475	212
Middle East	22	35	32
North America	22	45	17
Oceania	3	21	4
South America	43	61	53
Total authors	427	851	436
Number submissions	164	285	157

Table 3. Geographic locations of all authors of published papers (not only first or corresponding). Those figures were manually counted because it was an important criteria to understand the authorship of the journal.

Continents	2016	2017	2018	2019
Africa	4	14	11	5
Asia	57	67	41	40
Europe	141	278	146	135
Middle East	10	15	4	12
North America	10	26	11	23
Oceania	3	11	6	15
South America	33	35	40	49
Total	258	446	259	279
Number of articles published	87	135	99	100

Table 4. Articles published within *EJT* in which at least one author is affiliated to an African institution.

Years	Total number of articles published	Articles authored with an author affiliated to an African institution	Articles with a 1st author affiliated to an African institution
2016	87	3	1
2017	135	9	4
2018	99	7	4
2019	100	4	3
2020	175	6	3
Total	596	29 (5%)	15 (3%)

FAIR data: number of figures deposited in Biodiversity Literature Repository and treatments in and material citations in GBIF

11 979 figures have been extracted of which 11 565 or 97% are open access and FAIR in Biodiversity Literature Repository. From the 14 404 treatments liberated and available in TreatmentBank, 11 984 or 83% are available in Biodiversity Literature Repository including a DOI and extensive custom metadata. The remainder is being added once quality control issues have been resolved. 32 005 material citations have been extracted, of which 27 518 or 85% are in GBIF. This includes 9001/9354 (96%) through 2018, and 18 517 out of 22 651 (82%) from 2019 forward. The relatively higher percentage of the upload through 2018 is due to higher granularity of the output data and higher quality control standards enforced for the corpus from 2019 forward.

FAIR data is citing the original source and thus has a multiplication effect on the access to the original article, whereby an article with no treatments, figures and material citation has three more access points from Biodiversity Literature Repository, TB and GBIF, an average article 166 and the largest catalogue with 1202 treatments a maximum value of 7398 (Carneiro *et al.* 2014).

Timeline of production

This statistic is given thanks to the close monitoring of the *EJT* team, whose goal is to increase the rapidity of publication while improving the quality of production.

The publication timeline is traditionally divided into two sections: 1) the time dedicated to the peer-review process, revision process and the editorial decision; and 2) the time of production per se: editing, layout, structuring the paper for better data extraction, proofreading and finally publishing.

From September 2011 to September 2021, the timelines have evolved as shown in Table 5. The timeline of publication obviously correlates with the number of submissions. This timeline lengthened considerably in 2015 and particularly in 2016 due to the huge increase of submissions (see below) that can be explained by the fact that *EJT* had its first impact factor published in June 2015 (2014 Impact factor: 1.312). As a result of the longer timeline for production, the number of submissions decreased in 2016 and 2018. Such fluctuation in the number of submissions is a common pattern in academic journals. Thanks to the team's reinforcement by the new participating institutions (the Museo Nacional de Ciencias Naturales and the Real Jardín Botanico in Madrid joined *EJT* in 2016, followed by Naturalis in 2017, the Zoological Research Museum Alexander Koening (ZFMK) in Bonn in 2018 and the National Museum (NMCZ) in

Table 5. Timeline of publication in calendar weeks between the 1) peer review and revision process, and 2) editorial process and production.

Average timeline (in calendar weeks)	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
between submission and acceptance (1)	6	19.5	16	16	12	14	19	18	18	18	22
between acceptance and publication (2)	3.5	5.5	7	8	9	20	30	21	10	11	12
between submission and publication	9.5	25	23	24	21	34	49	39	28	29	34

Praha in 2019), the backlog was quickly reduced, and production time has returned to that observed in the years before the impact factor. This has kept the journal within the usual publication deadlines of a good high-level edited journal.

Number of submissions and rejection rate

Table 6 shows the data provided by the editorial team monitoring these figures. The rejection rate (number of rejected papers in a year divided by the number of submissions the same year) also correlates somewhat with the team's production capacity. In order to maintain an acceptable production timeline within the production constraints of a technical team that is not easily expandable, the journal has no other choice than to adopt a stringent editorial policy and apply its editorial scope more strictly. This has the disadvantage of making publication in *EJT* more difficult for authors, but has the advantage of greatly improving the quality of the published papers, as manuscripts are often greatly expanded and ameliorated in the process between first submission and acceptance.

Table 6. Number of submitted papers vs rejected papers from 2011 up to 9th Sept. 2021

	2011	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**
Number of papers submitted	28	42	55	48	164	285	157	157	255	368	249
Number of rejected papers	14	12	10	8	34	129	84	66	89	165	119
Rejection rate	50%	29%	18%	17%	21%	45%	54%	42%	35%	45%	48%

For comparison purposes only, *Zookeys* indicates a rejection of 25% for 2016–2017 (Erwin 2018), and rejection rate for papers published in *Zootaxa* during 2011 to 2020 varies from 1.9% to 60% depending on the group studied (Zhang 2021), and a similar rate of rejection as *Zookeys* (26.8%) was mentioned in the journal's website for 2001–2003, based on a pool of papers on Arachnida, Coleoptera, Diptera, Mollusca, Nematoda and Pisces.

Taxonomic data

Number of treatments

The first ten years of *EJT* saw the publication of 906 articles including 15 132 taxonomic treatments, 31 897 published PDF pages, 1098 tables, 19 255 treatment citations and cited 39 941 bibliographic references (Fig. 5). On average, it represents 1375 treatments published per year and 16.7 treatments per article. With an average of 35 pages per *EJT* paper, this means that more than 3.1 pages are dedicated per treatment. It reflects the scope of the journal and the fact that *EJT* tends to maximise monographic studies and revisions and to limit single-species descriptions, except when they come with a rich scientific context such as phylogenetic and detailed anatomical results and discussion. The representation of the number of papers and treatments per year of existence of the journal however suggests that it took five years, i.e., from 2011 to 2016, to reach the current average number of published articles (82 per year) and treatments (1299 per year) (Fig. 5).

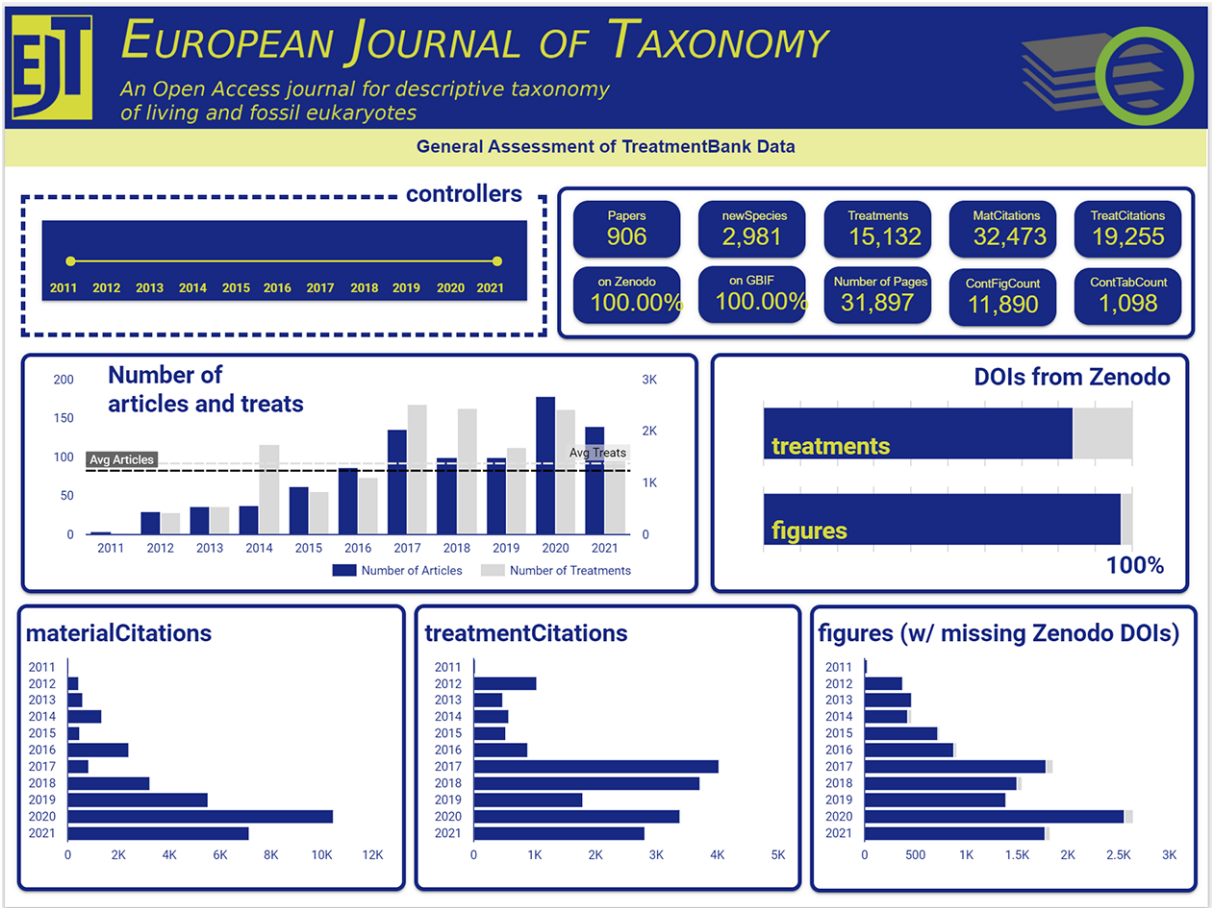

Fig. 5. Overview of *EJT* articles and major data liberated by Plazi and made FAIR on Biodiversity Literature Repository (source). The controllers are used to select a specific period.

Taxonomic overview

Analysing the taxonomic treatments gathered in the first 10 years of *EJT* and keeping in mind that it is a journal for descriptive taxonomy of eukaryotes (both living and fossil), there is an absolute predominance of animals (93.9%) over plants (5.7%). Fungi and Chromista are still poorly represented, whilst contributions on the other groups of Eukaryota have not yet been received.

Within animals, invertebrates dominate, headed by arthropods (75.6%), which with molluscs (7.7%) and annelids (2.2%) already account for 85.5%. Chordata are well-represented with 2.7%. Plant treatments almost exclusively concern vascular plants (Tracheophyta): 5.6 of 5.7% (Fig. 6).

It can be noted that the high number of invertebrates and arthropods can be easily explained by the percentage composition of these groups within the currently known biodiversity and by the huge number of taxa that still remain to be discovered according to estimates (Chapman 2009). The considerable percentage of contributions regarding chordates can be explained by the great attention traditionally given to this group of animals (Chapman 2009).

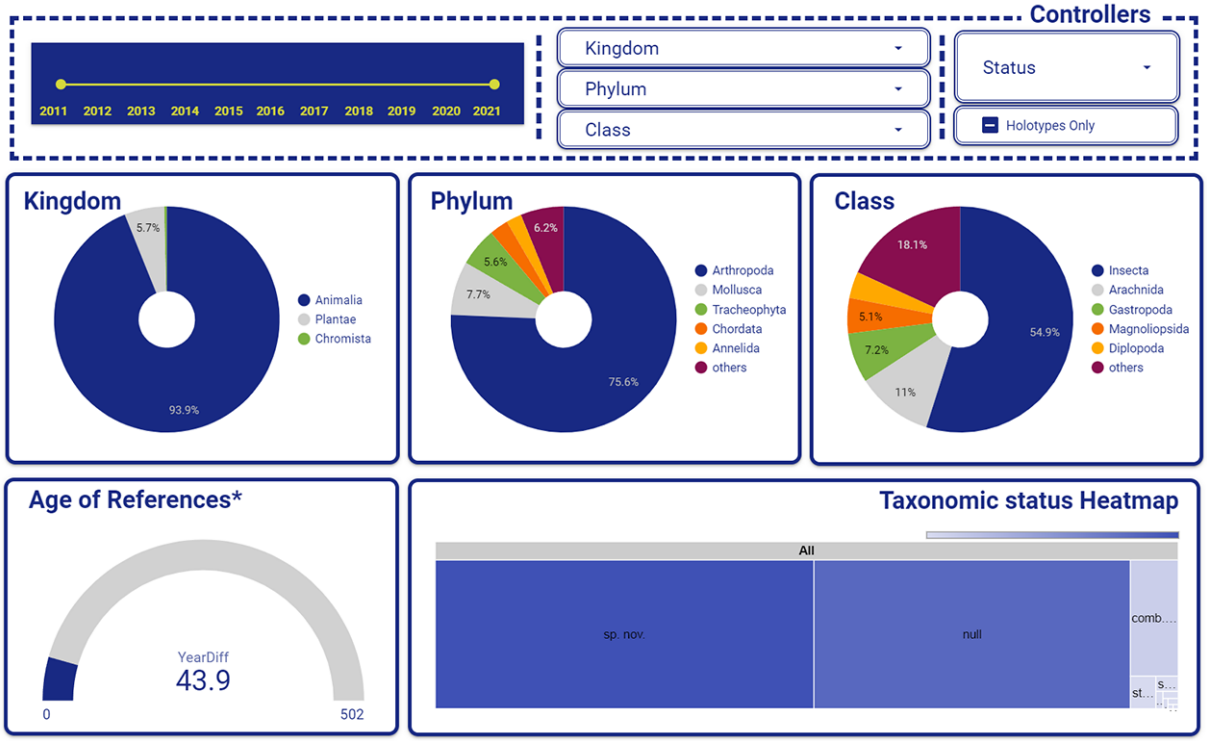

Fig. 6. Distribution of relative number of treatments by taxonomic rank. The average Age of Reference is the difference between the specimens cited in the material citations and the publications date.

Geographic assessment

Access to data in publications allows us to compare the geographic origin of authors through the affiliations, and the specimens and deposition of the specimens via the material citations.

Although the automated retrieval of data does not allow in each case its assignment to its geographic origin, and necessarily involves some simplifications (see Method section for further detail), interesting observations can be made by analysing the data obtained from Plazi (Fig. 7). Concerning the collecting countries (top ten positions), it is possible to point out that the most represented are African countries (e.g., D.R. Congo with more than 10 000 records; South Africa with more than 5000; Cameroon with more than 3500; and Ivory Coast with ca 2000). The Americas are also well-represented (e.g., Brazil with nearly 5000 records; USA with more than 2600; and Mexico with ca 2500), followed by some countries from Asia and from the Australasian region with, for example, the ca 2600 records from Indonesia and just under 2000 from Australia. Finally, China has 1850 records. A correlation with the highly diverse countries, especially those included in the tropical belt, appears immediately evident. However, the numbers also certainly reflect the composition of the collections hosted in the large natural history museums and their colonial past.

The countries where the specimens have been deposited is different with a predominance of USA institutions with more than 18 000 records and of those in European countries (i.e., UK and Germany each with more than 6000 records, France more than 4800, the Netherlands: ca 4000, and Austria with more than 2800). Relevant contributions then come from South Africa, Canada, Brazil, with more than 3000 records, and Australia with just over 2000 records. A correlation can be found with countries hosting

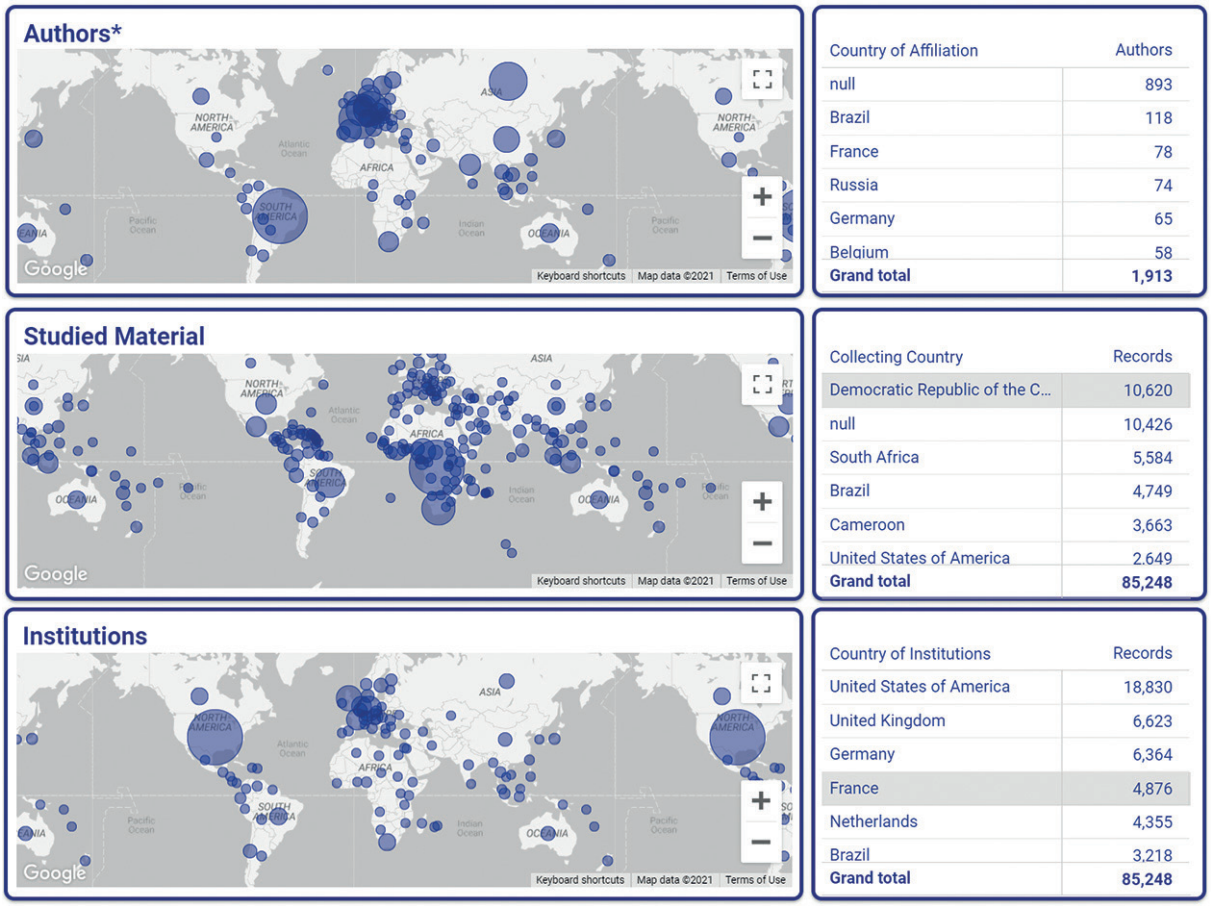

Fig. 7. Comparison of the location of the study author, the origin of the studied material and the institution or collection where the specimen referenced in the material citation is deposited.

important natural history museums and collections, but also supporting research initiatives occurring in some countries and promoted by the institutions.

Looking at holotypes only (Fig. 8), there is little difference between the location of the author and the deposition of the specimens, but the dominant place of the discovery of new species is now Indonesia.

New taxa and other nomenclatural acts

At the current state of data liberation, TreatmentBank includes 88% of the 3971 new described taxa kept in the records of *EJT*. The 3501 new taxa accessible through TreatmentBank consist of 90% of new described species (more than 3600), followed by genera (a little more than 300). The other taxonomic ranks account for much lower figures, as to be expected (Table 7). The number of new combinations (*combinatio nova*) – more than 500 – and the names proposed at new rank (*status novus*) – almost 60 – hosted on the articles published in the first ten years of *EJT* are also relevant (Table 8).

One of the most important goals of *EJT* continues to be the reduction of the taxonomic impediment through the publication of new taxa, because it adds to the understanding of the long tail of species about which very little to nothing is known yet, but which might disappear soon as part of the biodiversity

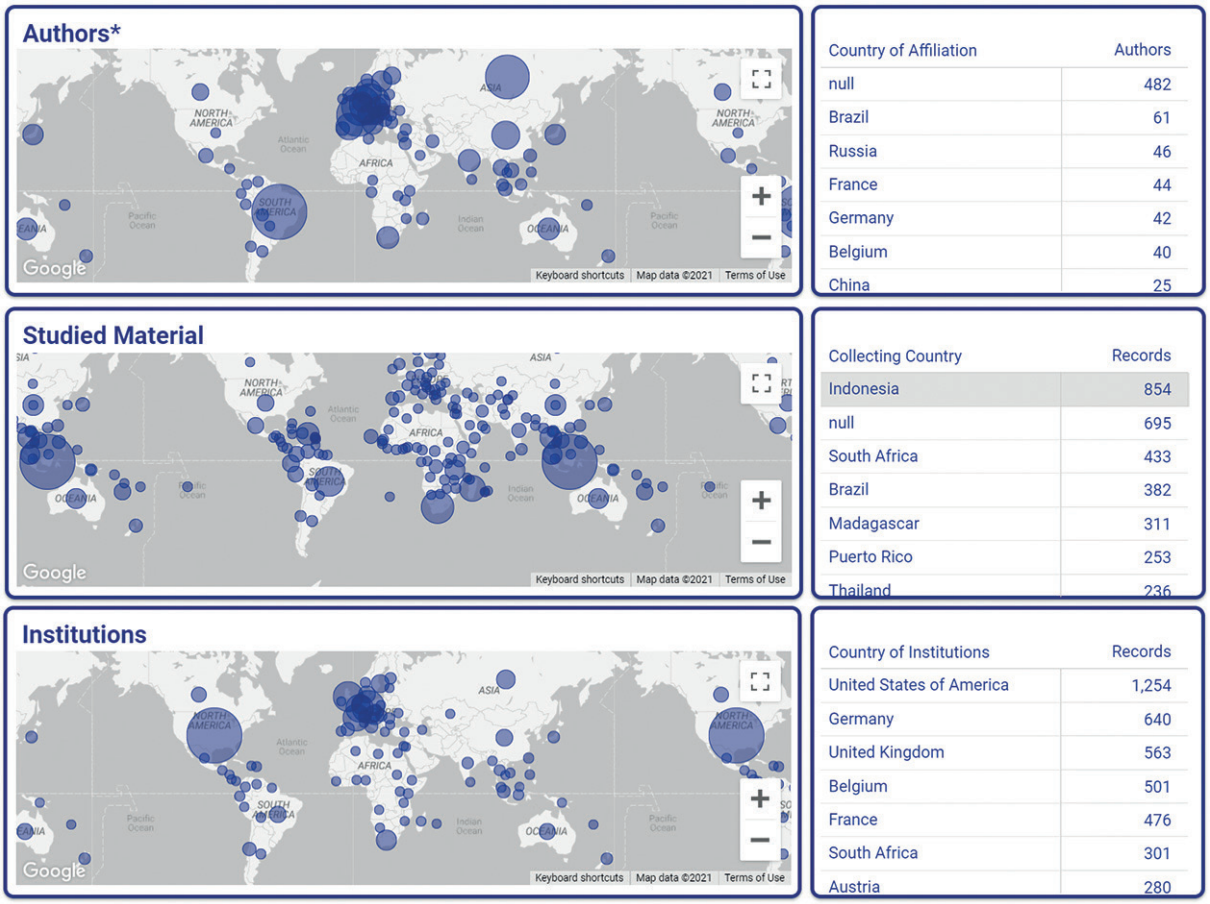

Fig. 8. Comparison of the location of the study author, the origin of the studied material and the institution or collection where the holotype specimen referenced in the material citation is deposited.

crisis. In 10 years of *EJT*, 3971 new taxa have been proposed (3501 according to TreatmentBank), including 3635 new species (3146 sp. nov. according to TreatmentBank), and numerous higher-rank taxa including 32 new families or subfamilies, and 304 new genera or sub-genera (268 gen. nov. according to TreatmentBank, 9 tribes, 6 subfamilies and 12 families). This represents more than 300 new taxa per year, all accessible via the publication, TreatmentBank, Biodiversity Literature Repository and GBIF.

Table 7 compares the figures obtained through the manual monitoring conducted by the *EJT* team in an internal database and the figures obtained by Plazi.

As with the issue of authorships described above, the results again diverge between the two types of analyses. It is pertinent to state here that at the time of writing, Plazi has already achieved 88% of the 'internal' results, for the same reason as with the origin of authors. For legacy publications, Plazi used retro-conversions of PDFs, resulting in variable granularity as older papers did not systematically include the precise criteria needed for full data processing.

The 19255 treatment citations – the citations of previous treatment in subsequent taxonomic treatments – include the taxonomic history essential to build the catalogue of life. The relationship between the nominate taxonomic name in a treatment and the names in the treatment citations can be typed when the author creates a new combination or synonymises a taxon, or is used to augment an existing treatment with new data.

Table 7. New taxa described per taxonomic rank, TreatmentBank records compared to the FileMaker database. The total shows that 88% of the new taxa described by *EJT* have been tagged as such in the retro-conversion process.

	% of total	TreatmentBank	FileMaker database
Family	0.3	12	32
Subfamily	0.15	6	NA
Tribe	0.26	9	NA
Genus	7.7	268	304
Subgenus	1	37	NA
Species	90	3146	3635
Subspecies	0.6	22	NA
Variety	0	1	NA
Total	88%	3501	3971

Table 8. Other nomenclatural acts.

New combination (combinatio nova)	511
Name at new rank (*status novus*)	58
New replacement name (*nomen novum*)	16
Invalid name (*nomen nudum)*	1
Total	563

Conclusion

The digital age provides a bright future for the dissemination of and access to research results for everybody, anywhere at any time, to integrate them in the research data life cycle and allows reproduction of the results. Hyperlinks to digital copies of the cited items save a huge amount of time. For natural history institutions, this offers the opportunity to provide access to the cited objects – the specimens – from research results based on their collections. When an article is available in different formats and its data is provided as open liberated FAIR data and reused in GBIF, this substantially multiplies the access to the article, in the case of *EJT* by a factor of 166 with a range from 3 to 7400, and thus highly increases the dissemination of the content. Furthermore, taxonomic publications play an integral role in providing links between specimens, their treatments, figures, gene sequences and various digital representations. They also play an important part in building the catalogue of life based on the treatment citations which could be automatically used to augment the catalogues with the new results, not just new species. To leverage this, linkage is already technically possible, or is in planning and will be implemented within the next couple of years. The closer collaboration and tighter integration of publishers with research infrastructures such as the CoL+, GBIF, European Nucleotide Archive (ENA), Biodiversity Literature Repository, and TreatmentBank is supported by the EU Horizon 2020 BiCIKL program and the Arcadia Fund.

The actual ongoing use of data from within publications is best exemplified by the 490 citations of data sets in GBIF (GBIF 2021) including data liberated from publications or OpenBioDiv (Dimitrova *et al.* 2021). The first of these publications is now processed in TreatmentBank and thus closes the cycle of the research data life (Fig. 1).

Publishing new articles that include all the links and identifiers to the cited objects is dependent on switching to formats such as XML. This makes it possible to embed not only the links to taxonomic names, specimens or gene sequences to either a reference catalogue or database, but also makes parts of them citable, for example each figure or treatment or individual material citation by adding a persistent identifier. Links to external vocabularies or reference databases such as the Biodiversity Information Standards (TDWG) Darwin Core allow machines to process the text and reuse respective elements. Well-known taxonomic publishers are well on the way in this direction, or are following suit like *EJT* and the MNHN Paris building an XML-first open source publishing workflow.

Even today, the understanding of knowledge included in scientific journals in biodiversity is based on human data analysis, creating databases for specific user questions, such as World Register of Marine Species (WoRMS), Catalogue of Life, World Spider Catalogue or an entire journal (e.g., Zootaxa: Garnett *et al.* 2020; Zhang 2021). This requires an extensive workforce; it cannot be easily replicated because of the huge effort needed to collect publications, many of which are not widely accessible due to closed access publishing.

From a different point of view, any person who contributes to build taxonomic catalogues has to find the article and search for the pertinent data, essentially creating a mental annotation of the text that is being copied, pasted and often interpreted before entering into the database. This manual effort to prepare cataloguing is transient and has to be repeated by successive scientists interested in the same data, because the source is cited but the access to the physical or digital copy is not provided. Even then, at best the PDF is provided, but more often only a DOI is provided because of the predominance of closed access publications. This means that everybody has to start again to obtain a PDF – one of the most time-consuming aspects of taxonomic research. Adding an extra step to store the annotations thus seems to make sense from this point of view.

The expectations in the digital age are that such analyses can be performed without handling each article manually. However, with the proper editing tools in place to populate new databases this information can be kept up to date as new data is added at the moment it is processed.

Fully-automated data extraction from unstructured text will, without a big effort, hardly ever be possible, explaining the difference between the numbers of new species accounted for manually by the *EJT* team and what Plazi produced. Looking at current practices of predominantly publishing in PDF, it is clear that this will accompany academic publishing for a long time. This can be mitigated partially by adopting publishing guidelines that aim at facilitating data extraction (Chester *et al.* 2019). Better though is to avoid producing more unstructured text and to make use of the opportunity offered by the digital age to make a substantial contribution to better understanding biodiversity and adding to a more comprehensive conservation of the biodiversity of planet Earth, such as demonstrated by Pensoft (Penev *et al.* 2010).

The important lesson from 900 *EJT* articles is that the costs to retro-digitize are almost insurmountable, making retro-conversion feasible only for a select corpus of relevant publications for research, agriculture, human health or conservation reasons. However, to ensure total coverage and FAIR-ization of the data contained in the articles, and to provide bidirectional links within the publication itself, *EJT* has to be published in semantically enhanced form in the future so that data in publications are immediately open and FAIR. This is best proven by its immediate reuse in GBIF. Indeed, another disadvantage of the retro-

conversion model is that the annotations performed and the FAIR data created by Plazi are not found nor linked to the original publication of the journal which impedes yielding the maximum impact of the data liberation process. A native XML approach would eliminate the discrepancies in the data, such as exemplified in Table 7.

The *EJT* team is currently working on a process that results in the direct production of native XML data rather than first going through a PDF version. Scientific journals that invest in new technologies – such as a workflow that produces directly articles in tagged XML (called XML-first based workflow) – will realize genuine added value as well as the certainty of full extraction of their data by Plazi. Journals that do not have the means to take this path will have to continue converting to XML after publication and must ensure that their handling is as rigorous and as consistent as possible to approach full extraction of their taxonomic data.

This work on legacy publications has highlighted a need for an infrastructure for institutional names, collections, and authors. Collection codes are inconsistent across treatments and it is difficult to check within the TreatmentBank process if the system recognizes the code as pointing to a particular collection, or an institution, and even when the full name of the collection appears in the source article, there does not seem to be a way to make use of this information using the GBIF Registry of Scientific Collections (https://www.gbif.org/grscicoll). The situation with the author country illustrates the urgency of this as well. The situation would be much more accurate if it would be possible to point to a collection/institution in a curated list. The use of ORCID for authors eases the situation to add and use a persistent identifier and thus contribute to ensuring the consistency of names. The journal team is thus involved in European working groups on persistent identifiers needed in taxonomy (Penev *et al.* 2021). This accuracy will be highly relevant for the institutions themselves, as well as for the stakeholders, because it will allow generic alternative metrics to measure the collections and its scientists output.

The next decade will be dedicated to new implementations. The new managing staff works as an Executive Committee and will prepare *EJT*'s future. A major achievement will be the possibility to publish in XML format directly. The new workflow, called XML-first, will provide greater ability to exchange data with databases and international platforms and can generate an XML-JATS version of the full article. *EJT* is involved in several projects that will enable the journal not only to produce an XML-First based workflow adapted to taxonomy but also to make it open and available to any independent taxonomic journal that wishes to produce its articles in XML.

Acknowledgements

The dedication of the team working for the journal is amazing and the authors would like to use this opportunity to thank all the people involved in the journal since its beginning, particularly:

Prof. Koen Martens, who first proposed this collaborative adventure and also served as its Editor-in-chief for the past ten years. The *EJT* team is profoundly grateful for his dedication to the journal.

The people who were involved in the description of the *EJT* business plan along with co-authors IG and LB: Daphne Duin, Steven Dessein, Koen Martens and Graham Higley.

The authors and *EJT* are grateful for the dedication of all the people that make it a successful journal: current Editor-in-chief FC; current and former Topical editors Nesrine Akkari, Rudy Jocqué, Frederik Leliaert, Thomas Jansen, Christian de Muizon, Gavin Broad, and TR; Section editors Thierry Backeljau, Helen Barber-James, Max Barclay, Gavin Broad, Jurate De Prins, Christopher Dietrich, Torbjørn Ekrem, Michel Louette, Felipe Polivanov Ottoni, Martin Vinther Sørensen, Daniel Stec, Didier VandenSpiegel,

and Peter Vd'ačný; Publication managers and Liaison officers IG, LB, and Niels Raes; desk editors Connie Baak, Natacha Beau, Chloe Chester, Danny Eibye-Jacobsen, Pepe Fernández, Kristiaan Hoedemakers, Eva-Maria Levermann, Alejandro Quintanar Sánchez, Radka Rosenbaumová, Marianne Salaün, Fariza Sissi, Charlotte Thionois and Jeroen Venderickx.

All projects carried out by the *European Journal of Taxonomy* are made possible by funding granted by the consortium of member institutions: Muséum national d'histoire naturelle, Paris, France; Meise Botanic Garden, Belgium; Royal Museum for Central Africa, Tervuren, Belgium; Royal Belgian Institute of Natural Sciences, Brussels, Belgium; Natural History Museum of Denmark, Copenhagen, Denmark; Naturalis Biodiversity Center, Leiden, the Netherlands; Museo Nacional de Ciencias Naturales-CSIC, Madrid, Spain; Real Jardín Botánico de Madrid CSIC, Spain; Zoological Research Museum Alexander Koenig, Bonn, Germany; National Museum, Prague, Czech Republic. Their support has made the journal one of the most successful European projects possible this past decade and we are profoundly grateful for it.

EJT would also like to thank CETAF for their continued support and endorsement of the journal.

Thank you to Emily Divinagracia for kindly editing the English. The authors are grateful to the three referees Susan Skomal, Quentin Groom and Jeremy Miller, whose comments and corrections greatly improved the manuscript.

References

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

Amsterdam Call for Action on Open Sciences. 2016. Available from https://www.government.nl/documents/reports/2016/04/04/amsterdam-call-for-action-on-open-science

Bénichou L., Dessein S., Duin D., Gérard I., Higley G. & Martens K. 2010. *EJT: European Journal of Taxonomy Business Plan*, report for the European Distributed Institute for Taxonomy, 22 p.

Bénichou L., Martens K., Higley G., Gérard I., Dessein S., Duin D. & Costello M.J. 2013. *European Journal of Taxonomy*: A public collaborative project in open access scholarly communication. *Scholarly and Research Communication* 4 (1): 010134. http://src-online.ca/index.php/src/article/view/37/114

BHL & Plazi 2021. Biodiversity Heritage Library and Plazi: Statement of Collaboration. https://doi.org/10.5281/zenodo.4958378

Budapest Open Access Initiative. 2002. The original Declaration and guidelines to make research free and available to anyone with internet access and promote advances in the sciences, medicine, and health. http://www.budapestopenaccessinitiative.org/read

Berlin Declaration 2013. Open Access to Knowledge in the Sciences and Humanities. https://openaccess.mpg.de/Berlin-Declaration

Bouchout Declaration for Open Biodiversity Knowledge Management. 2014. A contribution from the biodiversity community to Open Digital Science. Available from https://digital-strategy.ec.europa.eu/en/news/bouchout-declaration-contribution-biodiversity-community-open-digital-science [accessed 16 December 2021]

Carneiro M., Martins R., Landi M. & Costa F.O. 2014. Updated checklist of marine fishes (Chordata: Craniata) from Portugal and the proposed extension of the Portuguese continental shelf. *European Journal of Taxonomy* 73: 1–73. https://doi.org/10.5852/ejt.2014.73

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.* https://doi.org/10.5281/zenodo.3484285

Chapman A. 2009. *Numbers of Living Species in Australia and the World.* 2nd ed., Australian Biodiversity Information Services, Toowoomba, Australia, 84 p. https://www.environment.gov.au/science/abrs/publications/other/numbers-living-species/contents

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. https://doi.org/10.5852/ejt.2019.586

Côtez E., Mabille A., Chester C., Rocklin E., Deroin T., Desutter-Grandcolas L., Lesur J., Merle D., Robillard T. & Bénichou L. 2018. 1802–2018: a 220-year history of the Muséum periodicals. *Zoosystema* 40 (1): 1–41. https://doi.org/10.5252/zoosystema2018v40a1

DORA 2013. San Francisco Declaration on Research Assessment. https://sfdora.org/read/

EJT Analyses Data. 2021. Published in Biodiversity Literature Repository. https://doi.org/10.5281/zenodo.5703412

Erwin T., Stoev P. & Penev L. 2018. ZooKeys anniversary: 10 years of leadership toward open-access publishing of zoological data and establishment at Pensoft of like-minded sister journals across the biodiversity spectrum. *ZooKeys* 770: 1–8. https://doi.org/10.3897/zookeys.770.28105

Dimitrova M., Senderov V.E., 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

European Commission Recommendation on Access to and Preservation of Scientific Information. 2018. https://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservationscientific-information_en.pdf

Fawcett S., Agosti D., Cole S.R. & Wright D.F. (in press). Digital accessible knowledge: Mobilizing legacy data and the future of taxonomic publishing. *Bulletin of the Society of Systematic Biologists*.

Garnett S.T., Christidis L., Conix S., Costello M.J., Zachos F.E., Bánki O.S., Bao A., Barik S.K., Buckeridge J.S., Hobern D., Lien A., Montgomery N., Nikolaeva S., Pyle R.L., Thomson S.A., van Dijk P.P., Whalen A., Zhang Z.-Q. & Thiele K.R. 2020. Principles for creating a single authoritative list of the world's species. *PLoS Biol* 18 (7): e3000736. https://doi.org/10.1371/journal.pbio.3000736

GBIF 2020. New data-clustering feature aims to improve data quality and reveal cross-dataset connections. https://www.gbif.org/news/4U1dz8LygQvqIywiRIRpAU/new-data-clustering-feature-aims-to-improve-data-quality-and-reveal-cross-dataset-connections

GBIF 2021. Citation of treatment article data sets data published by Plazi.org taxonomic treatments database. https://www.gbif.org/resource/search?contentType=literature&publishingOrganizationKey=7ce8aef0-9e92-11dc-8738-b8a03c50a862

Penev L., Roberts D., Smith V.S., Agosti D. & Erwin T. 2010. Taxonomy shifts up a gear: New publishing tools to accelerate biodiversity research. *ZooKeys* 50: 1–4. http://doi.org/10.3897/zookeys.50.543

Penev L., Koureas D., Groom Q., Lanfear J., Agosti D., Casino A., Miller J., Arvanitidis C., Cochrane G., Barov B., Hobern D., Banki O., Addink W., Kõljalg U., Ruch P., Copas K., Mergen P., Güntsch A., Bénichou L. & Benito Gonzalez Lopez J. 2021. Towards Interlinked FAIR Biodiversity Knowledge: The BiCIKL perspective. *Biodiversity Information Science and Standards* 5: e74233. https://doi.org/10.3897/biss.5.74233

Plazi 2021. A new GBIF Plazi data issue feedback loop. Available from http://plazi.org/posts/data-issue-feedback-loop/ [accessed 16 December 2021]

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://doi.org/10.5281/zenodo.55665

Simoes F., Agosti D. & Guidoti M. 2021. Delivering Fit-for-Use Data: Quality control. *Biodiversity Information Science and Standards* 5: e75432. https://doi.org/10.3897/biss.5.75432

TDWG 2021. MaterialCitation. Available from https://dwc.tdwg.org/terms/#materialcitation [accessed 16 December 2021]

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

Zhang Z.-Q. 2021. Contributions of Zootaxa to biodiversity discovery: an overview of the first twenty years. *Zootaxa* 4979 (1): 6–16. https://doi.org/10.11646/zootaxa.4979.1.3

Manuscript received: 6 October 2021
Manuscript accepted: 6 December 2021
Published on: 17 December 2021
Topic editor: Frederik Leliaert
Desk editor: Marianne Salaün

Printed versions of all papers are also deposited in the libraries of the institutes that are members of the *EJT* consortium: Muséum national d'histoire naturelle, Paris, France; Meise Botanic Garden, Belgium; Royal Museum for Central Africa, Tervuren, Belgium; Royal Belgian Institute of Natural Sciences, Brussels, Belgium; Natural History Museum of Denmark, Copenhagen, Denmark; Naturalis Biodiversity Center, Leiden, the Netherlands; Museo Nacional de Ciencias Naturales-CSIC, Madrid, Spain; Real Jardín Botánico de Madrid CSIC, Spain; Zoological Research Museum Alexander Koenig, Bonn, Germany; National Museum, Prague, Czech Republic.