frontiers | Frontiers in Marine Science

TECHNOLOGY AND CODE
published: 22 July 2022
doi: 10.3389/fmars.2022.940844

Automating the Curation Process of Historical Literature on Marine Biodiversity Using Text Mining: The DECO Workflow

Savvas Paragkamian[1,2†], Georgia Sarafidou[2†], Dimitra Mavraki[2], Christina Pavloudi[2,3], Joana Beja[4], Menashè Eliezer[5], Marina Lipizer[5], Laura Boicenco[6], Leen Vandepitte[4], Ruben Perez-Perez[4], Haris Zafeiropoulos[1,2], Christos Arvanitidis[2,7], Evangelos Pafilis[2] and Vasilis Gerovasileiou[2,8]*

[1]Department of Biology, University of Crete, Heraklion, Greece, [2]Hellenic Centre for Marine Research (HCMR), Institute of Marine Biology, Biotechnology and Aquaculture (IMBBC), Heraklion, Greece, [3]Department of Biological Sciences, The George Washington University, Washington, DC, United States, [4]Flanders Marine Institute (VLIZ), Oostende, Belgium, [5]National Institute of Oceanography and Applied Geophysics (OGS), Trieste, Italy, [6]National Institute for Marine Research and Development "Grigore Antipa" (NIMRD), Constanta, Romania, [7]LifeWatch ERIC, Seville, Spain, [8]Department of Environment, Faculty of Environment, Ionian University, Zakynthos, Greece

OPEN ACCESS

Edited by:
Anne Chenuil,
Centre National de la Recherche Scientifique (CNRS), France

Reviewed by:
Halina Falfushynska,
Ternopil Volodymyr Hnatyuk National Pedagogical University, Ukraine
Javier Lloret,
Marine Biological Laboratory (MBL), United States

***Correspondence:**
Savvas Paragkamian
s.paragkamian@hcmr.gr

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Marine Ecosystem Ecology, a section of the journal Frontiers in Marine Science

Received: 10 May 2022
Accepted: 16 June 2022
Published: 22 July 2022

Citation:
Paragkamian S, Sarafidou G, Mavraki D, Pavloudi C, Beja J, Eliezer M, Lipizer M, Boicenco L, Vandepitte L, Perez-Perez R, Zafeiropoulos H, Arvanitidis C, Pafilis E and Gerovasileiou V (2022) Automating the Curation Process of Historical Literature on Marine Biodiversity Using Text Mining: The DECO Workflow. Front. Mar. Sci. 9:940844. doi: 10.3389/fmars.2022.940844

Historical biodiversity documents comprise an important link to the long-term data life cycle and provide useful insights on several aspects of biodiversity research and management. However, because of their historical context, they present specific challenges, primarily time- and effort-consuming in data curation. The data rescue process requires a multidisciplinary effort involving four tasks: (a) Document digitisation (b) Transcription, which involves text recognition and correction, and (c) Information Extraction, which is performed using text mining tools and involves the entity identification, their normalisation and their co-mentions in text. Finally, the extracted data go through (d) Publication to a data repository in a standardised format. Each of these tasks requires a dedicated multistep methodology with standards and procedures. During the past 8 years, Information Extraction (IE) tools have undergone remarkable advances, which created a landscape of various tools with distinct capabilities specific to biodiversity data. These tools recognise entities in text such as taxon names, localities, phenotypic traits and thus automate, accelerate and facilitate the curation process. Furthermore, they assist the normalisation and mapping of entities to specific identifiers. This work focuses on the IE step (c) from the marine historical biodiversity data perspective. It orchestrates IE tools and provides the curators with a unified view of the methodology; as a result the documentation of the strengths, limitations and dependencies of several tools was drafted. Additionally, the classification of tools into Graphical User Interface (web and standalone) applications and Command Line Interface ones enables the data curators to select the most suitable tool for their needs, according to their specific features. In addition, the high volume of already digitised marine documents that await curation is amassed and a demonstration of the methodology, with a new scalable, extendable and containerised tool, "DECO" (bioDivErsity data Curation programming wOrkflow) is presented. DECO's usage will provide a solid basis for future curation initiatives and an augmented degree of reliability

towards high value data products that allow for the connection between the past and the present, in marine biodiversity research.

Keywords: marine historical ecology, marine biodiversity data rescue, data archaeology, data curation, text mining, information extraction, scientific workflow, software containers

INTRODUCTION

Species' occurrence patterns across spatial and temporal scales are the cornerstone of ecological research (Levin, 1992). The compilation of both past and present marine data to a unified census is crucial to predict the future of ocean life (Ausubel, 1999; Anderson, 2006; Lo Brutto, 2021). This compilation has been attempted by big collaborative projects, like Census of Marine Life[1] (Vermeulen et al., 2013), that follow metadata standards and guidelines (Michener et al., 1997; Wilkinson et al., 2016) and modern web technologies (Michener, 2015). The project has resulted in the incorporation of census data from the past, i.e. historical data, to modern data platforms, such as the Ocean Biodiversity Information System (OBIS) (Klein et al., 2019), which feeds the Global Biodiversity Information Facility (GBIF) (GBIF, 2022). The transformation of historical data to modern standards is necessary for their rescue (data archaeology) from decay and inevitable loss (Bowker, 2000).

Historical data are usually found in the form of (a) historical literature and (b) specimens stored in biodiversity museum collections (Rainbow, 2009) (the digital transformation process and progress of specimens is reviewed by Nelson and Ellis, 2019). Historical biodiversity documents (also known as legacy, ancient or simply old documents) comprise literature from 1000 AD until 1960 and therefore are stored in an analogue and/or obsolete format (Lotze and Worm, 2009; Beja et al., 2022). These old documents can be found in institutional libraries, publications, books, expedition logbooks, project reports, newspapers (Faulwetter et al., 2016; Mavraki et al., 2016; Kwok, 2017) or other types of legacy formats (e.g. stored in floppy disks, microfilms or CDs).

From the scientific point of view, historical biodiversity data are as relevant as modern data (Griffin, 2019; Beja et al., 2022). They are valuable for studies on biodiversity loss (Stuart-Smith et al., 2015; Goethem and Zanden, 2021), as forming baseline studies for the design of future samplings (Rivera-Quiroz et al., 2020) and for predictions of future trends (Mouquet et al., 2015). Furthermore, historical data offer the kind of evidence needed for conservation policy and marine resource management, allowing for past patterns and processes to be compared with current ones (Fortibuoni et al., 2010; McClenachan et al., 2012; Costello et al., 2013; Engelhard et al., 2016). Hundreds of historical marine data held in documents have already been uploaded to OBIS, yet a Herculean effort is required to curate the thousands of available documents of the Biodiversity Heritage Library (BHL) (Gwinn and Rinaldo, 2009) and other repositories.

Adequate and interoperable metadata are equally necessary and have to be curated alongside data (Heidorn, 2008; Mouquet et al., 2015). In this context, standards and guidelines have been recently formulated in policies as Findable, Accessible, Interoperable and Reusable (FAIR) (meta)data (Wilkinson et al., 2016; Reiser et al., 2018). Identifiers and semantics are used to accomplish the interoperability and reusability of biodiversity data as well as the monitoring of their use (Mouquet et al., 2015). Indispensable to the curation process of marine data have been the standards of the Biodiversity Information Standards[2], more specifically Darwin Core (Wieczorek et al., 2012) and vocabularies such as those included in the International Commission on Zoological Nomenclature[3], the World Register of Marine Species[4] (WoRMS) (WoRMS Editorial Board, 2022), the Environmental Ontology[5] (ENVO) (Buttigieg et al., 2016) and Marine Regions[6] (Claus et al., 2014). These standards and vocabularies and their adoption by biodiversity initiatives like GBIF and OBIS align with the goal of marine biodiversity Linked Open Data and support their interoperability and reusability (Page, 2016; Penev et al., 2019; Zárate and Buckle, 2021).

The rescue process of historical biodiversity documents can be summarised in four tasks (**Figure 1**). The first task is the digitisation of the document, which involves locating and cataloguing the original data sources, imaging/scanning with specific equipment and standards and uploading them to digital libraries (Lin, 2006; Thompson and Richard, 2013). In the second task, the images are analysed with text recognition software, mainly through Optical Character Recognition (OCR) (for standards see Groom et al., 2019 and for reviews see Lyal, 2016 and Owen et al., 2020). Text recognition errors are then corrected manually by professionals or citizen scientists (Herrmann, 2020). The third task is named Information Extraction (IE) as it involves the steps of named entity recognition, mapping and normalisation of biodiversity information (Thessen et al., 2012). Here, the curators may compile a species' occurrence census enriched with metadata of the study, geolocation, environment, sampling methods and traits among others (Faulwetter et al., 2016). Lastly, the fourth task, is the data publishing to online biodiversity databases/repositories (Costello et al., 2013; Penev et al., 2017). Expert manual curation is a cross-cutting action through all the aforementioned tasks for quality control and stewardship (Vandepitte et al., 2015). This article focuses on the

[1] http://www.coml.org/

[2] https://www.tdwg.org/

[3] https://www.iczn.org/

[4] http://www.marinespecies.org/

[5] https://sites.google.com/site/environmentontology/home

[6] https://www.marineregions.org/

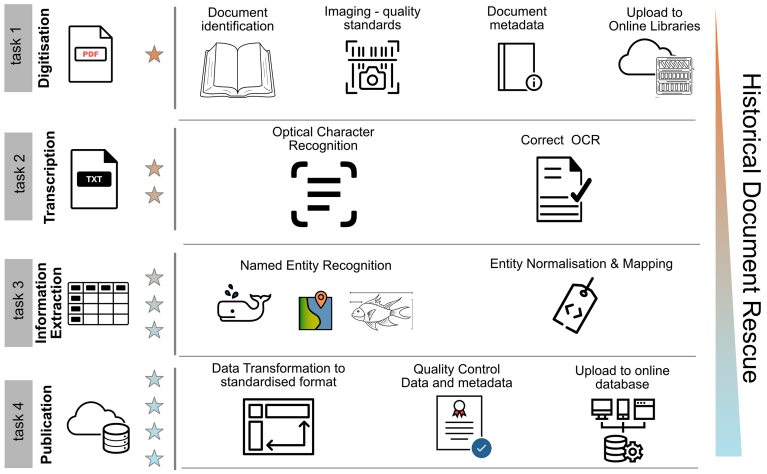


FIGURE 1 | Summarised process of historical document rescue. Four tasks are required to complete the data rescue process of biodiversity documents. Each of these has several steps, methodology, tools and standards. Curation is needed in every task, for tool handling and error correction. The stars represent the 5-star ranking system of Linked Data as introduced by W3C[7] (Heath and Bizer, 2011). Availability of information from historical data increases as the curation tasks are completed (as exemplified by the fan on the right). Icons used from the Noun Project released under CC BY: book by Oleksandr Panasovskyi, scanning by LAFS, Book info by Xinh Studio, Library by ibrandify, Scanner Text by Wolf Böse, Check form by allex, Whale by Alina Oleynik, Fish by Asmuh, tag code vigorn, pivot layout by paisan, Certificate by P Thanga Vignesh, web service by mynamepong.

tools and curation procedures encompassed in the third and fourth tasks described above.

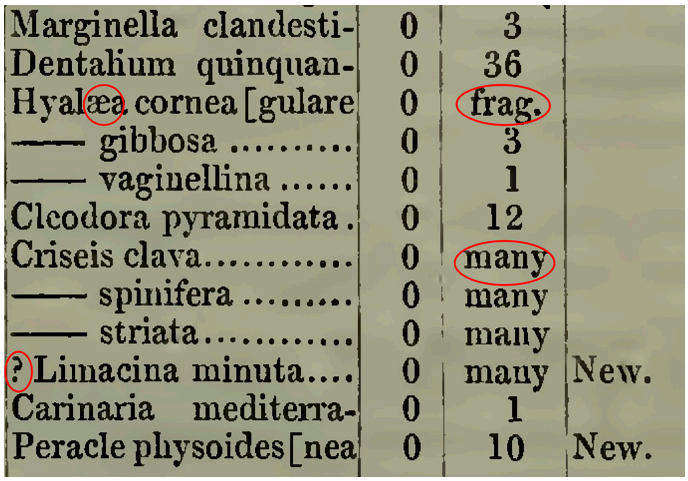
Several factors may turn the curation of historical documents into a serious challenge (Faulwetter et al., 2016; Beja et al., 2022). Errors from the first and second tasks, as presented in **Figure 1** (i.e. bad quality imaging, mis-recognised characters etc.) are propagated through the whole process. In terms of georeferencing constraints, location names or sampling points on an old map may be provided instead of the actual coordinates. Additionally, taxonomic constraints (e.g. old, currently unaccepted synonyms, lack of authority associated with the taxon names) combined with the absence of taxonomic literature or voucher specimens (e.g. identifier number for samples of natural history/expedition collections) require the taxonomists' assistance. Numerical measurement units often need to be converted to the International System of Units (SI system) (e.g. fathoms to metres) (Calder, 1982; Wieczorek et al., 2012). Old toponyms and political boundaries that have now changed should also be taken into consideration, as well as coordinates that now fall on land instead of in the sea, due to the changes in the coastline. Lastly, the use of languages other than English is quite common in old scientific publications, so multilingual curators are required. Some of the aforementioned issues are presented in **Figure 2**. Because of these limitations, the manual curation of data and metadata is mandatory when it comes to historical data (Faulwetter et al., 2016).

Manual curation, a tedious and multistep process, requires substantial effort for the correct interpretation of valuable historical information; however, text mining tools appear to be promising in assisting and accelerating this part of the curation process (Alex et al., 2008). Text mining is the automatic extraction of information from unstructured data (Hearst, 1999; Ananiadou and Mcnaught, 2005). These mining tools build upon standardised knowledge, vocabularies, dictionaries and perform multistep Natural Language Processes. Named Entity Recognition (NER) is a key step in this process for locating terms of interest in text (Perera et al., 2020). The entities of interest for biodiversity documents include: (1) taxon names, (2) people's names (Page, 2019a; Groom et al., 2020), (3) environments/habitats (Pafilis et al., 2015; Pafilis et al., 2017), (4) geolocations/localities (Alex et al., 2015; Stahlman and Sheffield, 2019), (5) phenotypic traits/morphological characteristics (Thessen et al., 2018), (6) physico-chemical variables, and (7) quantities, measurement units and/or values. Subsequent steps include the relation extraction between entities. Multiple tools have emerged to retrieve a single or a collection of these entities in the past few years (Batista-Navarro et al., 2017; Muñoz et al., 2019; Dimitrova et al., 2020; Le Guillarme and Thuiller, 2022).

The work described in this document has a threefold structure: (a) the abundance of marine historical literature digitised/available for curation is attempted to be estimated; (b) bioinformatics tools, focusing on automating and assisting the curation process for these documents, are compiled/reviewed. Two categories of such curation software are described: (i) the first one relies on web and standalone applications with Graphical User Interface (GUI) and the second (ii) combines Command Line Interface (CLI) programming libraries and software packages; lastly, (c) a demonstrator biodiversity data curation workflow, named DECO (bioDivErsity data Curation programming wOrkflow[8]), developed using programming tools, is presented.

[7]https://dvcs.w3.org/hg/gld/raw-file/default/glossary/index.html#x5-star-linked-open-data

[8] https://github.com/lab42open-team/deco



Marginella clandesti-	0	3	
Dentalium quinquan-	0	36	
Hyalæa cornea [gulare	0	frag.	
—— gibbosa	0	3	
—— vaginellina	0	1	
Cleodora pyramidata .	0	12	
Criseis clava.....	0	many	
—— spinifera	0	many	
—— striata.....	0	many	
?Limacina minuta....	0	many	New.
Carinaria mediterra-	0	1	
Peracle physoides [nea	0	10	New.

FIGURE 2 | Common problems encountered in historical data, such as old ligatures, absence of taxon names, ambiguous symbols, shortened words and descriptive information instead of numerical (page 185 from Forbes, 1844)

METHOD

Historical Literature Discovery

A search was conducted on BHL to amass the historical literature on BHL regarding marine biodiversity. Using the keywords "marine", "ocean", "fishery", "fisheries" and "sea" on the items' titles and their subjects (the scripts, results and documentation are available in this repository[9]) the documents available for information extraction were estimated. Subjects are categories provided for each title and multiple subjects can be assigned to each title. The items that were originally published before 1960 were selected, in order to include only historical documents, according to the definition included in the Introduction section. Furthermore, the taxon names on each page, which were identified by BHL using the Global Names parser tool (Mozzherin et al., 2022), were summarised for every document. Hence, summaries of the number of automatically identified taxon names were calculated along with the page number for each item. Additionally, OBIS' historical datasets originally published before 1960 were downloaded and analysed. This analysis provides an approximation of the size of available marine historical literature compared to the already rescued documents. All analysis scripts were written in GNU AWK programming language and the visualisation scripts were written in R using the ggplot2 library (Wickham, 2016).

Historical Document Rescue Methodology

Data curators thoroughly read each page of a document and insert the data into spreadsheets, mapping them to Darwin Core terms, adding metadata and creating a standard Darwin Core Archive[10]. This whole process, which is mostly manual, means reading the information (e.g. the occurrence of a specific taxon and its locality) and inputting it through typing to the corresponding cell of the data file. It is, as expected, a time- and resource-consuming procedure. Taxon names, traits, environments and localities can be identified as well and the transformation of these results to database identifiers (IDs), like Life Science Identifier[11] (LSID) of Aphia IDs[12], Encyclopedia of Life[13] (EOL) IDs (Parr et al., 2014), Marine regions gazetteer IDs, marine species traits[14] among others, can be facilitated through web applications and programming software. The Natural Environment Research Council[15] Vocabulary Server, developed and hosted by the British Oceanographic Data Centre[16] was used for mapping facts and additional measurements included in documents.

Tools assist curators in this process for the NER, Entity Mapping, data structure manipulation and finally data upload steps. Curation tools can be categorised as GUI applications (computer programs and web applications) and CLI applications (interconnected programming tools, libraries and packages) (**Figure 3**). As an example, multiple page documents can be searched for taxon names in seconds, with technologies that find synonyms and fuzzy search for the OCR transformation misspelling. The interconnection and guidance of these steps still requires human interaction and correction.

GUI applications are standalone applications or web applications, the latter support document upload and, once they are processed in a server, the results are delivered back to the user (Lamurias and Couto, 2019). CLI tools include

[9] https://github.com/savvas-paragkamian/historical-marine-literature
[10] https://manual.obis.org
[11] http://www.lsid.info/
[12] https://www.marinespecies.org/aphia.php?p=webservice
[13] https://eol.org/
[14] https://www.marinespecies.org/traits/
[15] https://www.bodc.ac.uk/resources/products/web_services/vocab/
[16] https://www.bodc.ac.uk/

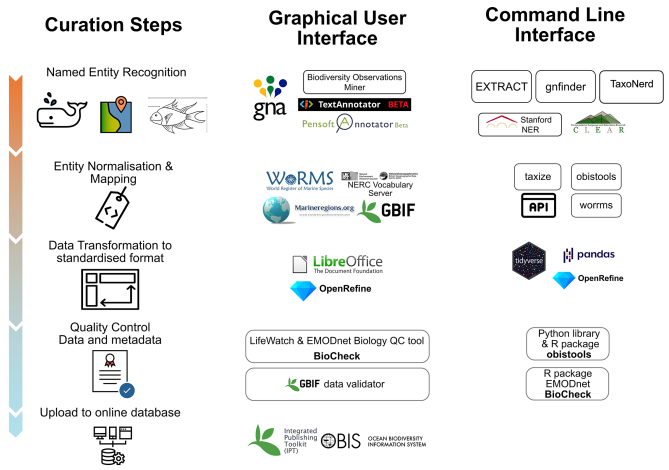
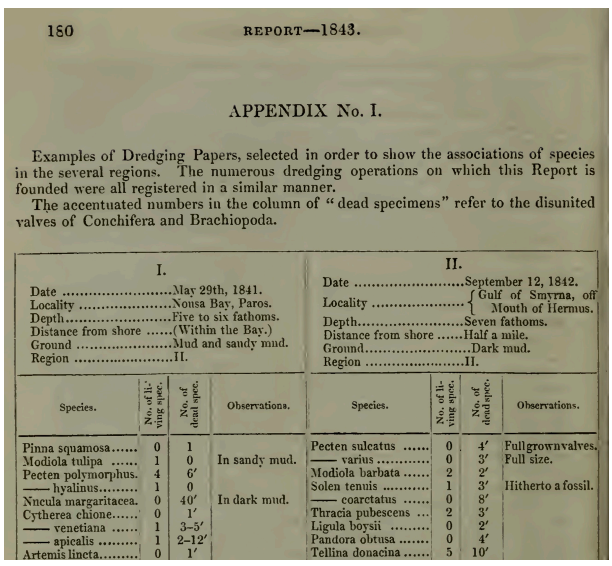


FIGURE 3 | The curation process of marine historical biodiversity documents: on the left column are the required steps starting from the scanned document (usually a PDF file) and ending with the data publishing step. Two approaches are presented: in the middle column are the GUI tools whereas on the right column are the CLI and/or the executable programming tools. Note that the list of given examples is non-exhaustive. Icons used from the Noun Project released under CC BY: Whale by Alina Oleynik, Fish by Asmuh, tag code by vigorn, pivot layout by paisan, Certificate by P Thanga Vignesh, web service by mynamepong

programming packages and libraries of any programming language in UNIX (Linux and Mac operating systems - OS) and Windows OS. Even though programming packages and libraries are fast and scalable they require familiarity and expertise in CLI and programming which, on the other hand, takes effort and time because of its initial learning curve. The CLI tools, Application Programming Interfaces (APIs) and programming packages chosen during this study are open-source, are in active development, can process many documents and can be combined with other tools in some of the considered steps.



180 REPORT—1843.

APPENDIX No. I.

Examples of Dredging Papers, selected in order to show the associations of species in the several regions. The numerous dredging operations on which this Report is founded were all registered in a similar manner.

The accentuated numbers in the column of "dead specimens" refer to the disunited valves of Conchifera and Brachiopoda.

I.

Date .....May 29th, 1841.
Locality .....Nonsa Bay, Paros.
Depth.....Five to six fathoms.
Distance from shore .....(Within the Bay.)
Ground .....Mud and sandy mud.
Region .....II.

Species.	No. of living spec.	No. of dead spec.	Observations.
Pinna squamosa.....	0	1	
Modiola tulipa	1	0	In sandy mud.
Pecten polymorphus.	4	6′	
—— hyalinus.....	1	0	
Nucula margaritacea.	0	40′	In dark mud.
Cytherea chione.....	0	1′	
—— venetiana	1	3–5′	
—— apicalis	1	2–12′	
Artemis lincta.....	0	1′	

II.

Date .....September 12, 1842.
Locality .....Gulf of Smyrna, off Mouth of Hermus.
Depth.....Seven fathoms.
Distance from shore .....Half a mile.
Ground.....Dark mud.
Region .....II.

Species.	No. of living spec.	No. of dead spec.	Observations.
Pecten sulcatus	0	4′	Full grown valves.
—— varius	0	3′	Full size.
Modiola barbata	2	2′	
Solen tenuis	1	3′	Hitherto a fossil.
—— coarctatus	0	8′	
Thracia pubescens ...	2	3′	
Ligula boysii	0	2′	
Pandora obtusa	0	4′	
Tellina donacina	5	10′	

FIGURE 4 | A screenshot of the dataset used where the structure of the data and metadata provided can be seen (page 180 from Forbes, 1844)

Case Study

The historical document "Report on the Mollusca and Radiata of the Aegean Sea: and on their Distribution, Considered as Bearing on Geology" by Forbes (1844) and its curated dataset were used as a case study for the tool usage description and evaluation (where applicable). More specifically, the six page long Appendix No. 1 (pages 180-185) document has been manually curated and published, thus serving as a golden standard (**Figure 4**). It was digitised and transcribed on 2009-04-22 by the Internet Archive[17] and on 2021-09-30 it was manually curated (Mavraki et al., 2021) and published in MedOBIS[18] (Arvanitidis et al., 2006). The rescue process resulted in a Darwin Core Archive file with 530 occurrence records, 17 unique sampling stations and 260 taxa, covering 217 species. The effort required from the information extraction task to data publishing was roughly 50 working days (8 hours per day) by a single data curator.

Tool Usability Evaluation

The web applications mentioned in this work were tested in November 2020 in two web browsers, Mozilla Firefox version 83 and Google Chrome version 87, both on Microsoft Windows 10 and MacOS 10.14.

Demonstrator

DECO was developed for the automation of biodiversity historical data curation. Its workflow combines image processing tools for scanned historical documents OCR with text mining technologies. It extracts biodiversity entities such as taxon names, environments as described in ENVO and tissue mentions. The extracted entities are further enriched with marine data

[17] https://archive.org/details/reportofbritisha43cor
[18] https://www.lifewatchgreece.eu/?q=content/medobis

identifiers from public APIs (e.g. WoRMS) and presented in a structured format as well as in report format with automated visualisation components. Furthermore, the workflow was implemented as a Docker container to ease its installation and its scalable application on large documents. DECO is under the GNU GPLv3 licence (for 3rd party components separate licences apply) and is available *via* the GitHub repository (https://github.com/lab42open-team/deco).

RESULTS

Historical Literature Discovery

Marine literature analysis on BHL holdings revealed that there are 1,627 different digital items that contain at least 100 distinct taxa to a maximum of 10,000 taxa, as identified automatically from the Global Names GNfinder tool. These items cover the period from 1558 to 1960, contain 648,927 pages, written in 10 different languages, 80% of which being English. An absolute estimation of historical marine data is difficult to be made as several more documents are stored locally in legacy formats.

The rescued historical marine data uploaded on OBIS are 223 datasets, published from 1753 to 1960. Hence, the manual curated literature is much lower than the available digitised documents. These rescued biogeographical datasets cover 46,000 species and 38 phyla that contain about 1.5 million occurrences at the species level.

Bioinformatics Tools Compilation and Review

This section describes the tools used in the curation workflow (**Figure 3**). In each step, the main up-to-date programming tools, web services and applications, used for the extraction of biodiversity data, are presented. These curation tools are listed, accompanied with features such as extracted information, input format and their interface in **Table 1**.

Named Entity Recognition

The Global Names Recognition and Discovery[19] (GNRD) tool, within Global Names Architecture[20] (GNA), is a web application used for the recognition of scientific names. It can use files such as PDF, images or Microsoft Office documents and one can still input URLs or even free-form text. It supports OCR transformation from PDF files using the tool Tesseract[21] and uses the GNfinder[22] discovery engine, in order to provide the list of names. It offers an API and can be installed locally. GNA is also used by the BHL platform to locate taxonomic names within the pages of its collections (Richard, 2020).

The test performed on the Forbes (1844) six-page PDF template provided 128 unique scientific names at species level, out of the 218 identified through the manual curation (**Figure S1**). WoRMS Aphia IDs (Vandepitte et al., 2015; Martín Míguez et al., 2019) are widely used and included in GNRD.

The Biodiversity Observation Miner[23] (BOM) is a web application based on R Shiny[24], also available on GitHub[25], that allows for the semi-automated discovery of biodiversity observations (e.g. biotic interactions, functional or behavioural traits and natural history descriptions) associated with the species scientific names (Muñoz et al., 2019). It uses the GNfinder discovery engine through the R package taxize[26] (Chamberlain and Szöcs, 2013). BOM is still under development (April 2022) and an OCR processed PDF file must be used as input. The novelty of this tool is the provision of text snippets (**Figure S2**) and the co-occurrence of words, accompanied with their count, to inform curators for terms that appear together in the document.

TextAnnotator[27], provided by the specialised information service BIOfid[28], is focused on information extraction about taxon names of vascular plants, birds, moths and butterflies, location and time mentioned in German texts (Driller et al., 2018; Driller et al., 2020). This could be extended to other environments, languages and taxonomic groups with the BIOfid Github page[29] serving as the starting point. The TextAnnotator - in beta version - accepts web pages or free text. Evidence of recent use of this tool was found in Driller et al. (2020).

The Pensoft Annotator[30] is another beta web application that works with ontologies (Dimitrova et al., 2020) (**Figure S3**). The Pensoft Annotator has Relation Ontology[31] (RO) and ENVO built in but it is extendable to any ontology with curation modifications for stopwords. The character limitation, however, can be expanded upon communication with the tool's administrators.

Taxonfinder[32] is a web application for the extraction of scientific names mentioned in web pages. It features an API that was used in BHL for large scale annotations of taxonomic names until 2019, when it was replaced by GNfinder (Richard, 2020).

The most notable NER tool, with CLI, for taxon names is the Global Names Finder (GNfinder) (Pyle, 2016; Mozzherin et al., 2022) which provides fuzzy search and is the underlying engine of most biodiversity text mining tools. It is in active development, deeming it a reliable tool for this work. The main command line tool is gnfinder find which returns two arrays (metadata and names). The metadata are the language, date of the execution of the command and total number of words. The data have one entry per identified string which contains the matched string, the returned name and the positional boundaries in character sequence.

[19] https://gnrd.globalnames.org/
[20] http://globalnames.org/
[21] https://github.com/tesseract-ocr/tesseract
[22] https://fgabriel1891.shinyapps.io/biodiversityobservationsminer/
[23] https://fgabriel1891.shinyapps.io/biodiversityobservationsminer/
[24] https://shiny.rstudio.com/
[25] https://github.com/fgabriel1891/BiodiversityObservationsMiner
[26] https://github.com/ropensci/taxize
[27] http://www.textannotator.texttechnologylab.org/
[28] https://biofid.de/en/
[29] https://github.com/FID-Biodiversity/BIOfid/tree/master/BIOfid-Dataset-NER
[30] https://annotator.pensoft.net/
[31] https://github.com/oborel/obo-relations
[32] http://taxonfinder.org/

TABLE 1 | Functions, interface and curation step of the tools tested in this work.

Tool	Curation Step	Input	Interface	Reference
Global Names Recognition and Discovery	NER - Taxon names	User query, Free text, PDF or image	WA, API, CLI	Pyle (2016)
BOM (Biodiversity Observations Miner)	OCR NER - Taxon names, Biotic interactions, Traits	User query, Free text, PDF	WA, API	Muñoz et al. (2019)
TextAnnotator	NER - Generic Annotations	User query, Free text	WA	Abrami et al. (2021)
Pensoft Annotator	NER - Annotation of free text with ontology terms Entity Mapping	User query, Free text	WA, API	Dimitrova et al. (2020)
Taxon Finder	NER - Taxon names	User query, Free text	WA, API	
EXTRACT	NER - Taxon names, Environments and Tissue	Free text	API, CLI	Pafilis et al. (2017)
TaxoNerd	NER - Taxon names	Free text, PDF, png	CLI	Le Guillarme and Thuiller (2022)
Stanford NER	NER - People, organisation, locality	Free text	CLI	Finkel et al. (2005)
Clear Earth	NER - Locality, unit, value, functional traits, taxon names	Free text	CLI	Thessen et al. (2018)
BioStor	Literature identification, NER - geolocation	Taxon names and other keywords	WA	Page (2011)
Marine Regions Gazetteer	Entity mapping	User input	WA, API	Claus et al. (2014)
Edinburgh geoparser	NER - geolocation Entity mapping	Free text	CLI	Alex et al. (2015)
Ontobee	Entity mapping	User input	WA	Xiang et al. (2011)
WoRMS taxon match	Entity mapping	Taxon list on comma separated/ spreadsheet file	WA, CLI, API	WoRMS Editorial Board (2022)
worrms R package	Entity mapping Data transformations	Taxon list: comma/tab separated file	CLI, API	Chamberlain (2020)
Taxize R package	Entity mapping Data transformations	Taxon list: comma/tab separated file	CLI, API	Chamberlain and Szöcs (2013)
GloBI nomer tool	Entity mapping Data transformations	Tab separated file	CLI	Poelen and Salim (2022)
OpenRefine	Data transformations, Quality control	Spreadsheet files, Comma/tab separated files, XML, RDF, JSON, SQL database	GUI app	Verborgh and De Wilde (2013)
LifeWatch Belgium & EMODnet Biology QC tool	Quality control	IPT or a DwC-A file	WA	
LifeWatch Belgium Data Services	Quality control, Entity mapping	Comma/tab separated file, spreadsheet excel file	WA	
EMODnetBiocheck	Quality Control	IPT, comma/tab separated file	CLI	De Pooter and Perez-Perez (2019)
GBIF Data Validator	Quality control	Comma separated, IPT or a DwC-A file	WA, CLI, API	
Obistools R package	Entity Mapping, Data transformations, Quality ontrol	Free text, comma/tab separated file	CLI	Provoost et al. (2019)
IPT server nodes	Quality control, Data Upload	Comma/tab separated file	GUI app	Robertson et al. (2014)
GoldenGate-Imagine	OCR, NER, Entity mapping	PDF	GUI app	Sautter et al. (2007)
DECO	OCR, NER, Entity mapping,	PDF, png, free text	CLI	This work

WA, Web Application; API, Application Programming Interface; CLI, Command Line Interface; GUI App, Graphical User Interface Application.

In order to simultaneously extract taxa, environment and tissue mentions, the tool EXTRACT[33] (Pafilis et al., 2017) implements the JensenLab tagger API (Jensen, 2016) with advanced dictionaries SPECIES-ORGANISMS[34] (Pafilis et al., 2013), ENVIRONMENTS[35] (Pafilis et al., 2015) and TISSUES[36] (Palasca et al., 2018). It returns NCBI Taxonomy IDs (Schoch et al., 2020), ENVO terms and BRENDA IDs[37], respectively to a file with 3 columns: tagged text, entity type and term ID. TaxoNERD (Le Guillarme and Thuiller, 2022), using Deep neural networks, scores higher than other NER tools on taxon name recognition based on golden standard corpora.

An important NER system is the Stanford NER[38] (Finkel et al., 2005) which recognises locations, persons and organisations in text. It has a generic scope but it can also assist in the curation of biodiversity data. The general tokenisation and normalisation procedures developed by the NLP Stanford team are the basis of many text mining tools. Additionally, the ClearEarth[39] project

[33] https://extract.jensenlab.org/
[34] https://species.jensenlab.org
[35] https://environments.jensenlab.org
[36] https://tissues.jensenlab.org/About
[37] https://www.brenda-enzymes.org/

[38] https://nlp.stanford.edu/software/CRF-NER.html
[39] http://github.com/ClearEarthProject/ClearEarthNLP

(Thessen et al., 2018) can tag biotic and abiotic entities, localities, units and values in text and is built using the ClearTK NLP toolkit[40] (Bethard et al., 2014). Upon installation it downloads multiple dictionaries and takes up to six gigabytes of space. It relies on Stanford NLP and other dependencies and provides a Python wrapper and a CLI.

A common constraint in historical documents is the lack of coordinates from the sampling areas, so the data curator should provide the coordinates using the toponyms given. There are tools that enable this procedure, such as Marine Gazetteer. BioStor-Lite map[41], which contains automated geolocation annotation of BHL documents (Page, 2019b), displays the points on the global map providing the user the ability to search for additional documents with selected points on the map or by drawing rectangles. The Edinburgh geoparser (Alex et al., 2015), a command line tool, recognises places in text and is one of very few tools to have this functionality. The Stanford NER system has been used as well (Stahlman and Sheffield, 2019) upon receiving training, for geolocation recognition.

Entity Normalisation and Mapping

Mapping the information retrieved from the NER tools to different IDs is crucial for cross-platform interoperability, ensuring a good output requires the mapping services to be up to date.

Taxon names can have multiple IDs depending on the platform, taxonomy common IDs, apart from the Linnaean system, are the LSID, NCBI taxonomy identifiers, EOL identifiers etc. For marine species LSIDs based on Aphia IDs are the most widely adopted.

Ontobee[42], a web server that links ontologies, is useful for the annotation of text to ontology IDs (Xiang et al., 2011). Curators can provide text snippets to Ontobee in order to retrieve ontology terms regarding environmental features (e.g. ENVO IDs), functional traits (e.g. PATO IDs[43] (Tan et al., 2022)) or other ontology terms of interest. Currently, the use of entire documents is not recommended.

The WoRMS Taxon match[44] tool matches the taxon list found against the World Register of Marine Species (WoRMS) taxon LSID. Geographic regions are confirmed with the use of the georeference tool developed for the Marine Gazetteer, users can enter the location name in the gazetteer search field of the web interface and the result's output includes the region's boundaries and the corresponding MRGID.

Most vocabulary servers provide APIs that map the different IDs. EMODnet Biology has adopted LSIDs for marine species based on Aphia IDs from the WoRMS vocabulary, which provides a dedicated API and an R package worrms (Chamberlain, 2020). Additionally, the R package taxize (Chamberlain and Szöcs, 2013) provides taxon mapping capabilities across many data sources (i.e. NCBI taxonomy, Integrated Taxonomic Information System, Encyclopedia of Life, WoRMS). Functions like get_eolid, get_nbnid, get_wormsid can perform mapping across rows of the taxon name of the case study. In addition, the GloBI[45] (Global Biotic Interactions) nomer tool[46] (Poelen and Salim, 2022) can also be used as it provides entity mapping functionality *via* CLI (Poelen et al., 2014).

Data Transformations

In this step, curators organise data according to the Darwin Core[47] standard and extensions, such as extended Measurement or Fact Extension[48], resulting in the creation of a Darwin Core Archive (see guidelines *via* the link[49]) with detailed sampling descriptors and terms based on controlled vocabularies.

When considering data transformations, curators tend to use GUI spreadsheet applications like Microsoft Excel, Google Sheets and LibreOffice Calc. OpenRefine[50] is a free, open source software that handles messy data and provides their transformation in various ways (Ham, 2013). The software's main goal is to provide data cleaning, fixing and analysing while also enhancing the interconnection between different datasets (Verborgh and De Wilde, 2013).

Automation can be used for this transformation through CLI tools like the R tidyverse[51] package suite, Python pandas[52] library and AWK programming language[53], among others. These tools support fast and scalable tabular and text data handling, manipulations, merging and filtering. The choice of tools depends on the users' familiarity, expertise and operating system.

Quality Control

Prior to publishing the dataset it is important to perform sanity checks and quality checks to ensure that the data comply with the Darwin Core Standards (Vandepitte et al., 2015). LifeWatch-EMODnetBiology QC tool[54] allows the use of the IPT URL or the dataset's DwC-A files and provides a list of the quality issues encountered, according to the EMODnet Biology criteria, as an output. It is available as a Web Application interface, based on RShiny, and as a R package[55] (De Pooter and Perez-Perez, 2019). LifeWatch Belgium Data Services'[56] has similar functionalities, providing a compilation of data services from plain text and spreadsheet files as input. The GBIF Data Validator[57] combines all the above mentioned options, in terms of input, and provides a detailed summary of issues encountered in data and metadata. Open Refine, is equipped with a few extensions that can also check for taxon names and reconcile them.

[40] http://cleartk.github.io/cleartk/
[41] https://biostor.org/map.php
[42] https://www.ontobee.org/
[43] https://github.com/pato-ontology
[44] http://www.marinespecies.org/aphia.php?p=match
[45] https://www.globalbioticinteractions.org/
[46] https://github.com/globalbioticinteractions/nomer
[47] https://dwc.tdwg.org/
[48] https://manual.obis.org
[49] https://www.gbif.org/tool/81282/darwin-core-archive-assistant
[50] https://openrefine.org/
[51] https://www.tidyverse.org
[52] https://pandas.pydata.org
[53] https://en.wikipedia.org/wiki/AWK
[54] https://rshiny.lifewatch.be/BioCheck/
[55] https://github.com/EMODnet/EMODnetBiocheck
[56] https://www.lifewatch.be/data-services/
[57] http://gbif.org/tools/data-validator

TABLE 2 | The platforms where the CLI workflow was tested.

OS	Source code - running time	Container - running time (minutes)	CPU	RAM (GB)
macOS Catalina 10.15.7	28 minutes	Docker - 33'	Intel(R) Core(TM) i5-4258U CPU @ 2.40GHz	8
Linux Ubuntu 18.04.5 LTS (Bionic Beaver)	20 minutes	Docker - 27'	Intel(R) Pentium(R) Dual-Core CPU T4200 @ 2.00GHz	4
Linux Debian server 4.9.0-8-amd64	—	Singularity - 20'	Intel(R) Xeon(R) Silver 4114 CPU @ 2.20GHz	4

Please note that running time can be affected by internet speed and stability due to API calls. The workflow uses open source tools and software libraries that are distributed across the major platforms; Linux, Mac and Windows.

The Obistools[58] R package (Provoost et al., 2019), the basis of the LifeWatch-EMODnetBiology QC tool, can be used to check the coordinate boundaries and calculate centroids in cases where the exact location is unknown. It also checks for dates' formats and events. It has comprehensive documentation and is in active development.

Upload to Database

The last step of the curation process is the publication of the standards' compliant formatted data, which is facilitated by the Integrated Publishing Toolkit[59] (IPT) software platform (Robertson et al., 2014).

Curators create an IPT resource entry with the aforementioned data and associated metadata, which is then uploaded to an IPT instance, e.g. the MedOBIS[60] Repository (Arvanitidis et al., 2006). In the case of MedOBIS, the IPT is subsequently harvested and made available by the central OBIS[61] system, thus being a strong example and supporter of the 'collect once, use many times' concept.

One-Stop-Shop Tools

The main all-in-one GUI computer program is Golden-GATE-imagine[62], an updated version of GoldenGATE editor (Sautter et al., 2007). This tool supports OCR, NER and entity mapping, as described in the various steps of the curator's workflow by providing annotations on PDF backed up by ontologies. It was developed by Plazi in 2015 and was last updated in 2016. Several recent biodiversity data related publications still report the use of it although it has not been updated since that time (Miller et al., 2019; Rivera-Quiroz and Miller, 2019; Agosti et al., 2020). Due to its open source nature, Golden-Gate-imagine can be further developed by any interested parties, as exemplified in GNfinder.

DECO: A Biodiversity Data Curation Programming Workflow

A CLI workflow named DECO developed to demonstrate the advantages of the CLI approach, is available *via* this GitHub repository[63]. DECO has connected different tools of the programming curation steps (**Figure 3**). The execution is *via* a single command with a user-provided PDF file and the output are the taxon names and records from WoRMS API, taxonomy NCBI IDs and ENVO terms from the Environmental Ontology. Complementary tools (i.e. Ghostscript[64], jq[65] and ImageMagick[66]) and UNIX commands are also called in a single Bash script which unifies the workflow. In order to simplify the setup procedure of the workflow a Docker container and a Singularity container were developed that include all the dependencies and the code. The code and both containers have been tested on Ubuntu, Mac and Debian server (**Table 2**). For a larger corpus of biodiversity historical data the recommendation is to use the Singularity container in a remote server or a High Performance Computing (HPC) cluster.

DISCUSSION

Data Rescue Landscape

The huge difference between rescued historical marine datasets uploaded on OBIS and the available digital items on BHL holdings reflects the challenges faced by curators and the minimal attention paid by the wider community, when compared to other data rescue activities (e.g. specimens, oceanographic data, etc.). Many publications lack basic metadata such as location, date, purpose or method of sampling. Tracing this information is limited as the data providers may (a) have forgotten these details, (b) be retired or (c) be deceased (Michener et al., 1997).

The project 'Census of Marine Life' included, among its initial objectives, the rescue of historical marine data. Since then, there have been ongoing efforts within the EMODnet Biology project and LifeWatch Research Infrastructure, among others. Similarly, initiatives like Global Oceanographic Data Archaeology and Rescue[67] (GODAR), Oceans Past Initiative[68] (OPI) and RECovery of Logbooks And International Marine data[69] (RECLAIM) (Wilkinson et al., 2011) rescue data from ship logs for oceanographic, climate and biodiversity data. More effort is however needed, as exemplified by museum specimen collections and herbaria digitisation (Mora et al., 2011; Wheeler et al., 2012). The museum specimen collections and herbaria digitisation has

[58] https://github.com/iobis/obistools
[59] https://www.gbif.org/ipt
[60] https://www.lifewatchgreece.eu/?q=content/medobis
[61] https://manual.obis.org
[62] https://github.com/plazi/GoldenGATE-Imagine
[63] https://github.com/lab42open-team/deco
[64] https://www.ghostscript.com/index.html
[65] https://stedolan.github.io/jq/
[66] https://imagemagick.org/index.php
[67] https://www.ncei.noaa.gov/products/ocean-climate-laboratory/global-oceanographic-data-archaeology-and-rescue
[68] https://oceanspast.org
[69] https://icoads.noaa.gov/reclaim/

multiple projects and infrastructures like Distributed System of Scientific Collections[70] (DiSSCo), Innovation and consolidation for large scale digitisation of natural heritage[71] (ICEDIG), Integrated Digitized Biocollections[72] (iDigBio) and Biodiversity Community Integrated Knowledge Library (BiCIKL) (Penev et al., 2022). Similar attention is required to rescue marine biodiversity data from historical documents that can contribute to a more complete global biodiversity synthesis (Heberling et al., 2021).

In the last few years, an upsurge in web applications development regarding the enhancement of biodiversity data digitisation has been observed. This is an indication of the need for such initiatives. Advancements in the field of OCR, text mining and information technology promise semi-automation and acceleration of the curator's work, which could transform the biodiversity curation field into an -omics like, interdisciplinary field that requires complementary skills of document handling, web technologies and text mining, to name but a few.

Interface Remarks

Web applications provide the advantage of visual aids (e.g. highlights of NER terms), which improve the evaluation easiness and intuitiveness when using their graphical interfaces. Emerging web development technologies like R Shiny, Flask[73] and Django[74] among others, have simplified the processes of web application development. These applications are powerful and effective in most cases but are siloed in functionality and extendability, they also have many software dependencies which increase instability, when not maintained in the long term.

CLI tools are a powerful way to implement scalable, reproducible and replicable workflows: scalable because the same code can be applied to multiple files (e.g. in this case, the various documents); reproducible and replicable because the code can be executed multiple times and with different types of documents, respectively. Furthermore, they usually have additional functionalities that have not been implemented in their web application counterparts. The difficulties regarding CLI tools' dependency and portability are being resolved with the rise of containerised applications which include all system requirements and are distributed through web repositories like Docker Hub[75], the downside is that without interactiveness they are cumbersome when assisting the curation process.

Sustainability

Tool usability relies on active development and continuous support and debugging. Sustainability is considered the main issue regarding the tools' functionality. An example is EnvMine (Tamames and de Lorenzo, 2010), a promising 2010 cutting edge tool which is no longer available. One-stop-shop purpose software applications for domain specific usage, like GoldenGate, are very helpful but require more effort to stay up to date with the integrated tools. Other tools are often out of date, as active development and contribution to reporting issues in open-source repositories, such as Github, is lacking, thus becoming obsolete and unsupported in only a few years from their first release.

Curation Step-Wise Remarks

The curators' role is invaluable in the data rescue process, as their domain specific expertise is far from becoming entirely automated. There are plenty of available digitised historical documents that are not curated in web libraries, such as BHL, the Belgian Marine Bibliography[76], Web of Science[77], Wiley Online Library[78] and Taylor & Francis Online[79], among others (Kearney, 2019). BHL provides "OCRed" documents and there are plenty of other tools that can tackle this process which are reviewed elsewhere (Owen et al., 2020), however OCR is a crucial limiting step in the workflows, in terms of the information transformed from image to text, because there are many cases that lead to mispelled or lost text; especially the case with handwritten text and poor quality images (Lyal, 2016).

Information extraction can be performed both on a small and a large scale. Named Entities are what most text mining tools extract. Taxon names recognition is the main function of the majority of the current tools and has matured significantly over the past decade, especially through the integration of multiple platforms with the GNA (Pyle, 2016). Environments and geolocations have strong background data, Environment Ontology terms (retrieved with the EXTRACT tool) and GeoNames[80]/Marineregions gazetteers, respectively. Geolocation mining, in particular, has not been adapted in biodiversity curation but there are generic tools (e.g. mordecai[81] - Halterman, 2017) that are able to be trained with gazetteers to extract approximate localities from text. Also extraction of sample location from maps is possible by first geolocating the historic map in Geographic Information Systems (Jenny and Hurni, 2011) and then using computer vision to find the locations' coordinates (Chiang et al., 2014). Characteristics of taxa, i.e. phenotypic traits, associated physico-chemical variables, units and the use of semantics to describe relations, are still under standardisation (Thessen et al., 2020) and NER prototypes have been made with ClearEarth and Pensoft Annotator, for example.

Entity mapping has also seen an important development because there are many open public APIs for vocabularies like those used in WoRMS, and Marine Regions and aggregators such as GBIF and OBIS, among others, and in some cases software packages (mostly in the R programming language). The task for Publication has its dedicated applications and tools with the CLI tools being able to perform quality control and deliver a preferred on-the-fly format.

[70] https://www.dissco.eu
[71] https://icedig.eu/
[72] https://www.idigbio.org/
[73] https://flask.palletsprojects.com/
[74] https://www.djangoproject.com/
[75] https://hub.docker.com/
[76] https://www.vliz.be/en/belgian-marine-bibliography
[77] https://www.webofknowledge.com
[78] https://onlinelibrary.wiley.com
[79] https://www.tandfonline.com/
[80] http://www.geonames.org
[81] https://github.com/openeventdata/mordecai

DECO

The CLI scientific workflow assembled in this paper, DECO, is a demonstration of EMODnet Biology's vision for biodiversity data rescue using programming tools. To the best of our knowledge, this is the first task-driven CLI that brings together state-of-the-art image processing, OCR tools, text mining technologies and Web APIs, in order to assist curators. By using programming interface and Command Line Tools the workflow is scalable, customisable and modular, meaning that more tools can be incorporated to, e.g. include the entities mentioned in the previous section. It is fast, may be used on a personal computer, and is available as a Docker and a Singularity container. The containerised versions of the workflow simplify the installation procedure and increase its stability, scalability and portability because they include all the necessary dependencies. This CLI scientific workflow promises a faster and high throughput processing that could be applied to any type of data, not only historical, thus contributing to the overall digitisation of biodiversity knowledge.

Future Outlook

Progress has been made in the advancement of the historical data rescue process, from digitisation platforms to standards, services and publication (Beja et al., 2022). To bridge the gap between tools and curators requires effort on both ends; namely the data curators and the tool developers. It is recommended that curators are trained in basic programming skills from which they and the historical data rescue process in general would benefit in the long term (Holinski et al., 2020). Regarding software development, important features are highlighted, like the use of multiple ontologies in Pensoft Annotator. This is a direction which should be further expanded to all entities of interest. Multidisciplinary cooperation between scientific communities and partners of tools, ontologies and databases is needed to accomplish this task (Bowker, 2000). An important example was set by GNA which advanced scientific names recognition significantly. In addition, the co-occurrence feature, that was present in Biodiversity Observation Miner, once expanded to other entities and associated with a scoring scheme will be a state-of-the-art text mining application that goes beyond NER to actually infer relations. The rise of deep neural networks is promising as well in all different tasks of Information Extraction, as seen in TaxoNERD (Le Guillarme and Thuiller, 2022). Lastly, the community is pushing to Semantic Publishing, FAIR completeness of new data and new taxonomic publishing guidelines to eliminate the need of text mining and curation in current publications (Penev et al., 2019; Fawcett et al., 2022).

The implementation of crowdsourced curation within citizen science projects for the historical biodiversity data is encouraged (Clavero and Revilla, 2014; Arnaboldi et al., 2020; Holinski et al., 2020). Practices like this are already in place in the digitisation of natural history collections and have been proved fruitful (Ellwood et al., 2015). EMODnet Biology's Phase IV will launch such a citizen science project for historical documents curation during the second half of 2022. Approaches from other fields of science that handle historical and old data, such as history, linguistics, archaeology would provide useful insights for the text mining of historical biodiversity data.

Concluding Remarks

Historical marine biodiversity data provide important and significant snapshots of the past that can help understand the current status of ocean ecosystems and predict future trends in face of the climate crisis. There is a wealth of historical documents that have been digitised yet, most of their data have not been rescued or published in online systems. To accelerate the tedious data rescue process it is essential that more curators become engaged, and tools for Information Extraction and Publication get improved to satisfy their needs. Tools like DECO and GoldenGATE demonstrate possible future directions for one-stop-shop applications for command line and graphical user interfaces, respectively. Research Infrastructures can play a pivotal role towards this goal. Last but not least, the community and funding bodies should prioritise the data rescue of these invaluable documents before their decay and inevitable loss.

DATA AVAILABILITY STATEMENT

DECO is available here: https://github.com/lab42open-team/deco. Historical marine literature analysis is here: https://github.com/savvas-paragkamian/historical-marine-literature. BHL, EMODnet Biology and OBIS data are available for download here https://about.biodiversitylibrary.org/tools-andservices/developer-and-data-tools/ and https://www.emodnetbiology.eu/toolbox/en/download/occurrence/explore and herehttps://obis.org/manual/access/, respectively. The digitised document of the "Report on the Mollusca andRadiata of the Aegean Sea, and on their distribution, considered as bearing on Geology. 13th Meeting of the British Association for the Advancement of Science, London, 1844" is available here: https://www.biodiversitylibrary.org/page/12920789. The curated dataset of the case study is available here (version 1.9 and above): http://ipt.medobis.eu/resource?r=mollusca_forbes.

AUTHOR CONTRIBUTIONS

Conceptualisation: CA, EP, and VG. Wrote first draft of the manuscript: SP, GS, DM, CP, CA, EP, and VG. Revised the manuscript: all. Web applications testing: SP, GS, ME, RP. Programming tools testing: SP, HZ, and EP. Code development and containerisation: SP and HZ. Work Package Leaders: VG and DM. Project coordinator: JB. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by EMODnet Biology Phase III (EASME/EMFF/2016/-1.3.1.2/Lot 5/SI2.750022 and EASME/EMFF/2017/1.3.1.2/02/SI2.789013) and Phase IV (EMFF/2019/1.3.1.9/Lot 6/SI2.837974). The European Marine Observation and Data Network (EMODnet) is financed by the European Union under Regulation (EU) No 508/2014 of the European Parliament and of the Council of 15 May 2014 on the European Maritime and Fisheries Fund. SP was supported also by EMODnet

Biology Phase IV and for different parts of his work he was supported from the project "Centre for the study and sustainable exploitation of Marine Biological Resources (CMBR)" (MIS 5002670), which is implemented under the Action "Reinforcement of the Research and Innovation Infrastructure," funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014–2020) and co-financed by Greece and the EU (European Regional Development Fund). GS received support from EMODnet Biology Phase III and Phase IV. SP and HZ received support from the Hellenic Foundation for Research and Innovation(HFRI) and the General Secretariat for Research and Innovation (GSRI), under grant agreement No. 241 (PREGO project). DM and VG have received support from LifeWatchGreece Research Infrastructure (Arvanitidis et al., 2016) and "Centre for the study and sustainable exploitation of Marine Biological Resources (CMBR)" (MIS 5002670), which is implemented under the Action "Reinforcement of the Research and Innovation Infrastructure," funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund). LB received support by EMODNET Biology, Phase IV. The work of LV is funded by Research Foundation - Flanders (FWO) as part of the Belgian contribution to LifeWatch. For different aspects of his work HZ received support from ELIXIR-GR: Managing and Analysing Life Sciences Data (MIS: 5002780) which is co-financed by Greece and the European Union - European Regional Development Fund. CA received support from LifeWatch ERIC. LifeWatch ERIC funded the publication fees. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmars.2022.940844/full#supplementary-material

REFERENCES

Abrami, G., Henlein, A., Lücking, A., Kett, A., Adeberg, P. and Mehler, A. (2021). Unleashing Annotations With TextAnnotator: Multimedia, Multi-Perspective Document Views for Ubiquitous Annotation. in *Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation* (Groningen, The Netherlands (online): Association for Computational Linguistics), 65–75. Available at: https://aclanthology.org/2021.isa-1.7

Agosti, D., Guidoti, M., Catapano, T., Ioannidis-Pantopikos, A. and Sautter, G. (2020). The Standards Behind the Scenes: Explaining Data From the Plazi Workflow. *Biodiversity. Inf. Sci. Standards.* 4, e59178. doi: 10.3897/biss.4.59178

Alex, B., Byrne, K., Grover, C. and Tobin, R. (2015). Adapting the Edinburgh Geoparser for Historical Georeferencing. *IJHAC* 9, 15–35. doi: 10.3366/ijhac.2015.0136

Alex, B., Grover, C., Haddow, B., Kabadjov, M., Klein, E., Matthews, M., et al. (2008). Assisted Curation: Does Text Mining Really Help? *Pac. Symp. Biocomput.* 556–567. doi: 10.1142/9789812776136_0054

Ananiadou, S. and Mcnaught, J. (2005). *Text Mining for Biology and Biomedicine* (USA: Artech House, Inc).

Anderson, K. (2006). Does History Count? *Endeavour* 30, 150–155. doi: 10.1016/j.endeavour.2006.11.002

Arnaboldi, V., Raciti, D., Van Auken, K., Chan, J. N., Müller, H.-M. and Sternberg, P. W. (2020). Text Mining Meets Community Curation: A Newly Designed Curation Platform to Improve Author Experience and Participation at WormBase. *Database* 2020. doi: 10.1093/database/baaa006

Arvanitidis, C., Chatzinikolaou, E., Gerovasileiou, V., Panteri, E., Bailly, N., Minadakis, N., et al. (2016). LifeWatchGreece: Construction and Operation of the National Research Infrastructure (ESFRI). *BDJ* 4, e10791. doi: 10.3897/BDJ.4.e10791

Arvanitidis, C., Valavanis, V., Eleftheriou A, D., Costello, M. J., Faulwetter, S., Gotsis, P., et al. (2006). MedOBIS: Biogeographic Information System for the Eastern Mediterranean and Black Sea. *Mar. Ecol. Prog. Ser.* 316, 225–230. doi: 10.3354/meps316225

Ausubel, J. H. (1999). GUEST EDITORIAL: Toward a Census of Marine Life. *Oceanography* 12, 4–5. doi: 10.5670/oceanog.1999.17

Batista-Navarro, R., Zerva, C., Nguyen, N. T. H. and Ananiadou, S. (2017). "A Text Mining-Based Framework for Constructing an RDF-Compliant Biodiversity Knowledge Repository," in *Information Management and Big Data*. Eds. Lossio-Ventura, J. A. and Alatrista-Salas, H. (Cham: Springer International Publishing), 30–42.

Beja, J., Vandepitte, L., Benson, A., Van de Putte, A., Lear, D., De Pooter, D., et al. (2022). "Chapter Two - Data services in ocean science with a focus on the biology", in Ocean Science Data, eds. G. Manzella and A. Novellino (Amsterdam, Netherlands: Elsevier), 67–129. doi: 10.1016/B978-0-12-823427-3.00006-2

Bethard, S., Ogren, P., and Becker, L. (2014). ClearTK 2.0: Design Patterns for Machine Learning in UIMA. in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) [Reykjavik, Iceland: European Language Resources Association (ELRA)], 3289–3293. Available at: http://www.lrec-conf.org/proceedings/lrec2014/pdf/218_Paper.pdf.

Bowker, G. C. (2000). Biodiversity Datadiversity. *Soc. Stud. Sci.* 30, 643–683. doi: 10.1177/030631200030005001

Buttigieg, P. L., Pafilis, E., Lewis, S. E., Schildhauer, M. P., Walls, R. L. and Mungall, C. J. (2016). The Environment Ontology in 2016: Bridging Domains With Increased Scope, Semantic Density, and Interoperation. *J. Biomed. Semantics.* 7, 57. doi: 10.1186/s13326-016-0097-6

Calder, W. A. (1982). A Proposal for the Standardization of Units and Symbols in Ecology. *Bull. Ecol. Soc. America* 63, 7–10.

Chamberlain, S. (2020) *Worrms: World Register of Marine Species (WoRMS) Client*. Available at: https://cran.r-project.org/package=worrms.

Chamberlain, S. A. and Szöcs, E. (2013). Taxize: Taxonomic Search and Retrieval in R. *F1000Res* 2. doi: 10.12688/f1000research.2-191.v2

Chiang, Y.-Y., Leyk, S. and Knoblock, C. A. (2014). A Survey of Digital Map Processing Techniques. *ACM Comput. Surv.* 47, 1–44. doi: 10.1145/2557423

Claus, S., De Hauwere, N., Vanhoorne, B., Deckers, P., Souza Dias, F., Hernandez, F., et al. (2014). Marine Regions: Towards a Global Standard for Georeferenced Marine Names and Boundaries. *null* 37, 99–125. doi: 10.1080/01490419.2014.902881

Clavero, M. and Revilla, E. (2014). Mine Centuries-Old Citizen Science. *Nature* 510, 35–35. doi: 10.1038/510035c

Costello, M. J., Michener, W. K., Gahegan, M., Zhang, Z.-Q. and Bourne, P. E. (2013). Biodiversity Data Should be Published, Cited, and Peer Reviewed. *Trends Ecol. Evol.* 28, 454–461. doi: 10.1016/j.tree.2013.05.002

De Pooter, D. and Perez-Perez, R. (2019) *EMODnetBiocheck: LifeWatch & EMODnet Biology QC Tool*. Available at: https://github.com/EMODnet/EMODnetBiocheck.

Dimitrova, M., Zhelezov, G., Georgiev, T. and Penev, L. (2020). The Pensoft Annotator: A New Tool for Text Annotation With Ontology Terms. *BISS* 4, e59042. doi: 10.3897/biss.4.59042

Driller, C., Koch, M., Abrami, G., Hemati, W., Lücking, A., Mehler, A., et al. (2020). Fast and Easy Access to Central European Biodiversity Data With BIOfid. *BISS* 4, e59157. doi: 10.3897/biss.4.59157

Driller, C., Koch, M., Schmidt, M., Weiland, C., Hörnschemeyer, T., Hickler, T., et al. (2018). Workflow and Current Achievements of BIOfid, an Information Service Mobilizing Biodiversity Data From Literature Sources. *Biodiversity. Inf. Sci. Standards.* 2, e25876. doi: 10.3897/biss.2.25876

Ellwood, E. R., Dunckel, B. A., Flemons, P., Guralnick, R., Nelson, G., Newman, G., et al. (2015). Accelerating the Digitization of Biodiversity Research Specimens Through Online Public Participation. *BioScience* 65, 383–396. doi: 10.1093/biosci/biv005

Engelhard, G. H., Thurstan, R. H., MacKenzie, B. R., Alleway, H. K., Bannister, R. C. A., Cardinale, M., et al. (2016). ICES Meets Marine Historical Ecology: Placing the History of Fish and Fisheries in Current Policy Context. *ICES J. Mar. Sci.* 73, 1386–1403. doi: 10.1093/icesjms/fsv219

Faulwetter, S., Pafilis, E., Fanini, L., Bailly, N., Agosti, D., Arvanitidis, C., et al. (2016). EMODnet Workshop on Mechanisms and Guidelines to Mobilise Historical Data Into Biogeographic Databases. *RIO* 2, e9774. doi: 10.3897/rio.2.e9774

Fawcett, S., Agosti, D., Cole, S. R. and Wright, D. F. (2022). Digital Accessible Knowledge: Mobilizing Legacy Data and the Future of Taxonomic Publishing. *Bull. Soc. Systematic. Biologists.* 1 (12). doi: 10.18061/bssb.v1i1.8296

Finkel, J. R., Grenager, T. and Manning, C. (2005). "Incorporating Non-Local Information Into Information Extraction Systems by Gibbs Sampling," in Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Ann Arbor, Michigan. 363–370. doi: 10.3115/1219840.1219885

Forbes, E. (1844). Report on the Mollusca and Radiata of the Aegean Sea, and on Their Distribution, Considered as Bearing on Geology. *Rep. Br. Assoc. Advancement. Sci.* 1843, 130–193.

Fortibuoni, T., Libralato, S., Raicevich, S., Giovanardi, O. and Solidoro, C. (2010). Coding Early Naturalists' Accounts Into Long-Term Fish Community Changes in the Adriatic Sea, (1800–2000). *PLos One* 5, e15502. doi: 10.1371/journal.pone.0015502

GBIF *The Global Biodiversity Information Facility GBIF: The Global Biodiversity Information Facility*. Available at: https://www.gbif.org/citation-guidelines (Accessed April 6, 2022).

Goethem, T.v. and Zanden, J.L.v. (2021) *Biodiversity Trends in a Historical Perspective*. Available at: https://www.oecd-ilibrary.org/content/component/2c94883d-en.

Griffin, E. (2019). Getting Necessary Historical Data Out of Deep Freeze. *Polar. Sci.* 21, 238–239. doi: 10.1016/j.polar.2019.05.008

Groom, Q., Dillen, M., Hardy, H., Phillips, S., Willemse, L. and Wu, Z. (2019). Improved Standardization of Transcribed Digital Specimen Data. *Database 2019*, baz129. doi: 10.1093/database/baz129

Groom, Q., Güntsch, A., Huybrechts, P., Kearney, N., Leachman, S., Nicolson, N., et al. (2020). People are Essential to Linking Biodiversity Data. *Database 2020*. 2020. doi: 10.1093/database/baaa072

Gwinn, N. E. and Rinaldo, C. (2009). The Biodiversity Heritage Library: Sharing Biodiversity Literature With the World. *IFLA. J.* 35, 25–34. doi: 10.1177/0340035208102032

Halterman, A. (2017). Mordecai: Full Text Geoparsing and Event Geocoding. *J. Open Source Software.* 2, 91. doi: 10.21105/joss.00091

Ham, K. (2013). OpenRefine (Version 2.5). Http://Openrefine.Org. Free, Open-Source Tool for Cleaning and Transforming Data. *J. Med. Libr. Assoc.* 101, 233–234. doi: 10.3163/1536-5050.101.3.020

Hearst, M. A. (1999). Untangling text data mining. in Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (*College Park, Maryland: Association for Computational Linguistics*), 3–10. doi: 10.3115/1034678.1034679

Heath, T. and Bizer, C. (2011). *Linked Data: Evolving the Web Into a Global Data Space*. 1st ed. San Rafael, California (USA): Morgan & Claypool Available at: doi: 10.2200/S00334ED1V01Y201102WBE001 [Accessed March 22, 2022]

Heberling, J.M., Miller, J. T., Noesgaard, D., Weingart, S. B. and Schigel, D. (2021). Data Integration Enables Global Biodiversity Synthesis. *Proc. Natl. Acad. Sci.* 118, e2018093118. doi: 10.1073/pnas.2018093118

Heidorn, P. B. (2008). Shedding Light on the Dark Data in the Long Tail of Science. *Library. Trends* 57, 280–299. doi: 10.1353/lib.0.0036

Herrmann, E. (2020). Building the Biodiversity Heritage Library's Technical Strategy. *BISS* 4, e59084. doi: 10.3897/biss.4.59084

Holinski, A., Burke, M., Morgan, S., McQuilton, P. and Palagi, P. (2020). Biocuration - Mapping Resources and Needs [Version 2; Peer Review: 2 Approved]. *F1000Research* 9. doi: 10.12688/f1000research.25413.2

Jenny, B. and Hurni, L. (2011). Studying Cartographic Heritage: Analysis and Visualization of Geometric Distortions. *Comput. Graphics* 35, 402–411. doi: 10.1016/j.cag.2011.01.005

Jensen, L. J. (2016). One Tagger, Many Uses: Illustrating the Power of Ontologies in Dictionary-Based Named Entity Recognition. *bioRxiv*, 067132. doi: 10.1101/067132

Kearney, N. (2019). It's Not Always FAIR: Choosing the Best Platform for Your Biodiversity Heritage Literature. *BISS* 3, e35493. doi: 10.3897/biss.3.35493

Klein, E., Appeltans, W., Provoost, P., Saeedi, H., Benson, A., Bajona, L., et al. (2019). OBIS Infrastructure, Lessons Learned, and Vision for the Future. *Front. Mar. Sci.* 6. doi: 10.3389/fmars.2019.00588

Kwok, R. (2017). Historical Data: Hidden in the Past. *Nature* 549, 419–421. doi: 10.1038/nj7672-419

Lamurias, A. and Couto, F. M. (2019). ""Text Mining for Bioinformatics Using Biomedical Literature,"" in *Encyclopedia of Bioinformatics and Computational Biology*. Eds. Ranganathan, S., Gribskov, M., Nakai, K. and Schönbach, C. (Oxford: Academic Press), 602–611. doi: 10.1016/B978-0-12-809633-8.20409-3

Le Guillarme, N. and Thuiller, W. (2022). TaxoNERD: Deep Neural Models for the Recognition of Taxonomic Entities in the Ecological and Evolutionary Literature. *Methods Ecol. Evol.* 13, 625–641. doi: 10.1111/2041-210X.13778

Levin, S. A. (1992). The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture. *Ecology* 73, 1943–1967. doi: 10.2307/1941447

Lin, X. F. (2006). Quality assurance in high volume document digitization: a survey. in Second International Conference on Document Image Analysis for Libraries (*DIAL'06*) (Lyon, France: IEEE), 312–319. doi: 10.1109/DIAL.2006.33

Lo Brutto, S. (2021). Historical and Current Diversity Patterns of Mediterranean Marine Species. *Diversity* 13. doi: 10.3390/d13040156

Lotze, H. K. and Worm, B. (2009). Historical Baselines for Large Marine Animals. *Trends Ecol. Evol.* 24, 254–262. doi: 10.1016/j.tree.2008.12.004

Lyal, C. H. C. (2016). Digitising Legacy Zoological Taxonomic Literature: Processes, Products and Using the Output. *ZK* 550, 189–206. doi: 10.3897/zookeys.550.9702

Martín Míguez, B., Novellino, A., Vinci, M., Claus, S., Calewaert, J.-B., Vallius, H., et al. (2019). The European Marine Observation and Data Network (EMODnet): Visions and Roles of the Gateway to Marine Data in Europe. *Front. Mar. Sci.* 6. doi: 10.3389/fmars.2019.00313

Mavraki, D., Fanini, L., Tsompanou, M., Gerovasileiou, V., Nikolopoulou, S., Chatzinikolaou, E., et al. (2016). Rescuing Biogeographic Legacy Data: The "Thor" Expedition, a Historical Oceanographic Expedition to the Mediterranean Sea. *Biodiversity. Data J.* 4, e11054. doi: 10.3897/BDJ.4.e11054

Mavraki, D., Sarafidou, G., Legaki, A., Nikolopoulou, S., and Gerovasileiou, V. (2021). Digitization of the dredging papers included in the Report on the Mollusca and Radiata of the Aegean Sea, and on their distribution, considered as bearing on Geology by Edward Forbes, 13th Meeting of the British Association for the Advancement of Science, London, 1844. *Heraklion*: Hellenic Center for Marine Research Available at: http://ipt.medobis.eu/resource?r=mollusca_forbes.

McClenachan, L., Ferretti, F. and Baum, J. K. (2012). From Archives to Conservation: Why Historical Data are Needed to Set Baselines for Marine Animals and Ecosystems. *Conserv. Lett.* 5, 349–359. doi: 10.1111/j.1755-263X.2012.00253.x

Michener, W. K. (2015). Ecological Data Sharing. *Ecol. Inf.* 29, 33–44. doi: 10.1016/j.ecoinf.2015.06.010

Michener, W. K., Brunt, J. W., Helly, J. J., Kirchner, T. B. and Stafford, S. G. (1997). Nongeospatial Metadata for the Ecological Sciences. *Ecol. Appl.* 7, 330–342. doi: 10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2

Miller, J.A.Â., Braumuller, Y., Kishor, P., Shorthouse, D. P., Dimitrova, M., Sautter, G., et al. (2019). Mobilizing Data From Taxonomic Literature for an Iconic Species (Dinosauria, Theropoda, Tyrannosaurus Rex). *Biodiversity. Inf. Sci. Standards.* 3, e37078. doi: 10.3897/biss.3.37078

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B. and Worm, B. (2011). How Many Species Are There on Earth and in the Ocean? *PLos Biol.* 9, e1001127. doi: 10.1371/journal.pbio.1001127

Mouquet, N., Lagadeuc, Y., Devictor, V., Doyen, L., Duputié, A., Eveillard, D., et al. (2015). REVIEW: Predictive Ecology in a Changing World. *J. Appl. Ecol.* 52, 1293–1310. doi: 10.1111/1365-2664.12482

Mozzherin, D., Myltsev, A. and Zalavadiya, H. (2022). *Gnames/Gnfinder: V0.18.3. Zenodo*. doi: 10.5281/zenodo.6378012

Muñoz, G., Kissling, W. D. and van Loon, E. E. (2019). Biodiversity Observations Miner: A Web Application to Unlock Primary Biodiversity Data From Published Literature. *Biodiversity. Data J.* 7, e28737. doi: 10.3897/BDJ.7.e28737

Nelson, G. and Ellis, S. (2019). The History and Impact of Digitization and Digital Data Mobilization on Biodiversity Research. *Philos. Trans. R. Soc. B.: Biol. Sci.* 374, 20170391. doi: 10.1098/rstb.2017.0391

Owen, D., Groom, Q., Hardisty, A., Leegwater, T., Livermore, L., van Walsum, M., et al. (2020). Towards a Scientific Workflow Featuring Natural Language Processing for the Digitisation of Natural History Collections. *Res. Ideas. Outcomes.* 6, e58030. doi: 10.3897/rio.6.e58030

Pafilis, E., Bērzinš, R., Arvanitidis, C. and Jensen, L. (2017). EXTRACT 2.0: Interactive Identification of Biological Entities Mentioned in Text to Assist Database Curation and Knowledge Extraction. *Biodiversity. Inf. Sci. Standards.* 1, e20152. doi: 10.3897/tdwgproceedings.1.20152

Pafilis, E., Frankild, S. P., Fanini, L., Faulwetter, S., Pavloudi, C., Vasileiadou, A., et al. (2013). The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text. *PLos One* 8, e65390. doi: 10.1371/journal.pone.0065390

Pafilis, E., Frankild, S. P., Schnetzer, J., Fanini, L., Faulwetter, S., Pavloudi, C., et al. (2015). ENVIRONMENTS and EOL: Identification of Environment Ontology Terms in Text and the Annotation of the Encyclopedia of Life. *Bioinformatics* 31, 1872–1874. doi: 10.1093/bioinformatics/btv045

Page, R. D. (2011). Extracting Scientific Articles From a Large Digital Archive: BioStor and the Biodiversity Heritage Library. *BMC Bioinf.* 12, 187. doi: 10.1186/1471-2105-12-187

Page, R. (2016). Towards a Biodiversity Knowledge Graph. *RIO* 2, e8767. doi: 10.3897/rio.2.e8767

Page, R. D. M. (2019a). Reconciling Author Names in Taxonomic and Publication Databases. *bioRxiv*, 870170. doi: 10.1101/870170

Page, R. (2019b). Text-mining BHL: towards new interfaces to the biodiversity literature. in Biodiversity_Next: SI33 - Improving access to hidden scientific data in the Biodiversity Heritage Library (Leiden - The Netherlands: Pensoft Publishers), e35013. doi: 10.3897/biss.3.35013

Palasca, O., Santos, A., Stolte, C., Gorodkin, J. and Jensen, L. J. (2018). TISSUES 2.0: An Integrative Web Resource on Mammalian Tissue Expression. *Database 2018*, bay003. doi: 10.1093/database/bay003

Parr, C. S., Wilson, N., Leary, P., Schulz, K. S., Lans, K., Walley, L., et al. (2014). The Encyclopedia of Life V2: Providing Global Access to Knowledge About Life on Earth. *BDJ* 2, e1079. doi: 10.3897/BDJ.2.e1079

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

Penev, L., Koureas, D., Groom, Q., Lanfear, J., Agosti, D., Casino, A., et al. (2022). Biodiversity Community Integrated Knowledge Library (BiCIKL). *RIO* 8, e81136. doi: 10.3897/rio.8.e81136

Penev, L., Mietchen, D., Chavan, V. S., Hagedorn, G., Smith, V. S., Shotton, D., et al. (2017). Strategies and Guidelines for Scholarly Publishing of Biodiversity Data. *RIO* 3, e12431. doi: 10.3897/rio.3.e12431

Perera, N., Dehmer, M. and Emmert-Streib, F. (2020). Named Entity Recognition and Relation Detection for Biomedical Information Extraction. *Front. Cell Dev. Biol.* 8. doi: 10.3389/fcell.2020.00673

Poelen, J. and Salim, J. A. (2022). *Globalbioticinteractions/Nomer:. Zenodo*. doi: 10.5281/zenodo.6478468

Poelen, J. H., Simons, J. D. and Mungall, C. J. (2014). Global Biotic Interactions: An Open Infrastructure to Share and Analyze Species-Interaction Datasets. *Ecol. Inf.* 24, 148–159. doi: 10.1016/j.ecoinf.2014.08.005

Provoost, S., Provoost, P. and Appeltans, W. (2019). *Iobis/Obistools: Version 0.0.9. Zenodo*. doi: 10.5281/zenodo.3338213

Pyle, R. L. (2016). Towards a Global Names Architecture: The Future of Indexing Scientific Names. *Zookeys*, 261–281. doi: 10.3897/zookeys.550.10009

Rainbow, P. S. (2009). Marine Biological Collections in the 21st Century. *Zoologica. Scripta.* 38, 33–40. doi: 10.1111/j.1463-6409.2007.00313.x

Reiser, L., Harper, L., Freeling, M., Han, B. and Luan, S. (2018). FAIR: A Call to Make Published Data More Findable, Accessible, Interoperable, and Reusable. *Mol. Plant* 11, 1105–1108. doi: 10.1016/j.molp.2018.07.005

Richard, J. (2020). Improving Taxonomic Name Finding in the Biodiversity Heritage Library. *Biodiversity. Inf. Sci. Standards.* 4, e58482. doi: 10.3897/biss.4.58482

Rivera-Quiroz, F. A. and Miller, J. (2019). Extracting Data From Legacy Taxonomic Literature: Applications for Planning Field Work. *Biodiversity. Inf. Sci. Standards.* 3, e37082. doi: 10.3897/biss.3.37082

Rivera-Quiroz, F. A., Petcharad, B. and Miller, J. A. (2020). Mining Data From Legacy Taxonomic Literature and Application for Sampling Spiders of the Teutamus Group (Araneae; Liocranidae) in Southeast Asia. *Sci. Rep.* 10, 15787. doi: 10.1038/s41598-020-72549-8

Robertson, T., Döring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., et al. (2014). The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet. *PLos One* 9, e102623. doi: 10.1371/journal.pone.0102623

Sautter, G., Böhm, K. and Agosti, D. (2007). Semi-Automated XML Markup of Biosystematic Legacy Literature With the GoldenGATE Editor. *Pac. Symp. Biocomput.* 12, 391–402.

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., et al. (2020). NCBI Taxonomy: A Comprehensive Update on Curation, Resources and Tools. Database *(*Oxford*)* 2020. doi: 10.1093/database/baaa062

Stahlman, G. R., and Sheffield, C. (2019). Geoparsing biodiversity heritage library collections: A preliminary exploration. in *iConference 2019 Proceedings*. (University of Maryland, College Park (UMD): iSchools). doi: 10.21900/iconf.2019.103357

Stuart-Smith, R. D., Edgar, G. J., Barrett, N. S., Kininmonth, S. J. and Bates, A. E. (2015). Thermal Biases and Vulnerability to Warming in the World's Marine Fauna. *Nature* 528, 88–92. doi: 10.1038/nature16144

Tamames, J. and de Lorenzo, V. (2010). EnvMine: A Text-Mining System for the Automatic Extraction of Contextual Information. *BMC Bioinf.* 11, 294. doi: 10.1186/1471-2105-11-294

Tan, S., Mungall, C., Vasilevsky, N., Matentzoglu, N., Osumi-Sutherland, D., Caron, A., et al. (2022). Pato-Ontology/Pato: 2022-02-20 Release. *Zenodo*. doi: 10.5281/zenodo.6190780

Thessen, A. E., Cui, H. and Mozzherin, D. (2012). Applications of Natural Language Processing in Biodiversity Science. *Adv. Bioinf.* 2012, 391574. doi: 10.1155/2012/391574

Thessen, A., Preciado, J., Jain, P., Martin, J., Palmer, M. and Bhat, R. (2018). Automated Trait Extraction Using ClearEarth, a Natural Language Processing System for Text Mining in Natural Sciences. *Biodiversity. Inf. Sci. Standards.* 2, e26080. doi: 10.3897/biss.2.26080

Thessen, A. E., Walls, R. L., Vogt, L., Singer, J., Warren, R., Buttigieg, P. L., et al. (2020). Transforming the Study of Organisms: Phenomic Data Models and Knowledge Bases. *PLos Comput. Biol.* 16, e1008376. doi: 10.1371/journal.pcbi.1008376

Thompson, K. and Richard, J. (2013). Moving Our Data to the Semantic Web: Leveraging a Content Management System to Create the Linked Open Library. *null* 13, 290–309. doi: 10.1080/19386389.2013.828551

Vandepitte, L., Bosch, S., Tyberghein, L., Waumans, F., Vanhoorne, B., Hernandez, F., et al. (2015). Fishing for Data and Sorting the Catch: Assessing the Data Quality, Completeness and Fitness for Use of Data in Marine Biogeographic Databases. *Database* 2015, bau125. doi: 10.1093/database/bau125

Verborgh, R. and De Wilde, M. (2013). *Using OpenRefine*. 1st ed. Eds. Birch, S., Gupta, S., Nayak, A. and Vairat, H. B. (Packt Publishing).

Vermeulen, N., Parker, J. N. and Penders, B. (2013). Understanding Life Together: A Brief History of Collaboration in Biology. *Endeavour* 37, 162–171. doi: 10.1016/j.endeavour.2013.03.001

Wheeler, Q. D., Knapp, S., Stevenson, D. W., Stevenson, J., Blum, S. D., Boom, B. M., et al. (2012). Mapping the Biosphere: Exploring Species to Understand the Origin, Organization and Sustainability of Biodiversity. *null* 10, 1–20. doi: 10.1080/14772000.2012.665095

Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis* (New York: Springer-Verlag). Available at: https://ggplot2.tidyverse.org.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., et al. (2012). Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLos One* 7, e29715. doi: 10.1371/journal.pone.0029715

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* 3, 160018. doi: 10.1038/sdata.2016.18

Wilkinson, C., Woodruff, S. D., Brohan, P., Claesson, S., Freeman, E., Koek, F., et al. (2011). Recovery of Logbooks and International Marine Data: The RECLAIM Project. *Int. J. Climatology.* 31, 968–979. doi: 10.1002/joc.2102

WoRMS Editorial Board (2022) *World Register of Marine Species* (VLIZ). Available at: https://www.marinespecies.org (Accessed April 13, 2022).

Xiang, Z., Mungall, C., Ruttenberg, A., and He, Y. (2011). *Ontobee: A Linked Data Server and Browser for Ontology Terms. in Proceedings of the 2nd International Conference on Biomedical Ontologies (ICBO) (Buffalo, NY, USA), 279–281*. Available at: http://ceur-ws.org/Vol-833/paper48.pdf.

Zárate, M. and Buckle, C. (2021). "LOBD: Linked Data Dashboard for Marine Biodiversity," in *Cloud Computing, Big Data & Emerging Topics*. Eds. Naiouf, M., Rucci, E., Chichizola, F. and De Giusti, L. (Cham: Springer International Publishing), 151–164.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.