



EMODnet



European Marine
Observation and
Data Network

EMODnet Lot n°V – Biology

EASME/EMFF/2017/1.3.1.2/02/SI2.789013

Start date of the project: 19/04/2019 - (24 months)

EMODnet Phase III

Scientific document on the design of the workflow of text mining technologies in data archaeology



Disclaimer

The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the EASME or of the European Commission. Neither the EASME, nor the European Commission, guarantee the accuracy of the data included in this study. Neither the EASME, the European Commission nor any person acting on the EASME's or on the European Commission's behalf may be held responsible for the use which may be made of the information.

Document info

Title [ref]*	Scientific document on the design of the workflow of text mining technologies in data archaeology
WP title [ref]*	WP3- Data archaeology
Task [ref]*	
Authors [affiliation]	Georgia Sarafidou (HCMR), Savvas Paragkaman (HCMR), Vasilis Gerovasileiou (HCMR), Evangelos Pafilis (HCMR), Dimitra Mavraki (HCMR), Christina Pavloudi (HCMR), Christos Arvanitidis (HCMR / LifeWatch ERIC), Joana Beja (VLIZ), Menashè Eliezer (OGS), Marina Lipizer (OGS), Laura Boicenco (NIMRD)
Dissemination level	Public
Submission date	07/12/2020
Deliverable due date	31/10/2020

*[ref] refers to the corresponding abbreviated name of the Deliverable (or WP, or Task...), if appropriate

Contents

Vision.....	5
1 Introduction	5
1.1 Importance of historical data	5
1.2 Challenges	6
1.3 Curation Process	7
2 Comparing curation tools: A case study.....	11
2.1 Identification of publications to-be-rescued and Scanning	11
2.2 Data used	11
2.3 Manual curation	12
2.4 Automated curation workflow tools	12
2.5 Web applications	12
2.6 Standalone Applications	12
2.7 Command Line Interface (CLI)	13
2.8 Step 1. Information retrieval	13
2.8.1 Information Retrieval - Web applications	13
2.8.2 Information Retrieval - Command Line Interface	13
2.9 Step 2. Optical Character Recognition	14
2.9.1 Optical Character Recognition - Standalone Applications	14
2.9.2 Optical Character Recognition - Web Applications	14
2.9.3 Optical Character Recognition - Command Line Interface	14
2.10 Step 3. Named Entity Recognition	15
2.10.1 Named Entity Recognition - Web applications	15
2.10.2 Named Entity Recognition - Command Line Interface	18
2.11 Step 4. Entity Mapping	19
2.11.1 Entity Mapping - Web applications	19
2.11.2 Entity Mapping - Command Line Interface	19
2.12 Step 5. Data structure manipulation	20
2.12.1 Data structure - Standalone Applications	20
2.12.2 Data structure – Command Line Interface	20
2.13 Step 6. Data upload	20
3 Concluding remarks	21

5 Acknowledgements	22
6 References	23
7 Appendix	28
7.1 Abbreviations	28
7.2 Tables	29

Vision

Curation of historical data is invaluable for long-term preservation and re-use purposes, therefore methods for improving, upgrading and accelerating this process must be further developed. This is, undoubtedly, a multidisciplinary task which involves Optical Character Recognition (OCR) of scanned documents, extraction of information such as diverse ecological data on species, location, environments among others, information mapping to standardised identifiers, insertion of data in a structured format and finally the uploading of results to a web platform. Keeping up-to-date with all the relevant technologies is a challenging and laborious process, yet indispensable in curation. In this report we present an update of the different options and state of the art approaches to this end.

1 Introduction

1.1 Importance of historical data

Historical data (also known as legacy, ancient, archaeological or simply old data) comprises past-periods' information stored in an analogue and/or obsolete format. For the purposes of this study, this definition will be used in the text whenever we refer to historical data. The data that are of interest in this report refer to circa 1950 and earlier surveys, which mainly cover the field of marine biodiversity. Such type of information can be found in institutional libraries, old publications, books, expedition logbooks, project reports, newspapers (Kwok, 2017) or other types of grey literature sources. There could also be data stored in floppy disks, microfilms or scattered sheets of paper, forgotten in office drawers, thus hard to retrieve and use.

Despite their value, historical data are often considered of lesser importance compared to more recent data. However, as Griffin (2019) stated: "New science is just as likely to emerge from old data as it is from modern data". Unfortunately, as time passes by, these (g)old documents tend to fade out due to exposure to light or humidity conditions or due to careless treatment, thus resulting in the loss of irreplaceable scientific data. This loss can be catastrophic, since they are unique biodiversity snapshots of the past. Therefore, it is high time such important information is rescued and securely stored and managed.

Regardless of their age or the format in which they currently exist, historical data provide very useful information, not only to infer past biodiversity patterns and processes, but also present and future ones. The changes observed in species occurrence can also be obtained from these data. This information, if scientifically consolidated and possibly accompanied by uncertainty evaluation, can be extremely valuable to address research on community changes, species migrations, biodiversity loss; all needed for conservation policy and marine resource management (Fortibuoni et al., 2010). This is particularly important for environmental management and monitoring, as the scientific community acquires a deeper understanding of the changes that have occurred and thus a more comprehensive conservation plan can be implemented when past patterns and processes are compared with current ones.

Furthermore, historical ecological data compiled from multiple sources, such as local gazetteers, published literature and unpublished reports can provide comprehensive information on species range shifts over time and space due to anthropogenic disturbance (Faulwetter et al., 2016; Mavraki et al., 2016). The presence of a lengthy time series and large scale data is crucial for assessments, efficient modelling and prediction of future trends (McClenachan et al., 2012). The use of historical data to assess the impact of climate change and anthropogenic disturbance is of utmost importance, especially nowadays that the effects of Climate Change are at the top of the scientific agenda and a major societal demand. Long-term historical data may help to overcome the uncertainties and thus provide scientists with comprehensive information on the long-

term effects of anthropogenic stress, disentangle its results on biodiversity from those of other natural sources of disturbance and provide answers to a multitude of scientific questions. They can, for example, be used for the estimation of valuable indices, such as the CTI (Community Temperature Index), and show the vulnerability of marine communities to global warming (Stuart-Smith et al., 2015), among others.

Another recent example which takes advantage of the use of such data is the work of Rivera-Quiroz et al. (2020) who designed their expeditions based on historical publications. This group of scientists curated 55 publications from the Biodiversity Heritage Library ([BHL](#)) and extracted information from past expeditions targeting the sampling of spiders in Southeast Asia. These data assisted them to choose the best season to conduct the sampling as well as the sampling sites, which resulted in the collection of more species (both as an absolute number and as total abundance) than all the previous studies combined. Last but not least, Clavero and Revilla (2014) highlighted the importance of citizen science and old reports for global biodiversity knowledge, if a multidisciplinary approach is accomplished.

1.2 Challenges

Several factors may turn the digitisation of historical data into a serious challenge. To begin with, there is a major difficulty in locating the original data sources if they are not already digitised. Digitisation is defined as “the process of converting analogue data about physical specimens to digital representation that includes electronic text, images and other forms. This digitisation is based on diverse aims, the needs of specific projects and the specific practices and workflows in different institutions, so the digitised output has a wide range of uses.” (Haston and Hardisty, 2020).

The lack of standardisation of data among different publications or even within the same publication is considered a problematic issue that curators encounter when dealing with historical data. A common example is unstructured information, such as important data hidden in free text rather than structured in tables (Ghazzawi, 1938). In addition, there are often inconsistencies between the tables and the main publication text, since data are often repeated in a (slightly) different way, contradicting the ones already mentioned. Many publications lack basic information on metadata such as location, date or method of sampling. A common issue, for example, is that a location name or a point on an old map may be provided instead of the actual coordinates (see Steinböck, 1937). In cases like this it is quite difficult to determine the exact point of sampling. Moreover, old toponyms and political boundaries that have now changed should be also taken into consideration, as well as coordinates that now fall on land instead of the sea, due to the changes in the coastline.

Typographic errors and misspellings are also frequent. Another problem is that quite often, inaccurate descriptive information is used instead of numerical data, such as “many”, “a lot”, “some”, “few”, instead of exact abundance numbers (eg. 1, 12 etc) (as in Forbes 1843). Additionally, the use of measurement units that need to be converted to the International System of Units (SI system) or other standardised units (e.g. fathoms instead of meters). Throughout the text one can also find ambiguous symbols that could be misinterpreted, especially when no legend is provided and there is no possibility to clarify what they are with the sampling operators if they are retired or deceased. In addition to the abovementioned, taxonomic inconsistency (e.g. unaccepted synonyms, absence of the genus’s name, absence of the taxonomic literature used, absence of voucher collections from the missions) is a regular limitation for historical data which requires careful curation and the involvement of the taxa specialists. Finally, the use of languages other than English is quite common in old publications. A curator should be able to have a good understanding of the text’s language in order to correctly extract the information needed. Even in English, characters such as “æ”, derived from Latin, are often seen in scientific names, which further complicates the clear understanding of the words by humans or machines. Some of the aforementioned issues are presented in Figure 1.

Marginella clandestini	0	3	
Dentalium quinquantum	0	36	
Hyalæa cornea [gularis]	0	frag.	
— gibbosa	0	3	
— vaginellina	0	1	
Cleodora pyramidata .	0	12	
Criseis clava.....	0	many	
— spinifera	0	many	
— striata.....	0	many	
? Linacina minuta....	0	many	New.
Carinaria mediterranea	0	1	
Peracle physoides [neapolitana]	0	10	New.

Figure 1. Common problems encountered in historical data, such as old diphthongs, absence of taxon names, ambiguous symbols, shortened words and descriptive information instead of numerical (from Forbes, 1843)

Regarding all these limitations (discussed more in (Faulwetter et al., 2016)), it is not always possible to perform accurate automatic curation of data or metadata when it comes to historical data. Manual curation, a tedious and multistep process, requires the curator's full attention for the correct interpretation of valuable historical information. However, novel technologies appear to be promising for the enhancement of the curation process.

1.3 Curation Process

"Data curation is the act of discovering a data source(s) of interest, cleaning and transforming the new data, semantically integrating it with other local data sources, and deduplicating the resulting composite" (Stonebraker et al., 2013). Regarding biodiversity historical data, curation standards should be met. Data curators initially identify and prioritise the available literature sources. In continuity, they digitise the selected documents with standardised procedures and equipment (Faulwetter et al., 2016) (Fig. 2). Then, they extract data and metadata with structured knowledge identifiers in order to become computer readable and manageable. Finally, they publish these data using controlled vocabularies. Recent upgrades of several Optical Character Recognition (OCR) software and text mining software have the potential to substantially improve the curators' process when dealing with the historical data challenges, data standardisation and process automation.

Integral part of digitisation of historical data is the transformation of scanned documents (i.e images) to text through the process of Optical Character Recognition (OCR). This is a crucial step as all the subsequent steps rely on its results. Complex formats (i.e. with tables, different fonts), handwritten text and poor image quality, make this task very challenging. Especially in handwritten text and poor-quality image editing is required to obtain better results. Innovations in OCR tools are emerging since the field of computer vision is advancing the use of deep neural networks (Long et al., 2020). The survey conducted by Owen et al. (2020) provided insight on the digitisation of natural history collections and especially of herbaria by performing tests on several state-of-the-art tools. They focused on OCR of handwritten text.

Indispensable to the curation process have been the standards of the International Commission on Zoological Nomenclature (ICZN) for zoological nomenclature, the World Register of Marine Species

([WoRMS](#)) for marine species taxonomy and the Environmental Ontology ([ENVO](#)) for environments among others. Unification of standards in tools for the search, classification and comparison for species names is facilitated by the Global Names Architecture (GNA) and for other entities is still ongoing. These tools and the structured knowledge they provide have enabled text mining technologies to be implemented in biodiversity texts and thus legacy literature. Text mining as defined by Hearst (1999) is “the automatic discovery of new, previously unknown, information from unstructured data”. This is often seen as “comprising three major tasks: information retrieval (IR, gathering relevant documents), information extraction (IE, extracting information of interest from these documents) and data mining (discovering new associations among the extracted pieces of information)” (Ananiadou and Mcnaught, 2006). Named Entity Recognition (NER) is a key step in such a process for locating terms of interest in text (e.g taxa, traits of organisms and environment types). There are text mining systems that rely on background knowledge for the identification and normalisation of entities through domain specific dictionaries. Conversely, others use solely statistical learning based on training data and can be applied in multiple domains (Perera et al., 2020).

Controlled vocabularies like the ones in [DarwinCore](#), identifiers and semantics are used to accomplish the interoperability for biodiversity data; taxonomy and nomenclature, environment type, organism traits and geolocation. In the same context, common units and data formats are being used. This best practice is a labour-intensive task, nevertheless crucial for data rescue and their integration with modern data, since by implementing all these steps, the indispensable value of historical data in biodiversity will be revealed. It is expected, of course, that more curation steps should follow, especially towards the “FAIR-ification” of the (meta)data (Findable, Accessible, Interoperable, Reusable; Reiser et al., 2018; Wilkinson et al., 2016) before the final submission to a repository. These steps are beyond the scope of the current document.

An ongoing effort towards the digitisation and publication of biodiversity historical data is being carried out within the EMODnet Biology Work Package 3 (WP3). More specifically, WP3 aims to fill the temporal and spatial gaps, in biodiversity knowledge, through the rescue of historical data and make them available through the [EMODnet portal](#). This is achieved by the implementation of long-term strategies, such as the continuous identification of historical data at risk and, subsequently, their harvest by EMODnet Biology. Faulwetter et al. (2016) provided a workflow (Fig. 2) that depicts the curation of the manual extraction of biodiversity data from legacy literature up to their final storage to [MedOBIS data repository](#), the Mediterranean node of the Ocean Biodiversity Information System (OBIS) and subsequent integration in [EMODnet Biology](#), [OBIS](#) and the Global Biodiversity Information Facility ([GBIF](#)) network.

This report focuses on the comparison of different tools and interfaces in order to automate and assist the curation process. Specifically, tools in terms of OCR and text mining technologies are tested and reviewed with the aim to design a workflow that can accommodate the need for automation and acceleration in digitising historical datasets and extracting their information. It is considered as an upgraded version of the previous one (Fig. 2).

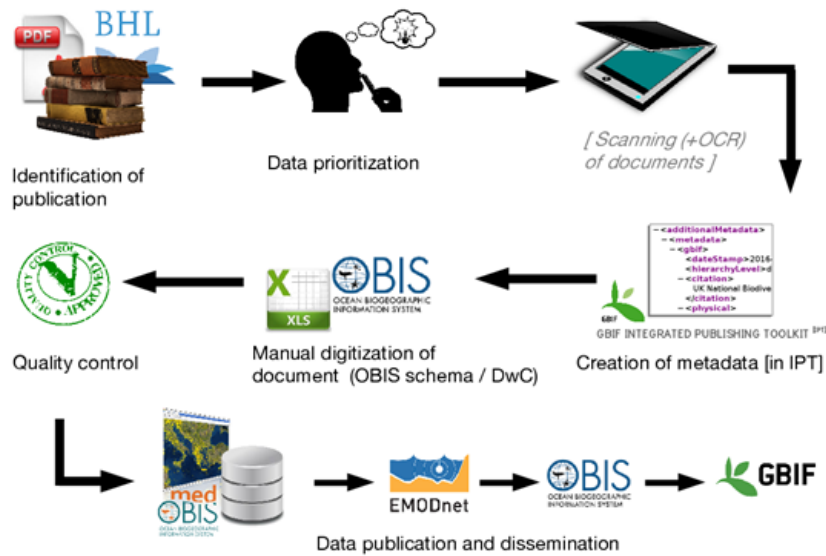


Figure 2. The original workflow (Faulwetter et al., 2016), depicting the process of manually extracting data from legacy literature, which is currently performed in EMODnet WP3

In particular the third and fourth steps are further subdivided, so that data curators have more detailed guidelines for the proper digitisation and processing of data. Two types of curation workflows are described (Fig. 3); one that relies on web applications and the other that combines programming libraries and packages. The latter is scalable, customisable and replicable but requires programming skills whereas the former is easy to implement through Graphical User Interfaces (GUI) at the expense of the previous advantages.

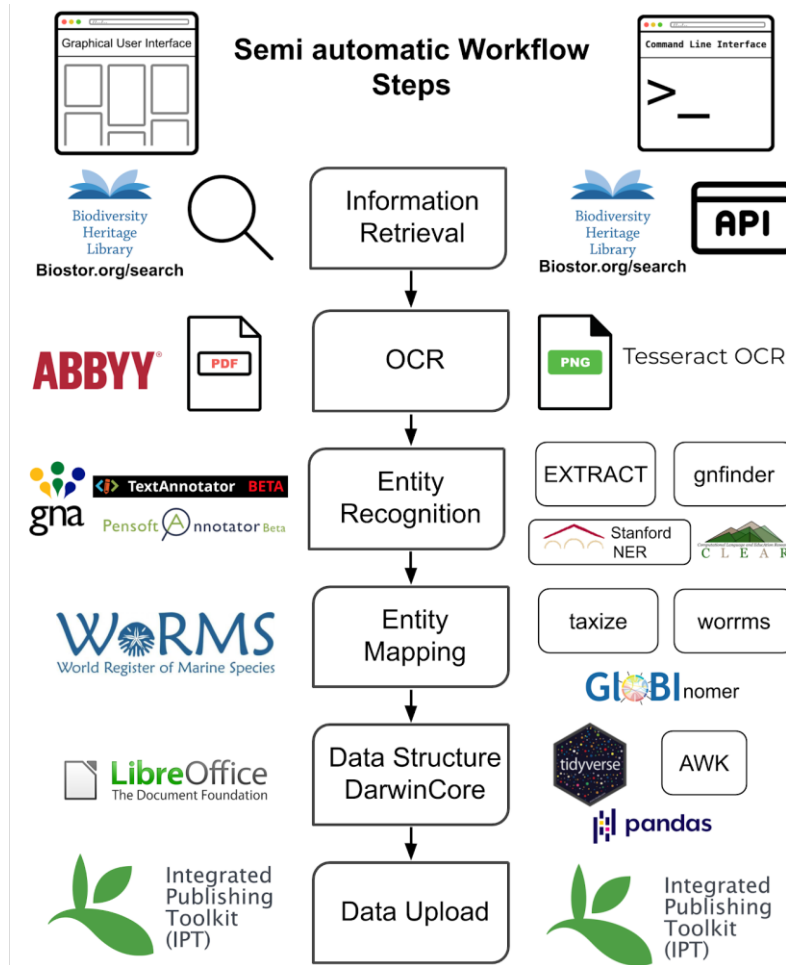


Figure 3. Workflow depicting the automation of the curation process of historical data

2 Comparing curation tools: A case study

In this section are presented the main up to date platforms, web services and applications that are used for the extraction of biodiversity data. In Table 1 (Appendix) the subsequent tools are listed accompanied with features such as extracted information, input format and their interface.

2.1 Identification of publications to-be-rescued and Scanning

The identification of the hard copy of the publication up to the scanning step requires human input and manual handling of the scanner. Not much has progressed in this area since the previous report by Faulwetter et al. (2016). The identification of the digitised literature can be facilitated from the [BHL](#), the [Belgian Marine Bibliography](#) (BMB) and other initiatives (Kearney, 2019). This step is inherently human curated and shared across all workflows.

2.2 Data used

In order to test the curation tools in terms of the digitisation process, the article “Report on the Mollusca and Radiata of the Aegean Sea: and on their Distribution, Considered as Bearing on Geology” by Forbes (1843) was used (Fig. 4). Taxon names and abundance of living and dead individuals were extracted from the Appendix 1 which is six pages long (p. 180-185) and includes data structured in table format (available on this [link](#)). Metadata (i.e date, locality, depth, distance from shore and substrate) were extracted from the text of the article. Data and metadata were manually curated and imported in a spreadsheet in November 2020. The spreadsheet with the taxon occurrences was used for the case of web applications comparison.

The aforementioned article is part of the Thirteenth Meeting of the British Association for the Advancement of Science which was scanned at 400 DPI resolution and uploaded to Biodiversity Heritage Library on 2009-04-22 by the [Internet Archive](#). The whole volume was processed for OCR with ABBYY FineReader 8.0. The high quality scanning, OCR processing and the manual curation constitute these data appropriate for the comparison of the available tools for automated curation.

I.				II.			
Date	May 29th, 1841.			Date	September 12, 1842.		
Locality	Nousa Bay, Paros.			Locality	{ Gulf of Smyrna, off Mouth of Hermus.		
Depth	Five to six fathoms.			Depth	Seven fathoms.		
Distance from shore	(Within the Bay.)			Distance from shore	Half a mile.		
Ground	Mud and sandy mud.			Ground	Dark mud.		
Region	II.			Region	II.		
Species.	No. of li- ving spec.	No. of dead spec.	Observations.	Species.	No. of li- ving spec.	No. of dead spec.	Observations.
Pinna squamosa	0	1	In sandy mud.	Pecten sulcatus	0	4'	Fullgrown valves.
Modiola tulipa	1	0		— varius	0	3'	Full size.
Pecten polymorphus.	4	6'	In dark mud.	Modiola barbata	2	2'	Hitherto a fossil.
— hyalinus	1	0		Solen tenuis	1	3'	
Nucula margaritacea.	0	40'		— coarctatus	0	8'	
Cytherea chione	0	1'		Thracia pubescens ...	2	3'	
— venetiana	1	3-5'		Ligula boysii	0	2'	
— apicalis	1	2-12'		Pandora obtusa	0	4'	
Artemis lineta	0	1'	Tellina donacina	5	10'		
Pullastra virginea ...	0	5'	Corbula nucleus	12	8-50'		
Venus verrucosa	0	5'	Cardium echinatum...	0	3'	Young speci- mens.	
Tellina donacina ...	0	1-3'	Artemis lineta	0	10'		
— balaustina	0	2'	Montacuta, sp.....	0	1	Much worn.	
Ligula boysii	0	2-10'	Nucula margaritacea..	5	30		

Figure 4. A screenshot of the dataset used showing the structure of the data and metadata provided (Forbes 1843)

2.3 Manual curation

In general, once the literature is identified, prioritised and scanned, the curator is responsible for extracting the necessary information from the files and organising it into a common format. For this reason, data curators read the document and digitise the data into IPT source files, mapping them to DarwinCore terms, adding metadata and creating a standard DarwinCore Archive. British Oceanographic Data Centre (BODC) vocabularies are being used for mapping the facts as required by EMODnet Biology. This whole process is mostly done manually, which means reading the information (e.g. the occurrence of a specific species) and inputting it through typing to the corresponding cell of the data file. It is, as expected, a time- and resource-consuming procedure. Specifically, as far as the aforementioned dataset is concerned, a rough estimation of the effort required from the digitisation step to the IPT upload step was a two weeks' (8 hours/day - 5 days/week) labour by one person.

2.4 Automated curation workflow tools

Automatic workflows assist curators with bulk text annotations in terms of species, environments and traits. To implement these workflows the following steps are required: information retrieval, OCR, NER, entity mapping and data structure manipulation. Multiple page documents can be searched for species mentions in seconds, with technologies that find synonyms and fuzzy search for the OCR transformation misspelling. The transformation of these results to database IDs, like LSID, Encyclopedia of Life (EOL) IDs among others, is easily facilitated through web services and programming software. The interconnection and guidance of these steps still requires manual input especially when using multiple web services.

2.5 Web applications

Web applications are used as tools for text mining purposes. These tools are promising because of their accessibility and their easy to use interface. By simply uploading documents, and after them having been processed in a server, the results are delivered back to the user (Lamurias, 2019). During the last years, an upsurge in web applications development regarding the enhancement of biodiversity data digitisation has been observed, indicating the need for such initiatives. The web applications mentioned in this report were tested in November 2020 in two web browsers, Mozilla Firefox version 83 and Google Chrome version 87 on Microsoft Windows 10.

2.6 Standalone Applications

The main all-in-one tool with a Graphical User Interface (GUI) is [Golden-Gate-imagine](#). This tool supports most of the steps of the curator's workflow by providing annotations on PDF backed up by ontologies. It was developed by [Plazi](#) in 2015 but it has not received any update since 2016. Taking this into account, however, several recent biodiversity data related publications have used it (Agosti et al., 2020; Miller et al., 2019; Rivera-Quiroz and Miller, 2019).

One-stop-shop purpose software applications for domain specific usage, like [GoldenGate](#), are very helpful but require dedicated developers, if they are to stay updated and relevant. This is generally the case for software with Graphical User Interface (GUI) because Operating Systems (OS) are constantly being updated thus making applications obsolete and unsupported only in a few years. This is the main reason to move to programming tools with a generic scope when applicable i.e. OCR.

2.7 Command Line Interface (CLI)

The CLI provides a simple way to connect different programming packages and libraries of any programming language in UNIX (Linux and Mac operating systems (OS)) and Windows OS. This is a powerful way to implement scalable, replicable and reproducible workflows: scalable because the same code can be applied in multiple files (in our case documents), reproducible and replicable because it can be executed multiple times and to different types of documents respectively. Even though programming packages and libraries are fast, scalable and easier to maintain, they require familiarity and expertise in CLI and programming which on the other hand, takes effort and time. Curators in the field of Biodiversity increasingly acquire such skills, especially those regarding the [R programming language](#).

The proposed CLI workflow includes open-source software based on CLI tools, APIs and programming packages. The tools we choose meet the subsequent criteria:

1. Open-source
2. To be in active development
3. Easy to scale to many documents
4. Combination of different tools in a few steps
5. Reproducibility, replicability and scalability

The tools used are distributed across the major platforms; Linux, Mac and Window. The presented code of each step is indicative of the commands needed to run the basic functionality of each tool. This code was tested on a Macbook Pro with 8gb Ram and Intel Core i5-4258U CPU at 2.40GHz. Complementary tools (i.e. [Ghostsript](#) and `cat`) and [Bash](#) commands are also used to provide a more complete view of the workflow.

2.8 Step 1. Information retrieval

2.8.1 Information Retrieval - Web applications

The Biodiversity Heritage Library ([BHL](#)) is a powerful tool that enhances the discovery of legacy literature, since it is an open access and user friendly web page, containing hundreds of thousands volumes from the 15th century until the present. The web resource [BioStor](#) facilitates article search and retrieve operations from the [BHL](#). It considers article-related metadata as well as OCR text in searches which makes the [BHL](#) articles more findable (Page, 2011). This tool enables the extraction of locality information from taxonomic papers through coordinated searching. This information is, however, not linked to the original location nor to the specimens from the source text, making it impossible to actually use the data as occurrence records (Page, 2019). The web service [BioNames](#) is a database that contains the names of species accompanied with their original publication. In some cases a phylogenetic tree is provided as well (Page, 2013) and there is also a direct connection from species to their Life Sciences Identifier (LSID).

2.8.2 Information Retrieval - Command Line Interface

To create a *corpus* it is possible to use the BioStor API and the [BHL](#) API. Both are based on REpresentational State Transfer (REST) API accepting HTTP queries in GET and POST form. In addition, [BHL](#) has developed an R package called `rbhl` that provides access through R functions.

In case that the PDF data are already processed for OCR from a trusted tool there is a way to extract the raw text from PDF using [Ghostsript](#) which is available in all platforms. Here we used the version 9.53.3. The following command will extract the text from the PDF file called `legacy-publication.pdf` and save it to a text file named `legacy-publication-all.txt`.

```
gs -sDEVICE=txtwrite -o legacy-publication-all.txt legacy-publication.pdf
```

If this is the case the next step can be omitted for the CLI workflow.

2.9 Step 2. Optical Character Recognition

There are plenty of OCR applications that vary in price, platform (OS, web application, CLI) and features but the underlying OCR engines mostly remain the same¹. Pre-processing of the files, i.e increase in contrast, noise reduction and conversion to black and white, before performing OCR has been suggested in some cases² such as handwritten text and/or low-quality scanning.

2.9.1 Optical Character Recognition - Standalone Applications

The most common OCR application is [Adobe Acrobat PRO](#) which is proprietary. The [Microsoft OneNote](#) application is free but has limited OCR functionality. The most advanced OCR platform is [ABBYY FineReader](#) which has best in class performance but in high price³.

Microsoft OneNote	ABBYY FineReader engine	OmniPage	Adobe Acrobat PRO
-----------------------------------	---	--------------------------	-----------------------------------

2.9.2 Optical Character Recognition - Web Applications

Cloud computing is used for OCR by big technological companies like Google, Amazon and Microsoft. They use their servers for the computationally expensive tasks of Deep Neural Networks learning algorithms. Thus, providing OCR as a web service through APIs or web applications. Recent exhaustive testing and evaluation of these tools indicate that ABBYY is one of the best OCR tools along with Google Vision for scanned documents⁴.

Google Vision	Amazon Cloud	Microsoft Azure Computer Vision	ABBYY Cloud Command line
-------------------------------	------------------------------	---	--

The [BHL](#) uses OCR to process all the page images in their collections so that the text contained within the images can be indexed and made searchable. Currently, the tool [ABBYY FineReader](#) version 11.0 is used through the web service [Internet Archive](#) for this purpose. It is underlined that the text is generated from automated OCR, without manual testing. Both the pdf and the text are provided.

2.9.3 Optical Character Recognition - Command Line Interface

The most common OCR engines with CLI are [Tesseract](#) and [OCRpus](#). The latter also contains image editing capabilities whereas the former just the OCR engine. In addition, [OCRpus](#) has to be trained in order to be used for OCR. [Kraken](#) and [Calamari](#) are based on [OCRpus](#) and have trained their models in order to be ready to use.

Tesseract	Calamari	Kraken	OCRopus
---------------------------	--------------------------	------------------------	-------------------------

¹ <https://source.opennews.org/articles/so-many-ocr-options/>

² <https://www.ocrsdk.com/documentation/hints-tips/image-recommendations/>

³ <https://medium.com/dida-machine-learning/comparison-of-ocr-tools-how-to-choose-the-best-tool-for-your-project-bd21fb9dce6b>

⁴ <https://source.opennews.org/articles/so-many-ocr-options/>

Tesseract is praised for its open sourced nature, easy to use interface and scanned text document performance^{5 6}. In addition, it is the main engine of many GUIs and Projects⁷. We chose the Tesseract tool version 4.1.1 for CLI workflow.

In order to use Tesseract an additional step is required; the transformation of the PDF files in single image per page format. This is possible with the command line tool [ImageMagick](#), available in all major platforms. A multipage PDF document can be easily exported as multiple single page images in the desirable format. In this particular example we chose PNG.

```
convert -density 400 legacy-publication.pdf -quality 100 folder/legacy-publication.png
```

The option `-density 400` is the dpi of the scanned document and the option `-quality 100` is to ensure the looseness transformation. As mentioned before, there are some cases that more advanced editing is required to obtain better OCR results. The document is now ready for OCR with Tesseract. We applied the `tesseract` command on all pages - PNG files - at once.

```
cd folder  
for f in *.png; do tesseract -l eng $f ${f%*.png}.txt; done
```

The `-l` flag stands for language which, in this example, was English. Tesseract identifies language as well as the DPIs (Dots Per Inch) of the document. The result of the above command was the creation of multiple text files (.txt) with OCR text.

We combined all these files into one document by:

```
cat *.txt > legacy-publication-all.txt
```

The file `legacy-publication-all.txt` is a complete OCR version of the input file `legacy-publication.pdf`.

2.10 Step 3. Named Entity Recognition

The majority of the text mining tools in this report are restricted to the Named Entity Recognition (NER) and information extraction of the species names within a document. It is indeed a very crucial step in the text mining field towards the biodiversity information extraction. Recognition of species scientific names has been advanced with major contributions from Global Names Architecture ([GNA](#)) tools. Nevertheless, there is a great need for the development of information extraction of entities such as abundance, biomass and organisms traits, sampling coordinates and sampling techniques.

2.10.1 Named Entity Recognition - Web applications

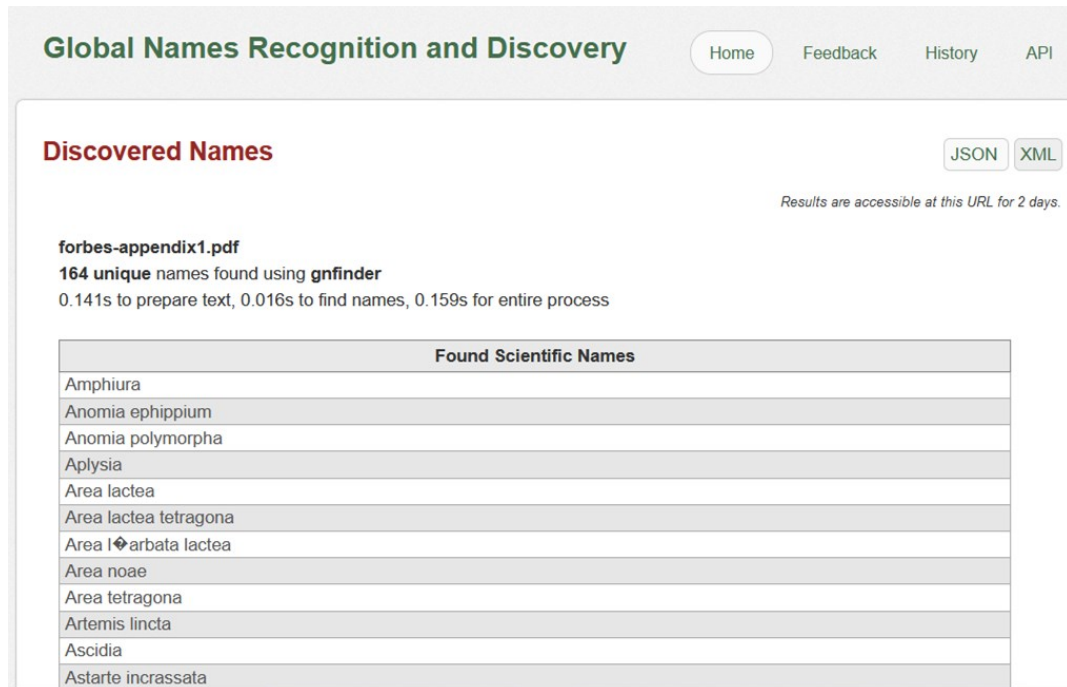
The [Global Names Recognition and Discovery \(GNRD\)](#) tool, within Global Names Architecture ([GNA](#)) is a web application regarding the recognition of scientific names. It can use files such as PDF, images or Microsoft Office documents and one can still input URLs or even freeform text for their search. The page performs OCR using the tool [Tesseract](#) and afterwards uses the [gnfinder](#) discovery engine in order to provide the names list. It offers an API, and can be installed locally based on an updated [Github resource](#). The [Global Names Architecture](#) is also used by the [BHL](#) platform in order to locate taxonomic names within the pages

⁵ <https://medium.com/dida-machine-learning/comparison-of-ocr-tools-how-to-choose-the-best-tool-for-your-project-bd21fb9dce6b>

⁶ <https://pdf.iskysoft.com/ocr-pdf/open-source-ocr.html>

⁷ <https://tesseract-ocr.github.io/tessdoc/User-Projects-%E2%80%93-3rdParty.html>

of the collections (Richard, 2020). The test performed (Fig. 5) on the six-page PDF template provided 128 unique scientific names in species level, out of the 240 identified through the manual curation.



Global Names Recognition and Discovery Home Feedback History API

Discovered Names JSON XML

Results are accessible at this URL for 2 days.

forbes-appendix1.pdf
164 unique names found using **gnfinder**
 0.141s to prepare text, 0.016s to find names, 0.159s for entire process

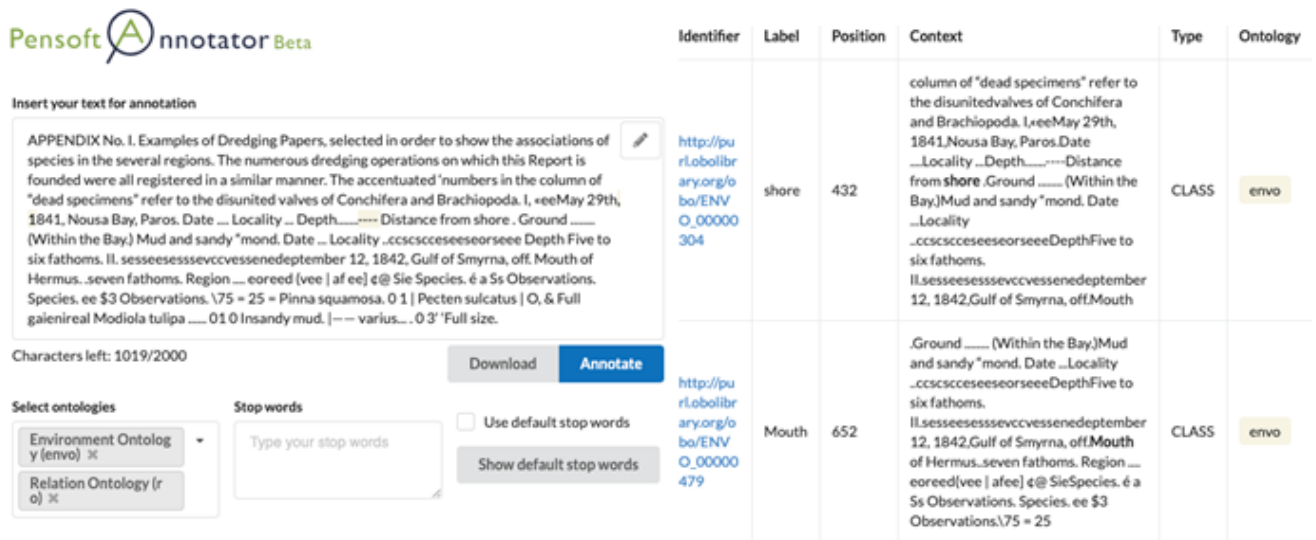
Found Scientific Names
Amphiura
Anomia ephippium
Anomia polymorpha
Aplysia
Area lactea
Area lactea tetragona
Area l ^o arbata lactea
Area noae
Area tetragona
Artemis lincta
Ascidia
Astarte incrassata

Figure 5. Screenshot of the web application GNRD performing NER

The [Biodiversity Observation Miner](#) is a web application based on [R shiny](#), also available on [GitHub](#), that allows 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 also uses the [gnfinder](#) discovery engine through the [R package taxize](#). The web application is also in Beta version and an OCR processed PDF file is used as input. The novelty of this tool is the provision of text snippets (Fig. 6) but limited functionality was occasionally observed, discussed in this [thread](#). In addition to snippets, this application provides the co-occurrence of words accompanied with their count. This way curators can be informed for terms that occur together in the document. This is an important step towards mining relationships between terms of interest in biodiversity data.

Figure 7. Screenshot of the web application [TextAnnotator](#) performing NER

The [Pensoft Annotator](#) is another Beta web application that works with ontologies (Fig. 8). Relation Ontology (RO) and ENVO are built in the Annotator but it is extendable to any ontology, though programming skills are required. It can annotate up to 2000 characters with the previous ontologies. This limitation is noted that it can be expandable upon communication.



Pensoft Annotator Beta

Insert your text for annotation

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, eeMay 29th, 1841, Nousa Bay, Paros. Date ... Locality ... Depth..... Distance from shore . Ground ____ (Within the Bay.) Mud and sandy"mond. Date ... Locality ..ccscscceeseorseee Depth Five to six fathoms. Il.sesseessesveccvssenedepteember 12, 1842, Gulf of Smyrna, off. Mouth of Hermus. seven fathoms. Region ... eoreed (vee | af ee] t@ Sie Species. é a Ss Observations. Species. ee \$3 Observations. \75 = 25 = Pinna squamosa. 0 1 | Pecten sulcatus | O, & Full gaieinreal Modiola tulipa ____ 01 0 Insandy mud. |--- varius... 0 3" Full size.

Characters left: 1019/2000

Select ontologies: Environment Ontology (envo) x, Relation Ontology (ro) x

Stop words: Type your stop words

Use default stop words: Show default stop words:

Identifier	Label	Position	Context	Type	Ontology
http://purl.org/bo/ENV/0.00000304	shore	432	column of "dead specimens" refer to the disunitedvalves of Conchifera and Brachiopoda. I, eeMay 29th, 1841, Nousa Bay, Paros. Date ... Locality ... Depth..... Distance from shore . Ground ____ (Within the Bay.)Mud and sandy"mond. Date ... Locality ..ccscscceeseorseeeDepthFive to six fathoms. Il.sesseessesveccvssenedepteember 12, 1842,Gulf of Smyrna, off.Mouth	CLASS	envo
http://purl.org/bo/ENV/0.00000479	Mouth	652	.Ground ____ (Within the Bay.)Mud and sandy"mond. Date ... Locality ..ccscscceeseorseeeDepthFive to six fathoms. Il.sesseessesveccvssenedepteember 12, 1842,Gulf of Smyrna, off.Mouth of Hermus. seven fathoms. Region ... eoreed(vee afee] t@ SieSpecies. é a Ss Observations. Species. ee \$3 Observations.\75 = 25	CLASS	envo

Figure 8. Screenshot of the web application [Pensoft Annotator](#) performing NER

The [Taxonfinder](#) is a web application regarding the extraction of scientific names. Nevertheless, the Github version is quite old (2014) and it works only with HTML URLs and not with PDF or other format of text. It features an API that was used in BHL for large scale annotations of taxonomic names until 2019 that was replaced by [gnfinder](#). Therefore it doesn't add any value to curators.

The [Ontobee](#), a web server that links ontologies, is useful for the annotation of text to ontology ids. For biodiversity data is most useful for environmental features (ENVO IDs) and functional traits ([PATO](#) IDs or other organism specific ontology). Text snippets must be used to function smoothly and not whole documents.

2.10.2 Named Entity Recognition - Command Line Interface

We applied multiple NER tools to retrieve information from biodiversity literature containing species names, traits, environments, location and geolocation data.

Some effort has been spent on extracting species names from text. The most notable is the [Global Names Architecture parser](#) (Pyle, 2016) which provides fuzzy search and is the underlying engine of most text mining tools about biodiversity. We found that it is in constant development, deeming it a reliable tool for this work. EMODnet Biology uses WoRMS, and it is included in [gnfinder](#), since it is based on [index.globalnames.org](#).

```
gnfinder find legacy-publication-all.txt > legacy-publication-all-gnfinder.json
```

The command line tool returns a .json (JavaScript Object Notation) file that has two arrays; metadata and names. Metadata are the language, date of the execution of the command and total words. The data have one entry per identified string which contains the matched string, the returned name and the positional boundaries in character sequence.

To extract the names we have used the command line tool for json manipulation called `jq` (<https://stedolan.github.io/jq/>). The following one-liner loads the file into `jq` and then selects the names of species only, removes the "" from the names and finally saves it in a single column `.tsv` file.

```
morelegacy-publication-all-gnfinder.json | jq '.names[] | {name: .name} | [.name] | @tsv' | sed 's/"//g' > legacy-publication-all-gnfinder.tsv
```

In order to simultaneously extract organisms, environments and tissues we have used the tool called [EXTRACT](#) (Pafilis *et al.*, 2017; Jensen, 2016). It implements the JensenLab tagger API (Jensen, 2016) with advanced dictionaries [SPECIES-ORGANISMS](#), [ENVIRONMENTS](#) (Pafilis *et al.*, 2015) and [TISSUES](#). It returns [NCBI](#) Taxonomy IDs, [ENVO](#) (Environment Ontology) terms and [BRENDA](#) IDs respectively.

```
./getEntities_EXTRACT_api.pl legacy-publication-all.txt > legacy-publication-all-extract.tsv
```

The result is a `tsv` file with 3 columns: tagged text, entity type and term id. We found it very fast, accurate and easy to use. The script `getEntities_EXTRACT_api.pl` is written in perl and is simple to comprehend.

Another important NER system is the Stanford NER (Finkel *et al.*, 2005) which recognises locations, persons and organisations in text. It has a generic scope but it can help in biodiversity data. The general tokenisation and normalisation procedures developed by the NLP Stanford team are the basis of many text mining tools.

Also, the ClearTK NLP toolkit (Bethard *et al.*, 2014) within the ClearEarth project (Thessen *et al.*, 2018) can be added as well which can tag from text entities like biotic, abiotic, locality, quality, unit and value. Upon installation it downloads multiple dictionaries and takes up to six gigabytes of space. It relies on Stanford NLP and other dependencies. Since it provides python wrapper and a command line interface it is possible to include it in the workflow.

2.11 Step 4. Entity Mapping

Species and higher taxonomies have multiple IDs depending on the platform. Mapping the information retrieved to different IDs is crucial for the cross-platform communications but can be tricky because the mapping service must be up to date. Some of the common IDs for taxonomy, apart from the Linnaean system, are the LSID, NCBI ID, EOL ID etc.

2.11.1 Entity Mapping - Web applications

[WoRMS Taxon match](#) that matches the species list found with the World Register of Marine Species ([WoRMS](#)) accepted scientific names and species LSID. Geographic regions are confirmed with the use of the georeference tool developed for the [Marine Gazetteer](#).

2.11.2 Entity Mapping - Command Line Interface

All these platforms provide APIs and the R package `Taxize` (Chamberlain and Szöcs, 2013) provides mapping capabilities across these and many more.

Simple functions like the one below can perform mapping easily across rows of the provided table.

```
get_eolid  
get_nbnid  
get_wormsid
```

In addition, the [GLOBI](#) (Global Biotic Interactions) nomer tool (<https://github.com/globalbioticinteractions/nomer>) provides mapping functionality in command line and python.

```
more species-list.tsv | nomer append > species-list-nomer.tsv
```

Especially for WoRMS there is an API, used by the [R worrms package](#) (Chamberlain, 2018) which provides the ability to match scientific names to Aphia IDs. This package is still supported

2.12 Step 5. Data structure manipulation

2.12.1 Data structure - Standalone Applications

In terms of data structure the applications Microsoft Excel or similar, such as LibreOffice and Apache OpenOffice are used. The data are organised based on the [Darwin Core](#) standard, where a Darwin Core Archive is created (see guidelines in this [link](#)).

2.12.2 Data structure – Command Line Interface

All the previous steps involve text handling and table manipulations, JSON files (depending on the tool), merging and filtering. These tasks are usually performed in R using the [tidyverse](#) package suite, in Python using the [pandas](#) library and in [AWK programming language](#). The choice of tools depends on the user's familiarity, expertise and operating system. We suggest using R because there are R packages available for most steps for a beginner who wants to start implementing the aforementioned procedures, and it's supported by active developers. Data source files or database tables based on Darwin Core terms are prepared. These files include also detailed sampling descriptors terms based on standard controlled vocabularies. Automation could be used for this preparation.

2.13 Step 6. Data upload

The [Integrated Publishing Toolkit \(IPT\)](#) tool is a free open source software written in Java which is used to publish and share biodiversity datasets through [GBIF](#). It bundles data and metadata in a Darwin Core Archive (DwC-A). The [IPT](#) can also be configured with a DataCite account in order to assign DOIs to datasets transforming it into a data repository. Data are made available using the DarwinCore standard and the extensions, such as [OBIS-ENV](#), that are in use by [GBIF](#) and [OBIS](#). Quality control of the published data is performed and the data flows to MedOBIS, then to the EMODnet portal and afterwards to [OBIS](#) and [GBIF](#). When the data curation is completed the data can be uploaded to the IPT.

3 Concluding remarks

Historical biodiversity data are of paramount importance and their rescue is a priority for EMODnet Biology's WP3. There are a large number of limitations during the digitisation process that further complicate a curator's work. Tools in the field of OCR and text mining promise semi-automation and acceleration of the process. Equally important to the development of these tools is the ability to sustain support and continue debugging both of which are the bottleneck of tool usability.

Generally, it is worth mentioning that curation assistance with specific tailored tools has been recognised by the biodiversity community. For the past 3 years many new tools have been developed that introduced innovative features that are promising. Albeit, active development and contribution to reporting issues of open-source repositories such as Github is lacking for many tools. The common patterns observed are active web page tools of text mining that accept PDF or text directly as input, are all still in beta version; they are not up to date and they often display error messages or long loading times. Also, the majority of the current tools provide information related to taxon names recognition. Despite that, the tools offer assistance and time saving, which is crucial for the curators. Either way, Command Line Interface provided a faster and larger scale data processing.

Therefore, it is highly recommended that curators are trained in basic programming skills which would benefit them and, in the long term, the historical data rescue process in general. The CLI code suggested here could be applied to any type of data, not only historical, and thus contribute to the digitisation of biodiversity knowledge overall. Finally, one of the promising future steps towards this direction would be the implementation of community curation procedures, where citizens and/or scientists are involved voluntarily in the process of data digitisation.

5 Acknowledgements

We are grateful to Leen Vandepitte (WP2) for their useful comments and the overall contribution to the fulfilment of this report.

6 References

- Agosti D, Guidoti M, Catapano T, Ioannidis-Pantopikou A, Sautter G (2020) The Standards behind the Scenes: Explaining data from the Plazi workflow. *Biodiversity Information Science and Standards* 4: e59178. <https://doi.org/10.3897/biss.4.59178>
- Ananiadou S, Mcnaught J (2005) *Text mining for biology and biomedicine*. Artech House, Inc., USA.
- Batista-Navarro R, Nguyen NTH, Soto AJ, Ulate W, Ananiadou S (2017) Argo as a platform for integrating distinct biodiversity analytics tools into workflows for building graph databases. *Biodiversity Information Science and Standards* 1: e20067. <https://doi.org/10.3897/tdwgproceedings.1.20067>
- Bethard S, Ogren P, Becker L (2014) ClearTK 2.0: Design Patterns for Machine Learning in UIMA. *LREC ... International Conference on Language Resources & Evaluation: [proceedings]*. International Conference on Language Resources and Evaluation 2014: 3289–3293.
- Chamberlain S (2020) *worrms: World Register of Marine Species (WoRMS) Client*. . manual Available from: <https://cran.r-project.org/package=worrms>.
- Chamberlain SA, Szöcs E (2013) *taxize: taxonomic search and retrieval in R*. *F1000Research* 2. <https://doi.org/10.12688/f1000research.2-191.v2>
- Clavero M, Revilla E (2014) Mine centuries-old citizen science. *Nature* 510: 35–35. <https://doi.org/10.1038/510035c>
- Driller C, Koch M, Abrami G, Hemati W, Lücking A, Mehler A, Pachzelt A, Kasperek G (2020) Fast and Easy Access to Central European Biodiversity Data with BIOfid. *Biodiversity Information Science and Standards* 4: e59157. <https://doi.org/10.3897/biss.4.59157>
- Driller C, Koch M, Schmidt M, Weiland C, Hörschemeyer T, Hickler T, Abrami G, Ahmed S, Gleim R, Hemati W, Uslu T, Mehler A, Pachzelt A, Rexhepi J, Risse T, Schuster J, Kasperek G, Hausinger A (2018) Workflow and Current Achievements of BIOfid, an Information Service Mobilizing Biodiversity Data from Literature Sources. *Biodiversity Information Science and Standards* 2: e25876. <https://doi.org/10.3897/biss.2.25876>
- Faulwetter S, Pafilis E, Fanini L, Bailly N, Agosti D, Arvanitidis C, Boicenco L, Capatano T, Claus S, Dekeyzer S, Georgiev T, Legaki A, Mavraki D, Oulas A, Papastefanou G, Penev L, Sautter G, Schigel D, Senderov V, Teaca A, Tsompanou M (2016) EMODnet Workshop on mechanisms and guidelines to mobilise historical data into biogeographic databases. *Research Ideas and Outcomes* 2: e9774. <https://doi.org/10.3897/rio.2.e9774>

- Finkel JR, Grenager T, 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). Association for Computational Linguistics, Ann Arbor, Michigan, 363–370. <https://doi.org/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. Reports of the British Association for the Advancement of Science for 1843: 130–193.
- Fortibuoni T, Libralato S, Raicevich S, Giovanardi O, Solidoro C (2010) Coding Early Naturalists' Accounts into Long-Term Fish Community Changes in the Adriatic Sea (1800–2000). PLOS ONE 5: e15502. <https://doi.org/10.1371/journal.pone.0015502>
- Ghazzawi FM (1938) Two Cladocera from the plankton: Plankton of the Egyptian Waters. In: Hydrobiology and Fisheries. Government Press, Cairo. Available from: https://www.lifewatchgreece.eu/sites/default/files//article_files/31.%20Two%20Cladocera%20from%20the%20Plankton.pdf.
- Griffin E (2019) Getting necessary historical data out of deep freeze. Polar Science 21: 238–239. <https://doi.org/10.1016/j.polar.2019.05.008>
- Haston E, Hardisty A (2020) An Introduction to the Minimum Information about a Digital Specimen (MIDS) Digitisation Standard. Biodiversity Information Science and Standards 4: e59214. <https://doi.org/10.3897/biss.4.59214>
- Hearst MA (1999) Untangling text data mining. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics -. Association for Computational Linguistics, College Park, Maryland, 3–10. <https://doi.org/10.3115/1034678.1034679>
- Jensen LJ (2016) One tagger, many uses: Illustrating the power of ontologies in dictionary-based named entity recognition. bioRxiv: 067132. <https://doi.org/10.1101/067132>
- Kearney N (2019) It's Not Always FAIR: Choosing the Best Platform for Your Biodiversity Heritage Literature. Biodiversity Information Science and Standards 3: e35493. <https://doi.org/10.3897/biss.3.35493>
- Kwok R (2017) Historical data: Hidden in the past. Nature 549: 419–421. <https://doi.org/10.1038/nj7672-419>
- Lamurias A, Couto FM (2019) Text Mining for Bioinformatics Using Biomedical Literature. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C (Eds), Encyclopedia of Bioinformatics and Computational Biology. Academic Press, Oxford, 602–611. <https://doi.org/10.1016/B978-0-12-809633-8.20409-3>

- Long S, He X, Yao C (2020) Scene Text Detection and Recognition: The Deep Learning Era. *International Journal of Computer Vision*. <https://doi.org/10.1007/s11263-020-01369-0>
- Mavraki D, Nikolopoulou S (2018) Digitization of Plankton of the Egyptian waters. Two Cladocera from the plankton. *Notes and Memories of the Hydrobiology and Fisheries Directorate of Egypt, Notes and Memories No 31*. <https://doi.org/10.15468/byglfb>
- Mavraki D, Fanini L, Tsompanou M, Gerovasileiou V, Nikolopoulou S, Chatzinikolaou E, Plaitis W, Faulwetter S (2016) Rescuing biogeographic legacy data: The “Thor” Expedition, a historical oceanographic expedition to the Mediterranean Sea. *Biodiversity Data Journal* 4: e11054. <https://doi.org/10.3897/BDJ.4.e11054>
- McClenachan L, Ferretti F, Baum JK (2012) From archives to conservation: why historical data are needed to set baselines for marine animals and ecosystems. *Conservation Letters* 5: 349–359. <https://doi.org/10.1111/j.1755-263X.2012.00253.x>
- Miller JA, Braumuller Y, Kishor P, Shorthouse DP, Dimitrova M, Sautter G, Agosti D (2019) Mobilizing data from taxonomic literature for an iconic species (dinosaurs, theropods, tyrannosaurus rex). *Biodiversity Information Science and Standards* 3: e37078. <https://doi.org/10.3897/biss.3.37078>
- Muñoz G, Kissling WD, van Loon EE (2019) Biodiversity Observations Miner: A web application to unlock primary biodiversity data from published literature. *Biodiversity Data Journal* 7: e28737. <https://doi.org/10.3897/BDJ.7.e28737>
- Owen D, Groom Q, Hardisty A, Leegwater T, Livermore L, van Walsum M, Wijkamp N, Spasić I (2020) Towards a scientific workflow featuring Natural Language Processing for the digitisation of natural history collections. *Research Ideas and Outcomes* 6: e58030. <https://doi.org/10.3897/rio.6.e58030>
- Pafilis E, Bērziņš R, Arvanitidis C, Jensen L (2017) EXTRACT 2.0: interactive identification of biological entities mentioned in text to assist database curation and knowledge extraction. *Biodiversity Information Science and Standards* 1: e20152. <https://doi.org/10.3897/tdwgproceedings.1.20152>
- Pafilis E, Frankild SP, Schnetzer J, Fanini L, Faulwetter S, Pavlodi C, Vasileiadou K, Leary P, Hammock J, Schulz K, Parr CS, Arvanitidis C, Jensen LJ (2015) ENVIRONMENTS and EOL: identification of Environment Ontology terms in text and the annotation of the Encyclopedia of Life. *Bioinformatics* 31: 1872–1874. <https://doi.org/10.1093/bioinformatics/btv045>
- Page R (2019) Text-mining BHL: Towards new interfaces to the biodiversity literature. In: *Biodiversity_Next*. Biodiversity Information Science and Standards. Available from: <http://eprints.gla.ac.uk/188463/>.

- Page R (2011) Extracting scientific articles from a large digital archive: BioStor and the Biodiversity Heritage Library. *BMC Bioinformatics* 12: 187. <https://doi.org/10.1186/1471-2105-12-187>
- Page R (2013) BioNames: linking taxonomy, texts, and trees. *PeerJ* 1: e190. <https://doi.org/10.7717/peerj.190>
- Penev L, Dimitrova M, Senderov V, Zhelezov G, Georgiev T, Stoev P, Simov K (2019) OpenBiodiv: A Knowledge Graph for Literature-Extracted Linked Open Data in Biodiversity Science. *Publications* 7: 38. <https://doi.org/10.3390/publications7020038>
- Perera N, Dehmer M, Emmert-Streib F (2020) Named Entity Recognition and Relation Detection for Biomedical Information Extraction. *Frontiers in Cell and Developmental Biology* 8. <https://doi.org/10.3389/fcell.2020.00673>
- Pyle RL (2016) Towards a Global Names Architecture: The future of indexing scientific names. *ZooKeys*: 261–281. <https://doi.org/10.3897/zookeys.550.10009>
- Reiser L, Harper L, Freeling M, Han B, Luan S (2018) FAIR: A Call to Make Published Data More Findable, Accessible, Interoperable, and Reusable. *Molecular Plant* 11: 1105–1108. <https://doi.org/10.1016/j.molp.2018.07.005>
- Richard J (2020) Improving Taxonomic Name Finding in the Biodiversity Heritage Library. *Biodiversity Information Science and Standards* 4: e58482. <https://doi.org/10.3897/biss.4.58482>
- Rivera-Quiroz FA, Petcharad B, Miller JA (2020) Mining data from legacy taxonomic literature and application for sampling spiders of the Teutamus group (Araneae; Liocranidae) in Southeast Asia. *Scientific Reports* 10: 15787. <https://doi.org/10.1038/s41598-020-72549-8>
- Steinböck O (1937) The fishery grounds near Alexandria. 14. Turbellaria. Ministry of Commerce and Industry, Egypt. Gov. Press, Cairo: 1–15.
- Stonebraker M, Bruckner D, Ilyas IF, Beskales G, Cherniack M, Zdonik S, Pagan A, Xu S (2013) Data curation at scale: The data tamer system. In: CIDR.
- Stuart-Smith RD, Edgar GJ, Barrett NS, Kininmonth SJ, Bates AE (2015) Thermal biases and vulnerability to warming in the world's marine fauna. *Nature* 528: 88–92. <https://doi.org/10.1038/nature16144>
- Thessen A, Preciado J, Jain P, Martin J, Palmer M, Bhat R (2018) Automated Trait Extraction using ClearEarth, a Natural Language Processing System for Text Mining in Natural Sciences. *Biodiversity Information Science and Standards* 2: e26080. <https://doi.org/10.3897/biss.2.26080>

Tsikopoulou I, Legaki A, Dimitriou PD, Avramidou E, Bailly N, Nikolopoulou S (2016) Digging for historical data on the occurrence of benthic macrofaunal species in the southeastern Mediterranean. *Biodiversity Data Journal*: e10071. <https://doi.org/10.3897/BDJ.4.e10071>

Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, 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>

7 Appendix

7.1 Abbreviations

API = Application Programming Interface

AWK = Aho Weinberger Kernighan

BHL = Biodiversity Heritage Library

BODC = British Oceanographic Data Centre

CLI = Command Line Interface

CTI = Community Temperature Index

DPI = Dots Per Inch

EMODnet = European Marine Observation and Data Network

ENVO = Environment Ontology

EOL = Encyclopedia Of Life

EM = Entity Mapping

FAIR = Findable Accessible Interoperable Reusable

GBIF = Global Biodiversity Information Facility

GLOBI = GLObal Biotic Interactions

GUI = Graphical User Interface

HTTP = HyperText Transfer Protocol

ID = Identifier

IR = Information Retrieval

IPT = Integrated Publishing Toolkit

JSON = JavaScript Object Notation

LSID = Life Sciences Identifier

MedOBIS = Mediterranean node of the Ocean Biodiversity Information System

NCBI = National Center for Biotechnology Information

NER = Named Entity Recognition

NLP = Natural Language Process

OBIS = Ocean Biodiversity Information System

OCR = Optical Character Recognition

OS = Operating System

PDF = Portable Document Format

PNG = Portable Network Graphics

REST = REpresentational State Transfer

WP3 = Work Package 3

URL = Uniform Resource Locator

SI = International System of Units

7.2 Tables

Table 1. Active text mining tools and their characteristics

Tool	Information extracted	Input	Interface	Steps
TextAnnotator	Generic Annotations	URL or Free Text	Web application	NER (3), EM (4)
BioNames	Taxa - phylogenetic relationships - link to original description	Scientific Names	Web application	EM (4)
BioStor	Literature	Scientific Names and other keywords	Web application	IR (1)
BOM (Biodiversity Observations Miner)	Scientific Names, Biotic interactions, Traits	PDF	Web application, Application Programming Interface (API)	NER (3), EM (4)
ClearEarth	Locality, unit, valu, functional traits, organisms' names	Free Text	CLI	NER (3)
Global Names Recognition and Discovery	Scientific names	Free Text, PDF or image	Web application, API, CLI	NER (3)
GoldenGate-Imagine	Mark up, enhance, extract text and data	PDF	GUI	OCR (2), NER (3), EM (4)
Taxon Finder	Scientific names	URL	Web application, API	NER (3)
Pensoft Annotator	Annotation of free text with ontology terms	Free Text	Web application, API	NER (3)
EXTRACT	Scientific Names, Environments and Tissue	Free Text	API, CLI	NER (3)
tesseract	Optical Character Recognition	image file	API, CLI	OCR (2)