Deliverable 4.4 (D4.4)

Workshop report including description of mechanisms and guidelines on mobilization of historical data into the systems M36

Project acronym:	EMODNET
Project name:	EMODNET: European Marine Observation and Data Networks
Call:	.....
Grant agreement:	.....
Project duration:	dd/mm/yyyy – dd/mm/mm (nn months)
Co-ordinator:	.....
Delivery date from Annex I:	M36 (Month yyyy)
Actual delivery date:	Mnn (Month yyyy)
Lead beneficiary:	.....
Authors:	Evangelos Pafilis, Lucia Fanini, Nicolas Bailly, Sarah Faulwetter, Dimitra Mavraki based on input from all "Participants"

Overview

Overall objective of EMODnet WP4 is to fill the spatial and temporal gaps in EMODnet species occurrence data availability by implementing data archaeology and rescue activities. To this end a workshop was organized in the Hellenic Center for Marine Research Crete (HCMR), Heraklion Crete, (8 - 9 June 2015) to access possible mechanisms and guidelines to mobilize legacy biodiversity data. Participants in this workshop were the data managers who actually implemented data archaeology and rescue activities.

The aforementioned is a two-part process of first identifying and locating occurrence data and then performing the steps required to incorporate them into a digital database, which further will be distributed through EurOBIS[1], and the EMODnet data portal[2].

In the context of EMODnet WP4, many old faunistic reports have been located which contain valuable occurrence data on marine species. The extraction of these data and their conversion into the OBIS format[3] (a Darwin Core[4] extension) is a slow and manual process.

During the HCMR's workshop the GoldenGATE-Imagine[5] software was demonstrated and participating data managers received training on how to semi-automate the previously mentioned tedious process.

Different types of legacy literature were explored such as expedition results, protocol logbooks and more biodiversity research articles. GoldenGATE-Imagine was used both for digital born files and for scanned image PDF files.

Via hands-on sessions the complete process was studied: starting from how to scan a document, to import it into GoldenGATE-Imagine, to mark different document sections as well as entities of interests (e.g. taxonomic mentions and location names), to upload the markup in the PLAZI server[6] and from there to retrieve the auto-generated Darwin Core Archives.

Finally, in addition to the hands-on sessions, extensive discussions among the data managers and the information technology experts resulted in the compilation reward-via-publication suggestions and best practices (e.g. in scanning documents) to the assistance of the data extraction process.

This report aims to present tools and state-of-art approaches in mobilizing of historical data, a hands-on evaluation of these tools by a group of data managers, a discussion on further improvements of such tools and downstream integration into literature and data repositories.

Achievements and current status

Completed on time

[1] http://www.eurobis.org/
[2] http://www.emodnet-biology.eu/portal/index.php
[3] http://www.iobis.org/data/schema-and-metadata
[4] http://rs.tdwg.org/dwc/
[5] https://github.com/plazi/GoldenGATE-Imagine
[6] http://plazi.org/wiki/Taxon_Search_Portal

Table of contents

1. Participants

1.	Adrian Teaca	GeoEcoMar, Romania	adrianxteaca_yahoo.com
2.	Aglaia Legaki	HCMR/Uni of Athens, Greece	aglalegaki_biol.uoa.gr
3.	Anastasis Oulas	HCMR, Greece	oulas _hcmr.gr
4.	Dimitra Mavraki	HCMR, Greece	dmavraki_hcmr.gr
5.	Dmitry Schigel	GBIF, Denmark	dschigel_gbif.org
6.	Donat Agosti	PLAZI, Switzerland	agosti _amnh.org
7.	Evangelos Pafilis	HCMR, Greece	pafilis _hcmr.gr
8.	Gabriella Papastefanou	HCMR/Uni of Athens, Greece	gabriella_papas_hotmail.com
9.	Guido Sautter	PLAZI/Uni Karlsruhe, Germany	gsautter_gmail.com
10.	Laura Boicenco	NIMRD, Romania	laura_boicenco_yahoo.com
11.	Lyubomir Penev	Pensoft Publishers, Bulgaria	lyubo.penev _gmail.com
12.	Marilena Tsompanou	HCMR, Greece	marilena-ts_hotmail.com
13.	Sarah Faulwetter	HCMR, Greece	sarifa_hcmr.gr
14.	Simon Claus	VLIZ, Belgium	simon.claus _vliz.be
15.	Stefanie Dekeyzer	VLIZ, Belgium	Stefanie.dekeyzer_vliz.be
16.	Teodor Georgiev	Pensoft Publishers, Bulgaria	preprint_pensoft.net
17.	Terry Capatano	PLAZI, USA	catapanoth_gmail.com
18.	Viktor Senderov	Pensoft Publishers, Bulgaria	datascience_pensoft.net

Figure 1: Workshop participants in front of the Hellenic Center for Marine Research, Cretaquarium entrance, 9th June 2015 (from left to right: participant 12, 5, 9, 15, 2, 18, 8, 11, 10, 1, 16, 6, 14, 17, 7, 4, 13)

2. Scientific background

The overall objective of EMODnet WP4 is to fill the spatial and temporal gaps in EMODnet species occurrence data availability by implementing data archaeology and rescue activities. This is a two-part process of first identifying and locating data and then performing the steps required to incorporate them into a digital database, which further will be distributed through EurOBIS, and the EMODnet data portal.

During this first part, many old faunistic reports have been located which contain valuable occurrence data on marine species. The extraction of these data and their conversion into OBIS format (a Darwin Core extension), is a time consuming, manual process. The tools presented in this workshop demonstrated a semi-automated process to extract these data, and at the same time create marked-up versions of the articles from which they derive.

Legacy literature includes a complementary set of observation and trait data for taxa that are not well represented in existing databases and which are considered an additional source for predictive models calculated in EU-BON (http://eubon.eu/). The problem is to extract this data and make it accessible to the EU-BON workflow. At the moment, Darwin Core Archives are used to export data harvested from legacy literature into GBIF, a supplier for EU-BON.

3. Extracting information from legacy literature: the manual procedure

3.1 Lifewatch Greece, EMODNET, and Lifewatch Belgium legacy literature data rescue

Legacy Literature Data Rescue activities are currently on-going in this EMODnet working group, as well as in the Lifewatch project. Lifewatch is the European e-Science Research Infrastructure for biodiversity and ecosystem research designed to provide advanced research and innovation capabilities on the complex biodiversity domain. The term “Research Infrastructure” refers to the strategic installation at a European/International level supplying facilities, resources and related services to the scientific and other user's communities to conduct top-level activities in their respective field of science. On the top of that, e-Science infrastructures capitalize existing resources, as well as data and data observatories from physical infrastructures, distributed centers and single research groups. A brief overview of the status at the time of the workshop is presented below:

A list of publications of species occurrence data which are currently being digitized by the Lifewatch Greece data managers, is presented below. These publications are mostly in the Biodiversity Heritage Library (BHL, www.biodiversitylibrary.org/), e.g. The Cumacea of the Puritan Expedition. Mitteilungen a. d. Zoologischen Station zu Neapel 14: 411–432. Calman, W.T., 1906: Scanned Document: http://biodiversitylibrary.org/page/9663476#page/431/mode/1up, Metadata: http://lifewww-00.her.hcmr.gr:8080/medobis/resource.do?r=pesche_abissali_puritan), but also in in-house repositories. These documents are primarily historical expedition records,

faunistic reports, logbooks of the Mediterranean Sea. Table 1 lists the progress status as of June 2015.

> 220 historical (pre-1945) publications / datasets identified
~70 of those chosen for digitization
> 50 annotated with metadata
~15 digitized and currently being quality-controlled

Table 1: Lifewatch Greece legacy literature data rescue progress

(based on a presentation by Dr. Sarah Faulwetter during the13th International Congress on the Zoogeography and Ecology of Greece and Adjacent Regions, Irakleio, Greece, October 7 to 11, 2015.)

In addition via four EMODnet small grants the digitization and integration process of datasets relating to the topics listed in Table 2 is in the final stage (please see individual reports by the grant holders):

Historical data on benthic macrofauna, demersal fish, and fish stomach content from the North Sea and Baltic Sea
Zooplankton Time series of France from 1966 onwards
Romanian Black Sea Phytoplankton data from 1956 - 1960
Romanian Black Sea Macrozoobenthos and Zooplankton

Table 2: EMODnet WG4 small-grant literature data rescue collections

Biological datasets identified using the Belgian Marine Bibliography (2012) · 199 selected data sources · 74 datasets described and archived	Publication years: - 1995 Data extracted: · > 1400 unique stations · > 4724 unique species · A total of 54 677 observation records
Biological datasets from Belgian-Kenyan research (2013) · 67 selected data sources · 67 datasets described and archived	
Phytoplankton data of the Belgian Part of the North Sea (2013-2014) Extraction focus: pigment & environmental variables, species observation data (plankton) · 41 selected data sources · 18 datasets described and archived	Publication years: 1968- 1981 Data extracted: · > 786 unique species · A total of 276510 biotic records · A total of 56.350 abiotic records Sources: Ijslandvaarten, Projekt Zee, Concerted Research Actions, Projekt Afvalwateren, Thesis : Smeets; Rabijns; Clarysse, De Block, Vanlangendonck, Robijns,...

Table 3: Lifewatch Belgium legacy literature data rescue information (based on a slide by Dr. Simon Claus presented on EMODnet WG4- EUBON Workshop, HCMR, 8-9 June, 2015).

3.2 Literature data extraction and digitization workflow

A summary of the manual occurrence data extraction from legacy literature is presented in Figure 2.

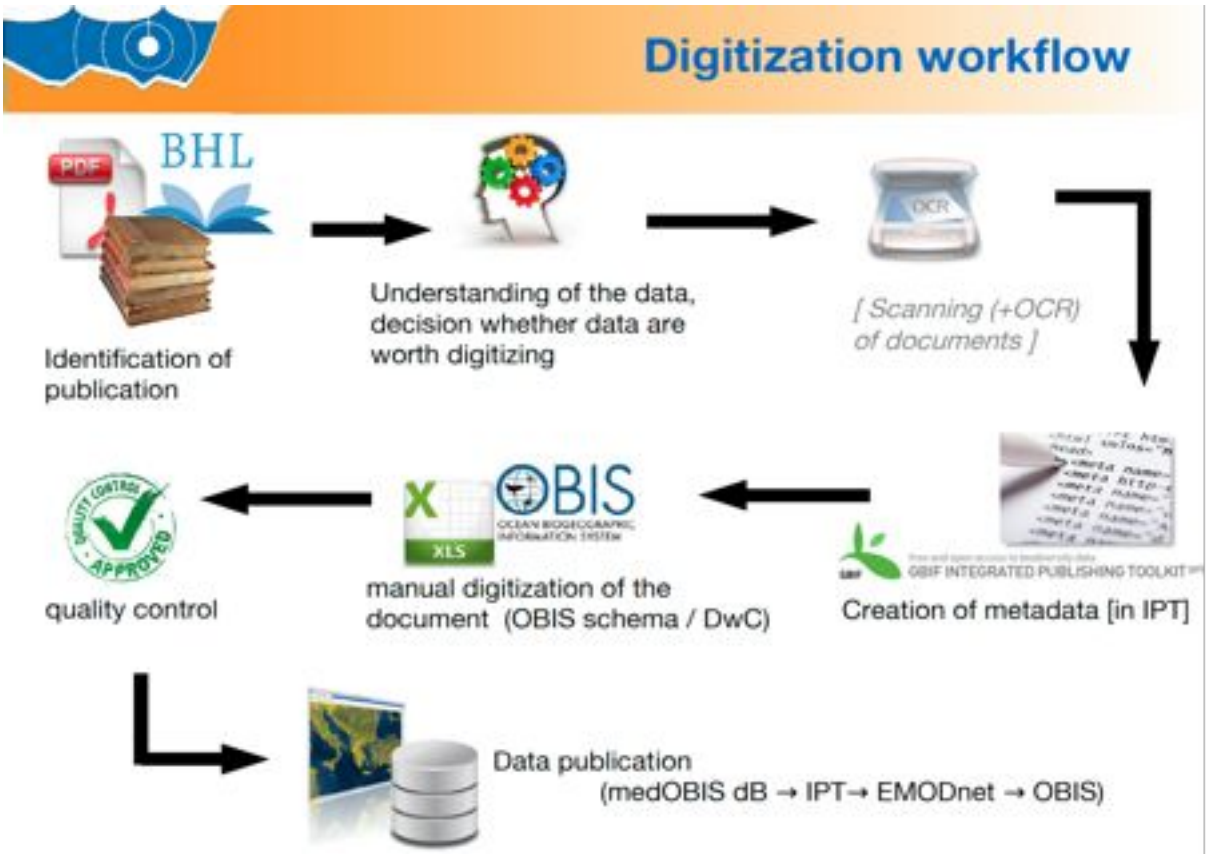


Figure 2: the manual data extraction from legacy literature workflow.
(based on a presentation by Dr. Sarah Faulwetter during the13th International Congress on the Zoogeography and Ecology of Greece and Adjacent Regions, Irakleio, Greece, October 7 to 11, 2015.)

The manual data extraction is is a seven-step procedure that starts with the identification and collection of the candidate literature for further processing. Document selection is mainly based on whether it contains adequate information on: taxonomic resolution, geographic resolution, temporal resolution, consistency of the information, presence/ absence vs. abundance, presence of additional information (e.g. sampling method). Another limited factor is the language that the document is written, which is related to the capability and knowledge of the data manager to translate it into english. Once a document has been selected and after it has been scanned; its metadata are extracted and registered by using the GBIF Integrated Publishing Tookit (IPT) repository. The document metadata cover among others the following sections: title & abstract, methods, taxonomic coverage, geographic coverage, temporal coverage, associated persons, usage rights. The next step of the workflow is the manual data occurrence extraction from the document. The extracted pieces of information are stored in a Darwin Core OBIS-compliant archive. Once extracted, the data occurrence information undergoes quality control. The latter includes: standardization of taxon names (according to the World Register of Marine Species), coordinates cross-checking, georeferencing and data consistency checks, for example in terms of time, depths and abundances). The last step of the workflow is its publication.

3.3 Common obstacles in the manual occurrence data extraction procedure

During the workshop an in-depth discussion, supported by example illustrations, revolved around collecting feedback from the data managers (i.e. the literature curators extracting species occurrence information) detailing the difficulties they are encountered with (Table 4).
First of all it is clear that most old publications are in languages other than English. As this issue is a challenge it self it has not discussed further during this workshop.

The manual extraction process of occurrence data is time-consuming, estimated ca. 7 hours for hand-digitizing 23 pages. While such work load is unavoidable, distractions are the first cause of loss of time and producing errors: typos, missing or mixed lines are frequent, and the attention level cannot be kept high for such long time (as a consequence, productivity reduces as time passes). Therefore we identified the basic problem in digitizing historical documents as the high demand in terms of time and repetitive work.

Further constraints identified are related to the complexity of datasets, i.e. not all the information is found in one table/figure or in the same document. Some examples: data related to the same record may be in other sources (such as station information, as often for the physical variables related to fauna abundance data). In that case there will be two or more documents to extract data from, simultaneously. Quite often abbreviations are used without any explanation. The latter can be extremely time consuming and frustrating.

3.3.1 Location information extraction

Information on locations is expressed in different ways, such as station names or named locations, with or without coordinates, shown in maps or in table (Figure 3). Location information cannot always be checked against a gazetteer, as often place names are given in their old form (e.g. Candia vs. Crete). In addition depth/(elevation for terrestrial) is an important location feature. In legacy literature often fathoms are used instead of meters (Figure 4), or described in prose like "shallow water". While units can be converted, there is not always agreement on the exact definition of descriptive terms such as “shallow”, “deep” and similar. Expressing the latter as a range or as Environment Ontology (http://www.environmentontology.org/) is an alternative.

It's also worth mentioning that as we deal with marine species; the precise location of sea/ocean sampling is a quite useful information. Quite often, the coastline of that period (do not forget that we deal with historical datasets) is not the same as it is now. This results many sampling locations to fall on land and not on sea. This has to be checked and noted down.

3.3.2 Data occurrence reporting and extraction

Data occurrences are reported in different formats and fashions. An occurrence might be reported as a simple presence/absence, as abundance (i.e. counts at a given location), as abundance with additional sex and life stage classification, as density (i.e. cells per litre); or as biomass (i.e. kg).

Within a manuscript, the information may be contained either within the text body (taxonomic section) or in a table, or in both (repetition of information). Sometimes, it may be

presented part in text and part in table. Data occurrences may be also found in different sections of a manuscript. The “Distribution“ section of the taxon description, whenever present, often reports new information mixed with other literature records. Data occurrence can be also expressed via complex sentences (e.g. "Taxon A was found here, while taxon B was not found, it was found instead in place XXX") and or via negation ("Argonauta was sought after, but not found").

As a matter of style, papers often contain first an extended classification of the encountered species, and then the actual "faunistic" section, which is basically a repetition of this classification, but with more information (e.g. descriptions, occurrences...).

Last but not least, information on time, place and methods is quite often separate from the observation records.

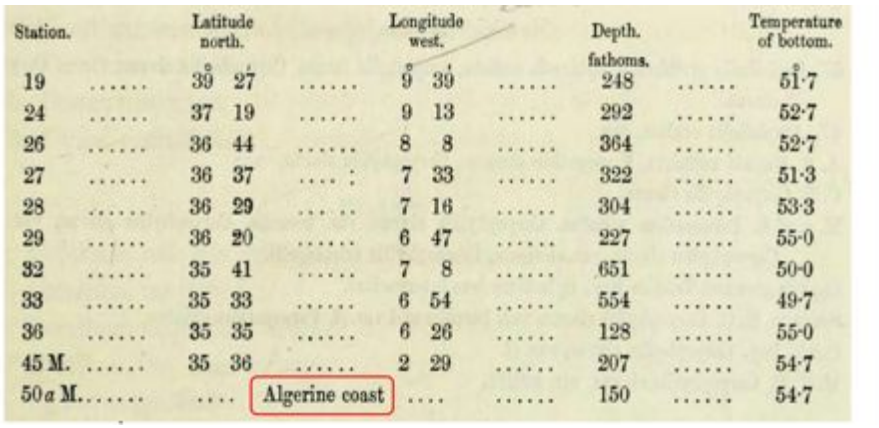
All the aforementioned constitute complexities that need to be overcome in the species data extraction processes. For example, presence/absence, abundance, and abundance with additional information need to be modeled, extracted and stored differently.

Complex sentences and negation delay the extraction process and might result in errors if they are overlooked. The need to combine information mentioned in different manuscript elements and sections, cause further delays.

It is pointed out that there are cases when data related to other expeditions need to be removed, as data extraction is expedition specific. Data extraction complexity also increases when additional but essential (e.g. sampling station details, methods, time) may exist in a different historical record that needs to be processed too.

Data managers, often use excel spreadsheets to compose the taxonomic information available. They create flat tables, containing all the taxonomic names and when species occurrence information is split into categories, each category is kept in a separate row of the spreadsheet (e.g. one row for occurrence of males at the given station, and one row for occurrence of females, both reporting the same station and related info (depth, temperature, etc.). If it is possible then every taxon is mapped to a specific sampling event with all the relevant information (date, location). Each taxon is also checked against WoRMS and the current accepted name is recorded by keeping the LSID. LSID is a way to name and locate pieces of information on the web. Essentially, an LSID is a unique identifier for some data, and the LSID protocol specifies a standard way to locate the data. An LSID is represented as a uniform resource name (URN) with the following format: URN:LSID:<Authority>:<Namespace>:<ObjectID>. In our case the taxon list of the provided dataset is matched against WoRMS and takes the following format: URN:LSID:marinespecies.org:taxname:number. It is common the taxonomic data to have typical taxonomic errors, such as misspelling, invalid names, inconsistencies, misidentifications. Data managers always keep the original name and add it to the database along with taxon remarks.

The debate about negative data is still ongoing. There is relevant information (e.g. related to alien and invasive species) that can be derived from negative data, however there is no clear workflow to extract such information from manuscripts and datasets.



Station.	Latitude north.	Longitude west.	Depth. fathoms.	Temperature of bottom.
19	39° 27′	9° 39′	248	51°·7
24	37 19	9 13	292	52·7
26	36 44	8 8	364	52·7
27	36 37	7 33	322	51·3
28	36 29	7 16	304	53·3
29	36 20	6 47	227	55·0
32	35 41	7 8	651	50·0
33	35 33	6 54	554	49·7
36	35 35	6 26	128	55·0
45 M.	35 36	2 29	207	54·7
50*a* M.	Algerine coast		150	54·7

Figure 3: Stations without coordinates is a common case, as well as older depth measurement units, for example fathoms (based on a slide by: Mrs. Aglaia Legaki, Mrs. Gabriella Papastefanou, Mrs. Marilena Tsompanou presented on EMODnet WG4- EUBON Workshop, HCMR, 8-9 June, 2015).

Slow and tedious process, Lifewatch Greece literature curators reported a rate of approximately 3 pages per hour
Most old publications are in languages other than English
Historical datasets are not always available in a good format (photocopies or scanned documents of low quality), so information cannot be easily read or extracted by software
Georeference is rather difficult. Information on locations can be either stations (with or without coordinates, either as a map or as a table), or a named location, or both (see Figure 3 for a scanned document example)
Occurrence information might be either simple presences or abundances (counts) at a location, the latter sometimes split into sex & life stages
Information can be either free-text (taxonomic section) or in a table, or both (repetition of information), or part in text and part in table (see Figure 4 for a scanned document example)
Important species occurrence information (e.g. sampling station details) may exist in a different historical record that needs to be processed too. Ditto for time, and methods
Authors are inconsistent in species name spelling. In addition, contradictions among authors (even in the same collection of document) cause important delays in the data extraction
In relevance to the expedition reports: data mentions to other expeditions need to be removed as data extraction is expedition specific
Rescuing data directly from their primary record (e.g. protocol logs) in certain case can be more important than digitizing the article in which they are published
Depth /(elevation for terrestrial) is important. Often fathoms are given instead of metres, or verbatim info such as "shallow water" (see Figure 3 for a scanned document example)
Location information cannot always be checked against a gazeteer, as often place names are given in their old form (e.g. Candia vs. Crete)
Cumulative info: treating e.g. all taxa of a certain family in one paragraph, but without structure (e.g. "Taxon A was found here, while taxon B was not found, it was found instead in place <another location>")
Actual observation information can be difficult to distinguish from literature-derived information, especially in the "distribution" section of the taxon description
Papers often contain first an extended classification of the encountered species, and then the actual "faunistic" section, which is basically a repetition of this classification, but with more information (e.g. descriptions, occurrences...)

Table 4: data managers' feedback summary on the manual literature digitization procedure

4. Issues in semi-automating the digitization process

A group of data managers, prior to the workshop, experimented with the use of optical character recognition (OCR) to semi-automate the legacy literature digitization process. A reflection on the experience gained and example cases of issues they encountered were presented to all and facilitated further discussion. Their feedback falls in two basic categories. The first relates to the quality of the optical character recognition and its application in reading legacy literature documents. The second refers to the occurrence information extraction and problems that may arise in the automated mining of document due to its authoring style and its contents.

4.1 Optical Character Recognition issues

Historical datasets most of the times are not available in a good format. It is common that these documents are available either as photocopies or scanned documents of low quality. This hinders OCR software from correctly recognizing characters in a scanned document (Figure 4).

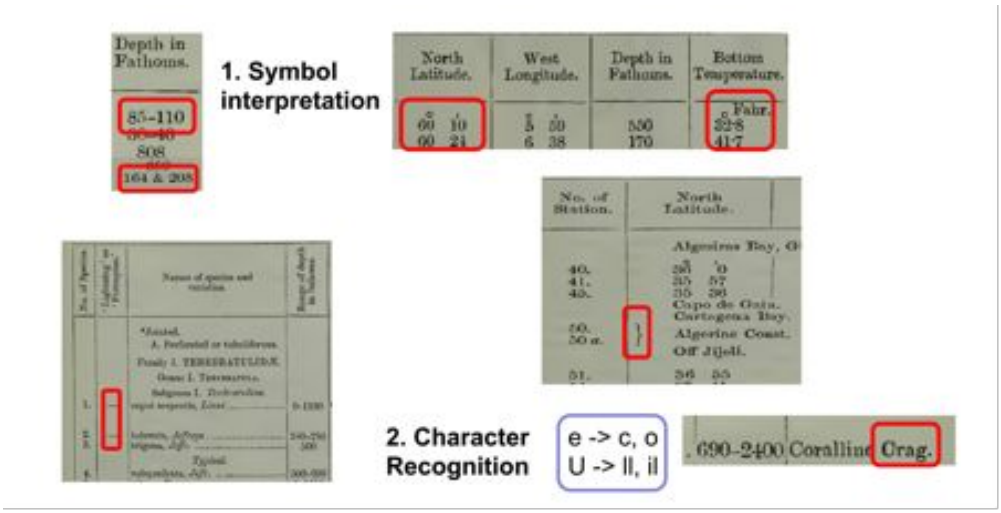
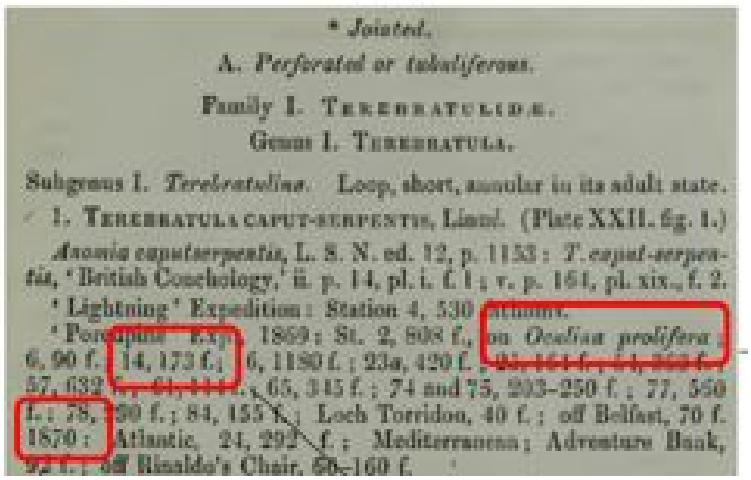


Figure 4: Symbol interpretation, such as correctly interpreting the meaning of “hyphen” as a range or an absence, “degree” e.g. as temperature or geographical coordinate, and “brackets”, and character recognition (e.g. misinterpreting an “e” as “c” or “o”) are two common issues when attempting to automatically digitize a legacy literature document (based on a slide by: Mrs. Aglaia Legaki, Mrs. Gabriella Papastefanou, Mrs. Marilena Tsompanou presented on EMODnet WG4-EUBON Workshop, HCMR, 8-9 June, 2015).

4.2 Automated occurrence information extraction issues

As explained in section 3.3.2 and reported in Table 4, biodiversity legacy document comprise mixed context: complex occurrence statements, negations, and references to background

knowledge and to other expeditions. These can lead to false positive species-location associations and to incorrect occurrence extraction. Such ambiguity would still be present even in the case of an ideal, 100% correct, OCR system. Expert's (data manager in our case) judgment is required to select the expedition specific data occurrences. Suggestion of species-location associations might accelerate the aforementioned step.


* *Jointed.*

A. *Perforated or tubuliferous.*

Family 1. TEREBRATULIDÆ.

Genus 1. TEREBRATULA.

Subgenus 1. *Terebratulina.* Loop, short, annular in its adult state.

1. TEREBRATULA CAPUT-SERPENTIS, Linné. (Plate XXII. fig. 1.)

Anomia caputserpentis, L. S. N. ed. 12, p. 1153: *T. caput-serpentis*, 'British Conchology,' ii. p. 14, pl. i. f. 1; v. p. 164, pl. xix., f. 2.

'Lightning' Expedition: Station 4, 530 fathoms.

'Porcupine' Exp., 1869: St. 2, 508 f., on *Oculina prolifera*; 6, 90 f.; 14, 173 f.; 6, 1180 f.; 23a, 420 f.; [illegible]

57, 633 f.; [illegible] 65, 345 f.; 74 and 75, 293–250 f.; 77, 550 f.; 78, [illegible]50 f.; 84, 155 f.; Loch Torridon, 40 f.; off Belfast, 70 f.

1870: Atlantic, 24, 292 f.; Mediterranean; Adventure Bank, [illegible] off Rinaldo's Chair, 95–160 f.

Figure 5. Mixed content issues that can lead to incorrect species-occurrence extraction. (based on a slide by: Mrs. Aglaia Legaki, Mrs. Gabriella Papastefanou, Mrs. Marilena Tsompanou presented on EMODnet WG4- EUBON Workshop, HCMR, 8-9 June, 2015).

As you can see at figure 5., *Oculina prolifera* (coral, Order: Scleractinia) might be recognized as an occurrence among branchiopods like Terebratulide. Assigning "14, 173f" to its correct station, depth, and distinguish this from "1870" (year of the second expedition); might required extra work in *ad hoc* software training and customization.

5. The software assisted document annotation process and data publication

To gain experience with automated methods for species occurrence extraction and data publishing, the PLAZI document annotation (http://plazi.org/) pipeline and PENSOFT's Biodiversity Data Journal (BDJ, http://biodiversitydatajournal.com/) were presented to the Lifewatch and EMODNET data managers.

5.1. PLAZI rationale and taxonomic treatments

Plazi is an association supporting and promoting the development and service of persistent and openly accessible digital taxonomic literature. To this end Plazi:

- Maintains a digital taxonomic literature repository to enable archiving, accessing and disseminating of taxonomic treatments and included data (see section 5.2.3).
- Enhances submitted taxonomic treatments by creating TaxonX and Taxpub XML versions (see section 5.2.2).

- Participate in the development of new models for publishing taxonomic treatments in order to maximize interoperability with other relevant cyberinfrastructure components (e.g., name servers, biodiversity resources, etc...)
- Advocate and educate about the vital importance of maintaining free and open access to scientific discourse and data (see the The Bouchout Declaration for Open Biodiversity Knowledge Management, http://www.bouchoutdeclaration.org/declaration/)
- Develops software required for the semi-automated markup of biodiversity literature documents (see section 5.2.1).

5.1.1 Taxonomic Treatment

Nothing in biology makes sense except in the light of treatments, Donat Agosti

It is estimated that in the several hundred million pages of printed species descriptions available in legacy biodiversity literature, 1.8M species descriptions are included therein. It's is PLAZI's aim to extract such pieces of information and make them available to the community in a searchable, interlinked and easy to nativgate fashion.

A Taxon Treatment is key concept and vehicle to such endeavor.

From a practical point of view a "Treatment" is a well-defined part of an article that defines the particular usage of a scientific name by an authority at a given time (a page(s) in a publication).

A Taxon Treatment (also referred to as "taxonomic treatment" in this report) is instantiation of the above for the case to organism taxa. Thus a taxonomic treatment can be seen as the scientific description of a taxon including a Latinized name of the nominate taxon, followed by one or several elements such as references to older literature citing this taxon and putting it in relation (nov.comb, syn., etc.), a description (a verbatim morphological description; that is why the element is not called description but treatment), distribution (a summary of the materials citated), materials citation (including references to the original specimen or observations used for the analysis), biology, ecology, host-relationships, etymology, etc. semantically described clause of data describing a taxon (source: "What is a treatment?" http://biosyscontext.blogspot.gr/2011/02/what-is-treatment-on-way-to-define-or.html)

From a legal and information dissemination point of view, a taxonomic treatment is a descriptive clause of text extracted from the literature. Thus, it constitutes an observation. As such, according to Swiss law (PLAZI's home) it is not copyrightable irrespective of the copyright status of the literature that the biodiversity researcher extracted it from.

A clear example is the taxonomic treatment of *Leucon longirostris* G.O. Sars Treatment extracted from the The Cumacea of the Puritan Expedition. Mitteilungen a. d. Zoologischen Station zu Neapel 14: 411–432. Calman,W.T.,1906, one of the document EMDONET legacy literature documents.

5.2 PLAZI pipeline components and auxiliary resources

5.2.1 GoldenGATE-Imagine

GoldenGATE-Imagine (GGI) document editor is an environment for marking up, enhancing, and extracting text and data from PDF files. It has specific enhancements for articles containing descriptions of taxa ("taxonomic treatments") in the field of Biological Systematics, but its core features may be used for general purposes as well.

GoldenGATE-Imagine (GGI) opens PDF documents, extracting or rendering page images, performing OCR or decoding embedded fonts where required, and finally segmenting pages into columns, blocks, paragraphs, and lines. Afterwards, it offers a wide range of (semi-)automated tools and manual markup and editing functionality.

Tools include:

- automated document structuring, comprising
- elimination of page headers
- extraction of figures and tables, together with their respective captions
- detection of footnotes
- detection of headings and their hierarchy
- document metadata extraction
- detection and parsing of bibliographic references, using RefParse
- markup of citations and linking to the corresponding bibliographic references
- detection, atomization, and reconciliation of taxonomic names, backed by Catalog of Life, GBIF, and IPNI
- markup of taxonomic treatments and their inner structure
- extraction and parsing of occurrence records
- tagging of trait terms, backed by ontologies.

At any stage of document processing, users can choose to export documents as XML. As soon as document metadata is extracted and taxon names and treatments are marked, they can also export a DarwinCore Archive or upload the document to the Plazi treatment repository, which allows them to share their markup work with the public. Further, users can export tables and figures, together with their captions

(text source: https://github.com/plazi/GoldenGATE-Imagine, sofware manual: https://docs.google.com/document/d/1mRSK4g0AVS1L4lbTYKH0Sq7JRX5qEWfPsYGNcWJE6A/edit?pli=1)

5.2.2 TaxonX Schema

Once a literature curator has defined the contents of a taxonomic treatment the latter have to be extracted and comminuted to repositories holding and indexing such clauses of information.

TaxonX, as a flexible and lightweight schema, facilitates such communication step by offering developers an agree-upon taxon treatment model with which they may package the extracted text (“encoding”).

In particular via Taxonx developers and data managers can:

- Create open, stable, persistent, full text digital surrogates of taxonomic treatments
- Identify taxonomic treatments and their major structural components to enable networked reference and citation

- Identify lower level textual data such scientific names, localities, morphological characters, and bibliographic citations to facilitate their extraction by, and integration with external applications and resources
- Study and describe the structure of systematics publications by creating few typical corpora of literature, such as entire journal (eg AMNH Novitates), across taxa (e.g all ant systematics papers post 1995), or faunistic (e.g. all ant systematics paper covering Madagascar ranging from 1758 to 2006)

(text source and more information: http://plazi.org/wiki/TaxonX_Schema)

5.2.3 The Biodiversity Literature Repository within Zenodo

The next step towards making the extracted taxonomic treatment available to the community is the upload of the former to an appropriate public literature repository.

Zenodo (https://zenodo.org/) is an open research home, enabling researchers to share and preserve any research outputs in any size, any format and from any science. Withing Zenodo, the Biodiversity Literature Repository (https://zenodo.org/collection/user-biosyslit, BLR: Figure 6) is a collection related to bio-systematics. Its goal is to provide

- open access to publications cited in publications or in combination with scientific names
- a digital object identifier (DOI) to enable citation of the publications including direct access to its digital representation.

Additional search functionality is available including searches in CrossRef, DataCite, PubMed, RefBank, GNUB and Mendeley.

A guideline document on how to upload literature document to BLR is available: https://drive.google.com/file/d/0B_yrQwn4yBySX3JkTV9RZzZfNUU/view

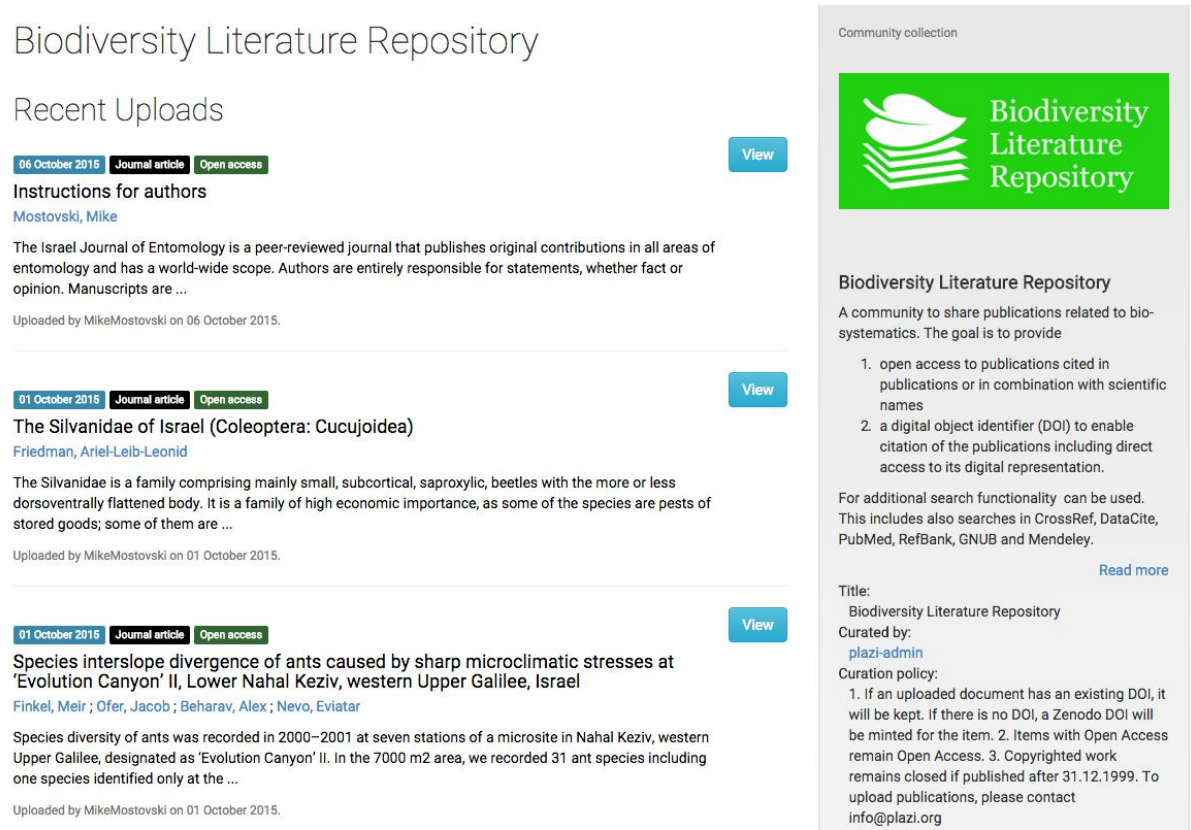


Figure 6: Biodiversity related articles and instructions to the authors available the Biodiversity Literature Repository home page (https://zenodo.org/collection/user-biosyslit)

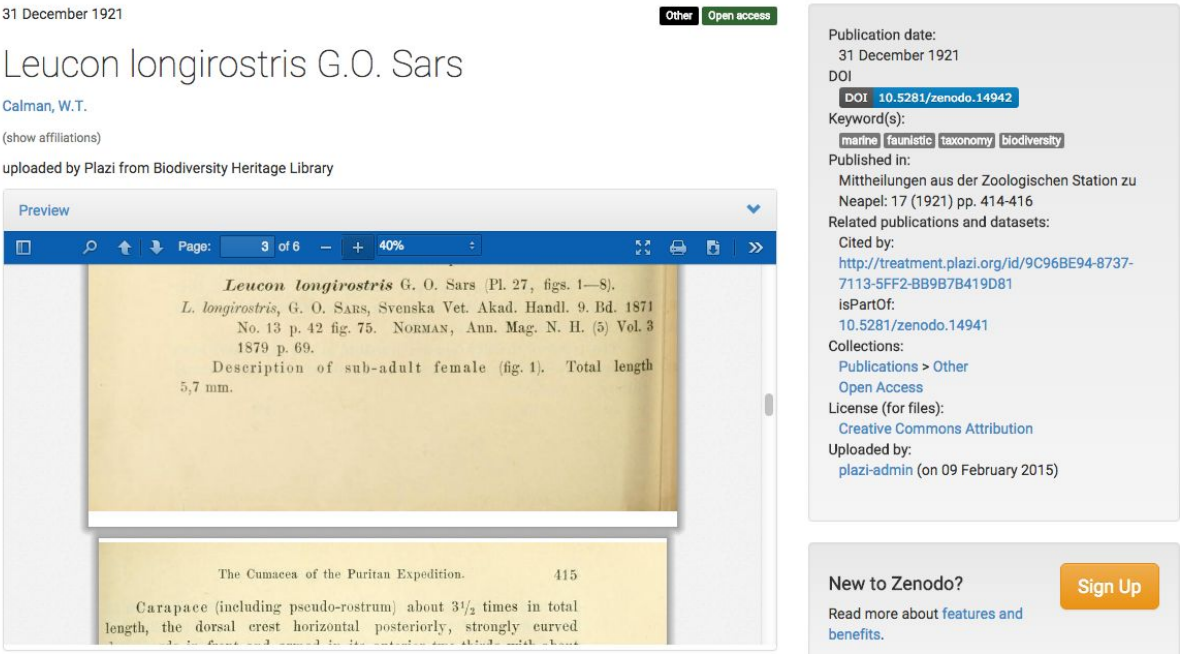


Figure 7: The Cumacea of the Puritan Expedition. Mitteilungen a. d. Zoologischen Station zu Neapel 14: 411–432. Calman,W.T.,1906 is avaiable in BLR as https://zenodo.org/record/14941. *Leucon longirostris* G.O. Sars (shown above) extracted from this expedition document is also avaible in BLR: https://zenodo.org/record/14942. Both links are assigned unique DOIs and are also retrievable as http://dx.doi.org/10.5281/zenodo.14941 , http://dx.doi.org/10.5281/zenodo.14942 accordingly)

5.2.5 GoldenGate-Imagine Tutorial

The GoldenGATE-Imagine tutorial followed in this workshop is based on the the GoldenGATE-Imagine Manual available in: https://docs.google.com/document/d/1mRSK4g0AVS1L4lbTYKH0Sq-7JRX5qEWfPsYGNcWJE6A/edit#

5.3 Data papers and the Biodiversity Data Journal

Data papers (https://en.wikipedia.org/wiki/Data_paper) are "scholarly publication of a searchable metadata document describing a particular on-line accessible dataset, or a group of datasets, published in accordance to the standard academic practices" (Chavan, V., & Penev, L. 2011, doi:10.1186/1471-2105-12-S15-S2). Their objective is to enable "information on the what, where, why, how and who of the data", Callaghan, S., *et al.*, 2012, doi:10.2218/ijdc.v7i1.218).

Given the the previous two definition, a *data paper* could complement a legacy-literature-extracted species occurrence dataset release in an *ad hoc* repository such as GBIF and OBIS, increase outreach and facilitated retrievability (see also section 8.3, and Figure 9).

The Biodiversity Data Journal (BDJ, http://biodiversitydatajournal.com/) builds around such *data paper* concept and aims to consitute both a workflow and an infrastructure to mobilize, review, publish, store, disseminate, make interoperable, collate and re-use data through the act of scholarly publishing.

In particular BDJ is a community peer-reviewed, open-access, comprehensive online platform, designed to accelerate publishing, dissemination and sharing of biodiversity-related data of any kind. All structural elements of the articles – text, morphological descriptions, occurrences, data tables, etc. – are treated and stored as data, in accordance with the Data Publishing Policies and Guidelines of Pensoft Publishers (http://www.pensoft.net/J_FILES/Pensoft_Data_Publishing_Policies_and_Guidelines.pdf) .

The journal publishes papers in biodiversity science containing taxonomic, floristic/faunistic, morphological, genomic, phylogenetic, ecological or environmental data on any taxon of any geological age from any part of the world with no lower or upper limit to manuscript size. For example:

- Single taxon treatments and acts
- Local/regional and habitat checklists, sampling reports, N2K inventories
- Ecological and biological observations on species and communities
- Identification keys
- Data papers describing data
- Descriptions of software tools and workflows

(text source: http://biodiversitydatajournal.com/about#Focus-and-Scope)

6. Reflection on the EMODNET WG4 legacy document pilot annotation

Appendix 9.1 lists legacy literature documents that were attempted to be digitized with GoldenGATE-Imagine.

Below are some overall remarks based on the data manager feedback on PLAZI/GGI

- 48% of the PLAZI names are not in GBIF. This implies there is great potential for new contributions such as the Lifewatch/EMODNET legacy literature data rescue
- a really useful GGI is the data table extraction
- there were problems with OCR documents (e.g. loading, processing) including BHL retrieved PDF files
- the identification and correction of errors due to distraction also needs at least double-check of the data, possibly by another person.
- an occurrence-extraction specific version of GGI could be spinned out
- a possible improvement of GGI could then be its adaptation to open e.g. a .zip file with a bunch of image files resulting from the scanning.
- OCR effort, can be pushed from 5 down to 2- minutes per page with experience/GGI improvements

- marking up has a learning curve, it pays off more the longer your document is

7. Evaluation of the semi-automated annotation pipeline

Data managers were asked to evaluate the semi-automated annotation pipeline evaluated based on a subjective measure via a system usability questionnaire. The questionnaire can be found in http://www.biocreative.org/tasks/biocreative-v/track-user-interactive-task/ and is also available in Appendix 9.2 (citation information is included therein).

Questionnaire evaluation:
Users present at the workshop evaluated different aspects (G1-G6) of GoldenGATE-Imagine system. Evaluations were provided on a Likert scale (1-5), with no need of reversion (i.e. all questions were formulated with a positive statement).

Given the low sample size (N = 7 complete questionnaires returned), and the spread between experienced and novice users, results are only presented in a descriptive way, as summaries by group of questions. Scores are presented below.

Due to high occurrence of Non-Applicable (NA) answers (more than 50%), the following questions/groups of questions were not evaluated (in red in the legend):
. The group of questions regarding comparison to similar systems (G2). NAs were due to lack of experience of similar systems, even from interviewed with sufficient experience (1-3 years) in the task.
. The question No3 (“documentation and help”) of G5. The other questions of the same group received instead and answer from all the interviewed.

G1. Overall reaction	**G2. Overall comparison to similar systems**
G1.Q1. Please rate your experience with the system	G2.Q1. System is easy to use
G1.Q2. How would you rate the system	G2.Q2. Satisfaction with use
G1.Q3. Would you recommend the system to others?	G2.Q3. Power to help complete task
	G2.Q4. Flexibility in modes to use
G1 score: 84 (range 21-105)	NA due to lack of experience of similar systems

G3. System's ability to help complete tasks	G4. Design of application
G3.Q1. Speed: the system decreases the time it takes to reach my curation goal	G4.Q1. Ease of reading text
G3.Q2. Effectiveness: the system helps me get closer to my curation goal	G4.Q2. Use of highlighting
G3.Q3. Efficiency: with this system I can be both fast and effective	G4.Q3. Organisation of information
	G4.Q4. Sequence of screens
G3 score: 45 (range 12-60, due to NAs)	G4 score: 92 (range 28 - 140)

G5. Learning to use the application	**G6. Usability**
G5.Q1. Learning to operate the application	G6.Q1 Speed
G5.Q2. Remembering features	G6.Q2 Reliability
G5.Q3. Documentation and help	G6.Q3 Consistency
G5.Q4. Straightforwardness of use	G6.Q4 Ease of correcting mistakes
	G6.Q5 Error messages
G5 score: 57 (range 24 – 120)	G6 score: 105 (range 35-175)

Table 5. A summary of the GoldenGateImagine software evaluation. Positive scores have been achieved in the question groups 1, 3 and 4; neutral score in the question group 6 and negative score in group 5. The complete questionnaire is available in Appendix 9.2. Please note that due to the small sample size (7 complete questionnaires returned) the above results are considered indicative only.

8. Overall discussion and feedback

The following emerged from the meeting, where data managers compared their hands-on experiences. By taking note of all the pitfalls towards digitization and possible solutions to overcome them, coming from good practices, we hope to provide insights for further developments and more efficient work. Issues are presented along with respective solutions/mitigation proposed.

8.1 OCR best practices and BHL scanned document retrieval

OCR misreading in old documents is very common. In some cases it makes more sense to retype than to edit the image.

Best practice in this case is to try to get a good scan as much as possible. If outsourcing of the document scanning to a specialized company this it is suggested this option is explored further. For in-house document scanning some practical tips are listed below.

Non-necessary parts of the document can be omitted, for the sake of relevant ones: spending a initial time in looking at the document and locating the points of interest can save time later and allow the data manager to work on high quality scans.

Even if background documents may have different characteristics, making generalization difficult, a few general guidelines for scanning can be derived from successful experiences:

For older book pages (19th and 20th century) we find that capturing and OCRing in color gives more accurate results than grayscale or bitonal. The files can always be converted to bitonal after OCR (if necessary for storage limitations).

For book digitization images should be captured at a minimum of 400 ppi (at 100% item size). If the font size is particularly small or complicated, we will capture at 600 ppi. But 400 ppi is the minimum (The Digital Imaging Lab experience).

If a 35mm camera is available (16, 24 or 36 megapixels), the frame could be filled as much as possible and then down sampled to 400 ppi. This usually give a sharper and more detailed image than capturing at 400 ppi the objects original size (Dave Ortiz pers.comm.).

In summary, suggested specifications would be:

Scanning Mode	Scanning Resolution	Output Format	Image	Color Depth
RGB color	400 ppi (at 100% of objects size)	TIFF		48 bit

Table 6. Recommended OCR book scanning specifications

A suggested procedure is to retrieve a scanned document from BHL . In that casesomeone could visit the corresponding web page on the Internet Archive (http://archive.org) to retrieve the JP2 (jpeg2000) version of the scanned book (Figure 8). From the image file he could then create the PDF.

Figure 8: *Top*: to retrieve a scanned BHL book document from BHL click on the "Download Contents" icon on the top-right and select to browse the corresponding web page on the Internet Archive ("View at Internet Archive"). *Bottom*: there you may find the link to the jpeg2000 (JP2) image on the bottom right. The web pages shown in this example are:

top: http://biodiversitylibrary.org/page/9663476#page/431/mode/1up,
bottom: https://archive.org/details/mittheilungenaus17staz

From the experience gained through this workshop; the scanning process itself seems to be a bottleneck in terms of the quality, size, resolution etc. of a scan. These factors seemed to be crucial for the quality of the OCR process. It would be good to get some basic information on a "scanning best practice", this is also important for the success of GoldenGATE-Imagine - otherwise users might experience frustration. Due to all the constraints discussed above, it is quite obvious that not all is doable: some documents are simply too difficult and in such cese impossible to proceed with an automated way.

Best practice would then be to seek advice at the start-phase. Sample documents can instead be collected and classified: simple to complex, doable to non-possible, through different document types. Seek advice at the start-phase. The PLAZI team would be happy to offer feedback to the community requests[d2] .

8.2 "Reward" of data managers

Having to deal with such huge amount of work and constraints affecting the speed and efficiency, data managers should be given incentives to pursue on their efforts. Publishing the outcomes of their work and being cited - when the extracted data are used in other analyses - could be the most obvious incentive. This will allow old papers to be newly shareable and searchable, offering baselines for current research (including outreach the marine community - OBIS is a well respected broker to this end). Credit should be given to people who made this possible.

A high quality sampling report can comprise an exemplar paper describing the legacy documents, the rescue methodology, and the data extraction process along with the results. In addition to publishing the species occurrence in the GBIF/OBIS repositories, linking the results to PLAZI taxonomic treatments and to occurrence repositories could add value and strengthen the outreach of the extracted datasets.

A publication as data paper (e.g. in BDJ) is also recommended. An integrated workflow, e.g. from annotation in GGI, to publishing in BDJ could assist this process. Emphasis should not only be given in the initial publication of a dataset, but also in the ability to incrementally include annotations, corrections, and additional elements (e.g. tables, maps) once these have been established.

An exemplar data publication could be the Danish Mediterranean Expedition (Thor), currently digitized by the Lifewatch Greece data managers, as well as the Egypt Expedition, already published at the 13th International Congress on the Zoogeography and Ecology of Greece and Adjacent Regions, Irakleio, Greece, October 7 to 11, 2015

8.3 Data publication landscape

Data papers are strongly recommended giving the emerging success of open data (Figure 9). Although the issue of peer-reviewed should be clarified in this respect: BDJ (section 5.3) is considered a peer reviewed scientific journal, but reviewing this kind of data is still debated. Moreover, there is a different approach from different funders: some of them do not consider (yet) data papers as peer-reviewed papers. Even when undergoing peer-review, for datasets including taxonomic information you need taxon experts to act as reviewers. To facilitate this task,

PLAZI's dashboards can be generated out of a data paper, to be used as a QC tool for the reviewers; a devoted component in the BDJ publishing tool can also be added.

It is moreover advisable to have someone used to work on the kind of data digitized to perform a further control, answering the question: would these data be of use for your models? e.g. an ecologist or an oceanographer (i.e. data related to Climate Change are useful if down to the second decimal).

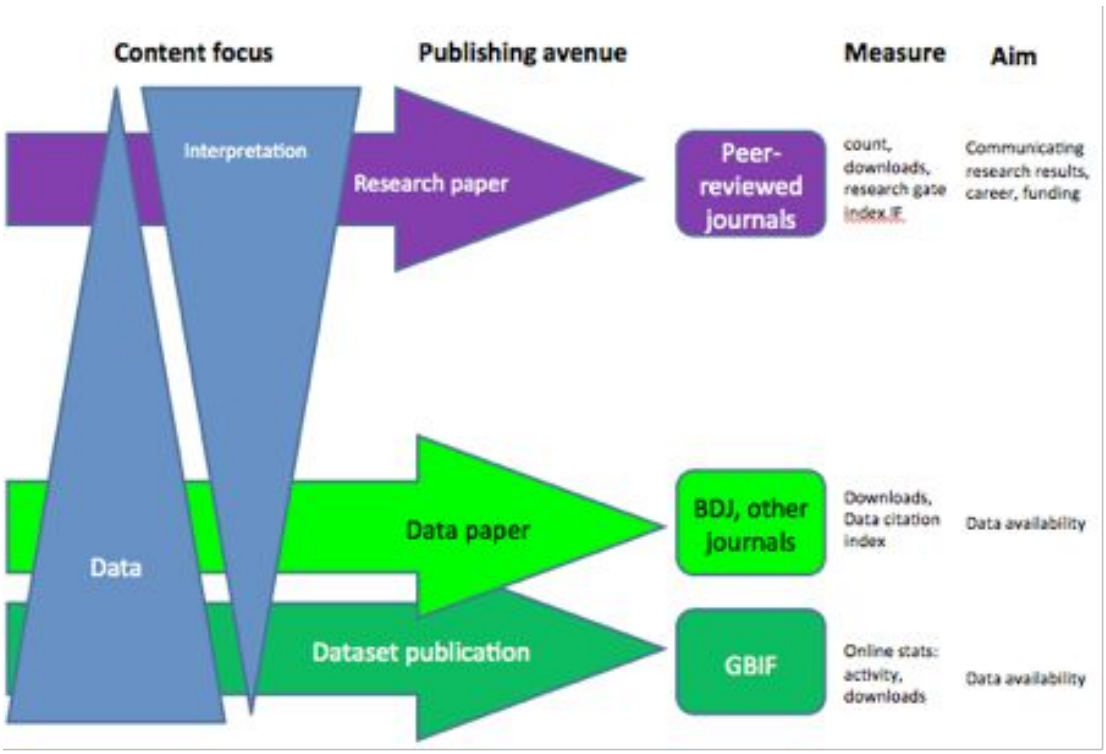


Figure 9: Open Data: an emerging landscape of data and other academic publications (courtesy: Dr. Dmitry Schigel)

9. APPENDIX

9.1 Hands-on session datasets

1. The cumacea of the puritan expedition, http://biodiversitylibrary.org/page/9663476
2. A description of the Madreporaria dredged up during the expeditions of H.M.S. 'Porcupine' in 1869 and 1870. (1873) http://www.biodiversitylibrary.org/page/31661501#page/397/mode/1up
3. On the mollusca procured during the 'Lighting' and 'Porcupine' expeditions,1868-70. part I (1878) http://www.biodiversitylibrary.org/item/90438#page/487/mode/1up
4. Laackmann, H., 1913. Adriatische Tintinnodeen. - Sitz. K. Akad. Wiss. Wien Math. nat. Klasse 122: 123-163 http://biodiversitylibrary.org/page/36050965
5. VIII. On the Annelida of the 'Porcupine' expeditions of 1869 and 1870.By W.C. McIntosh(1875) http://www.biodiversitylibrary.org/item/91779#page/543/mode/1up

9.2 System usability questionnaire

Original source of the BioCreative IV Interactive Annotation Task, system usability evaluation questionnaire: http://ir.cis.udel.edu/biocreative/survey2.html (BioCreative IV Interactive Task. Sherri Matis-Mitchell, Phoebe Roberts, Catalina O. Tudor and Cecilia N. Arighi. *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop vol. 1*, 190-203,
Described in: http://www.biocreative.org/media/store/files/2013/bc4_v1_27.pdf

Please rate the usability of the system. Please respond to all items below as best you can.
Top of Form
Your name:
Your email:
System you are assessing:
How would you rate your experience level for this task?

- novice to the task (less than one year)
- sufficient experience (1-3 years)
- expert (more than 3 years)

Where could this system fit into your curation workflow?
What changes would be needed to make it fit into your curation workflow?

Overall reaction

Please rate your experience with the system.	How would you rate the system?	Would you recommend this system to others?
very negative	very bad	not at all
negative	bad	if no other option
neutral	neutral	neutrally
positive	good	positively
very positive	very good	enthusiastically
comments?	comments?	**comments?**

Overall comparison to similar systems

System is easy to use: (compared to similar systems)	Satisfaction with use: (compared to similar systems)	Power to help complete task: (compared to similar systems)	Flexibility in modes of use: (compared to similar systems)
much harder	very frustrating	completely inadequa	completely inflexible
harder	frustrating	less powerful	less flexible
about the same	about the same	about the same	about the same
easier	satisfying	more powerful	more flexible

much easier	very satisfying	sufficient	highly flexible
NA	NA	NA	NA
comments?	comments?	comments?	**comments?**

System's ability to help complete tasks

Speed: the system decreases the time it takes to reach my curation goal:	Effectiveness: the system helps me get closer to my curation goal:	Efficiency: with this system I can be both fast and effective:
strongly disagree	strongly disagree	strongly disagree
disagree	disagree	disagree
neutral	neutral	neutral
agree	agree	agree
strongly agree	strongly agree	strongly agree
NA	NA	NA
comments?	comments?	**comments?**

Design of application

Ease of reading text?	Use of highlighting:	Organization of information:	Sequence of screens:
very hard	unhelpful and distracting	very confusing	very confusing
hard	unhelpful	confusing	confusing
neutral	not distracting	intuitive	intuitive
easy	helpful sometimes	helpful	helpful
very easy	very helpful	very helpful	very helpful
NA	NA	NA	NA
comments?	comments?	comments?	**comments?**

Learning to use the application:

Learning to operate application:	Remembering features:	Documentation and help:	Straightforwardness of use:
very hard	very hard	very confusing	hard even after help
hard	hard	confusing	hard
neutral	neutral	neutral	neutral
easy	easy	helpful	easy after help
very easy	very easy	very helpful	completely intuitive

NA NA NA NA

comments? comments? comments? **comments?**

Usability:

peed: Reliability: Consistency: Ease of correcting mistakes Error messages:

very slow very unreliable very inconsistent very hard very unhelpfu

slow unreliable some inconsistencies hard unhelpful

acceptable adequate neutral adequate

fast reliable adequate easy helpful

very fast completely reliable mostly consistent very easy very helpful

NA completely consistent NA NA

comments? NA comments? comments?

comments? NA

comments?