EMODnet

European Marine Observation and Data Network

EMODnet Biology

EMFF/2019/1.3.1.9/Lot 6/SI2.837974

EMODnet Phase IV

D2.3: Report on efforts undertaken in rescuing historical data through citizen science

Disclaimer[1]

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	D 2.3: Report on efforts undertaken in rescuing historical data through citizen science
WP title	WP2: Access to marine biological data
Task	Task 1: Maintain and improve a common method of access to data held in repositories
Authors	Leen Vandepitte, Dimitra Mavraki, Ruben Perez Perez, Georgia Sarafidou, Savvas Paragkamian, Vasilis Gerovasileiou, Laura Marquez
Dissemination level	Public
Submission date	30/11/2022
Deliverable due date	15/10/2022

[1] The disclaimer is needed when the document is published

Contents

Report on efforts undertaken in rescuing historical data through citizen science

The main objective for WP2 is covered in Task 1: Maintain and improve a common method of access to data held in repositories. The data made available during EMODnet Biology Phase IV primarily includes the following groups: macroalgae, angiosperms, benthos, birds, fish, mammals, phytoplankton, reptiles and zooplankton. The geographical scope focuses on the European seas, more specifically defined in six regions: Arctic, Atlantic, Baltic Sea, Black Sea, Mediterranean Sea and North Sea, including their coastal and estuarine zones.
In Phase IV, the Work Packages on 'Data access to marine biological data' and 'Data archaeology and rescue' were merged. As processing of historical data mostly starts with digitization (in this report the term digitization refers to the transcription process, and not to the process of scanning), two historical datasets - in paper form- were selected as a test-case to evaluate the feasibility and ease of engaging volunteers to rescue a selection of the identified historical biodiversity data through existing Citizen Science (CS) platforms.

Historical biodiversity documents comprise an important link to the long-term data life cycle and provide useful insights on several aspects of biodiversity research and management. One of the main goals of rescuing historical marine biodiversity data is to better understand the ocean's past, in order to predict the future of ocean life. The Deliverable 2.3 of WP2 in Phase IV entitled "Report on efforts undertaken in rescuing historical data through citizen science" started in M12 and aims at involving citizen scientists in this journey of knowledge. As there is an enormous amount of (meta)data waiting to be "FAIRified" (Findable, Accessible, Interoperable and Reusable), citizen science transcription efforts seem to be a promising resource towards this end. In the present report the experience of setting up the activities and familiarizing the volunteers with the historical data rescue concepts are described and analyzed.

1 Exploration of existing citizen science platforms

As several online citizen science platforms already exist, a first task consisted of discovering these platforms, and exploring their capabilities and functionalities. Exploration and evaluation of each platform was performed from both the perspective of a data manager, as well as the perspective of a volunteer, also referred to as the citizen scientist.

Table 1 Platforms explored and a short evaluation of their main aspects, listed in alphabetical order.

Platform	Evaluation
Cartoscope	CS platform for image labeling/not serving our purpose
Citizen Cyberlab	Team promoting the CS concept/not serving our purpose
Citizen Science Grid	No longer active webpage
CrowdCrafting	CS project builder/paying service
Digivol	Crowdsourcing transcription project for Australian terrestrial collections/ not serving our purpose
DoeDat	CS project builder platform, free of charge/serves our purpose
Epicollect5	CS project builder, available only as an app
Fromthepage	Transcription software/paying service
iNaturalist	Species observation platform/not serving our purpose
SciStarter	Project finder platform/not serving our purpose
Spotteron	CS application designer/paying service
Transkribus	Digitisation and transcription software & webpage/not serving our purpose
World Community Grid	Donation of unused computer power/not serving our purpose
Zooniverse	CS project builder platform, free of charge/serves our purpose

Table 1 presents all the evaluated CS platforms (14 in total). For every single platform, targeted questions were raised, considering the data manager's and citizen scientist's perspective. These questions are presented below and provide a holistic approach, covering the needs of the efforts undertaken to engage volunteers to historical data rescue.

From the data manager's perspective, the questions raised are:

- Does the platform offer unlimited usage (more than one dataset at the same time)?

- Could more than one team be responsible for the project?
- Does it enhance the collaboration of a community of citizens?
- Does it offer notifications for the progress of the project?
- Does the platform provide data export capabilities in a feasible way?
- Could the project be provided in multiple languages?
- Does the platform offer a version control (ie a time stamp on input/changes) of the project?
- Does it provide a difficulty ranking of the dataset?

From the citizen scientist's perspective:

- Is the platform fun to use/ is it attractive?
- Is there any reward/feedback given?
- Is it available in various operating systems?

From both users' perspective:

- Is the platform free of charge?
- Does it offer helpdesk possibilities?
- Does it provide options for manuals/tutorials?

A full overview (table-format) of the evaluation can be found in Addendum. Based on the evaluation, two platforms were chosen as the most appropriate to be tested extensively, by publishing a project on them and inviting citizen scientists to do the transcription. These platforms were **Zooniverse** and **DoeDat**. For both platforms, comments and experience gained are described in more detail below.

2 Dataset selection

The decision on the platforms to be evaluated, was followed by a decision on which historical datasets are considered more suitable for volunteers to participate in this test. Based on input from all WP2 partners for this exercise, a list containing 13 possible datasets was compiled and analyzed. Both HCMR and VLIZ teams tested a platform each, Zooniverse and DoeDat respectively, and it was decided to select the most appropriate dataset for each case, based on a number of facts such as geographical and temporal scope, language of the text, as well as the structure of the publication (purely text-based or with tables).

The HCMR team, for the Zooniverse platform, selected the "Report on the Danish Oceanographical expeditions 1908-1910 to the Mediterranean and adjacent seas. Vol II Biology. A.8 Lepadogaster By Frederic Guitel (1919)" dataset. The Danish Oceanographic Expedition is a series of historical datasets, which are available in the HCMR library in Athens, Greece. The digitization effort of these datasets began with the introductory volume in 2016 (Mavraki *et al.*, 2016) and it continues up to now with the digitization of more reports belonging to this historical expedition. The original text of the report is in French, but for the needs of this CS initiative, an English guide was provided to the volunteers.

The selected data paper for DoeDat came from the Mercator training ship expedition data series, which were still awaiting digitization. From the series, the 'Pisces' (fish) report was selected, due to its language (English), average length and clear structure of the publication itself.

3 Zooniverse

3.1 Background

The Zooniverse (Fig. 1) is considered as "the world's largest and most popular platform for people-powered research" since it hosts the largest collection of online citizen science projects in the world. It is "a collaboration between the University of Oxford, Chicago's Adler Planetarium, the University of Minnesota – Twin Cities (UMN), hundreds of researchers, and over 2 million participants from around the world". In Zooniverse thousands of volunteers join citizen science projects in order to contribute in their own way to the scientific community, with no prior expertise. The projects may concern fields such as astronomy, ecology, humanities, physics, and beyond. One important aspect of keeping this community engaged is the "Talk" board of the platform, where any registered user, whichever their identity may be (researcher, project member, volunteer, support member, etc.) may open a thread and start a conversation about anything regarding the platform. It works like social media, but for the Zooniverse users.

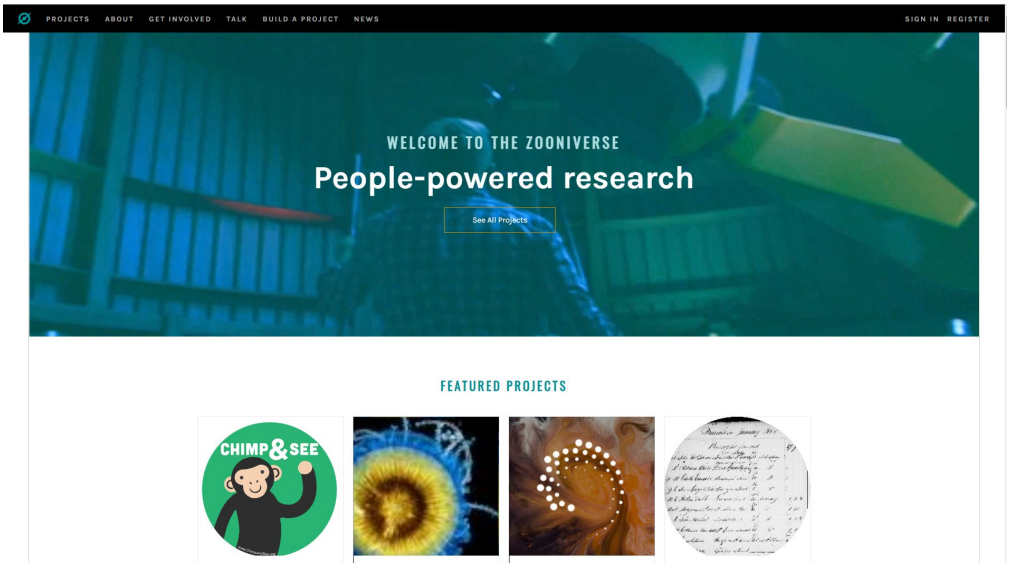


Figure 1. Zooniverse platform homepage screenshot.

As far as the creation of a project is concerned, a user-friendly "*Project Builder*'' is provided by the platform, which is accompanied with a detailed tutorial for the relevant Data Management Team (DMT) to consult; without any cost. There are numerous ways in which a project can be built with the option among four tasks or the combination of them. Many of the Biology and Nature related projects are about species identification, organism count and historical document/label transcription (e.g. Notes from Nature). Further down in the document it is described how the HCMR DMT project was built.

3.2 Preparation - “digging up the oceans’ past” project

3.2.1 Building the project

The project entitled “digging up the oceans’ past” (from now on “dutop”) was built by the HCMR DMT within the Zooniverse platform. The team had great flexibility on the building and formatting of this CS project. The platform offered several basic tabs (shown on the left handside of Fig. 2), each of which included a number of fields that were filled in and personalized according to the project’s needs and based on the extensive guidelines provided. Within these fields, various characteristics of the project were moderated. The interaction between the HCMR DMT and the Zooniverse support team and volunteers has been crucial for the development and improvement of the project.

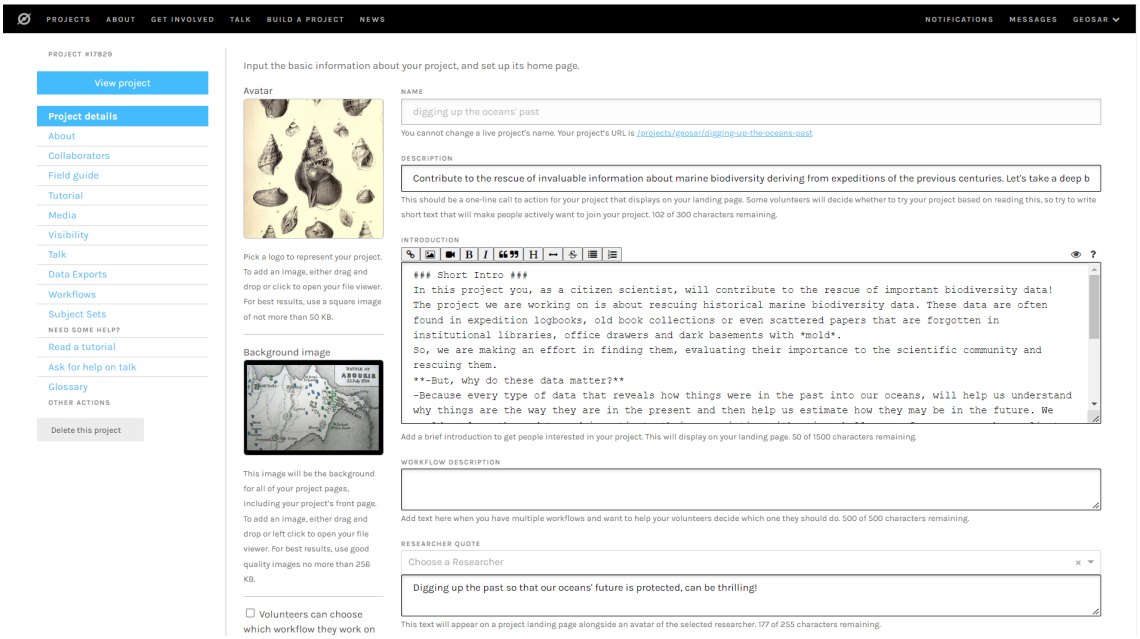


Figure 2. Screenshot of one of the tabs of the project builder from the Zooniverse platform

One of the most crucial parts of the project was the “Workflow” building. As the project builder clarifies “A workflow is the sequence of tasks that you’re asking volunteers to perform. For example, you might want to ask volunteers to answer questions about your images, or to mark features in your images, or both”. There are four options of tasks on which the workflow may be built upon (Fig. 3, left) or combinations of them. For transcription projects, the most preferable is the “Text” task, where “the volunteer writes free-form text into a dialogue box”.

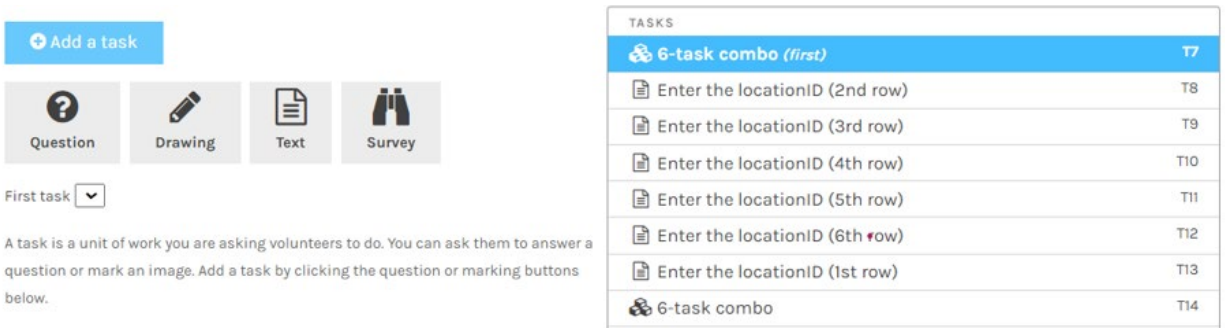


Figure 3. Options of tasks of the project's workflow (left) and combo text task selected for "dutop" (right)

General information about the project and the team were given in the tab "About". A step by step "Tutorial" was created to enhance the proper transcription and more details about the project were included in the "Field Guide". The main language was English, but there was also the option of the project translation to any other language by a translator provided by the HCMR DMT. All of the project's "Media" (e.g. photos or videos included in the field guide and tutorial) were uploaded to the corresponding tab and in continuity they were used wherever needed through Markdown markup language. The "Subject Sets" tab is where the subjects (i.e. the images) that are about to be transcribed are uploaded. The targeted keywords data rescue, historical texts, ocean, sea, marine biodiversity, plants and animals were used under the general categories of Biology, Nature and History to better describe the project.

3.2.2 Handling the dataset

The selected dataset for the project - in short "Lepadogaster" - was formatted in a table, consisting of 6 columns and 30 rows (Fig. 4). The table was divided into 5 equal parts with 6 rows and 6 columns in each part, which eventually formed a separate subject in the terms of Zooniverse (Fig. 5). This breaking down enhances the experience of a volunteer since in general smaller steps/tasks are more preferable as they are easier to handle and faster to finish (micro-volunteerism). Then, each subject had to be transcribed according to the guidelines provided in the project's tutorial.

Numéro de la Station de pêche	Date de la pêche	Situation géographique de la Station	Profondeur en mètres	Longueur du câble en mètres	Animaux capturés
289	5/IX 1904	58°44′ N 3°21′ W	95	20	1 individu
289	5/IX 1904	58°44′ N 3°21′ W	95	60	1 —
289	5/IX 1904	58°44′ N 3°21′ W	95	100	2 —
290	5/IX 1904	58°08′ N 2°24′ W	70	50	2 —
290	5/IX 1904	58°08′ N 2°24′ W	75	90	4 —
95	27/VI 1905	49°56′ N 5°00′ W	74	25	1 —
96	27/VI 1905	50°15′ N 4°19′ W	50	65	3 —
97	29/VI 1905	50°17′ N 3°14′ W	60	25	3 —
97	29/VI 1905	50°17′ N 3°14′ W	60	75	29 —
98	29/VI 1905	50°32′ N 1°05′ W	60	40	1 —
99	30/VI 1905	50°43′ N 0°43′ W	41	65	1 —
168	2/IX 1905	58°42′ N 6°13′ W	110	65	1 —
168	2/IX 1905	58°42′ N 6°13′ W	110	120	1 —
169	2/IX 1905	58°43′ N 3°30′ W	75	25	8 —
169	2/IX 1905	58°43′ N 3°30′ W	75	65	5 —
161	21/VIII 1906	51°00′ N 1°07′ E	33	20	1 —
162	21/VIII 1906	50°30′ N 0°12′ W	34	100	15 —
163	22/VIII 1906	50°21′ N 2°00′ W	49	25	2 —
163	22/VIII 1906	50°21′ N 2°00′ W	49	90	10 individus
164	23/VIII 1906	50°14′ N 4°24′ W	60	25	1 —
196	14/IX 1906	49°24′ N 3°25′ W	76	65	39 —
196	14/IX 1906	49°24′ N 3°25′ W	76	120	16 —
198	15/IX 1906	50°12′ N 0°10′ W	45	25	4 —
198	15/IX 1906	50°12′ N 0°10′ W	45	65	29 —
200	16/IX 1906	51°35′ N 2°23′ E	33	60	3 —
98	6/VIII 1908	58°48′ N 3°28′ W	90	150	5 —
62	22/II 1909	35°45′ N 5°59′ W	58	25	2 —
63	22/II 1909	35°50′ N 6°03′ W	490	600	2 —
96	23/VI 1910	35°48′ N 5°58′ W	185	65	2 —
248	29/IX 1910	49°52′ N 2°20′ W	>100	65	11 —

Figure 4. The entire “Lepadogaster” dataset as provided in the original publication.

Numéro de la Station de pêche	Date de la pêche	Situation géographique de la Station	Profondeur en mètres	Longueur du câble en mètres	Animaux capturés
168	2/IX 1905	58°42′ N 6°13′ W	110	120	1 —
169	2/IX 1905	58°43′ N 3°30′ W	75	25	8 —
169	2/IX 1905	58°43′ N 3°30′ W	75	65	5 —
161	21/VIII 1906	51°00′ N 1°07′ E	33	20	1 —
162	21/VIII 1906	50°30′ N 0°12′ W	34	100	15 —
163	22/VIII 1906	50°21′ N 2°00′ W	49	25	2 —

Figure 5. One of the 5 subjects of the project (Lepadogaster3).

One of the main steps that data managers have to follow when working with datasets is to standardize data according to Darwin Core. For this reason, a brief introduction to the Darwin Core terms was provided in the “Tutorial” and “Field Guide” sections and the column headers were presented as Darwin Core terms throughout the workflow. For example, the “Numéro de la Station de pêche'' was asked to be transcribed as “locationID” (Fig. 6, on the right). The workflow was the plain transcription of all the data and metadata given in the table. A “help task” tab was provided at the end of each task giving extra guidelines to the users. The data and metadata obtained were: location, date, coordinates, depth, sampling effort and abundance (Fig. 6).

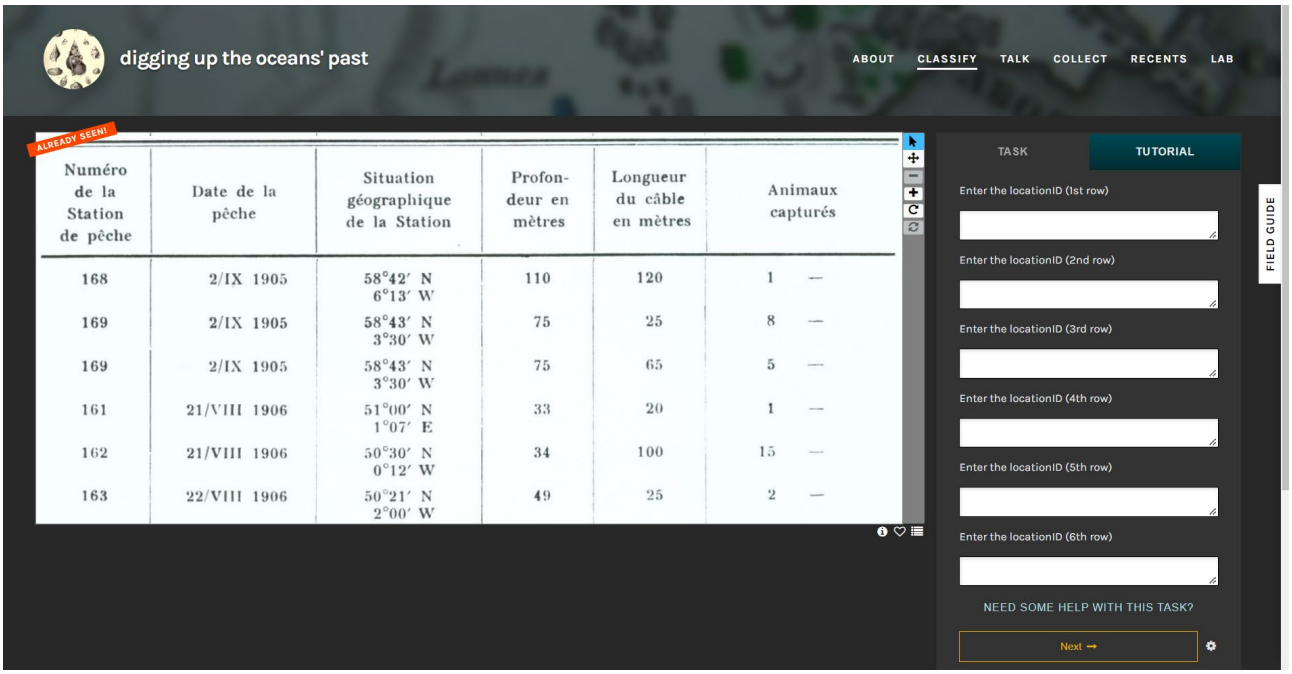


Figure 6. Digging up the oceans' past classification window screenshot with the Lepadogaster3 subject

Initially, the first versions of the workflow were based on a row by row transcription approach; each task was about one row of the table and the volunteers transcribed the data of the columns. The volunteers were prompted to format the date and coordinates according to the Darwin Core Standard. However, after receiving feedback from the "Talk" board of the platform, the approach was changed to column by column, as it was considered to be more practical to write the same data type across rows. Therefore, each column constituted a separate task, where all rows had to be transcribed in order to move to the next column (next task). In addition, the date and coordinates formatting were also removed as it was thought it discouraged volunteers and a plain transcription was opted for instead. In continuity, the workflow was tested by a number of "Testers" defined by the HCMR DMT. The feedback given (oral communication) was incorporated so as to suit the practicality of the workflow.

A project hosted in the Zooniverse platform is actually completed as soon as all of the uploaded subjects are classified as many times as the retirement limit indicates. The retirement limit is the number of people that are needed to classify one subject before it is considered complete and it could range from 1 to 100: "Once a subject has reached the retirement limit it will no longer be shown to any volunteers". The retirement limit in dutop was set initially on 10 classifications per subject but it was lowered to 5 classifications, since it was made known that in similar transcription projects the average retirement limit is usually 3 classifications per subject, even on handwritten texts which are considered a lot more difficult to transcribe. This is the reason why some subjects have more than 5 classifications; 40 in total (see Citizen Science Efforts section).

3.3 Dissemination

Several actions towards the project's dissemination took place. The dutop's URL was posted on social media, i.e. Twitter and Facebook through the EurOBIS and IMBBC (HCMR) accounts. It was also posted on the EMODnet Biology webpage, on the news section of EMODnet, OBIS and Lifewatch Greece webpages. In addition, it was included in the European Citizen Science Association (ECSA) newsletter and in the Citizen Science project finder, Sci-Starter. In the Platform Evaluation section another considerable option of dissemination for future steps is discussed.

3.4 Citizen scientist efforts

In total 40 classifications were submitted for all 5 subjects by 22 transcribers, 6 of which were subscribed to the platform and 16 not, within a timespan from 06-09-2022 to 17-10-2022 (date of first and last classification). The project was completed as soon as the retirement limit was set on 5 classifications per subject (they had already received as many by that time).

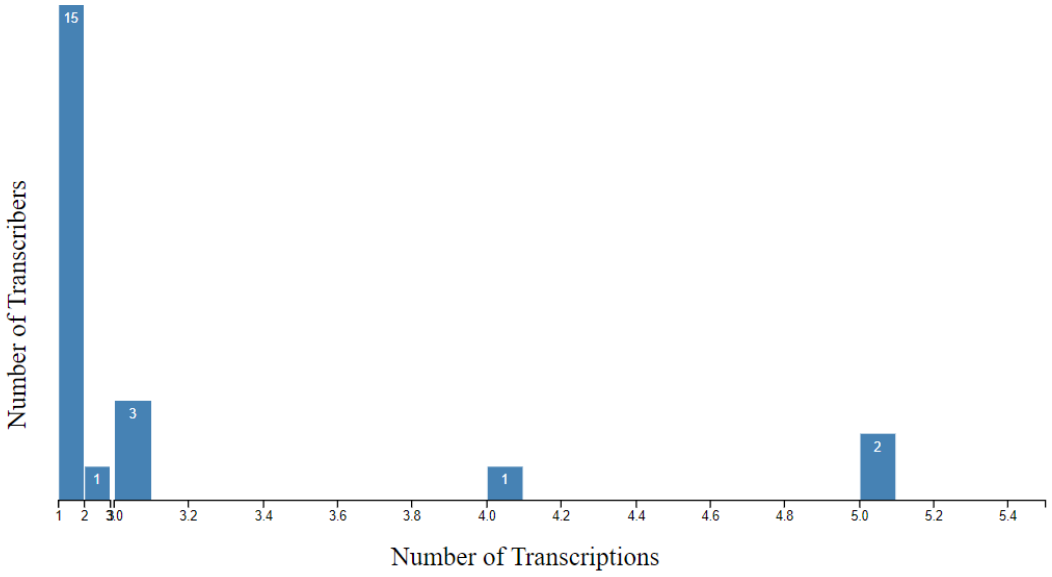


Figure 7. Number of transcriptions per transcriber number.

As shown in figure 7, the majority of the transcribers (15) contributed with one classification (one subject) while two out of 22 completed all 5 subjects. This highlights the collaborative effort of many people that may produce sufficient results. It is conceivable that other volunteers started but never finished the entire classification and these efforts were not recorded.

The timestamps that each volunteer started and finished classifying a subject were recorded, so the duration of the subject classification was calculated. The minimum duration recorded for one subject was 1 minute and 53 seconds, while the maximum was more than a day. This means that

the time invested on one subject cannot be really used as an effort estimation indicator, since the volunteer may have been interrupted/disrupted by a number of factors during their interaction with the project. Nevertheless, if the classifications of more than one day are removed (3 out of 40) the average time taken for a subject to be finished is 24 minutes and 55 seconds. It is highlighted though that this time should not be correlated with the difficulty of the task given, due to the above mentioned reasons. What is worth mentioning is that the minimum time recorded for a subject could indeed be used as an indicator of the effort per subject, for a possibly focused and uninterrupted volunteer.

In general, the workflow was considered easy and the volunteers responded really well. Most of the differentiations were observed in the coordinates task, where the volunteers used symbols other than the ones indicated ("*", "o", or space were used instead of the "°" symbol) but these were not considered as errors. Regarding the date task, there was the highest number of errors, where the "/" symbol was omitted, the "IX" was typed as "1X" or "XI", the year was missing or some numbers were typed incorrectly. On the other hand, in the same task, one volunteer converted the roman date directly to the current date system.
As far as the other tasks were concerned (location, depth, sampling effort, abundance) there was only one error ("11" was typed instead of "1"). It is interesting that in one of the depth classifications where the volunteer had to type ">100", the ">" symbol was omitted. This shows the significance of instructing the volunteers that in historical datasets every symbol matters and indicates something valuable. Finally, only 3 out of the 40 classifications seemed to be puzzling; in one "0" was typed in all fields, in the second all fields were blank and in the third only the first row of each task was completed (incorrectly). This is something common in CS projects; where volunteers, when facing a difficulty in understanding the procedure, prefer to give up efforts rather than giving a second try.

3.5 Post-processing

The export file of the classifications that is provided by the Zooniverse platform was a csv file with all the classifications attributes (e.g. user_id, time, workflow, annotations, subjects) and associated metadata included as rows. The metadata and annotations fields of the table contains all the information in key-value pairs, i.e. json format. Hence, handling this file for cleaning and aggregating the results was not friendly for a curator without programming skills. Zooniverse has an organization on the GitHub platform with repositories containing scripts and tutorials about using the Panoptes Application Programming Interface (API) and analyzing data.

Aggregation and cleaning of the data was performed with a custom script provided by a Zooniverse volunteer, Peter Mason. Additionally, a script provided here was adapted according to the "dutop" needs for the reconciliation of the data. In particular, the first step was to aggregate all answers to the corresponding subjects. Next, the cleaning included unifying the different

symbols used by the volunteers due to keyboard differences and issues mentioned above (mostly in the coordinates task). The final step was to reconcile all these answers so that a consensus was reached and the "best" option decided for each transcription. After this process, Table 2 was created, and including all the data about to be formatted and uploaded to the EMODnet Biology portal.

Table 2 Transcription reconciled results of the subject Lepadogaster3

#Row	locationID	verbatimEventDate	verbatimLatitude|verbatimLongitude	verbatimDepth	samplingEffort	individualCount
1	168	2/IX 1905	58°42´N|6°13´W	110	120	1
2	169	2/IX 1905	58°43´N|3°30´W	75	25	8
3	169	2/IX 1905	58°43´N|3°30´W	75	65	5
4	161	21/VIII 1906	51°00´N|1°07´E	33	20	1
5	162	21/VIII 1906	50°30´N|0°12´W	34	100	15
6	163	22/VIII 1906	50°21´N|2°00´W	49	25	2

3.6 Platform evaluation

As far as the Zooniverse platform is concerned, the general experience was positive. The potential it provides is obvious considering the number of engaged volunteers. There are a number of issues that should be considered though, in order to improve the overall efforts of incorporating the citizen scientists into the rescue process. As far as the visibility of the project is concerned, there are two pre-requisites demanded by the platform:

1. To include content in the FAQs section
2. To have more than 100 subjects (1 subject = 1 image)

When these pre-requisites are fulfilled, the project may then undergo an initial review by a member of the Zooniverse platform and afterwards by multiple beta testers, who are engaged Zooniverse volunteers. In this way it is ensured that the results produced are valid and sufficient for the research goals that are set by the DMT. The project could then become a "Zooniverse Project", which has a number of benefits, such as being findable by name or category through the platform's search tabs and being diffused to thousands of volunteers through the Zooniverse's newsletter. The "dutop" project did not fulfill the second pre-requisite, as it regarded a small

dataset with just 5 subjects. It is, therefore, recommended that if future efforts are undertaken in this platform, that the datasets are as large as possible. In this way, the review and improvement of the project is ensured and at the same time the workflows and tutorials that need to be created are as few as possible, saving up the data manager's time. It should be noted that there is no need in creating a new project; the new dataset may be perfectly incorporated into the already existing project.

Another point that is interesting to consider is that volunteers seem to be attracted by handwritten documents. This of course raises the difficulty status of the CS project and directly influences the quality of the results provided. Nevertheless, in the future, handwritten datasets could be also included in the project. As far as the typed documents are concerned, an alternative would be having the document go through OCR and afterwards ask the volunteers to correct the outcome.

Overall, the platform is considered more than suitable regarding the citizen science efforts in rescuing historical datasets. The HCMR DMT has indeed invested a considerable amount of time in learning how the platform works and meeting its requirements. However, in a long-term approach, this would be beneficial for the flow of several historical and rescued datasets in the EMODnet platform, as long as the abovementioned recommendations are considered.

This publication uses data generated via the Zooniverse.org platform, development of which is funded by generous support, including a Global Impact Award from Google, and by a grant from the Alfred P. Sloan Foundation.

4 DoeDat

4.1 Background

DoeDat was funded by the Flemish Government under a project called DOE! (Digitale Ontsluiting Erfgoedcollecties - Digital Access to Cultural Heritage Collections). DoeDat is all about creating data and “doe dat”, means “do that” in Dutch.

The platform is managed by the Meise Botanic Gardens and its main purpose is to help Meise in their mission to document and digitize their collections, while giving the public the possibility to take an active part in the process, contributing to making data from historical biological collections more easily accessible for a broader community of scientists and other citizens alike. The advantage of DoeDat is that it is not only suitable for the digitization of specimen collections, but it is flexible enough to also deal with other paper-based information such as historical data publications.

The usage of DoeDat has been free of charge for this trial. The Meise Botanical Garden is coordinating the DiSSCo Flanders (Distributed System of Scientific Collections) project, in which VLIZ is a partner. DiSSCo Flanders closely follows the progress of DiSSCo Prepare and aligns with the objectives of the international DiSSCo infrastructure. DiSSCo is a new world-class Research Infrastructure for the physical and digital curation of European natural science collections under common management and access policies. To make the collections more visible and used, their data and media should become more Findable, Accessible, Interoperable and Reusable (FAIR).

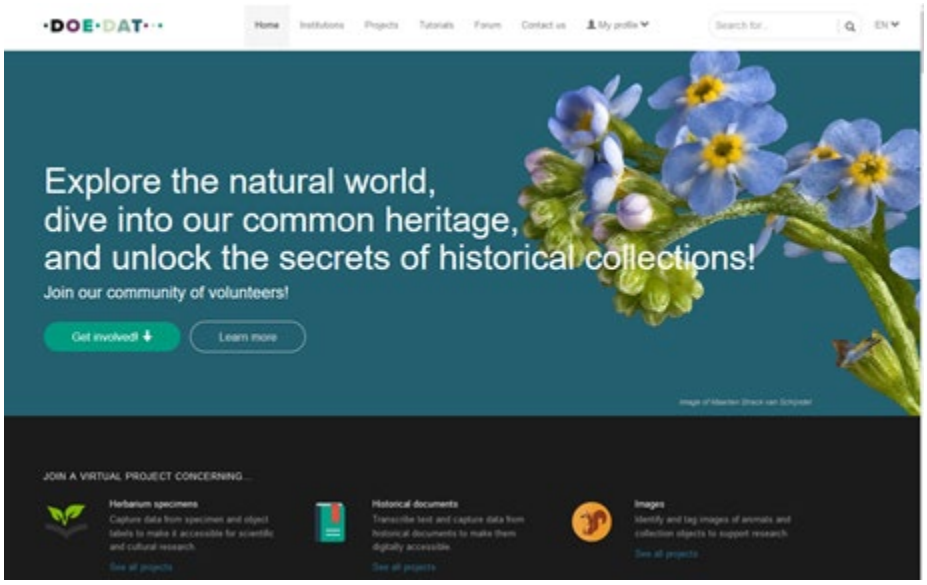


Figure 8. DoeDat platform

4.2 Preparation - “Mercator training ship expedition data” project

Several steps were undertaken to create a project on the DoeDat platform, enabling volunteers to smoothly assist in the data digitization, as well as making sure that the VLIZ DMT could transfer the dataset to the EMODnet Biology portal as effectively as possible.

Throughout the process, the DoeDat team at Meise was indispensable. They were very helpful in assisting VLIZ to get to know the details of DoeDat, as well as providing expert-advice on how to best create user manuals and communicate on this initiative, to reach as many potential volunteers as possible.

As DoeDat is managed by Meise, all information on the platform needs to adhere to the national language rules and regulations. Practically - as the working language within the EMODnet Biology project is English - this implied that all information not only needed to be available in English, but also in the three official languages within Belgium: French, German & Dutch. For translation to all four languages, the VLIZ DMT was able to rely on native speakers within the institute as well as its wider network.

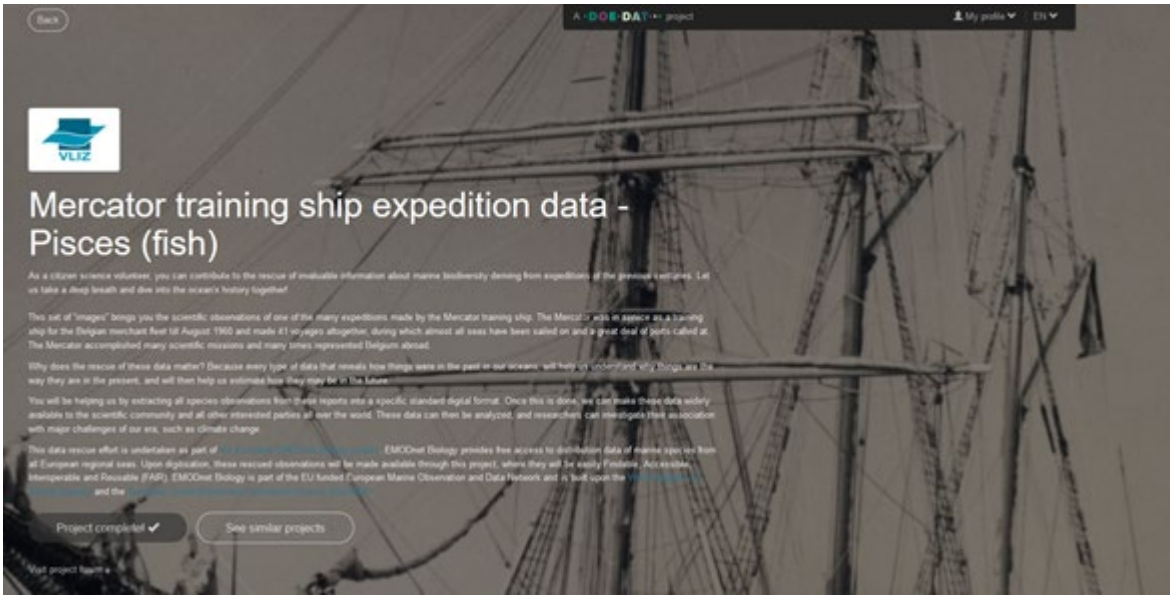


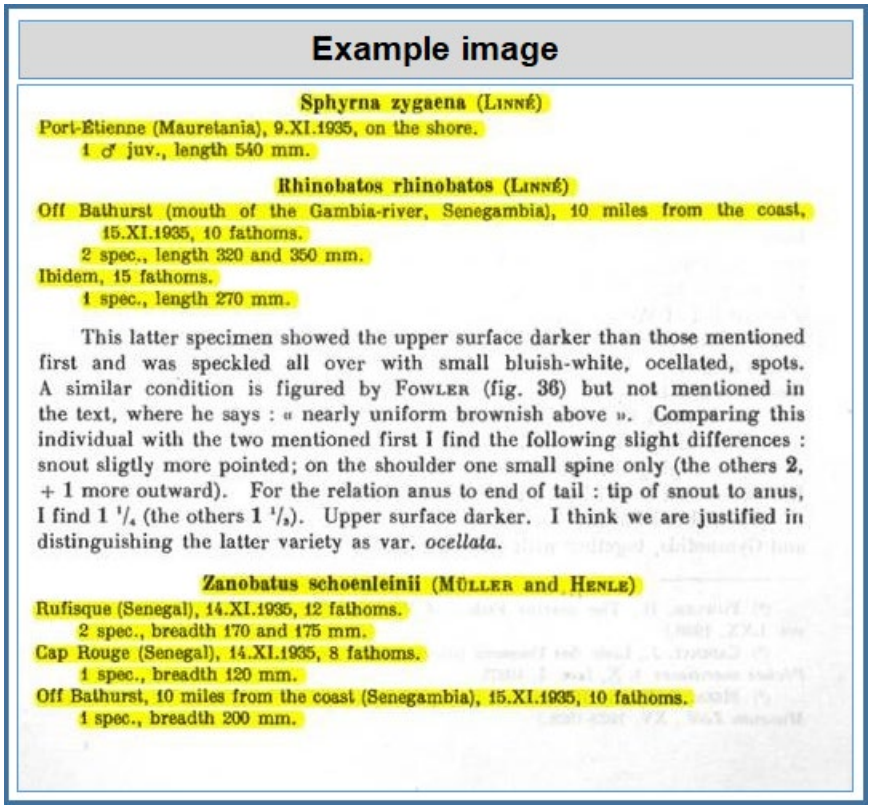
Figure 9. Entry page for the DoeDat ‘Mercator training ship expedition data’ project

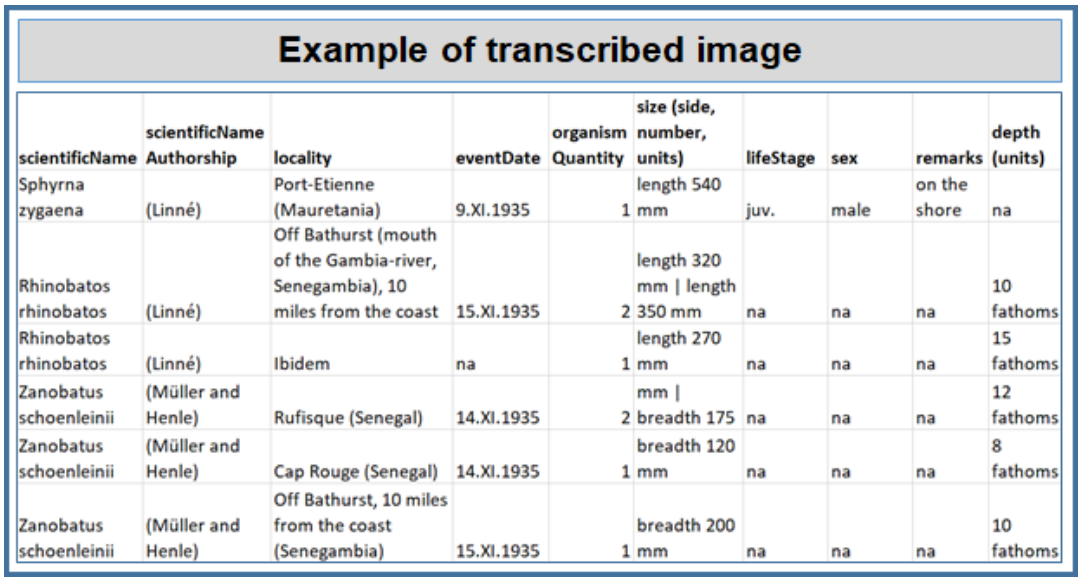
4.3 Dissemination

As soon as the project was activated on the DoeDat platform, a wide communication action was undertaken to make people aware of this, and to attract citizen scientists to work on it. As the DoeDat platform already has a vast network of volunteers, this has greatly improved the ease of communication and recruitment. The project’s URL was posted on social media, i.e. Twitter and Facebook through the EurOBIS and IMBBC (HCMR) accounts. It was also posted on the EMODnet Biology, EMODnet and on the OBIS websites.

4.4 Citizen scientist efforts

On the DoeDat platform, the "Mercator training ship expedition data - Pisces (fish)" consisted of 36 separate tasks, each corresponding to a page of the original publication. In total, four volunteers have fully transcribed the data and information from the original text to a table format, compliant with the Darwin Core format used as a data standard format for all data published in EMODnet Biology. The project itself was fully transcribed within 7 days, after which it could be downloaded by the DMT for further processing.





Example of transcribed image

scientificName	scientificName Authorship	locality	eventDate	organism Quantity	size (side, number, units)	lifeStage	sex	remarks	depth (units)
Sphyrna zygaena	(Linné)	Port-Etienne (Mauretania)	9.XI.1935	1	length 540 mm	juv.	male	on the shore	na
Rhinobatos rhinobatos	(Linné)	Off Bathurst (mouth of the Gambia-river, Senegambia), 10 miles from the coast	15.XI.1935	2	length 320 mm | length 350 mm	na	na	na	10 fathoms
Rhinobatos rhinobatos	(Linné)	Ibidem	na	1	length 270 mm	na	na	na	15 fathoms
Zanobatus schoenleinii	(Müller and Henle)	Rufisque (Senegal)	14.XI.1935	2	mm | breadth 175	na	na	na	12 fathoms
Zanobatus schoenleinii	(Müller and Henle)	Cap Rouge (Senegal)	14.XI.1935	1	breadth 120 mm	na	na	na	8 fathoms
Zanobatus schoenleinii	(Müller and Henle)	Off Bathurst, 10 miles from the coast (Senegambia)	15.XI.1935	1	breadth 200 mm	na	na	na	10 fathoms

Figure 10. Extracts from the manual created specifically for the transcription of this dataset through DoeDat. Through these examples, the volunteers were offered a clear view on what the original page looks like (top) and how it should be transcribed to a table format (bottom).

4.5 Post-processing

Upon completion of the transcription on the platform, the VLIZ DMT validated and downloaded the results and the resulting tables were run through the standard quality control and formatting steps, similar to what is done for all other datasets. The whole process took 2 days, and included small corrections in the transcribed data.

The completeness and quality of the work done by the citizen scientists was of a very high level, making it relatively easy for the DMT to bring the dataset to its final format for upload into the EurOBIS database, the data system behind EMODnet Biology.

During the August 2022 harvest, the dataset was published through the EMODnet Biology portal: https://www.emodnet-biology.eu/portal/index.php?dasid=8107. Within its metadata, the usage of the DoeDat platform was acknowledged. As instructed by DoeDat, the platform as a whole, rather than the names of the individual volunteers, was mentioned.

4.6 Platform evaluation

The VLIZ DMT evaluates the DoeDat platform as very well suited for transcription of marine biodiversity data, although its original and main focus is the transcription of specimen and object labels. The end-results of this trial were of very high quality, especially thanks to the active and enthusiastic input of, and collaboration with, the DoeDat management team.

Considering the Belgian and Meise language requirements, the effort of providing a manual in 4 languages is not straightforward. Input by native speakers was highly appreciated and needed for high-quality, clear and unambiguous guidelines for the volunteers.

Due to the nature of the platform and the way information is made available for transcription to its volunteers, it is not always straightforward to fit a publication into its layout. This aspect can potentially complicate the compilation of a clear manual for possible future transcription projects within DoeDat.

5 Summary

5.1 Direct results for EMODnet Biology

The use of the citizen science platforms Zooniverse and DoeDat has proven to be beneficial to mobilize paper-based historical data to digital form in order to be published by EMODnet Biology. It has also given us the opportunity to reach out and inform citizen scientists on the EMODnet Biology project and has allowed the contribute of people external to the project consortium. The data from the DoeDat project are already available through EMODnet Biology (Fig 11). The data from the Zooniverse project will be available in early 2023.

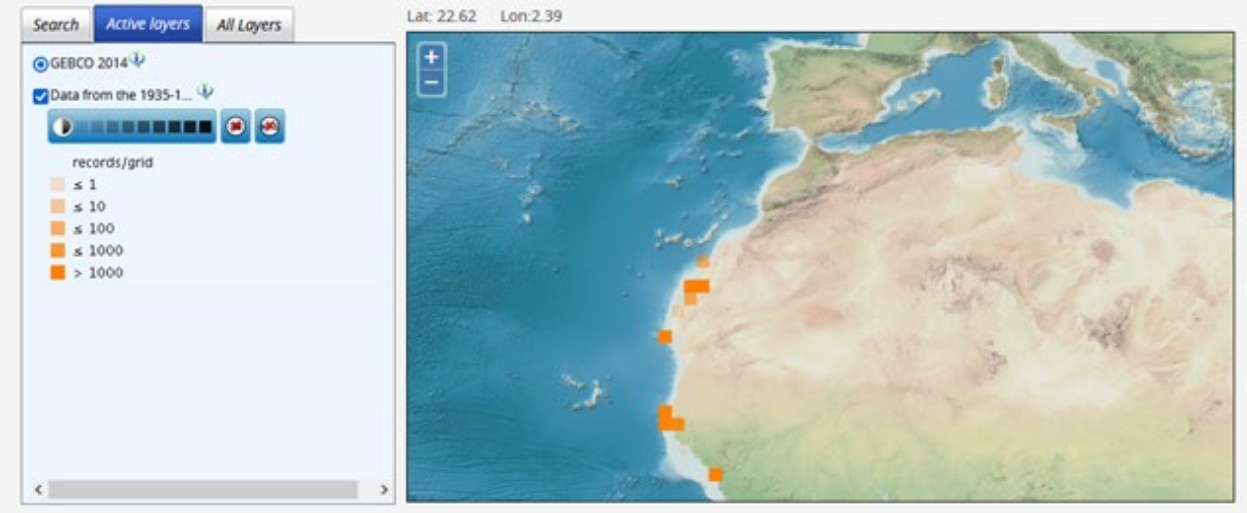


Figure 11. Data from the 1935-1936 cruises of the Belgian training ship Mercator 3 Pisces collection

5.2 Future outlook

Citizen science platforms have proven their usefulness and strength, even long before a trial was done within the EMODnet Biology project. The work described in this document focused on historical data, in need of transcription from publications, to help fill identified gaps in space and time within European marine waters.

Even though it was a successful trial, with the two datasets swiftly being digitized and published via EMODnet Biology, future activities need to be carefully considered and planned. The setup of the projects in both platforms was, as expected, quite time consuming, with roughly, 2-3 weeks of time invested by the VLIZ DMT to get the project ready on DoeDat and about 1 month of the HCMR DMT for the Zooniverse platform. This investment included getting the documents ready for transcription, creation of clear and easy-to-follow guidelines (in 4 languages for the DoeDat

platform), follow-up of volunteer questions and interaction with the platform's teams and running final checks on format and quality for publication via EMODnet Biology.

As EMODnet Biology is a European initiative, future digitization efforts via citizen science platforms, should be pursued only through the Zooniverse, mainly due to the fact that the guidelines can be published in only one language, thus decreasing the effort of the DMT in setting up a project. The experience acquired might prove useful in providing support to other partners or organisations that wish to pursue similar activities that can help address not only the gaps in (both historical and rescue) data but also engagement with the wider civil society. The Zooniverse platform, as indicated in the introduction of this report, is a very potent and interactive tool for citizen science initiatives. It should be highlighted though that datasets as big and as homogenous (structured in tables) as possible will be preferred for future efforts of EMODnet Biology. Based on work done in the previous phase of the EMODnet Biology, the number of historical datasets provided currently by the Biodiversity Heritage Library (BHL) that concern marine species and include at least 100 taxa is 1627. Emphasis should be given on evaluating these datasets -or any other potential dataset that fulfills the aforementioned criteria (size and homogeneity) - at the time of prioritizing the next phase's rescuing activities.

6 Glossary

Darwin Core: a glossary of terms intended to facilitate the sharing of information about biological diversity by providing identifiers, labels, and definitions. Darwin Core is primarily based on taxa, their occurrence in nature as documented by observations, specimens, samples, and related information.

Field Guide: a field guide is a place to store general project-specific information that volunteers will need to understand in order to complete classifications and talk about what they're seeing. It's available anywhere in your project. It's different from the tutorial in that the information is generally more about the science behind it, and is a way of sharing knowledge with your volunteers. Field guides are optional and generally contain more information than tutorials.

Project: a project is a way for the volunteer community to engage with a specific research goal or question, using data provided by the researchers. This gives the researcher data to work with and helps progress science.

Project builder: the web tool that researchers use to create Zooniverse projects. There is documentation to help with this process on the Project Builder page.

Subject: the chunk of data/thing a volunteer on a Zooniverse project is being presented with and asked to review and analyze. It typically is an image, graph, photo, audio recording, video, or a collection of these different things.

Talk: the object-orientated discussion tool associated with a project. Talk enables volunteers to comment on the subjects they've reviewed and promotes discussion amongst the volunteer community. Talk is also a place where the research team and project volunteers can interact. Talk has a series of message boards for longer discussions. Additionally, each subject has a dedicated page on its project Talk where a registered volunteer can write a comment, add searchable Twitter-like hashtags, or link multiple subjects together into groups called collections.

Task: a task could be listing how many of a particular thing a volunteer sees in an image and then drawing circles around them, identifying the various animals they can see in an image or identifying whereabouts in an image something is. There are a wide variety of tools to help create a wide variety of different tasks in the Project Builder tool. One or more tasks make up a workflow.

Testers: Testers are people who can view and classify on your project to give feedback while it's still private. They cannot access the project builder. You can add testers to a project you own through the Collaborators section of the Project Builder.

Transcription: the process of recognizing text in an image and converting it into a computer readable format.

Tutorial: a very brief walk-through explaining the main goals and aims of your project. It quickly introduces and explains to the volunteer how to do the requested tasks. This is created in the project builder and is presented to first-time volunteers of your project. Project tutorials are optional.

Volunteer or Citizen Scientist: a member of the public who is participating in and contributing to a Citizen Science project.

Workflow: a series of tasks and assessments that a volunteer is asked to do when presented with data in a Citizen Science's project classification interface. This can be either one task or multiple tasks, depending on the project.

7 Acronyms

CS: Citizen Science

DMT: Data Management Team

EurOBIS: European Node of OBIS

HCMR: Hellenic Centre for Marine Research

IMBBC: Institute of Marine Biology, Biotechnology and Aquaculture

OBIS: Ocean Biodiversity Information System

VLIZ: Vlaams Instituut voor de Zee (Flanders Marine Institute)

8 References

Mavraki D, Fanini L, Tsompanou M, Gerovasileiou V, Nikolopoulou S, Chatzinikolaou E, Plaitis W, Faulwetter S. Rescuing biogeographic legacy data: The "Thor'' Expedition, a historical oceanographic expedition to the Mediterranean Sea. Biodivers Data J. 2016 Dec 22;(4):e11054. doi: 10.3897/BDJ.4.e11054. PMID: 28174510; PMCID: PMC5267529.

9 Acknowledgments

We would like to thank Peter Mason (volunteer of the Zooniverse platform) for his help on processing the data and the full DoeDat team at the Botanical Garden of Meise for their assistance and valuable input to get our project online.

10 Addendum

The link opens the table with the complete assessment for all platforms evaluated. Platforms that only offered paid services were rejected from the very beginning, so no further evaluation was made for these.

Table 3 The entire dataset of the publication as resulted by the Zooniverse Platform

Filename	location1	location2	location3	location4	location5	location6
lepadogaster 1.jpg	289	289	289	290	290	95
slepadogaster 2.jpg	96	97	97	98	99	168
lepadogaster 3.jpg	168	169	169	161	162	163
lepadogaster 4.jpg	163	164	196	196	198	198
lepadogaster 5.jpg	200	98	62	63	96	248
lepadogaster 1.jpg	5/IX 1904	5/IX 1904	5/IX 1904	5/IX 1904	5/IX 1904	27/VI 1905
lepadogaster 2.jpg	27/VI 1905	29/VI 1905	29/VI 1905	29/VI 1905	30/VI 1905	2/IX 1905
lepadogaster 3.jpg	2/IX 1905	2/IX 1905	2/IX 1905	21/VIII 1906	21/VIII 1906	22/VIII 1906
lepadogaster 4.jpg	22/VIII 1906	23/VIII 1906	14/IX 1906	14/IX 1906	15/IX 1906	15/IX 1906
lepadogaster 5.jpg	16/IX 1906	6/VIII 1908	22/II 1909	22/II 1909	23/VI 1910	29/IX 1910
lepadogaster 1.jpg	58°44´N|3°21´W	58°44´N|3°21´W	58°44´N|3°21´W	58°08´N|2°24´W	58°08´N|2°24´W	49°56´N|5°00´W
lepadogaster 2.jpg	50°15´N|4°19´W	50°17´N|3°14´W	50°17´N|3°14´W	50°32´N|1°05´W	50°43´N|0°43´W	58°42´N|6°13´W
lepadogaster 3.jpg	58°42´N|6°13´W	58°43´N|3°30´W	58°43´N|3°30´W	51°00´N|1°07´E	50°30´N|0°12´W	50°21´N|2°00´W

Filename	location1	location2	location3	location4	location5	location6
lepadogaster 4.jpg	50°21´N|2°00´W	50°14´N|4°24´W	49°24´N|3°25´W	49°24´N|3°25´W	50°12´N|0°10´W	50°12´N|0°10´W
lepadogaster 5.jpg	51°35´N|2°23´E	58°48´N|3°28´W	35°45´N|5°59´W	35°50´N|6°03´W	35°48´N|5°58´W	49°52´N|2°20´W
lepadogaster 1.jpg	95	95	95	70	75	74
lepadogaster 2.jpg	50	60	60	60	41	110
lepadogaster 3.jpg	110	75	75	33	34	49
lepadogaster 4.jpg	49	60	76	76	45	45
lepadogaster 5.jpg	33	90	58	490	185	>100
lepadogaster 1.jpg	20	60	100	50	90	25
lepadogaster 2.jpg	65	25	75	40	65	65
lepadogaster 3.jpg	120	25	65	20	100	25
lepadogaster 4.jpg	90	25	65	120	25	65
lepadogaster 5.jpg	60	150	25	600	65	65
lepadogaster 1.jpg	1	1	2	2	4	1

Filename	location1	location2	location3	location4	location5	location6
lepadogaster 2.jpg	3	3	29	1	1	1
lepadogaster 3.jpg	1	8	5	1	15	2
lepadogaster 4.jpg	10	1	39	16	4	29
lepadogaster 5.jpg	3	5	2	2	2	11

Zooniverse data cleaning and flatten script by Peter Mason

"""This script was written in Python 3.7 "out of the box" and should run without any added packages."""
Developer Peter Mason

```
import csv
import json
import operator
import os
import re

# csv.field_size_limit(sys.maxsize)

# File location section Peter:
#directory = r'~/Downloads/Oceans_past'  # modify this to match your directory structure
location = 'digging-up-the-oceans-past-classifications.csv'
out_location = 'flatten_digging-up-the-oceans-past_classifications.csv'
sorted_location = 'flatten_digging-up-the-oceans-past_class_sorted.csv'

reg_1 = re.compile(r'\d\d?[° ]*\d\d?[´ ]*[NS][ ]?[|]?')


# Function definitions needed for any blocks.
def include(class_record):
    if int(class_record['workflow_id']) in [20922]:
        pass
    else:
        return False
    if float(class_record['workflow_version']) >= 141.0:
        pass  # replace '001.01' with first version of the workflow to include.
    else:
        return False
    return True

def clean_lat_long(text):
    clean_text = text.replace('\n', ' ').replace('  ', ' ') \
        .replace('*', '°').replace('o', '°') \
        .replace("'", '´').replace("''", '´').replace("´ ", '´') \
        .replace("I", '|')

    if reg_1.search(clean_text):
        clean_text = clean_text.replace(reg_1.search(clean_text).group(0),
                                        (reg_1.search(clean_text).group(0).replace('|', '') + '|')
                                        )
        clean_text = clean_text.replace('  ', ' ').replace(' |', '|').replace('| ', '|').replace(' ', '°')
    print(text, clean_text)
```

```
    return clean_text


def clean_count(text):
    text = text.replace(' individus', '').replace(' individu', '')
    return text

# Set up the output file structure with desired fields:
# prepare the output file and write the header
with open(out_location, 'w', newline='', encoding='utf-8') as file:
    fieldnames = ['classification_id',
                  'subject_id',
                  'user_name',
                  'workflow_id',
                  'workflow_version',
                  'Filename'
                  ]
    fieldnames.extend(['location' + str(i) for i in range(1, 7)])
    fieldnames.extend(['date' + str(i) for i in range(1, 7)])
    fieldnames.extend([r'lat/long' + str(i) for i in range(1, 7)])
    fieldnames.extend(['depth' + str(i) for i in range(1, 7)])
    fieldnames.extend(['sampleEffort' + str(i) for i in range(1, 7)])
    fieldnames.extend(['count' + str(i) for i in range(1, 7)])
    writer = csv.DictWriter(file, fieldnames=fieldnames)
    writer.writeheader()

    # this area for initializing counters, status lists and loading pick lists into memory:
    i = 0
    j = 0

    #  open the zooniverse data file using dictreader,  and load the more complex json strings as python objects
    with open(location, encoding='utf-8') as f:
        r = csv.DictReader(f)
        for row in r:
            # useful for debugging - set the number of record to process at a low number ~1000
            # if i == 500:
            #     break
            i += 1
            if i % 5000 == 0:
                print('.', end='')
            if include(row) is True:
                j += 1
                annotations = json.loads(row['annotations'])
                subject_data = json.loads(row['subject_data'])

                # this is the area the various blocks of code will be inserted to preform additional general
                # tasks, to flatten the annotations field, or test the data for various conditions.

                # pull metadata from the subject data field
                metadata = subject_data[(row['subject_ids'])]
                try:
                    filename = metadata['Filename']
                except KeyError:
```

```
    filename = ''

# reset the field variables for each new row
loc = ['' for i in range(0, 6)]  # location
date = ['' for i in range(0, 6)]  # date
lat_long = ['' for i in range(0, 6)]  # lat/long
depth = ['' for i in range(0, 6)]  # depth
sample = ['' for i in range(0, 6)]  # sampleEffort
count = ['' for i in range(0, 6)]  # location

# loop over the tasks
for task in annotations:

    # Free Transcription locations
    try:
        if task['task'] == 'T7':
            for sub_task in task['value']:
                if sub_task['task'] == 'T13':
                    loc[0] = sub_task['value']
                for i1 in range(1, 6):
                    if sub_task['task'] == 'T' + str(i1 + 7):
                        loc[i1] = sub_task['value'].replace('\n', ' ')
    except KeyError:
        pass

    # Free Transcription dates
    try:
        if task['task'] == 'T14':
            for sub_task in task['value']:
                for i2 in range(0, 6):
                    if sub_task['task'] == 'T' + str(i2 + 15):
                        date[i2] = sub_task['value'].replace('\n', ' ')
    except KeyError:
        pass

    # Free Transcription lat/long
    try:
        if task['task'] == 'T21':
            for sub_task in task['value']:
                for i3 in range(0, 6):
                    if sub_task['task'] == 'T' + str(i3 + 22):
                        lat_long[i3] = clean_lat_long(sub_task['value'])
    except KeyError:
        pass

    # Free Transcription depth
    try:
        if task['task'] == 'T28':
            for sub_task in task['value']:
                for i4 in range(0, 6):
                    if sub_task['task'] == 'T' + str(i4 + 29):
                        depth[i4] = sub_task['value'].replace('\n', ' ')
    except KeyError:
```

```
                pass

            # Free Transcription sampleEffort
            try:
                if task['task'] == 'T35':
                    for sub_task in task['value']:
                        for i5 in range(0, 6):
                            if sub_task['task'] == 'T' + str(i5 + 36):
                                sample[i5] = sub_task['value'].replace('\n', ' ')
            except KeyError:
                pass

            # Free Transcription count
            try:
                if task['task'] == 'T42':
                    for sub_task in task['value']:
                        for i6 in range(0, 6):
                            if sub_task['task'] == 'T' + str(i6 + 43):
                                count[i6] = clean_count(sub_task['value'])
            except KeyError:
                pass
        # This set up the writer to match the field names above and the variable names of their values:
        new_row = {'classification_id': row['classification_id'],
                   'subject_id': row['subject_ids'],
                   'user_name': row['user_name'],
                   'workflow_id': row['workflow_id'],
                   'workflow_version': row['workflow_version'],
                   'Filename': filename
                   }
        for j1 in range(1, 7):
            new_row['location' + str(j1)] = loc[j1 - 1]
        for j2 in range(1, 7):
            new_row['date' + str(j2)] = date[j2 - 1]
        for j3 in range(1, 7):
            new_row[r'lat/long' + str(j3)] = lat_long[j3 - 1]
        for j4 in range(1, 7):
            new_row['depth' + str(j4)] = depth[j4 - 1]
        for j5 in range(1, 7):
            new_row['sampleEffort' + str(j5)] = sample[j5 - 1]
        for j6 in range(1, 7):
            new_row['count' + str(j6)] = count[j6 - 1]
        writer.writerow(new_row)

    # This area prints some basic process info and status
    print('\n', i, 'lines read and inspected', j, 'records processed and copied')


# This section defines a sort function. Note the last parameter is the field to sort by where fields
# are numbered starting from '0'
def sort_file(input_file, output_file_sorted, field, reverse, clean):
    #  This allows a sort of the output file on a specific field.
    with open(input_file, 'r', encoding='utf-8') as in_file:
        in_put = csv.reader(in_file, dialect='excel')
```

```
        headers = in_put.__next__()
        sort = sorted(in_put, key=operator.itemgetter(field), reverse=reverse)
        with open(output_file_sorted, 'w', newline='', encoding='utf-8') as out_file:
            write_sorted = csv.writer(out_file, delimiter=',')
            write_sorted.writerow(headers)
            sort_counter = 0
            for line in sort:
                write_sorted.writerow(line)
                sort_counter += 1
    if clean:  # clean up temporary file
        try:
            os.remove(input_file)
        except OSError:
            print('temp file not found and deleted')
    return sort_counter


print(sort_file(out_location, sorted_location, 1, False, True), 'lines sorted and written')
```