Grant Agreement Number: 777483 / **Acronym:** ICEDIG
Call: H2020-INFRADEV-2017-1 / **Type of Action:** RIA
Start Date: 01 Jan 2018 / **Duration:** 27 months

REFERENCES:
MILESTONE MS28

Work package **WP 5** / Task **5.2** Lead: MNHN
Deadline M14

ICEDIG.EU *Innovation and consolidation for large scale digitisation of natural heritage*

SPECIFICATION OF DATA EXCHANGE FORMAT FOR TRANSCRIPTION PLATFORMS

MILESTONE MS28

Authors: Gwenaël Le Bras[1], Simon Chagnoux[1] and Mathias Dillen[2]

1- Muséum National d'Histoire Naturelle (Paris - France)
2- Agentschap Plantentuin Meise (Meise - Belgium)

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

ICEDIG.EU

Content

Introduction 2

Preparing data for citizen science 4

Requirements 4

Packing data in a Darwin Core Archive 5

What is a Darwin Core Archive? 5

A simple DwC exchange archive 7

Archive structure in details 8

Example 1: Data sent from the CMS 9

Example 2: Data sent from the CMS 10

In case several specimens are present on the same image 11

Structuring citizen science outputs 12

Packing data in a DwC Archive 13

Basic information 13

Common additional data 14

Optional additional data 15

Specific cases: no information or uncertain information 16

Archive structure 17

Example 3: Data sent from the CS platform 20

Example 4: Data sent from the CS platform 21

Giving feedback on transcriber activity 23

Methodology 24

1- Inventory of transcription platforms 24

2- Choice of a basic data model 24

3- Adaptation of the DwC model structure to our specific requirements 24

4- Inventory of terms currently used on the major CS platforms and DwC correspondence 25

5- Categorization of DwC terms following their use in collection 25

6- Realisation of an illustrative archive of data import to a CS platform 27

7- Realisation of a notional archive of data exportation from a CS platform 27

8- Crossing information with data quality working group 28

9- Redaction of the specification document 28

Conclusion 29

Appendix 1: The descriptor file in details 30

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

References 32

Introduction

Cataloguing specimens has been one of the core activities in Natural History collections for centuries. In modern times, databases have progressively replaced paper books. A major change happened during the last decade, during which some institutions have completed massive industrial digitization projects, producing millions of images. The pace of this process is still increasing and the planned DiSSCo infrastructure in Europe will provide the scientific community with access to a large number of specimen images and data.

These imaging techniques are not only producing more data, they also drastically changed the workflows, allowing the general public to be involved in the documentation of natural specimens. Over half a million of labels have already been transcribed by volunteers on a dozen of platforms. The ICEDIG project reviewed those platforms (MS26: *Evaluation of existing volunteer transcription systems* available online (Le Bras and Chagnoux 2018)) and concluded that there is no "best transcription platform": Dynamism of communities, although difficult to quantify, is more important than any specific feature of each website. So DiSSCo will not embed one single platform but will have to interoperate with a growing ecosystem of platforms with different publics, practices and languages.

However, the transcription requirements for specimens are similar enough amongst institutions to agree on a common protocol to exchange data between their databases and transcription platforms. The present document is proposing such a protocol.

We aimed at writing this report in such a way that it can be useful today for both platform administrators and less-technical collection managers, but still generic enough to build DiSSCo services upon it. We have tried not to invent anything new and stick as much as we could to Biodiversity Information Standards (https://www.tdwg.org/). We hope that within the next few years most platforms will implement this specification and become "DiSSCo compatible"

The protocol covers two data flows (Figure 1):

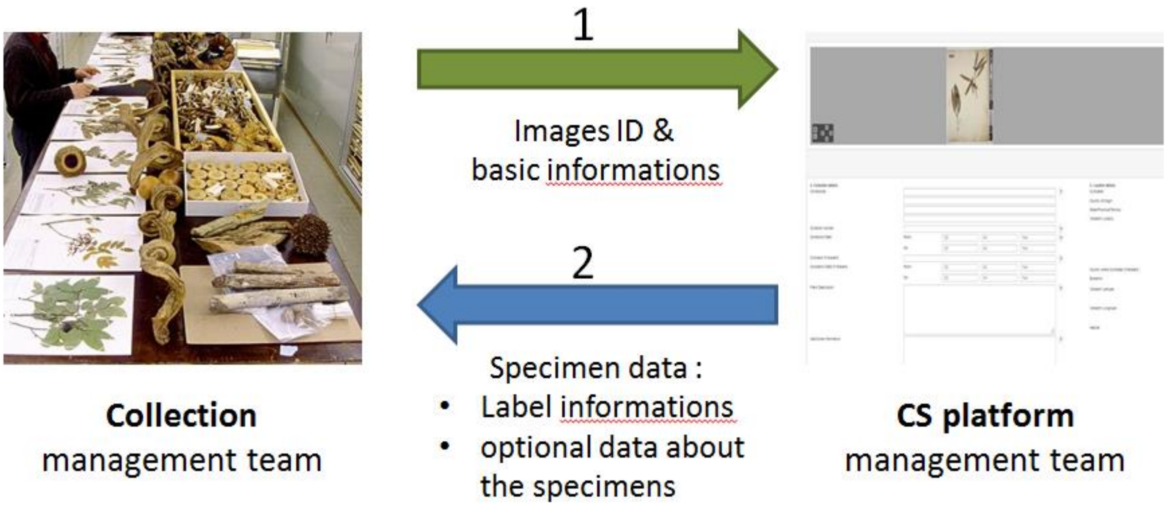

Figure 1: Basic mapping of the Data flow over a citizen science project

The first one described in the chapter “Preparing data for citizen science” offers a simple way to send a set of images and basic information on them from a collection database to a transcription platform.

The second flow delivers transcribed data back to the collections. Detailed in chapter “Structuring citizen science outputs”, it addresses issues shared with ICEDIG Task 4.3 (Interoperability with Collection Management Systems (CMS)), which we have been closely working with despite our more prescriptive approach.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Preparing data for citizen science

Requirements

Planning a citizen science (CS) project requires several steps that will condition the data sent:

- **Have a subject for the project.** One has to be able to explain clearly to the public why this project is being held. This subject can be either a scientific purpose (i.e. the study of a particular Genus, or the flora from a particular area for instance), or a mission about beautiful specimens for the pleasure of the users. It needs in any case to be explainable clearly. A transcription pilot to test data quality held on several different platforms, made us conclude that true motivations of a mission can't be hidden from the users (otherwise they start to imagine what these motivations are). Further results from this transcription pilot (pilot 2) can be found in the report online (Phillips et al. 2019)
- **Select a platform.** Each platform has their own requirements (i.e. number of specimens per project, project description files and images, image format). This information has to be decided before starting to design the project. To help select the most appropriate CS platform to collection holding institutions needs, an *evaluation of existing volunteer transcription systems* was realized in the frame of ICEDIG work package 5.2 on citizen science transcription platform. It can be found online (Le Bras and Chagnoux 2018).
- **A unique standardized way to identify the specimens (catalog number) is used for the collection, and the collection itself is identified.**
- **Every specimen to be part of the project has been imaged.** One to several images can be done for each specimen, depending on particular needs.
- **Labels are clearly readable at least on one image per specimen.** As the information to be transcribed is the target of the project, the labels are a basic requirement. Unless the specimen is of very little interest to the general audience (i.e. dry fungi), the specimen needs to be clearly visible as well on at least one image per specimen in order to keep the user's interest (Le Bras and Chagnoux 2018).
- **Images of the specimens to be included in the projects are available online at a specific Uniform Resource Identifier (URI)**. In order to alleviate the data exchange, the images will not be transferred with the archive. The receiving system will retrieve them from the network. To do so, the images have to be available through a dedicated service, and available at a distinct URI. Images should be available online at least for the duration of the project, and preferentially permanently. For the institutions not able to maintain such a permanent service, ICEDIG will provide an evaluation of Zenodo infrastructure (deliverable D6.3 of the subtask 6.3.3 dedicated to the Zenodo infrastructure and due for end of july 2019). This long-term repository offers the possibility to host online in case no institutional server is available.
- **Licencing for the images should be established** prior to the publishing of images on a third-party CS platform. Any published licence can be referenced here. We suggest to follow, as much as possible the GBIF data licencing terms (https://www.gbif.org/terms). The choice is given to the institutions to choose between the following licences of creative commons (https://creativecommons.org/):
    - CC0: https://creativecommons.org/publicdomain/zero/1.0/
    - CC BY: https://creativecommons.org/licenses/by/4.0/

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

  - CC BY-NC: https://creativecommons.org/licenses/by-nc/4.0/

  The citation of the licencing on the extract has to be done using the URL redirecting to the full description of the licence terms on the creative commons' website. The version we cite here is the latest to date, yet we invite each institute to check for newer versions while deciding their licencing policy.

- **A taxon or scientific name should correspond to each specimen to be part of the project.** This name can be the name the specimen is filed under. In case the specimen is not determined to species level, it is possible to indicate the lowest taxon rank it was determined to.

Within the frame of its WP6, ICEDIG is categorizing levels of minimum information standards for digital specimens (MIDS). These standards should be published soon. The requirements of our digital specimen to be able to go through a citizen science project is a MIDS level 1 with options.

Once these requirements are met, the main activity of preparing the project will be to select the images to transcribe, and to write a description of the project. The images and data should then be structured prior to be sent. The next chapter describes that structure.

The procedure hereunder described is as simple as possible, so that any collection manager will be able to follow it. Consequently, no particular IT knowledge is required, other than being able to realise an extract from the local collection management system, and to process it with text editing software and/or spreadsheet tools.

Packing data in a Darwin Core Archive

To exchange data, we will pack them into an archive based on Darwin Core Archive.

What is a Darwin Core Archive?

A Darwin Core Archive is a biodiversity dataset using a list of standardized terms named Darwin Core (DwC). It is widely used to exchange data about species occurrence, taxon checklists, sampling events or collection specimen data. Created between 1998 and 2009 by the Taxonomic Databases Working Group (https://www.tdwg.org/), it has become a major data standard used for many of the main biodiversity science projects such as the Global Biodiversity Information Facility (https://www.gbif.org/) or the Encyclopedia of Life (https://www.eol.org/). Every institution committing data to these projects are regularly producing DwC archives.

A DwC Archive is a simple dataset easy to read on every computer. It is often a simplification of an existing complex database, and is made to share data between different databases. Created to ease biodiversity data exchange, it has become over its years of use a stable and strong data standard.

DwC Archive is the most appropriate way to exchange data in our situation.

Figure 2 depicts a common DwC archive. It is based on a collection data sharing standard, used to share data about collection specimens. More precisely, it is the scheme of DwC archive-building used by the MNHN to share data about the collection it holds to the GBIF.

A DwC archive constitutes a .zip folder containing two types of files:

- .xml files. These are the descriptor files. Descriptors show the metadata of the dataset. There are two of them:
  - meta.xml describes the structure of the dataset itself for a proper reading of the data by a computer. It mentions the encoding, the delimitation character of the values in the data files, the field order, gives links to definitions of the DwC terms used, etc.
  - eml.xml. This file gives a description for human reading of the content of the dataset, in order for users to be able to make good use of the data. It will give information about the persons in charge of managing the data, a text describing the dataset, date of constitution of the dataset. Also, in the case of a specimen collection, the area the specimen originates from, a date (range)when the specimens were collected, etc.
- text files (.txt). These are the data files. They gather the information of the dataset in a separated values file. Usually, this is a tab separated value file, but it can also be comma separated or semicolon separated.
  - A DwC archive contain at least one data file: the core. On the Figure 2 example, the core datafile is “occurrence.txt”, and it contains the basic information about the specimen itself and its collection event. Each row within this file has a unique identifier, which works as its primary key.
  - Usually, the core datafile is accompanied by one or several additional data files. On Figure 2,there are 4 of them: Reference, Identification, Identifier and Multimedia. Within these additional data files, each line refers to one specific line in the core file through this last identifier (secondary key).

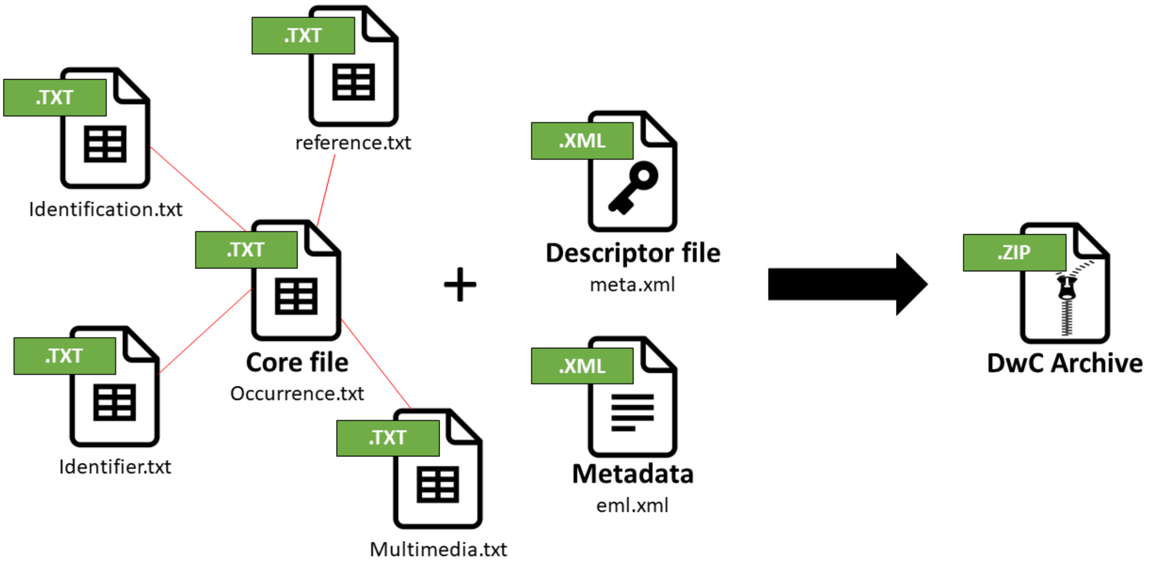

Figure 2: Common Darwin Core Archive composition

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

A simple DwC exchange archive

Data transferred to a CS platform have to be altogether easy to prepare by the collection holder, easy to process by the platform managing team and contain all the required information to set up a CS transcription project. In order to ease the creation of the mission, the data included have to be as complete as possible. We will use here a simplified DwC archive package with only two files:

- **a core datafile** based on the images information
- **a descriptor xml**

No need here to set an eml file, as it will mostly contain mission characteristics that could be sent in an easier way to the platform by mail or informal discussion. As each platform has its own sets of attributes for a transcription project, fixing arbitrary fields for a single record is an unnecessary burden.

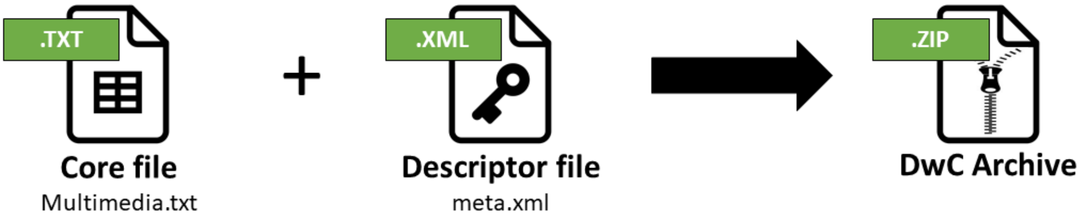

Figure 3: Simplified DwC based Archive for data input on a CS platform from a CMS

The information to include in the core file will be:

- **The unique identifier of the image/media** (*typically URI* - DwC: associatedMedia**)**. This is the URI from where the platform management can get the image from. It is preferably a permalink, but in case no permalinks are available, it should be valid for at least the duration of the CS project.
- **The unique identifier of the specimen** (DwC: occurrenceID). This should ideally be a permalink. These identifiers are to be used to link the newly produced data to the correct specimens. In the case of several images per specimen, this information will allow the CS management team to link all the images of the specimen to the correct entry. The Consortium of European Taxonomic Facilities (CETaF) has worked on unique identifiers in order to help institutions to set their own. More information about this is available on the CETAF website through a poster (Güntsch et al. n.d.) or on the dedicated wiki (https://cetafidentifiers.biowikifarm.net/wiki/Main_Page).
- **The media type** (*usually stillImage* - DwC: type). This is to be useful for the CS platform, as the media has to be handled differently depending on this type.

- **The media format** (DwC: format). CS platforms usually need the input images to be in a certain format. Describing here the format of the image stored and available on the URI will allow them to convert the images if necessary.
- **The licence under which the media should be used** (DwC: licence). This is the licencing chosen as described above.
- **The code of the institution** holding the specimen (DwC: institutionCode).
- **The collection code** the specimen is part of (DwC: collectionCode)
- **The catalog number of the specimen** within the collection the specimen is part of (DwC: catalogNumber).
- **A scientific name** corresponding to the specimen, e.g.: the one the specimen is stored under (DwC: scientificName). In case no taxon name can be linked to the specimen (i.e. a non-determined specimen), a value should be included to state clearly that the absence of data is not due to an informatic technical issue, but rather a determination issue. For instance, the MNHN chose “Insertae sedis” in our data example. The meaning of this value can be easily found by the CS user on the web.

Archive structure in details

The archive to send data to a CS platform is constituted of two files, meta.xml and multimedia.txt, as depicted above (Figure 3). To help picture this archive structure, we built up an illustrative archive (Le Bras 2019a) displaying the specimens used for the trans-institutional and trans-platform pilot project held in the frame of ICEDIG (pilot 2 on the relevant online report (Phillips et al. 2019)). It is possible to re-use the descriptor file (meta.xml) from the example by respecting the following rules:

- Encoding of the files has to be UTF8. This encoding format allows for correct coding of most world language characters, and is supported on most machines. Consequently, it is the most appropriate encoding system for our purpose.
- Files have to be delimited by tabs (\t)
- Lines have to be delimited by the line feed character (\n)
- No field enclosure characters will be used
- Headers have to be one single row
- If needed, new columns should be added at the end

For more details about the descriptor file structure, a description is provided in Appendix 1: The descriptor file in details.

multimedia.txt: This file contains the actual data from our archive. In our example, this file is a tab separated value file, created by doing a copy/paste from spreadsheet software into a text editor software. As mentioned above, the first column (id) is a copy of the second one (associatedMedia). This first column is the primary key of our data (it is then no actual data). The remaining 9 columns are containing relevant data described above. Each column corresponds to a <field/> entry in the data.xml file:

- associatedMedia
- occurrenceID

- type
- format
- licence
- institutionCode
- collectionCode
- catalogNumber
- scientificName

Example 1: Data sent from the CMS

Case for the simple specimen P03558024

For readability reasons, we here framed our fields with quotation marks (“). A tabular version of this example is available online (https://doi.org/10.5281/zenodo.2579686).

```
ID = "http://mediaphoto.mnhn.fr/media/1441365719281fbgNH3QOftOJIz09"
associatedMedia = "http://mediaphoto.mnhn.fr/media/1441365719281fbgNH3QOftOJIz09"
occurrenceID = "http://coldb.mnhn.fr/catalognumber/mnhn/p/p03558024"
type = "StillImage"
format = "image/jpeg"
licence = "http://creativecommons.org/licenses/by/4.0/legalcode"
institutionCode = "MNHN"
collectionCode = "P"
catalogNumber = "P03558024"
scientificName = "Castanopsis acuminatissima (Blume) A.DC."
```

Example 2: Data sent from the CMS

Case for the multi-imaged vertebrate specimen MNHN-ZO-2013-152

For readability reasons, we here framed our fields with quotation marks (“). A tabular version of this example is available online (https://doi.org/10.5281/zenodo.2579738).

This specimen of razorbill has 5 images recorded in GBIF. If it was to be sent to a CS transcription platform, there would be 5 lines describing it in the relevant multimedia datafile filled as follows. This is a common case for zoological or paleontological specimens. In our file, a line in the document should correspond to each image sent, and the same information needs to be repeated for occurrenceID, institutionCode, collectionCode, catalogNumber and scientificName. This will allow the CS platform to link several images to the same specimen.

ID = "http://mediaphoto.mnhn.fr/media/1432022935007Ijp7LVEZylb7BUyF"
associatedMedia = "http://mediaphoto.mnhn.fr/media/1432022935007Ijp7LVEZylb7BUyF"
occurrenceID = "http://coldb.mnhn.fr/catalognumber/mnhn/zo/2013-152"
type = "StillImage"
format = "image/jpeg"
licence = "http://creativecommons.org/licenses/by/4.0/legalcode"
institutionCode = "MNHN"
collectionCode = "ZO"
catalogNumber = "2013-152"
scientificName = "Alca torda Linnaeus, 1758"

ID = "http://mediaphoto.mnhn.fr/media/1432022936311Lia8CCKdSuOY52v7"
associatedMedia = "http://mediaphoto.mnhn.fr/media/1432022936311Lia8CCKdSuOY52v7"
occurrenceID = "http://coldb.mnhn.fr/catalognumber/mnhn/zo/2013-152"
type = "StillImage"
format = "image/jpeg"
licence = "http://creativecommons.org/licenses/by/4.0/legalcode"
institutionCode = "MNHN"
collectionCode = "ZO"
catalogNumber = "2013-152"
scentificName = "Alca torda Linnaeus, 1758"

ID = "http://mediaphoto.mnhn.fr/media/1432022937102nn5mviWGcYEw5eln"
associatedMedia = "http://mediaphoto.mnhn.fr/media/1432022937102nn5mviWGcYEw5eln"
occurrenceID = "http://coldb.mnhn.fr/catalognumber/mnhn/zo/2013-152"
type = "StillImage"
format = "image/jpeg"
licence = "http://creativecommons.org/licenses/by/4.0/legalcode"
institutionCode = "MNHN"
collectionCode = "ZO"
catalogNumber = "2013-152"

```
scentificName = "Alca torda Linnaeus, 1758"

ID = "http://mediaphoto.mnhn.fr/media/14320229378493SF5trxl8WGLFJG1"
associatedMedia = "http://mediaphoto.mnhn.fr/media/14320229378493SF5trxl8WGLFJG1"
occurrenceID = "http://coldb.mnhn.fr/catalognumber/mnhn/zo/2013-152"
type = "StillImage"
format = "image/jpeg"
licence = "http://creativecommons.org/licenses/by/4.0/legalcode"
institutionCode = "MNHN"
collectionCode = "ZO"
catalogNumber = "2013-152"
scentificName = "Alca torda Linnaeus, 1758"

ID = "http://mediaphoto.mnhn.fr/media/143202296244433EpMw3CYLIHKp9l"
associatedMedia = "http://mediaphoto.mnhn.fr/media/143202296244433EpMw3CYLIHKp9l"
occurrenceID = "http://coldb.mnhn.fr/catalognumber/mnhn/zo/2013-152"
type = "StillImage"
format = "image/jpeg"
licence = "http://creativecommons.org/licenses/by/4.0/legalcode"
institutionCode = "MNHN"
collectionCode = "ZO"
catalogNumber = "2013-152"
scentificName = "Alca torda Linnaeus, 1758"
```

In case several specimens are present on the same image

This case can occur for paleontology specimens or herbarium sheets with small specimens attached together on the same sheet. In this case, it is important the specimen can be easily identifiable, and that each one can be clearly linked to its catalog number. The image has then to be duplicated for each specimen recognised on it. One image URI can correspond only to one specimen. Consequently if 3 specimens are present on one image, the image has to be duplicated into three images, each getting a distinct URI which will be linked to a distinct specimen.

Although not impossible, this is a tricky situation even for professional digitising teams, so we suggest to avoid as much as possible such complicated cases in CS transcription project.

An alternative to image duplication could have been to handle the id with distinct value from the associatedMedia.

Structuring citizen science outputs

Once the transcription project is completed, the data needs to be sent back to the curating institution facility. To do so, we will use again a DwC archive. This time our archive will be centred on the specimen information (occurrence for DwC) as depicted in Figure 4. As more data are linked to the specimen, this part is a bit more technical, and requires a better understanding of the DwC archive used and biodiversity databases. However, it is important for both sides of the exchange to understand how it is constituted. The descriptions below address both the collection manager and platform operator. For the collection manager, this document will help him understand how it has been built, in order for him to understand how to treat the archive. For the platform operator, this document offers precision on the DwC terms used and the format of the data stored under DwC terms that needs to be followed.

This step is usually done by the platform operator. He should prepare an export that is conform to the present specification. Of course, automatization of that export by a compiling script or even as a feature of the web application will be the ideal target. The later integration of the data into a collection management system usually necessitates formatting the data to the schema of the systems. The format hereunder was developed in order for this part to be as easy as possible, but it still requires databasing skills. The ideal solution will be to have an import feature in most CMSs based on the described protocol. But at an initial stage, semi-automated methods such as SQL scripts or Open refine (http://openrefine.org/) manipulation can be used.

The format we describe below applies for most of the information included in actual transcription projects. We did have to make compromises in some cases to match platform and DwC characteristics, which are detailed in the methodology chapter.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Packing data in a DwC Archive

Prior to packing the data into an archive, the data should be formatted to fit a common standard define below. We categorised the field/terms used for describing a specimen in function of their use. We defined 3 categories as follow (cf. Methodology):

1. **The basic information**: these are the terms giving the most important information about a specimen. They are answers to the questions where/when/what/by who. They constitute the information that will be most commonly used to describe the specimen, cite it in literature and find it through search engines. These information fields are the firsts ones asked for from citizen scientists. Consequently, these values have to be transcribed if available on the labels. A digitisation cannot be considered complete if one of these fields are left blank (cf. below for the cases with no information available).
2. **The common additional data:** These are the information fields precising the previous ones. As such, they will be used as a complement for basic queries on a search engine.
3. **The optional additional data**: These are the information fields used for specific research projects or fields.

We used here the DwC terms to present them. All these terms can be found online at http://rs.tdwg.org/dwc/ .

Basic information

These data are the very basic ones describing a specimen. They are the ones most commonly searched for in a database, and the ones used to describe the specimen in literature. As such these fields are mandatory in a CS project output (note that the collection number is mandatory only for botany).

- **institutionCode:** the code of the institution holding the specimen.
- **collectionCode:** the code of the collection the specimen is part of.
- **catalogNumber:** the catalogue number of the specimen within the collection the specimen is part of.
- **recordedBy:** the name of the individual(s) collecting/capturing the specimen. The name should be in the format “Name, I.”. This format is the most commonly used in existing databases (cf. Methodology). In case several people are mentioned as collector/capturer of the specimen, the names should be transcribed in the same order as on the specimen label, and separated by a “;”. Example: “Bonpland, A.J.A.; von Humboldt, F.W.H.A.”. There is a work in progress to propose standards for assigning unique IDs to people, but more time will be needed before platform implementation and validation by TDWG.
- **fieldNumber:** the field number attribute to the specimen collection event. It is the number given to the collecting event. This field doesn’t apply to all collection specimen. It is a particularity to botany. However, in botany, it is a mandatory field (collection number). In this field, do not use space characters.
- **eventDate:** The date format will follow the norm ISO 8601 (https://en.wikipedia.org/wiki/ISO_8601) to format the dates, in order for most systems to

correctly interpret the given information. The date precision will be to the day (YYYY-MM-DD). In case the collection date is a date range, the information will be put in as described by the ISO norm, with a starting date and an ending date. In case the collection/capture date mentioned is not precise to the day (just a month/year, or just a year), it will be coded as a range. Examples:

  - A specimen collected/captured on the 5th of december 2018: eventDate= “2018-12-05”
  - A specimen collected/captured in December 2018 (no more precision): eventDate: “2018-12-01/2018-12-31”
  - A specimen collected/captured during a mission between the 5th of march 1865 and the 23rd march 1865 (no more precision): eventDate= “1865-03-05/1865-03-23”.

- **countryCode:** the code of the country where the collection took place. Use will be made of the norm ISO 3166-1 alpha-2 (https://en.wikipedia.org/wiki/ISO_3166-1), which is broadly used and interpretable by most systems.
- **country:** The name of the country in full, for easy human reading of the dataset. **scientificName:** The taxon name linked to the specimen. On some platforms, other names can be added to the specimen data. It should then fit into this same field in a different identification row (see below).
- **verbatimLocality:** a verbatim text name of the locality.

Common additional data

- **stateProvince:** the first level of administrative layer below the country one. If possible in a formatted way. Use will be made of the norm ISO 3166-2 (https://en.wikipedia.org/wiki/ISO_3166-2), as it is the most complete referential available.
- **county:** the level of administrative layer below the stateProvince. If possible, in a formatted way.
- **decimalLatitude/decimalLongitude:** The latitude and longitude coordinate where the collection/capture took place, in decimal format. This field is the one that should be used to store mapping application outcomes.
- **coordinatePrecision:** The coordinate precision when the mapping application allows to produce an incertitude.
- **verbatimCoordinates:** Full text data can be included in this field. This field will contain the information manually transcribed by the CS user (i.e. transcription of coordinates present on a sheet).
- **minimumElevationInMeters/maximumElevationInMeters:** The minimum/maximum elevation where the collection/capture took place, in meters (conversion into this unit accepted by the International System of Units has to be made in the case the sheet mentions elevation in feet).
- **verbatimElevation:** For the case no elevation in meters are available, or the relevant elevation fields in the CS database are not formatted to receive information only in meter, it is possible to include this field in order not to lose the information.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

- **establishmentMeans**: The process by which the biological individual(s) represented in the occurrence became established at the location. Use needs to be made of controlled vocabulary here (managed, native, invasive, introduced).
- **identifiedBy:** the name of the scientist who named the specimen. This field will be formatted like recordedBy.
- **dateIdentified:** The date the specimen was identified. As for the eventDate, use will be made of the norm ISO 8601 (https://en.wikipedia.org/wiki/ISO_8601) to format the dates. In case the date precision is not to the day, but to the month or to the year, no interval will be used here, but respectively the YYYY-MM or the YYYY format.
- **modified:** This field records when the data was last modified (automatically implemented by the platform application). As for dateIdentified and eventDate, use will be made of the norm ISO 8601 (https://en.wikipedia.org/wiki/ISO_8601).

Optional additional data

- **habitat:** a verbatim rendition of the habitat the specimen was collected/captured from.
- **occurrenceRemarks:** a verbatim field to gather all information relative to the specimen that cannot fit in the previous fields, such as conservation state. This non-format field can contain loads of different information and be difficult to exploit.
- **organismRemarks:** a verbatim field to gather all information relative to the organism that cannot fit in the previous fields, such as a morphologic particularity. This non-format field can contain loads of different information and be difficult to exploit.
- **taxonRemarks:** a verbatim field to gather all information relative to the taxon that cannot fit in the previous fields, such as the use of this taxon for traditional medical purpose, or food in general. This non-format field can contain loads of different information and be difficult to exploit.
- **Multifield scientific name:** For some uses, it is necessary to have the scientific name split into taxonomical ranks. Although the name should be concatenated into scientific name, it is possible to use the following split. Although not recommended here, if agreed upon between the collection management team and the CS platform, it is also possible to include the rank level above the genus one.
  - **genus**
  - **specificEpithet**
  - **taxonRank:** in case of infraspecificEpithet, this field has to be used to precise the taxon rank of the infraspecific names given in infraspecificEpithet (example “subsp.” or “var.”).
  - **infraspecificEpithet:** this field has to be linked to the presence of data in taxonRank.
  - **scientificNameAuthorship:** scientific authority that described the considered taxon, formatted following the relevant code (botanical (Turland et al. 2018) or zoological (International Commission on Zoological Nomenclature 1999)). For botany, this should follow the IPNI abbreviations based on Brummit works (http://www.ipni.org/ipni/authorsearchpage.do).
- **Geological context:** a paleontological specimen CS project would need to include terms about the geological context the specimen was collected from. Depending on the questions asked

from the citizen scientists, a range of specific terms are available on the TDWG wiki (http://rs.tdwg.org/dwc/terms/GeologicalContext).

- **vernacularName:** a vernacular name of the taxon if available.
- **otherCatalogNumbers:** any other catalogue number than the official one, or number associated to a subcollection the specimen is part of, for example.

Specific cases: no information or uncertain information

The absence of information in a field can mean several things:

- nobody transcribed data relevant to this specific field
- no relevant information is available on the specimen itself

These two cases should be different in their resolution. To resolve the first one, the specimen can be included in a new mission with this specific question asked. In the second case however, putting the specimen through another CS project is pointless. Transcription platforms often include a check-box in case no data are available on the labels. When no data are present on the label (that's to say, when a CS volunteer checked the "no data" box), use of the code *n/a* has to be made in the relevant field of the DwC archive.

Some CS platforms give their users the possibility to quote a checkbox "I'm not sure" in order to express their uncertainty on some transcription made. To record this in the data archive, a question mark in square brackets will be added to the end of the relevant field character chain (" [?]").

Example:

on the platform:

collector = "von Humboldt, F.W.H.A."
not_sure_checkbox = true

in the DwC archive:

recordedBy = "von Humboldt, F.W.H.A. [?]"

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Archive structure

As the data the output archive contains describes a specimen, and not only the images associated to it, it will be anchored on the occurrence and this time have several associated files:

- **identification.txt**: the CS project may include the transcription of other determinations than the scientific name from the input file. Consequently, the identification data have to be in a separated data file.
- **multimedia.txt:** this data file could be produced even if the initial data were already including them. This to facilitate the work of troubleshooting in case of issues with the data.
- in case of specific projects, such as measurements of a specimen or segmentation of an image together with transcription of the labels on it, it is possible to include tables as *measurements.txt* or *segmentation.txt* to compile relevant data. These cases are not developed further here, as we concentrated on the transcription process. To make use of them, both sides exchanging data will have to agree on terms. Terms for measurements can be found on the relevant page of the TDWG website (https://terms.tdwg.org/wiki/Darwin_Core_Measurement_or_Fact). DwC allows for the creation of new terms if needed, after agreement by both exchanging sides, such as those that one can imagine in *segmentation.txt* to transmit information about an image segmentation project.

The described archive here will be constituted of a core data file: occurrence.txt, two data tables called identification.txt and multimedia.txt, and meta.xml containing the machine-readable metadata.

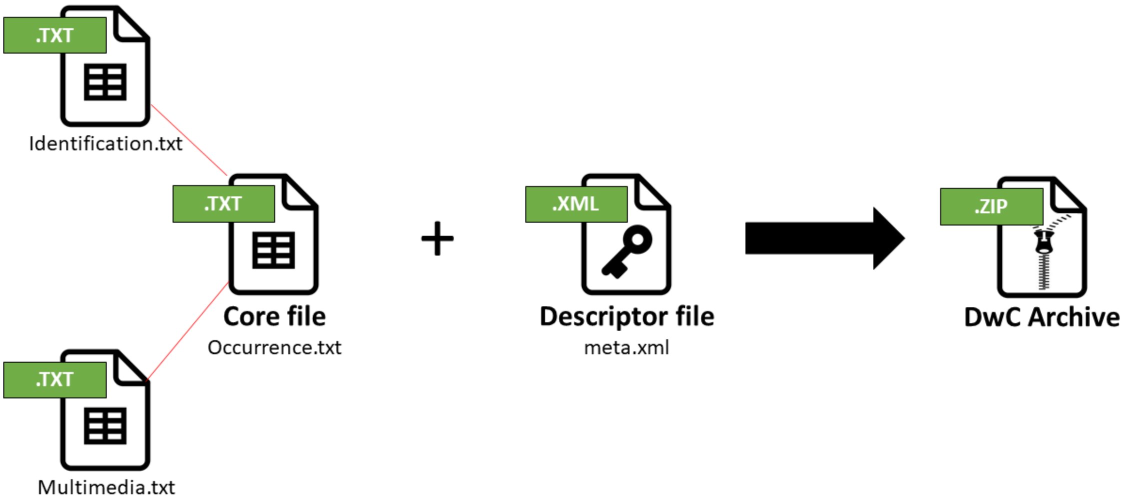

Figure 4: Simple DwC based Archive for data output from a CS platform to a CMS

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

meta.xml: the metadata of the archive. This file's function is to technically describe the archive to allow machines to process it. It is an .xml file. The file uses the following tags:

- <archive> see Appendix 1: The descriptor file in details
- <core> *within the <archive> tag*. see Appendix 1: The descriptor file in details
- <extension> *within the <archive> tag.* This tag functions as the <core> one, but stores secundary data (data linked to the <core> one through secondary keys).
- <file> *within the <core> and <extension> tag.* see Appendix 1: The descriptor file in details
- <location> *within the <file> tag*. see Appendix 1: The descriptor file in details. In this case the core is named "occurrence.txt" and the extensions "identification.txt" and "multimedia.txt". All three are located in the same folder as the meta.xml (the archive folder).
- <id/> *within the <core> tag*. This non-including tag defines the very first column of the core datafile. It contains the primary key of our data
- <coreid/> *within the <extension> tag*. This non-including tag defines the very first column of the core datafile. It contains the secundary key of our data, that refers to the primary keys (id) from the core.
- <field/> *within the <file> tag.* see Appendix 1: The descriptor file in details.

occurrence.txt: This datafile is the core of our document. As such, the id's in the first column are the unique identifier of our data. As there should exist an institutional unique identifier for each natural history collection specimen, the ID of occurrence will be the occurrenceID. The core datafile then contains all the information concerning the specimen, organism and collection/capture event. The file has to contain at least the mandatory ones and the basic informations:

- **id** (technically mandatory): As said above this is a column for machine reading. The choice has to be made of a unique identifier per specimen. We chose occurrenceID in our example.
- **modified** (optional): as the data have been modified during the CS project duration, it could be useful to give back that information to the collection.
- **occurrenceID** (mandatory): This column is for human reading. Even if it bears the exact same information as ID column, it has to be repeated.
- **institutionCode** (mandatory)
- **collectionCode** (mandatory)
- **catalogNumber** (mandatory)
- **recordedBy** (basis of record)
- **recordNumber** (basis of record)
- **eventDate** (basis of record)
- **countryCode** (basis of record)
- **verbatimLocality** (basis of record)
- **stateProvince** (optional common)
- **county** (optional common)
- **decimalLatitude** (optional common)

- **decimalLongitude** (optional common)
- **coordinatePrecision** (optional common)
- **verbatimCoordinates** (optional common)
- **minimumElevationInMeters** (optional common)
- **maximumElevationInMeters** (optional common)
- **verbatimElevation** (optional common)
- **establishmentMeans** (optional common)
- **habitat** (optional)
- **occurrenceRemarks** (optional)
- **organismRemarks** (optional)
- **otherCatalogNumbers** (optional)

identification.txt: This extension data file contains the data pertaining to the identification and the taxon the specimens have been identified as. Each specimen must have at least one identification (= correspond to a line in this datafile). The separate data file for the identification allows us to get several identifications for the same specimen. Consequently, it is possible to get several lines with the same coreID in the identification file, each of these lines corresponding to a single identification of the specimen. However, each identification line should correspond to one and only one line in the occurrence file.

- **coreID** (technically mandatory): This is the secundary key linking the data from this data file to the occurrence data file.
- **modified** (strongly suggested)
- **scientificName** (basis of record)
- **identifiedBy** (optional)
- **dateIdentified** (optional)
- **taxonRemarks** (optional)
- **genus** (optional)
- **specificEpithet** (optional)
- **taxonRank** (optional)
- **infraspecificEpithet** (optional)
- **scientificNameAuthorship** (optional)
- **vernacularName** (optional)

multimedia.txt: this data file contains the data pertaining to the images used for the CS project. Same as for identification.txt, it is possible to have several images for the same specimen (=several lines in the datafile with the same coreID/occurrenceID, but with a different associatedMedia). This data file is presented here to allow troubleshooting in case of mismatch during the production or the data exchange.

- **coreID** (technically mandatory): same as for identification.txt.
- **associatedMedia** (mandatory)
- **type** (mandatory)
- **format** (mandatory)

- **licence** (mandatory)

As for the data sent to the CS platform, we created an illustrative archive (Le Bras 2019b) with CS project data. We used here the data produced on Les Herbonautes for the pilot project held within ICEDIG WP4.2 (pilot project 2 on the report (Phillips et al. 2019)). As for the input archive, it is possible to re-use the meta.xml, however a modification of the terms in the mission result may have to be done. The formating rules to follow to use the archive as working base are the same as above.

Example 3: Data sent from the CS platform

Case for the simple specimen P03558024

A tabular version of this example is available online, in which the two data files are made into separate sheets into the same spreadsheet (https://doi.org/10.5281/zenodo.2579753).

Remarks: In this example, empty fields are specified by null. The field number of this specimen was not transcribed although it is clearly visible on the image (13488). The information reflects it was not transcribed by leaving the field empty. Same applies for dateIdentified, which is as well clearly visible on the label (1965-05-03). On the other hand, no elevation was specified on the label, and the mention n/a indicates someone did notice the lack of information on the label and reported it in the dataset. Same applies for the name of the identifier (in facts, specialists knowing the collection will know M. Debray is here himself the identifier, but that is deduction we can hardly ask from volunteers).

occurrence.txt (1 line)

```
id = "http://coldb.mnhn.fr/catalognumber/mnhn/p/p03558024"
modified = "2015-09-04"
occurrenceID = "http://coldb.mnhn.fr/catalognumber/mnhn/p/p03558024"
institutionCode = "MNHN"
collectionCode = "P"
catalogNumber = "P03558024"
recordedBy = "Debray, M."
fieldNumber = null
eventDate = "1964-07-28"
countryCode = "FR"
verbatimLocality = "Côtes-du-Nord : Perros-Guirec à Ploumanac'h"
stateProvince = "FR-E"
county = "FR-22"
decimalLatitude = 48.83698
decimalLongitude = -3.4831
coordinatePrecision = null
verbatimCoordinates = "48° 50' 13.128'' N ; 3° 28' 59.16'' O"
minimumElevationInMeters = "n/a"
```

```
maximumElevationInMeters = "n/a"
verbatimElevation = "n/a"
```

identification.txt (1 line)

```
coreID = "http://coldb.mnhn.fr/catalognumber/mnhn/p/p03558024"
scientificName = "Thymus polytrichus A.Kern. ex Borbás"
identifiedBy = "n/a"
dateIdentified = null
```

multimedia.txt (1 line)

```
coreID = "http://coldb.mnhn.fr/catalognumber/mnhn/p/p03558024"
associatedMedia = "http://mediaphoto.mnhn.fr/media/1441365719281fbgNH3QOftOJIz09"
type = "StillImage"
format = "image/jpeg"
licence = "http://creativecommons.org/licenses/by/4.0/legalcode"
```

Example 4: Data sent from the CS platform

Case for the multi-determined specimen P01978557

A tabular version of this example is available online, in which the two data files are made into separate sheets into the same spreadsheet (https://doi.org/10.5281/zenodo.2579768).

Remarks: In this example, three lines in the identification documents refer to a specimen with three different identifications (action of identification). The date of collection, as often in old specimens, was not specified to the day, so the information was treated as a range (between the 1st of February 1889 and the 28th of February 1889. On the other hand, Kok made his identification in June 2018, and the month is treated as such following ISO 8601.

occurrence.txt (1 line)

```
id = "http://coldb.mnhn.fr/catalognumber/mnhn/p/p01978557"
modified = "2018-07-17"
occurrenceID = "http://coldb.mnhn.fr/catalognumber/mnhn/p/p01978557"
institutionCode = "MNHN"
collectionCode = "P"
catalogNumber = "P01978557"
recordedBy = "Balansa, B."
fieldNumber = "2414"
eventDate = "1889-02-01/1889-02-28"
countryCode = "VN"
verbatimLocality = "Hanoï, dans les jardins"
```

stateProvince = "VN-HN"
county = null
decimalLatitude = 21.02776
decimalLongitude = 105.83416
coordinatePrecision = null
verbatimCoordinates = null
minimumElevationInMeters = "n/a"
maximumElevationInMeters = "n/a"
verbatimElevation = "n/a"

identification.txt (3 lines)

- coreID = "http://coldb.mnhn.fr/catalognumber/mnhn/p/p01978557"
  scientificName = "Cinnamomum tonkinense (Lecomte) A.Chev."
  identifiedBy = "Kok"
  dateIdentified = "2018-06"

- coreID = "http://coldb.mnhn.fr/catalognumber/mnhn/p/p01978557"
  scientificName = "Cinnamomum tonkinense (Lecomte) A.Chev."
  identifiedBy = "Kostermans"
  dateIdentified = "n/a"

- coreID = "http://coldb.mnhn.fr/catalognumber/mnhn/p/p01978557"
  scientificName = "Cinnamomum albiflorum Nees var. tonkinensis Lecomte"
  identifiedBy = "Kok"
  dateIdentified = "2018-06"

multimedia.txt (1 line)

coreID = "http://coldb.mnhn.fr/catalognumber/mnhn/p/p01978557"
associatedMedia = "http://mediaphoto.mnhn.fr/media/14413029065481L6TwhTj5Lagqxn4"
type = "StillImage"
format = "image/jpeg"
licence = "http://creativecommons.org/licenses/by/4.0/legalcode"

Giving feedback on transcriber activity

During the transcription process, the present document specifies no data exchanges between transcription platform and institutional CMS or future DiSSCo infrastructure.

A dashboard displaying the digitization progress of European collections is under study. As a first step, during the ICEDIG project, the design of such a DiSSCo Dashboard is focusing on reflecting the state of member's collections and digital catalogues. Citizen science was out of scope. However, having the results of their work displayed outside of the platform is a major incentive for transcriber's communities.

A global dashboard for transcription platforms has been set up temporary for the WeDigBio event. On February 2019 it was still available on https://wedigbio.org/. IDigBio, the United States of America program to facilitate national collection digitization, launched WeDigBio in 2014. This event is annual, lasts 4 days, and aims to highlight and encourage biodiversity collections transcription by citizens worldwide. A dashboard has been developed to display on the event website the worldwide activity linked to the project.

Some interoperability was needed during WeDigBio event. A draft protocol is documented on https://github.com/iDigBio/wedigbio-dashboard. This event-driven dashboard can serve as a starting point for designing a more general protocol.

Implementation of such a protocol in platforms and a DiSSCo Dashboard will allow to measure the role of volunteers in transcription and foster participation.

Methodology

During the process of design of exchange protocols, our major concern was to deliver a simple solution. As simplicity was not always simple when facing the diversity of Natural History collections, practices and transcription platforms we had to make several trade-offs.

This chapter details each step of the design process.

1- Inventory of transcription platforms

Prior to work on this document, we made an inventory of the existing major transcription systems dedicated to natural history collections. This *evaluation of existing volunteer transcription systems* is available online (Le Bras and Chagnoux 2018).

2- Choice of a basic data model

The first step in order to write this specification was to choose the biodiversity standard to start from. We wanted it to fit the following requirements:

- Remain simple
- Being a commonly used solution. This is especially important because:
    - There are more chance the exchanging parties will know about the data model already
    - The platform technical teams might already be able to create automated imports/exports using the standard (so the data model can be applied straight ahead)
    - Documentation about the data model exists
    - People already worked on and thought about the data model, its limitations, and how to improve it
    - The model is then stronger
- Being adapted to specific needs of biodiversity collections databases
- Being open source
- Being costless to use, and having no particular software involved in its use
- Using light files to transfer the information

Darwin Core Archive appeared to us as being the best possible solution, as it met all of our requirements.

It was then proposed during ICEDIG all-hands meeting in Meise on the 5th of december 2018, and our proposition was validated by the 5.2 working group.

3- Adaptation of the DwC model structure to our specific requirements

In order to facilitate the use of DwC we decided to get simplified structures for our archive. This included:

- The abandoning of the EML descriptive file. Indeed, the project description requirements significantly differ from one platform to another. It seemed to us way more complicated to fit these different requirements in a standard. We then left it to the two parties to exchange data on this subject as they already do.
- The data sent from the CMS to the platform was to be limited and focusing on the images.
- The data from the platform to the CMS was to be focused on the occurrence.

These basic schemes were then proposed during ICEDIG all-hands meeting in Meise on the 5th of December 2018, and our propositions were validated by the 5.2 working group.

4- Inventory of terms currently used on the major CS platforms and DwC correspondence

The first step was to inventory the existing fields on the major CS platforms (Herbonautes/Herbonauten, Doedat/Digivol, Zooniverse). As could be expected, for each platform dealing with the same types of data, the vast majorities of the fields were common from one platform to the other, with only slight differences.

A correspondence in DwC terminology was then sought for each CS platform when possible.

The first issues were then identified with particular fields on some platforms, such as the Belgium national geographical grid cells systems (IFBL) codes been asked for on Doedat, for instance. If such concepts cannot fit in existing fields like verbatimCoordinates, the decision was made to leave this issue to be discussed between the CS platform and the collection management team on a case to case basis.

5- Categorization of DwC terms following their use in collection

The list of terms was then categorized into three categories:

- **Mandatory for setting a CS mission**. This list was buildt by listing the minimal information about a specimen that appears on a CS platform prior to its transcription, and also based on the experience of what is needed to build a mission on les Herbonautes.
- **Basic information**. This list was built by considering the information given in a summary of results from search engines used by institutional collection platforms, GBIF, Jstor and others. This list was then compared to information used to cite a specimen in literature.
- **Additional information.** This list contained all the information about the specimen that can be found on a CS platform and that fits our list of terms. As it contains a lot of terms, we later divided it in two subparts, mainly in order to ease readability of the document. This division is suggestive, based on authors' experiences of collection databases, but exists purely for an easier understanding, and it doesn't affect the specifications themselves. These parts are:
    - **Common additional data.** The additional data that can be used for sorting the specimens in a database.

- **Optional additional data.** The fields in which the data will be usually input in a way to specify the information, but which are difficult to query by a search engine.

The following list was made:

- mandatory for setting a CS mission
  - associatedMedia: the unique identifier of the image/media
  - occurrenceID: the unique identifier of the specimen corresponding to the image
  - type: the type of the media (usually stillImage)
  - format: the format of the media
  - licence: the licence under which the media should be used
  - institutionCode: the code of the institution holding the specimen
  - collectionCode: the collection code the specimen is part of
  - catalogNumber: the catalog number of the specimen
  - scientificName: a scientific name corresponding to the specimen.
- basic information
  - recordedBy
  - recordNumber
  - eventDate
  - countryCode
  - verbatimLocality
- optional fields
  - stateProvince
  - county
  - decimalLatitude
  - decimalLongitude
  - coordinatePrecision
  - verbatimCoordinates
  - verbatimElevation
  - minimumElevationInMeters
  - maximumElevationInMeters
  - establishmentMeans
  - habitat
  - occurrenceRemarks
  - organismRemarks
  - otherCatalogNumbers

This list was later crossed with the minimum information standard for digital specimens working document (version 0.5) set in the frame of work package 6 of ICEDIG.

6- Realisation of an illustrative archive of data import to a CS platform

In order to test the archive realization step by step, we created one based on the pilot conducted by ICEDIG for WP4.2. This mission constituted of specimens from 7 institutions within Europe, with different languages involved and should represent a wider case study. More information on the data used in this pilot can be found in the data paper (Dillen et al. 2019). As is the case for most transcription projects however, this project is constituted only of herbaria specimens. Indeed, due to greater technical difficulty, zoological and even more paleontological collections are imaged to a much lower extent than botanical ones. At the time of writing this document, no CS paleontological project was being run. Consequently, there was no reference we could build on. Issues from different collection types were hypothesized based on authors' experience, and DwC plasticity can be expected to allow relatively easy troubleshooting in case of issues due to zoological or paleontological particularities.

7- Realisation of a notional archive of data exportation from a CS platform

We then build a notional archive of export based on the pilot project data produced on les Herbonautes (mission held from 22/06/2018 to 10/10/2018). This raised several issues:

- How to indicate the absence of data on the specimen (to differentiate it from "no data has been produced"). On les Herbonautes, there is a "no information" checkbox to be checked by transcribers in the case there is no data available. This type of checkbox is as well present on other CS systems. We then decided to use "n/a" to distinguish it from null.
- Although never quoted in the final data export from our mission, the systems give the possibility to the citizen scientist to express their uncertainty by quoting a checkbox for "I'm not sure". This sort of checkbox exists on most of the CS platforms. We decided to add " [?]" to the end of the relevant field content in case the checkbox is quoted.
- The collector/capturer names (RecordedBy) are very often spread between several columns, especially in the case of multiple collectors (first collector in a column, others in a second column). Although it is basic information about the specimen, it is very difficult to have it standardized. Before deciding a format, we compared data from different CS platforms and on GBIF. We then chose the name format which was most common: the family name first immediately followed by a comma, a space, and the initial(s) each followed by a point. (Name, F.). In order to ease the separation of the names and to distinguish between the comma that separates the initials from the surname, a semicolon will be used between the names of different persons.
- Several formats of dates are in use on the eventDate. The ISO 8601 https://en.wikipedia.org/wiki/ISO_8601 norm provides a solution to all the issues met (range, lack of precision). The other date format columns (dateIdentified and Modified) should be formatted in ISO 8601 normally as exposed above.
- Sometimes, no scientific name can be linked to the specimen (i.e. a non-determined specimen). An indication should then be included to state clearly that the absence of data is

not due to an informatic technical issue, but rather a determination issue. We chose here "Insertae sedis", as the meaning of this value can be easily found by the CS user on the web.

Part of these different solutions were discussed on 5th December 2018 at the second ICEDIG All-Hands meeting by the WP5.2 group.

8- Crossing information with data quality working group

The ICEDIG Report on new methods for data quality assurance, verification and enrichment (Phillips et al. 2019) did compare data quality from different crowdsourcing platforms and other resources. It allowed us to confirm the choices made above about standardisation for recordedBy, eventDate, and fieldNumber. It helped us understand the most common issues met within the data in order to confirm the solutions we proposed. Data standard solutions described here are therefore considered to facilitate data interoperability as well as data quality.

9- Redaction of the specification document

We then gathered all the information here and structured the first version of this document. This document was later completed with remarks from the ICEDIG community and some external partners (WeDigBio and BGBM).

Conclusion

The protocol described in this document aims at facilitating data exchanges and, as such, at facilitating natural history collections digitization through citizen science platforms. It was elaborated with the concern of keeping things as simple as possible.

The proposed protocol seems simple enough to be implemented soon by major platforms. We hope that the document is clear enough to allow collection managers to prepare the images for citizen science. We also hope that the document is precise enough for implementation by platform administrators.

Independently of implementation, the transcription by volunteers will continue in the next coming months and years. Having several protocol compatible platforms in the European landscape will allow better interoperability, which will both open citizen science to institutions with no transcription platform, and leverage transcription capacity by taking advantage of the specific strength of each community, the most obvious one being language proficiency.

With these advances in interoperability, the integration of transcription platform into DiSSCo should be relatively straightforward, when the infrastructure will be up and running in a few years time.

Appendix 1: The descriptor file in details

From one archive to another, meta.xml keeps the same structure. Its function is to technically describe the archive to allow machines to read and process it. It is an .xml file. This file uses the following tags:

- <archive> This tag includes an attribute describing the whole archive format:
  - xmlns (XML NameSpace) indicate which "xml language" your archive is written. In our example, it is a DwC archive, basic terms of which are available on "http://rs.tdwg.org/dwc/text/"
- <core> *within the <archive> tag*. this refers to the central file of our archive (here multimedia.txt). As we only have one datafile, it is the one being described here. This tag includes attributes describing the core format:
  - encoding: the character encoding of the file (in our example it is UTF8).
  - fieldsTerminatedBy: which character are used to separate fields within your archive. By default, if saved in common spreadsheet software ";", if copy/paste from a spreadsheet software onto a notepad solution "\t". In our example "\t".
  - linesTerminatedBy: which character are used to go on the next line within the core of your archive. By default, in common spreadsheet software generated files "\n". In our example "\n".
  - fieldsEnclosedBy: which character are used to frame your field content (depending on your procedure on spreadsheet software can be framed by " or not). In our example, there are none.
  - ignoreHeaderLines: This define the size (in row) of the headers in the datafile. In our example, one single line describes the contents of each columns. To note that the headers in the datafile are only for human reading. The machine takes the information about the columns from what will be given in the tags <id/> and <field/> described here under.
  - rowType: where to find the information about the core format. In our example, it is the DwC extension for simple multimedia file (http://rs.gbif.org/terms/1.0/Multimedia)
- <file> *within the <core> tag.* This tag refers to the datafile basic informations
- <location> *within the <file> tag*. this mention the relative location of the core file (relatively to the meta.xml one). In our case the core is named “multimedia.txt” and located in the same folder as the meta.xml.
- <id/> *within the <core> tag*. This non-including tag define the very first column of our datafile. Its attribute is
  - index: locate the content in the row. For <id/> the index is 0 as it is not properly data for the computer: the computer considers it as the primary key of our core data (the unique identifier of each row). In our example, we used the associatedMedia content, as it is a unique identifier for our images.
- <field/> *within the <core> tag.* This describe the proper content of each column of our datafile. Each column containing data in our file should correspond to a <field/> entry. Its attributes are:

- index: defining the position of the column after the ID. The column with index=1 is then the first column after the id (that’s to say, for a human eye, the actual second column, or the one spreadsheet software usually defines as the column “B”). <field/> should be ordered by growing index.
- term: gives the URL of a descriptive of the content.

References

Dillen M, Groom Q, Chagnoux S, Güntsch A, Hardisty A, Haston E, Livermore L, Runnel V, Schulman L, Willemse L, Wu Z, Phillips S (2019) A benchmark dataset of herbarium specimen images with label data. Biodiversity Data Journal 7. doi: 10.3897/BDJ.7.e31817

Güntsch A, Hagedorn G, Hyam R, Röpert D (CETAF stable identifiers for specimens. CETAF-ISTC Available from: http://cetaf.org/sites/default/files/cetaf-istc_stable_identifiers_poster50x70.pdf.

International Commission on Zoological Nomenclature (1999) International code of zoological nomenclature. 4th ed. Ride WDL, International Trust for Zoological Nomenclature, Natural History Museum (London, England), International Union of Biological Sciences (Eds). International Trust for Zoological Nomenclature, c/o Natural History Museum, London, 306 pp.

Le Bras G (2019a) Illustrative Darwin core archive to input data on a citizen science platform from a collection management system. doi: 10.5281/zenodo.2579778

Le Bras G (2019b) Illustrative Darwin core archive to output data from a citizen science platform to a collection management system. doi: 10.5281/zenodo.2579782

Le Bras G, Chagnoux S (2018) Evaluation of Existing Volunteer Transcription Systems. Zenodo doi: 10.5281/zenodo.2578938

Phillips S, Dillen M, Groom Q, Green L, Weech M-H, Wijkamp N (2019) Report on New Methods for Data Quality Assurance, Verification and Enrichment. ICEDIG/DiSSCo. Deliverable 4.2 Available from: https://icedig.eu/sites/default/files/deliverable_d4.2_icedig_data_quality_in_transcription.pdf.

Turland N, Wiersema J, Barrie F, Greuter W, Hawksworth D, Herendeen P, Knapp S, Kusber W-H, Li D-Z, Marhold K, May T, McNeill J, Monro A, Prado J, Price M, Smith G eds. (2018) 159 International Code of Nomenclature for algae, fungi, and plants. Koeltz Botanical Books. doi: 10.12705/Code.2018