Biodiversity Data Mobilization Course

GBIF Secretariat

Версия 12, May 2021

Содержание

Course description . . . 1
- Audience . . . 1
- Prerequisites . . . 1
- Learning objectives . . . 1
- Certification . . . 2

Files for download . . . 2
- Videos . . . 2
- Exercise data . . . 3
- Exercise sheets . . . 3

Foundations . . . 3
- Terminology . . . 3
  - Definitions . . . 3
  - Software . . . 3
  - Structures . . . 3
- Data quality . . . 3
- Documentation . . . 7
- Digitization Workflows . . . 8
- Software tools . . . 9
- Install OpenRefine . . . 9
  - Installation Requirements . . . 10
  - Installation on MS Windows . . . 10
  - Installation on Mac . . . 12
  - Installation on Linux . . . 18
- Foundations review . . . 20

Use Case I - Herbarium Specimens . . . 23
- Scenario . . . 23
  - Description . . . 24
  - Data collection . . . 25
  - Dataset description . . . 25
- Exercises . . . 25
  - Planning . . . 25
  - Data capture . . . 25
  - Data management . . . 26
  - Data publishing . . . 26
- Exercise sheet . . . 26

Planning . . . 26
- Resourcing . . . 26
- Organizing . . . 26
- Exercise 1a-c . . . 26
  - Exercise 1a . . . 27
  - Exercise 1b . . . 28
  - Exercise 1c . . . 28
- Review . . . 29

Data capture . . . 29
Standards and Darwin Core . . . 30
Data origins and types . . . 30
Data capture, processing and quality . . . 30
Exercise 2 . . . 30
Review . . . 31
Data management . . . 37
Principles of data management . . . 37
Data management tools . . . 37
Exercise 3a-c . . . 37
Exercise 3a . . . 37
Exercise 3b . . . 38
Exercise 3c . . . 38
Exercise tips . . . 39
Validation checks . . . 39
Helpful tools . . . 39
Review . . . 40
Data publishing . . . 41
Data publishing concepts . . . 41
IPT overview . . . 41
Training IPT installations . . . 41
IPT demonstration . . . 41
Exercise 4 . . . 41
Review . . . 42
Assessment and certification . . . 43
Planning rubric . . . 43
Data capture rubric . . . 44
Data management rubric . . . 46
Data publishing rubric . . . 48
Use Case II - Invasive species . . . 49
Scenario . . . 50
Description . . . 50
Data collection . . . 51
Digital data description . . . 53
Invasives exercise sheet . . . 53
Exercise 1 . . . 53
Exercise 1a . . . 53
Exercise 1b . . . 53
Exercise 2 . . . 54
Exercise 3 . . . 54
Exercise 4 . . . 55
Use Case II - Lepidoptera sightings . . . 55
Scenario . . . 55
Description . . . 56
Data collection . . . 57

Analogue data capture example . . . 57
Digital data description . . . 58
Lepidoptera exercise sheet . . . 59
Exercise 1 . . . 59
Exercise 1a . . . 60
Exercise 1b . . . 60
Exercise 2 . . . 61
Exercise 3 . . . 61
Exercise 4 . . . 61
Use Case III - Birds from literature . . . 62
Scenario . . . 62
Description . . . 63
Original data collection . . . 63
Analogue data description . . . 64
Scanned and translated data description . . . 64
Digital data description . . . 65
Birds from literature exercise sheet . . . 65
Exercise 1 . . . 65
Exercise 2 . . . 66
Exercise 3 . . . 67
Exercise 4 . . . 67
Final assignments . . . 67
USE CASE II . . . 67
USE CASE III . . . 68
Assignment submission . . . 68
Course evaluation . . . 68
Key documentation . . . 68
Darwin Core . . . 68
Data publishing . . . 68
Data publishing: IPT . . . 69
Digitization . . . 69
GBIF . . . 70
Georeferencing . . . 70
Invasive Species . . . 70
Living Atlases . . . 70
Miscellaneous . . . 70
OpenRefine . . . 70
Planning/Collaboration . . . 71
Quality . . . 71
Sensitive species . . . 71
Taxonomy . . . 71
Glossary . . . 72
Appendix: Data papers . . . 74
Appendix: Solutions . . . 75
Foundations review solutions . . . 75

Planning review solutions . . . 78
Data capture review solutions . . . 79
Data management review solutions . . . 80
Data publishing review solutions . . . 81
Use Case I suggested solution . . . 82
Acknowledgements . . . 83
Course design and instruction . . . 83
Translators . . . 83
French . . . 83
Portuguese . . . 83
Spanish . . . 84
Resources . . . 84
Resource support . . . 84
Colophon . . . 84
Suggested citation . . . 84
Contributors . . . 84
Licence . . . 85
Persistent URI . . . 85
Document control . . . 85

Course description

This course enables participants to plan and implement biodiversity data mobilization efforts effectively using accepted community standards. Its aim is to increase the volume, richness and quality of the data published through the GBIF network. This course was first developed as part of the Biodiversity Information for Development (BID) programme funded by the European Union.

Topics include:

- Project management
- Data capture
- Data management
- Data publishing

This course is comprised of video instruction paired with quizzes and practical exercises. When offered as an onsite or virtual workshop, group work and social interaction are encouraged.

Audience

This course is designed for individuals who work as researchers or technicians in biodiversity research or policy institutions. The instruction provided is particularly useful for those who have a need or desire to plan mobilization projects and/or mobilize biodiversity information for their respective institutions.

Prerequisites

1. Introduction to GBIF course
2. Additionally, to make best use of the activities around this course, the participants should possess the following skills and knowledge:
    - Basic skills in computer and internet use, and, in particular, in the use of spreadsheets.
    - Basic knowledge about geography and biodiversity informatics: geography and mapping concepts, basic taxonomy and nomenclature rules.
    - Willingness to disseminate the knowledge learned in the workshop with partners and collaborators in your project by adapting the biodiversity data mobilization training materials to specific contexts and languages while maintaining their instructional value.
    - A good command of English. While efforts are made to provide materials in other languages, instruction/videos will be in English.

Learning objectives

- Learn key concepts (foundations) of biodiversity informatics, particular to biodiversity digital data management.
- Receive an introduction to the Darwin Core Standard and its components.
- Learn to understand the different stages for planning a mobilization project and how to adapt stages to a specific project.

- Evaluate a data mobilization strategy to identify potential gaps, inefficiencies and pitfalls.
- Develop a data mobilization strategy customized to a given institutional framework.
- Learn to identify the types of data and how to best capture relevant information using best practices, existing softwares, tools and techniques.
- Use software tools designed to facilitate biodiversity data capture to produce digital biodiversity data from analogue sources.
- Learn data quality concepts and receive an introduction to tools used for standardizing data, validating data, and cleaning data.
- Use software tools to evaluate the fitness-for-use of a biodiversity dataset
- Use software tools designed for (biodiversity) data cleaning.
- Learn the process of making biodiversity data freely available online, also known as data publishing, utilizing GBIF's Integrated Publishing Toolkit (IPT).
- Define the publishable data types and subtypes (if any) for a biodiversity dataset.
- Use the GBIF IPT to publish biodiversity datasets using the appropriate extensions.
- Capacitate others in the planning, capture, management and publishing of biodiversity data.

Certification

Upon successful completion of the course and successful assessment of assignments (by trainers and mentors), participants have the opportunity to receive an official certification in the form of an Open Badge. See Assessment and Certification for more details.

Files for download

All files for the course may be downloaded from this page. Or if you prefer, all files are linked individually throughout the course as they occur in the curriculum. The video files are embedded throughout the course, as well, and play from YouTube. Subtitles are available when playing from YouTube for most videos. If you have difficulty accessing the embedded videos, please download the mp4 files to play them locally on your computer.

Videos

The videos are narrated in English. Subtitles are not available for the downloaded videos.

Foundations1.zip (73.7 MB)

Foundations2.zip (90.2 MB)

Planning.zip (51.3 MB)

Capture.zip (63.1 MB)

Management.zip (30.2 MB)

Publishing.zip (77.9 MB)

Appendix-Data-Papers.zip (97.5 MB)

Exercise data

This compressed file (ZIP 37.7 MB) contains the exercise data for all the use cases.

Exercise sheets

This compressed file (ZIP 1.4 MB) contains the exercise sheets for all the use cases. The exercises sheets are written in English and should be completed in English.

Foundations

This module includes instruction to ensure all participants are at the same level before delving into the data mobilization topics. You will receive an introduction to the language, terminology and definitions for some of the basic concepts, functions and processes that you are going to be putting into use during the rest of the course. You will also receive an introduction to data quality and learn the importance of documentation. Lastly, you will be asked to install the OpenRefine software as part of this module.

Terminology

Definitions

In this video (12:02), you will review terminology used in this course. If you are unable to watch the embedded video, you can download it locally. (MP4 - 38.5 MB)

▶ https://www.youtube.com/watch?v=FZAF5Sy8Nsc *(YouTube video)*

Software

In this video (05:58), you will review examples of the different types of applications and software available in the world of biodiversity mobilization informatics. If you are unable to watch the embedded video, you can download it locally. (MP4 - 18.9 MB)

▶ https://www.youtube.com/watch?v=vYfDIgBBKXY *(YouTube video)*

Structures

In this video (13:10), you will review the field and data types that hold data, the structures that help to organize and protect that data and what these mean for the integrity and security of your data. If you are unable to watch the embedded video, you can download it locally. (MP4 - 38.8 MB)

▶ https://www.youtube.com/watch?v=msnVbZvIy2E *(YouTube video)*

Data quality

In this video (12:26), you will review terminology used in this course. If you are

unable to watch the embedded video, you can download it locally. (MP4 - 44.5 MB)

https://www.youtube.com/watch?v=5o7TcS2K7Cw *(YouTube video)*

Below you will find a selected reading from Arthur Chapman's guide "Principles of data quality". Full document, references and translations can be found on GBIF.org.

Before a detailed discussion on data quality and its application to species-occurrence data can take place, there are a number of concepts that need to be defined and described. These include the term data quality itself, the terms accuracy and precision that are often misapplied, and what we mean by primary species data and species-occurrence data.

Species-occurrence data

Species-occurrence data is used here to include specimen label data attached to specimens or lots housed in museums and herbaria, observational data and environmental survey data. In general, the data are what we term "point-based", although line (transect data from environmental surveys, collections along a river), polygon (observations from within a defined area such as a national park) and grid data (observations or survey records from a regular grid) are also included. In general we are talking about georeferenced data – i.e. records with geographic references that tie them to a particular place in space – whether with a georeferenced coordinate (e.g. latitude and longitude, UTM) or not (textual description of a locality, altitude, depth) – and time (date, time of day).

In general the data are also tied to a taxonomic name, but unidentified collections may also be included. The term has occasionally been used interchangeably with the term "primary species data".

Primary species data

"Primary species data" is used to describe raw collection data and data without any spatial attributes. It includes taxonomic and nomenclatural data without spatial attributes, such as names, taxa and taxonomic concepts without associated geographic references.

Accuracy and Precision

Accuracy and precision are regularly confused and the differences are not generally understood.

Accuracy refers to the closeness of measured values, observations or

estimates to the real or true value (or to a value that is accepted as being true – for example, the coordinates of a survey control point).

Precision (or Resolution) can be divided into two main types. Statistical precision is the closeness with which repeated observations conform to themselves. They have nothing to do with their relationship to the true value, and may have high precision, but low accuracy. Numerical precision is the number of significant digits that an observation is recorded in and has become far more obvious with the advent of computers. For example a database may output a decimal latitude/longitude record to 10 decimal places – i.e. ca .01 mm when in reality the record has a resolution no greater than 10-100 m (3-4 decimal places). This often leads to a false impression of both the resolution and the accuracy.

These terms – accuracy and precision – can also be applied to non-spatial data as well as to spatial data. For example, a collection may have an identification to subspecies level (i.e. have high precision), but be the wrong taxon (i.e. have low accuracy), or be identified only to Family level (high accuracy, but low precision).

Data quality

Data quality is multidimensional, and involves data management, modelling and analysis, quality control and assurance, storage and presentation. As independently stated by Chrisman (1991) and Strong et al. (1997), data quality is related to use and cannot be assessed independently of the user. In a database, the data have no actual quality or value (Dalcin 2004); they only have potential value that is realized only when someone uses the data to do something useful. Information quality relates to its ability to satisfy its customers and to meet customers' needs (English 1999).

Redman (2001), suggested that for data to be fit for use they must be accessible, accurate, timely, complete, consistent with other sources, relevant, comprehensive, provide a proper level of detail, be easy to read and easy to interpret.

One issue that a data custodian may need to consider is what may need to be done with the database to increase its usability to a wider audience (i.e. increase its potential use or relevance) and thus make it fit for a wider range of purposes. There will be a trade off in this between the increased usability and the amount of effort required to add extra functionality and usability. This may require such things as atomizing data fields, adding geo-referencing information, etc.

Quality Assurance/ Quality Control

The difference between quality control and quality assurance is not always clear. Taulbee (1996) makes the distinction between Quality Control and Quality Assurance and stresses that one cannot exist without the other if quality goals are to be met. She defines Quality Control as a judgement of quality based on internal standards, processes and procedures established to control and monitor quality; and Quality Assurance as a judgement of quality based on standards external to the process and is the reviewing of the activities and quality control processes to insure that the final products meet predetermined standards of quality.

In a more business-oriented approach, Redman (2001) defines Quality Assurance as “those activities that are designed to produce defect-free information products to meet the most important needs of the most important customers, at the lowest possible cost”.

How these terms are to be applied in practice is not clear, and in most cases the terms seem to be largely used synonymously to describe the overall practice of data quality management.

Uncertainty

Uncertainty may be thought of as a “measure of the incompleteness of one’s knowledge or information about an unknown quantity whose true value could be established if a perfect measuring device were available” (Cullen and Frey 1999). Uncertainty is a property of the observer’s understanding of the data, and is more about the observer than the data per se. There is always uncertainty in data; the difficulty is in recording, understanding and visualizing that uncertainty so that others can also understand it. Uncertainty is a key term in understanding risk and risk assessment.

Error

Error encompasses both the imprecision of data and their inaccuracies. There are many factors that contribute to error. Error is generally seen as being either random or systematic. Random error tends to refer to deviation from the true state in a random manner. Systematic error or bias arises from a uniform shift in values and is sometimes described as having ‘relative accuracy’ in the cartographic world (Chrisman 1991). In determining ‘fitness for use’ systematic error may be acceptable for some applications, and unfit for others.

An example may be the use of a different geodetic datum1 – where, if

used throughout the analysis, may not cause any major problems. Problems will arise though where an analysis uses data from different sources and with different biases – for example data sources that use different geodetic datums, or where identifications may have been carried out using an earlier version of a nomenclatural code.

“Because error is inescapable, it should be recognized as a fundamental dimension of data” (Chrisman 1991). Only when error is included in a representation of the data is it possible to answer questions about limitations in the data, and even limitations in current knowledge. Known errors in the three dimensions of space, attribute and time need to be measured, calculated, recorded and documented.

Validation and Cleaning

Validation is a process used to determine if data are inaccurate, incomplete, or unreasonable. The process may include format checks, completeness checks, reasonableness checks, limit checks, review of the data to identify outliers (geographic, statistical, temporal or environmental) or other errors, and assessment of data by subject area experts (e.g. taxonomic specialists). These processes usually result in flagging, documenting and subsequent checking of suspect records. Validation checks may also involve checking for compliance against applicable standards, rules, and conventions. A key stage in data validation and cleaning is to identify the root causes of the errors detected and to focus on preventing those errors from re-occurring (Redman 2001).

Data cleaning refers to the process of “fixing” errors in the data that have been identified during the validation process. The term is synonymous with “data cleansing”, although some use data cleansing to encompass both data validation and data cleaning. It is important in the data cleaning process that data is not inadvertently lost, and changes to existing information be carried out very carefully. It is often better to retain both the old (original data) and the new (corrected data) side by side in the database so that if mistakes are made in the cleaning process, the original information can be recovered.

Documentation

In this video (09:47), we will provide an overview of the importance of documentation as it relates to data management and data publishing. You will learn about data mapping, data relationships and metadata. If you are unable to watch the embedded video, you can download it locally. (MP4 - 29.2 MB)

▶ https://www.youtube.com/watch?v=Z5-SYImGRGc *(YouTube video)*

Digitization Workflows

This video (07:20) on Digitization Workflows identifies five clusters (or stages) in the process of digitizing natural history collection objects using digital images, and these stages can be easily adapted to other biodiversity data sources. If you are unable to watch the embedded video, you can download it locally. (MP4 - 26.8 MB)

▶ https://vimeo.com/120369455 *(Vimeo video)*

As the video highlights, digitization protocols vary from institution to institution, but it is essential that the chosen protocol is agreed, documented and respected.

We do not teach digitization, per se, during the workshop, as it can easily stand as a week-long course on its own, instead we focus on basic introduction to biodiversity data capture. However, we want to provide you with resources on digitization as we know many are interested in this.

There are many ways to organize digitization efforts and so digitization can seem daunting to begin with. It is important to remember that in most cases someone else has already tried to digitize the same types of specimens and objects that you are planning to. In this exercise we introduce you to some practical digitization workflow resources to help get you started. These will also form the basis for work we will do in the workshop on selecting, modifying and assessing workflows.

Some steps in the process may include:

- **Pre-digitization curation and staging**: This includes the preparation of the data source for the digitization process, including the assignment of unique identifiers that will help to refer to the source without error and to keep all derived information together.
- **Image capture**: This includes a fair amount of planning, not only on the image capture itself (e.g. definition of the work sequence, selection of adequate hardware), but also on how and where the images will be stored and handled.
- **Image processing**: This includes quality control, file conversion, etc.
- **Electronic data capture**: The core of the digitization process, includes capturing key information in a database. The video highlights that the most common method of entering the information is through a keyboard, but more and more institutions are turning to advanced data entry technologies.
- **Georeferencing**: Geographical information is very important for biodiversity analysis, so digitization projects should seek to extract the most accurate geographical information possible.

Integrated Digitized Biocollections (iDigBio) is the coordination centre for the United States National Resource for Advancing Digitization of Biodiversity Collections (ADBC). They lead a nation-wide effort to make data and images for millions of biological specimens available in a standard electronic format for the research community, government agencies, students, educators, and the general public. They have produced several videos that discuss the digitization process.

There are other videos in the iDigBio series that you may be interested in, if you wish to learn more about specific workflows for different specimen types:

- “Digitizing Wet Collections” (4:34 mins) https://vimeo.com/120369690
- “Imaging Workflows for the Digitization of Dry-preserved Vertebrate Specimens” (7:25 mins)

https://vimeo.com/160615629

- "Digitizing Herbarium Specimens" (7:34 mins) https://vimeo.com/120369768

Software tools

Review software tools used in biodiversity informatics

During the course activities, we'll demonstrate and work with many different software tools related to data digitization, data quality and transformation. You probably already use several of them in your daily work.

Community trainers, mentors and former course participants have compiled a list with information about biodiversity informatics software tools. It provides links for their main websites, a key facts and a summary of strong and weak points.

Download Software-database-EN.xlsx. (23 KB)

When analysing biodiversity software that you have not used before, you need to consider how you would adapt it for your purposes. You will find below a list with which you can start your evaluation. They are inspired by the chapter "characteristics of a good database solution" of the GBIF manual "Initiating a Digitisation Project":

- **Price**: One of the most determining factors. Beware of other costs beyond the price of the software license, such as hardware needed to run it, maintenance, upgrades, and the expertise to run it.
- **Functionality**: You need to have clarity on what do you expect the software to achieve, and make sure it does it efficiently. Do not get distracted by additional functionality that can make the software more complex unnecessarily.
- **Stability**: Some solutions have been in the market for long and are supported by solid institutions or companies are more likely to be bug-free and/or have good systems in place to solve any issues arising. It will also make more likely to be updated and ported to more modern operating systems.
- **Scalability**: Some software performs very well when demoed out-of-the-box, but its performance degrades after some time or when using them with larger amounts of data or when several users access it simultaneously. Check the opinions of other users online.
- **Integration**: Make sure that the software accepts and produces the data formats that you use and need. Data transformation is a time consuming task.
- **Language support**: it is essential that everyone using the software can understand its interface, and the documentation that will make possible its use.
- **Documentation and technical support**: make sure to explore the existing documentation and support mechanisms. You can be sure that at some point you will need it.
- **Learning curve**: Some software may require specific training to learn how to use it, while others are more intuitive and can be learned while using them, supported by in-line help systems.

Install OpenRefine

Install software required for activities later in the course

OpenRefine is a tool with a set of features for working with tabular data that improves the overall quality of a dataset. It is an application that runs on your own computer as a small web server, and in order to use it your web browser should point at that web server. So, think of OpenRefine as a personal and private web application.

We will use OpenRefine during the data mobilization portion of the course, especially during the practical exercises. It will be necessary to install OpenRefine on your laptop. If you are a skilled computer user, you can follow these steps to install the software on your computer. If you are not confident, please ask for help. Refer to the OpenRefine download page for more details.

Administrative passwords may be required to install software.

Installation Requirements

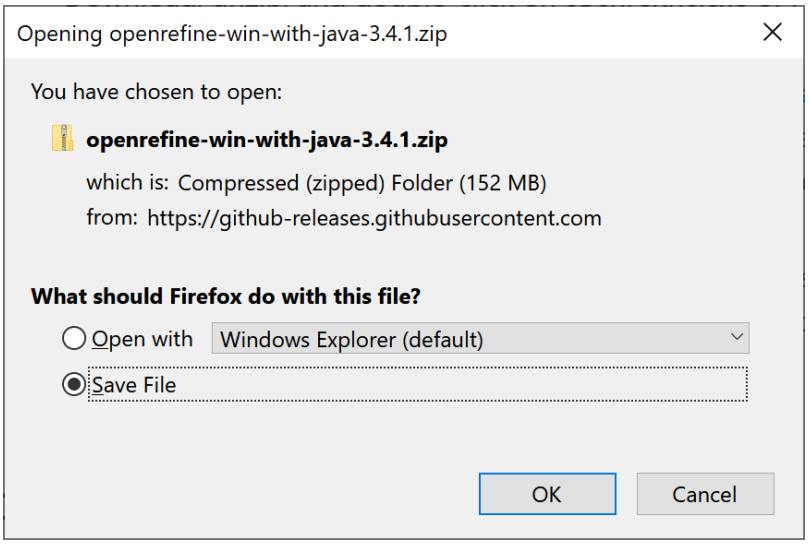
1. Linux users only: Java JRE installed.
2. Google Chrome, Microsoft Edge or Mozilla Firefox installed. Internet Explorer is not supported.

The latest stable release is OpenRefine 3.4.1, released on September 24, 2020. Detailed installation instructions are available at https://docs.openrefine.org/manual/installing/.

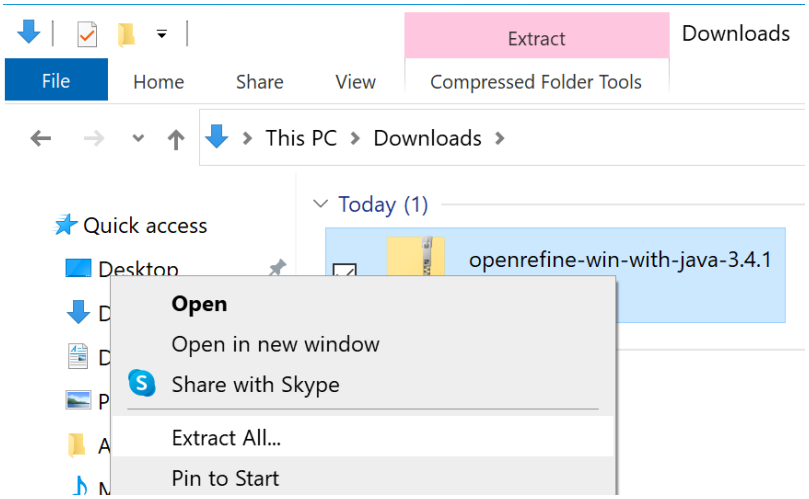
Installation on MS Windows

1. Download the Windows kit **with embedded Java**. Choose to save the file rather than open it.
2. Find the downloaded file. Right click it, and choose "Extract all…". Unzip, and double-click on openrefine.exe or refine.bat if the former does not work.
3. A command window will appear (don't close it) and soon after a new web browser window will show the application.

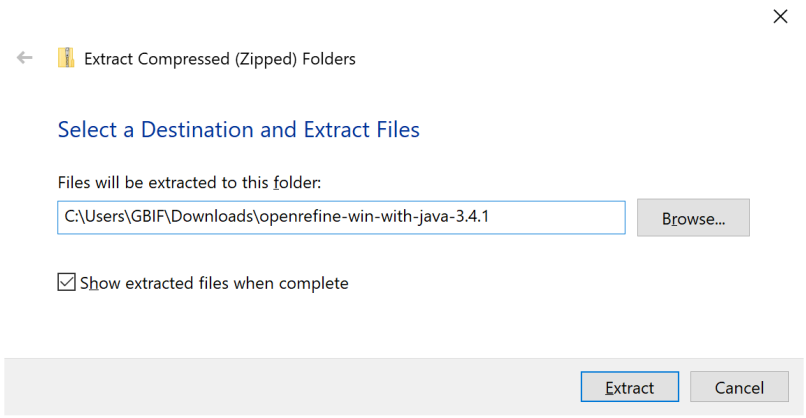
▼ *Detailed instructions for MS Windows (click to expand)*

1. 

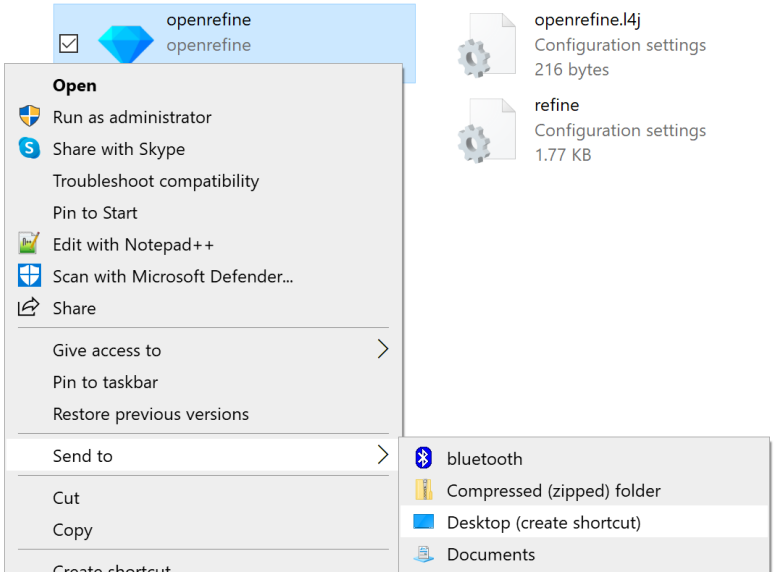
Download the Windows kit **with embedded Java**. Choose to save the file rather than open it.

2. 

Find the file you downloaded. Right click it, and choose "Extract All..."

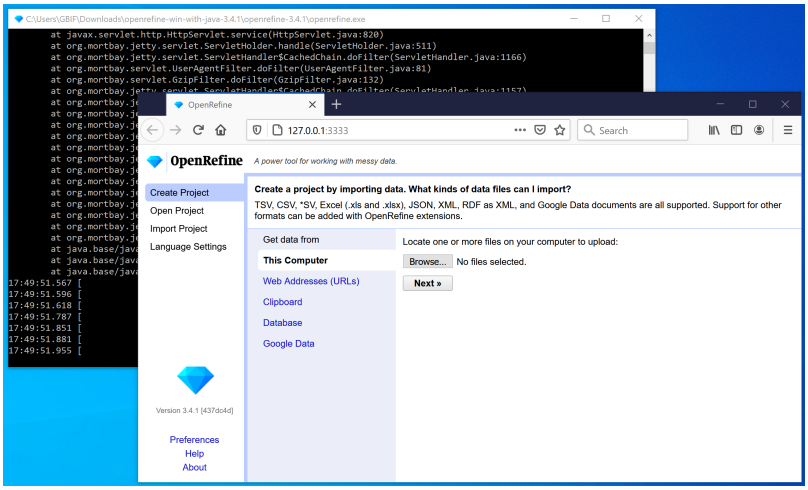
3. 

Click "Extract"

4. 

Find the extracted files. Optionally, right click "openrefine" and choose "Send to → Desktop (create shortcut)" to create a shortcut on your desktop. Then double click "openrefine"

5.



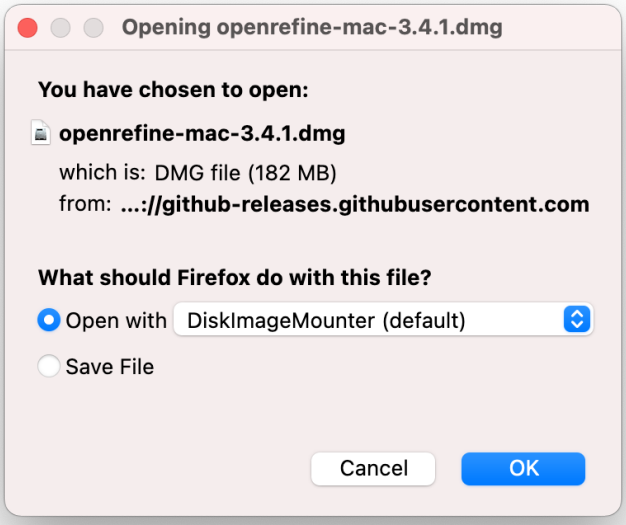
A black console window opens, and a short time later the browser opens. OpenRefine is now ready to use.

Installation on Mac

1. Download the Mac kit.
2. Download, open, drag icon into the Applications folder. You do not need to install Java separately.
3. Double click on it and a new web browser window will show the application.

▼ *Detailed instructions for Mac (click to expand)*

1.

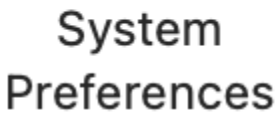


Download the Mac kit, and choose to open it.

2.

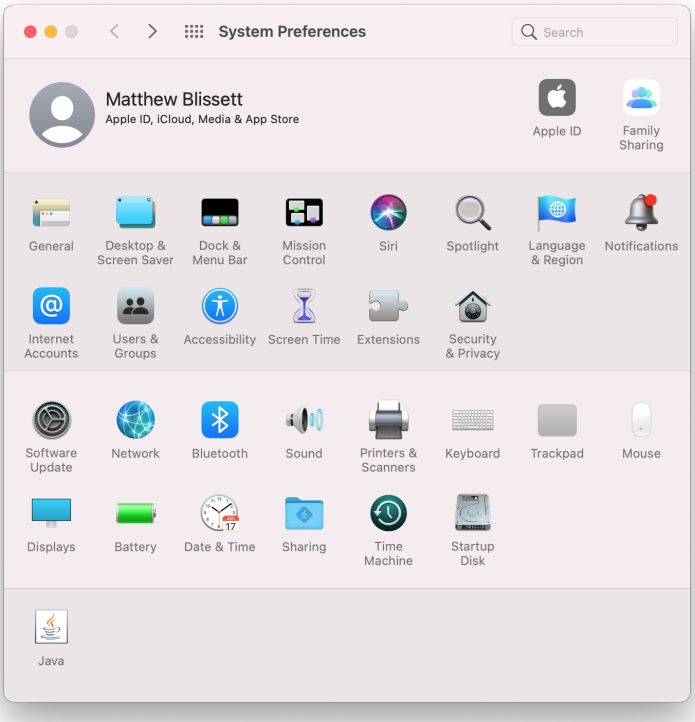
A warning is shown. Click "OK".

3. 



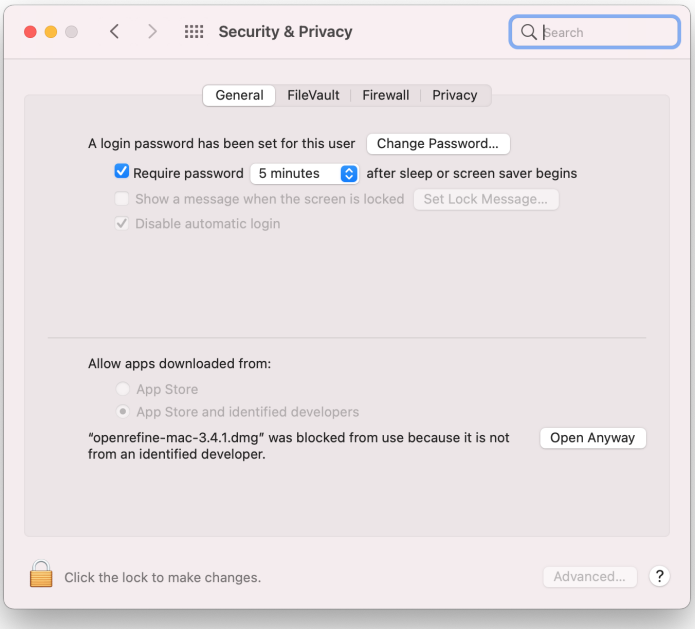
Open System Preferences.

4.



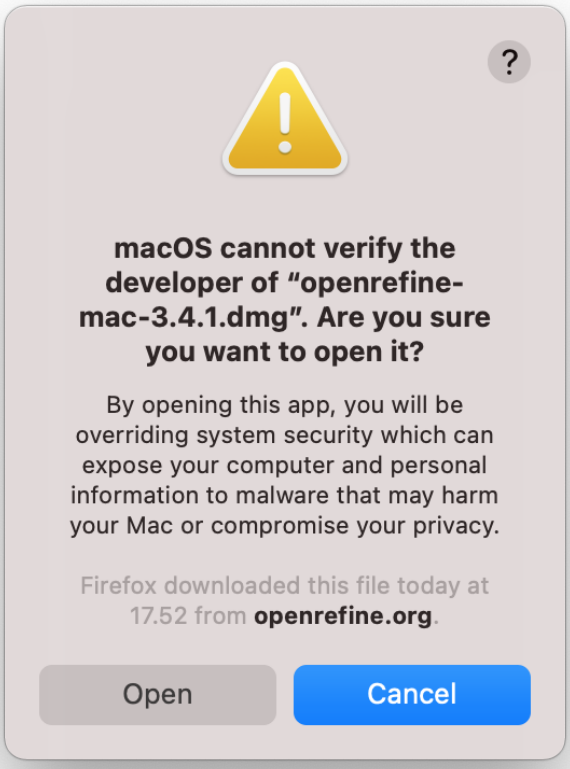
Open Security & Privacy

5.

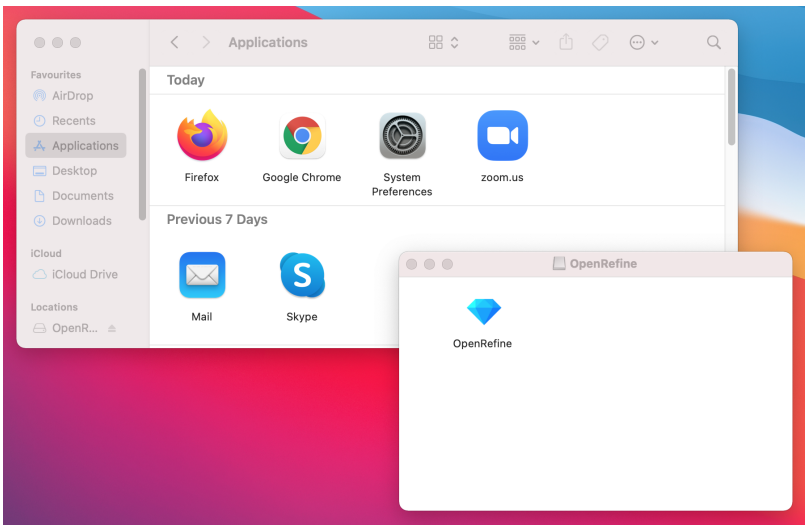


Choose "Open Anyway" at the bottom.

6.

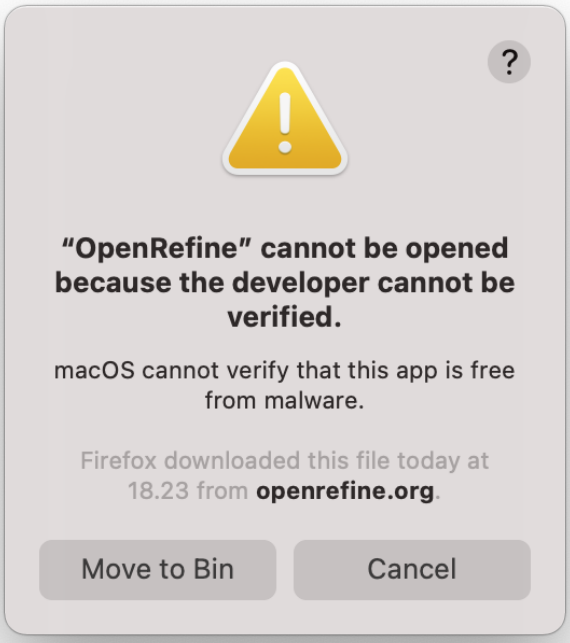


Choose "Open"

7. 

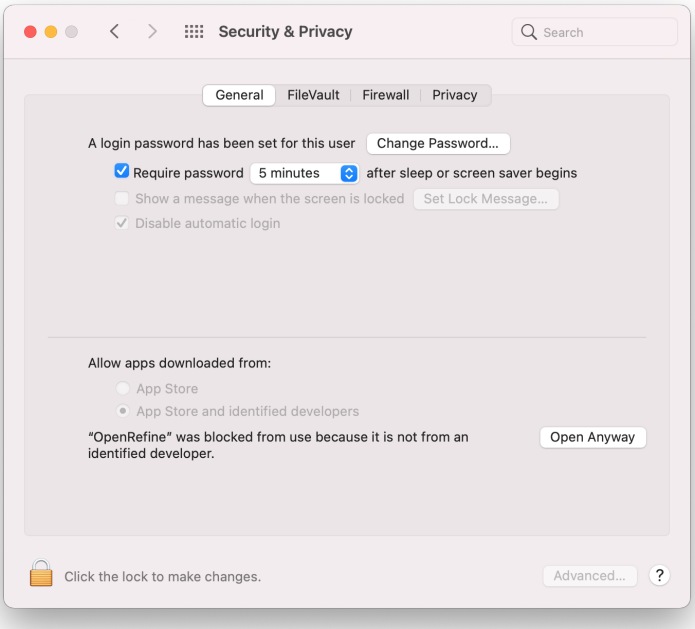
Finally, the application archive is opened! Drag it to your Applications folder.

8.



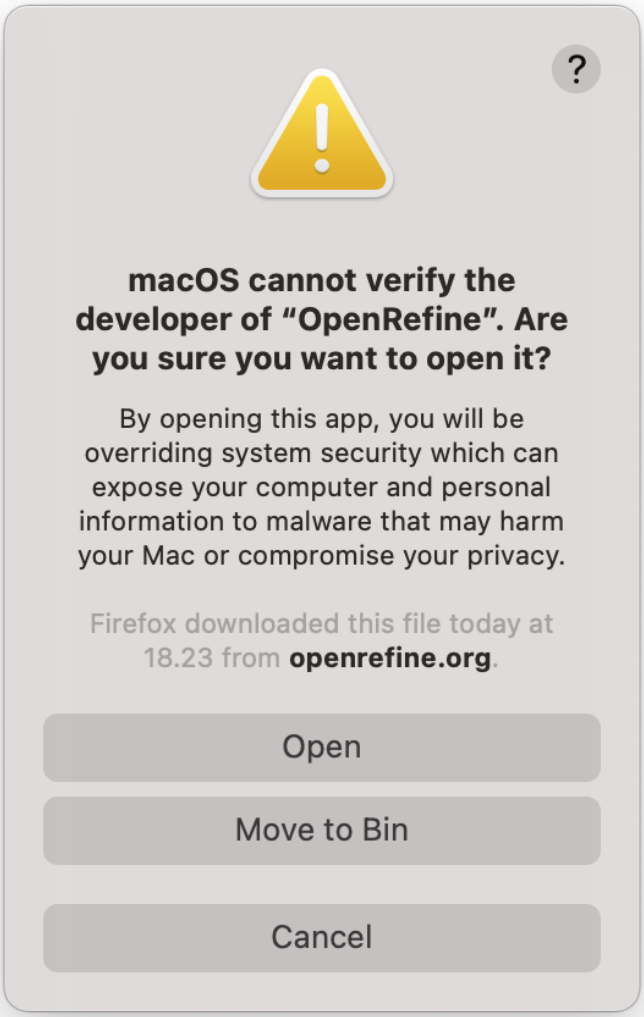
Double-click the OpenRefine icon. Another security warning appears!

9.



Go back to "Security & Privacy" and click "Open Anyway" — again.

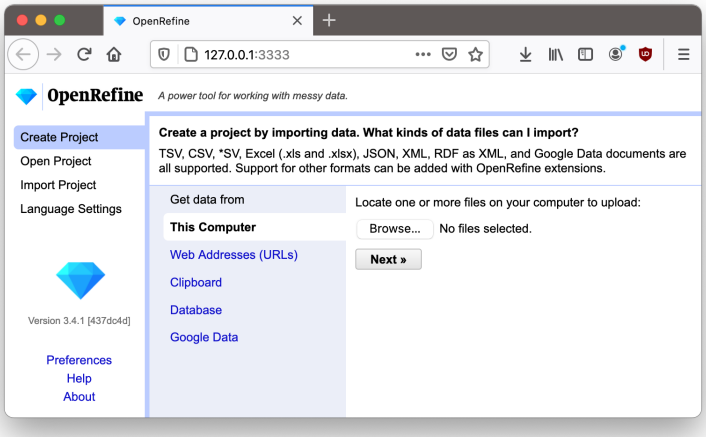
10.



(To avoid these warnings, the OpenRefine developers would need to pay Apple.)

Click "Open".

11.



Finally! The application is running.

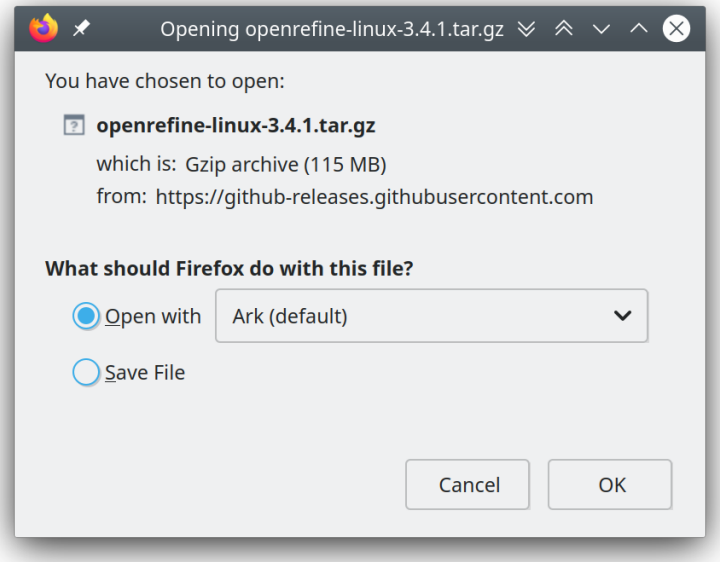
Installation on Linux

1. Download the Linux kit.
2. Download, extract, then type ./refine to start. This requires Java to be installed on your computer.

▼ *Detailed instructions for Linux (click to expand)*

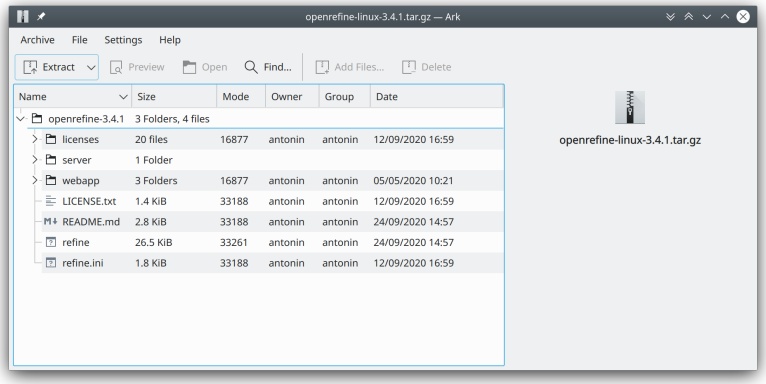
These instructions are for KDE (e.g. Kubuntu, SuSE), but the process is similar for Gnome (e.g. Ubuntu, Red Hat, CentOS).

1.



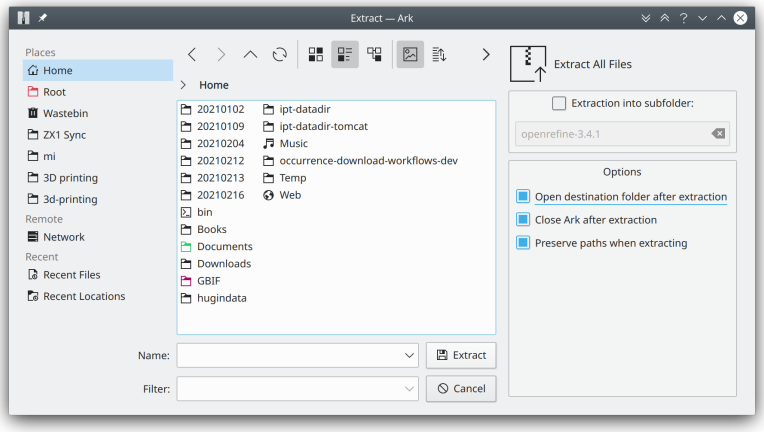
Download the Linux kit. Open it.

2.



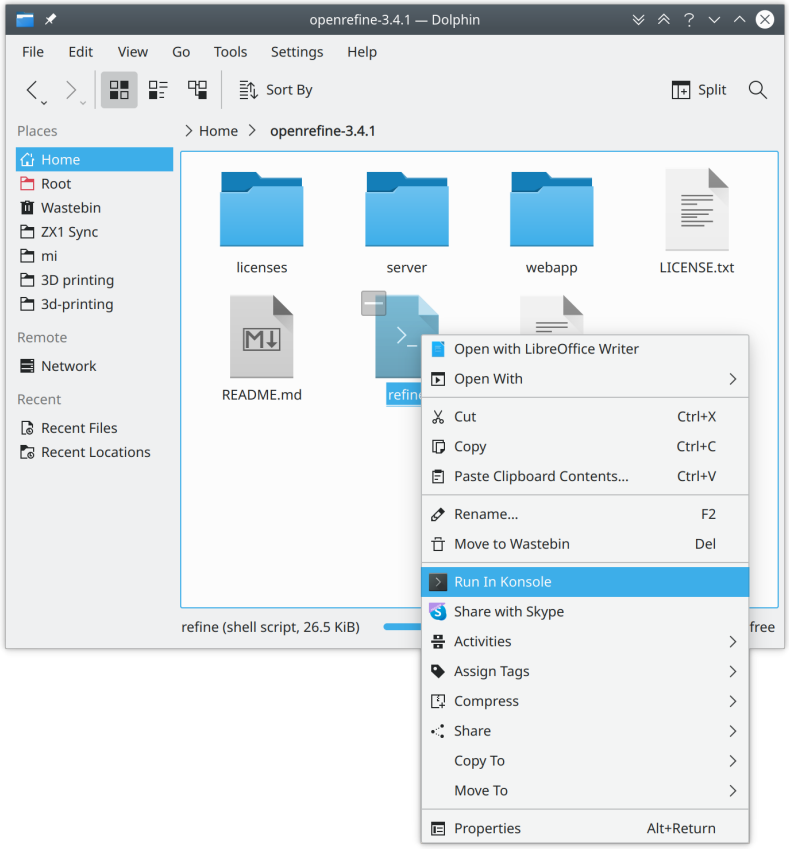
Click "Extract" to unpack the downloaded application.

3.



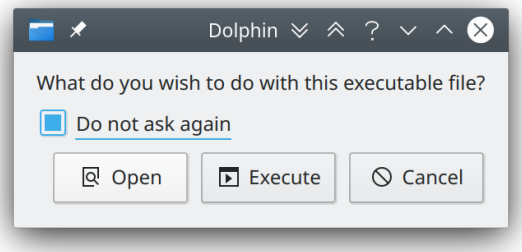
Choose a suitable place. I also selected "Open destination folder after extraction" and "Close Ark after extraction"

4.



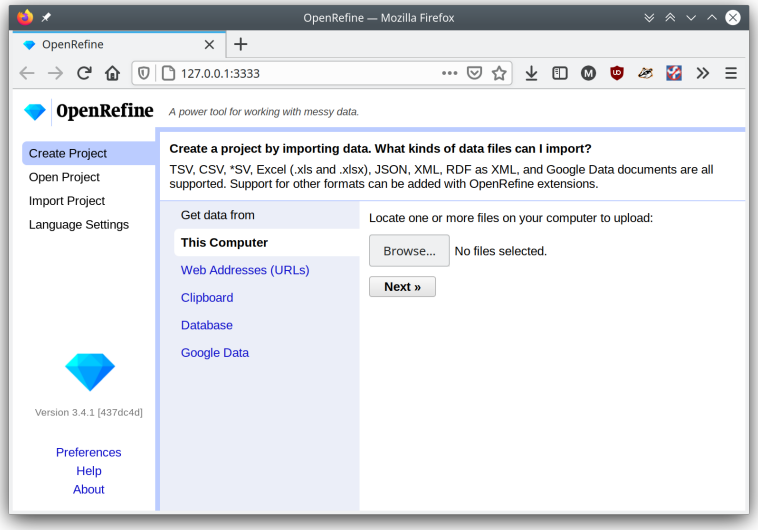
Right click "refine" and choose "Run in Konsole". This is needed so you can safely exit OpenRefine later, by closing the Konsole window.

5.



Confirm that you wish to execute the downloaded application.

6.



OpenRefine is now running.

Foundations review

Quiz yourself on the concepts learned in this section.

1. For the given statement, input the correct term (database, database language, database program)
   - combines and presents functions and features for manipulating data, together in a unified interface
     _____
   - structured and organized collection of data and/or information held on a computer
     _____
   - the way by which a human communicates with a computer
     _____

2. If you open a data file and see the following, what would you suspect is the issue?

   `□tre, ou ne pas □tre, c□est l□ la question.`

   - ☐ Nothing
   - ☐ It is corrupt
   - ☐ The wrong encoding was used to open the file
   - ☐ The sender used a weird font

3. For the given software, input the type of software (data capture, data management, data cleaning, data publishing).
   - Integrated Publishing Toolkit (IPT)
     _____
   - Specify
     _____ AND _____
   - iNaturalist

_____

- OpenRefine

  _____

4. For the given example, input the correct data type (binary, boolean, float, integer, long integer, text, unstructured text)

   - 1236975

     _____
   - 01101111

     _____
   - We walked 5 miles down the road west from the post office in the center of town. We then went 2 miles north on a dirt path to the river. Then we continued west along the river for another 5 miles.

     _____
   - 1024

     _____
   - 29.0

     _____
   - Yes/No

     _____
   - 6 rabbits were observed

     _____

5. Which of these terms describes a "field/column name"?

   - ☐ Assigned
   - ☐ Descriptive
   - ☐ Identifying
   - ☐ Readable
   - ☐ Unique
   - ☐ User-interface

6. Which of these terms describes a "field label"?

   - ☐ Assigned
   - ☐ Descriptive
   - ☐ Identifying
   - ☐ Readable
   - ☐ Unique
   - ☐ User-interface

7. For each statement, input the correct structure (row, column, table)

   - All data refers to a SINGLE concept.

     _____
   - An attribute has the SAME field/data type for every record.

_____

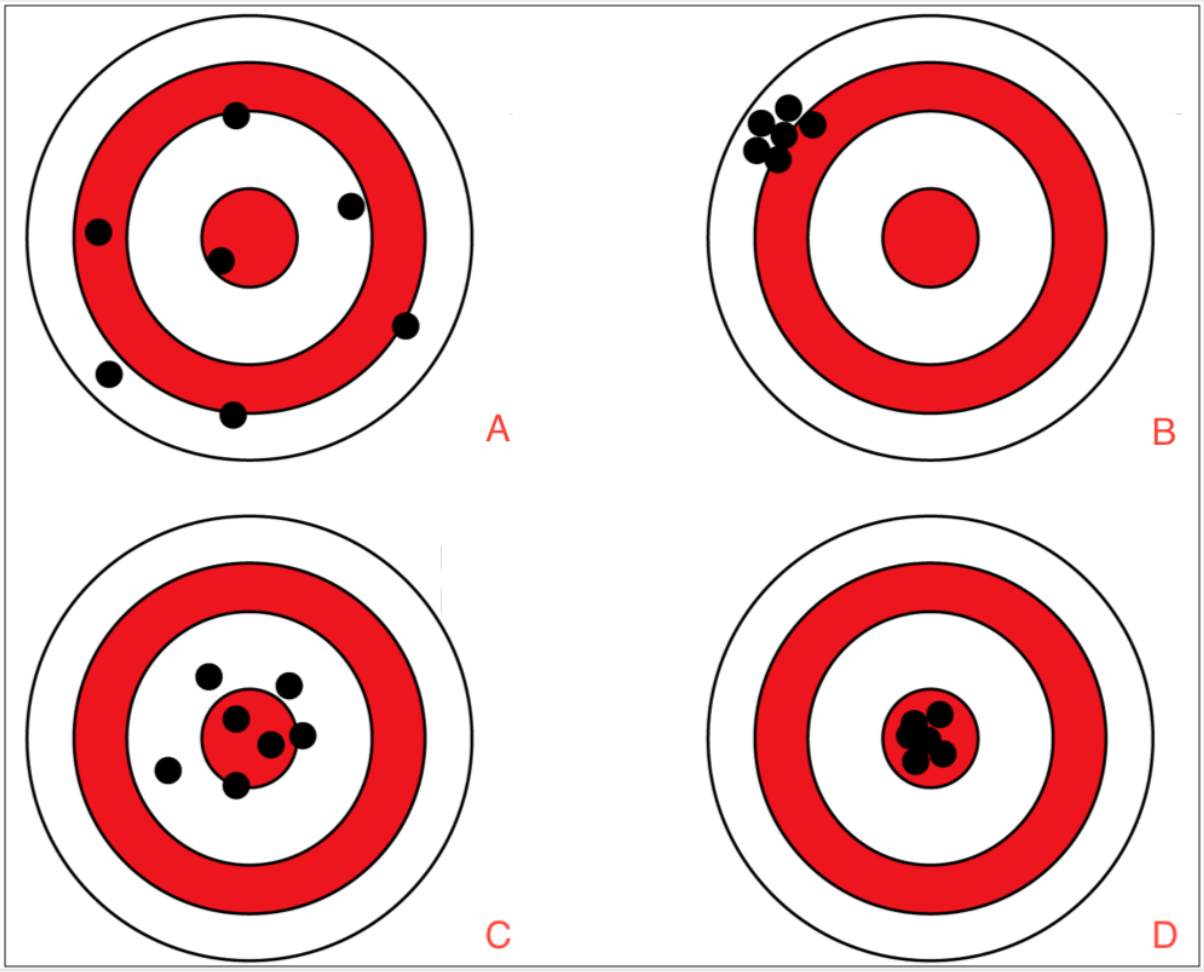
- Attributes of a record ALWAYS stay together.

_____

8. Who determines the fitness for use of your data?

- ☐ The museum or department director
- ☐ The users of the data for research or education
- ☐ The collector of the data in the field
- ☐ The person entering the data into the database

9. For the given statements, input the matching image. (A, B, C, D)



- High accuracy, low precision

_____

- Low accuracy, high precision

_____

- High accuracy, high precision

_____

- Low accuracy, low precision

_____

10. Identify the data relationships where Dataset B needs to be merged into Dataset A (0:1, 1:0, 1:1, 1:∞, ∞:1, ∞:∞). Not all the relationships are used.

- Collector field exists in both dataset A and B

———————

- Country field only exists in dataset B

  ———————
- Name field exists in dataset A, but dataset B contains First Name and Last Name fields

  ———————
- ID field exists in both dataset A and B

  ———————
- Elevation exists in dataset A, but not in dataset B

  ———————
- Date exists in dataset A, but Day, Month, and Year are separate fields in dataset B

  ———————

11. Metadata is important because (select the TRUE statements):

- ☐ it allows users to determine if a dataset is fit for their use.
- ☐ it allows you to share exact coordinates for each occurrence.
- ☐ it allows you to know under which legal terms the reuse of data is permitted.
- ☐ it also applies to all supplemental and associated materials, including images, video, and other media.
- ☐ it allows you know about the institution's next exhibition/opening hours.

Use Case I - Herbarium Specimens

Familiarize yourself with the use case scenario.

Use Case I is a practice use case for the planning, data capture, data management and data publishing modules. It is recommended that you download the exercise sheet (MS Word 345 KB) so that you can make notes as you work through the exercises. A suggested solution will be provided in the solution appendix. Use Case I is not graded.

Scenario

A Data Mobilization Project in a Regional Herbarium

Eriocaulon bilobatum Morong *collected in Guatemala by Rapid Reference Collection (RRC) | Field Museum of Natural History - Keller Action Science Center (licensed under* CC-BY-NC 4.0*)*

This narrative was developed as a basis for practical exercises in the biodiversity data mobilization course and the exercise concept and content was developed by Alberto González-Talaván, Néstor Beltrán, Nicolas Noé, Sharon Grant. The data are from a real dataset, but have been modified for the purposes of these exercises. It is a fictionalized scenario and is meant only for instructional purposes.

Description

The University of White Plains is a well recognized tertiary education institution in Guatemala and a national reference for biodiversity research. The Plant Biology Department keeps a medium-size herbarium containing approximately 80,000 specimens collected in and around the local region and dating from the mid-20th century to the present. The collection includes important specimens including types and endemics.

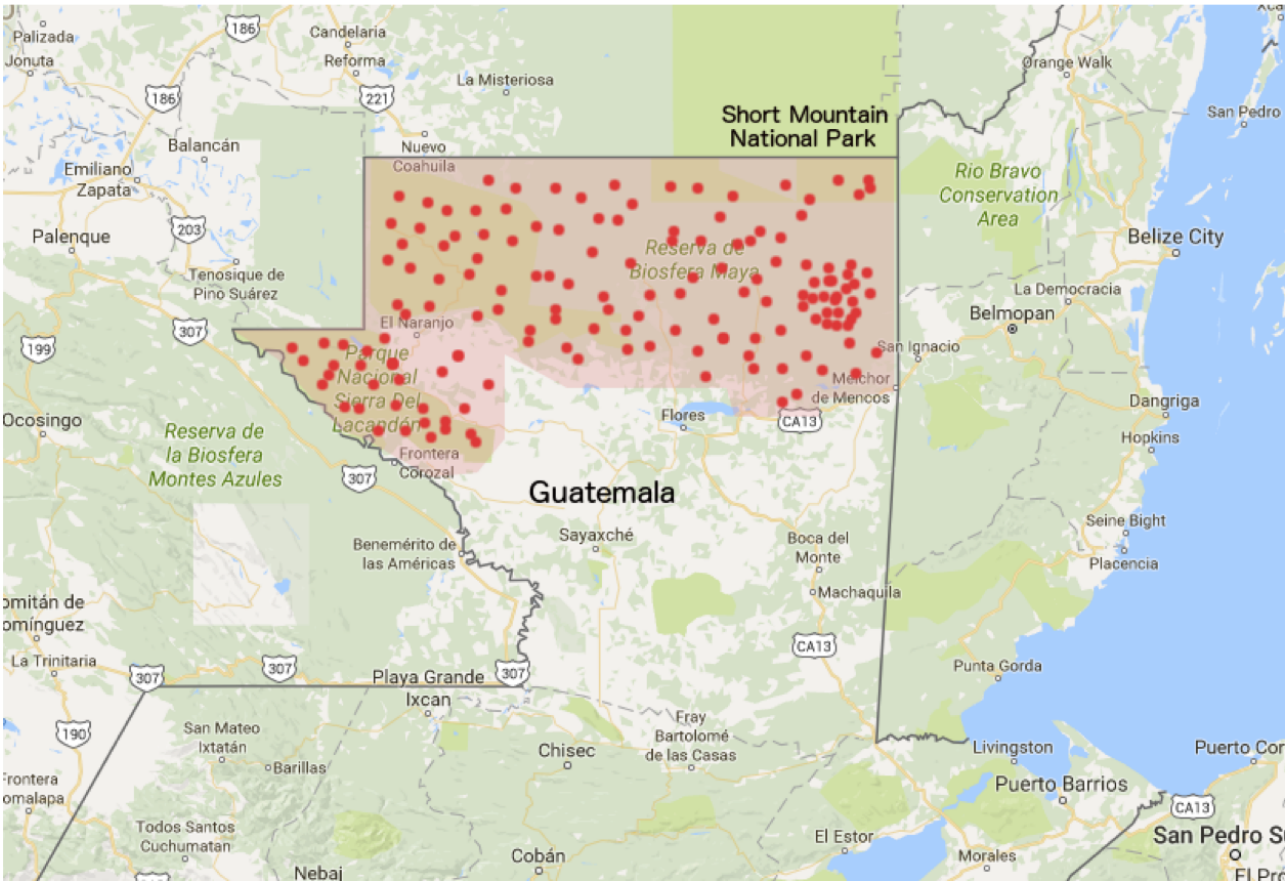
Currently, the care of the collection is assigned to the Professor of Plant Systematics, who performs curatorial tasks as well as their regular research and teaching work. The Departmental Admin is responsible for day-to-day administrative tasks for the herbarium such as purchasing consumables including paper and labels. Faculty staff and students collaborating within the department occasionally work on and update the identifications on the specimens and there are two retired botanists who regularly volunteer in the collection assisting the Professor to prepare loans.

The University already has an online search for its library collections which is maintained by the University's central web-team on externally hosted servers. No natural history specimens are currently served via this platform.

The Head of the Plant Biology Department has recently secured a 50,000USD, two year grant to image and publish the botanical collections information online. The team wants to use this opportunity to establish a permanent digitization and publishing protocol that will give higher visibility

to the herbarium and continue attracting funding.

Data collection



The herbarium comprises approximately 80,000 specimens collected from 1960 to 2015 concentrated mainly in the biodiversity hot spot – Short Mountain National Park. The collection is still growing as a result of exchanges, donations and several active research projects. For each specimen, information about collector, time, date, location and taxonomy are documented. So far no quality control measures have been carried out on the data and there has been no systematic imaging of specimens.

Dataset description

The professor has a simple non-relational database on his computer which serves as an index of many but not all specimens. Any images of the specimens that exist are held locally by the individual researchers that took them.

Exercises

The individual exercises for this use case are located in the corresponding modules.

Planning

Exercise 1a-c

Data capture

Exercise 2

Data management

Exercise 3a-c

Data publishing

Exercise 4

Exercise sheet

UC1-Herbarium-Exercise_EN.docx (MS Word 345 KB)

Planning

In the planning module, you will review key project planning stages and will learn how to create a viable workflow. Additionally, you will create an idealized project plan/workflow based on USE CASE I. You will identify goals, tasks, key stakeholders and roles and will assign specific tasks to stages.

Resourcing

In this video (15:09), you will learn how to define the elements that will have an impact on a project and will look at the interactions between the elements. If you are unable to watch the embedded video, you can download it locally. (MP4 - 25.1 MB)

▶ https://www.youtube.com/watch?v=VRvUdMjd93c *(YouTube video)*

Organizing

This video (15:52) focuses on tasks and the ways that they can be organized in order to prepare practical workable plans and clear documentation. If you are unable to watch the embedded video, you can download it locally. (MP4 - 26.2 MB)

▶ https://www.youtube.com/watch?v=uhhK6B2Vwls *(YouTube video)*

Exercise 1a-c

For this activity, you will be using a virtual tabletop platform. Review the exercise instructions below.

References

- Role definitions
- Stage definitions

To help you to work through the planning exercises for all use cases, we have made available five virtual playing card tables.

Because these are shared spaces, we ask that you follow these rules so that everyone who wants to

use them can:

1. Clean up after yourself. If you use a card table, please be sure to return it to it's clean and original state when you are done. This means putting all of the cards away and removing/deleting any new cards that you have added to the deck. You can do this by clicking the recall buttons below each deck of cards.
2. If it's busy, try another table. If you see that a card table is in use, please use one of the other card table links. If you are unable to find an empty card table, alert us at training@gbif.org so that we can check the tables to be sure that any tables were not left unattended (or if we need to create an additional table for you).
3. Don't leave a table unattended. Please do not leave your work on a table sitting out for long periods of time. If you need to step away for an hour or two, that is acceptable, but please do not leave the cards unattended for more than 4 hours. If you need to leave your set up for an extended period of time, please use the eLearning platform to notify the group and trainers which link you are using, why you need to leave your cards, and when you will return to complete your work.
4. Take screen captures often. We cannot guarantee that your cards will not be reset accidentally by someone from the workshop or by someone from your own project. You cannot reset the cards once they have been recalled and reset. If you want to submit a screenshot with your exercise please do.
5. The trainers reserve the right to reset the card tables. If we observe that one or more tables contain cards that have not been put away and have not been used for more than four hours without any notifications in the Forum on the eLearning platform, we will clear and reset the table for others to use.

Gameroom links (each table will open in a new window/tab):

BLUE: https://playingcards.io/ndqppx

GREEN: https://playingcards.io/zpwe9j

ORANGE: https://playingcards.io/868jgh

PURPLE: https://playingcards.io/g6khb7

RED: https://playingcards.io/3pad4w

YELLOW: https://playingcards.io/w488nb

Much appreciation and thanks to Jwalant Patel and Eric Ma for finding and helping to create the online playing tables for this exercise and to Kate Webbink for artistic expertise. The online tabletop platform is provided by https://playingcards.io/.

Exercise 1a

Read USE CASE I (if you haven't already).

Using the cards select the goals that mostly closely match those the project outlines and then choose the tasks that would need to be carried out to complete them. Next, identify the people/resources that the project has available to it and assign role cards to them as appropriate. Lastly, assign these to their stakeholder groups and affiliations.

1. Review the GOALS cards with the group, select and lay out the ones that fit the use case.

2. Review the TASK cards with the group.
3. Assign the TASK cards to each of the GOAL cards selected from the use case description.
4. Identify the institutions and people mentioned in the use case text and make note of them.
5. Lay out the AFFILIATIONS cards on the table.
6. Review the STAKEHOLDER cards, identify any mentioned in the use case and then decide which AFFILIATION they belong to.
7. Review the ROLE cards, identify any mentioned in the use case and decide which STAKEHOLDER group they belong to.
8. Make notes to cards as appropriate.
9. Once the cards are assigned take pictures/screenshots.
10. Use the exercise sheet to provide your answers.

Question

Are there resources or goals missing from the cards that you feel are critical to the successful completion of the project? Note these on the answer sheet.

Exercise 1b

Using the stakeholder and goal analysis from exercise 1a, develop a workflow using the STAGE cards.

1. Re-read the use case introduction, if necessary.
2. Decide which TASKS for which each ROLE will be responsible.
3. Review the STAGE cards and order the TASKS appropriately.
4. Use the previously downloaded exercise sheet to provide your answers.

Questions

- Are there obvious bottlenecks in the workflow? For example: Are there too many tasks for a particular role/resource?
- What issues do you think would be important to the successful mobilization of data from the points of view of each stakeholder/role? For example: What are the deliverables? Thinking about the general project timeframe, are they realistic?
- Consolidate notes and prioritize in order of importance.
- If you have time you can explore different combinations as different scenarios are possible in different contexts or even try to document the situation for your own project.

Exercise 1c

This exercise should be used when the course is taught virtually or onsite as a group activity.

After the exercises the presenter of each group will:

1. Present any missing stakeholders and/or tasks saying why they were added.
2. Highlight the two most critical issues/topics that the group has identified.

Potential discussion points across groups:

- What similarities and differences can be seen in the flows?
- Were there common issues that came up across the groups?

Review

Quiz yourself on the concepts learned in this section.

1. What is the order of the five PMBoK Process Groupings?

   - ☐ Planning, Initiating, Monitoring and Controlling, Executing, Closing
   - ☐ Initiating, Planning, Executing, Monitoring and Controlling, Closing
   - ☐ Initiating, Planning, Executing, Closing, Monitoring and Controlling
   - ☐ Initiating, Planning, Monitoring and Controlling, Executing, Closing

2. What are the types of deliverables? (multiple correct answers)

   - ☐ Stated
   - ☐ Implied
   - ☐ Estimated
   - ☐ Direct
   - ☐ Indirect
   - ☐ Guesses

3. What is a bottleneck?

   - ☐ a blockage that delays development or progress
   - ☐ a space where something or someone is missing
   - ☐ a problem, or situation that prevents somebody from doing something, or that makes something impossible.

4. Which are examples of mobilization tasks? (multiple correct answers)

   - ☐ Affiliation
   - ☐ Publishing
   - ☐ Imaging
   - ☐ Georeferencing
   - ☐ Increased Public Awareness

Data capture

In this module, you will learn about the concept of standards, in particular, the Darwin Core Standard and its components. You will also learn the types of primary biodiversity data and how to best share that information within GBIF. Lastly, you will

review principles of data quality in the context of data capture and will learn about data quality and coherence (especially on subjects such as georeferencing, dates, names and taxa cross-checking).

Standards and Darwin Core

In this video (15:37), you will learn how you interact with standards every day. Then you will be introduced to Biodiversity Information Standards, including the Darwin Core Standard with which you will continue to use throughout this course. If you are unable to watch the embedded video, you can download it locally. (MP4 - 27 MB)

▶ https://www.youtube.com/watch?v=S02PJHPsRAs *(YouTube video)*

Data origins and types

In this video (10:45), you will review primary biodiversity data that can be shared within GBIF. If you are unable to watch the embedded video, you can download it locally. (MP4 - 19 MB)

▶ https://www.youtube.com/watch?v=wKeOveydjsw *(YouTube video)*

Questions

- Is your data type different than you originally thought?
- With what kind of data do you work?
- How would you publish your data to GBIF (using which Core and/or extension)?

Data capture, processing and quality

In this video (09:11), you will explore the principles of data quality applied to data capture, specifically when capturing data from collection labels, fieldwork notebooks, spreadsheets, etc. If you are unable to watch the embedded video, you can download it locally. (MP4 - 19 MB)

▶ https://www.youtube.com/watch?v=QkDJIkmwBMA *(YouTube video)*

Exercise 2

For this activity, you will complete an exercise simulating data capture. You will begin to work with Darwin Core terms and make decisions on data that is needed for your organization/project and you will consider which of that data will be shared later during publication.

Read USE CASE I (if you haven't already).

Imagine that you are the person assigned to transcribe the data found on the herbarium sheets.

1. Download UC1-2-base-material.zip. (34.4 MB). There are 10 images. Two images per specimen for

a total of five specimens. The herbarium sheets are in Spanish (data may come to you in various means and in other languages than your own), but you should be able to recognize the data contained in the fields on the labels. Remember to use both images per record to compile the information.

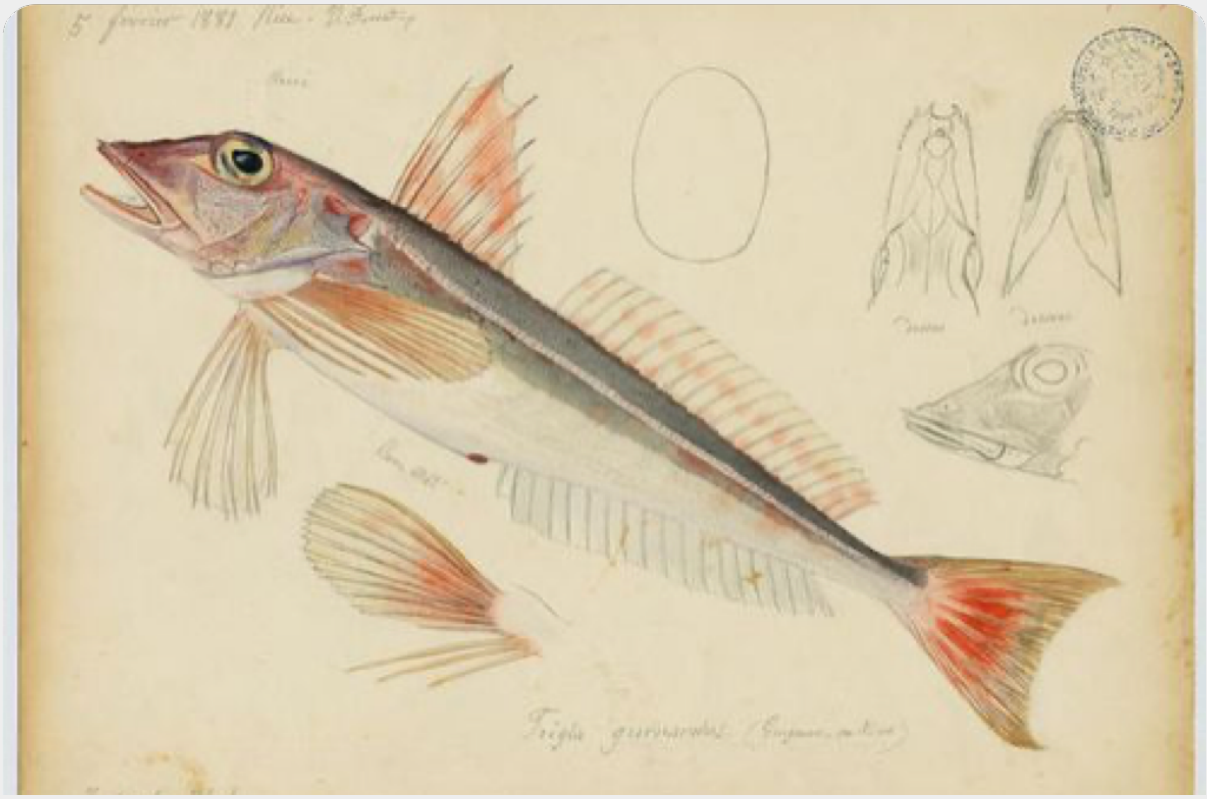
2. Download the spreadsheet template: UC1-2-occurrence-template.xlsx (57.3 KB) to transcribe the information found on each of the images for the five specimens.
3. Use the previously downloaded exercise sheet to provide your answers.

you may need to add fields to the spreadsheet as you may be able to capture more information from the labels than was planned for in the template.

Review

Quiz yourself on the concepts learned in this section.

1. What dataset type(s) would you choose for an ichthyology collection?



Eutrigla gurnardus (Linnaeus, 1758) | *Muséum d'histoire naturelle de Nice*

- ☐ occurrence
- ☐ checklist
- ☐ sampling event

2. What dataset type(s) would you choose for a list of invasive species?

Water hyacinth (Eichhornia crassipes) observed in Bourail, New Caledonia, where it is an introduced and invasive species by GRIIS. Photo by gérard (2016) licensed under CC BY-SA 2.0

- ☐ occurrence
- ☐ checklist
- ☐ sampling event

3. What dataset type(s) would you choose for the flora and fauna of an environmental impact study?

   Environmental impact assessment studies are done by experts in order to assess the biodiversity and biotopes of a given area, before, during and after it is affected by human activities (road works, wind turbines, mining, building construction, etc.).

Entomologist chasing butterflies *by Matthieu Gauvain (CC-BY-SA)*

- ☐ occurrence
- ☐ checklist
- ☐ sampling event

4. What dataset type(s) would you choose for bird tracking data?

   Bird-tracking data are recorded using specific devices, such as GPS trackers mounted on live birds, thus allowing scientists to track their migratory routes or breeding sites.

Griffin vulture *observed at Gamla Nature Reserve by גיזונית - MinoZig (CC0)*

- ☐ occurrence
- ☐ checklist
- ☐ sampling event

5. What dataset type(s) would you choose for insect trap data?

Insect trap *by miheco (CC-BY-SA)*

- ☐ occurrence

- ☐ checklist
- ☐ sampling event

6. What dataset type(s) would you choose for national park management data?

   Data acquired in the context of protected areas management (such as national parks but also smaller nature reserves) can be diverse and have different origins: botanical surveys, tagged animals tracking, observations from rangers and guards, and even 'citizen science' data or data inferred from pictures shared on social medias.

Sri Lankan elephants *observed by pen_ash.*

- ☐ occurrence
- ☐ checklist
- ☐ sampling event

7. What dataset type(s) would you choose for a citizen science bioblitz?

   Citizen science data are often collected through thematic fieldwork days known as a "bioblitz." Volunteers typically gather in a given area and spend the day trying to observe and identify as many species as they can in this area.

   Data from each participant are captured and merged into the citizen science programme's data capture or data management tool.

Looking for birds with park staff *by US National Park Service (authorized reuse on google image search)*

- ☐ occurrence
- ☐ checklist
- ☐ sampling event

8. What dataset type(s) would you choose for a regional species list?

Black rhino *observed at the Magdeburg Zoo in Germany by Mani300*

- ☐ occurrence
- ☐ checklist
- ☐ sampling event

Data management

In this module, you will review the main concepts, related tools and best practices for data management, particularly, data cleaning and standardization.

Principles of data management

In this video (09:49), you will review an important set of principles necessary to improve data through the processes of data cleaning. If you are unable to watch the embedded video, you can download it locally. (MP4 - 16.6 MB)

▶ https://www.youtube.com/watch?v=4ijm1cJeVHE *(YouTube video)*

Data management tools

In this video (06:42), you will learn about a variety of tools that you can use to improve the quality of your data. If you are unable to watch the embedded video, you can download it locally. (MP4 - 10.3 MB)

▶ https://www.youtube.com/watch?v=Ru3vWiYU3gw *(YouTube video)*

Exercise 3a-c

For these exercises, you will perform technical and consistency validation checks, improve data with different tools, and learn how to use OpenRefine.

Read USE CASE I (if you haven't already).

Your institution is part of the "Global Poales Association (GPA)". This association has secured funding to publish an up-to-date flora on the group and has requested your herbarium to participate and provide any high quality records you may have on this order of plants. The order is well represented in your collection so you think you could contribute substantially to this effort.

Exercise 3a

Validation checks

In this exercise we will focus on technical errors and perform a basic validation check to identify technical errors. Refer to Validation checks for information on the types of errors.

1. Download UC1-3ab-data-cleaning.csv. (207.5 KB)
2. Import the CSV file in Excel using the Excel wizard. See Excel-tips-EN.pdf (PDF, 7 MB) for import instructions for your operating system (Windows, Mac, Linux).

3. Find and correct the errors manually.
4. Use the previously downloaded exercise sheet to provide your answers.

Exercise 3b

Other data management tools

The GPA association has given you a checklist of data quality elements to verify:

- All plant names (full name) are correctly spelled
- All plant names belong to the order
- All records have coordinates
- All coordinates are inside the country stated and converted to decimal format
- All dates are in the proper column and in the format YYYY-MM-DD

The three categories of errors are:

- Nomenclatural errors
- Format errors
- Geographic errors / outliers

1. Refer to Helpful tools in order to complete the exercise. You are not limited to these tools, you may use any tools you like.
2. Use the same file from the previous exercise.
3. Make the correction ONLY for the Eriocaulaceae family (so you may want to filter the data)
4. Correct the errors found in the dataset used in exercise 3a (previous exercise), using the tools of your choice, and document the changes you perform in the exercise sheet.
5. Correct the entire file if you have time.
6. Use the previously downloaded exercise sheet to provide your answers.

Exercise 3c

In this video (03:27), you will learn about OpenRefine. You can use OpenRefine to standardize and improve the quality of your data. If you are unable to watch the embedded video, you can download it locally. (MP4 - 3.8 MB)

▶ https://www.youtube.com/watch?v=_YFw_bfwc3Y *(YouTube video)*

OpenRefine

In this exercise we use OpenRefine to improve the quality of a dataset by using the default features, existing web services and regular expressions.

1. Download UC1-3c-open-refine.csv. (207.5 KB)
2. Download and complete the exercises in OpenRefine-Exercise3c-EN.pdf. (PDF, 1.1 MB) Also available in French and Spanish.
3. Use the previously downloaded exercise sheet to provide your answers.

Exercise tips

Validation checks

Technical errors Relatively simple, often able to be automated, **checks against the integrity of the data**. These may indicate incorrect exports, data mapping, field slippage (e.g. moving 1 column to the right) or data missing at the source.

- **Completeness**: Whether all the data and metadata is available – are all fields present, are all fields filled out?
- **Bounds**: For example, are days given in the range 1-31 (depending on month)
- **Data type**: For example, does the Date field contain a date or a number?
- **Data format**: For example, are Dates provided as 01/01/2010 or 01/Jan/10?

Consistency errors

Application of real-world rules to the data. These may indicate incorrect data entry from older records, transcription errors or post processing. Some are complex to implement and **require reference data sets to check against**. E.g. a list of known collectors and collecting habits. These rules can be gathered from data users and analysts.

- **Taxonomic**: For example, if identified to species level, have a binomial scientific name and entries in genus and species fields been provided?
- **Currency**: Are dates of collection, identification, update and digitization consistent?
- **Outliers**: Detect outliers, but remember that not all outliers are necessarily errors. For example, compare against a known species range, or known environmental range (but remember that outliers may be misidentifications, rather than incorrect coordinates).
- **Geographic**: Are the coordinates within the identified locality or region? For example, are there any terrestrial occurrences in the sea or marine occurrences on land?
- **Collecting patterns**: Does the occurrence detail match the known collecting patterns of the organization or collector? Do any records appear to have been created after a collector has died (could this possibly be a different collector with a similar name)? For example, are any mammal records attributed to a bird watching group?
- **Accuracy and precision**: For example, are any georeferenced records indicating very high precision or accuracy from a pre-GPS (or pre-accurate GPS) collecting period?
- **Collecting methods**: Different survey methods (e.g. transects and area surveys) have particular characteristics. Are the records consistent with the method provided?

Helpful tools

- **GBIF Name Parser**: https://www.gbif.org/tools/name-parser
- **Global Names Resolver**: http://resolver.globalnames.org
- **Catalogue of Life name match**: https://data.catalogueoflife.org/tools/name-match
- **TNRS**: https://tnrs.biendata.org/
- **WoRMS**: https://www.marinespecies.org/aphia.php?p=match
- **InfoXY**: http://splink.cria.org.br/infoxy?criaLANG=en
- **Georeferencing Calculator**: http://georeferencing.org/georefcalculator/gc.html

- **Canadensys coordinate conversion**: http://data.canadensys.net/tools/coordinates
- **Canadensys date parsing**: http://data.canadensys.net/tools/dates
- **Google Maps**: https://maps.google.com/

Review

Quiz yourself on the concepts learned in this section.

1. Why is it best to clean your data?
   - ☐ to make them as fit for use as possible
   - ☐ to achieve your data quality goals
   - ☐ data should be cleaned by the users, not the providers
2. How should you organize your data cleaning workflow?
   - ☐ work alone, you know your data best
   - ☐ ask your colleagues for expertise
   - ☐ work at an institutional level to harmonize data quality workflows
3. Which is best:
   - ☐ prevent errors from occurring
   - ☐ correct errors as soon as you find them in your database or spreadsheet
   - ☐ not cleaning errors but documenting them as you go, so people who reuse your data know where they are
4. Whose responsibility is data quality?
   - ☐ The person(s) who record data on the field
   - ☐ The data transcribers
   - ☐ The database manager
   - ☐ Everyone involved in the management of data
   - ☐ The people who use your data
   - ☐ GBIF
5. Which tools can be used to clean your data ?
   - ☐ Excel & other spreadsheets management tools
   - ☐ OpenRefine
   - ☐ Your database software
   - ☐ Online tools such as Scientific Names Resolver or Google Maps

Data publishing

In this module, you will learn about data publishing concepts, including the IPT, cores and extensions, and the importance of licenses, metadata, mandatory fields and hosting of datasets.

Data publishing concepts

In this video (11:45), you will learn about data publishing concepts and will receive an introduction to the Integrated Publishing Toolkit (IPT). If you are unable to watch the embedded video, you can download it locally. (MP4 - 20 MB)

▶ https://www.youtube.com/watch?v=b900d9ukjSQ *(YouTube video)*

IPT overview

In this video (06:56), you will receive an overview of the IPT data publishing interface. If you are unable to watch the embedded video, you can download it locally. (MP4 - 8.7 MB)

▶ https://www.youtube.com/watch?v=gHXsaN_JWeI *(YouTube video)*

Training IPT installations

If you have not already been provided a login, please contact training@gbif.org and you will be provided with a login and password on the course IPT.

https://training-ipt-a.gbif.org/

https://training-ipt-b.gbif.org/

https://training-ipt-c.gbif.org/

IPT demonstration

In this video (24:16), you will learn how to publish an occurrence dataset using an IPT. If you are unable to watch the embedded video, you can download it locally. (MP4 - 52.6 MB)

▶ https://www.youtube.com/watch?v=eDH9IoTrMVE *(YouTube video)*

Exercise 4

In this exercise, you will publish an occurrence dataset using the IPT.

Read USE CASE I (if you haven't already).

Data publishing

After cleaning the data in the Poales dataset, the team considers that publishing the data online

through the GBIF network could be a good way to make this effort visible. You have been requested to lead that publishing work, based on the dataset.

1. At this point, you need an account on one of the course IPTs. If you have not already been provided a login, please contact training@gbif.org and you will be provided with a login and password on one of the course IPTs.
2. Download UC1-4-poales-publishing.csv. (233.5 KB)
3. Use the assigned IPT installation and publish this file.
4. Use the previously downloaded exercise sheet to provide your answers.

Review

Quiz yourself on the concepts learned in this section.

1. What does data publishing mean in the context of GBIF?
   - ☐ Exporting a csv file of your cleaned data that you can share with your colleagues
   - ☐ Writing an article describing your data, and the protocol(s) you used to collect, capture and clean them
   - ☐ Making your biodiversity dataset(s) publicly accessible and discoverable in a standardized format
2. What is an IPT?
   - ☐ a tool that helps you manage & correct your data
   - ☐ a tool that helps you publish your data to GBIF
   - ☐ a tool that helps you produce a Data paper
3. Which Creative Commons licences and waivers are recommended by GBIF for data publication?
   - ☐ CC-BY, CC-BY-SA and CC-BY-ND
   - ☐ CC0, CC-BY and CC-BY-NC
   - ☐ CC0, CC-BY and CC-BY-SA
4. What are the three Cores from which you can choose for an IPT resource?
   - ☐ Metadata Core, Occurrence Core, Multimedia Core
   - ☐ Taxon Core, Collection Core, MeasurementOrFact Core
   - ☐ Occurrence Core, Taxon Core, Event Core
5. How many Extensions files can a dataset have?
   - ☐ zero
   - ☐ one
   - ☐ as many as needed

Assessment and certification

In this module, you will review the criteria to be used to assess assignments and to achieve certification.

Upon successful completion of the course and successful assessment of assignments (by trainers and mentors), participants have the opportunity to receive an official certification in the form of an Open Badge.

An overall score of 2.5-2.9 earns a BASIC Biodiversity Data Mobilization badge

An overall score of 3.0-4.0 earns an ADVANCED Biodiversity Data Mobilization badge

Participants are required to submit (in English) Use Case II (choice between two options) and Use Case III and each Use Case is scored against the educational rubrics for the course. The rubrics define the skills and performance levels on which the learning objectives for the course are built.

Review the rubrics to ensure understanding of the skills that will be assessed for certification.

Planning rubric

Planning

Skills	Beginning performance 1	Developing performance 2	Accomplished performance 3	Outstanding performance 4
A. Knowledge of the different elements that are part of a biodiversity data mobilization plan	Shows understanding of a few roles and tasks required in a solid mobilization plan, but is not able to differentiate who performs each task in the plan or mixes up the roles.	Understands many of the roles and tasks needed in a mobilization plan, but still misses links and interactions between them.	Understands the key tasks and roles needed in a mobilization plan and how they interact.	Can identify additional roles and task that may be needed for specific situations

Skills	Beginning performance 1	Developing performance 2	Accomplished performance 3	Outstanding performance 4
B. Capacity to apply the different elements of a biodiversity data mobilization plan to a given institutional context (e.g. their own)	Can see how a few of the tasks and roles relate to a given institutional context, but has difficulties when the context changes even slightly from the generic references used.	Can see how most of the tasks and roles can be applied to a given institutional context, but still has gaps when the situation is different to the generic references used.	Can identify all relevant tasks and roles and how they can be translated to a given institutional context, even if the situation is not identical to the one in the generic references used.	Can apply the different elements of a mobilization plan creatively, incorporating new elements not found in the generic references used.
C. Ability to write/compose a clear planning document	Can gather raw material that would be the basis for a plan (e.g. as bullet points), but has difficulties translating this into a narrative.	Can construct a basic work plan, but lacks coherence between its elements.	Can write a coherent and well articulated document which includes all relevant elements of a work plan.	Can write concise plans, with effective summaries and increasing levels of detail
D. Capacity to evaluate a biodiversity data mobilization plan	Can identify a few elements of a generic mobilization plan in a given example, but has difficulties matching them to tasks and roles.	Can identify most elements of a mobilization plan and match them to tasks and roles when they are identical to references used. Has difficulties assessing the feasibility of some elements of the plan.	Can recognize all existing elements in a mobilization plan, and can identify gaps, duplicated efforts, and inconsistencies. Can evaluate the quality of the individual components. Can evaluate potential points of success and weak points of the plan as a whole.	Can suggest solutions to fix issues identified in the plan evaluated.

Data capture rubric

Data capture

Skills	Beginning performance 1	Developing performance 2	Accomplished performance 3	Outstanding performance 4
A. Ability to identify the type of digital data that can be extracted from a source of biodiversity data (i.e. that can be published using the GBIF network)	Can identify only the most evident data types from common sources of biodiversity data (e.g. occurrences from natural history collection specimens). Shows little understanding of potential for online publishing using GBIF.	Can frequently identify correctly, at least one digital data type that can be extracted of common sources of data. Has difficulty identifying which ones can be currently published using GBIF.	Can always identify one (or more) types of digital data that can be extracted from common sources of data. Can identify which one of those types can be currently published using GBIF.	Can always identify one or more types of digital data that can be extracted from common and uncommon sources of data. Can identify which one of those types can be currently published using GBIF and which ones are under discussion. Can identify data cores and extensions used for publishing those data types.
B. Capacity to extract relevant information from a source of biodiversity data into simple data structures (e.g. spreadsheets) that follows international standards	Can only extract large pieces of obvious information (e.g. all geographic information as a single unit) which are evident in the data source. Shows little knowledge of current standards for recording biodiversity data.	Can retrieve several information items from the data source (but not all) and can disaggregate them into meaningful pieces. Shows some basic knowledge of the most common standards (e.g. DwC) and the most used data fields in those standards.	Can identify all valuable information in a data source, and extract the mandatory elements in a standard data structure (e.g. a spreadsheet based on Simple DwC). Can identify missing information and infer from existing information (e.g. derive a country name from a province).	Can identify all valuable information in a complex data source, and divide it into meaningful pieces which then translate directly into international standards. Can identify critical information missing in the source and infer it from the existing data or from additional information about the source (metadata).

Skills	Beginning performance 1	Developing performance 2	Accomplished performance 3	Outstanding performance 4
C. Ability to understand and apply basic principles of data quality to the data capture process	Shows limited understanding of how applying simple data quality principles can have a large impact on the final product, preventing additional required cleaning afterwards.	Knows some of most generic principles of data quality (e.g. avoid misspellings) but shows limited knowledge on how to apply more specific principles to the data capture process.	Knows all the basic principles of data quality and how to apply these in simple ways to the data capture process. Uses formats consistently during the data capture process (e.g. in dates, country names). Documents all procedures and changes connected to data quality in a simple manner.	Shows good knowledge of all common principles of data quality and how to use them to improve the data capture process. Uses data formats consistently and can use gazetteers, reference lists, or software-specific features to improve quality from the original. Documents clearly all changes and decisions taken in connection to data quality.

Data management rubric

Data management

Skills	Beginning performance 1	Developing performance 2	Accomplished performance 3	Outstanding performance 4
A. Capacity to assess the quality (i.e. identify issues and their types) of a biodiversity dataset.	Only uses visual checks to analyse quality. Cannot differentiate between types of errors. Can detect missing values in required fields and severe data inconsistencies.	Can only use very basic techniques (e.g. sorting) to analyse data quality. Can detect mismatches between field names and content. Can consistently identify technical errors, but only the most typical consistency errors in a dataset.	Can use specific tools and techniques to assess quality. Recognizes the minimum level of disaggregation/normalization needed for common use and publishing. Can consistently identify technical errors and most of the consistency errors in a dataset.	Uses a systematic approach to dataset analysis covering all major data domains. Can consistently identify both technical and consistency errors in a dataset. Can use other sources of data (e.g. metadata or other datasets) to identify or infer consistency errors in a dataset.

Skills	Beginning performance 1	Developing performance 2	Accomplished performance 3	Outstanding performance 4
B. Capacity to perform data format correction.	Can only make corrections manually in the tables. Shows generic knowledge about use of format types in digital data (e.g. dates, strings, numbers)	Can identify at least one specific tool to automatically correct format errors, but can only use it in specific cases. Otherwise, uses simple mechanisms (e.g. 'find & replace') to solve issues.	Can use at least one tool to automatically correct format errors.	Can use advanced features of more than one tool to correct format errors.
C. Capacity to perform nomenclatural data correction.	Can only make corrections manually in the tables. Only uses personal knowledge of known taxonomic groups.	Can identify at least one specific tool to automatically correct nomenclatural errors, but can only use it in specific cases. Otherwise, uses simple mechanisms (e.g. 'find & replace') to solve issues.	Can use at least one tool to automatically correct nomenclatural errors. Can find and use suitable reference nomenclatural information for the taxonomic groups with which (s)he usually works.	Can use more than one tool to correct nomenclatural errors. Can find and use suitable reference nomenclatural information for taxonomic groups outside of his/her areas of expertise.
D. Capacity to perform geographical data correction.	Can only make corrections manually in the tables. Only uses personal knowledge of known geographical areas.	Can identify at least one specific tool to map and/or automatically correct errors in geographical information, but can only use it in specific cases. Otherwise, uses simple mechanisms (e.g. 'find & replace3') to solve issues.	Can use at least one tool to map and/or automatically correct errors in geographical information. Can find and use suitable reference geographical information in a suitable format for the areas with which (s)he usually works.	Can use more than one tool to map and/or automatically correct errors in geographical information. Can find and use reference geographical information in a suitable format for areas outside of his/her areas of expertise.

Skills	Beginning performance 1	Developing performance 2	Accomplished performance 3	Outstanding performance 4
E. Capacity to use specific software (e.g. OpenRefine) as tools for data cleaning.	Can identify at least one data cleaning tool. Can identify the main features of a data cleaning tool (e.g. OpenRefine).	Can identify multiple data cleaning tools. Can use one or a few of the basic features of data cleaning software to clean a dataset (e.g. create an OpenRefine project, use faceting, filtering, clustering or reconciling).	Can use all the basic features of a data cleaning software to clean a dataset (e.g. in OpenRefine: faceting, filtering, clustering, reconciling).	Can use the advanced features of one or more data cleaning software packages to clean datasets (e.g. in OpenRefine: use API, regular expressions, Google Refine Expression Language).
F. Capacity to document data transformation procedures.	Seldom describes any changes made while curating, formatting, or transforming data.	Describes changes made most of the time. Doesn't describe changes consistently or fully (e.g. describes the change, but not the author).	Always remembers to describe changes made. Always describes changes consistently, so that all edits of the same type can be easily identified.	Can accurately and consistently describe changes made in a repeatable way.

Data publishing rubric

Data publishing

Skills	Beginning performance 1	Developing performance 2	Accomplished performance 3	Outstanding performance 4
A. Knowledge about biodiversity information (BDI) data standards.	Shows limited or no knowledge about BDI data standards and which of those data standards are accepted by GBIF.	Can identify BDI standards and knows which ones are accepted by GBIF, but does not know where to find information on how to use them. Cannot identify which terms are mandatory.	Knows the BDI standards accepted by GBIF. Can find a list of the accepted data cores and extensions. Publishes datasets according to the required and/or recommended GBIF standards for data and metadata terms and knows how to find the definitions of the terms.	Shows understanding about the characteristics and limitations of the various BDI standards.

Skills	Beginning performance 1	Developing performance 2	Accomplished performance 3	Outstanding performance 4
B. Capacity to analyse the suitability of a biodiversity dataset for publishing through GBIF.	Shows limited or no knowledge of the formal criteria that a dataset needs to meet to be publishable through GBIF.	Knows the formal criteria that a dataset needs to meet to be publishable through GBIF, but cannot assess if a given dataset meets them.	Can correctly assess if a dataset can be currently published through GBIF. Can assign at least one valid data type (=core) to a dataset based on the description provided by the data holder and after having analysed the dataset.	Can identify more than one publishing option for a dataset (where possible).
C. IPT use: capacity to produce/analyse high quality metadata.	Shows limited or no knowledge about the characteristics of good metadata.	Knows the characteristics of good metadata, but has difficulties recognizing them.	Knows the characteristics of good metadata and how to recognize them. Can produce recommendations on how to improve existing metadata.	Knows the characteristics of high-quality metadata and how to produce them.
D. IPT use: capacity to upload/connect data and map it to existing cores & extensions.	Can upload single-file datasets into IPT, but does not succeed to map them to any core.	Can only upload single-file datasets into IPT and map to a single type of core with no extensions.	Can upload multiple files into an IPT as part of a single dataset and map them correctly to a core and at least one extension. Can use the IPT constant value feature.	Can upload multiple files into an IPT as part of a single dataset and map them correctly to a core and multiple extensions. Can use the IPT data translation feature.
E. IPT use: capacity to use the tool to publish and register datasets.	Can view a published dataset and associated metadata on an IPT. Can download a DwC-A file from an IPT. Can navigate a registered dataset from the IPT to the GBIF portal.	Can update an existing, published dataset by uploading a new source file. Can republish the file, error free.	Can successfully publish and register a new dataset. Can understand and act upon publishing error messages in IPT.	Shows understanding of dataset versioning in IPT.

Use Case II - Invasive species

Familiarize yourself with the use case scenario.

You have a choice for Use Case II between two scenarios:

- Invasive species checklist
- Lepidoptera sampling

Your choice for Use Case II will be graded.

Scenario

Tracking invasive species

Leucaena leucocephala (Lam.) de Wit *observed in Hawaii by Sharon Grant (licensed under* CC-BY-NC 4.0*)*

This narrative was developed as a basis for practical exercises in the biodiversity data mobilization course and the exercise concept and content was developed by Sharon Grant, John Wieczorek, David Bloom and Laura Anne Russell.

It is a fictionalized scenario based on a real dataset and is meant only for instructional purposes. The original dataset is attributed to Simpson A (2016). Big Island Invasive Species Committee - Pest Reports - 2005-2010. Version 4.1. United States Geological Survey. Occurrence Dataset accessed via GBIF.org on 2017-07-13.

Description

The Hawaii Invasive Species Council (HISC) received a federal grant to collaborate with high schools as part of Hawaii's statewide curriculum on Invasive Species to increase local knowledge about invasive species, increase data collection and produce annotated checklists for under reported areas. A full-time Project Manager is employed to oversee the project. All funds and allocations are managed by the HISC Fiscal Associate.

The Manager of each island's Invasive Species Committee (ISC) received a sub-award to set up local

education programs and collect data. The programs trained high school students to become Student Mentors and to facilitate image and data collection by members of the local community. The Big Island Invasive Species Committee (BIISC) received a further sub-award to extend their central database to accommodate each ISC's data, provide participating schools with their own websites, and maintain a single, searchable data portal to serve government, public, and scholarly research efforts.

Two schools on each island were selected because they were located in areas where knowledge and documentation of invasive species assessment was poor or non-existent. Teachers worked with their local Invasive Species Council (ISC) Outreach Associate to create teaching materials detailing 21 important invasive plant species, including how to identify each species' life stages and the most effective control methods.

The Graduate School of the University of Hawaii in Maui (UHM) runs a course in community outreach. Four botany students from the University, as a part of their final course assessment, are validating the identifications from the images and descriptions submitted by each high school to their local ISC.

Data collection

Students from each high school organized a series of day-long community surveys in their local neighbourhoods. Participants, guided by local ISC Early Detection Technicians and Student Mentors, visited various locations where they were given photo guides and assigned a route to follow during collection events. Along each route, they were tasked with identifying the target species and taking 1-3 photos of them using GPS-enabled mobile phones.

Details, describing every observation of the 21 invasive species of interest, were recorded using a digital data collection form during each community collection event. Participants uploaded the images captured on mobile phones and were encouraged to click their locations using a Google map, embedded in the form, to assign latitude and longitude to each observation. The form's design was based on the HISC pest reporting form.

REPORTER INFORMATION

Report Number: T07222017172139

First Name: | Last Name:

Email: e.g. pestreport@643pest.com | Phone: e.g. 808-123-4567 or 8081234567

PEST SIGHTING INFORMATION

NOTE: Asterisk [*] and red label color indicate a required field.

*Date of Pest Sighting:

*Pest Name:

*Pest Description:
(Plant: size; flower, foliage, or fruit color / scent / orientation; habitat)
(Insect / animal: size; color; plant / host found on or nearby; habitat)

*Island of Pest Sighting:
(Please choose an island before entering information into the Location field).
--Select Island--

Location of Pest Sighting:
(Street address, cross streets, city, mile marker, place name or general area)

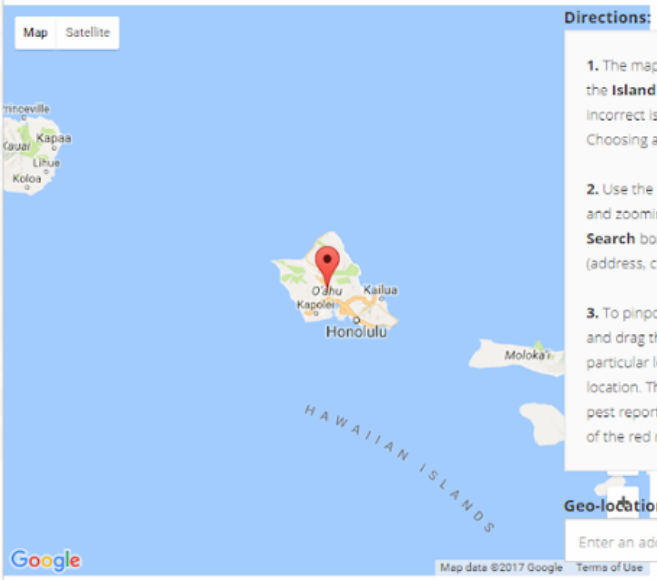
Additional Comments:

Image Upload (*Allowed File Types: .jpg, .png, .gif | Upload up to 3 images*):

Drag & drop files here ...

Browse ...

PEST SIGHTING LOCATION DETAILS



Directions:

1. The map will have automatically moved to the island chosen from the **Island** drop-down list above. If an island was not chosen or the incorrect island was chosen, please go back and make a new selection. Choosing an island is a **mandatory** step.

2. Use the **map tools** at left (Zoom In, Zoom Out, Pan) for help finding and zooming in to a desired area or location. Use the **Geo-location Search** box below to search for and zoom in to a specific geo-location (address, city, place name) in Hawaii.

3. To pinpoint the **exact location of the pest sighting**: a) click on and drag the red map marker to a particular location; or b) click on any particular location on the map to move the red map marker to that location. The **coordinates (Latitide, Longitude)** submitted with the pest report are shown in the boxes below and reflect the final position of the red map marker.

Geo-location Search:
Enter an address, city, place name...

Marker Coordinates:
Latitude | Longitude

Digital data description

A database, created and hosted by the Computing Department at UHM, holds the imagery and data from the online form, but these data are not accessible publicly. Data were exported as comma separated value (.csv) files and given to the four UHM graduate students for taxonomic validation using the images and descriptions submitted. The BIISC GIS Analyst used the Google Maps coordinates and image EXIF data to check observations for quality and to add any missing georeferences. Student mentors renamed all image files to match the observation number for cross referencing later at BIISC.

Invasives exercise sheet

Download the exercise sheet. (MS Word, 342 KB)

Exercise 1

Planning

You are the local ISC Manager and, because of the success of the lesson plans and community surveys, 10 more schools on your island would like to set up their own projects the following year. You would like to accommodate them, but your ISC funding will expire at the end of this year. HISC has indicated that they will look favourably on a small grant application to expand your programs in the following year and BIISC has offered support.

Exercise 1a

Analyse the financial implications of expanding the number of schools

1. Evaluate the following options to expand the number of participating schools. You can only select TWO of these options, so you need to choose wisely.
2. Use the exercise sheet to provide your answers.

Options

1. Pay extra summer interns to work at the local ISC to coordinate surveys .
2. Offer financial support to BIISC to set up websites for each new school.
3. Offer financial compensation to the graduate students. You will not be able to pay all four of them the equivalent of a regular salary, but could cover the costs of part time positions for two of them.
4. Contract a software company to build a database that can automatically ingest data directly from the online form. The system will include an admin interface to allow data manipulation and csv exports.
5. Fund four public outreach activities (e.g., a BioBlitz) to promote awareness in the communities and increase volunteer participation.
6. Prepare and carry out a reusable training a course for the teachers at the schools to teach them how to prepare data for submission to BIISC.

Exercise 1b

Assign roles

The new project has the following people available for data processing and mobilization.

1. Assign roles to maximize the efficiency of the data processing and transformation to produce data of the highest quality as efficiently as possible?
2. Use the exercise sheet to provide your answers.

Roles

- BIISC GIS Analyst: Advanced computer use, GIS and data analysis tools.
- ISC Manager: Good computer skills.
- ISC Outreach Associate: Good field identification skills; Basic computer use. Social media expert.
- Student Mentor 1: Basic taxonomic knowledge. Basic computer use.
- Student Mentor 2: Basic taxonomic knowledge. Basic computer use.
- Botany Student 1: Advanced taxonomic knowledge. Programming skills.
- Botany Student 2: Advanced taxonomic knowledge.
- Botany Student 3: Advanced taxonomic knowledge.
- Botany student 4: Advanced taxonomic knowledge.

Exercise 2

Data capture

The BIISC is now planning to make all of the data from the project publicly available by publishing datasets to GBIF. As the BIISC's Outreach Assistant, you must identify the relevant Darwin Core fields to accommodate the data from the online form. You've noticed that additional data describing species and locations have been added to the data form by the graduate students performing validations. To accommodate this data, you need to extend the data structure to aggregate the data from the online form with the added taxonomy and georeferences.

1. Download UC2-IS-2-ForCapture.csv. (7.1 MB)
2. Using the downloaded dataset, produce a spreadsheet as example of the extended data structure and the fields you've identified as relevant for Darwin Core.
3. Use the exercise sheet to provide your answers and submit the spreadsheet.

Exercise 3

Data management

Over the summer, interns at the HISC main office created checklists from the original occurrence data that were collected and augmented from the online form. Taking the role of the HSC Project Leader, you must now carry out final quality checks prior to publication.

1. Download UC2-IS-3-ForCleaning.xlsx. (156 KB)
2. Evaluate the dataset and identify which types of errors are present.
3. Identify possible ways to correct those issues and perform those corrections for as many of the errors as you can.
4. Use the exercise sheet to provide your answers and submit the spreadsheet.

Exercise 4

Data publishing

The HISC is now ready to publish the checklist data and associated occurrences to GBIF. For this exercise, you will take the role of the Project Leader. Your responsibility: publishing the cleaned checklist data and associated occurrences online through the GBIF network.

1. Download UC2-IS-4-ForPublication.xlsx. (99 KB)
2. Use the previously provided IPT installation to publish the given dataset.
3. Use the exercise sheet to provide your answers and link to the published dataset.

Use Case II - Lepidoptera sightings

Familiarize yourself with the use case scenario.

You have a choice for Use Case II between two scenarios:

- Invasive species checklist
- Lepidoptera sampling

Your choice for Use Case II will be graded.

Scenario

Sampling of Lepidoptera across Countries

Papilio machaon Linnaeus, 1758 *observed in Israel by רניו רמוע (licensed under* CC-BY-NC 4.0*)*

This narrative was developed as a basis for practical exercises in the biodiversity data mobilization course and the exercise concept and content was developed by Alberto González-Talaván, based on previous work by Alberto González-Talaván, Danny Vélez, Larissa Smirnova, Laura Russell, Mélianie Raymond and Nicolas Noé. It is a fictionalized scenario and is meant only for instructional purposes.

Description

The International Butterfly Amateur Network (IBAN) has been providing a framework for national amateur observational groups to capture data about the occurrence of butterflies (Lepidoptera) since 2009. An extensive network of amateur observers use a standard protocol based on Pollard walks to capture this information on paper sheets that they send to their national office. Some of these offices digitize this information into spreadsheets, but others do not have the human resources to do this and they send the paper logs to the IBAN for processing. IBAN produces an annual report based on the sightings provided by these national members, with updated distribution maps and analysis of population trends for some key species.

The IBAN headquarters is mainly staffed with volunteers. With the increasing popularity of citizen science and the general interest in butterflies as a charismatic group of organisms, more and more data are received every year and the paper data sheets quickly pile up undigitized. The IBAN steering committee is trying to identify a more efficient and agile workflow for the creation of digital data because they would like to start publishing these data online regularly. They would also like to start processing digital pictures that their volunteers are already capturing with mobile devices like phones and tablets. Their ultimate objective is to raise the profile of the network and strengthen collaborations with local and regional governments to influence conservation policies for Lepidoptera in the countries involved.

There is currently no formal agreement between IBAN and the amateurs capturing data, to cover the ways in which the data can be used, for example. The steering committee has some concerns that when they start publishing the data online, they will have to formalize this arrangement.

Data collection

The recommended protocol —Pollard walks— is based on transects that range between 300 and 600 m in length, divided into 50 m sections. Each transect should cover a single habitat type.

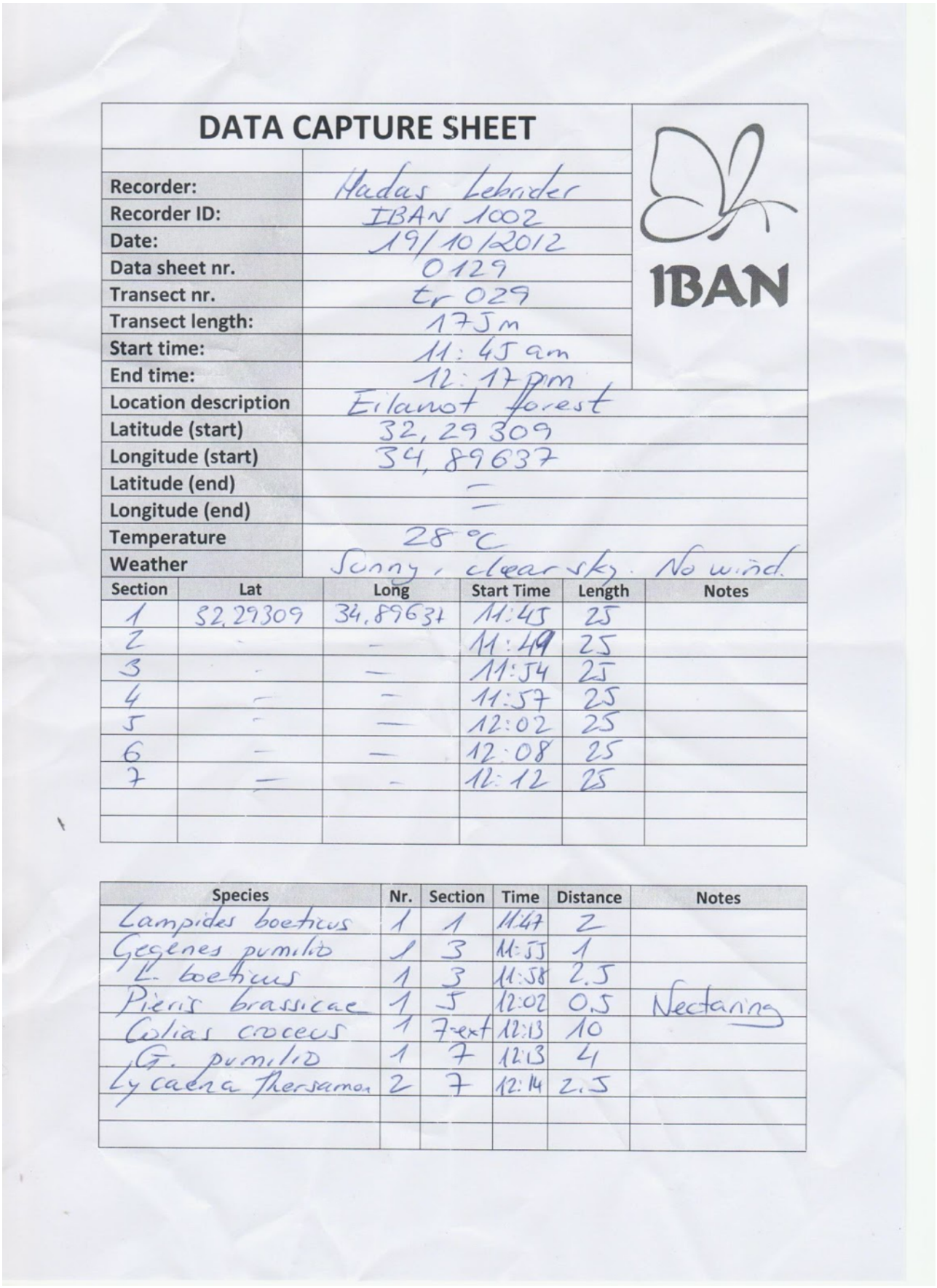
In each visit, transect-walkers have to count all species of Lepidoptera that can be seen within 5 m of the transect line. Special behaviours (egg laying or nectaring), as well as developmental stage (e.g., larvae or eggs), should be recorded as well.

For most countries, these sampling efforts happen once every two weeks from the beginning of October to the end of June.

There are quality control measures in place: every reported record is flagged "Pending approval". Record status is only changed to "Approved" after verification by a designated taxonomic expert. Species spotted out of their regular season or distribution area are flagged for additional verification.

Time of day and weather conditions are recorded at the beginning of the transect. Along the transect, the number of individuals of every species seen is counted. Un-identified species are counted and recorded either by family or as a predefined complex of two or three similar species. Butterflies seen outside the 5 meter range are recorded as “Extra+the number of the nearest section” (e.g. 5-extra). The end time of the transect is also recorded.

Analogue data capture example



DATA CAPTURE SHEET

IBAN

Recorder:	Hadas Lebiush
Recorder ID:	IBAN 1002
Date:	19/10/2012
Data sheet nr.	0129
Transect nr.	tr 029
Transect length:	175 m
Start time:	11:45 am
End time:	12:17 pm
Location description	Eilanot forest
Latitude (start)	32,29309
Longitude (start)	34,89637
Latitude (end)	–
Longitude (end)	–
Temperature	28 °C
Weather	Sunny, clear sky. No wind.

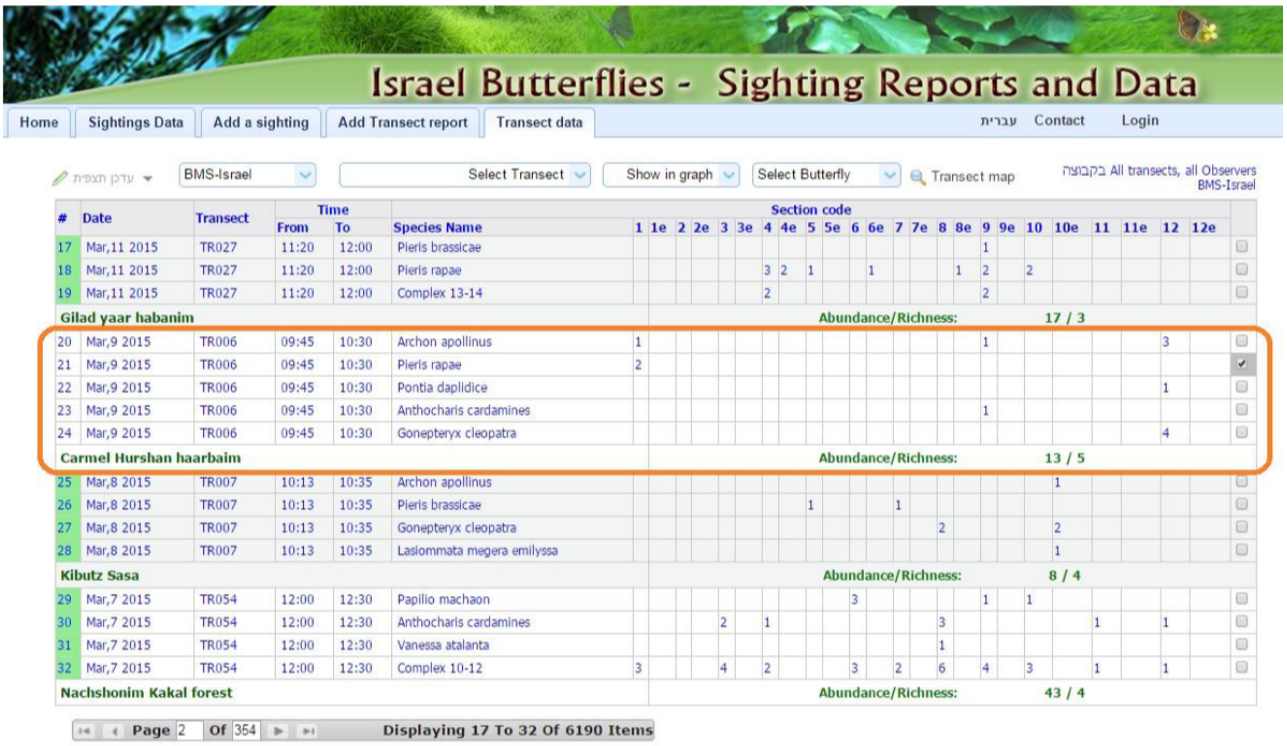
Section	Lat	Long	Start Time	Length	Notes
1	32,29309	34,89637	11:45	25	
2	–	–	11:49	25	
3	–	–	11:54	25	
4	–	–	11:57	25	
5	–	–	12:02	25	
6	–	–	12:08	25	
7	–	–	12:12	25	

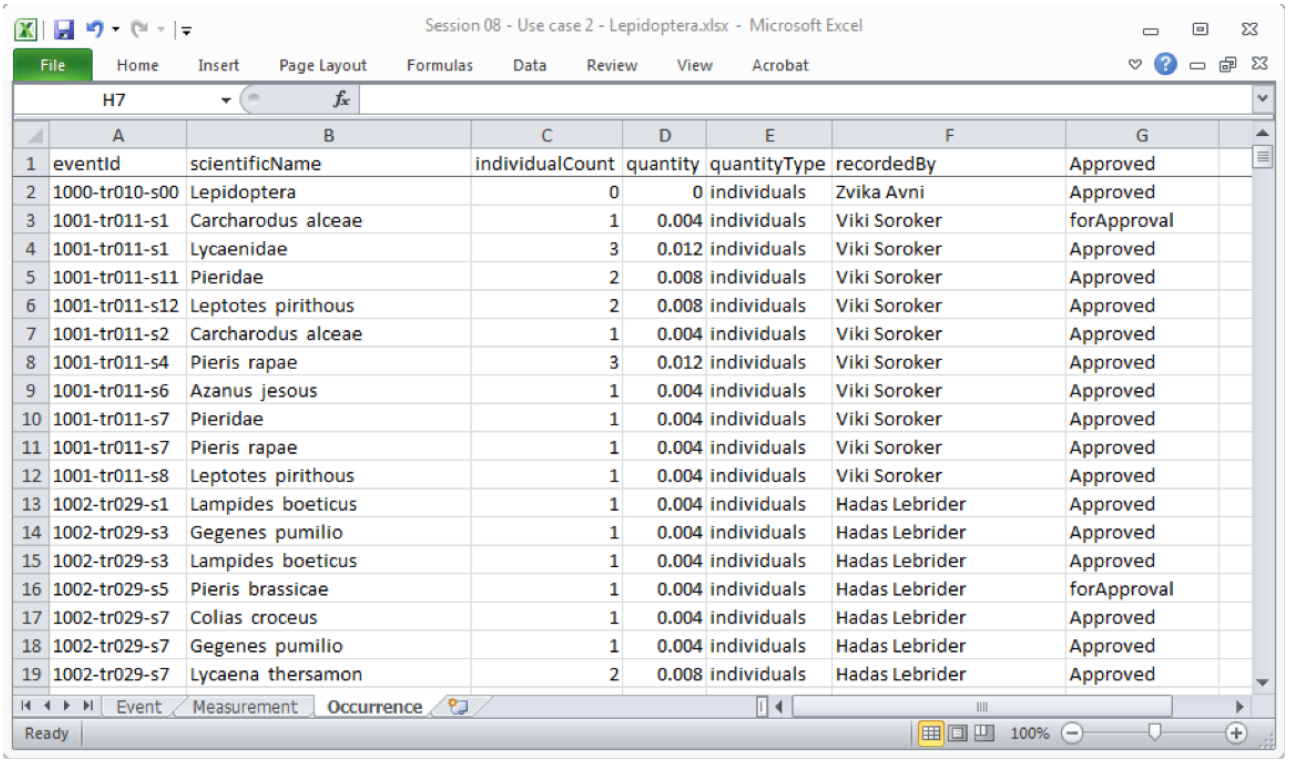
Species	Nr.	Section	Time	Distance	Notes
Lampides boeticus	1	1	11:47	2	
Gegenes pumilio	1	3	11:55	1	
L. boeticus	1	3	11:58	2.5	
Pieris brassicae	1	5	12:02	0.5	Nectaring
Colias croceus	1	7-ext	12:13	10	
G. pumilio	1	7	12:13	4	
Lycaena thersamon	2	7	12:14	2.5	

Digital data description

Some national offices use groups of volunteers to digitize the paper logs and produce digital spreadsheets. The spreadsheets are very simple and include three datasheets. One captures the information linked to the sampling efforts, the second the weather conditions and the third the

species encountered and the number of individuals observed by the amateur.





	A	B	C	D	E	F	G
1	eventId	scientificName	individualCount	quantity	quantityType	recordedBy	Approved
2	1000-tr010-s00	Lepidoptera	0	0	individuals	Zvika Avni	Approved
3	1001-tr011-s1	Carcharodus alceae	1	0.004	individuals	Viki Soroker	forApproval
4	1001-tr011-s1	Lycaenidae	3	0.012	individuals	Viki Soroker	Approved
5	1001-tr011-s11	Pieridae	2	0.008	individuals	Viki Soroker	Approved
6	1001-tr011-s12	Leptotes pirithous	2	0.008	individuals	Viki Soroker	Approved
7	1001-tr011-s2	Carcharodus alceae	1	0.004	individuals	Viki Soroker	Approved
8	1001-tr011-s4	Pieris rapae	3	0.012	individuals	Viki Soroker	Approved
9	1001-tr011-s6	Azanus jesous	1	0.004	individuals	Viki Soroker	Approved
10	1001-tr011-s7	Pieridae	1	0.004	individuals	Viki Soroker	Approved
11	1001-tr011-s7	Pieris rapae	1	0.004	individuals	Viki Soroker	Approved
12	1001-tr011-s8	Leptotes pirithous	1	0.004	individuals	Viki Soroker	Approved
13	1002-tr029-s1	Lampides boeticus	1	0.004	individuals	Hadas Lebrider	Approved
14	1002-tr029-s3	Gegenes pumilio	1	0.004	individuals	Hadas Lebrider	Approved
15	1002-tr029-s3	Lampides boeticus	1	0.004	individuals	Hadas Lebrider	Approved
16	1002-tr029-s5	Pieris brassicae	1	0.004	individuals	Hadas Lebrider	forApproval
17	1002-tr029-s7	Colias croceus	1	0.004	individuals	Hadas Lebrider	Approved
18	1002-tr029-s7	Gegenes pumilio	1	0.004	individuals	Hadas Lebrider	Approved
19	1002-tr029-s7	Lycaena thersamon	2	0.008	individuals	Hadas Lebrider	Approved

Lepidoptera exercise sheet

Download the exercise sheet. (MS Word, 342 KB)

Exercise 1

Planning

The volume of analogue data (paper logs) arriving at the IBAN headquarters will soon exceed their

capacity to digitize, so the steering committee has decided to reconsider the current approach to this area of their work which has grown unmanaged for the last few years. To date, this is how work has been organized:

- The paper logs arrive via post. The secretary opens the packages and collates the logs.
- There are five volunteers with basic computer skills using two shared computers to digitize the paper logs. These volunteers are also citizen scientists themselves, so they are familiar with the taxonomy of the order Lepidoptera, and with the species occurring in the country where the headquarters of IBAN are located.
- The digitizers come and go whenever they have time so they usually check for computer availability via phone. Sometimes there are time clashes and some have to go home as the two computers are busy, and sometimes the two computers are unused.
- When they digitize, they usually pick one paper log at a time from the pile, and digitize it (if they can). Common problems that occur are that:
  - the digitizer does not know the species observed (misspellings occur),
  - the digitizer does not know the area where the sampling has occurred,
  - the digitizer cannot read the handwriting or the language in which some of the comments are written.
- A single taxonomic expert gets all the digitized tables and produces the report and distribution maps based on them. Normally she needs to discard around 15% of the digitized data because of inconsistencies, misspellings or other errors that she does not have the time to check.

Exercise 1a

Analyze the financial component of their new digitization plan

The steering committee is analyzing the following options for their new digitization plan, all of which have financial implications on their already reduced budget. They know they can only implement TWO of these options, so they need to choose wisely. Use the exercise sheet to provide a recommendation on which two options they should select and explain why you chose them.

1. Option 1: Buy three more computers so all digitizers can work simultaneously.
2. Option 2: Offer financial support to the national offices to buy flatbed scanners and send/share the logs electronically instead of by post.
3. Option 3: Offer financial compensation to the digitizers. They will not be able to pay all five of them the equivalent of a regular salary, but could cover the costs of part time positions for three of the volunteers.
4. Option 4: Purchase existing biodiversity digitization software in English, which comes with taxonomic entry check and in-built aids to correct geographical information.
5. Option 5: Contract a software development company to develop customized digitization software. For the same price of the commercial software, the developers will provide a solution in the local language, which will match the original data schema perfectly and will also provide a web data portal to expose the results of the digitization effort.
6. Option 6: Organize a course for the five digitizers to improve their skills in taxonomy, computer use and biodiversity informatics standards.

Exercise 1b

Assign roles

These are the human resources available for this digitization effort. How would you assign roles to maximize the efficiency of the digitization process and produce data of the highest quality possible? Use the exercise sheet to provide your answers.

1. Administrative assistant. No taxonomic knowledge. Basic computer use. Can read 3 languages.
2. Volunteer 1. Basic taxonomic knowledge. Basic computer use.
3. Volunteer 2. Basic taxonomic knowledge. Basic computer use.
4. Volunteer 3. Basic taxonomic knowledge. Basic computer use. Can read 3 languages.
5. Volunteer 4. Basic taxonomic knowledge. Basic computer use. Can read 3 languages.
6. Volunteer 5. Basic taxonomic knowledge. Advanced computer use (including GIS and data analysis tools).
7. Taxonomic expert. Advanced taxonomic knowledge. Advanced computer use (including GIS and data analysis tools).

Exercise 2

Data capture

Imagine you are one of the volunteers digitizing the paper logs received at the IBAN headquarters. You have received two paper logs.

1. Download logs 1 and 2 UC2-LS-2-ForCapture.zip. (943 KB)
2. What data structure would you use to reflect the data in these logs?
3. Create a spreadsheet using this structure and the data from the logs.
4. Use the exercise sheet to provide your answers and submit the spreadsheet created in the previous step.

Exercise 3

Data management

Taking the role of one of the volunteers with advanced computer skills, imagine you have been assigned the responsibility for data quality issues. Your main task is to reduce the amount of data that is currently discarded (around a 15%) before processing due to errors and inconsistencies. You have received a dataset as the raw product of the digitization effort.

1. Download UC2-LS-3-ForCleaning.xlsx. (44 KB)
2. Evaluate the dataset and identify which types of errors are present.
3. Identify possible ways to correct those issues, and perform those corrections for as many of the errors present as you can.
4. Use the exercise sheet to provide your answers and submit the spreadsheet.

Exercise 4

Data publishing

For this exercise, you will take the role of the taxonomic expert collaborating with IBAN at their headquarters. Some of your previous responsibilities (writing the annual report, and producing the

base distribution maps) have been handed over to the volunteers, and you have now been given a new responsibility: publishing the cleaned data online through the GBIF network. The volunteer in charge of data quality has provided a dataset to be published.

1. Download UC2-LS-4-ForPublication.xlsx. (58 KB)
2. Use the previously provided IPT installation to publish the given dataset.
3. Use the exercise sheet to provide your answers and link to the published dataset.

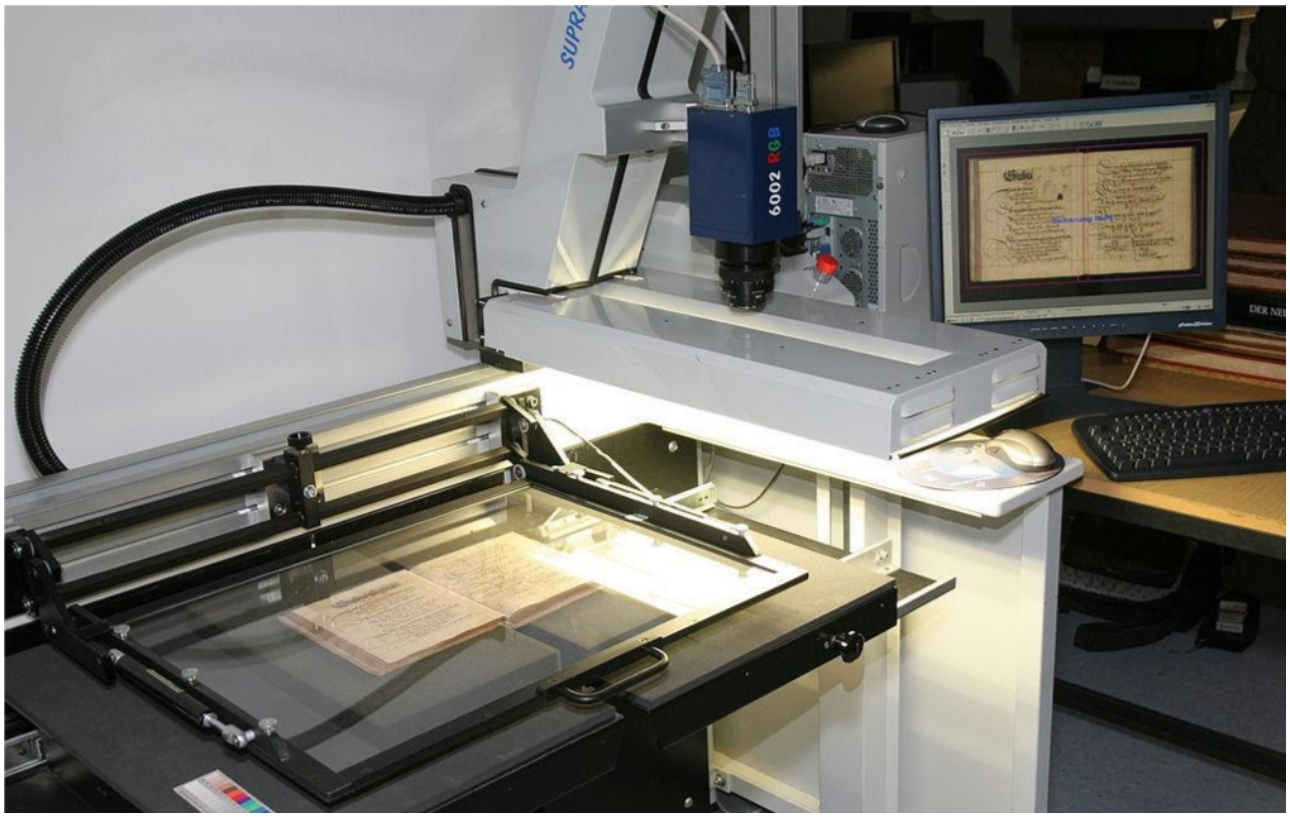
Use Case III - Birds from literature

Familiarize yourself with the use case scenario.

Use Case III will be graded.

Scenario

Data Mobilization Project from Literature "Birds fallen at Danish Lighthouses, 1883–1939"



High resolution scanner *for book digitization project by Heiko Hornig (licensed under* CC BY-SA 2.5*)*

This narrative was developed as a basis for practical exercises in the biodiversity data mobilization course and the exercise concept and content was developed by Alberto González-Talaván, Andrea Hahn, Laura Russell and Sharon Grant. It is based upon a previous adaptation by Alberto González-Talaván, Danny Vélez, Larissa Smirnova, Laura Russell, Mélianie Raymond and Nicolas Noé.

It is a fictionalized scenario based on a real project and dataset and is meant only for instructional purposes. The original project and the original dataset are attributed to the Danish GBIF Node, DanBIF.

Description

The Natural History Museum of Denmark (NHM-DK) is a research centre associated with the University of Copenhagen. Their library is a member of the national library association who recently received state funding to make available online the resources held by its members. The NHM-DK would like to begin digitization of the field notebooks, journal publications and books held in their library, some of which have significant historic value.

After a short consultation with their regular partners, NHM-DK received a suggestion from the Head of the management office of the Nordjylland National Park. They would like the contents of a particular classic literature compilation digitized for a project they are running: 'Birds at the Danish Lighthouses, 1883–1939' (In Danish, 'Fuglene ved de danske Fyr, 1883–1939'). They want to use any occurrence data recorded in those books from two lighthouses (Lodbjerg Fyr and Hanstholm Fyr) for an on-site exhibition project.

The NHM-DK has started discussions with their national GBIF node, DanBIF, about the mobilization of the information contained in these volumes, namely to preserve their contents for the future and provide online access for everyone. With the involvement of DanBIF, there is intent to publish and register the resulting extracted data with GBIF. As GBIF requires a license be applied with all published data, the museum has decided to publish the data with a Creative Commons license allowing use of data with attribution (CC-BY).

The IT services required are provided by the Technology Unit of the University of Copenhagen, as for all museum digital projects.

The NHM-DK deputy director, who is coordinating this piece of work has developed a general outline for the work:

1. The museum will carry out the digitization of the literature using two library staff members trained in the use of the library scanner to scan delicate volumes. They will also extract text from the scans through OCR (Optical Character Recognition) software.
2. Three volunteers from the Copenhagen Ornithological Society (COS) who regularly collaborate with the museum and are familiar with the birds of the region have been enlisted to assist and will complete the transfer of data from the scanned PDFs into spreadsheet format. They will need to go to the museum and use the computers available in the library to gain access to the files stored in the museum intranet (private network).
3. The Ornithology Curator in the NHM-DK Bird Department will lead the team responsible for taxonomic checking, data curation, cleaning, format and transformation, and will oversee the entry of metadata for the published dataset. The team includes a collaborating researcher from Sweden, and two postdoctoral students. They have been selected for this task because they are used to working with digital biodiversity data. They will all use their own work computers.
4. The DanBIF Node Manager will ensure that the institution is adequately registered in GBIF as a data publisher and that the deputy director and the ornithology curator have the proper credentials and access to DanBIF's IPT instance to upload and publish the data.

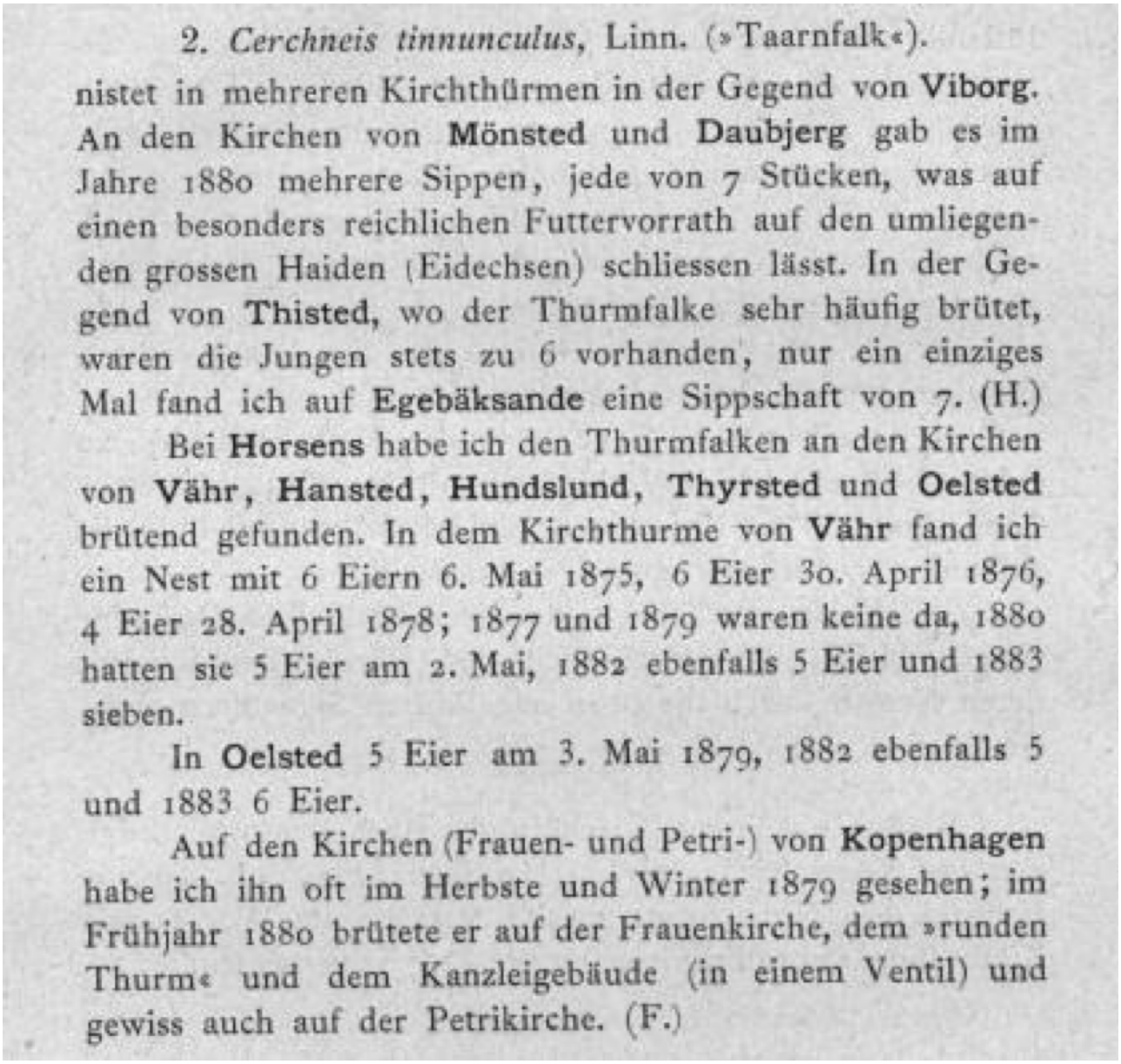
Original data collection

In the period 1883-1939, there were 45 lighthouses and lightships functioning in Denmark. These lighthouses were used by several species of birds during the nights of the bird migration period from the years 1886 through 1939. The presence and activities of these birds were recorded, especially by the keepers of these lighthouses who also collected specimens that were sent to the museum in Copenhagen. These birds were carefully preserved and catalogued by collection managers at the museum and the specimens can still be found there today. Observations of weather conditions during

the nights when birds were observed by the keepers were also documented.

Analogue data description

This is an example of the description of a series of species observations from one of the books (in German, except the common name for the species which is provided in Danish).



2. *Cerchneis tinnunculus*, Linn. (»Taarnfalk«).

nistet in mehreren Kirchthürmen in der Gegend von **Viborg**. An den Kirchen von **Mönsted** und **Daubjerg** gab es im Jahre 1880 mehrere Sippen, jede von 7 Stücken, was auf einen besonders reichlichen Futtervorrath auf den umliegenden grossen Haiden (Eidechsen) schliessen lässt. In der Gegend von **Thisted**, wo der Thurmfalke sehr häufig brütet, waren die Jungen stets zu 6 vorhanden, nur ein einziges Mal fand ich auf **Egebäksande** eine Sippschaft von 7. (H.)

Bei **Horsens** habe ich den Thurmfalken an den Kirchen von **Vähr**, **Hansted**, **Hundslund**, **Thyrsted** und **Oelsted** brütend gefunden. In dem Kirchthurme von **Vähr** fand ich ein Nest mit 6 Eiern 6. Mai 1875, 6 Eier 30. April 1876, 4 Eier 28. April 1878; 1877 und 1879 waren keine da, 1880 hatten sie 5 Eier am 2. Mai, 1882 ebenfalls 5 Eier und 1883 sieben.

In **Oelsted** 5 Eier am 3. Mai 1879, 1882 ebenfalls 5 und 1883 6 Eier.

Auf den Kirchen (Frauen- und Petri-) von **Kopenhagen** habe ich ihn oft im Herbste und Winter 1879 gesehen; im Frühjahr 1880 brütete er auf der Frauenkirche, dem »runden Thurm« und dem Kanzleigebäude (in einem Ventil) und gewiss auch auf der Petrikirche. (F.)

Scanned and translated data description

This is an example of the scanned and translated output from the analogue example above.

Output of the OCR software	Translation into English
2. *Cerchneis tinnunculus*, Linn. (»Taarnfalk«) nistet in mehreren Kirchthürmen in der Gegend von **Viborg**.	2. *Cerchneis tinnunculus*, Linn. ("Taarnfalk") nests in several steeples around **Viborg**.
An den Kirchen von **Mönsted** und **Daubjerg** gab es im Jahre 1880 mehrere Sippen, jede von 7 Stücken, was auf einen besonders reichlichen Futtervorrath auf den umliegenden grossen Haiden (Eidechsen) schliessen lässt. In der Gegend von **Thisted**, wo der Thurmfalke sehr häufig brütet, waren die Jungen stets zu 6 vorhanden, nur ein einziges Mal fand ich auf **Egebäksande** eine Sippschaft von 7. (H.)	At the churches of **Mönsted** and **Daubjerg** there were several family groups in 1880, each of 7 individuals, suggesting a particularly abundant source of food on the surrounding heather (lizards). In the area of **Thisted**, where the tower falcon broods very often, young were always present in broods of 6, only once did I find a group of 7 on **Egebäksande** (H.)
Bei **Horsens** habe ich den Thurmfalken an den Kirchen von **Vähr**, **Hansted**, **Hundslund**, **Thyrsted** und **Oelsted** brütend gefunden. In dem Kirchthurme von **Vähr** fand ich ein Nest mit 6 Eiern 6. Mai 1875, 6 Eier 30. April 1876, 4 Eier 28. April 1878; 1877 und 1879 waren keine da, 1880 hatten sie 5 Eier am 2. Mai, 1882 ebenfalls 5 Eier und 1883 sieben.	In **Horsens** I found kestrels brooding on the churches of **Vähr**, **Hansted**, **Hundslund**, **Thyrsted** and **Oelsted**. In the steeple of **Vähr** I found a nest with 6 eggs on 6 May 1875, 6 eggs on 30 April 1876, 4 eggs on 28 April 1878; in 1877 and 1879 there were none, on 2 May 1880 they had 5 eggs, in 1882 also 5 eggs, and in 1883 seven.
In **Oelsted** 5 Eier am 3. Mai 1879, 1882 ebenfalls 5 und 1883 6 Eier.	In **Oelsted** 5 eggs on 3 May 1879, in 1882 also 5, and in 1883 6 eggs.
Auf den Kirchen (Frauen- und Petri-) von **Kopenhagen** habe ich ihn oft im Herbste und Winter 1879 gesehen; im Frühjahr 1880 brütete er auf der Frauenkirche, dem »runden Thurm« und dem Kanzleigebäude (in einem Ventil) und gewiss auch auf der Petrikirche. (F.)	On the churches (Our Lady's and St. Peter's) of **Copenhagen**, I have often seen it in the autumn and winter of 1879; in spring 1880, it brooded on Our Lady's church, the "round tower" and the law firm building (in a valve) and certainly also on St. Peter's Church. (F.)

Digital data description

Studying the extract from the book, the volunteers from the Copenhagen Ornithology Society suggest extracting the following data from the scanned and translated text:

- Scientific name as appearing in the book
- Common name(s) in Danish as appearing in the book
- Locality
- Year/month/day
- Observed number of individuals
- Sex
- Lifestage
- Remarks
- URL of the digitized book page in which the occurrence is provided

Birds from literature exercise sheet

Download the exercise sheet. (MS Word, 342 KB)

Exercise 1

Planning

The team needs to develop a sustainable workflow to digitize literature resources, extract any valuable biodiversity information on them and publish it online via GBIF. They need to develop a plan

that can be sustained in the future once the funding from the national library association is over.

The scenario section of this use case includes a brief description of the workflow conceived by the deputy director. Based on the workflow and the accompanying text complete the following:

1. Identify the different stakeholders participating in this project
2. Identify their affiliation and assign each of them to a stakeholder group
3. Identify the roles associated to them and assign the tasks for which they are currently responsible
4. Perform a critical analysis of the workflow, identify potential risks and gaps, and suggest ways to improve the workflow, maximize the efficiency of the digitization project and produce data of the highest quality possible.
5. Use the exercise sheet to provide your answers.

Exercise 2

Data capture

The scans and character recognition (OCR) of the books have been completed. Occurrence data must now be extracted from those sources and compiled in a spreadsheet format.

The original data was in German and, to make it more widely usable when published online, the project manager would like to make it available in English.

> 2. Cerchneis tinnunculus, Linn. ("Taarnfalk") nests in several steeples around Viborg.
>
> At the churches of Mönsted and Daubjerg there were several family groups in 1880, each of 7 individuals, suggesting a particularly abundant source of food on the surrounding heather (lizards). In the area of Thisted, where the tower falcon broods very often, young were always present in broods of 6, only once did I find a group of 7 on Egebäksande (H.)
>
> In Horsens I found kestrels brooding on the churches of Vähr, Hansted, Hundslund, Thyrsted and Oelsted. In the steeple of Vähr I found a nest with 6 eggs on 6 May 1875, 6 eggs on 30 April 1876, 4 eggs on 28 April 1878; in 1877 and 1879 there were none, on 2 May 1880 they had 5 eggs, in 1882 also 5 eggs, and in 1883 seven.
>
> In Oelsted 5 eggs on 3 May 1879, in 1882 also 5, and in 1883 6 eggs.
>
> On the churches (Our Lady's and St. Peter's) of Copenhagen, I have often seen it in the autumn and winter of 1879; in spring 1880, it brooded on Our Lady's church, the "round tower" and the law firm building (in a valve) and certainly also on St. Peter's Church. (F.)

1. Take the role of a volunteer charged with transforming the translated text into a spreadsheet as individual occurrences. The occurrences will need unique numbers assigned to them.
2. Create a spreadsheet using the data fields listed in the Digital data description using data found in the example above recorded by: Chr. Fr. Lütken.
3. Use the exercise sheet to provide your answers and submit the spreadsheet created in the previous step.

In the examples used, the individual occurrences do not always contain data to complete all of the columns in the spreadsheet.

Exercise 3

Data management

Data has now been compiled into a spreadsheet format by the volunteers from the Copenhagen Ornithological Society. Taking the role of the Ornithology Curator in the Bird Department, you have been assigned the responsibility for data quality issues on the dataset.

Through retrospective georeferencing, coordinates have been added to the dataset along with the locality, but no other higher geography. Since all the observations were made in Denmark, continent and country can easily be added. Additionally, only the scientific name was provided. Higher taxonomy can be derived utilizing software tools such as OpenRefine. You are also aware that there are typographic errors that were made by the digitizers.

1. Download UC3-DL-3-ForCleaning.zip. (45 KB)
2. Identify and correct any invalid years.
3. Verify and correct taxonomy.
4. Verify coordinates are correct for the two given localities. Correct any that are not. Coordinates should be in decimal format.
5. Add any data for missing elements that can be derived using the available data
6. Remember to keep the original information provided and document your changes and assumptions as part of the individual records and the metadata.
7. Use the exercise sheet to provide your answers and submit the cleaned text file extracted from the step 1.

dataset should contain only years 1883-1939

Exercise 4

Data publishing

For this exercise, you will take the role of the person responsible for publishing the cleaned data online via the GBIF network. You have been supplied with a multimedia file and an identification history file that should be published along with the observations. The staff member in charge of data quality has provided cleaned datasets for you to publish.

1. Download UC3-DL-4-ForPublication.zip. (65 KB)
2. Use the previously provided IPT installation to publish the given dataset.
3. Use the exercise sheet to provide your answers and link to the published dataset.

Final assignments

For your final activities, you will complete and submit Use Case II and Use Case III for evaluation.

USE CASE II

There are two options for USE CASE II (Invasive species OR Lepidoptera sightings). You only need to

select one of them for your assignment.

Required files for submission:

- completed exercise sheet (MS Word Doc or similar is acceptable)
- data capture spreadsheet (MS Excel, csv, txt or similar is acceptable)
- cleaned/standardized dataset (MS Excel, csv, txt or similar is acceptable)

USE CASE III

There is only one option for USE CASE III (Birds from literature).

Required files for submission:

- completed exercise sheet (MS Word Doc or similar is acceptable)
- data capture spreadsheet (MS Excel, csv, txt or similar is acceptable)
- cleaned/standardized dataset (MS Excel, csv, txt or similar is acceptable)

Include the use case and exercise number along with your name on all files submitted. For example, Russell-UC2-IS-exercise-sheet.docx, Russell-UC2-IS-2.xlsx, Russell-UC2-IS-3.xlsx. **All files must be submitted in English**. Contact training@gbif.org if you have any questions.

Assignment submission

Assignments can be submitted from the online (HTML) version of the course.

Course evaluation

Complete the course evaluation

Key documentation

The following references provide further detail on the topics covered in this course. All links open in a new window/tab.

Darwin Core

- Darwin Core Terms: A quick reference guide
- Simple DarwinCore
- Darwin Core Questions & Answers
- Darwin Core extensions registered with GBIF

Data publishing

- Quick guide to publishing data through GBIF

- How to publish biodiversity data through GBIF.org
- Become a data publisher with GBIF
- Best Practices for Publishing Biodiversity Data from Environmental Impact Assessments
  GBIF Secretariat & IAIA: International Association for Impact Assessment (2020).
- Guidance for private companies to become data publishers through GBIF: Template document to support the internal authorization process to become a GBIF publisher
  Rui Figueira, Pedro Beja, Cristina Villaverde, Miguel Vega, Katia Cezón, Tainan Messina, Anne-Sophie Archambeau, Rukaya Johaadien, Dag Endresen & Dairo Escobar (2020).
- Publishing DNA-derived data through biodiversity data platforms
  Anders F. Andersson, Andrew Bissett, Anders G. Finstad, Frode Fossøy, Marie Grosjean, Michael Hope, Thomas S. Jeppesen, Urmas Kõljalg, Daniel Lundin, R. Henrik Nilsson, Maria Prager, Cecilie Svenningsen & Dmitry Schigel (2020).
- Classes of datasets supported by GBIF
- GBIF data quality requirements for publishing
- GBIF data licenses
- Checklist core templates
- Occurrence core templates
- Sampling event core templates
- Sampling event data best practices
- Sharing images, sounds and videos on GBIF
- Data papers
- Published data papers

Data publishing: IPT

- The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet
  Robertson et al. (2014)
- To install IPT or not to install IPT
- IPT data hosting centres
- IPT Install / Set up webinar
- Installing the IPT video
- IPT in Practice demonstration video

Digitization

- iDigBio Digitization Resources
- iDigBio Collections Digitization Workflows
- iDigBio Digitization Workflows and Protocols
- iDigBio specimen image capture guide
- Canadensys 10-step guide to managing images with your biodiversity data

GBIF

- What is GBIF
- Strategic Plan
- Become a member
- Science Review
- Establishing an Effective GBIF Participant Node: Concepts and general considerations
  GBIF Secretariat (2019).

Georeferencing

- Georeferencing Best Practices
  Arthur D. Chapman & John R. Wieczorek (2020).
- Georeferencing Quick Reference Guide
  Paula F. Zermoglio, Arthur D. Chapman, John R. Wieczorek, Maria Celeste Luna & David A. Bloom (2020).
- Georeferencing Calculator Manual
  David A. Bloom, John R. Wieczorek & Paula F. Zermoglio (2020).
- Georeferencing resources

Invasive Species

- GRISS - Global Register of Introduced and Invasive Species
- TrIAS - Tracking Invasive Alien Species

Living Atlases

- Living Atlases
- ALA key technical documentation

Miscellaneous

- VertNet Guide to opening text files in Excel
- VertNet data licensing guide

OpenRefine

- OpenRefine documentation
- OpenRefine regular expressions
- Guía para la limpieza de datos sobre biodiversidad con OpenRefine
  Paula F. Zermoglio, Camila A. Plata Corredor, John R. Wieczorek, Ricardo Ortiz Gallego & Leonardo Buitrago (2021).
- Using Google Refine and taxonomic databases (EOL, NCBI, uBio, WORMS) to clean messy data
  iPhylo blog post. Rod Page 2012.
- Reconciling author names using Open Refine and VIAF

iPhylo blog post. Rod Page 2013.

- Validating scientific names with the GBIF Portal web service API
  Guest post was written by Gaurav Vaidya, Victoria Tersigni and Robert Guralnick 2013.
- iDigBio Cleaning data with OpenRefine
  iDigBio 2013.
- Have We Got the Names “Right”?
  Canadensys 2014.
- Cleaning data with OpenRefine
  Desmet and Brosens 2016 TDWG.
- EasyOpen Redlist
  Querying the IUCN Red List, using a species list, OpenRefine, and some pre-written code. Olly Griffin July 2019.

Planning/Collaboration

- Agile methodology
- What is SCRUM
- SCRUM Framework
- Kanban methodology
- Scrum Guide
- GitHub

Quality

- Principles of Data Quality
  Arthur Chapman 2005.
- Principles and Methods of Data Cleaning: Primary Species and Species-Occurrence Data
  Arthur Chapman 2005.
- Be careful with dates in Excel
  DataOne 2014.
- Character encoding for beginners
- MVZ Guide for Recording Localities in Field Notes

Sensitive species

- Current Best Practices for Generalizing Sensitive Species Occurrence Data
  Arthur D. Chapman 2020.

Taxonomy

- GBIF checklist datasets and data gaps
- GBIF Labs - Names Parser
- GBIF Labs - Species Matching
- Global Names Resolver

- Marine Name Matching Strategy for taxonomic quality control
- Nomenmatch

Glossary

ALA

Atlas of Living Australia. The Australian node of GBIF, who developed an open source data portal now widely used within the GBIF community & partners for biodiversity national portals.

API

Application Programming Interface. A set of clearly defined methods of communication between various software components.

BID

Biodiversity Information for Development. An EU funded project co-ordinated by GBIF whose aim is to increase data mobilization capacity in the Africa, Caribbean and Pacific regions.

BIFA

Biodiversity Fund for Asia.

CC Licences

Creative Commons. These are a series of licenses set up by the Creative Commons organization that enable sharing and reuse of creativity and knowledge through the provision of free legal tools. Three of them can be assigned to GBIF-shared datasets: CC0, CC BY and CC BY-NC.

Controlled Vocabulary

This is a restricted set of terms that are used as possible values for a given field. One can think of it as a lookup list or dropdown for a particular field. For example the DwC field basisOfRecord should only contain one of these values: "PreservedSpecimen", "FossilSpecimen", "LivingSpecimen", "HumanObservation", "MachineObservation". We would say that list of values is a controlled vocabulary for that field.

DwC

Darwin Core is a biodiversity data standard, maintained by TDWG & widely used within the GBIF community and partners. It is a set of standardized terms (or field names) and their definitions, which are used to share biodiversity information.

DOI

Digital Object Identifier. A persistent identifier or handle used to uniquely identify objects. DOIs are in wide use mainly to identify academic, professional, and government information, such as journal articles, research reports and data sets, and official publications.

DwC-A

Darwin Core Archive. A compressed (zipped) file containing all the information needed to share with GBIF, for a particular resource. Each zip contains three types of files:

1. the actual data, in one or more text files: occurrence.txt/event.txt/measurmentoffact.txt etc
2. a mapping file: rtf.xml
3. a metadata (EML) file: eml.xml When you publish using the IPT, it creates a Darwin Core Archive, which is shared with GBIF. Also, when you download data from the GBIF website you can choose a DwC-A format as well.

GUID

Globally Unique IDentifier

IPT

Integrated Publishing Toolkit. It is a free and open source web application (software) for publishing biodiversity data. The software itself lives on a server (either at your institution or elsewhere) that must have access to the internet 24/7. It is used to create and handle Darwin Core Archive files that can be shared and used by anyone including GBIF.

Loan

In the context of natural history collections, this is the procedure of lending specimens between institutions.

LSID

Life Sciences Identifier. They are persistent, globally unique identifiers for biological objects.

Data Publishing

With regards to GBIF we have a very specific definition of data publishing. It refers to making biodiversity datasets publicly accessible and discoverable, in a standardized form, via an access point, typically a web address (a URL).

Resource

A Resource is the collective term used to refer to a particular dataset and its metadata once it has been uploaded to an IPT instance.

TDWG

Taxonomic Databases Working Group, now renamed Biodiversity Information Standards.

URN

Uniform Resource Number

UUID

Universally Unique Identifier

Appendix: Data papers

A data paper is a peer reviewed document describing a dataset, published in a peer reviewed journal. It takes effort to prepare, curate and describe data. Data papers provide recognition for this effort by means of a scholarly article. We do not cover how to create data papers in this course, however, as an optional activity, you can watch this video (51:51) presented by Lizanne Roxburgh. In this video, you will learn more about publishing data papers. If you are unable to watch the embedded video, you can download it locally. (MP4 - 99.2 MB)

▶ https://vimeo.com/265350948 *(Vimeo video)*

You can read more about data papers on GBIF.org.

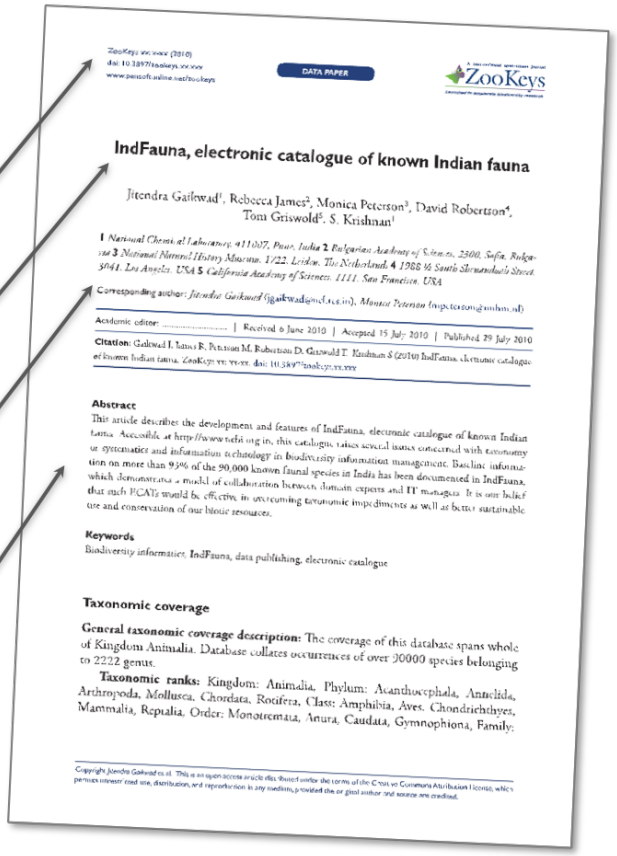
What it is: A scholarly publication of searchable metadata – a document describing a dataset, or a group of datasets

DOI: indexation and citation

Promote and publicize existence of data

Provide scholarly credit to data publishers through citable journal publications

Describe the data in a structured human-readable form



View the links below to see a data paper as it appears on the IPT, on GBIF.org and in the Journal. Each are cross-linked.

- Journal: https://doi.org/10.3897/BDJ.2.e4244
- GBIF: https://www.gbif.org/dataset/3c9e2297-bf20-4827-928e-7c7eefd9432c
- IPT: http://130.120.204.55:8080/ipt/resource.do?r=harvestmen_of_french_guiana

Appendix: Solutions

This appendix contains the answers and additional information to all of the review quizzes. Additionally, this section contains a suggested solution to USE CASE I.

Foundations review solutions

For the given statement, input the correct term (database, database language, database program)

- combines and presents functions and features for manipulating data, together in a unified interface
  database program
- structured and organized collection of data and/or information held on a computer
  database

- the way by which a human communicates with a computer
  database language

If you open a data file and see the following, what would you suspect is the issue?

```
tre, ou ne pas tre, c est l la question.
```

- The wrong encoding was used to open the file

For the given software, input the type of software (data capture, data management, data cleaning, data publishing).

- Integrated Publishing Toolkit (IPT)
  data publishing
- Specify
  data capture AND **data management**
- iNaturalist
  data capture
- OpenRefine
  data cleaning

For the given example, input the correct data type (binary, boolean, float, integer, long integer, text, unstructured text)

- 1236975
  long integer
- 01101111
  binary
- We walked 5 miles down the road west from the post office in the center of town. We then went 2 miles north on a dirt path to the river. Then we continued west along the river for another 5 miles.
  unstructured text
- 1024
  integer
- 29.0
  float
- Yes/No
  boolean
- 6 rabbits were observed
  text

Which of these terms describes a "field/column name"?

- Assigned
- Identifying
- Unique

Which of these terms describes a "field label"?

- Descriptive
- Readable

- User-interface

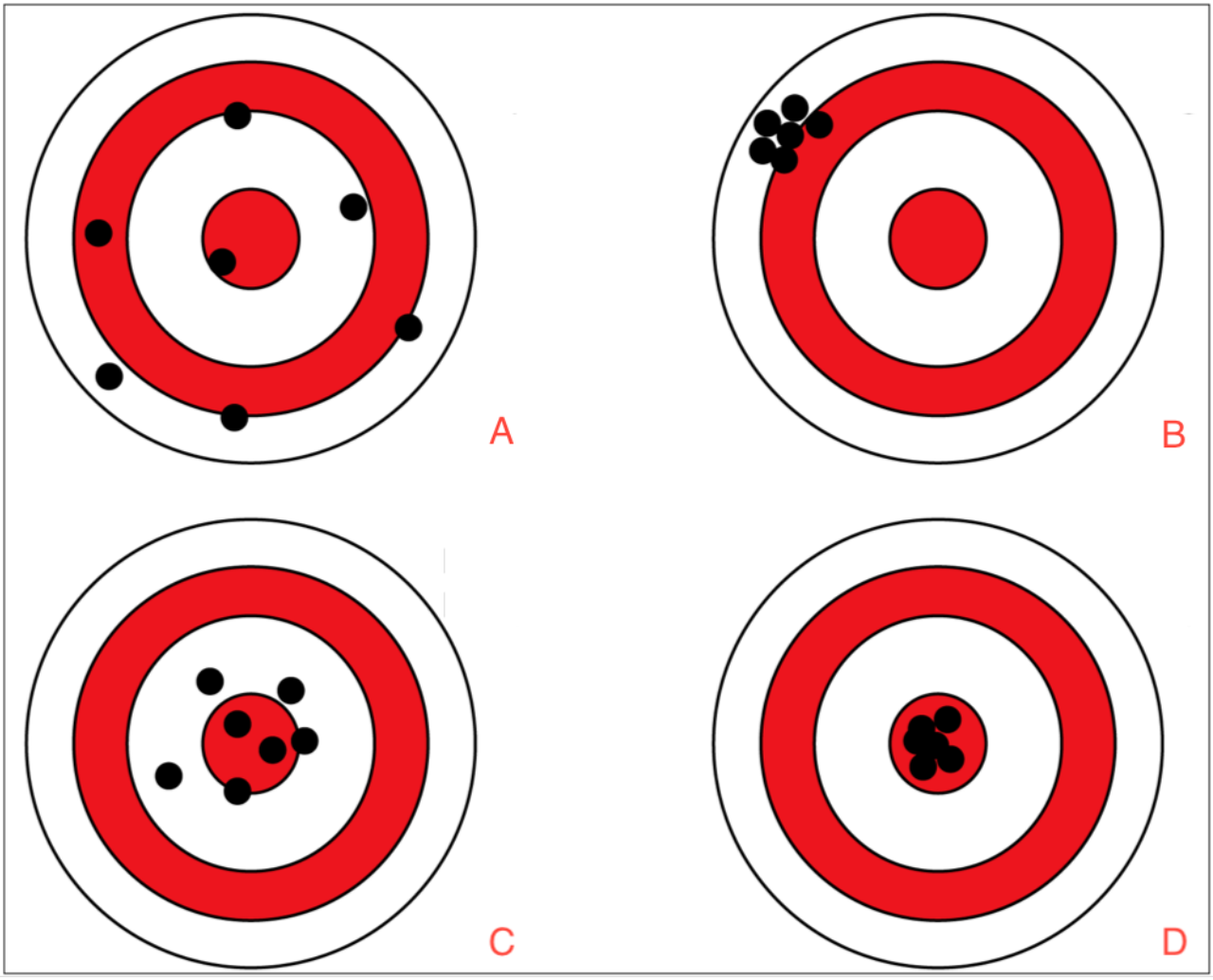
For each statement, input the correct structure (row, column, table)

- All data refers to a SINGLE concept.
  table
- An attribute has the SAME field/data type for every record.
  column
- Attributes of a record ALWAYS stay together.
  row

Who determines the fitness for use of your data?

- The users of the data for research or education

For the given statements, input the matching image. (A, B, C, D)



- High accuracy, low precision
  C
- Low accuracy, high precision
  B
- High accuracy, high precision
  D
- Low accuracy, low precision
  A

Identify the data relationships where Dataset B needs to be merged into Dataset A (0:1, 1:0, 1:1, 1:∞, ∞:1, ∞:∞). Not all the relationships used.

- Collector field exists in both dataset A and B
  1:1
- Country field only exists in dataset B
  0:1
- Name field exists in dataset A, but dataset B contains First Name and Last Name fields
  1:∞
- ID field exists in both dataset A and B
  1:1
- Elevation exists in dataset A, but not in dataset B
  1:0
- Date exists in dataset A, but Day, Month, and Year are separate fields in dataset B
  1:∞

Metadata is important because (select the TRUE statements)

- it allows users to determine if a dataset is fit for their use.
- it allows you to know under which legal terms the reuse of data is permitted.

Planning review solutions

What is the order of the five PMBoK Process Groupings?

- Initiating, Planning, Executing, Monitoring and Controlling, Closing

Learn More: https://quizlet.com/306742513/1-introduction-pmbok-guide-6th-edition-flash-cards/

What are the types of deliverable?

- Stated - YES
- Implied - YES
- Estimated - NO
- Direct - YES
- Indirect - YES
- Guesses - NO

What is a bottleneck?

- a blockage that delays development or progress - YES a space where something or someone is missing - NO, THIS IS A GAP
- a problem, or situation that prevents somebody from doing something, or that makes something impossible. NO, THIS IS A BARRIER

Which are examples of mobilization tasks?

- Affiliation - NO, This is a Resource Type
- Publishing - YES

- Imaging - YES
- Georeferencing - YES
- Increased Public Awareness - NO, This is an implied goal.

Data capture review solutions

What dataset type(s) would you choose for an ichthyology collection?

- occurrence
  Most of the time, specimens from collection databases are shared as occurrence data. Each occurrence (specimen or group of specimens) has its own unique identifier (sometimes derived from its catalogue number in the source collection) and the Darwin Core fields used to share them within GBIF describe each specimen: scientific name, the date it was collected on the field, who collected and/or identified it, where, etc. Each collection can have more than one specimen from a same species, as long as each specimen is identified by a unique ID.
- checklist
  It is also possible to create and share a taxonomical checklist derived from a collection database; in this case, it is recommended to share the checklist as a taxonomical dataset, with the occurrence (specimen) list associated with it by using the Occurrence core as an extension to the Taxon Core on the GBIF IPT.

What dataset type(s) would you choose for a list of invasive species?

- occurrence
  Some data publishers will share occurrence datasets coming from studies or programs tracking specimens from some specific invasive species; when the data focuses on individuals instead of the invasive species, in general, they can be shared as occurrence data.
- checklist
  Invasive species can be tracked and monitored at different scales (regional, national, thematic…); as this type of dataset focuses more on the species and their distribution across a given geographical scope, they are mainly shared as taxonomical datasets within GBIF (see GRIIS search results).

What dataset type(s) would you choose for the flora and fauna of an environmental impact study?

- occurrence
  Data are recorded by naturalists on the field and can be shared as simple occurrence datasets.
- sampling event
  They can also be shared as event datasets if standardized protocols (such as vegetation plots, transects, traps…) are used to collect the data.

What dataset type(s) would you choose for bird tracking data?

- occurrence
  These data are shared as occurrence datasets: ideally, each bird is identified with its organismID, and each occurrence (GPS ping) has its own occurrenceID, which is useful to track the different GPS locations of the same bird over the scope of the tracking programme or project. (See example)

What dataset type(s) would you choose for insect trap data?

- occurrence
  Although such data can be shared as simple occurrence datasets, it is best if they're shared as event datasets, where the location, identifier and contents of each trap can be better detailed.
- sampling event
  Insect traps (as well as other traps such as pitfall traps, malaise traps…) are typically used in monitoring programmes to check the presence (or absence) of some species and/or assess their specific abundance. Using the "eventID" field to identify each trap allows the users to get all of the specimens collected within each trap. The same logic applies to other field protocols such as transects, plots, remote cameras, etc.: by using the Event Core instead of the Occurrence core, you'll be able to share much more information about the context of the data collection, and allow users to better understand (and even replicate) your work.

What dataset type(s) would you choose for national park management data?

- occurrence
  record individuals of species
- checklist
  It is important to know how many species are present in the park/reserve perimeter and their conservation status.
- sampling event
  check and track the populations

What dataset type(s) would you choose for a citizen science bioblitz?

- occurrence
  Bioblitz datasets are mainly shared as occurrence datasets.
- sampling event
  Depending on the citizen science programme, specific sampling protocols might be used by the volunteers, in which case, the data can be shared as an event dataset.

What dataset type(s) would you choose for a regional species list?

- checklist
  Geographical or thematic species lists are often used to share information about the species present in a given area; most of the time, these lists also mention the distribution of each species as well as their conservation status in this area. Regional species lists can give a useful insight into a region's biodiversity and habitats, and need to be shared as taxonomical datasets, with or without associated occurrences.

Data management review solutions

Why is it best to clean your data?

- to make them as fit for use as possible
- to achieve your data quality goals

You should always aim to manage and publish data with the highest possible quality. This will improve your day-to-day work (it is easier to work with organized and clean data), as well as the work of potential re-users of your data, who need to understand them and trust their source before using them.

How should you organize your data cleaning workflow?

- ask your colleagues for expertise
- work at an institutional level to harmonize data quality workflows

Nobody is expected to know everything about biodiversity data; you should seek help and advice from your colleagues or other knowledgeable people, and ensure that you're applying the good practices recommended by your institution as you clean your data.

Which is best:

- prevent errors from occurring
- correct errors as soon as you find them in your database or spreadsheet

The best way to avoid spreading errors in your data is to prevent them from occurring at the start of the data collecting/recording process.

Of course, mistakes are unavoidable so you should also clean them as soon as you find them, and document the cleaning process.

If you don't have the time or resources to properly clean your data, it is best to wait before you can do so instead of publishing erroneous data that might confuse people.

Whose responsibility is data quality?

- Everyone involved in the management of data

Every person involved in your data management workflow is at least partly responsible for their quality, from the field technicians to the database manager(s).

People who might later use your data can inform you of any remaining error in your data, and should use them responsibly for their own research, but the initial data quality is not their responsibility.

GBIF can perform automatic checks on your data (e.g. detection of missing values, geographic outliers, unknown scientific names) but should not be held responsible for errors that occurred earlier in the data management process.

Which tools can be used to clean your data?

- Excel & other spreadsheets management tools
- OpenRefine
- Your database software
- Online tools such as Scientific Names Resolver or Google Maps

All kinds of tools can be used to clean your data, but you should identify which ones will answer your needs in terms of taxonomic resolving, georeferencing, deleting duplicates, and so on. You can find helpful tools listed in the data management section.

Data publishing review solutions

What does data publishing mean in the context of GBIF?

- Making your biodiversity dataset(s) publicly accessible and discoverable in a standardized format

Data publishing within GBIF means making your biodiversity dataset(s) publicly accessible in a

standardized format (most of the time, Darwin Core), so that it can be discovered and reused by other people.

What is an IPT?

- a tool that helps you publish your data to GBIF
- a tool that helps you produce a Data paper

The IPT (Integrated Publishing Toolkit) is a Java-coded software that allows you to upload and publish data to GBIF. It is not to be used as a data management or data cleaning tool.

The IPT can also help you with the process of writing and submitting a data paper, thanks to the EML file it generates automatically when you fill in the metadata for your data resource.

Which Creative Commons licences and waivers are recommended by GBIF for data publication?

- CC0, CC-BY and CC-BY-NC

The Creative Commons licences and waivers recommended to publish your dataset(s) to GBIF are CC0, CC-BY and CC-BY-NC. They are widely recognized licenses and/or waivers that align with international open-data requirements for data sharing and re-use.

Please note that you should only choose CC0 or CC-BY waiver/license for your BID-related dataset(s).

What are the three Cores from which you can choose for an IPT resource?

- Occurrence Core, Taxon Core, Event Core

You can choose one of the three following Cores for each of your IPT resources: Occurrence, Taxon or Event Core.

The Darwin Core standard also allows you to link extensions to your chosen Core, such as SimpleMultimedia or MeasurementOrFact.

The metadata are filled in a separate section of the IPT and are shared using the EML standard, not the Darwin Core (which is used for data only).

How many Extensions files can a dataset have?

- as many as needed

Once you have chosen a Core for your IPT resource, you can add Darwin core extensions to it. You can add only one or several extensions, depending on the type of Core you chose, and which extensions are compatible with it.

Extensions are not mandatory (you can publish a dataset without any extension) but can be useful if you want to share additional information that you could not map with your chosen Core.

Use Case I suggested solution

suggested solution (PDF 144 KB)

Acknowledgements

Course design and instruction

The success of this course depends heavily on the support provided to participants from GBIF's network of capacity enhancement mentors. Visit the GBIF page on capacity enhancement mentoring to read more about these individuals and their contributions.

The following individuals are recognized for their significant contributions to the course design, materials and instruction:

- Nestor Beltran*
- David Bloom
- Katia Cezón*
- Dag Endresen
- Alberto González-Talaván*
- Sharon Grant*
- Marie-Elise Lecoq
- Sophie Pamerlon*
- Nicolas Noé*
- Mélianie Raymond*
- Laura Anne Russell*
- John Wieczorek
- Paula Zermoglio

*Originators of the curriculum

Special acknowledgement to Arthur Chapman for the reuse of his materials on Data Quality.

Translators

French

- Maxime Coupremanne
- Jaures Gbètoho
- Marie Grosjean
- Patricia Mergen
- Sophie Pamerlon
- Andry Jean Marc Rakotomanjaka
- Y. Sabastian Wirsiy

Portuguese

- Rui Figueira

- Clara Baringo Fonseca
- Keila Elizabeth Macfadem Juarez
- Tainan Messina

Spanish

- Leonardo Buitrago
- Victor Chocho
- Camila Plata
- Anabela Plos
- William Ulate
- Paula Zermoglio

Resources

- The online tabletop platform is provided by PlayingCards.io. Much appreciation and thanks to Jwalant Patel and Eric Ma for finding and helping to create the online playing tables and to Kate Webbink for artistic expertise.
- Icons used in this course are made by Freepik from www.flaticon.com
- OpenRefine
- Integrated Publishing Toolkit

Resource support

- Belgium Biodiversity Platform
- GBIF France
- GBIF Norway
- GBIF Spain
- SiB Colombia
- The Field Museum
- VertNet

Colophon

Suggested citation

GBIF Secretariat (2021) GBIF Biodiversity Data Mobilization Course. 12th edition. GBIF Secretariat: Copenhagen. https://doi.org/10.35035/ce-c6cr-6w42. [Date of course.]

Contributors

The *GBIF Biodiversity Data Mobilization Course* was originally developed as part of Biodiversity Information Development, a programme funded by the European Union. The original curriculum was created by Nestor Beltran, Sharon Grant, Nicolas Nóe, Sophie Pamerlon, Alberto González-Talaván,

Mélianie Raymond, Laura Anne Russell and Katia Cezón, with additional contributions by GBIF trainers, mentors and students.

Licence

Persistent URI

https://doi.org/10.35035/ce-c6cr-6w42

Document control

12th edition, May 2021