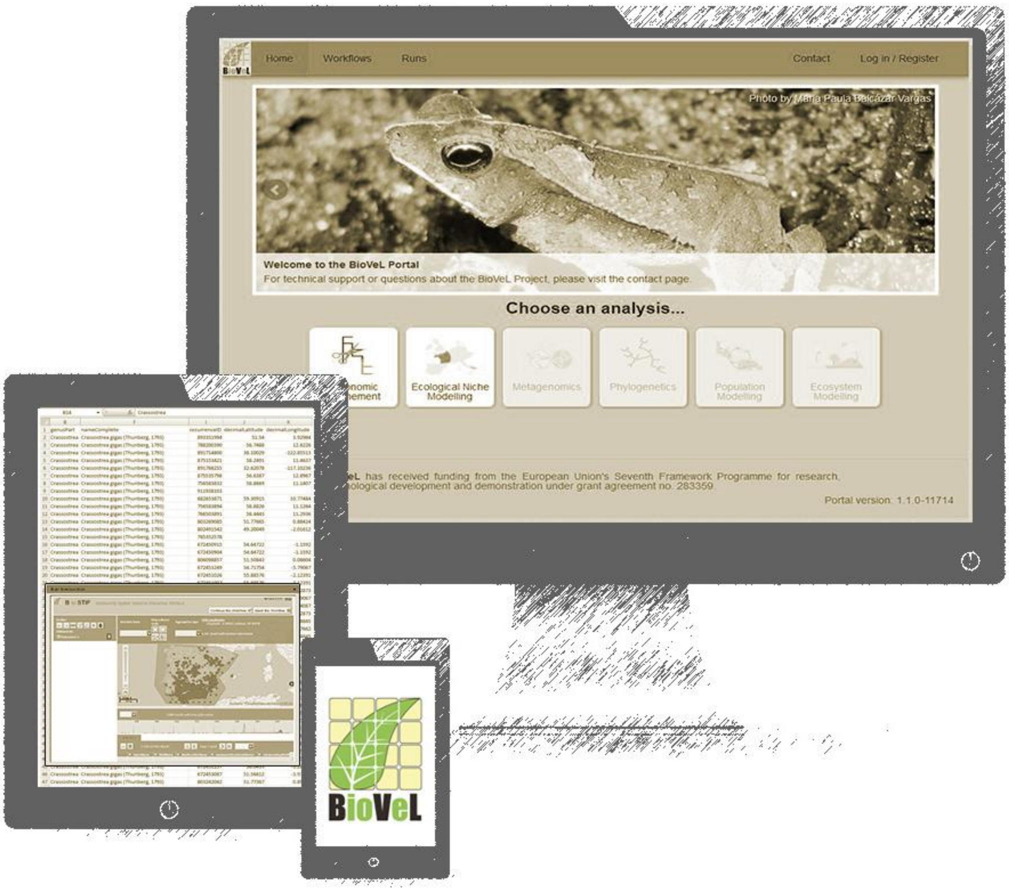
GBIF
www.gbif.org

Data Refinement Using the BioVeL Portal

González-Talaván, A., Mathew, C., Obst, M. & Paymal, E.

Version 1.0



Welcome to the BioVeL Portal
For technical support or questions about the BioVeL Project, please visit the contact page.

eL has received funding from the European Union's Seventh Framework Programme for research, ological development and demonstration under grant agreement no. 283359.

November 2014

Suggested citation:
González-Talaván, A., Mathew, C., Obst, M. & Paymal, E. (2014). Data Refinement Using the BioVeL Portal. Copenhagen: Global Biodiversity Information Facility. Available online at http://www.gbif.org/orc/?doc_id=5983.

ISBN number: not yet assigned.
Persistent URI: http://www.gbif.org/orc/?doc_id=5983.
Language: English.

This document is the product of a collaboration between the EC FP7 project Biodiversity Virtual e-Laboratory (BioVeL) and the Global Biodiversity Information Facility (GBIF). BioVeL has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 283359.

Disclaimer:
We welcome comments about this manual at info@gbif.org. Please direct your comments about the BioVeL Portal directly to the project representatives at support@biovel.eu. GBIF's characterizations and descriptions of the BioVeL data refinement mechanisms do not represent specific endorsements.

Document Control

Version	Description	Date of release	Authors (in alphabetic order)
1.0	First public release	November 2014	González-Talaván, A., Mathew, C., Obst, M. & Paymal, E. Edited by González-Talaván, A. Language review by Copas, K.

Cover Art Credit: *GBIF Secretariat, 2014. Based on an image by mocho1, obtained via freeimages.com.*

Executive summary

This document is a practical guide on how to assess the quality of biodiversity datasets, like those accessible through the GBIF Network using BioVeL online workflows portal. The manual takes a practical tutorial-based approach that the reader can repeat using the sample datasets provided. A complement of practical exercises based on real-case scenarios should help users attain the skills demonstrated in the tutorials. The manual also provides a generic introduction about data quality and the use of workflows for those who wish to familiarize themselves with the basic theory behind these practices.

About the publishers

The production of this manual has been possible thanks to:

The Biodiversity Virtual eLaboratory (BioVeL)

BioVeL is a virtual e-laboratory that supports research on biodiversity issues using large amounts of data from cross-disciplinary sources. BioVeL outlines step-wise computerized 'workflows'—or a series of data analyses for processing data from one's own research or from existing sources. A researcher can build custom workflows by selecting and applying successive 'services', or data processing techniques.

BioVeL cuts down research time and overhead expenses by making a library of existing workflows available for reuse and providing access to a worldwide network of expert scientists who develop, support and use workflows and services. This virtual e-laboratory pools interests and shared-knowledge on biodiversity research, and fosters an international community of researchers and partners on biodiversity issues. At the core of BioVeL is a consortium of 15 partners from nine countries, as well as an outer circle of 'Friends of BioVeL'.

Learn More

- Project-related information
- BioVeL portal
- Catalogue of services
- BioVeL Wiki
- BioVeL area on myExperiment
- Workbench for designing new workflows

GBIF: the Global Biodiversity Information Facility

The Global Biodiversity Information Facility (GBIF) is an international open data infrastructure, funded by governments. It allows anyone, anywhere to access worldwide evidence of species' existence via the Internet.

By encouraging and helping institutions to publish data according to common standards, GBIF enables research not possible before, and informs better decisions to conserve and sustainably use the biological resources of the planet.

GBIF operates through a network of nodes, coordinating the biodiversity information facilities of Participant countries and organizations, collaborating with each other and the global Secretariat to share skills, experiences and technical capacity.

GBIF's vision is ‘A world in which biodiversity information is freely and universally available for science, society and a sustainable future’.

Learn More

- GBIF.org

Table of Contents

1. Background information

We would like to start our work by introducing the guide itself, describing its contents and its structure so you can better evaluate how to make best use of it.

This guide is a compilation of tutorials, exercises and recommendations around the use of the BioVeL online portal (http://portal.biovel.eu), which can analyse and improve the relative quality of existing biodiversity data, such as those published through the GBIF network. While the BioVeL portal offers many other functions, in particular for modelling (ecological niches, populations, ecosystem functioning) and for phylogenetics and metagenomics, this manual focuses on the ones more closely related to data quality.

This text is the result of the collaboration between BioVeL and GBIF and hopes to contribute to improved management and curation of digital biodiversity information.

Send your comments or suggestions to the editorial teams at BioVeL and GBIF.

1.1. How to use this manual

This consists of three main sections:

- An introduction to the concepts used in the guide related to biodiversity data quality and workflows in chapters 2 and 3.
- A series of tutorials and practical examples that demonstrate the different functions of the portal in chapters 4, 5, 6 and 8.
- An analysis of how users can apply different techniques to the different domains of biodiversity data in chapter 7.

The guide’s modular design will allow you to consult only the sections that you need at any moment. The chapters need not be read in sequence. If you are already familiar with common concepts about biodiversity data quality, starting with chapter 4 will permit you to get into the portal immediately. If you encounter a particular issue while working with your data, consult the section of chapter 7 corresponding to your data domain and explore your options for addressing it.

The supplementary exercises that follow tutorials 4 to 11 will improve your understanding of how to use the procedures while giving you an opportunity to test your newly acquired skills. We would recommend you to go through the exercises, as certain details can only be fully explained through these practical cases.

Users can repeat all the tutorials and exercises using the provided instructions and datasets. The example datasets follow the CSV (comma-separated values) format, with controlled header terms, which are needed for the successful execution of certain exercises. You can also use these files as templates for your own analyses. More details can be found on the corresponding BioVeL wiki page.

You can find below the list of files that will be used in this guide. You can either download them now and save them in your computer for later, or wait till they are required by the relevant exercises (links are also provided there):

- Exercise File 1: BioVeL-GBIF_BPG_File1_105OR.csv. 105 occurrence records of aquatic organisms recorded in Sweden, in CSV format. These diverse data points include scientific names, geographical coordinates, time, altitude and habitat types. Available in http://links.gbif.org/biovel_f1.
- Exercise File 2: BioVeL-GBIF_BPG_File2_CrassostreaGigas.csv. A very simple data input file in CSV format with a single name column with one name: *Crassostrea gigas*, the pacific oyster. Available in http://links.gbif.org/biovel_f2.

1.2. Target audience

This manual will help those wishing to learn how to assess the quality of a given biodiversity dataset, and to improve its relative quality through automatic corrections. This may include biodiversity data managers, staff associated to natural history collections (curators and technicians), staff working in biodiversity projects (e.g., researchers, practitioners, data collectors) and natural resource managers, among many others.

The manual assumes that you have a basic understanding of taxonomic concepts, specifically scientific names / synonyms and species occurrence records. Basic computer skills—file management, basic office and internet tools—are also needed to complete the exercises successfully.

1.3. Do you need help?

If you encounter difficulties while doing some of the exercises or you have questions that are not covered by this manual, you can look for help by:

- Visiting the BioVeL community discussion forum where you can connect with other users of the platform. The forum also contains answers to earlier questions posted by fellow users. The top menu of the BioVeL portal always includes a link to the forum.
- Contacting BioVeL user support.
- Joining the GBIF/TDWG biodiversity data quality interest group. This online community dedicated to issues around data quality and fitness for use is a good place to start with generic data-quality related issues.

2. Introduction to biodiversity data quality

Data quality is an abiding concern of biodiversity data holders and users, but problems related to it became more evident once networks such as GBIF started aggregating and publishing large collections of data to the web.

The community of data collectors and users has made continuous efforts to address data quality issues through all levels of the data management chain, from data collection to data use. This manual itself contributes to these efforts, by helping those working with biodiversity datasets to evaluate their relative quality and to perform automated routines that improve quality.

Along the years, the concept of 'data quality' has been progressively taken over by that of 'fitness for use', as the community acknowledged that the quality of the data is best measured in relation to its intended use. This makes it essential to define which level of quality or precision is required for any analysis we wish to perform in the data (e.g., which geographical or taxonomic definition), and to have mechanisms to evaluate whether a given data record or dataset meets our requirements.

To learn more about the concept of quality and fitness for use in relation to biodiversity data, we would recommend you to review these two resources published by GBIF:

- 'Principles of Data Quality', by A. Chapman (2005).
- 'GBIF Position Paper on Future Directions and Recommendations for Enhancing Fitness-for-Use Across the GBIF Network' by A. W Hill, J. Otegui, A. H. Ariño and R. P. Guralnick (2010).

2.1. Defining data quality requirements

The list of data quality requirements for your study is one of the first things any researcher or analyst needs to define. These criteria, which are specific to each study, will determine which data are fit for your own particular use. Most will derive from the methodology you plan to follow (e.g., some analysis require a certain precision in the geographic information), but they can also come from the type of organisms studied, their ecology, the scale of the study, etc.

Consult the documentation associated to the software or protocol you plan to use to help you define your specific criteria. The GBIF manual mentioned above—'Principles of Data Quality'—provides useful generic guidance on data quality issues related to the different dimensions of biodiversity data.

2.2. Initial dataset selection and filtering

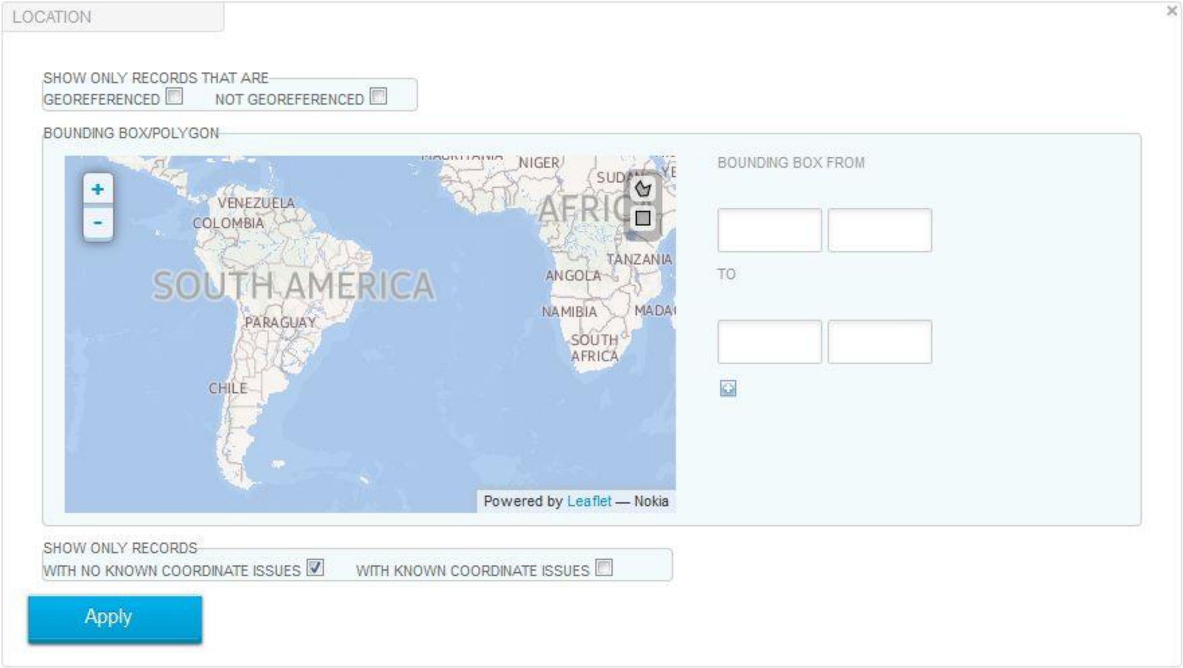
A key step for avoiding problems with data quality later is to select the right dataset from the start. When using data available online, it is important to know what (if any) mechanisms the data aggregator provides to select and filter records based on data quality and precision.

GBIF analyses data during indexing, detecting and recording errors related to 42 potential problems (as of June 2014). These checks are applied to

- geographic information (coordinates, geodetic datum, depth, altitude and country information)
- taxonomic information (concordance with the GBIF backbone taxonomy)
- temporal information (dates of collection, identification and modification),
- basis of record (concordance with a controlled vocabulary that includes preserved specimen, observation, etc.)
- biological type status (concordance with a controlled vocabulary that includes isotype, holotype, etc.)
- URL/URI consistency (for multimedia and references links).

In Tutorial 4 you can learn how to use the BioVeL portal to automatically retrieve occurrence records from the GBIF network. This exercise will not allow to filter occurrences according to the results of the indexing checks. If you are interested in manipulating the data in this way, you will need to extract the data manually from GBIF.org.

To do so, visit the occurrences area on GBIF.org (http://www.gbif.org/occurrence), and run a query that excludes records with coordinate issues using the 'location' filter (see image below)



LOCATION

SHOW ONLY RECORDS THAT ARE
GEOREFERENCED □ NOT GEOREFERENCED □

BOUNDING BOX/POLYGON

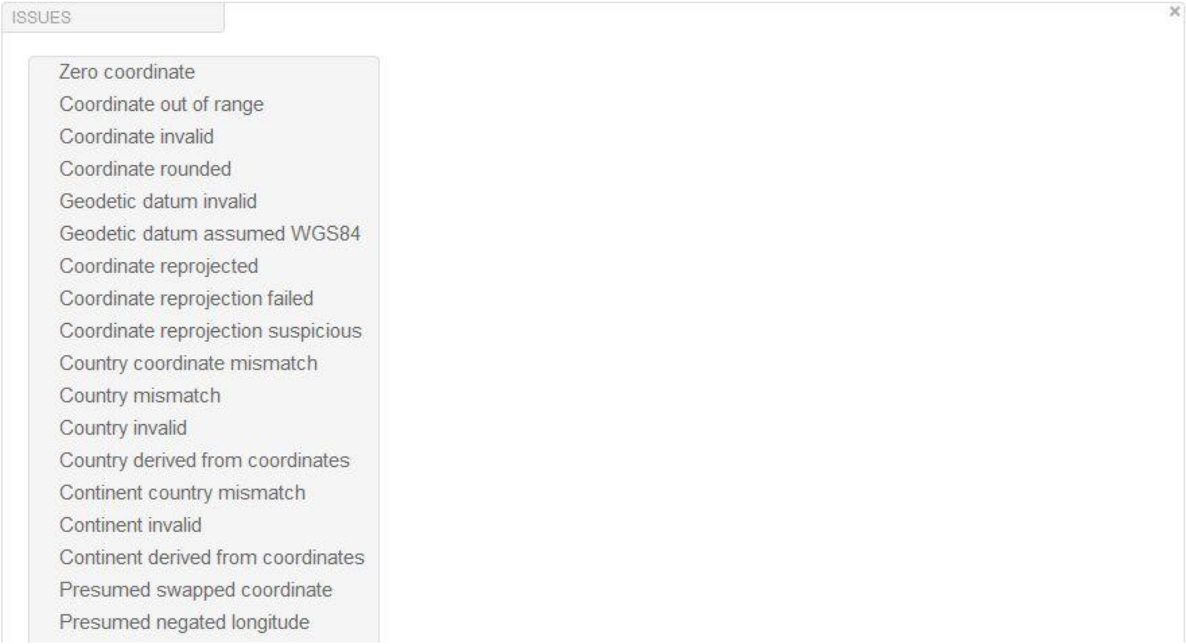
BOUNDING BOX FROM

TO

SHOW ONLY RECORDS
WITH NO KNOWN COORDINATE ISSUES ☑ WITH KNOWN COORDINATE ISSUES □

This filter can flag occurrences using the field 'hasGeospatialIssues'. Should you need to access and filter data via other methods, like GBIF's web services, make queries with this selection for similar results.

Please note that the process of analysing and marking records with issues is an automated one; it is possible that records suitable for your particular use may be marked as problematic (false positives) or the other way around (false negatives). An additional 'Issues' filter at GBIF.org allows you to investigate problematic records so you can evaluate whether to include them in your analysis or not.


ISSUES
×
Zero coordinate
Coordinate out of range
Coordinate invalid
Coordinate rounded
Geodetic datum invalid
Geodetic datum assumed WGS84
Coordinate reprojected
Coordinate reprojection failed
Coordinate reprojection suspicious
Country coordinate mismatch
Country mismatch
Country invalid
Country derived from coordinates
Continent country mismatch
Continent invalid
Continent derived from coordinates
Presumed swapped coordinate
Presumed negated longitude

If your potential dataset includes multiple types of issues, we recommend downloading one category at a time to stay aware of any possible interpretation difficulties.

2.3. Data quality assessment

Once you have selected your initial dataset and defined your own fitness-for-use criteria, you need to assess which records comply with them.

Different kinds of errors present different degrees of 'detectability', depending on how much does a particular error make the record stand out from what we consider correct data. In fact, some of the errors are only detectable when we analyse the dataset as a whole.

Many of the operations described in this manual will help you to identify records that meet your quality requirements. Some qualifying categories you may want to consider are[1]:

- **Missing data values** where data should occur (e.g., if a species epithet is included in its field, the genus should be included as well)
- **Incorrect data values,** including those that do not comply with the agreed list of possible values for a field in a given context (e.g., misspellings)

[1] Extracted from 'Principles and Methods of Data Cleaning', Chapman, A., 2005.

- **Nonatomic data values** that should be split across multiple fields but are lumped into a single one (e.g., country information included in the locality field)
- **Domain schizophrenia**, using some data fields in unintended ways (e.g., in a species field: '*globulus?*', '*sp. nov.*', '*to be determined*')
- **Duplicate occurrences** of records that have been introduced twice, with both entries referring to exactly the same occurrence (e.g., using different reference taxonomies at the time of digitization)
- **Inconsistent data values**, which can occur when multiple sources contribute to the dataset using different criteria (e.g., data recorded in different formats or languages) or when data derives from different sources (a common problem when obtaining data from data aggregators like GBIF)

You can apply these categories more or less rigidly depending on the type of information you need. Chapter 7 provides a detailed summary of the most frequently appearances they make in the different data domains that make up an occurrence record.

2.4. Data refinement and cleaning

The fact that a record is affected by a particular kind of error does not necessarily require its exclusion from your study. Data refinement and cleaning aims precisely at increasing the quality of data records so that they are fit for a given purpose.

In the same way that some errors are easier to detect than others, some are easier to correct. Some authors refer to this quality as the 'resolvability' of an error.

This manual focuses on the use of the BioVeL portal to help fix some of these errors automatically. Extreme care is required to ensure that automatic data processes are applied correctly and to the correct subset of the dataset.

When performing data refinement, you should always respect these general **recommendations**:

1. **Preserve the original data** (sometimes also referred as verbatim data). In the event of problems, you can always return to these data to analyse what happened and run your analysis again. As a general rule, always add information instead of deleting or replacing it.
2. **Document any changes**: what was changed, which protocol/criteria were followed, when and by whom. Also describe any uncertainties or alternative interpretations that you have considered and discarded (and why).
3. Document the result of your data quality checks **no matter the result**: a positive result is also valuable information for future users.
4. Whenever possible, **contact the original data publisher** and alert them about the result of any data quality checks that you have performed in their data.

By following these recommendations, you will contribute to increasing public information about the quality of the available data, while keeping you and other users from repeating work you have already performed on the same data.

For general guidance on how to define data refining operations to include in your work, consult the GBIF manual 'Principles and Methods of Data Cleaning' by A. Chapman (2005).

3. Introduction to workflows

This chapter presents a brief introduction to workflows and to the infrastructure that supports them.

3.1. General introduction to workflows

The quantity and heterogeneity of data in the biodiversity sciences have given rise to many distributed resources. Scientists in biodiversity research often need to combine data from different resources, analyse them or apply transformations. As described above, researchers also need to standardize data formats and ensure the quality of the data, checking for misspellings, aberrations, etc.

Workflows are a series of computational tasks for a range of analytical purposes. They consist of modularised software units (‘services’) that researchers can link, repeat, share, reuse and repurpose. BioVeL workflows, for example, offer a practical solution to respond to the needs of biodiversity data processing, including data refinement, cleaning and analysis.

3.2. Workflow runs

Users of the BioVeL portal, execute a workflow as a ‘workflow run’—a simple web interface that provides access to a pool of ready-made workflows. Chapter 4 provides a more thorough introduction to the portal and how it uses workflows, but in short, it allows you to manage, share and save workflow results, and to monitor and interact with running workflows while changing parameters and directing your analyses. You can upload your own workflows and run those too.

A ‘sub-workflow’ is a set of services leading to a set of results within a larger workflow.

3.3. Designing and constructing workflows

Researchers can discover new tools and resources (‘services’) through the BiodiversityCatalogue and then combine them as steps in a workflow. Components are modularised units that are well-documented fragments of workflow to perform specific tasks (like a sub-workflow does). They are designed to be used as steps in other workflows. A ‘plug-and-play’ approach simplifies workflow construction.

The Taverna Workbench provides a graphical environment where researchers can design and construct new analysis protocols as workflows, or customise workflows expressing existing protocols, before they are deployed and shared through the BioVeL portal.

3.4. Sharing and discovering workflows

myExperiment (http://www.myexperiment.org/) is a social network and repository where users can publish, share, and discover workflows and workflow components. You can search and discover workflows from myExperiment through both the Taverna workbench and the BioVeL portal. This means you have access to a large pool of existing methods to reuse or customise. BioVeL workflows are also to be found in myExperiment at http://www.myexperiment.org/groups/643.html.

In the following chapters we will see how to use workflows in a practical way in the BioVeL portal.

4. Introduction to the BioVeL portal

The BioVeL portal (http://portal.biovel.eu) is a powerful virtual analysis workbench, designed by scientists for scientists. The BioVeL portal enables users to conduct complex analyses routinely in a web-based environment. Features of the portal include: tools for discovering, organizing and managing your analyses; batch processing; session management, which allow returning to your work where you left it off; and tools for sharing your results between collaborators.

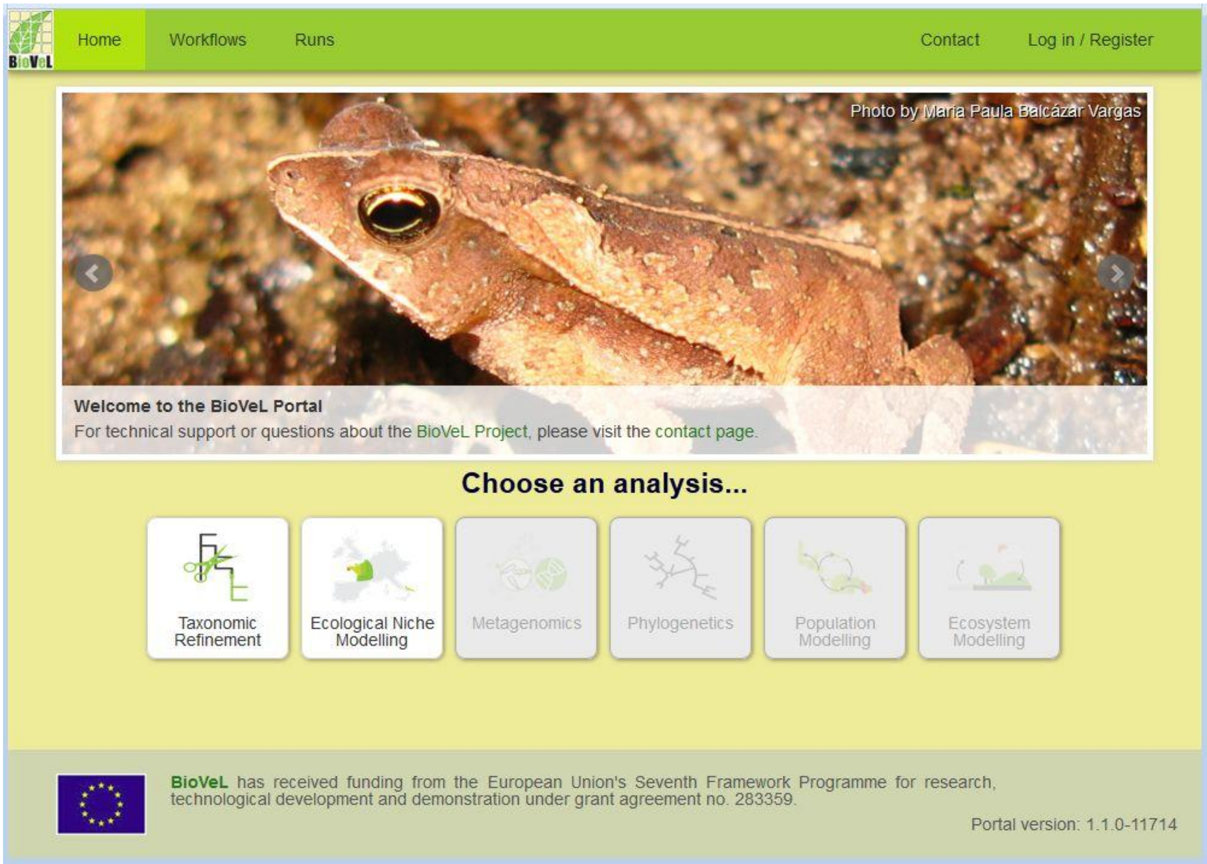


Photo by Maria Paula Balcázar Vargas

Welcome to the BioVeL Portal

For technical support or questions about the BioVeL Project, please visit the contact page.

BioVeL has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 283359.

Portal version: 1.1.0-11714

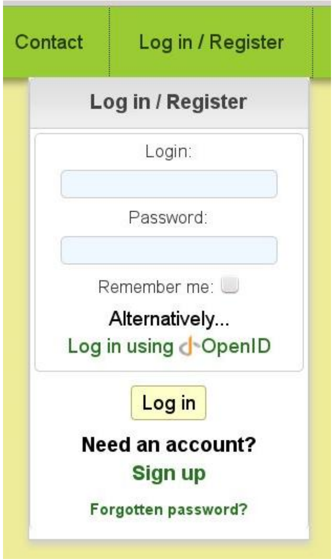
In this chapter we will have the opportunity to get familiar with the different features of the portal. We will see how to create runs (= instances) of a workflow, and how to manage them. We will see the different elements of a workflow run before we start applying them to specific cases of data refinement.

4.1. Tutorial 1: Getting familiar with the BioVeL portal

Our first tutorial introduces the portal's basic controls while preparing us for the next exercises. It should take around five minutes to complete it.

To keep the work that you perform in the BioVeL portal organized, you need to create an account by following these steps:

1. Launch a web browser and visit http://portal.biovel.eu. The developers of the tool recommend using Mozilla Firefox for a better user experience.
2. Click on 'Log in / Register' link found at the top right corner of the page.
3. On the drop down, click on the 'Sign up' link.
4. On the sign up page, fill in the required information and click the 'Register' button.



Once logged into the site, you can familiarize yourself with the basic layout of the portal and explore its functions by clicking on the various menu items and buttons. Through the main menu on the top left, you'll find the portal's two main elements: workflows and workflow runs.

Have a quick look at the different **workflows** within the various service sets. For each one you will find a description on its intended use and the problems they try to solve.

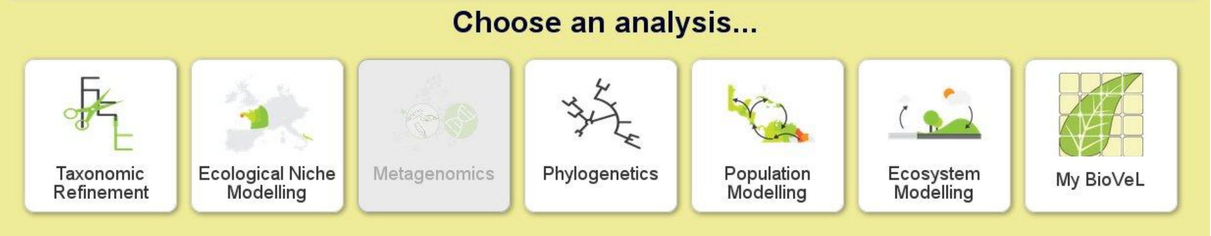
As this is the first time that you log in, there should be no **runs** associated to your account, but you can see also run instances by other users.

To finish, have a look at the **'My BioVeL'** area, where your current and past work will be displayed.

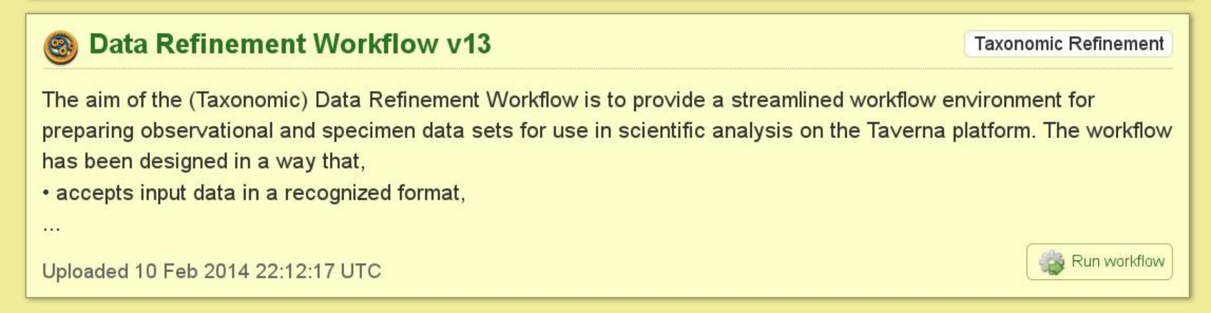
4.2. Tutorial 2: Initializing a workflow run

We will start our work with workflows by creating a simple workflow run. This tutorial should take around five minutes to complete.

1. Login in to the BioVeL portal (http://portal.biovel.eu) and choose the 'Taxonomic Refinement' analysis in the centre of the home page. You can also use the 'Workflows' menu on the top left and then select 'Taxonomic Refinement' in the menu on the left of the new page that opens.



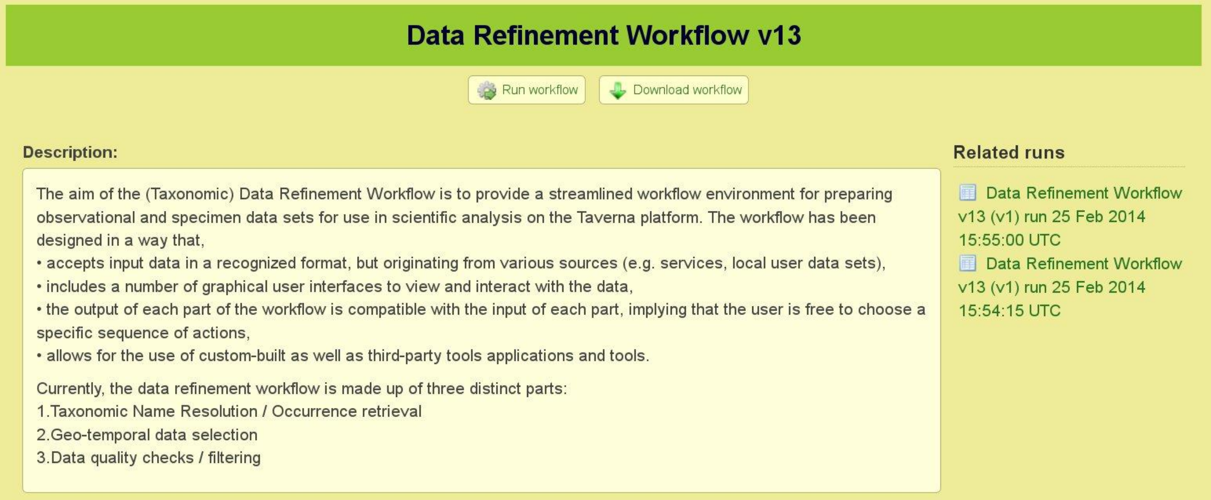
2. Upon being redirected to the workflows page, you'll see a short description for each available workflow. Choose the workflow named 'Data Refinement Workflow' by clicking on its name (you can also directly run the workflow using the 'Run workflow' button at the bottom-right of the workflow box).



The aim of the (Taxonomic) Data Refinement Workflow is to provide a streamlined workflow environment for preparing observational and specimen data sets for use in scientific analysis on the Taverna platform. The workflow has been designed in a way that,
• accepts input data in a recognized format,
...

Uploaded 10 Feb 2014 22:12:17 UTC

3. On the 'Data Refinement workflow' page, you can find more information about this particular workflow. Click on the 'Run workflow' button at the top to start an instance of this workflow.



Description:

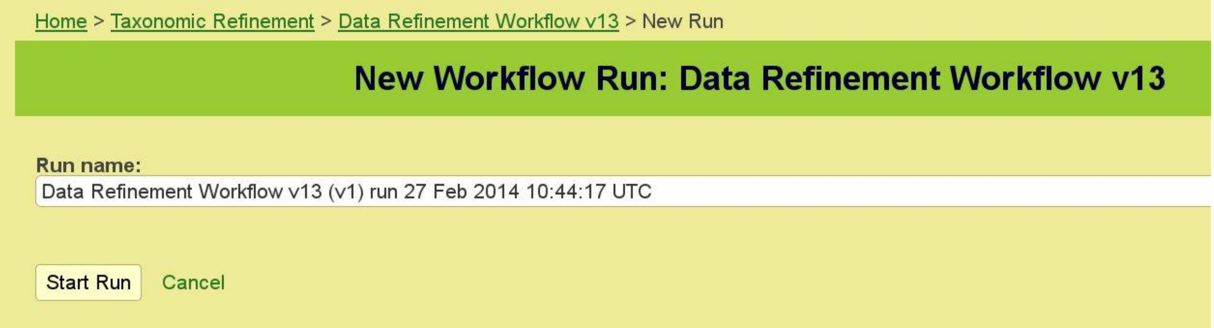
The aim of the (Taxonomic) Data Refinement Workflow is to provide a streamlined workflow environment for preparing observational and specimen data sets for use in scientific analysis on the Taverna platform. The workflow has been designed in a way that,
• accepts input data in a recognized format, but originating from various sources (e.g. services, local user data sets),
• includes a number of graphical user interfaces to view and interact with the data,
• the output of each part of the workflow is compatible with the input of each part, implying that the user is free to choose a specific sequence of actions,
• allows for the use of custom-built as well as third-party tools applications and tools.

Currently, the data refinement workflow is made up of three distinct parts:
1.Taxonomic Name Resolution / Occurrence retrieval
2.Geo-temporal data selection
3.Data quality checks / filtering

Related runs

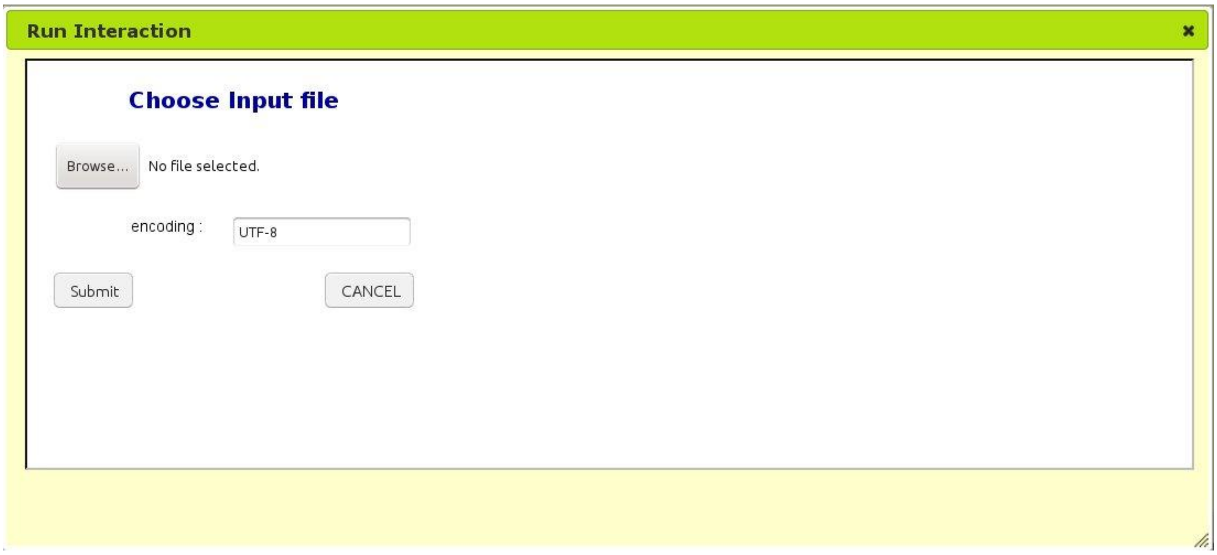
Data Refinement Workflow v13 (v1) run 25 Feb 2014 15:55:00 UTC
Data Refinement Workflow v13 (v1) run 25 Feb 2014 15:54:15 UTC

4. On the next page you can edit the name of the workflow run. Preferably choose a name that helps you clearly identify it later. To start the run click on the 'Start Run' button.


Home > Taxonomic Refinement > Data Refinement Workflow v13 > New Run
New Workflow Run: Data Refinement Workflow v13
Start Run
Cancel

Run name: Data Refinement Workflow v13 (v1) run 27 Feb 2014 10:44:17 UTC

5. The workflow run is then started (it may take a moment). You should see the first interaction page within a popup panel as the one showed below (press the browser refresh button if the interaction does not appear after a minute or two).
   Right now we will not continue defining an input file. Just close the interaction window using the cross button in the top right corner. That will maintain the workflow run waiting for user input.

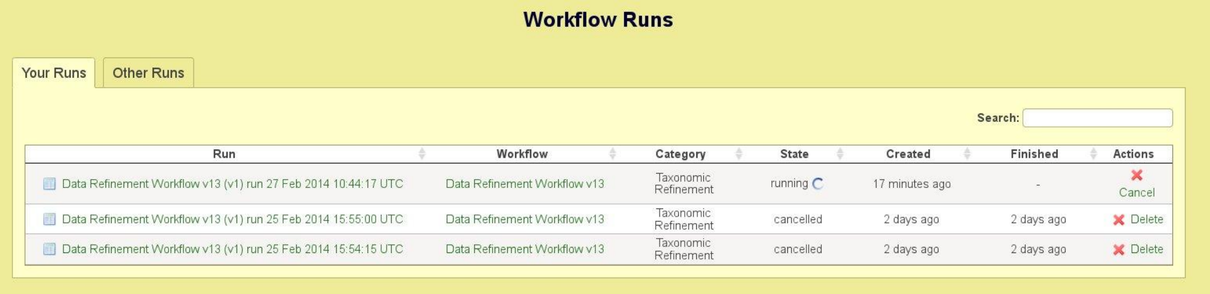


4.3. Tutorial 3: Managing workflow runs

The BioVeL portal enables researchers to manage numerous workflow runs under a single user profile. In addition, it supports asynchronous workflow execution, which allows users to log out of the portal while running a workflow and then log in at a later time to continue the workflow. Following this tutorial should take less than five minutes.

To take this tutorial you need to have previously created a run: please refer to Tutorial 2 if you need help on that.

1. If you have been previously working on the portal and are logged into the site, click in the 'Log out' button at the top-right corner to log out. Also close the web browser as if you had finished your work for the moment and to end the current browser session.
2. Launch the web browser again and go to the BioVeL portal (http://portal.biovel.eu/). Log in again using your credentials.
3. Go to the runs area by clicking on the 'Runs' item in the top right menu. You will find there a list with any runs that you started earlier, along with additional information like their current status. If you click on the name of one of the elements it will take you back to the interaction page resulting from the last task assigned to the workflow before you logged out of the portal. You can continue your work from there or switch to another workflow run by going back to the runs page.



Run	Workflow	Category	State	Created	Finished	Actions
Data Refinement Workflow v13 (v1) run 27 Feb 2014 10:44:17 UTC	Data Refinement Workflow v13	Taxonomic Refinement	running	17 minutes ago	-	✖ Cancel
Data Refinement Workflow v13 (v1) run 25 Feb 2014 15:55:00 UTC	Data Refinement Workflow v13	Taxonomic Refinement	cancelled	2 days ago	2 days ago	✖ Delete
Data Refinement Workflow v13 (v1) run 25 Feb 2014 15:54:15 UTC	Data Refinement Workflow v13	Taxonomic Refinement	cancelled	2 days ago	2 days ago	✖ Delete

4. Make sure you cancel any workflow runs that you don't need using the link on right. It will free up resources in the system. If you are following these tutorials in sequence, cancel and delete the run we created in the Tutorial 2.

5. The taxonomic data refinement workflow: basic use

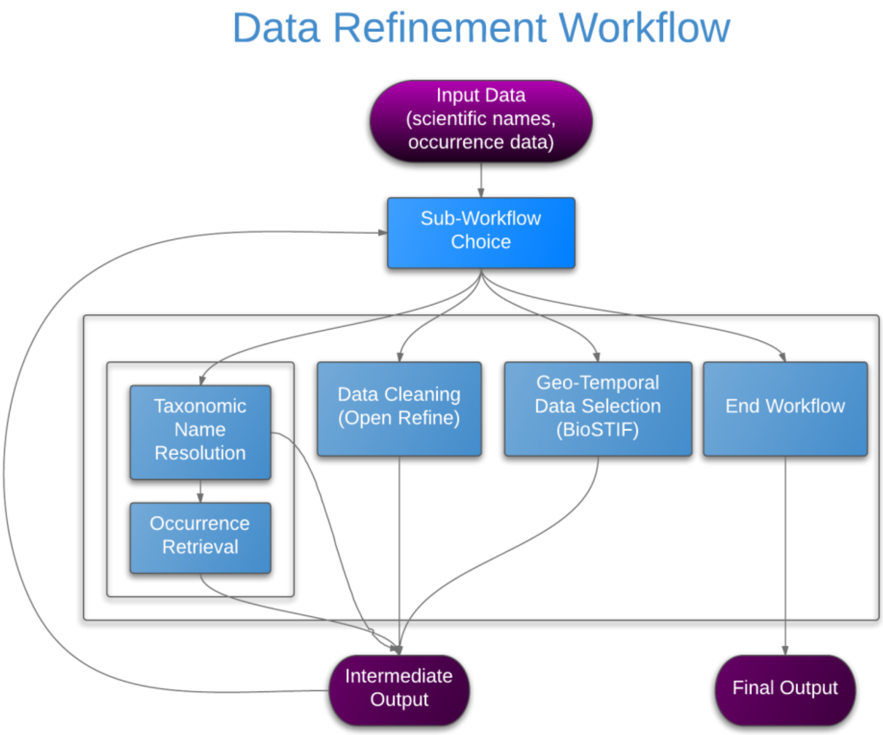
The aim of the taxonomic data refinement workflow is to provide a streamlined environment for preparing species occurrence datasets (i.e., based on observations or collection specimens) for use in scientific analysis.

At the time of writing this manual, the workflow is made up of three distinct parts:

1. **Synonym expansion / occurrence retrieval**: users can expand a list of scientific names to include synonyms and retrieve occurrence data for them. Both the synonym expansion and occurrence retrieval are built on generic frameworks that allow for the inclusion of multiple sources.
2. **Geo-temporal data selection**: users can select or exclude data records based on geographical and temporal criteria. Geographic selections can be made by drawing areas on a map or filtering data based on geo-markers like country or latitude/longitude. Records related to specific time periods can also be isolated using time-based filtering. The web-based BioSTIF client provides these functionalities.
3. **Data quality checks / filtering**: users can apply a set of data quality and data integrity checks to the selected data. The main interface for this phase uses OpenRefine along with a custom-made extension for biodiversity data to access various local and external functionalities.

The three tutorials in this chapter (4 to 6) introduce these three parts and some of the basic functionality of the sub-workflows that represent them.

More information about this workflow can be found on the corresponding BioVeL Wiki page.


Data Refinement Workflow
Input Data (scientific names, occurrence data)
Sub-Workflow Choice
Taxonomic Name Resolution
Occurrence Retrieval
Data Cleaning (Open Refine)
Geo-Temporal Data Selection (BioSTIF)
End Workflow
Intermediate Output
Final Output

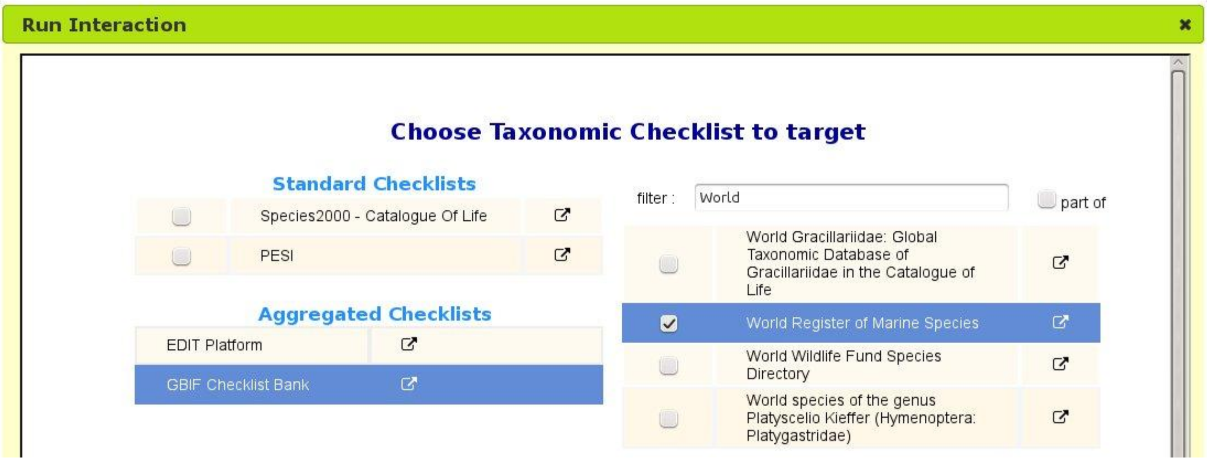
5.1. Tutorial 4: Synonym expansion and retrieval of occurrence records

The first sub-workflow of the data refinement workflow help users to retrieve valid taxonomic synonyms for one or several species names, and then use the list of synonyms to retrieve occurrence records from data repositories such as the GBIF network. It may take up to 10 minutes to complete this tutorial.

The ability to resolve scientific names to their corresponding taxonomic concept information and to retrieve species occurrence are recurring requirements in many fields of research related to biodiversity like ecology, phylogenetics, modelling, etc. This section of the workflow tries to simplify these two tasks as a first step towards compiling data that is fit for use in a given analysis.

As an input to this sub-workflow, you need to have a file with your initial list of scientific names. A very simple CSV file with a header 'nameComplete' would suffice (if other fields exist, they will be ignored in this sub-workflow). In this tutorial, use the Exercise File 2: BioVeL-GBIF_BPG_File2_CrassostreaGigas.csv which contains a single name: *Crassostrea gigas*. Make sure you have downloaded and saved it in your computer and that you know its location in your hard drive.

1. Create a 'Data Refinement Workflow' run as explained in Tutorial 2.
2. On the 'Choose Input File' dialogue, click the 'Browse' button and select the example data file mentioned above. Click on 'Submit'. The data will be then uploaded to the portal.
3. Once the upload is finished, a new dialogue window titled 'Choose Sub-Workflow ' will appear. Please choose 'Taxonomic Name Resolution / Occurrence Retrieval' and click on 'OK'. Please be aware that the BioSTIF application may take some time to load, especially if you are loading a large number of records (>50.000 records).
4. A new interaction window will appear, offering you to select which checklists you want to use to search for synonyms of your list of taxa. Choose the appropriate taxonomic checklist for your species of interest. You can choose more than one checklist at a time but we recommend sticking to 2-3 checklists per workflow run.
   Checklist repositories such as the GBIF checklist bank offer thousands of checklists. We would recommend searching for the checklists you are interested in first, using the more complete interface of the GBIF.org site, and then use the 'filter' text box in the dialogue back in the BioVeL portal to locate the exact checklist you want to use by its name.
   For this tutorial, please choose the 'GBIF Checklist Bank' and 'World Register of Marine Species' as the individual checklist. If the checklist that you have selected has usage restrictions or a usage agreement associated to it, you would have to estate your agreement with the usage conditions (otherwise you won't be allowed to use the resource). Click on 'OK' when you are done. Be aware that, depending on your screen resolution, the 'OK' button may not be visible till you scroll down in the window.



Run Interaction

Choose Taxonomic Checklist to target

Standard Checklists

- ☐ Species2000 - Catalogue Of Life
- ☐ PESI

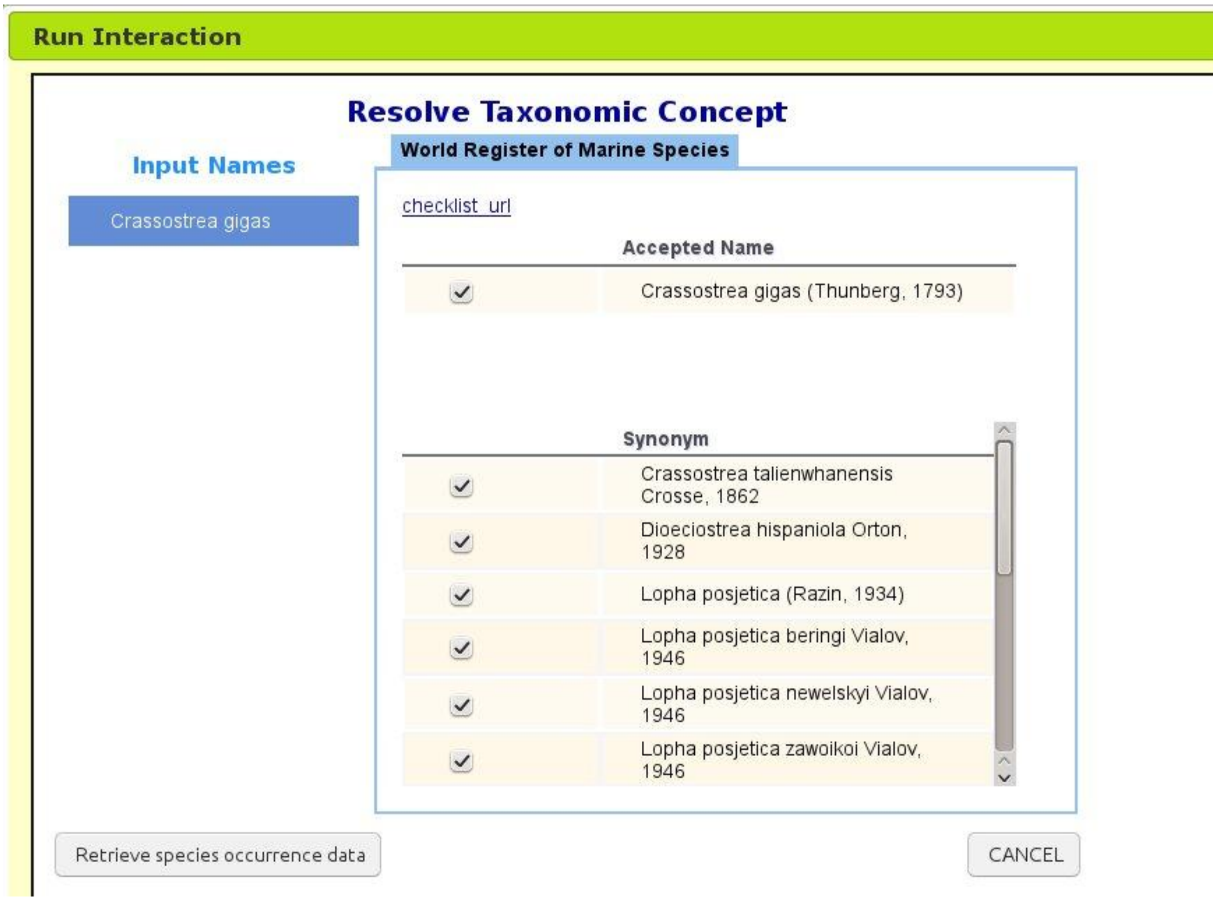
Aggregated Checklists

- EDIT Platform
- GBIF Checklist Bank

filter : World ☐ part of

- ☐ World Gracillariidae: Global Taxonomic Database of Gracillariidae in the Catalogue of Life
- ☑ World Register of Marine Species
- ☐ World Wildlife Fund Species Directory
- ☐ World species of the genus Platyscelio Kieffer (Hymenoptera: Platygastridae)

5. The next interaction window shows you the result of the name resolution with the given target checklist(s). You have the opportunity now to select which of the registered synonyms you want to use to retrieve occurrence data. Use the checkbox on the left of each name to include or exclude the name from the list, according to your own criteria. Click on 'Retrieve species occurrence data' when you are done.



Run Interaction

Resolve Taxonomic Concept

Input Names

Crassostrea gigas

World Register of Marine Species

checklist_url

Accepted Name

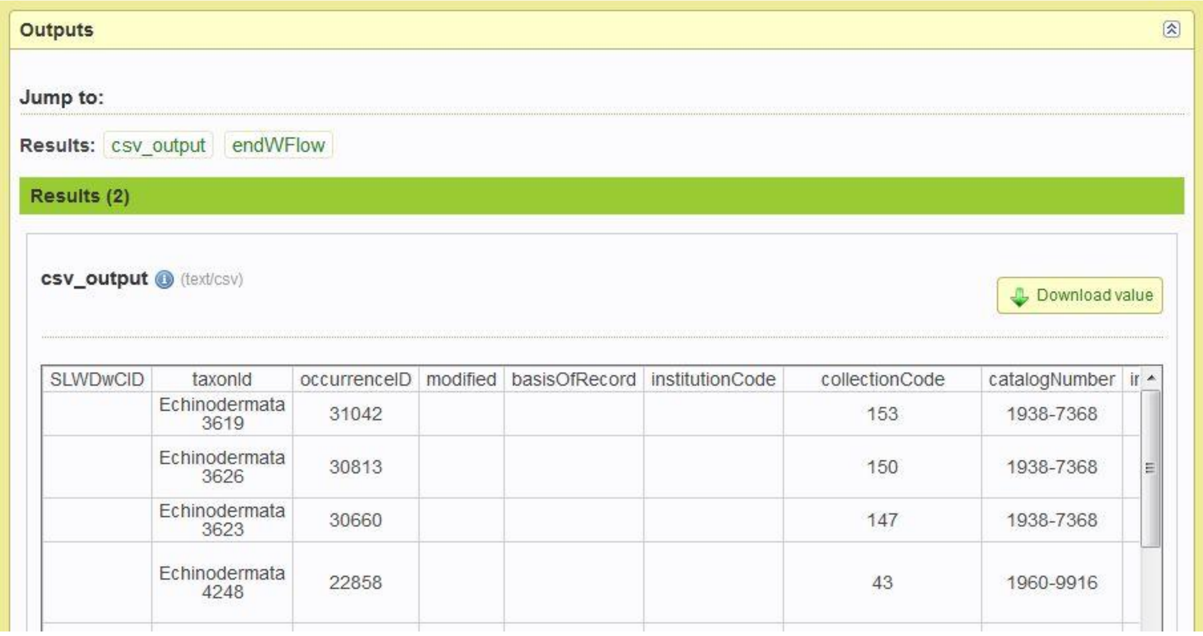
- ☑ Crassostrea gigas (Thunberg, 1793)

Synonym

- ☑ Crassostrea talienwhanensis Crosse, 1862
- ☑ Dioeciostrea hispaniola Orton, 1928
- ☑ Lopha posjetica (Razin, 1934)
- ☑ Lopha posjetica beringi Vialov, 1946
- ☑ Lopha posjetica newelskyi Vialov, 1946
- ☑ Lopha posjetica zawoikoi Vialov, 1946

6. You will be presented then with a choice of species occurrence data banks to target. As in the taxonomic checklists, if the occurrence bank has usage conditions, you will have to estate your agreement with them before you can use that bank.
Choose the 'GBIF Occurrence Bank' and press 'OK'.
This will send a request to the GBIF occurrence web service and retrieve occurrences for all the names marked in the list.

7. When the retrieval is done, you will be sent back to the 'Choose Sub-Workflow' interaction dialog. Choose 'End Workflow'.
8. The summary page about this workflow run shows the results of your query in the 'Outputs' box. At the time of writing this manual, we obtained more than 1600 occurrence records for that species.
9. You can download the dataset by clicking on the 'Download value' button on the right of the 'csv_output' entry in the 'Outputs' box. In more complex workflows, you will be able to access in the same way the products of all the different steps defined in your workflow.



Outputs

Jump to:

Results: csv_output endWFlow

Results (2)

csv_output (text/csv)

SLWDwCID	taxonId	occurrenceID	modified	basisOfRecord	institutionCode	collectionCode	catalogNumber	ir
	Echinodermata 3619	31042				153	1938-7368	
	Echinodermata 3626	30813				150	1938-7368	
	Echinodermata 3623	30660				147	1938-7368	
	Echinodermata 4248	22858				43	1960-9916	

10. You can also click on 'download all results' on the top of the page to obtain a zip file containing all the results together.

Exercises 01 and 02

Here you have some exercises that you can use to check your skills using what you have learnt with this tutorial. Please remember that the answers to all the exercises are provided in chapter 8.

EXERCISE 1: You receive an email from a fellow researcher working with earthworms, requesting your help in finding occurrence data about *Octolasion cyaneum* (Savigny, 1826). He does not give you much more information, so you decide to use the BioVeL portal to make sure that you download the maximum number of occurrences using all the known synonyms in at least two big catalogues.

Create an input file with this name, and find out: How many synonyms does this species have when using the GBIF Backbone Taxonomy? How about the Catalogue of Life?

EXERCISE 2: One of your friends, a botanist working on mycorrhizas (a symbiotic interaction between vascular plant roots and fungi) is travelling to the Iberian Peninsula and she is interested in the genus *Retama* Raf. She would like to take with her a list of all the occurrences for the two species occurring in the area: *Retama sphaerocarpa* and *Retama monosperma*. How would you proceed to generate such a dataset using the BioVeL portal? Make sure you make your query using all the known synonyms in the GBIF Backbone Taxonomy.

5.2. Tutorial 5: Data visualization and filtering using BioSTIF

In this tutorial we will take you through a workflow run from beginning to end (i.e., from launching the workflow to saving results). In this simple example, we will use only the BioSTIF application (Biodiversity Spatial Temporal Interactive interFace) to examine basic geospatial filtering over an example dataset. This tutorial can take from 10 to 15 minutes, depending on how much time you want to spend exploring the application interface.

BioSTIF provides visual filtering and selection of biodiversity data containing both geographical coordinates AND time information. Since the tool presents itself as a web service, it is easily integrated in workflows that require such functionality. Learn more about BioSTIF's features on the BioVeL Wiki.

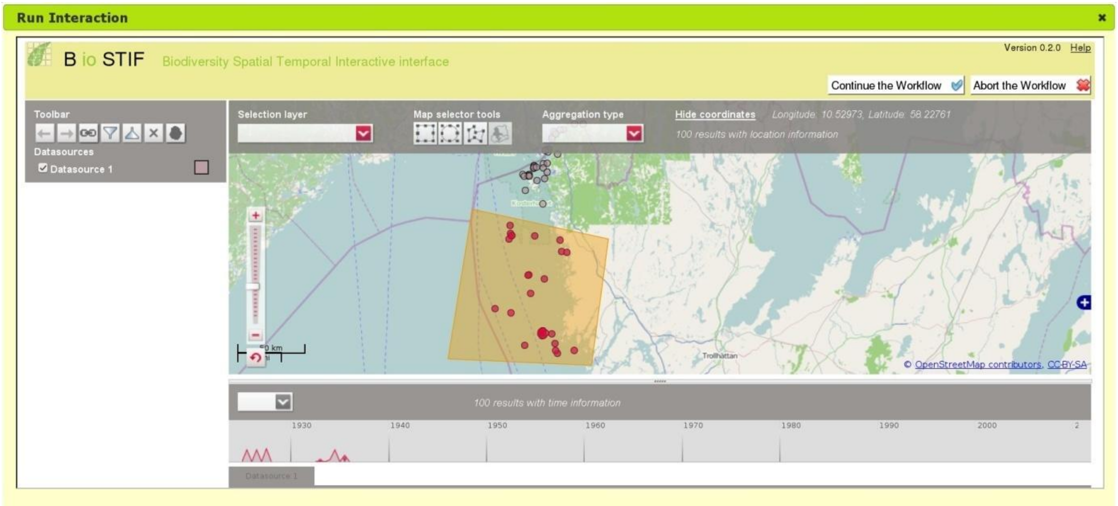
If you have not done already, download the Exercise File 1: BioVeL-GBIF_BPG_File1_105OR.csv and save it in your computer. We will use this example of a dataset of marine animal occurrences in CSV format (Comma Separated Values). It contains a heterogeneous group of data records coming from different projects working on Sweden. Each record includes taxonomic and geographical information, time and depth of collection, and habitat types.

IMPORTANT NOTE: When using your own data, make sure that the file is encoded in the UTF-8 format to avoid encoding problems. You can usually select a spreadsheet's encoding format in your application's 'save as' dialogue. It is best to avoid blank spaces in your file names.

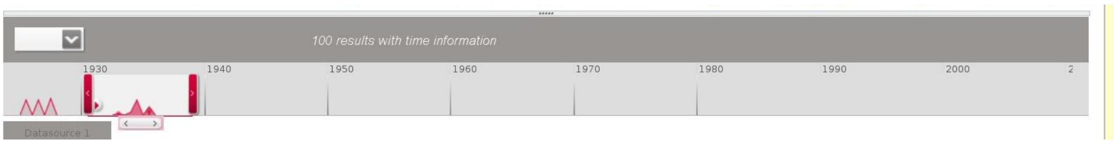
Using the following simple workflow, we can extract subsets of a given dataset using geographical or temporal criteria. This type of selections can be a task by itself, or it can be one of the initial steps before applying other data refinement techniques.

1. Create a 'Data Refinement Workflow' run as explained in Tutorial 2.
2. On the 'Choose Input File' interaction dialog, click the 'Browse' button and select the example data file. Click on 'Submit'. The data will be then uploaded to the portal.
3. Once the upload is finished, a new dialogue titled 'Choose Sub-Workflow ' will appear. Please choose 'Data Selection (BioSTIF)' and click on 'OK'. Please be aware that the BioSTIF application may take some time to load, especially if you are loading a large number of records (>50.000 records).
4. Use some time to get familiar with the interface. On the left column you will see a toolbar with buttons that perform operations related to the filters. The wider column on the right shows a map, a time series and a table showing records selected at a given moment.
5. Now it is time to apply some filters to the data:

- To define a geographical filter: Choose one of the 'Map selector tools' available in the top of the map: 'Square selection', 'Circle selection' or 'Polygon selection'. Draw the figure by clicking on the map. Double click to seal the polygon in the case of the free-form polygon.



- To define a temporal selection you can use the timeline below the map, by double clicking on the timeline and dragging the period range.



6. After each selection, a new set of buttons will appear in the Toolbar area on the left-hand column. You can use them to include or exclude the selected records from the final dataset, discard the selection or save it as a URL. You can also undo the last filter operation by clicking on the left-pointing arrow.
7. To help you with the selection, you can show different layers in the base map by using the 'selection layer' dropdown menu on the top right of the map.
8. Finish the selection process by clicking the 'Continue the workflow' button at the top-right. That will save your results and return you to the 'Choose Sub-Workflow' dialogue window.
9. Select 'End workflow' in the sub-workflow selection list. That will take you to a summary page about this workflow run. You can browse here the results of your selection in the 'Outputs' box. You can download the subset of records that you selected by clicking on the 'Download value' button on the right of the 'csv_output' entry in the 'Outputs' box.

Exercises 03 and 04

We suggest you some exercises here, to check you new skills using the BioSTIF application. Please use the same example dataset used in Tutorial 5 (Exercise File 1: BioVeL-GBIF_BPG_File1_105OR.csv) and remember that the answer to all the exercises are provided in chapter 8.

EXERCISE 03: A fellow researcher from your institution is writing an article about biological expeditions in the early 20th century and would like to know how many entries you have in your database recorded before 1940. She would like to evaluate whether the volume of georeferenced data is sufficient to generate a map to illustrate the article. Can you help her?

EXERCISE 04: A regional government is interested to know how many records you have that are located north of Strömstad. They are interested in obtaining a copy of those records, if possible.

5.3. Tutorial 6: Basic data filtering and cleaning using OpenRefine

OpenRefine (http://openrefine.org/) is a data analysis tool originally developed by Google (under the name of Google Refine), now transformed into an open-source project whose development, documentation and promotion are fully supported by volunteers.

The BioVeL portal integrates OpenRefine as part of the data refinement workflow, so researchers can use this powerful tool to define data quality assessments and transformation as part of their workflows. It includes a tailor-made BioVeL extension that allows remote data access (e.g., to the GBIF name checklists) while providing a relatively easy-to-use interface.

Nearly all the procedures explained in this manual can be performed directly in OpenRefine, if you wish to do so.

This tutorial will take around 15-20 minutes to complete. We will get familiar with the basic operations within OpenRefine. Some of those operations are:

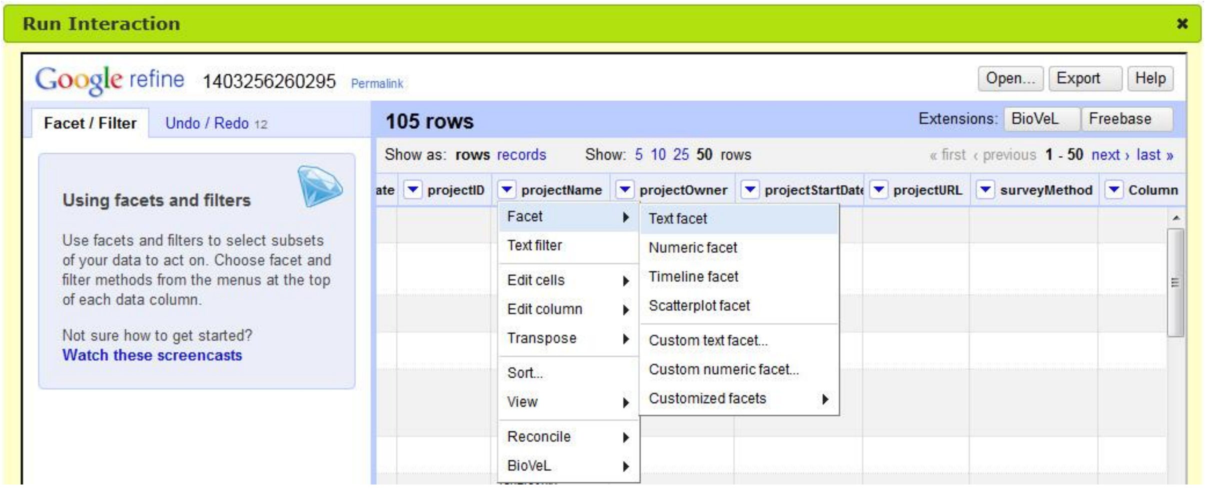
1. Extract metrics from the contents of a dataset
2. Find misspellings, duplicates and incorrect, inconsistent or missing data values
3. Correct errors in multiple records in one operation
4. Extract subsets of data according to different criteria
5. Reuse data filtering and transformation operations to apply them on similar data or to share them with other users

OpenRefine is based on the use of facets and filters that aggregate and present the contents of a dataset. In the OpenRefine wiki you can read this about facets: "*Typically, you create a facet on a particular column. The facet summarizes the cells in that column to give you a big picture on that column, and allows you to filter to some subset of rows for which their cells in that column satisfy some constraint*".

Be aware that all the modifications that we do inside OpenRefine apply only to the copy that OpenRefine is working with, so your original data remains unmodified.

But let's see the tool functionality through an example.

1. Create a workflow run using the Exercise File 1: BioVeL-GBIF_BPG_File1_105OR.csv. Stop when you are prompted to select a sub-workflow. Choose 'Data Quality (Google Refine)' this time and click on 'OK'.
2. We will start with a simple exercise to **extract metrics** using facets to analyse how many projects are represented in our dataset and the proportion of records in each project. Find the 'projectName' field and use the arrow to the left of its name to select 'Facet' and then 'Text facet'.

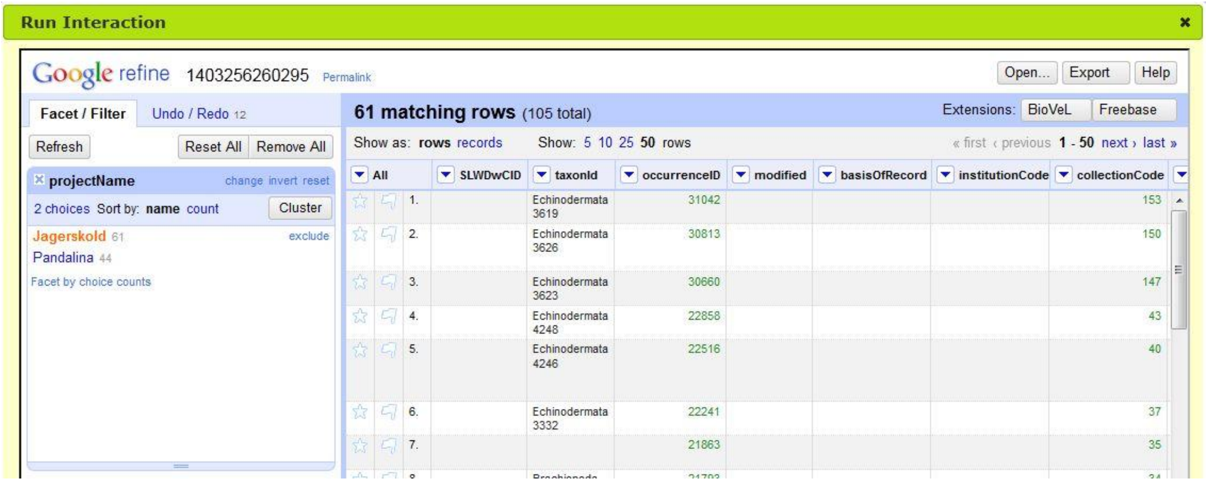


Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
Watch these screencasts

3. A new box appears in the left-hand column named after the field with a list of the different options available in that field (two in this case). We can see that there are 61 elements in the project named 'Jagerskold' and 44 in 'Pandalina'. If you click in one of the project names the view on the right only shows the records that match with that project. This is an easy way to **filter records**. Select 'Jagerskold' as we will use only the records from that project. You can deactivate, modify or invert these filters by clicking on the corresponding links on the right of the facet name.



Jagerskold 61
Pandalina 44

61 matching rows (105 total)

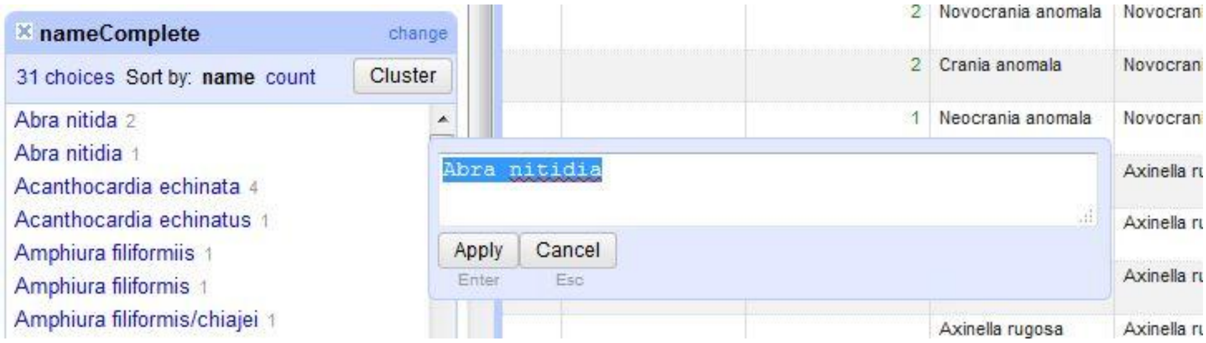
All	SLWDwCID	taxonId	occurrenceID	modified	basisOfRecord	institutionCode	collectionCode
1.		Echinodermata 3619	31042				153
2.		Echinodermata 3626	30813				150
3.		Echinodermata 3623	30660				147
4.		Echinodermata 4248	22858				43
5.		Echinodermata 4246	22516				40
6.		Echinodermata 3332	22241				37
7.			21863				35

4. Let's add another facet to the analysis. Go to the field 'nameComplete' and create another text facet based on this field. We can see that there are many more options in this facet: up to 31 different names for that project. If you browse through them you will probably immediately spot some misspellings and invalid entries. Click on one of them, and you will see which records are assigned to that scientific name for the

<u>project selected in the previously defined facet</u>: that means that selections made through facets add to each other. In this way you can create **more complex filters**.

5. We will start now with simple data cleaning operations. As you can see there are two spelling variations for the species epithet of *Abra nitida* (O. F. Mueller, 1776): 'nitida' and 'nitidia'. Let's assume that this is part of a correction exercise done right after data capture and that we can safely **change the record** (in later tutorials we will do a more correct procedure of data cleaning where we will capture the original information and document the changes made).

   When you put the mouse pointer over the name, an option to 'edit' appears on the right. If you click on it, you have the option to edit the name. When you click on 'Apply', <u>ALL the instances</u> of that name will be changed (although in our example we only have one occurrence of that name).

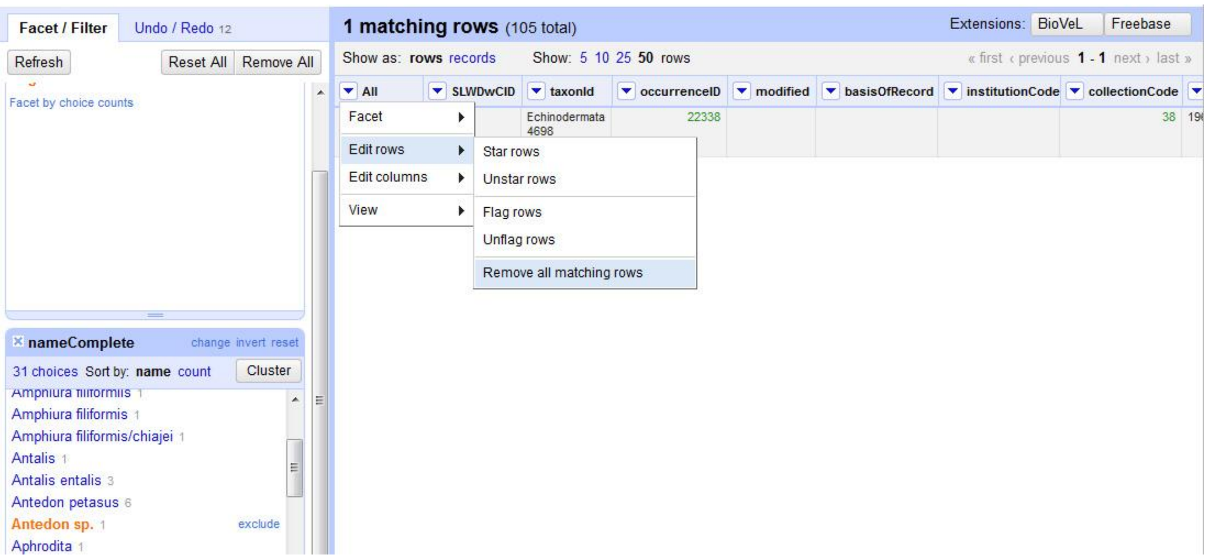


Abra nitida 2
Abra nitidia 1
Acanthocardia echinata 4
Acanthocardia echinatus 1
Amphiura filiformiis 1
Amphiura filiformis 1
Amphiura filiformis/chiajei 1

Abra nitidia

2	Novocrania anomala	Novocran
2	Crania anomala	Novocran
1	Neocrania anomala	Novocran
		Axinella ru
		Axinella ru
		Axinella ru
	Axinella rugosa	Axinella ru

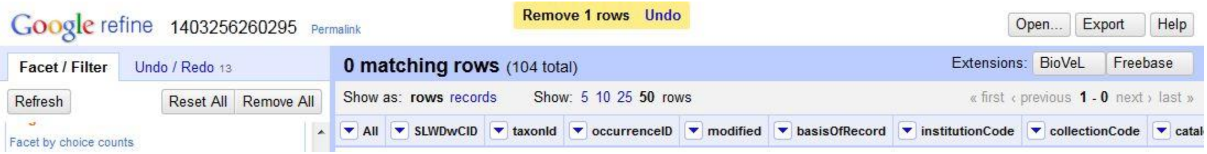
6. In other cases, you may want to **remove records** that do not comply with one of your data quality criteria. Again, let's assume that you want only records of organisms identified at least at the species level. Several records are identified at only the genus level, one of them is ' *Antedon* sp.'. Click on the name and you will see only one record displayed in the list of the right. Go to the first column in that list, labelled as 'All'. Select 'Edit rows' and then 'Remove all matching rows'. Those records are now excluded from the dataset loaded in OpenRefine. Remember (as mentioned above) that OpenRefine does not modify the original dataset that you uploaded to the BioVeL portal.



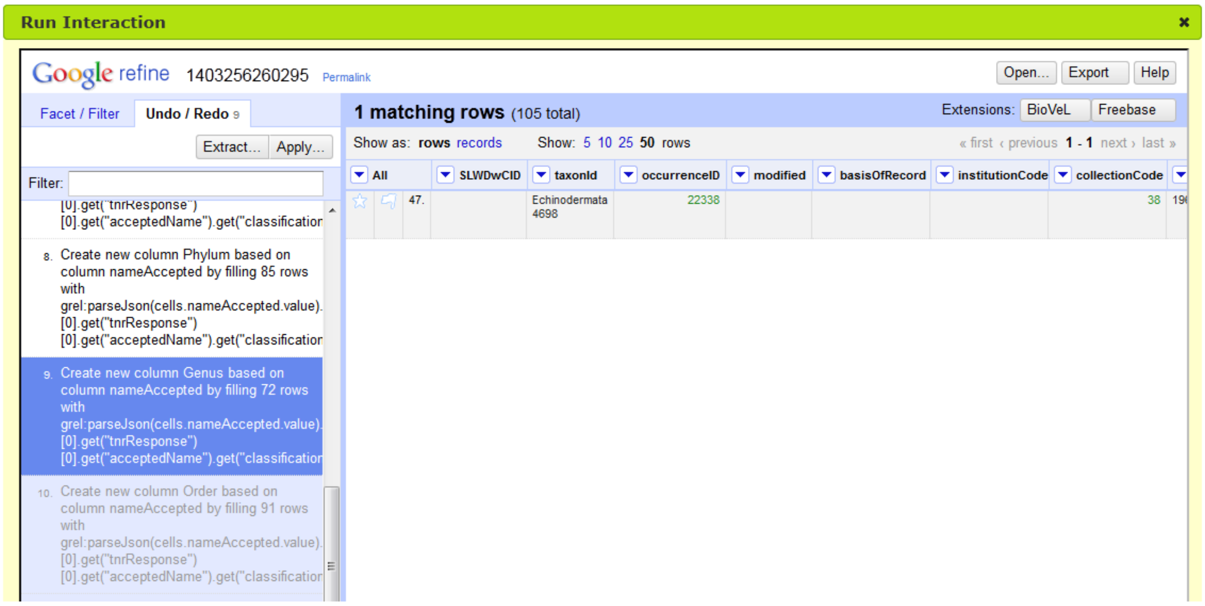
Echinodermata 4698 | 22338 | 38 | 19

Amphiura filiformiis 1
Amphiura filiformis 1
Amphiura filiformis/chiajei 1
Antalis 1
Antalis entalis 3
Antedon petasus 6
Antedon sp. 1
Aphrodita 1

7. One of the nice features of OpenRefine is that you can undo every action, even a deletion such as the one you just did. The quickest way to do so is to use the 'Undo' link that appears on the top of the OpenRefine window in a small yellow box after performing an action:


Google refine 1403256260295 Permalink
Remove 1 rows Undo
Open... Export Help
Facet / Filter
Undo / Redo 13
Refresh
Reset All Remove All
Facet by choice counts
0 matching rows (104 total)
Extensions: BioVeL Freebase
Show as: rows records Show: 5 10 25 50 rows
« first ‹ previous 1 - 0 next › last »
All SLWDwCID taxonId occurrenceID modified basisOfRecord institutionCode collectionCode catal

Otherwise, there is a second tab that we have still not explored in the left-hand column named 'Undo/Redo'. Clicking on one of the entries of the history of changes will restore the data to its state when that operation was performed.


Run Interaction
Google refine 1403256260295 Permalink
Open... Export Help
Facet / Filter
Undo / Redo 9
Extract... Apply...
Filter:
1 matching rows (105 total)
Extensions: BioVeL Freebase
Show as: rows records Show: 5 10 25 50 rows
« first ‹ previous 1 - 1 next › last »
All SLWDwCID taxonId occurrenceID modified basisOfRecord institutionCode collectionCode
47. Echinodermata 4698 22338 38 19

[0].get("tnrResponse")
[0].get("acceptedName").get("classification

8. Create new column Phylum based on column nameAccepted by filling 85 rows with grel:parseJson(cells.nameAccepted.value).[0].get("tnrResponse")[0].get("acceptedName").get("classification

9. Create new column Genus based on column nameAccepted by filling 72 rows with grel:parseJson(cells.nameAccepted.value).[0].get("tnrResponse")[0].get("acceptedName").get("classification

10. Create new column Order based on column nameAccepted by filling 91 rows with grel:parseJson(cells.nameAccepted.value).[0].get("tnrResponse")[0].get("acceptedName").get("classification

Using the 'Extract...' button will give you the opportunity to obtain some lines of code describing the operation(s) performed. That code can be reused through the 'Apply' button (e.g., for repeating the same operation on the same dataset later, or performing it in another dataset with the same structure).

8. That completes the OpenRefine operations we will perform in this tutorial. To send the resulting dataset to our workflow run in the BioVeL portal, we have to use the button labelled 'BioVeL' on the top-right part of the window, in the area labelled 'Extensions'. Click on the button, and select 'Save updates and return to the workflow'. That will close OpenRefine and send us back to the 'Choose Sub-Workflow' dialog.

9. You can now perform other operations on the dataset, or choose 'End Workflow' and finish the exercise here. As in the example included in Tutorial 5, you can download the results of your workflow run and use it in other contexts.

Exercises 05 to 08

Here you can find some exercises that will help you to practice the skills that you have just acquired using OpenRefine. Are you in? Please use the Exercise File 1: BioVeL-GBIF_BPG_File1_105OR.csv and remember that the answer to all the exercises are provided in chapter 8.

EXERCISE 05: As part of the data cleaning protocol that you are following, you need to register some metrics about the original dataset before making any modifications to it. Can you find out how many different species names does the data file contain?

EXERCISE 06: A fellow researcher has asked you to share all the records you have in your dataset about *Abra nitida* as a CSV file. Can you generate such a file using OpenRefine? How many rows does the exported file contain?

EXERCISE 07: Prior to publishing the dataset on the Internet you have been asked to provide some metadata describing its contents. Can you find the depth range of the records included in your dataset?

EXERCISE 08: You are part of a project with several partners working in different parts of the country collecting data using the same spreadsheet template. You would like to provide your project colleagues with OpenRefine operations that remove all records that do not have both geographic coordinates from their datasets. Which operations would you perform? What would you send to your colleagues?

6. The taxonomic data refinement workflow: advanced use

In this chapter we will learn richer data quality assessment and cleaning methods. As the methods become more complex, we will also specialize more and more: some of the procedures included in this chapter are specifically designed for specific domains that are commonly part of occurrence data (e.g., taxonomy and nomenclature, geospatial information). Chapter 7 will provide guidance if you wish to know more about applying these techniques to other domains.

6.1. Tutorial 7: Cleaning and refining data, spelling errors

This tutorial focuses on correcting spelling errors by clustering similar name strings in OpenRefine. It will take around 20 minutes to complete.

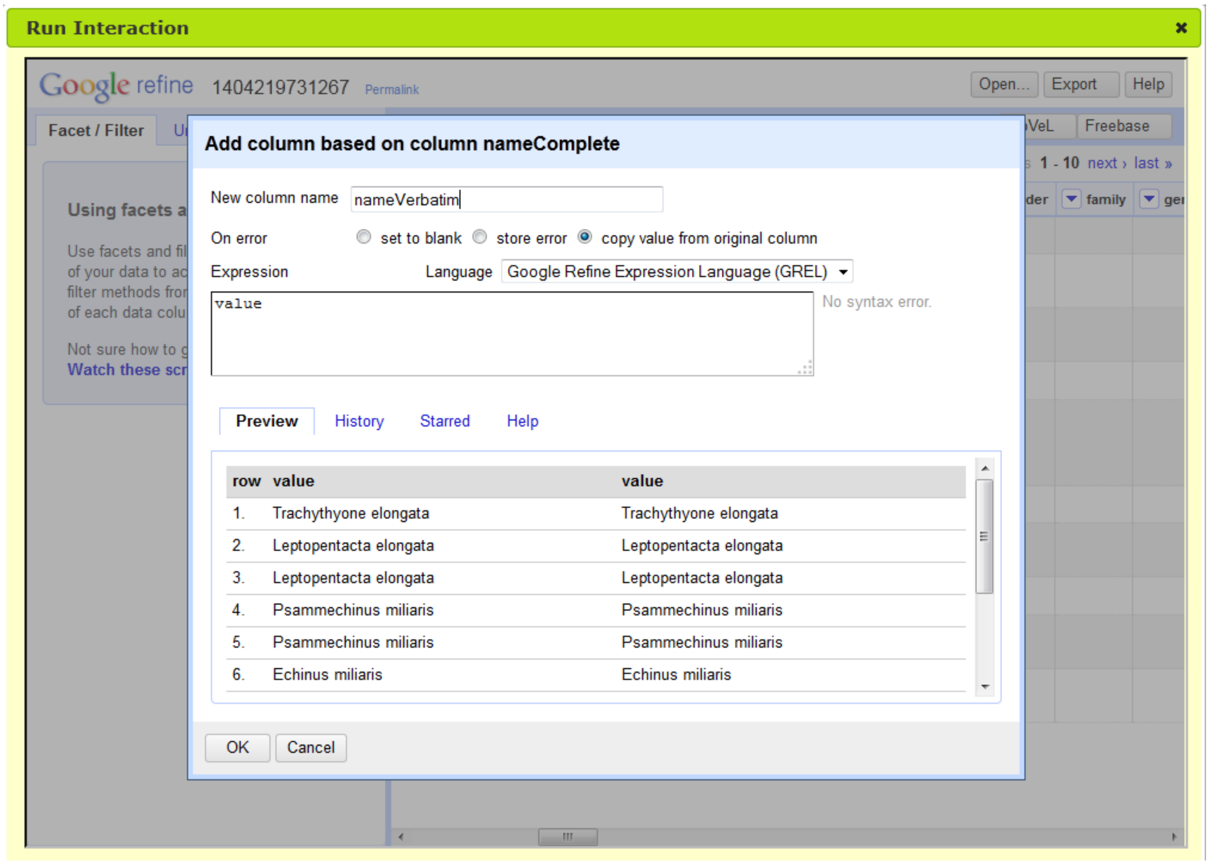
According to OpenRefine documentation, clustering refers to the operation of '*finding groups of different values that might be alternative representations of the same thing*'. Different methods can produce clusters with different execution times. The examples provided here are presented from the quickest to the slowest. If you are working with a large dataset, you may want to test the methods in order.

Clustering is a very effective way to detect and correct errors in taxonomic information, but it can be applied to any other field whose contents are restricted to a limited number of correct options.

While in some cases it is obvious which value is correct, to perform corrections properly it is important to have the appropriate reference list of valid values. Such lists could be taxonomic checklists, controlled vocabularies for certain fields, geographical gazetteers for the area studied, etc.

Let's get started:

1. Start a workflow run and load the Exercise File 1: BioVeL-GBIF_BPG_File1_105OR.csv data file. In the sub-workflow selection dialog, choose 'Data Quality (Google Refine)'.
2. Create a duplicate of the 'nameComplete' field to keep the original data and name it 'nameVerbatim'. To do that, use the arrow close to the column name and select 'Edit column' and 'Add column based on this column'. A dialogue window will open where we can define both the name of the new column as well as any transformation the data requires before copying it to the new column. Since we want a literal copy, we only need to provide the name for the new column, and may select the option 'copy value from original column' in case any errors occur while copying.



Add column based on column nameComplete

New column name: nameVerbatim

On error: ○ set to blank ○ store error ● copy value from original column

Expression — Language: Google Refine Expression Language (GREL)

```
value
```

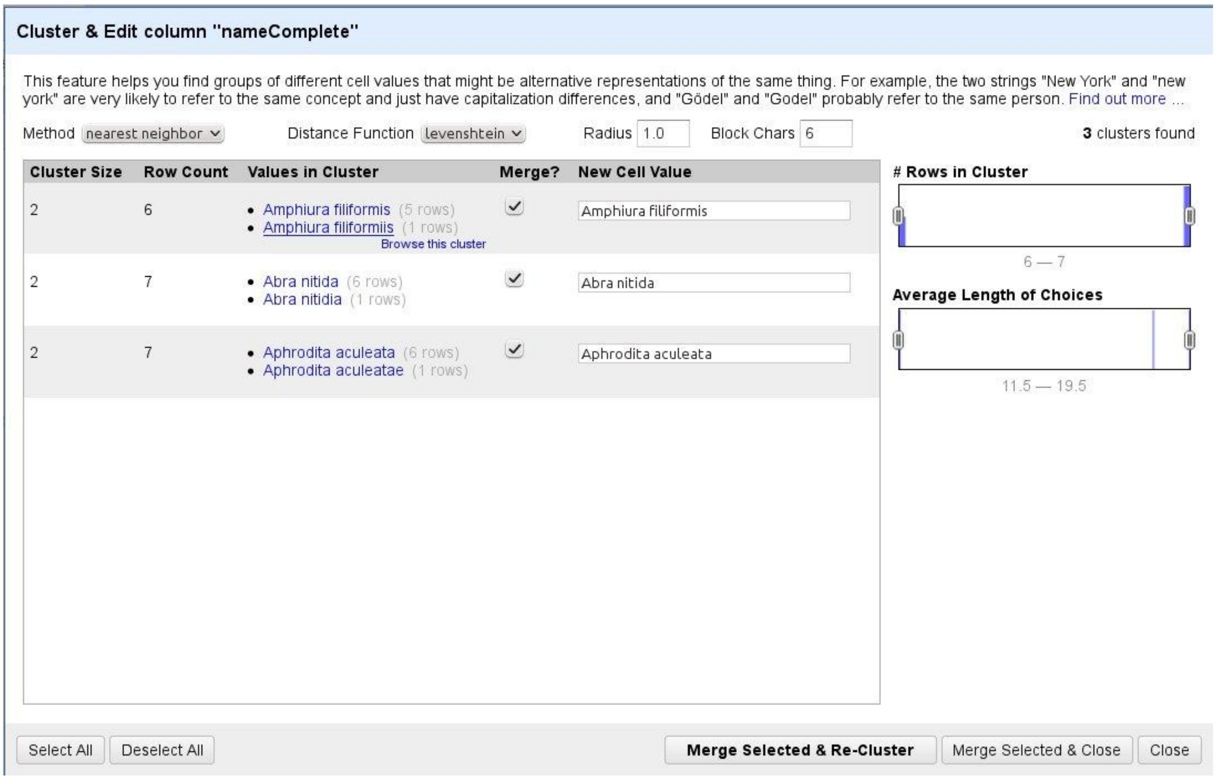
No syntax error.

Preview | History | Starred | Help

row	value	value
1.	Trachythyone elongata	Trachythyone elongata
2.	Leptopentacta elongata	Leptopentacta elongata
3.	Leptopentacta elongata	Leptopentacta elongata
4.	Psammechinus miliaris	Psammechinus miliaris
5.	Psammechinus miliaris	Psammechinus miliaris
6.	Echinus miliaris	Echinus miliaris

OK | Cancel

3. Create a new field (if it does not exist already) called 'taxonRemarks' to store information about the changes performed on the records. There should be columns on the right end of the dataset named 'column' + a number that you can rename and use. Just use the arrow close to its name and choose 'Edit column' and 'Rename this column'.
4. Create a text facet based on the field 'nameComplete'.
5. In the facet box, you will find a button labelled 'Cluster'. Click on it and choose the 'nearest neighbour 'clustering method from the 'Method' drop down.



Cluster & Edit column "nameComplete"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

Method nearest neighbor Distance Function levenshtein Radius 1.0 Block Chars 6 **3** clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	6	• Amphiura filiformis (5 rows) • Amphiura filiformiis (1 rows) Browse this cluster	✓	Amphiura filiformis
2	7	• Abra nitida (6 rows) • Abra nitidia (1 rows)	✓	Abra nitida
2	7	• Aphrodita aculeata (6 rows) • Aphrodita aculeatae (1 rows)	✓	Aphrodita aculeata

This instructs OpenRefine to try to group alternatives that could be referring to the same entity. We are offered now the option to substitute the less frequent values by the most frequent ones. For that we need to select the checkbox in the 'Merge?' column. You should make sure that the suggestion listed for each case in 'New Cell Value' column is the correct assumption. Otherwise you can define which one is the final correct value in the same text box.

6. In case you need to make individual changes or describe the changes you are about to perform (i.e., in the 'taxonRemarks' field that we just created), you can use the link 'browse this cluster' that appears when you hover your mouse pointer over the elements in the 'Values in Cluster' column. This command will load the records into the data-editing interface where you can make any modifications needed.
7. When you are done defining the changes that you would like to perform automatically, click on 'Merge Selected & Re-cluster'. After making the correction, double-check whether there are other clusters that need correcting. Click 'Close' when you are done.
8. You can now use the 'BioVeL' button on the top right to 'Save updates and return to workflow', and continue your workflow with other operations or finish it and download your results, as in previous tutorials.

Exercise 09

Here you can find an extra exercise for you to try your new skills. Please use the Exercise File 1: BioVeL-GBIF_BPG_File1_105OR.csv and remember that the answer to all the exercises are provided in chapter 8.

EXERCISE 09: The communications department of your research institute want to create a poster about the field expeditions by researchers from the institute. They ask for your help to clean the data related to the collectors from the example dataset, so they can plot the data into a map based on the collector teams. Can you use the OpenRefine clustering function to clean that information? What is the final list of collectors and the number of collection events associated to them? Note: due to data aggregation, the surname of one of the researchers (Jägerskiöld) is wrong in the database—make sure that you correct that, too.

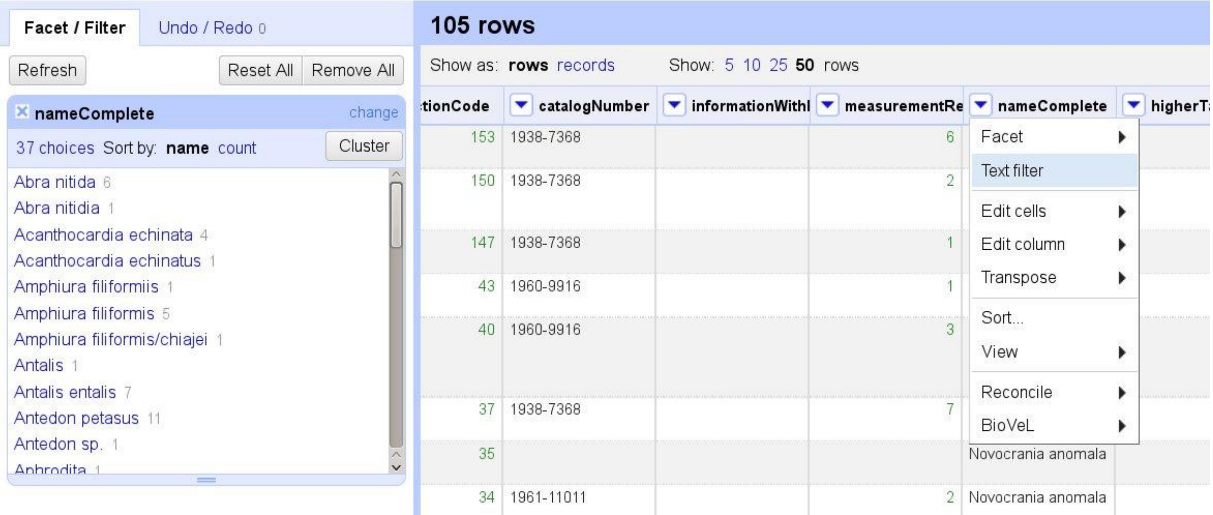
6.2. Tutorial 8: Filtering records by a list of elements

This tutorial will show how to filter records very easily using a list of criteria at the same time. You can use this technique to identify elements that should not be in the dataset. It should take no more than five minutes to complete if you have gone through the earlier OpenRefine tutorials.

In our example we are going to illustrate one of the common situations in taxonomic databases, which we call 'domain schizophrenia': when data fields are used for a different purpose than what they were created for. This is one of the cases where you can easily detect which are the most common invalid insertions and create a rule that you can re-use all the time.

In our example, we will mark (or delete) those records with species names without epithet (e.g., *Novocrania*, *Astarte*, etc), or with abbreviations (*Novocrania* cf. *anomala*, *Astarte* sp.).

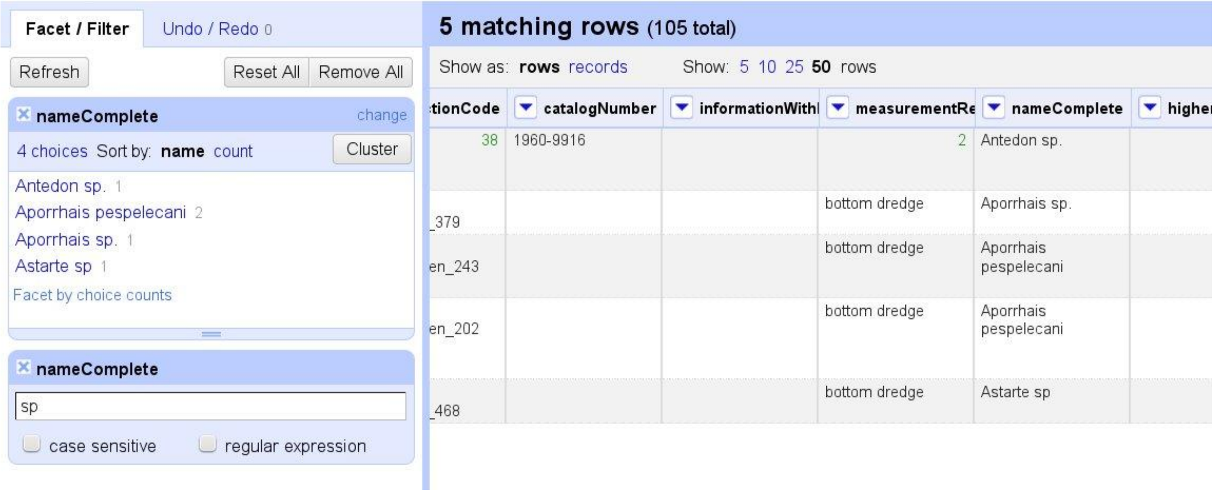
1. Start a workflow run and load the Exercise File 1: BioVeL-GBIF_BPG_File1_105OR.csv data file. In the sub-workflow selection dialog, choose 'Data Quality (Google Refine)'.
2. Find the column associated to 'nameComplete' and use the arrow close to the name to choose 'Text filter'.



name	count
Abra nitida	6
Abra nitidia	1
Acanthocardia echinata	4
Acanthocardia echinatus	1
Amphiura filiformiis	1
Amphiura filiformis	5
Amphiura filiformis/chiajei	1
Antalis	1
Antalis entalis	7
Antedon petasus	11
Antedon sp.	1
Aphrodita	1

tionCode	catalogNumber	informationWithl	measurementRe	nameComplete	higherT
153	1938-7368		6		
150	1938-7368		2		
147	1938-7368		1		
43	1960-9916		1		
40	1960-9916		3		
37	1938-7368		7		
35				Novocrania anomala	
34	1961-11011		2	Novocrania anomala	

3. Insert 'cf', '/', 'sp.', 'sp' in the text box to filter out all occurrences defined at the genus level or with an ambiguous taxonomic assignment.



nameComplete change

4 choices Sort by: name count Cluster

- Antedon sp. 1
- Aporrhais pespelecani 2
- Aporrhais sp. 1
- Astarte sp 1

Facet by choice counts

nameComplete

sp

□ case sensitive □ regular expression

5 matching rows (105 total)

Show as: rows records Show: 5 10 25 50 rows

tionCode	catalogNumber	informationWithl	measurementRe	nameComplete	highe
38	1960-9916		2	Antedon sp.	
_379			bottom dredge	Aporrhais sp.	
en_243			bottom dredge	Aporrhais pespelecani	
en_202			bottom dredge	Aporrhais pespelecani	
_468			bottom dredge	Astarte sp	

4. Now you may choose to mark those records or remove them from this copy of the dataset (following the steps described in Tutorial 6). Always remember to keep a copy of the original or 'verbatim' data and to describe the outcome of all your operations related to data quality.

Exercise 10

This new way of using filters open new possibilities for data checking. Are you up for some extra practice? Please use the Exercise File 1: BioVeL-GBIF_BPG_File1_105OR.csv and remember that the answer to all the exercises are provided in chapter 8.

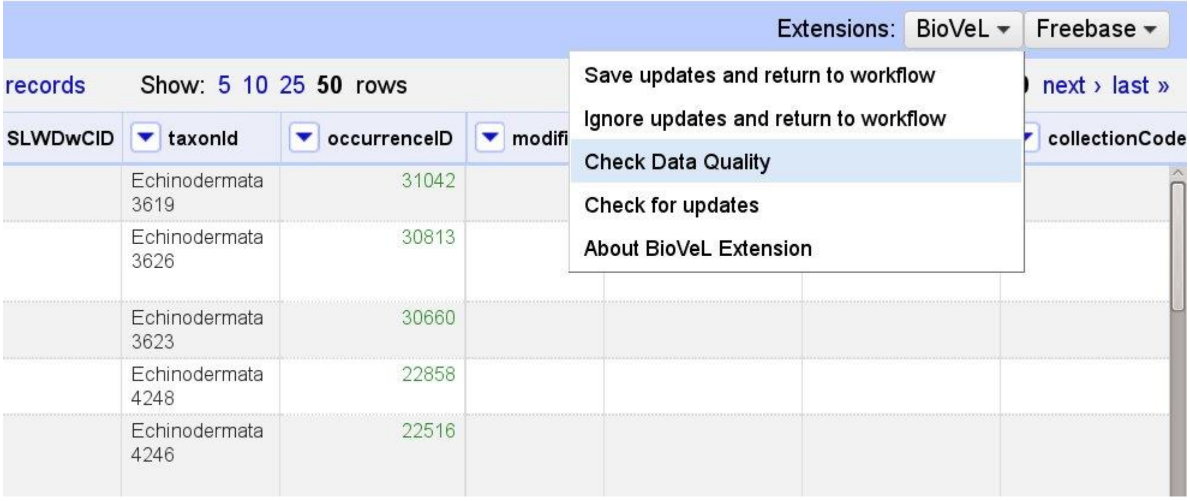
EXERCISE 10: A construction firm planning to build new infrastructures along the coast of Sweden has hired you to do some preparatory work for the necessary environmental impact assessment. The first step is to produce a checklist for the species present in the area. They send you a dataset produced in previous projects. Use OpenRefine clustering to produce the cleanest possible list of species. How many entries would have that list? Note: Consider species with 'cf' reference (= identification not confirmed) as correctly identified.

6.3. Tutorial 9: Seeking incomplete coordinate pairs

This tutorial will make use of a specific functionality of the BioVeL OpenRefine extension to filter out latitude/longitude records with incorrect or missing geographic information. It should take around five minutes to complete.

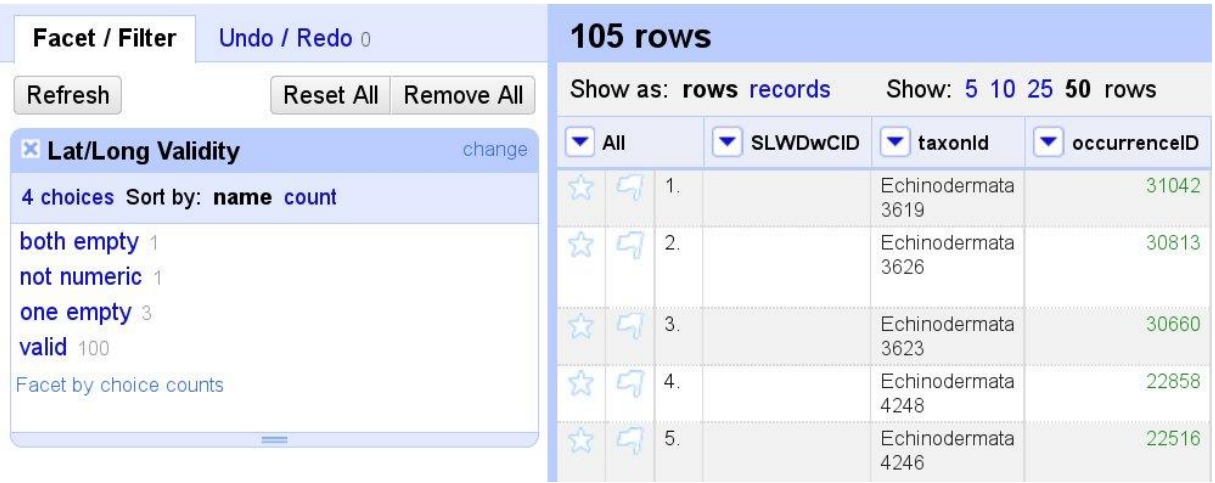
Taxonomic datasets seldom have complete geographic references for all observations in the data file. Typically, there will be some records where the latitude and/or longitude information is missing, or they are non-numeric. Here we show you how to mark or remove such records from your data file.

1. Start a workflow run and load the Exercise File 1: BioVeL-GBIF_BPG_File1_105OR.csv data file. In the sub-workflow selection dialog, choose 'Data Quality (Google Refine)'.
2. Click on the BioVeL button on the top-right and select 'Check Data Quality'.


Extensions: BioVeL ▾ Freebase ▾
records Show: 5 10 25 50 rows
next › last »
Save updates and return to workflow
Ignore updates and return to workflow
Check Data Quality
Check for updates
About BioVeL Extension

SLWDwCID	taxonId	occurrenceID	modifi	collectionCode
	Echinodermata 3619	31042		
	Echinodermata 3626	30813		
	Echinodermata 3623	30660		
	Echinodermata 4248	22858		
	Echinodermata 4246	22516		

3. Choose the 'Latitude / Longitude Check' in the popup dialog.
4. When the calculation is finished, we will see another way to identify potentially problematic records. The options found are presented as a box in the 'Facet / Filter' column of the left.



Lat/Long Validity	
both empty	1
not numeric	1
one empty	3
valid	100

All	SLWDwCID	taxonId	occurrenceID
1.		Echinodermata 3619	31042
2.		Echinodermata 3626	30813
3.		Echinodermata 3623	30660
4.		Echinodermata 4248	22858
5.		Echinodermata 4246	22516

5. Now you may choose whether you want to mark/edit those records (e.g., with a note in the 'georeferenceRemarks' field) or to remove them altogether from this copy of the dataset (following the steps described in Tutorial 6). Always remember to follow the recommendations listed in section 2.4 of this manual when making modifications to your original dataset.

Exercise 11

Please use the Exercise File 1: BioVeL-GBIF_BPG_File1_105OR.csv and remember that the answer to all the exercises are provided in chapter 8.

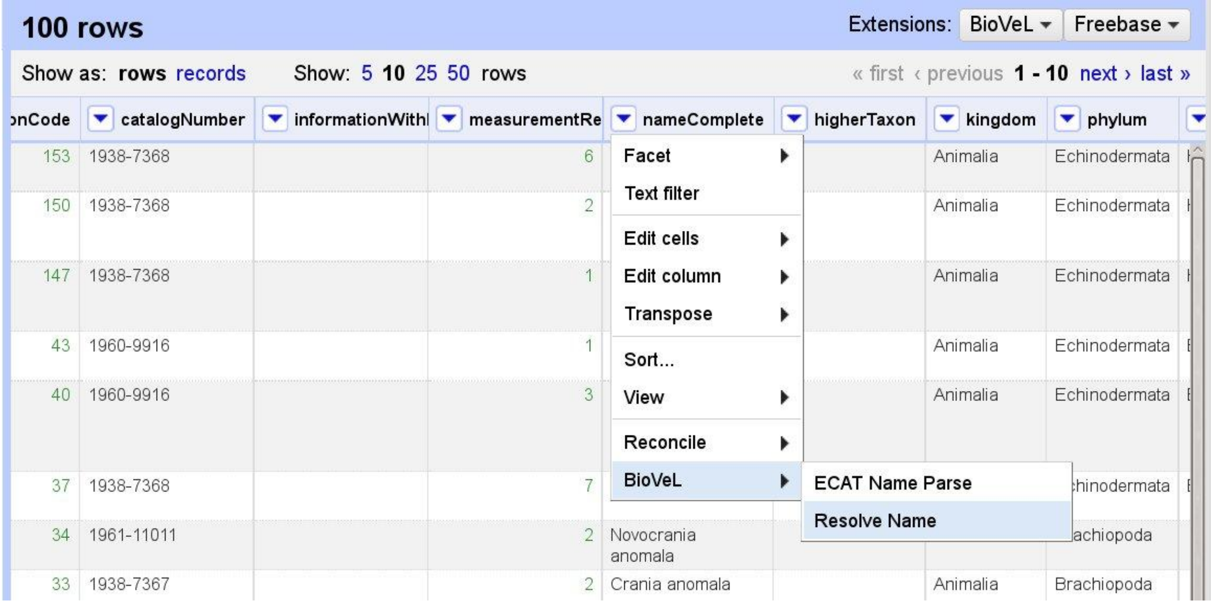
EXERCISE 11: You are part of a research group working on very sensitive environments, so you need to restrict your field trips to the area to a minimum: every record already captured counts! You are asked to check the occurrence dataset and try to fix any existing issues with coordinates so you can use as many records as possible in your study. Use the 'Latitude / Longitude Check' feature of the BioVeL extension and review your alternatives for fixing any detected issues.

6.4. Tutorial 10: Completing missing taxonomic information

In this tutorial, we will learn how to use the BioVeL extension for OpenRefine to lookup names against a target name checklist and to retrieve missing and/or incorrect information regarding rank, classification, accepted name, etc. Depending on the number of records you want to work with and the speed of your network connection, it will take a minimum of 10 minutes to complete.

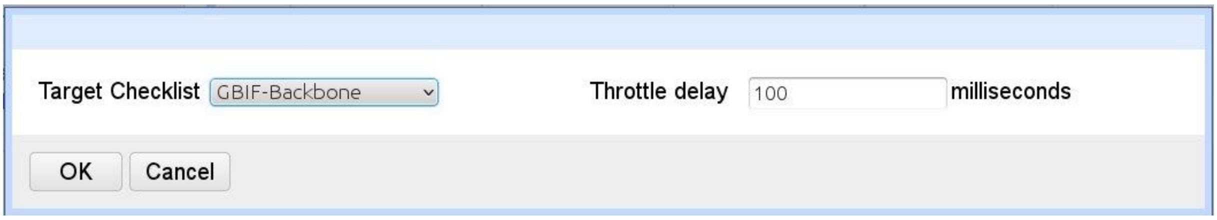
Data files often contain several name spellings for the same species, especially when the data are compiled from different sources (e.g., historical data from museums combined with your own data). Using the BioVeL extension for OpenRefine, we can make all our names to align with the accepted names in a selected checklist (in this example, the GBIF Backbone Taxonomy).

1. Start a workflow run and load the Exercise File 1: BioVeL-GBIF_BPG_File1_105OR.csv. In the sub-workflow selection dialog, choose 'Data Quality (Google Refine)'.
2. Once the OpenRefine interface is loaded, go to the 'nameComplete' column and, using the arrow to the left of the field name, select 'BioVeL' and 'Resolve Name'.



nCode	catalogNumber	informationWithl	measurementRe	nameComplete	higherTaxon	kingdom	phylum
153	1938-7368		6			Animalia	Echinodermata
150	1938-7368		2			Animalia	Echinodermata
147	1938-7368		1			Animalia	Echinodermata
43	1960-9916		1			Animalia	Echinodermata
40	1960-9916		3			Animalia	Echinodermata
37	1938-7368		7				hinodermata
34	1961-11011		2	Novocrania anomala			achiopoda
33	1938-7367		2	Crania anomala		Animalia	Brachiopoda

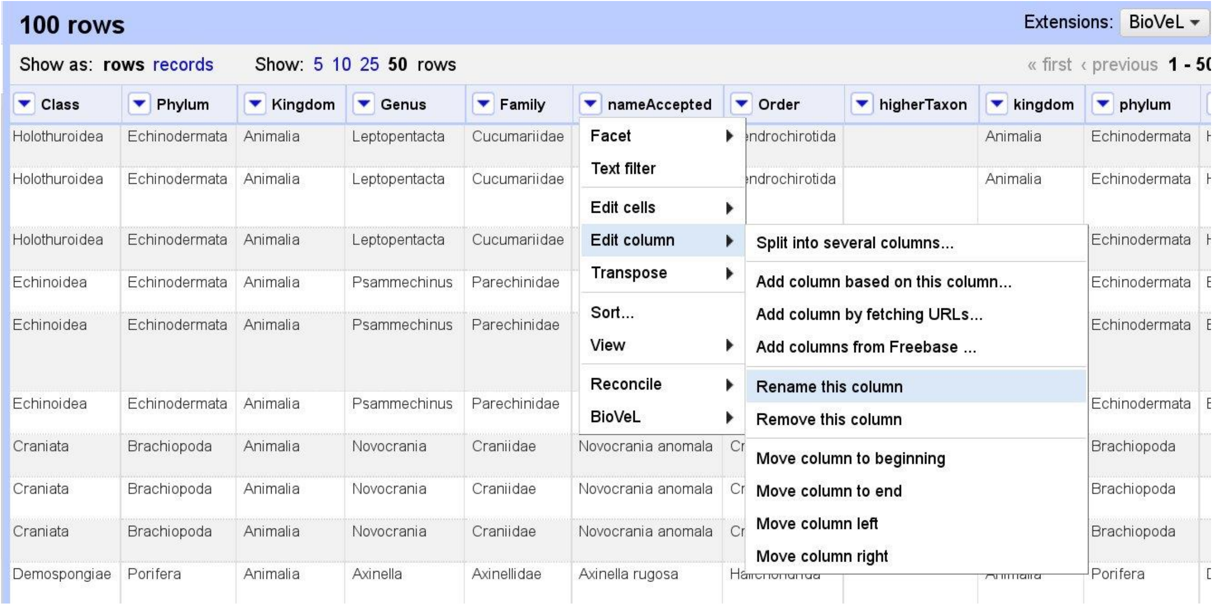
3. A pop-up window will appear that lets you choose the target checklist you wish to use to check your data against. Select 'GBIF-Backbone' and click 'OK'. Depending on the size of your dataset, you may have to wait for several minutes.


Target Checklist: GBIF-Backbone

Throttle delay: 100 milliseconds

4. Once the operation finishes, you’ll see new columns in the dataset: 'nameAccepted', and those related to taxonomic classification (Phylum, Genus, Family, etc), and to authorship.

5. If you apply a text facet on both name lists 'nameComplete' and 'nameAccepted', you will see how many changes were made in the database. You will find similar situations if your compare any of the taxonomic classification fields with their original values.
6. If you want to continue working with this dataset through the BioVeL portal, you may want to rename some of the fields, as the scripts expect certain data to be in certain fields. It would be a good idea to rename the original 'nameComplete' into something like 'nameVerbatim', and then rename the new 'nameAccepted' into 'nameComplete', as 'NameComplete' is a key field for the BioVeL portal.
   To rename a field/column in OpenRefine, go to the column and use the arrow on the title to select 'Edit column' and then 'Rename this column'.



Class	Phylum	Kingdom	Genus	Family	nameAccepted	Order	higherTaxon	kingdom	phylum
Holothuroidea	Echinodermata	Animalia	Leptopentacta	Cucumariidae		ndrochirotida		Animalia	Echinodermata
Holothuroidea	Echinodermata	Animalia	Leptopentacta	Cucumariidae		ndrochirotida		Animalia	Echinodermata
Holothuroidea	Echinodermata	Animalia	Leptopentacta	Cucumariidae					Echinodermata
Echinoidea	Echinodermata	Animalia	Psammechinus	Parechinidae					Echinodermata
Echinoidea	Echinodermata	Animalia	Psammechinus	Parechinidae					Echinodermata
Echinoidea	Echinodermata	Animalia	Psammechinus	Parechinidae					Echinodermata
Craniata	Brachiopoda	Animalia	Novocrania	Craniidae	Novocrania anomala	Cr			Brachiopoda
Craniata	Brachiopoda	Animalia	Novocrania	Craniidae	Novocrania anomala	Cr			Brachiopoda
Craniata	Brachiopoda	Animalia	Novocrania	Craniidae	Novocrania anomala	Cr			Brachiopoda
Demospongiae	Porifera	Animalia	Axinella	Axinellidae	Axinella rugosa	Halichondrida		Animalia	Porifera

7. Always remember to follow the recommendations listed in section 2.4 of this manual when making modifications to your original dataset.

Exercise 12

This feature of the BioVeL extension can be really useful when we have to work with heterogeneous datasets. We suggest you an additional exercise so you can practice a bit more. Show us what you have learned! Please use the Exercise File 1: BioVeL-GBIF_BPG_File1_105OR.csv and remember that the answer to all the exercises are provided in chapter 8.

EXERCISE 12: You work in a group within the Ministry of Environment, which is conducting research on a parasite endangering the populations of echinoderms in a neighbouring country because it has been recently observed close to your country's borders. To start your work, you are provided with the example occurrence dataset.

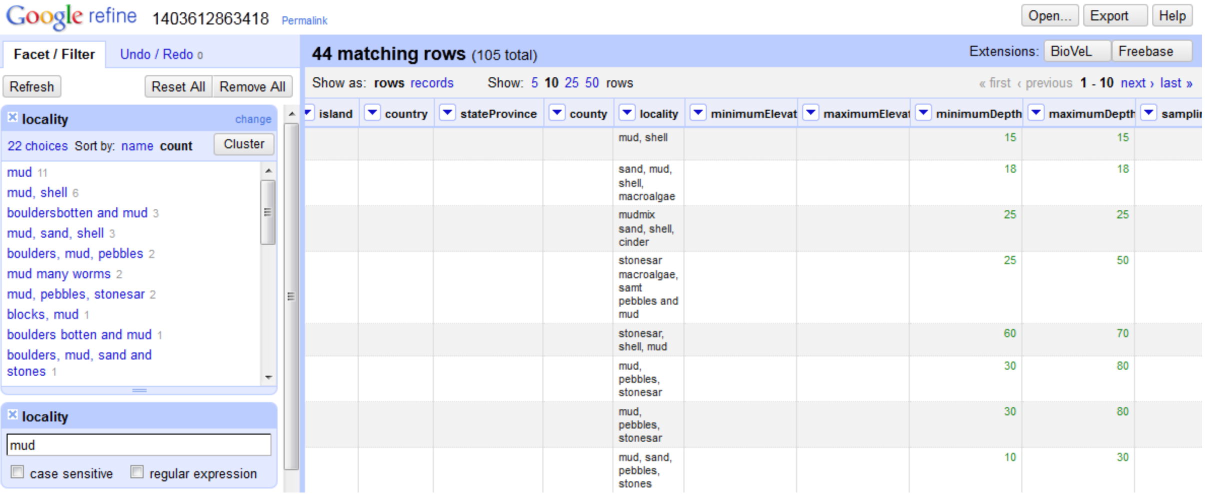
You are not an expert on echinoderms, so the first thing you would like to do is to obtain a clean list of the valid species names for echinoderms present in the area. Use the GBIF name resolution service through the BioVeL OpenRefine extension to refine the information in the dataset and try to prepare a list as clean as you can from it.

6.5. Tutorial 11: Filtering using descriptive data

Many biodiversity datasets contain descriptive information, like ecological information, habitat type, sampling methods, etc. OpenRefine allows you to use this information by selecting all records that share some of the elements (e.g., from shallow water, soft bottom habitats).

In this tutorial, you will learn how to select and filter data according to ecological criteria and generate ecologically comparable data sets. It will take around five minutes to complete if you have gone through some of the previous tutorials using OpenRefine.

1. Start a workflow run and load the Exercise File 1: BioVeL-GBIF_BPG_File1_105OR.csv data file. In the sub-workflow selection dialog, choose 'Data Quality (Google Refine)'.
2. Apply a text facet on the column 'locality'.
3. Apply a text filter in the field 'locality' for the keyword 'mud'. You can see in the corresponding facet that only the options that include 'mud' on them are displayed.



44 matching rows (105 total)

locality (change)

22 choices Sort by: name count — Cluster

- mud 11
- mud, shell 6
- bouldersbotten and mud 3
- mud, sand, shell 3
- boulders, mud, pebbles 2
- mud many worms 2
- mud, pebbles, stonesar 2
- blocks, mud 1
- boulders botten and mud 1
- boulders, mud, sand and stones 1

locality: mud

□ case sensitive □ regular expression

island	country	stateProvince	county	locality	minimumElevat	maximumElevat	minimumDepth	maximumDepth	samplin
				mud, shell			15	15	
				sand, mud, shell, macroalgae			18	18	
				mudmix sand, shell, cinder			25	25	
				stonesar macroalgae, samt pebbles and mud			25	50	
				stonesar, shell, mud			60	70	
				mud, pebbles, stonesar			30	80	
				mud, pebbles, stonesar			30	80	
				mud, sand, pebbles, stones			10	30	

4. Apply now a numeric facet in 'maximumDepthInMeters'.
5. Use the slider in the facet to restrict the maximum depth between 0 and 30 meters.
6. The resulting dataset should now include only organisms living in muddy substrate up to 30 meters of depth.
7. You can now extract that information, for example, for use in external software for further analysis, for review by a specialist or expert on such environments, or for evaluation of the resulting list of organisms against their ecological requirements.

Exercise 13

For the last exercise we are going to try a more complex situation where we will need to apply several techniques learnt in this manual. You can find a suggested solution in Chapter 8.

EXERCISE 13: You are a botanist specialized on the family *Euphorbiaceae*, and you will attend a conference in Tenerife in the Canary Islands (Spain) soon. You know about the *tabaibas* (arborescent euphorbias) growing in the islands, and you would like to prepare yourself for a field trip in nearby Güimar.

Use the BioVeL portal to download data from the GBIF repository for the following species of tabaibas: *Euphorbia aphylla*, *E. atropurpurea*, *E. balsamifera*, *E. berthelotii*, *E. bourgeauana*, *E. bravoana*, *E. lamarckii*, *E. lambii*, *E. leptocaula* and *E. regis-jubae*. Use the GBIF Backbone Taxonomy to check for synonyms prior to downloading the file. Restrict your results to those with coordinates and within the geographic area of the Canary Islands.

According to your list, which species are known to occur in the area of Güimar? Do you need to reserve time for additional trips in the islands, or will the trip to the conference along be enough?

7. Data refinement analysis per domain

In this chapter we would like to review how the procedures that we have seen in previous chapters apply to certain domains of occurrence data. We will consider the following:

1. taxonomic and nomenclatural information
2. geospatial information
3. temporal information (related to dates and time)
4. information about people
5. descriptive data (e.g., narrative descriptions of habitats) and
6. indexes and identifiers.

At the end of each section, we will highlight the relevant Darwin Core[2] classes for data quality checks in each domain, using some of the most frequently used terms as examples.

7.1. Taxonomic and nomenclatural information

As the most evident link to the biological context, taxonomic information has traditionally been considered one of the most important parts of an occurrence record. It is also very frequently the first criteria used to query and access information from a database, meaning that the accuracy of the nomenclatural information is critical for the accessibility of the records.

When discussing data quality, one key restriction about taxonomy and nomenclature is that a comprehensive catalogue of all species present on earth still does not exist. For the task of checking our data, we must work instead with a wide selection of reference checklists, whose completeness and accuracy depend on the taxonomic group we are working with, the geographical location, and even the preference of a given researcher or institution. To perform data quality assessment and cleaning operations, you need to be clear about which are the taxonomic references for your institution or project—these references are frequently referred to as 'authority files'.

Nomenclatural and taxonomic information can suffer from many of the types of errors that we listed in section 2.3: missing data values, incorrect data values, Nonatomic data values (very common in databases designed in isolation), domain schizophrenia and inconsistent data values. We will go over each of them as each requires different tools.

- **Missing data values**. Even if values are missing from some of the higher hierarchy levels, we can fill them automatically, provided that we accept one of the classifications available online, like the Catalogue of Life or the GBIF Backbone Taxonomy. Check Tutorial 10 to see how to perform this operation.
  If your data are missing part (or all) of its main taxonomic information, you will probably need to make a big effort to complete or restore the information. That task

[2] The Darwin Core standard is one of the most frequently used for the exchange of primary biodiversity data.

will require returning to the data source or analyzing the contents of other fields. Such a specialized research work is only advisable if the record has a critical or unique value for your study.

- **Incorrect values** may be more or less difficult to identify, depending on the nature of the error. In Tutorials 6 and 7 we have seen how to use facets and clusters to help us find and correct misspellings.

  Values not considered accepted for your taxonomic reference of choice may be more or less easy to spot, depending again if the checklist is available online. If you are using the Catalogue of Life or the GBIF Backbone Taxonomy as a reference, you can use the procedure described in Tutorial 10 to find 'accepted' names for your original data.

  There is no automatic procedure for detecting errors at the level of the biological identification. If you are working with data records that represent physical evidence (e.g., a collection specimen, photograph, or audio recording), you could request that an expert review it in hope of re-identifying the organism. In other cases, you can analyse factors like the quality of match between the location or the ecological information and the expected geographical distribution for the species. But in any case, solving these errors is a specialized work not covered here.
- **Nonatomic data values** are relatively common in databases developed in isolation. The 'ECAT Name Parser' included in the BioVeL OpenRefine extension targets these errors specifically. It automatically analyses the selected column and splits its components into appropriately separate fields.
- **Domain schizophrenia**. This problem affects taxonomic and nomenclatural information more intensely than the rest of domains. A number of marks and annotations that experts include with the name information lead computer systems to treat them as different entities.

  It is important to have clear policies on how to act in each of the most frequent cases. What is the minimum level of definition that you will allow? What do you do when the expert indicates uncertainty in the identification? As always, documenting your assumptions from the start is critical. Always favour annotation and marking over deletion, especially if you plan to share your data with other groups.

  In Tutorial 8, we specifically used filters to find records that include these annotations.
- **Inconsistent data values** can easily appear in the nomenclatural information if you are using data that comes from different sources. As with incorrect data, you can automatically correct such errors by passing the data through one of the reference checklists published online and connected to the BioVeL portal. Please check Tutorial 10 to learn how to use the BioVeL OpenRefine extension to give consistency to your dataset.

If you would like to have more information about taxonomic and nomenclatural data from the point of view of data quality, check the chapter with the same name at 'Principles of Data Quality' by A. Chapman (2005).

In relation to the Darwin Core Standard, data quality checks mainly affect the fields included in the 'Class Taxon'. Some of the most common terms that you would like to check are: scientificName, kingdom, phylum, class, order, family, genus, subgenus, specificEpithet, infraspecificEpithet, taxonRank and scientificNameAuthorship.

7.2. Geospatial information

Geospatial information is the other critical domain where inaccuracies and mistakes make the data fit for fewer uses. Fortunately, the management of geospatial information is not specific to biodiversity data, so many web resources can help assess and improve the quality/accuracy of this component.

The recommendations that we have just given for the nomenclatural and taxonomic domain also apply for the geospatial information: it is important, before starting any data quality related activities, to get hold of any relevant reference lists you will use to check your data against: gazetteers, list of administrative unit abbreviations and correct names, digital and paper maps, etc.

You have to take into account that geospatial information is recorded in many different fields and formats in biodiversity datasets. Locality information is usually recorded as free-text. Information about countries, provinces or geographical datums can be systematized through controlled lists. Geographical coordinates and accuracy information are usually numerical (depending on the format in which they have been recorded). Based on the nature of the information under consideration, you have different options for data quality checking and refinement.

Geospatial information can suffer from several types of the errors listed in section 2.3: missing data values, incorrect data values, nonatomic data values and inconsistent data values. Let's have a look at how this usually happens:

- **Missing data values**: The most frequent issues found in biodiversity datasets—in particular, in those including historical records—are the lack of geographical coordinates, the lack of the datum in which those coordinates are expressed, and the low accuracy of the geographic information.
  One of the early GBIF manuals tackled the issue of retrospective assignment of geographical coordinates to records based on textual descriptions of localities, a process also known as **georeferencing**. You can use what you learnt in Tutorial 9 to locate records with one or both of the coordinate fields missing. But if you plan to start working on georeferencing systematically, we recommend consulting the GBIF manual on the topic: 'Guide to Best Practices for Georeferencing' by A. Chapman and J. Wieczorek (2006).

Certain missing data values can be inferred if other information is present (e.g., if we have geographical coordinates we most likely can find out which country they fall into). But country boundaries and names also change over time, so automatic correction should take this into account.

Use the generic skills learnt in Tutorials 6 and 7 to locate and automatically correct issues with missing values in those cases.

- **Incorrect data values**. Identifying incorrect data values can be tricky as they can be more difficult to detect with the naked eye, and it usually involves[3]:
  - Checking the information against a gazetteer that covers that geographic area at the time of the collection of the information
  - Checking the information against other information present in the record
  - Checking the information against a reference database or in an external GIS system
  - Checking for outliers in the geographical or ecological space

  As explained in Tutorial 5, you can use the BioSTIF application to filter records spatially and, for example, extract any outliers for more thorough examination. It can also help to extract certain subsets of data for additional data checks. Just remember that BioSTIF only works with data records that have both geographical coordinates AND temporal data associated to them.

  To check outliers in the ecological space you can use the options to filter using descriptive information described in Tutorial 11.

  Otherwise, you can still use the generic skills learnt in Tutorials 6 and 7 to locate and automatically correct incorrect values.
- **Nonatomic data values**. This problem also occurs frequently, when all the locality and geographical information appears in a single field. While some information (e.g., country) can be easily recognized and extracted, other information such as habitat information, altitude/depth, etc., can be more difficult to parse and divide.

  While none of the tutorials included in this manual target the parsing of geographical information, you can use filters like the ones seen in Tutorial 8 to help you select and fill fields automatically based the content of aggregated fields.
- **Inconsistent data values**. Geographical information can also suffer from inconsistencies like using different reference systems (i.e., geographical datum), different names of abbreviations for the same entity, etc.

  As explained in Tutorial 6, using facets can help you to locate and correct inconsistent values automatically, when used in combination with good reference lists.

  In other cases, you will have to perform more complex data transformations that go beyond the current capabilities of the BioVeL portal (e.g., datum transformation). Consult other web resources or contact other colleagues working in the same area through the channels listed in section 1.3 for more information.

[3] Extracted from 'Principles and Methods of Data Cleaning', Chapman, 2005.

For a more in-depth analysis of data quality implications in relation to geospatial information, we would like to recommend you to check the chapter 'Spatial data' in the manual 'Principles of Data Quality' by A. D. Chapman (2005).

In relation to the Darwin Core Standard, these checks mainly affect the fields included in the 'Class dcterms:Location', as per the Darwin Core official documentation. Some terms most frequently affected by the errors listed above are higherGeography, continent, waterBody, islandGroup, island, country, countryCode, stateProvince, county, municipality, locality, verbatimLocality, verbatimElevation, minimumElevationInMeters, maximumElevationInMeters, verbatimDepth, minimumDepthInMeters, maximumDepthInMeters, minimumDistanceAboveSurfaceInMeters, maximumDistanceAboveSurfaceInMeters, locationAccordingTo, locationRemarks, verbatimCoordinates, verbatimLatitude, verbatimLongitude, verbatimCoordinateSystem, verbatimSRS, decimalLatitude, decimalLongitude, geodeticDatum, coordinateUncertaintyInMeters, coordinatePrecision, pointRadiusSpatialFit, footprintWKT, footprintSRS and footprintSpatialFit.

7.3. Temporal information

Temporal information in biodiversity occurrence data is usually restricted to information about the moment of collection, and in some cases the time of identification and georeferencing.

In this case, it is relatively easy to detect the most common problems:

- **Missing data values**: If the temporal information is missing, two main routes might fill the gap. If the information omitted from a record was deleted accidentally, you can use other records from the same collector and the same locality to determine whether it was gathered at the same time. You can also study other external resources (e.g., journals from the collector defining when he/she visited certain areas) to fill the temporal information missing.
  The BioSTIF application performs time-related filters, if all your records also have geographical coordinates. Check Tutorial 5 for more information.
  Generic filters as those demonstrated in Tutorials 6 and 8 can help you analyse other data domains in your record and find collections from the same time.
- **Incorrect data values**. Incorrect data values on the temporal information are easier to spot if the field content does not comply with a date/time format. To detect records that have a valid format but are incorrect, you must use tests similar to those for the geographical information like identifying outliers and checking them against the general project/collection timeline, the collector's active period, etc.
  As in the previous case, you can use the BioSTIF application to perform time-related filters, if all the records you are interested in also have geographical coordinates. Remember that you can find information about BioSTIF in Tutorial 5.

Facets can help you to find some of these outliers; Tutorial 6 can help you learn to use them.

- **Inconsistent data values**. In the case of temporal information, inconsistencies can come in heterogeneous datasets from the different way dates are noted in different parts of the world. If you suspect that this can be a problem in your dataset, check for invalid date formats in combination with the different data origins.

If you are interested in knowing more about dates data and quality issues, you should check the article 'On the dates of GBIF mobilised primary biodiversity records' by Otegui et al. (2013).

In relation to the Darwin Core Standard, these checks mainly affect a subset of the fields included in the 'Class Event' and several other time-related fields from the Darwin Core official documentation: eventDate, eventTime, startDayOfYear, endDayOfYear, year, month, day, verbatimEventDate and georeferencedDate.

7.4. Information about people

Biodiversity databases are full with of references to people: species original describers, information collectors or data managers, to name a few. Unfortunately, not all the references are made in the same way, even when referring to the same person.

The information about people who described a given species represents an exception, as citations are regulated by the codes of nomenclature. Because they are considered part of the scientific names, you can use the techniques described in section 7.1 about taxonomic and nomenclatural information to clean data or correct errors.

Information about collectors, managers, technicians, and others is far less systematic. But establishing consistency with these references helps support other data refining procedures and enable certain types of analysis (e.g., natural sciences history studies). Clusters are especially useful for finding different spellings and ways of citing the same person. Please check Tutorial 7 to learn more about clusters. You can also use the generic filtering skills learnt in Tutorials 6 and 8 to locate and correct incorrect values automatically.

In the Darwin Core standard, information about people is distributed among classes, but you can find them in the following individual fields: scientificNameAuthorship, recordedBy or georeferencedBy.

7.5. Other fields with associated reference lists

You can find many fields in a biodiversity dataset that should be filled using controlled list of elements. Unfortunately, unless the software used to capture the information enforces it, users have a strong tendency towards creativity when filling forms and databases.

The Darwin Core standard does not enforce any reference list for its constituent fields, but it does include recommendations on preferred values for many of them. Check the online documentation about the standard in the TDWG site or check the reference guide 'Darwin Core Quick Reference Guide' published by GBIF (2010).

Once you have a reference list for use in specific fields, you can use clusters and filters to find elements not included in your list. Please review Tutorials 6 to 8 if you need help in locating and automatically correcting incorrect values.

This type of Darwin Core terms are spread across all classes. Some examples include occurrenceStatus, geodeticDatum, georeferenceVerificationStatus, taxonomicStatus and nomenclaturalStatus.

7.6. Descriptive information

Descriptive information includes any field whose contents use free text. This usually includes descriptions of the habitat, interactions with other species, etc., but could also involve fields that store information related to the management of the record (e.g., remarks about previous data cleaning processes).

Most of these fields do not have a reference list to check the content of these fields against, so it is more difficult to evaluate and automatically improve their contents. However, some of them are critical for completing information missing in other fields, as we have seen in other sections of this chapter. You can best determine how much effort you want (or can) put into the quality of the contents of these fields.

Tutorial 11 specifically targets filters and data cleaning procedures involving descriptive information.

In relation to the Darwin Core Standard, many terms in different classes include descriptive information. Some examples include occurrenceRemarks, sex, lifeStage, reproductiveCondition, behavior, georeferenceProtocol, georeferenceSources, georeferenceRemarks or taxonRemarks.

7.7. Indexes and identifiers

Some of the fields included in a biodiversity dataset are index numbers and identifiers that link the information to other sources.

Some identifiers are supposed to be unique, so there should never be two records sharing exactly the same index (e.g., the exact same catalogue number).

You can detect duplicates by using simple facets as explained in Tutorial 6, and then sorting these facets by count (you will want to check any number higher than 1).

In the case of the indexes and links to external resources, one good practice is to ensure consistency so users can follow those indexes to find the extra information. To do so, you will need to find the external source, then either check the records individually or obtain a list of the valid indexes that you can check your data against.

Some of the Darwin Core standard terms where you can find indexes and identifiers are occurrenceID, catalogNumber, individualID or geologicalContextID. As in previous cases, you can find them distributed across classes.

8. Answers to the suggested exercises

Exercise 01

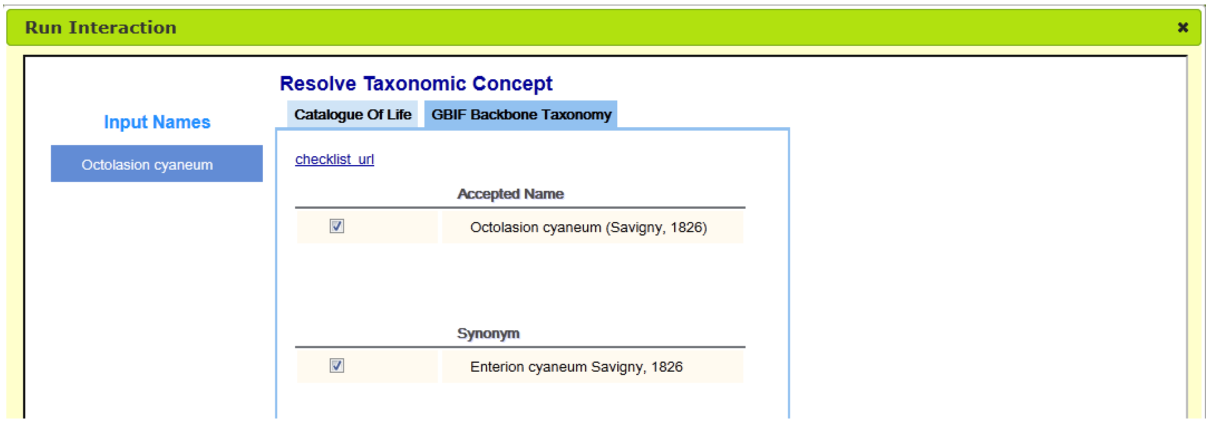
At present, only one synonym is registered in the GBIF Backbone Taxonomy for *Octolasion cyaneum* (Savigny, 1826): *Enterion cyaneum* Savigny, 1826. The Catalogue of Life does not include any synonym for that taxon.

This exercise is very similar to the one we have done in Tutorial 4. Create a CSV file with one column. In the first row include 'nameComplete' and in the second row the name of the taxon '*Octolasion cyaneum*'. You can also make a copy of the file used during that tutorial and substitute the scientific name there with '*Octolasion cyaneum*'.

Create a workflow run, and upload your CSV input file with your species name on it. Choose the 'Taxonomic Name Resolution / Occurrence Retrieval' sub-workflow.

In the dialogue 'Choose Taxonomic Checklist to target', select 'Species2000 - Catalogue of Life' from the 'Standard Checklists'. Then click on 'GBIF Checklist Bank' under 'Aggregated checklists', and write 'GBIF' in the filter textbox. The option 'GBIF Backbone Taxonomy' should be one of the first ones to appear. Click the checkbox to the left of the checklist name, and accept the GBIF user agreement. Click on 'OK'.

In the resulting search results, each checklist is represented by a separate tab.



Run Interaction ✖

Input Names

Octolasion cyaneum

Resolve Taxonomic Concept

Catalogue Of Life | GBIF Backbone Taxonomy

checklist_url

	Accepted Name
☑	Octolasion cyaneum (Savigny, 1826)

	Synonym
☑	Enterion cyaneum Savigny, 1826

Exercise 02

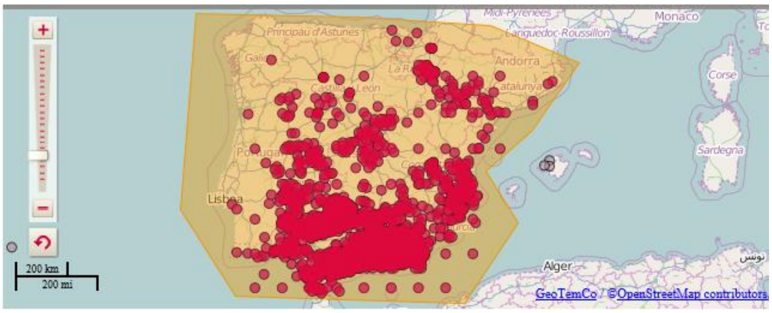
This exercise is very similar to the previous one. At present, there are 8,016 records for these species in the GBIF network, and 6,342 georeferenced ones for the Iberian Peninsula. There was 1 synonym for *R. sphaerocarpa* (*Lygos sphaerocarpa* (L.) Heywood) and 3 for *R. monosperma* (*Genista monosperma* (L.) Lam., *Lygos monosperma* (L.) Heywood and *Retama rhodorhizoides* Webb & Berthel.) in the GBIF backbone taxonomy.

To obtain these results, create a workflow run and upload your CSV input file with the two original species names on it. Choose the 'Taxonomic Name Resolution / Occurrence Retrieval' sub-workflow.

In the dialogue 'Choose Taxonomic Checklist to target', select only 'GBIF Checklist Bank' under 'Aggregated checklists' and write 'GBIF' in the filter textbox. The option 'GBIF Backbone Taxonomy' should be one of the first ones to appear. Click the checkbox to the left of the checklist name, and accept the GBIF user agreement. Click on 'OK'. You will see the synonyms in the dialogue box that follows.

If you leave all the options selected, click on the button 'Retrieve species occurrence data', and select 'GBIF occurrence bank' in the following window. Read and accept the GBIF data user agreement and click on 'OK'.

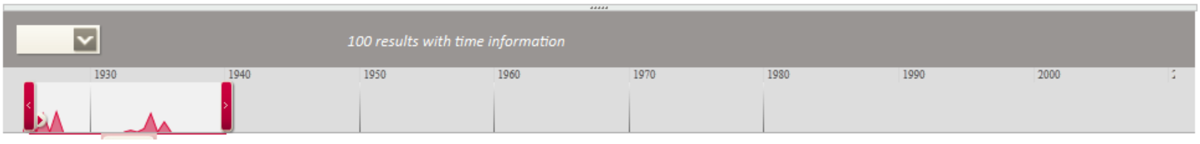
Once the query is completed, you will return to the 'Choose subworkflow' dialogue. You can now download the results as a CSV file (as described at the end of Tutorial 4). To filter the records using geographical criteria, review Tutorial 5. As you can see in the image below, your colleagues are not going to have any trouble to find those species!



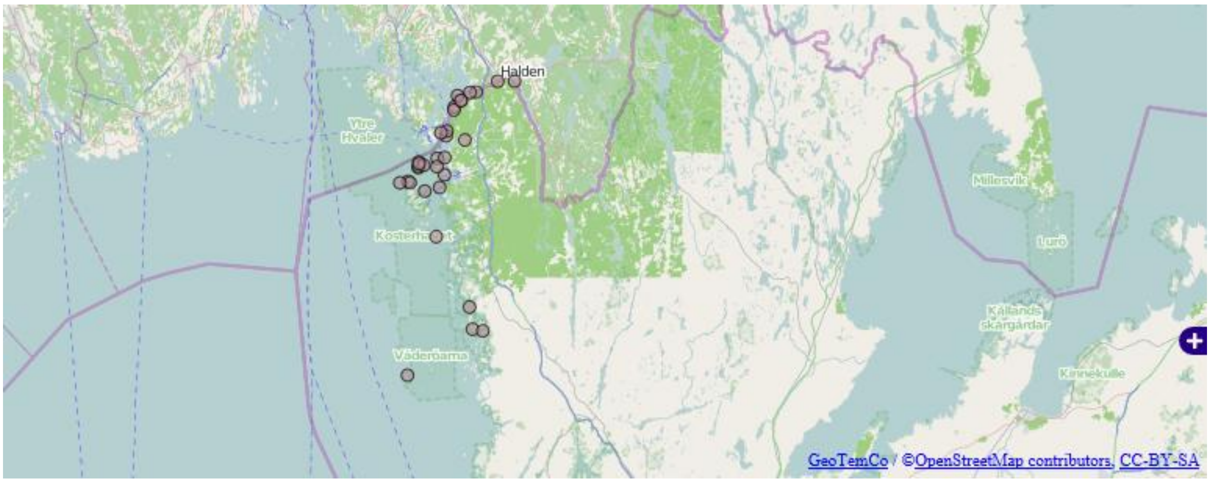
Exercise 03

The correct answer is 54.

Click on the temporal slider in 1940 and slide all the way to the left.



Once you apply the filter by clicking on 'remove unselected items', you can already see that the legend on the top of the map changes into '54 results with location information'.

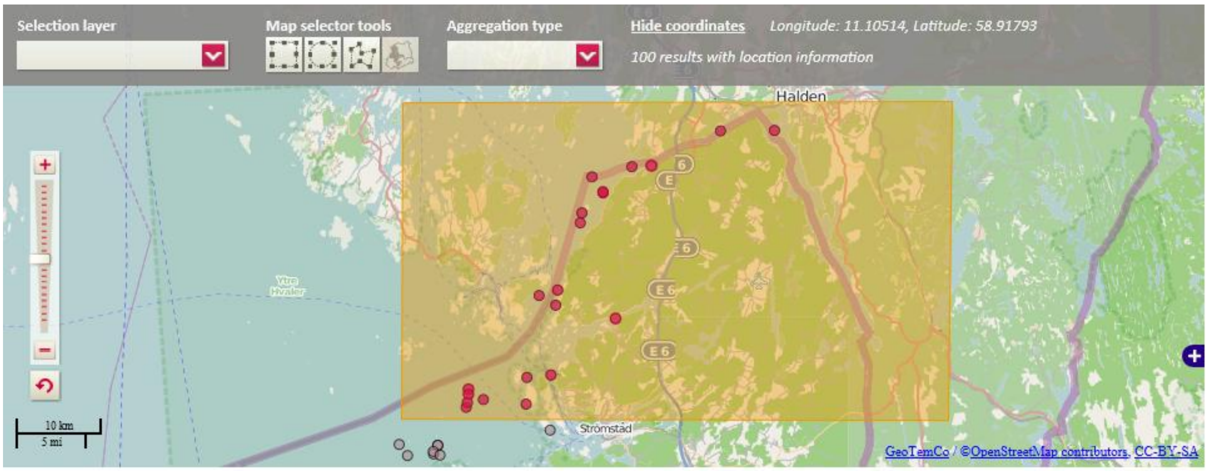


You can even include a screen capture of the map in BioSTIF for your colleague to evaluate if this point density would be suitable for the use she had in mind.

Exercise 04

The correct answer is 44 (from those records which include coordinates and temporal information).

First locate Strömstad in the map (You can use an external mapping search engine such as Google Maps to get an idea where the locality is). Explore the whole dataset at different zoom levels in the map in BioSTIF to make sure that you can see all the records in the dataset north of Strömstad. Use one of the map selection tools to define a figure that includes all the points.



Longitude: 11.10514, Latitude: 58.91793

100 results with location information

Once you apply the filter by clicking on 'remove unselected items', you can already see that the legend on the top of the map changes into '44 results with location information'.

Click on 'continue the workflow', and select 'End Workflow' in the 'Choose Sub-Workflow' dialogue. Download the csv_output.csv file and zip it. You can now attach that file to your answer to the regional government that issued the original question.

Exercise 05

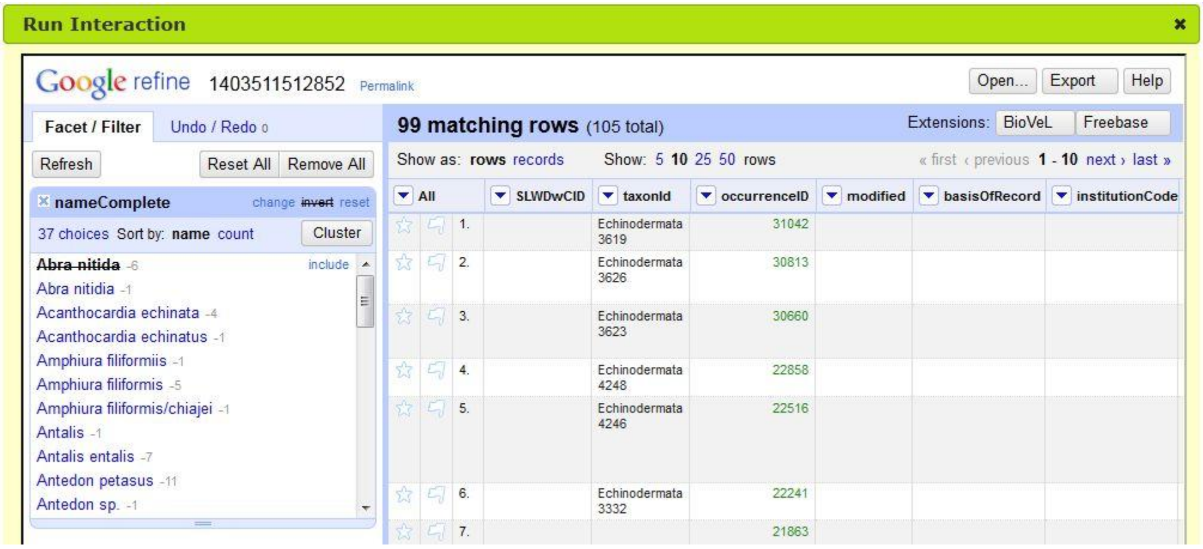
The correct answer is 37. This is an easy one to find out! Start the workflow run, load the data file and start OpenRefine. Create a text facet using the field 'nameComplete' and check the header of the facet box in the left-hand column: you will see that it reads '37 choices'.

Be careful if you are following the same steps as in Tutorial 5, as in that case we applied a first filter to work with only one of the projects. That is why the number is lower in that case.

Exercise 06

The exported file should have 6 rows if you strictly go for text match, and 7 if you want to correct the record assigned to '*Abra nitidia*', a misspelling. The steps to follow are similar to those used in Tutorial 5: Go to the BioVeL portal, start a workflow run, import the example data and select the sub-workflow 'Data Quality (Google Refine)'. Once loaded, create a text facet based on the field 'nameComplete'.

If you noted the misspelling and want to correct it, this is the moment—we will not do it in our example. In the facet box, look for '*Abra nitida*' and click on the name, which selects the records associated to that taxon. Back in the facet box, you'll now see an option to 'exclude' in the top right corner. Clicking on it will invert the selection: all the records not matching the taxon name will be selected.



37 choices Sort by: **name** count

~~**Abra nitida**~~ 6
Abra nitidia 1
Acanthocardia echinata 4
Acanthocardia echinatus 1
Amphiura filiformiis 1
Amphiura filiformis 5
Amphiura filiformis/chiajei 1
Antalis 1
Antalis entalis 7
Antedon petasus 11
Antedon sp. 1

99 matching rows (105 total)

Show as: **rows** records Show: 5 **10** 25 50 rows « first ‹ previous **1 - 10** next › last »

All	SLWDwCID	taxonId	occurrenceID	modified	basisOfRecord	institutionCode
1.		Echinodermata 3619	31042			
2.		Echinodermata 3626	30813			
3.		Echinodermata 3623	30660			
4.		Echinodermata 4248	22858			
5.		Echinodermata 4246	22516			
6.		Echinodermata 3332	22241			
7.			21863			

Now you can go to the 'All' column on the right window and use the arrow to select 'Edit rows' and 'Remove all matching rows'. To view the file's remaining records, you can simply close the facet box by clicking on the button with the cross to the left of the facet name.

From here, you have several alternatives: the quickest one would be to use the 'Export' button on the top right, and select 'Comma Separated Value'. The file will be available for download.

If you want to use the BioVeL portal functionality, use the 'BioVeL' button instead and choose 'Save updates and return to the workflow'. Back in the BioVeL portal, finish the workflow run and download the resulting file.

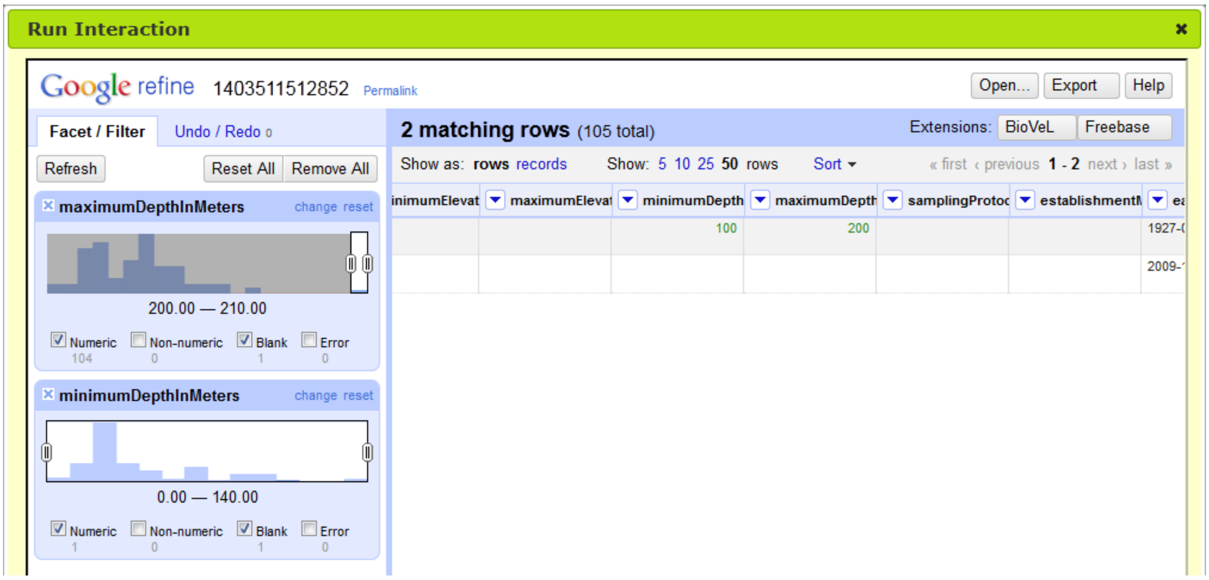
Exercise 07

The depth range of organism occurrences recorded on the dataset is 0-200m.

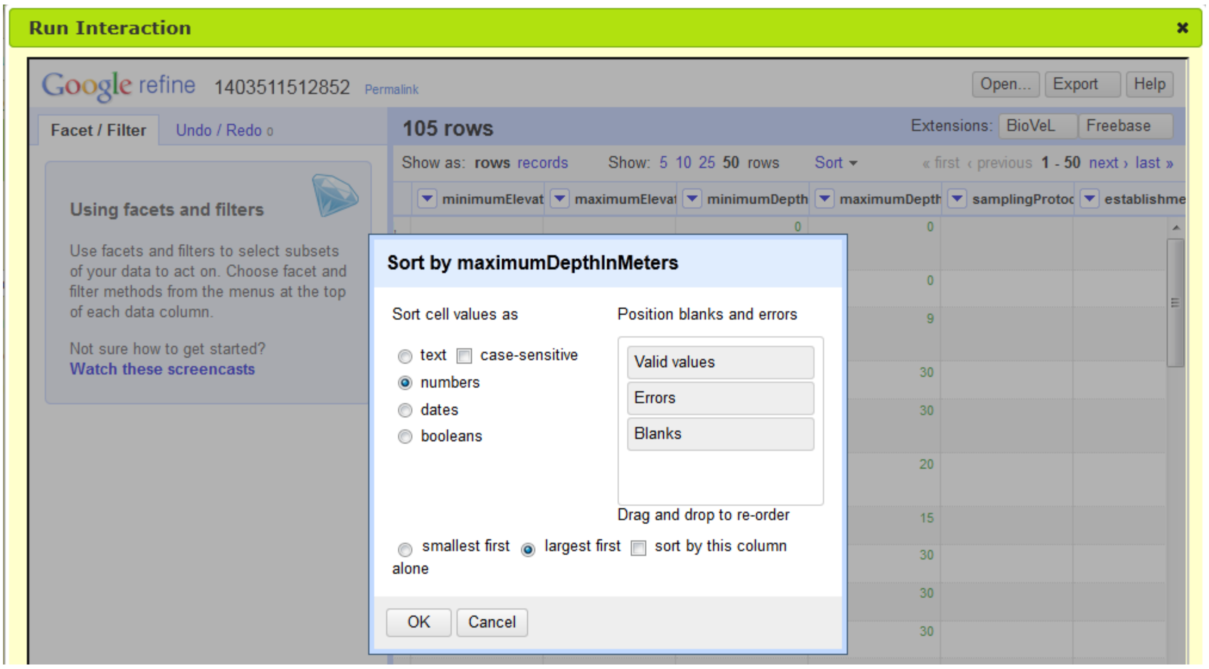
We start this exercise as all the others in this section: go to the BioVeL portal, start a workflow run, import the example data and select the sub-workflow 'Data Quality (Google Refine)'.

You have several options to find the requested information. You can create facets using the fields 'minimumDepthInMeters' and 'maximumDepthInMeters', but be aware that:

- If you use text facets, the results will be ordered alphabetically, so the last value will not be necessarily the highest one. In the case of the 'maximumDepthInMeters', the option '200' appears right after '20'.
- If you use numeric facets, you will view a graph divided in 10-meter intervals. In the case of 'maximumDepthinMeters', the graph will have a legend that could be misleading: '0.00-210.00'. If you use the sliders to see the upper category, you would see that the maximum depth registered is indeed 200 and not 210 meters.


Run Interaction
Google refine 1403511512852 Permalink
Open... Export Help
Facet / Filter Undo / Redo 0
Refresh Reset All Remove All
2 matching rows (105 total)
Extensions: BioVeL Freebase
Show as: rows records Show: 5 10 25 50 rows Sort
« first ‹ previous 1 - 2 next › last »
inimumElevat maximumElevat minimumDepth maximumDepth samplingProtoc establishmentM ea
100 200 1927-(
2009-
maximumDepthInMeters change reset
200.00 — 210.00
☑ Numeric 104 □ Non-numeric 0 ☑ Blank 1 □ Error 0
minimumDepthInMeters change reset
0.00 — 140.00
☑ Numeric 1 □ Non-numeric 0 ☑ Blank 1 □ Error 0

A quicker method is simply sort the table based on those fields. Use the same arrow to the left of the field name and select 'Sort'. Choose 'numbers' in the 'Sort cell values as' section, and 'smallest first' or 'largest first' depending on the field you are using to sort.


Run Interaction
Google refine 1403511512852 Permalink
Open... Export Help
Facet / Filter Undo / Redo 0
105 rows
Extensions: BioVeL Freebase
Show as: rows records Show: 5 10 25 50 rows Sort
« first ‹ previous 1 - 50 next › last »
minimumElevat maximumElevat minimumDepth maximumDepth samplingProtoc establishme
Using facets and filters
Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.
Not sure how to get started?
Watch these screencasts
0 0 0 9 30 30 20 15 30 30 30

Sort by maximumDepthInMeters

Sort cell values as
○ text □ case-sensitive
● numbers
○ dates
○ booleans

Position blanks and errors
Valid values
Errors
Blanks
Drag and drop to re-order

○ smallest first ● largest first □ sort by this column alone

OK Cancel

After two of these sort operations, you will have the information you sought: 0 to 200 meters.

Exercise 08

To remove the records without values in the two coordinate fields 'decimalLatitude' and 'decimalLongitude', you can also use facets.

Go to the BioVeL portal, start a workflow run, import the example data and select the sub-workflow 'Data Quality (Google Refine)'.

Create a text facet for the first of the two above-named fields. In the list of available options for the facet, choose the last option '(blank)'. Now go to the right window. Go to the 'All' column (the first one), and use the arrow to select 'Edit rows' and 'Remove all matching rows'. Close that facet window and repeat the operation with the other field. The remaining dataset should have 101 rows.

Go now to the 'Undo / Redo' area by clicking on the link on the top left part of the window and click on 'Extract...'. Make sure that the two 'remove...' operations are listed and marked.

Copy the text on the right and send it in your message to your partners. It should look similar to this:

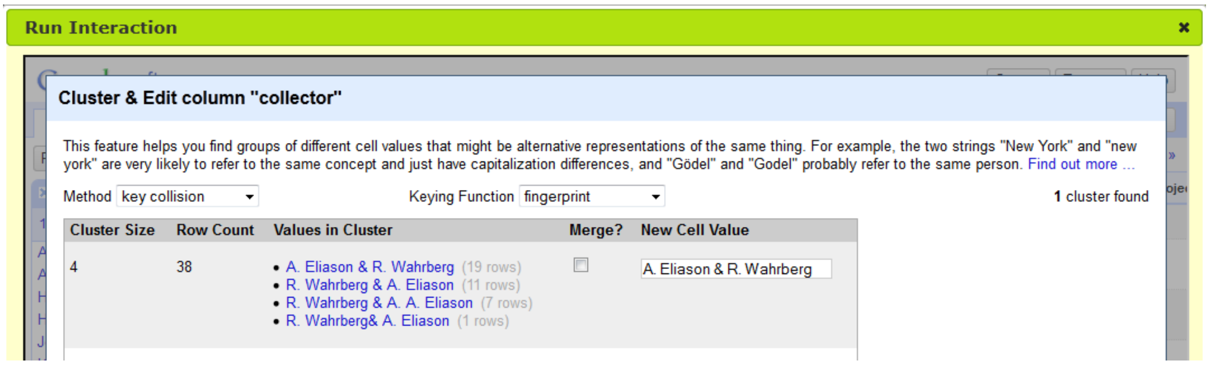
```
[
  {
    "op": "core/row-removal",
    "description": "Remove rows",
    "engineConfig": {
      "facets": [
        {
          "invert": false,
          "expression": "value",
          "selectError": false,
          "omitError": false,
          "selectBlank": true,
          "name": "decimalLatitude",
          "omitBlank": false,
          "columnName": "decimalLatitude",
          "type": "list",
          "selection": []
        }
      ],
      "mode": "row-based"
    }
  },
  {
    "op": "core/row-removal",
    "description": "Remove rows",
    "engineConfig": {
      "facets": [
        {
          "invert": false,
          "expression": "value",
          "selectError": false,
          "omitError": false,
          "selectBlank": true,
          "name": "decimalLongitude",
          "omitBlank": false,
          "columnName": "decimalLongitude",
          "type": "list",
          "selection": []
        }
      ],
      "mode": "row-based"
    }
  }
]
```

Exercise 09

After our refinements, we ended up with ten options. Here's how:

Start a workflow run with the example dataset and start the 'Data Quality (Google Refine)' subworkflow. Duplicate the column 'collector' and name the new column 'collectorVerbatim', in case you need to return to the original data.

Create a text facet based on the field 'collector' and then click on 'Cluster'. In the default method 'key collision', you'll see the following result:



Cluster & Edit column "collector"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

Method: key collision   Keying Function: fingerprint   **1** cluster found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
4	38	• A. Eliason & R. Wahrberg (19 rows) • R. Wahrberg & A. Eliason (11 rows) • R. Wahrberg & A. A. Eliason (7 rows) • R. Wahrberg& A. Eliason (1 rows)	□	A. Eliason & R. Wahrberg

Merge all 38 existing alternatives into 'A. Eliason & R. Wahrberg'. There are no other evident clusters.

In the list, you can identify the problematic record you were alerted about: 'L.A. J_gerski_ld'. You can edit its 23 appearances automatically here. Just click on the 'edit' link on the right of the name, and change it to 'L.A. Jägerskiöld'.

The resulting list, sorted by number of occurrences looks like this:

- A. Eliason & R. Wahrberg (38 occurrences)
- L.A. Jägerskiöld (23 occurrences)
- Hans G. Hansson (14 occurrences)
- Jon-Arne Sneli (11 occurrences)
- Maj Persson (5 occurrences)
- Thomas Dahlgren (3 occurrences)
- Kennet Lundin (2 occurrences)
- Linda Ottosson (2 occurrences)
- Anna Karlsson (1 occurrence)

- Helena Wiklund (1 occurrence)

Exercise 10

Our neatest list contains 24 entries. How many did yours have?

We started a workflow run with the example dataset and start the 'Data Quality (Google Refine)' subworkflow. Following the same steps as in Tutorial 7, use a nearest neighbour cluster to automatically correct the three misspellings displayed: '*Amphiura filiformiis*', '*Abra nitidia*' and '*Aphrodita aculeatae*'. You can use a canonical name cluster to remove the existing 'cf.' references automatically.

This operation leaves 35 names. As the requirement is to produce a list of species, you can check for records defined at the genus level and remove them, as explained in Tutorial 8.

That further reduces the list to 27 names. If you browse the list you will still notice inconsistencies, such as two names separated by a bar '/', and some misspellings that have not been detected by the clustering (e.g., *Acanthocardia echinatus*). If you correct/remove those, you will end with the following 24 options:

- *Abra nitida*
- *Acanthocardia echinata*
- *Amphiura filiformis*
- *Antalis entalis*
- *Antedon petasus*
- *Aphrodita aculeata*
- *Apomatus globifer*
- *Aporrhais pespelecani*
- *Arctica islandica*
- *Astarte montagui*
- *Astarte sulcata*
- *Asterias rubens*
- *Astropecten irregularis*
- *Axinella rugosa*
- *Crania anomala*
- *Echinus miliaris*
- *Leptopentacta elongata*
- *Mytilus borealis*
- *Mytilus edulis*
- *Neocrania anomala*
- *Novocrania anomala*
- *Phakellia rugosa*
- *Psammechinus miliaris*
- *Trachythyone elongata*

We cannot reduce this list more without using taxonomic checklists. If you want to continue the process, you should check Tutorial 10.

Exercise 11

This is an exercise based on research and reflection. Follow Tutorial 9 to find out the current problems of the dataset: one record with both coordinates empty, three with one of them empty and one that is not numeric.

Before making any changes in the dataset, we will duplicate 'decimalLatitude' and 'decimalLongitude' into 'decimalLatitudeVerbatim' and 'decimalLongitudeVerbatim' respectively. If the column 'georeferenceRemarks' does not exist, create it.

In a dataset as this one, we can try to find similarities between records. If several records share the collector, locality description and date/time of collection, it seems reasonable to assume that they all belong to the same collection effort, allowing you to add or correct the existing information on the basis of that information.

Let's review an example. Looking at the record marked as 'not numeric' coordinates, you will see a longitude value of '11.222ac333'. The latitude looks okay (59.08468333), so we can try to use it to find similar records. Show all the records and filter by that value in 'decimalLatitude'. We obtain a result of 6 records, all collected by 'A. Eliason & R. Wahrberg' the 31 July 1925 for the same location and the same depth. It seems safe to assume that there was some problem with that record in particular and that the correct value for the longitude is the same as the rest of the records: '11.22273333'.

You can apply exactly the same principle for the records marked as having only one coordinate. For the one having no coordinates, we can try filtering by date/time using the 'earliestDateCollected' and then compare the rest of the fields. It seems to match with other records collected at latitude 58.95081667 and longitude 11.02076667.

What we just did can be considered an example of georeferencing. You may want to check the section on geospatial information in section 7.2 for more information about this.

When making this kind of exercises, always apply your common sense, keep the original data AND document your changes and assumptions. We created the field 'georeferenceRemarks' for that purpose!

Exercise 12

After doing some data cleaning over the names information, we ended up with a list of six echinoderms. How many did you get?

We started a workflow run in the BioVeL portal, loaded the example file and chose the 'Data Quality (Google Refine)' subworkflow. We do not need to create new fields to store the verbatim data, as this procedure actually preserves them. What we did is to create the column 'taxonRemarks' to document the potential changes made.

Using the arrow in the 'completeName' column and selecting 'BioVeL' and 'Resolve Name' we launch the process. We selected 'GBIF-Backbone' as the list to check our dataset against.

Once we obtain a result, we have an 'nameAccepted' column. Now we can create a facet based on that field to check its contents. To restrict our list to the echinoderms, we need to filter to match 'Echinodermata' based on the 'phylum' column (beware of the capital letters: the original field was called 'Phylum'). There are 40 records in our example that have been assigned to that phylum.

The final list of accepted species should look something like this:

- *Amphiura filiformis*
- *Antedon petasus*

- *Asterias rubens*
- *Astropecten irregularis*
- *Leptopentacta elongata*
- *Psammechinus miliaris*

Exercise 13

This is a very thorough exercise we are suggesting here, right? Let's get started!

We need a CSV file with all the names included in the exercise. You can re-use the Exercise File 2: BioVeL-GBIF_BPG_File2_CrassostreaGigas.csv and simply substitute the content by the list of names provided.

Start a workflow run, and upload the CSV file. Start with the 'Taxonomic Name Resolution / Occurrence retrieval' sub-workflow. Under the 'Aggregated Checklists', select 'GBIF Checklist Bank' and then choose the 'GBIF Backbone Taxonomy' from the list that appears on the right.

In the next step you will be offered a number of synonyms for your species. Leave all of them selected and proceed to the occurrence retrieval. Select the 'GBIF Occurrence Bank' as source, and proceed with the download of records.

Back into the sub-workflow selection, go to 'Data Selection (BioSTIF)'. Using the polygon selection tool in the map, make a polygon that includes the islands and the area surrounding them (use Google Maps or a similar mapping application if you need help to locate the islands or their extension). At the time of preparing this exercise, that left us 1413 records from the original download. Go back to the workflow.

To finish the exercise, select the 'Data Quality (Google Refine)' sub-workflow. We need to filter our results for the area near Güimar. The quick filter reveals no records for that locality... which is strange. If you explore the contents of the field using facets or just browsing, you will realise that there are indeed several records for 'Guimar', without the umlauted ü. Use that string to filter.

A single text filter on locality for 'Guimar' now gives indeed results (49 when we did this exercise). If you create a text facet on the 'nameComplete' field, you will obtain the list of species cited for Güimar according to the GBIF network. We got the following names, sorted by the number of occurrences (most frequent first):

- *Euphorbia obtusifolia* subsp. *regis-jubae*;
- *Euphorbia balsamifera* Aiton (including *E. balsamifera* subsp. *balsamifera*);
- *Euphorbia lamarckii* Sweet (also cited as *Euphorbia obtusifolia* Poir.); and
- *Euphorbia atropurpurea* Brouss. ex Willd.

So six out of the 10 species you are interested in are not present in the area of Güimar. If you want to make the most of your visit, you should probably plan more trips while you are there!

Acknowledgements

We would like to thank the members of the GBIF community and Secretariat who have actively collaborated in the review of the contents of this manual [*to be completed with the names of the actual contributors to the public review*].

We would also like to thank all the members of the BioVeL project team who have worked together to turn into reality the vision of a virtual laboratory providing tools designed to process data for research on biodiversity.

Glossary

- **BioSTIF**: BioVeL Spatio-Temporal Interactive interFace (Web Tool/Interaction Service).
- **Data cleaning**: Detecting and correcting incomplete, incorrect or irrelevant records from a dataset.
- **DRW**: Data Refinement Workflow.
- **e-Laboratory** : Computer environment supporting the research on biodiversity issues using large amounts of data from cross-disciplinary data and computational sources.
- **ENM**: Ecological Niche Modeling.
- **GBIF**: Global Biodiversity Information Facility.
- **LSID**: Life Science IDentifier.
- **Occurrence** : The category of information pertaining to evidence of an occurrence in nature, in a collection, or in a dataset (specimen, observation, etc.).
- **OpenRefine** : Tool for cleaning tabular data - used interchangeably with GoogleRefine. See chapters 7 and 8 for more information about the tool.
- **Synonym**: In scientific nomenclature, a scientific name that applies to a taxon that now has a different scientific name. The correct name depends on the taxonomic viewpoint.
- **Taxon**: A taxonomic group or unit. Strictly spoken a group of organisms proceeding from a single ancestor (so a monophyletic branch/cluster/clade in a phylogeny). Could be of any rank (species, genus, order, class, etc).
- **Taxonomy**: The academic discipline of defining groups of biological organisms on the basis of shared characteristics and giving names to those groups.
- **Taverna**: Workflow Management System used in BioVeL. http://www.taverna.org.uk/.
- **TDWG/BIS**: International body established to define standards for use in biological data projects. Its official name is Biodiversity Information Standards (BIS) but it is still widely referred by its old acronym TDWG (Taxonomic Databases Working Group).
- **Workflow** : Automation of a business process, in whole or part, during which documents, information, or tasks are passed from one participant to another for action, according to a set of procedural rules. (ISO/DIS 19119). In Taverna, it is the series of data analysis (steps) to process data, be that from one's own research and/or from existing sources.
- **Workflow run**: An execution of a single workflow instance. This information includes what input data was provided.

References

- Chapman, A. D. 2005. **Principles of Data Quality**, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. ISBN 87-92020-03-8. Available online at http://www.gbif.org/orc/?doc_id=1229.
- Chapman, A. D. 2005. **Principles and Methods of Data Cleaning - Primary Species and Species-Occurrence Data**, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. Available online at http://www.gbif.org/orc/?doc_id=1262.
- Chapman, A.D. and J. Wieczorek (eds). 2006. **Guide to Best Practices for Georeferencing**. Copenhagen: Global Biodiversity Information Facility. Available online at http://www.gbif.org/orc/?doc_id=1288.
- GBIF. 2010. **GBIF Position Paper on Future Directions and Recommendations for Enhancing Fitness-for-Use Across the GBIF Network**, version 1.0. Authored by Hill, A. W., Otegui, J., Ariño, A. H., and R. P. Guralnick. 2010. Copenhagen: Global Biodiversity Information Facility, 25 pp. ISBN: 87-92020-11-9. Available online at http://www.gbif.org/orc/?doc_id=2777.
- Otegui, J., Ariño, A. H., Chavan, V. & Gaiji, S. 2013. **On the dates of GBIF mobilised primary biodiversity records**. Biodiversity Informatics, 8, 2013, pp. 173-184. Available online at https://journals.ku.edu/index.php/jbi/article/view/4125.