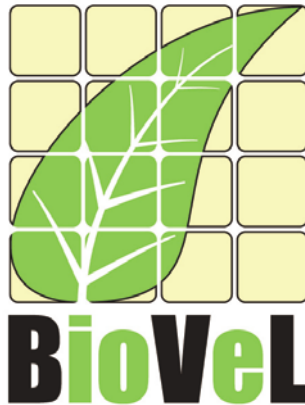Taxonomic Data Refinement (Integrated) Workflow

BioVeL – Biodiversity Virtual e-Laboratory

Workflow Documentation

**Taxonomic Data Refinement
(Integrated) Workflow
for local execution**

November 2012

Capacities Programme of Framework 7: EC e-Infrastructure Programme –
e-Science Environments - INFRA-2011-1.2.1

Grant Agreement No: 283359
Project Co-ordinator: Mr Alex Hardisty
Project Homepage: http://www.biovel.eu

BioVel Documentation

Taxonomic Data Refinement (Integrated) Workflow

1 Description

The Taxonomic Data Refinement Workflow provides an environment for preparing observational and specimen data sets for use in scientific analyses such as: species distribution analysis,species richness and diversity studies, species occurrence studiesh, historical analysis, and other spatio-temporal analyses.

This pack contains:

- The Taxonomic Data Refinement (Integrated) Workflow (.t2flow file)

- Its (version) dependent libraries

- Relevant documentation, including diagrammatic representation of the workflow

- Example data files , including input and output data.

2 General

2.1 Name of the workflow and MyExperiment identifier

Name: **Taxonomic Data Refinement (Integrated) Workflow**

Download: The workflow pack can be downloaded at myExperiment under the following address http://www.myexperiment.org/packs/267.html

2.2 Date, version and licensing

Last updated: 31st October 2012

Version: Data Refinement Workflow v8

Licensing: Creative Commons Attribution ShareAlike CC-BY- SA

2.3 How to cite this workflow

To report work that has made use of this workflow, please add the following credit acknowledgement to your research publication:

The results reported in this publication come from processing data (<*personal source or others--cite which, e.g. GBIF>*) through BioVeL workflows and services (www.biovel.eu). The taxonomic data refinement workflows were run on <*date of the workflow run>*. BioVeL is funded by the EU's Seventh Framework Program, grant no. 283359.

3 Scientific specifications

3.1 Keywords

Taxonomy, Species richness and diversity, Species occurrence, Species distribution analysis, Taxonomic data cleaning and refinement, Historical analysis, Taxonomic Name Resolution, Synonym expansion, Geo-temporal data selection and filtering, spatio-temporal analysis, Occurrence retrieval, Data quality and filtering

3.2 Scientific workflow description

The aim of the Taxonomic Data Refinement Workflow (Fig. 1) is to provide a seamlessly connected environment for preparing observational and specimen data sets for use in scientific analysis. The workflow accepts input data in a recognized format, and these data can be combined from various sources (e.g. occurrence retrieval services, local user data sets). The workflow includes a number of graphical user interfaces to view and interact with the data, while the output of each part of the workflow is compatible with the input of each part. This implies that the user is free to choose any specific sequence of actions and repetition of steps. The construction of the workflow also allows for custom-built as well as third-party tools applications to easily be integrated into the workflow. Currently, the data refinement workflow is composed of three distinct parts:

1. Taxonomic Name Resolution / Occurrence retrieval. Here users can resolve a list of scientific names using taxonomic checklists. This process results in the retrieval of taxonomic information related to each scientific name, including related synonyms as well as other concept information like rank, classification, etc. The resulting information can then be used to retrieve occurrence data or saved as a list of expanded names. The synonym expansion and occurrence retrieval are built on generic frameworks allowing the inclusion of multiple sources.

2. Geo-temporal data selection. Here users select, filter, and refine data records according to spatial and temporal criteria. Geographical selection can be done by drawing polygons, circles, rectangles, etc., on a map as well as by filtering data based on geo-markers (e.g. country, latitude/longitude). Records relating 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. Here the user can apply a set of data quality and data integrity rules on the data matrix. This allows users to perform data-specific cleaning and filtering. This tool is developed based on 'Google Refine' as the central interface for accessing the various local as well as external functionalities.

BioVel Documentation

Taxonomic Data Refinement (Integrated) Workflow
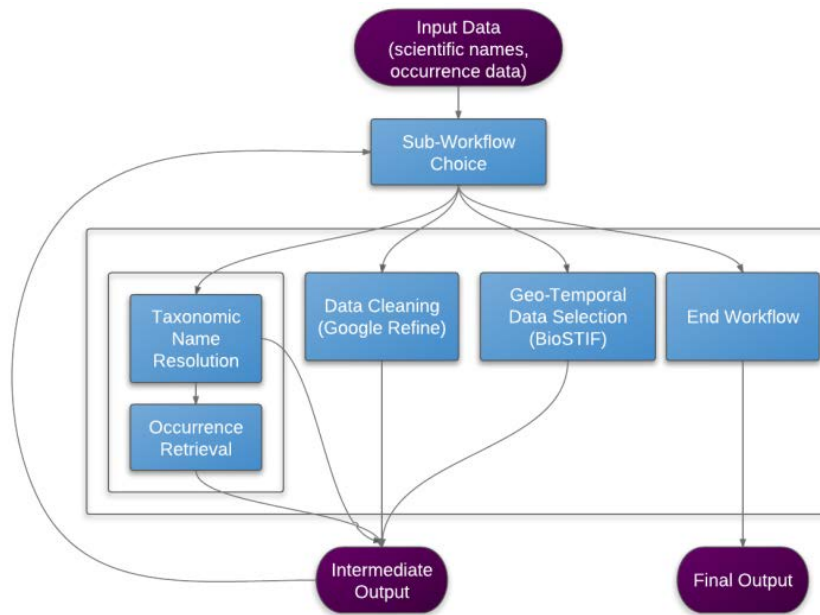
Fig 1. The components of the data cleaning workflow.

For more detailed description of the functions, please visit the [*Tutorial*](#) *section.*

4   Technical specifications

4.1 Execution environment and installation requirements

The workflow has been developed to be run in the automated workflow environment – Taverna[1]. In its current form, the workflow file (with the .t2flow extension) can be loaded and executed in the workbench variant of Taverna and has been tested with version 2.4.

The workflow is dependent on external libraries (written in JAVA[2]) as well as an instance of the Google Refine[3] server which needs to run locally.

---

[1] http://www.taverna.org.uk/

[2] http://www.java.com

[3] http://code.google.com/p/google-refine/

NOTE:

1. The requirement of running the Google Refine server locally will be optional in future releases since a remote version of the server will be provided as default.

2. Future release will provide the possibility to run the entire workflow remotely, as a web application and will not require the installation of any local elements.

4.2 Taverna installation, including updates and plugins

4.2.1 Taverna installation

You need to have the Taverna workbench (v2.4) installed on your computer. For installation files and instructions, please go to http://www.taverna.org.uk/download/workbench/2-4/

Please make sure you install the updates under Advanced -> Updates and plugins.

4.2.2 Interaction Plug-in

In addition, you need to install the Taverna Interaction plug-in. Please note that the Plugin is a "work in progress". Comments, suggestions and especially bug reports are welcomed.

4.2.2.1 Installation of the plug-in:

- Add http://www.mygrid.org.uk/taverna/internal/biovel/240/ as a plugin site - see http://www.mygrid.org.uk/dev/wiki/display/taverna/Plugins

- Install the Interaction Service Plugin

- Restart Taverna

4.2.2.2 Configuring the Interaction preferences:

- Once you restart Taverna
- Goto Preferences (under File or Taverna) -> Interaction and update only if the following conditions apply:
    1. If port 8080 is already by some other process on your machine then specify the port on which you want Jetty (the feed and HTML server) to run, or

    2. If you have an external server available then specify the URL for another machine and the relative webdav and Atom feed.

- Restart Taverna.

4.2.3 Dependent Libraries

The following dependencies need to be installed before running the workflow. The dependent libraries (jars) should be copied to the Taverna local lib directory. You can find it using the Taverna Workbench. Go to the menu "Advanced" -> "Show logs and data folder" and find there the "lib" folder. The libraries to be copied are in the workflow pack and include,

- DCWorkflow.jar (current) : includes all utility / gui classes with functionality used in the workflow.

- opencsv-2.3.jar : from the opencsv project, used for parsing CSV data.

- json-simple-1.1.1.jar : from the json-simple project, used for parsing json.

4.2.4 Google Refine Installation instructions

Full Install: This is the complete install of Google Refine along with the BioVeL Extension. After downloading the patched version, the installation procedure is similar to the standard Google Refine installation.The workflow requires Google Refine to be running at http://127.0.0.1:3333 (which is the default).

WARNING! When deleting other Google Refine installations on your machine please note that you may lose uploaded data, so please export any data before the deletion. The OS' supported are,

- Microsoft Windows (tested on Windows 7 64bit) : [link].

- Mac OSX (tested on 10.6.8): [link].

Extension Install: If you have already installed Google Refine and would like to continue to use this installation, you can download only the BioVeL extension and install it. The extension can be found here as a .zip file which must be extracted and installed as a Google Refine extension.

4.3 Description of the operations, functions and interactions in the workflow

This section describes the Data Refinement Workflow in terms of the functionality provided by the workflow.

4.3.1 Input

Due to the flexible design of the workflow, the output of each sub-workflow is compatible with the input of each section (including the input to the workflow itself). Currently, the workflow supports only the CSV format (with controlled header terms defined here. The reason for this is to focus more on building functionality and providing more options.

4.3.2 Intermediate / Final Output

The output of each sub-workflow as well as the final output of the integrated workflow (depending on the sub-workflow chosen) could be one of the following two types,

4.3.2.1 Taxon Data

This type of data consists of information related to taxa, including classification, status, rank, etc.

nameComplete	authorship	taxonomic_statu	rank	kingdom	phylum	class	order	family	genus
Daphnia		accepted name	Genus	Animalia	Arthropoda	Branchiopoda	Diplostraca	Daphniidae	
Daphnia			genus	Metazoa	Arthropoda	Branchiopoda	Diplostraca	Daphniidae	Daphnia
Daphnia O. F. MÃƒÆ'Ã‚¼ller, 1785		accepted	genus	Animalia	Arthropoda	Branchiopoda Latreille, 1817	Diplostraca Gerstaecker, 1866	Daphniidae Straus, 1820	Daphnia O. F. MÃƒÆ'Ã‚¼ller, 1785
Eois		accepted name	Genus	Animalia	Arthropoda	Insecta	Lepidoptera	Geometridae	
Eois			genus	Metazoa	Arthropoda	Insecta	Lepidoptera	Geometridae	Eois
Eois HÃƒÆ'Ã‚¼bner, 1818		accepted	genus	Animalia	Arthropoda	Insecta	Lepidoptera	Geometridae Leach, 1815	Eois HÃƒÆ'Ã‚¼bner, 1818
Hymenoptera		accepted name	Order	Animalia	Arthropoda	Insecta			
Hymenoptera			order	Metazoa	Arthropoda	Insecta	Hymenoptera		
Hymenoptera		accepted	order	Animalia	Arthropoda	Insecta	Hymenoptera		

4.3.2.2 Species Occurrence Data

This type of data consists of information related to the physical occurrence of specific species, including geographical location (latitude, longitude, depth among others), time range, etc.

taxonName	occurrenceID	decimalLatitude	decimalLongitu	earliestDateCol	latestDateColle	coordinateUnc	country	collector	fieldNotes	locality
Tapes japonica	40833298	35.4492	139.642				Japan			Kanagawa Prefectural Office
Tapes japonica	144534824	21.29	-157.915556	1981-11-14	1981-11-14		United States			Oahu Island, Mamala Bay
Orconectes limosus	55225700	51.6581	4.4627				NL			Tonnekreek-oost
Orconectes limosus	55225745	52.5241	4.9623				NL			Sloot, parallel aan Meeuwstraat, tpv loo
Orconectes limosus	55225771	52.5605	6.2064				NL			Watergang 9.0 Tolhuislanden ov Spoorijn
Orconectes limosus	55225643	51.9286	5.8756				NL			Instroming Linge nabij Elsterbrug
Orconectes limosus	55225763	51.6216	4.3912				NL			Derriekreek
Orconectes limosus	55225835	52.0586	5.0409				NL			NIEUWEGEIN
Orconectes limosus	55226583	51.6929	5.8273				NL			Sambeekse uitwatering
Orconectes limosus	55226639	52.0659	6.4115				NL			Veengoot Reerinkweg Veldhoek
Orconectes limosus	55226535	51.8112	4.7735				NL			Kavelsloot-sloot loodrecht op noordelijk
Orconectes limosus	55225712	52.5123	6.3916				NL			Regge Ommen
Orconectes limosus	55225724	51.7146	4.8812				NL			west van A27,sloot langs IGB fabr.

4.3.3 Workflow Run

The workflow begins with asking the user for an input file and then performs the following steps.

4.3.3.1 Sub-Workflow Chooser

At the beginning of the workflow and after the completion of each sub-workflow the user is presented with a list of sub-workflows available for execution as well as the choice to end the workflow when desired.
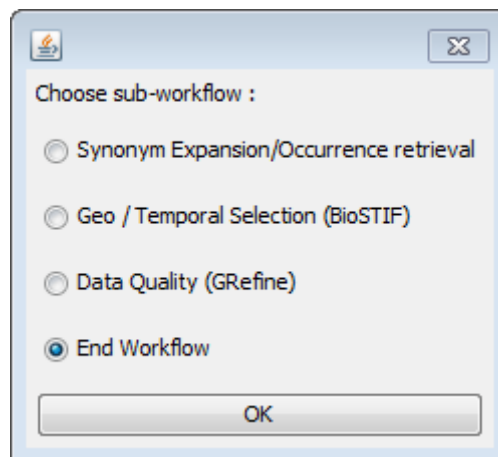
This leads to one of the following workflows to be executed,

4.3.3.2 Synonym Expansion / Occurrence Retrieval
4.3.3.3 Data Quality / Cleaning
4.3.3.4 Data Selection

4.3.4 Dependencies

The external tools / libriaries use in the workflow include,

- GBIF ECAT Name Parser : This utility is a scientific name parsing tool which parses name strings into its individual atomic parts. The tool has been provided by GBIF as a web service as well as a .jar library. The library has been included in the BioVeL Googel Refine extension and is redistributed under the Apache 2.0 licence.

4.4 Web service documentation and specification

This workflow uses Web services from 3rd party Service Providers like GBIF and Species 2000 / ITIS Catalogue of Life. Each provider maintains a page of detailed information concerning the services they offer. In cases where these web pages do not provide adequate information or answers to problems the organisations concerned can be contacted directly by email:

- For GBIF Web Services, email contact: <to be added - Alex has asked GBIF how/who>
- For Species 2000 / ITIS Catalogue of Life Advanced Services, email contact: support@sp2000.org

5   Functional annotation of the workflow in Taverna

5.1 Annotation Synonym Expansion / Occurrence retrieval

The input/output to/from this sub-workflow follows the format already discussed, with the additional requirement that the column 'nameComplete' is mandatory for the input.

This sub-workflow uses the scientific names provided in the 'nameComplete' column in the input and retrieves taxonomic information related to these names by targetting checklists which are chosen by the user. This information includes information like classification, rank , synonyms, etc. The user is then given the option to:

- use the de-duplicated list of the scientific names and the synonyms of these names to retrieve occurrences or

- return from the sub-workflow with conceptual information related to the names

In the case of retrieving occurrences, the user is given a choice of species occurrence databanks (currently only GBIF is a stable option) to target, resulting in occurrence data as output.

5.2 **Annotation BioSTIF**

The input/output to/from this sub-workflow follows the format already [discussed](#), with the additional requirement that the columns 'decimalLatitude' / 'decimalLongitude' is mandatory for the input.

This sub-workflow takes as input species occurrence data, loads it into a remote BioSTIF server and opens up the geo-temporal web (BioSTIF) client in a web browser with a map-based view of the data. The user can then select data by isolating specific regions using standard / custom polygons or by choosing time ranges using the time selection tool.

Once the data is selected the user can return to Taverna with the intermediate 'selected' output data.

5.3 **Annotation GoogleRefine**

The input/output to/from this sub-workflow follows the format already [discussed](#).

This sub-workflow takes as input both taxon related data as well as species occurrence data, loads the data into an already running (local) Google Refine server and opens up the Google Refine web client with a tabular view of the data. The user can then choose to filter / select / remove data as required using existing functionality as well as functionality added on by the BioVeL extension.

When the data has been modified as required, the user can return to Taverna with the intermediate 'cleaned' output data.

6 Tutorial

6.1 General

In this tutorial, you will use a list of species to retrieve synonyms from accepted species checklist banks, and thereafter retrieve species occurrence data from a GBIF for selected synonyms. Thereafter you will inspect, format, clean, and refine the data in a spatio-temporal interface (BioSTIF) as well as in a tabular interface (GRefine). Finally you will be able to export your modified data set. All three data processing tools can be used independently.

Please note: The workflow as well as the tutorial are a beta version, and are likely to have failures. Furthermore, we only explain a few functions here, while many new functions will be added and also made easier in the coming months. We hope you will still find it interesting to work with this workflow and not give up in despair. We arelooking forward to your feedback at support@biovel.eu*.*

6.2 Prepare input file

This is the data mobilisation part to generate the input as described earlier in the [input](#) sub-section. The sample input file used in this tutorial is available [here](#).

6.3 Upload your data

1. Start Taverna application
2. Start GoogleRefine application
3. File/Open Workflow/ load Data Cleaning Workflow
4. Press 'Run the workflow' (green arrow)
5. Choose Input file
6. Choose a sub-workflow (Fig 1a). You have the following options:

- Run synonym expansion and occurrence retrieval on your species list, go to 7
- Interact with the data in the spatio-temporal interface (BioSTIF), go to 11
- Interact with the data in the Data Quality selection interface (GoogleRefine), go to 16
- Export the data in excel format (.csv), go to 24
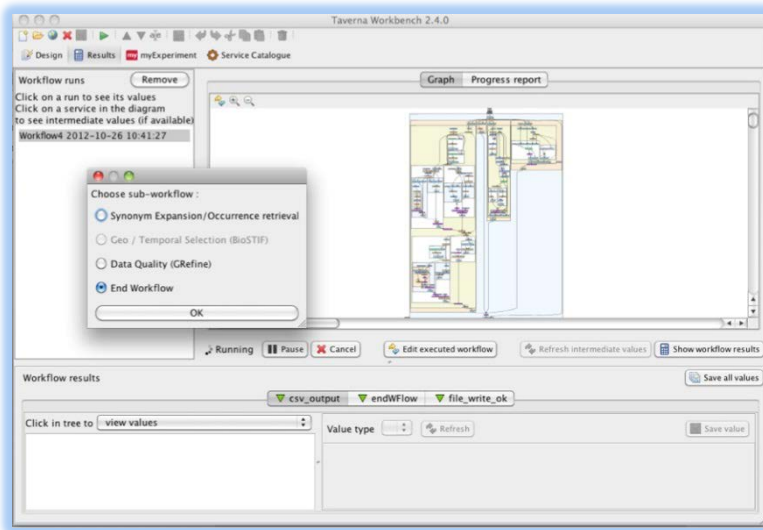
Fig 1a. Choose your sub-workflow.

6.4 **Synonym Expansion / Occurrence Retrieval**

7. Choose Synonym expansion / Occurrence retrieval and press OK (Fig. 1b).
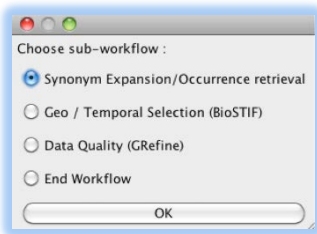
Fig 1b.

8. Choose a Taxonomic Checklist Bank (Fig 2). You will be prompted to accept the Data Use Agreement of the respective data repositories.
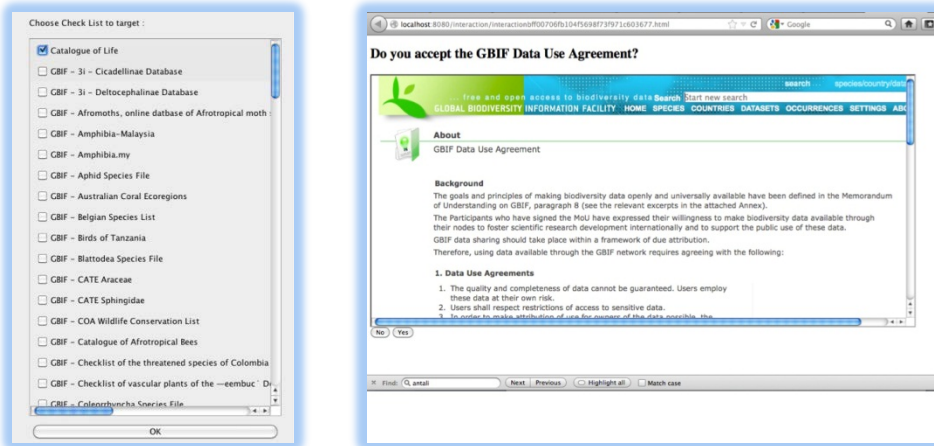
Fig 2.

9. Choose the synonyms for occurrence retrieval, both for each species and for each Taxonomic Checklist Bank (Fig 3). Press 'Find occurrences for these names', and thereafter choose and Data Provider. *Please note, at this point only GBIF works, while SLW (Swedish LifeWatch) data access is still in the testing phase*. You will be prompted to accept the Data Use Agreement of the respective data repositories.
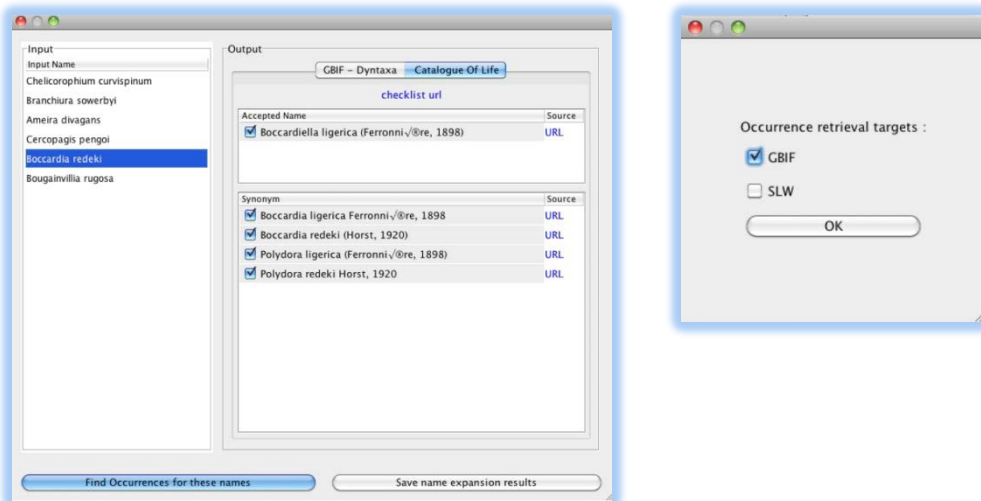
Fig 3.

10. Choose how to proceed with the retrieved data (Fig 4). You have the following options:
    - Interact with the data in the spatio-temporal interface (BioSTIF), go to 11
    - Interact with the data in the Data Quality / Cleaning interface (GoogleRefine), go to 16
    - Export the data in excel format (.csv), go to 24

6.5 **Geo-temporal Selection**

11. Choose Geo / Temporal Selection (BioSTIF) and press OK (Fig 4).
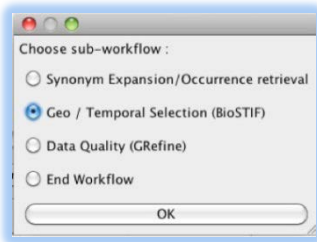
   Fig 4.

12. Make geographical refinements (Fig 5). Zoom and select areas, e.g. select all occurrences from Europe using the polygon function (Fig 5b, arrow 1), then press the filter button (Fig 5b, arrow 2). You can also select all records outside of the polygon by pressing the inverted filter button (Fig 5b, arrow 3).
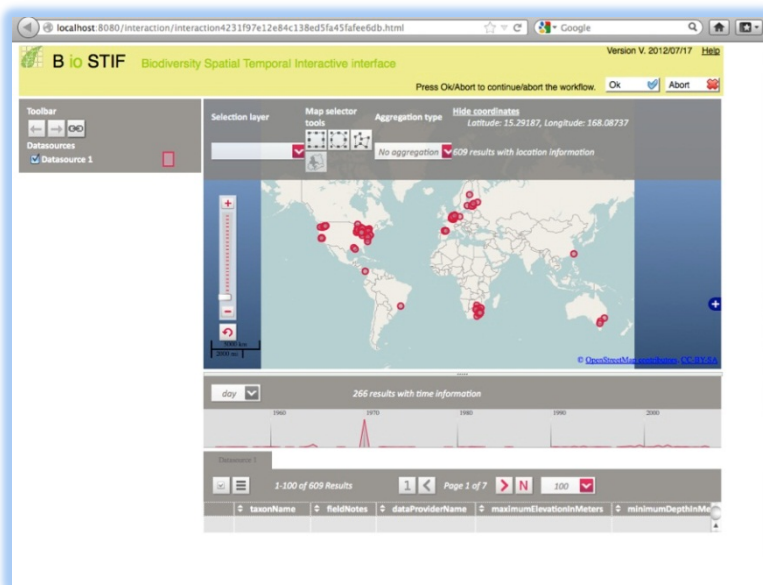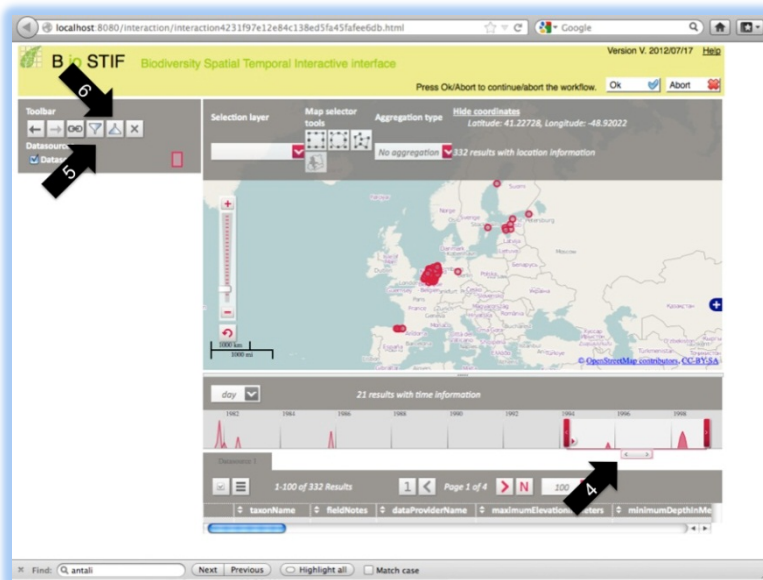

   Fig 5a.
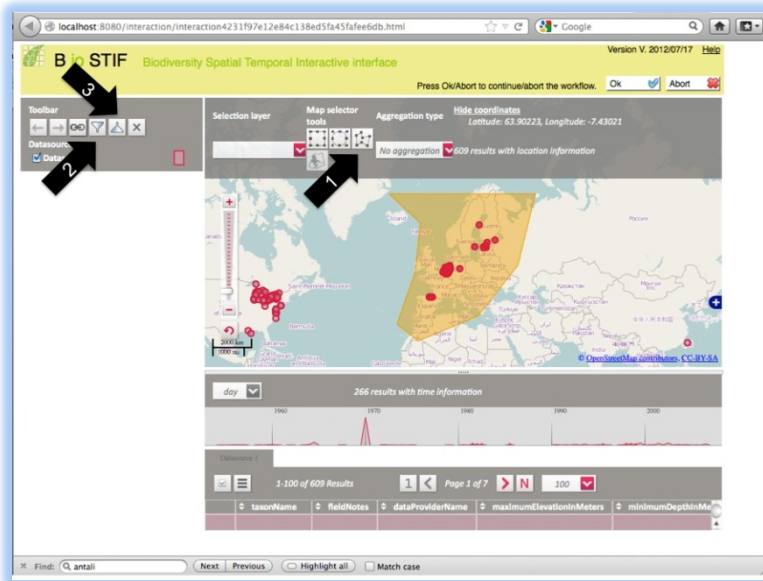
Fig 5b.

13. Make temporal refinements. Use your mouse button on the time chart to select records according to their age (Fig. 6, arrow 4). Once the time window is selected press the filter button (Fig 6, arrow 5). You can also select all records outside the time window by pressing the inverted filter button (Fig 6, arrow 6).

Fig 6.

14. Inspect records. You can activate single records by clicking on them. Then you will see detailed information displayed in a pop-up window (Fig 7a). You can also select multiple records, e.g. by using the polygon function (Fig 7b, arrow 7), then press the 'select all active elements' button (Fig 7b, arrow 8), and then inspect all information associated with the records (e.g. depth, altitude, collector, institute, habitat, etc) in the table widget (Fig 7b, arrow 9).
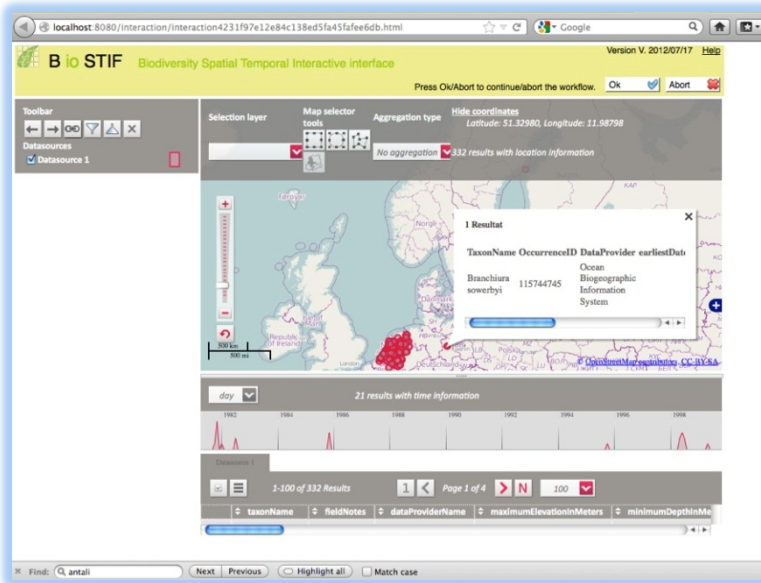
Fig 7a.
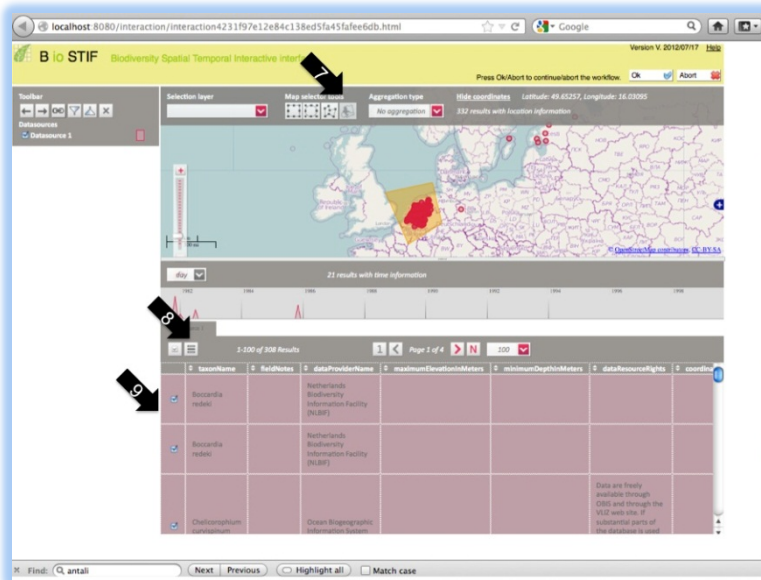
Fig 7b.

15. End the geospatial interaction with your data. Before you leave the BioSTIF interface you can use the chain button (Fig 8, arrow 10) to save (i.e. by copy/paste) this data page, and later access the same data in BioSTIF (without Taverna). If you are done selecting and inspecting your data press the OK button (Fig 8, arrow 11). You then have the following options:

- Interact with the data in the Data Quality selection interface (GoogleRefine), go to 16
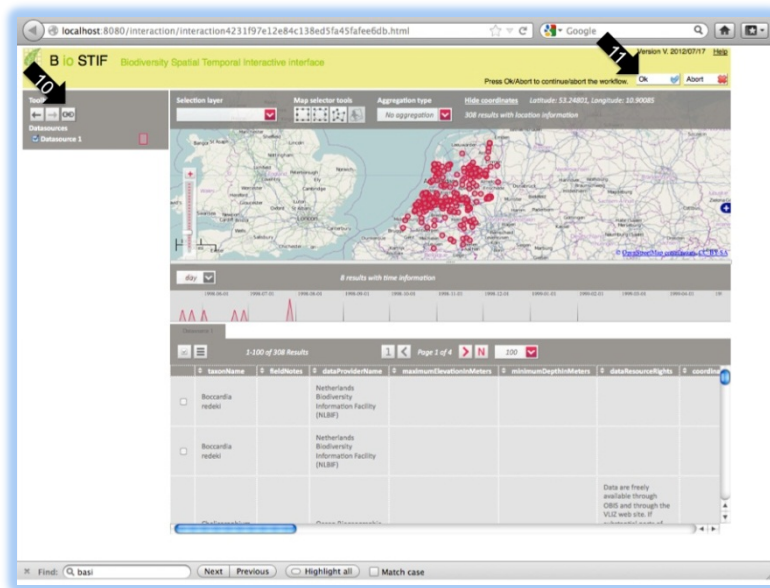- Export the data in excel format (.csv), go to 24

Fig 8.

6.6 Data Quality / Cleaning

16. Choose Data Quality interface (GoogleRefine) and press OK (Fig 9.).
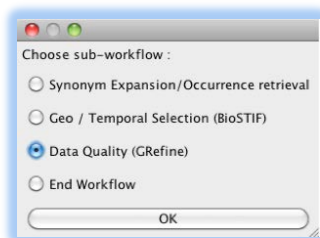
Fig.9.

17. In GoogleRefine you can trace and restore all changes you make to the data set by using the Undo/Redo function (Fig. 10, arrow 12). The size of the full data set is indicated as number of rows in you data matrix (Fig. 10, arrow 13). You can choose the number of your representative records displayed in the GoogleRefine interface (Fig. 10, arrow 14).
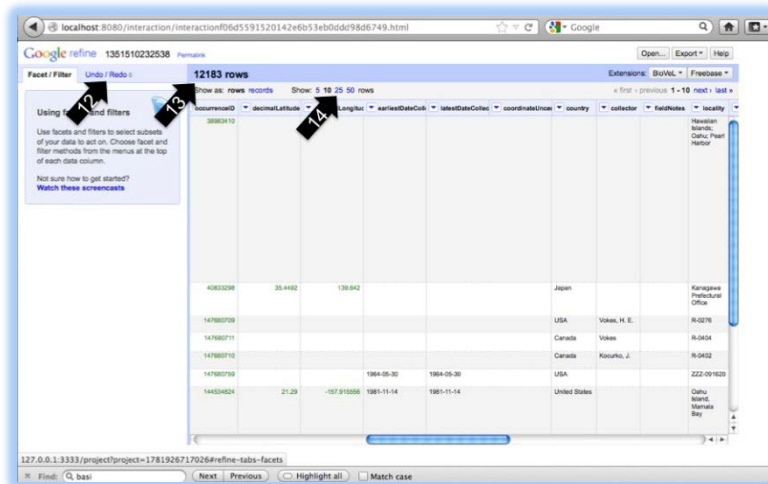

Fig 10.

18. Format dates. To transform all dates in your matrix to the same format go to 'Edit Columns/CommonTransforms/To date' (Fig. 11a, arrow 15). Use then the Transform option to format dates and geographic references (Fig. 11b, arrow 16). For the date change, paste the command *toString(value,"dd/MM/yyyy")* in the transform box of the date column (Fig. 11b, arrow 17). This will convert the date to dd/MM/yyyy format. So 23 May 2012 becomes 23/05/2012.
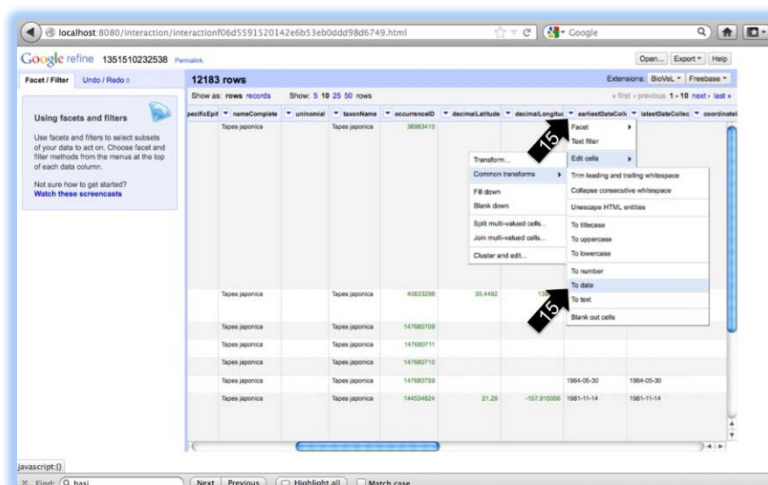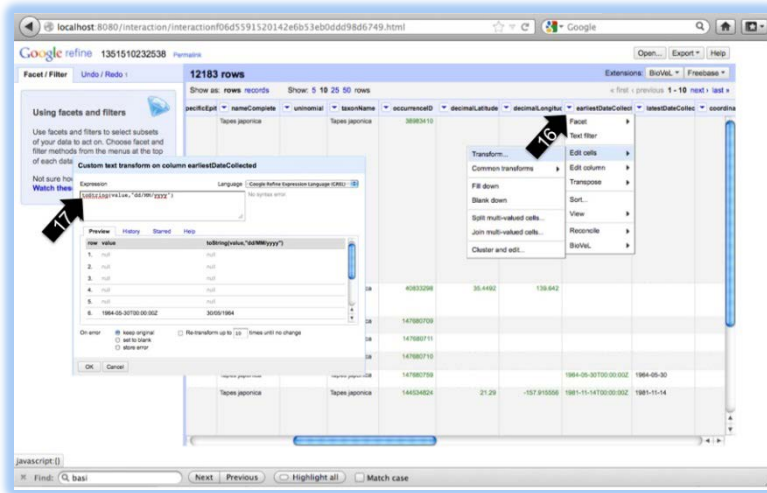
Fig 11a.

Fig 11b.

19. Formatting lattitude / longitude from gradual to decimal values. To transform your lattitude / longitude go to each of the georeference Column/EditCells/Transform (Fig. 11b, arrow 16) and paste in the window *replace(value,",",".")*. Thereafter repeat run the same step with the command *toNumber(value.split("°")[0])+ toNumber(value.split("°")[1])/60.* Thereafter go to each of the georeference Column/EditColumn/RenameThisColumn and replace the name with *decimalLatitude* or *decimalLongitude.* Then go to the Undo/Redo tab (Fig. 12, arrow 18), press Apply (Fig 12, arrow 19), and paste the contents of the file script_1_latlong_correction.txt (you find it in the example data folder) into the command box. Press Perform Operations.
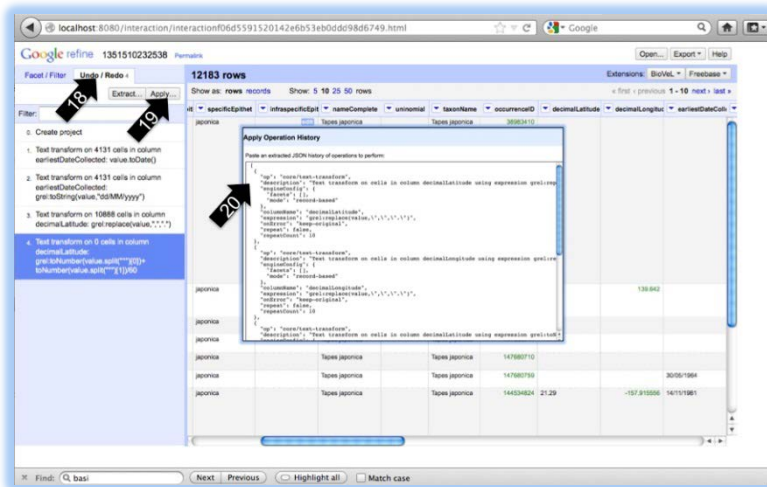
Fig 12.

20. Taxonomic Data Cleaning. Choose Name column/Facet/TextFacet (Fig 13a, arrow 21) to overview the frequency of each species in your data set (Fig. 13a, arrow 22). *Please note, this facet view can be applied to every column in your data set, e.g. habitat descriptions, taxonomic classifications, etc.* You can order the list according to name (Fig 13a, arrow 23) or frequency (Fig 13a, arrow 24). To find and correct spell errors, use the cluster option (Fig 13a, arrow 25). In the cluster mode, apply a range of cluster methods (Fig 13b, arrow 26) and Keying function (Fig 13b, arrow 27) to detect name clusters. Correct names by clicking on an underlined name in the list. The corrected name will then appear in the name window to the right (Fig. 13b, arrow 28).
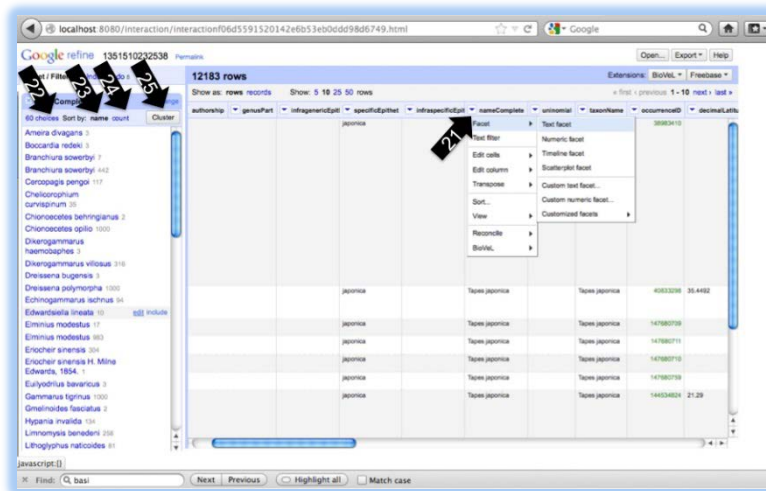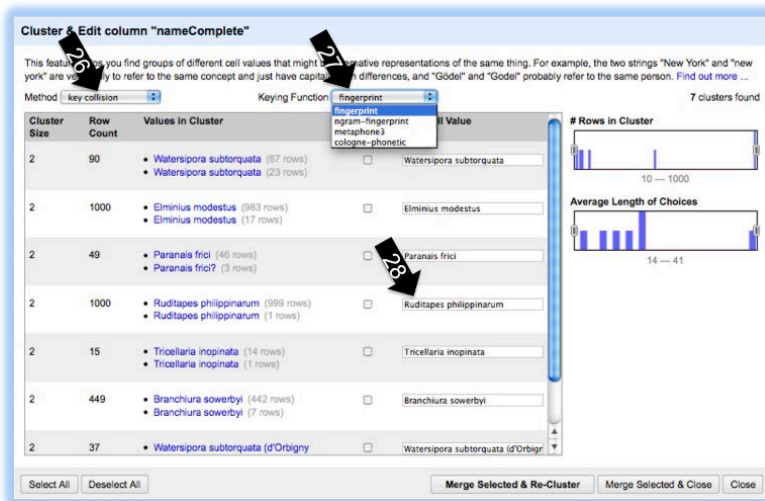

Fig 13a.

Fig 13b.

21. Taxonomic Data Refinement. Choose a taxon Column/Facet/TextFacet to overview the frequency of each taxon, e.g. family, class, phylum (Fig 14, arrow 29). Refine the taxonomic contents of your data set by selecting a taxonomic group (Fig 14, arrow 30), and then All/EditRows/RemoveAllMatchingRows (Fig 14, arrow 31).
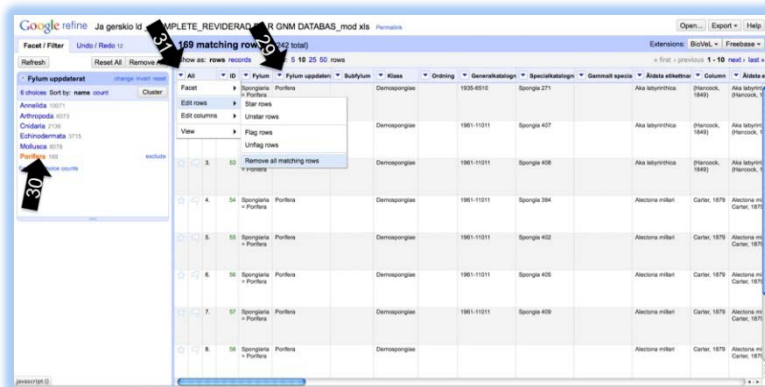

Fig 14. *Please note, this screenshot was produced with personal data because the example data did not contain taxonomic classifications of the species.*

22. Ecological refinement. Choose a habitat description or collection method Column/Facet/TextFacet to overview the frequency of each habitat or method. Refine the contents of your data set by selecting a habitat or collection method, and then All/EditRows/RemoveAllMatchingRows.

23. End the Data Quality selection. Choose the Extension button BioVeL/SaveToTaverna to export the data set shown in your GoogleRefine window (Fig 15, arrow 32). You then have the following options (Fig 16):
    - Interact with the data in the spatio-temporal interface (BioSTIF), go to 11
    - Export the data in excel format (.csv), go to 24
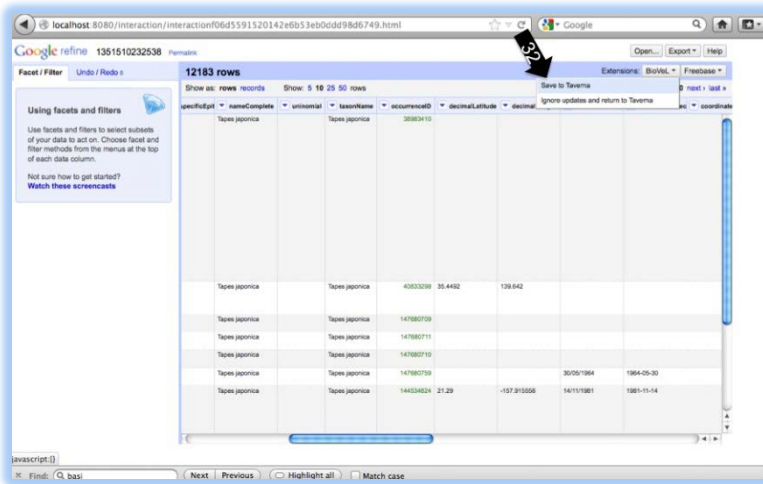

Fig 15.

6.7 Data Export

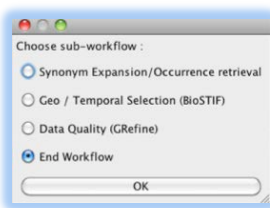24. Choose End Workflow and press OK (Fig 16)

Fig 16.

25. You will be prompted to select an output file, which has to be a text file in CSV format. You can also retrieve all output data directly from the Taverna workbench by clicking on value (Fig 17, arrow 33) and copy/paste the contents of the result window (Fig 17, arrow 34) into a text file.

BioVel Documentation

Taxonomic Data Refinement (Integrated) Workflow

Fig 17.

7 Support

For support and help with this workflow, please write to support@biovel.eu

For definitions of technical and biological terms, please visit the BioVeL glossary page:
http://www.biovel.eu/index.php/support/glossary

8   References

8.1   Acknowledgements

8.1.1 Workflow design and implementation

Workflow Design / Development : Cherian Mathew[1], Vera Hernandez Ernst [2], Anton Güntsch[1]
Scientific Feedback / Testing : Matthias Obst[3], Sarah Bourlat[3]

[1]Botanic Garden & Botanical Museum Berlin-Dahlem, Königin-Luise-Str.6-8, 14195 Berlin, Germany

[2]Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS, Schloss Birlinghoven, 53754 Sankt Augustin, Germany

[3]Göteborg University, Department of Biological and Environmental Sciences, Box 463, SE-405 30 Göteborg, Sweden

8.1.2 Project funding

The workflow described in this documentation has been designed and implemented as part of the BioVeL project.

BioVeL is funded by the European Commission 7th Framework Programme (FP7) as part of its e-Infrastructures activity. Under FP7, the e-Infrastructures activity is part of the Research Infrastructures programme, funded under the FP7 'Capacities' Specific Programme. It focuses on the further development and evolution of the high-capacity and high-performance communication network (GÉANT), distributed computing infrastructures (grids and clouds), supercomputer infrastructures, simulation software, scientific data infrastructures, e-Science services as well as on the adoption of e-Infrastructures by user communities.