B. Mitschang et al. (Hrsg.): BTW 2017 – Workshopband,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 117

Interactive Data Exploration for Geoscience

Christian Beilschmidt[1], Johannes Drönner[1], Michael Mattig[1], Marco Schmidt[2], Christian Authmann[1], Aidin Niamir[2], Thomas Hickler[2,3], Bernhard Seeger[1]

Abstract: Data-driven research requires interactive systems supporting fast and intuitive data exploration. An important component is the user interface that facilitates this process. In biodiversity research, data is commonly of spatio-temporal nature. This poses unique opportunities for visual analytics approaches. In this paper we present the core concepts of the web-based front end of our VAT (Visualization, Analysis and Transformation) system, a distributed geo-processing application. We present the results of a user study and highlight unique features for the management of time and the generalization of data.

Keywords: Visualization, Biodiversity, Scientific Workflows

1 Introduction

Recently, research has become increasingly data-driven. Researchers often form new ideas by exploring large databases and identifying interesting patterns, instead of collecting data with a concrete hypothesis already in mind. Visual analytics plays an important role in this approach. It provides the necessary tools for researchers in order to facilitate effective data exploration. In biodiversity research a large fraction of the data is inherently spatio-temporal, i.e. the position of objects can be represented in a coordinate system at a certain point in time. This makes it especially appealing for a visual analytics approach, as spatial data can naturally be visualized on a map.

While data-driven research offers many new scientific opportunities, it also poses challenges for users regarding data integration, cleansing, filtering and lineage. First, there is a multitude of publicly available heterogeneous data which users want to combine, possibly also with their own data. There are different types of data, e.g. vector and raster data, and different reference systems for space and time. Second, data often appear as time series and computations need to take this into account in order to produce valid results. Third, the size of individual data sets poses challenges as high resolution raster images easily exceed hundreds of gigabytes. It is not feasible to store and process such data on regular desktop hardware using standard software. Fourth, specific subsets of data have often quality issues. An appropriate visualization has to support identifying errors and relevant subsets. Fifth, the flexible composition of workflows via a user-friendly interface makes it difficult

[1] University of Marburg, Dept. of Mathematics and Computer Science, Hans-Meerwein-Str., 35032 Marburg, {authmanc, beilschmidt, droenner, mattig, seeger}@mathematik.uni-marburg.de
[2] Senckenberg Biodiversity and Climate Research Centre (BiK-F), Senckenberganlage 25, 60325 Frankfurt am Main, {firstname.lastname}@senckenberg.de
[3] Department of Physical Geography, Goehte University, Altenhöferallee 1, 60438 Frankfurt am Main, Germany

for users to keep track of the data lineage. In particular, this poses challenges in correctly citing the source data. To the best of our knowledge, we are not aware of a single system addressing all these challenges for biodiversity science.

Our Visualization, Analysis and Transformation system (VAT) [Au15a; Au15b] aims to support such an interactive data exploration for biodiversity data. It consists of a back end for low-latency geo processing called MAPPING (Marburg's Analysis, Processing and Provenance of Information for Networked Geographics) and a web front end called WAVE (Workflow, Analysis and Visualization Editor). The main purpose of this system is to pre-process data and to export results for further analysis in custom tools. The user interface is of utmost importance in order to effectively enable data-driven science on large and heterogeneous data for scientific users with little background in information technology.

Hidden behind the user interface are so-called exploratory workflows representing a composition of atomic scientific tasks. WAVE creates these workflows on the basis of an intuitive user interface. Moreover, workflows can process data with provided building blocks for investigating different approaches. When users obtain meaningful results, all steps that led to their computation are available as workflows. These workflows can be stored, shared and adjusted for similar use cases on different data sets.

The VAT system is already in use in two ongoing projects. GFBio[4] [D+14] is a national data infrastructure for German biodiversity research projects. It offers an archive for long-term access in order to facilitate data sharing and re-usage. VAT provides added value services for exploring and processing the data sets of the GFBio archives. Idessa[5] deals with sustainable range-land management in South African savannas. Here, the VAT system is used as a toolbox for implementing a web-based decision support system for farmers to avoid land degradation.

The main contributions of this paper are: we present an interface for exploratory workflow creation, effective data generalization and previews, linked time series computations, and automatic provenance and citation tracking.

The rest of the paper is structured as follows. First, we discuss briefly in Section 2 the motivation of building a new system by shortly presenting other work in this area. Then, we introduce the design of the user interface of VAT and describe some aspects in more detail, like our mechanisms for data generalization. Section 4 presents an evaluation of our interface by discussing a user study. Finally, Section 5 concludes the paper.

2 Related Work

There is an ongoing scientific interest in interactive map applications [AAG03; Ro13]. This stresses the importance of immediate responses of operations. Current visual analytic approaches [S+13] follow this approach but lack the ability to track workflow provenance and modify configurations.

[4] www.gfbio.org
[5] www.idessa.org

Typically, data processing in geo sciences is either done using scripting languages like R or Geographic Information Systems (GIS) like QGIS[6]. Writing R programs requires knowledge of the language and the required packages. The development takes time and the processing speed is limited. GIS offer a graphical user interface that requires less programming skills. However, desktop GIS also suffer from slow processing as they are limited to local resources and do not exploit modern hardware sufficiently well [Au15b]. Workflow builders for GIS do not support an exploratory usage. To the best of our knowledge there exists no GIS that support time series as a core concept.

Web-based applications allow for ubiquitous access and are able to provide more processing power than desktop applications. There are specialized applications like Map Of Life [JMG12] that aim to solve specific use cases very well. More general functionality is provided by cloud-based GIS like CartoDB[7] and GIS Cloud[8]. However, the processing capabilities of these systems is still limited as they mainly focus on map creation rather than scientific processing tasks.

Workflow systems like Taverna [W+13], Kepler [A+04] and Pegasus [M+13] are building workflows upfront, with the goal of executing them on multiple data sets. This is contrary to our approach of exploratory workflows that are transparently built in the background during data exploration. Additionally, they do not allow web-based workflow creation, offer little geo functionality and limited processing throughput [Au15b].

3 WAVE: A Workflow, Analysis and Visualization Editor

This section describes our system's user interface. It starts with an overview of the fundamentals. Then, it discusses different concepts for solving the challenges introduced in the Introduction. This covers methods for data generalization and temporal support as well as project and citation management.

3.1 Basic Principles

The fundamental idea for WAVE is to offer an intuitive web-based user interface for interactive exploration of biodiversity data. Users can select data from a data repository and upload custom files. They perform operations for filtering or enriching data by combining them with other sources of information. Finally, they export data for further analysis in their preferred custom tools on a different system. The complete workflow of computations is always accessible and allows a reproduction of results anytime later.

Biodiversity data is mostly spatio-temporal. The system thus supports temporal collections of data from one of the following types: points, lines, polygons and rasters. Examples for such data are occurrences of species as points, roads as lines, country borders as polygons

[6] www.qgis.com
[7] www.carto.com
[8] www.giscloud.com

and environmental variables like elevation as rasters. The system treats every collection as a time series. Points, lines and polygons are all homogeneous object collections. Every object has an associated time interval expressing its validity. In a raster every cell shares the same temporal validity.

Computations on the available data are specified as workflows. They describe a dataflow from source collections over processing operators to a final result. A JSON representation makes the workflows storable and shareable. WAVE builds a workflow on-the-fly in the background when a user performs actions on selected data. It thus keeps track of all the applied processing steps. A query consists of a workflow, a spatial window and a reference time that selects the temporal valid objects. Queries are processed on MAPPING, the back end of the VAT system.

The central part of the data visualization is a panel with a map that consists of multiple layers of geographic data and shows them in their spatial and temporal context. The map supports different projections and allows for panning and zooming. It is linked to a data table that contains further information about non-spatial attributes. All data, either from sources or results of computations, are represented as separate layers. Users can also specify the order in which layers are drawn on the map. Layers refer to a workflow that is part of the query processed in MAPPING. They also serve as inputs for operators in order to create a new workflow (c.f. the operator list above the map panel in Figure 1 for examples). Beside layers, a workflow can also output plots, e.g. histograms or scatter-plots.

Layers and plots are linked to a component that allows the selection of the temporal reference. Here, users specify the point in time that is transferred to MAPPING as part of the spatio-temporal context. A change of the temporal reference triggers a re-computation of all layers and plots. WAVE supports a video mode for which the user specifies a time interval. Then, the computation slides over the interval, continuously producing the outputs of layers and plots. This effectively visualizes the changes in the data over time.

Adding a new layer consists of either selecting data from a source or applying an operator on one or more existing layers. In order to allow users to easily combine their own data with important environmental information, WAVE offers an interface to access a repository of raster and vector data hosted by MAPPING. In addition, users may upload their own data represented in the popular CSV data format. Operators allow the selection of appropriate data sets as inputs. One of the benefits of WAVE is that there is a check whether inputs fit to operators. If not, it tries to transform the data, e.g. adapt to the coordinate system, to make it compatible. By automatically applying such transformations, users can easily integrate heterogeneous data sets in their scientific workflows.

Figure 1 shows a map of an application containing three layers and two plots. Here, import dialogues were used to gather repository and custom data of African and Forest Elephants. The user applied filters for cleansing the vector occurrence points with a polygon and removed outliers by applying numeric range filters. Additionally, combination operators added environmental data from rasters to the vector occurrence points. There is support to create statistics by either using built-in operators or calling user-defined R scripts. It is

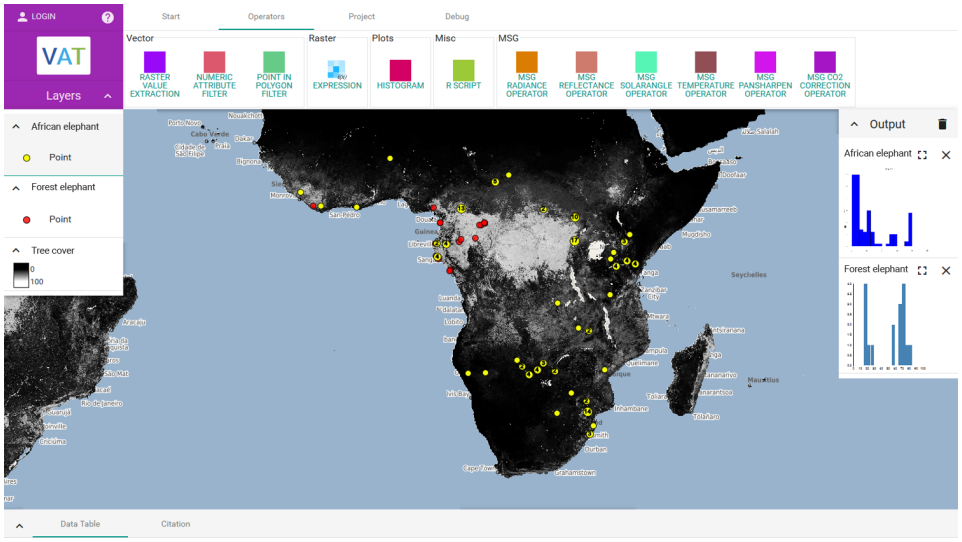

Fig. 1: Screenshot of the use case result

possible to export the output of our workflows as files. A more convenient approach is our R package that supports calling the workflows directly from a user program.

Users can work in multiple projects, each of them offers its specific data sets, workflows, layers and plots. The auto-save feature of WAVE ensures that projects are always up-to-date. A flexible rights management allows sharing projects within a team. In addition, projects can be published to other users with the option to restrict the right of changing the workflows and other aspects.

3.2 Data Generalization

There are two major objectives when interactively visualizing data for the user: One is to compute results in a near-realtime fashion for the user's exploration experience. The other is to provide an abstraction of the data that facilitates detecting interesting patterns. Data generalization can address both concerns.

The generalization of raster data is possible by aggregating multiple adjacent cells and representing them in a lower resolution. This requires less storage but comes at the expense of losing information. However, the amount of visible cells is naturally limited by the amount of pixels on the user's screen. Thus, it is sufficient to output raster images in this resolution for previews. Moreover, it is also sufficient to use source rasters and intermediate results of queries in this resolution instead of restricting the aggregation only to the results. This allows us to compute preview results with low latency. Users can afterwards trigger the computation in full resolution to produce scientifically valid results. In addition, the user can increase the accuracy of the data processing and the data visualization incrementally by simply zooming into interesting areas.

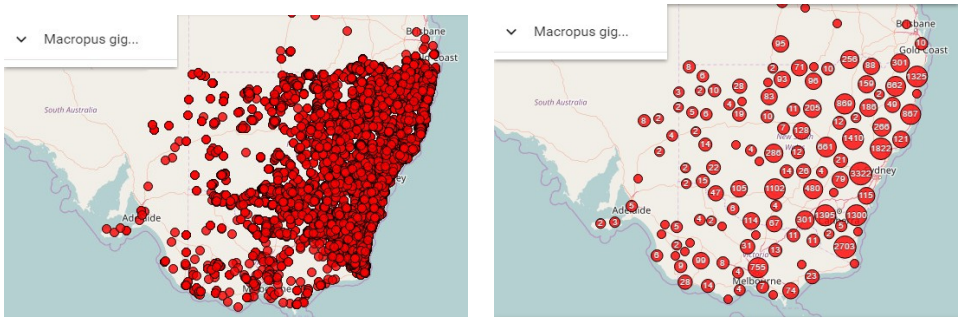

Fig. 2: Clustering a set of kangaroo observations in Australia to improve density information

For the generalization of vector data, a first popular approach is to transform vector data into a raster, but this goes along with a loss of attribute information. A second approach is to apply simplification techniques. Beside of being expensive, this also causes semantic changes of queries. We will examine these techniques in our future work.

WAVE offers an approach to generalizing big point sets to speed up their visualization and to identify cluster patterns. Displaying each point with its associated attributes exceeds the capabilities of current browsers on modern hardware even for sizes of less than one million points. Additionally, the size of transferring the data in the GeoJSON standard format stresses the internet connection of mobile devices. An example is 23,039 kangaroo (Macropus giganteus) occurrence points from GBIF[9], c.f. the left hand side of Figure 2. The uncompressed size with 20 common attributes is ~15 MB even for this relatively small data set. Furthermore, it is hard to recognize in the original plot that there is a very dense population of kangaroos in the south of Canberra. WAVE uses an adapted tree implementation of the hierarchical method developed by Jänicke, et al. [Jä12] to cluster data for the purpose of visualization. This allows combining nearby data points dependent on the zoom level and map resolution. By zooming in, the user gets a smaller excerpt of the map in more detail such that clusters break up and more details reveal. We represent the clusters as circles with logarithmically scaled area based on the number of included points. Additionally, the circles contain the number of points as labels. As circles are non-overlapping it is easy to identify clusters, c.f. the right hand side of Figure 2.

The tabular view is linked to the map and reacts to changes that occur in the layer as well. The view only has the possibility to present the clustered point data. Otherwise we would have to transfer all data which conflicts with data compression. Therefore, the table shows exactly the same data points as the map in the same resolution. The basic idea is to keep the attributes of the original data and to report an aggregate derived from the data in the cluster. For numeric attributes, we use the mean and standard deviation that can be computed in linear time. By zooming in, the amount of points in a cluster decreases and so does the standard deviation. This means the information gets more exact by diving into the data. For textual attributes, WAVE keeps a small number of representative points (typically three to five) for each cluster. Among the many options for selecting representative points,

[9] Global Biodiversity Information Facility: www.gbif.org

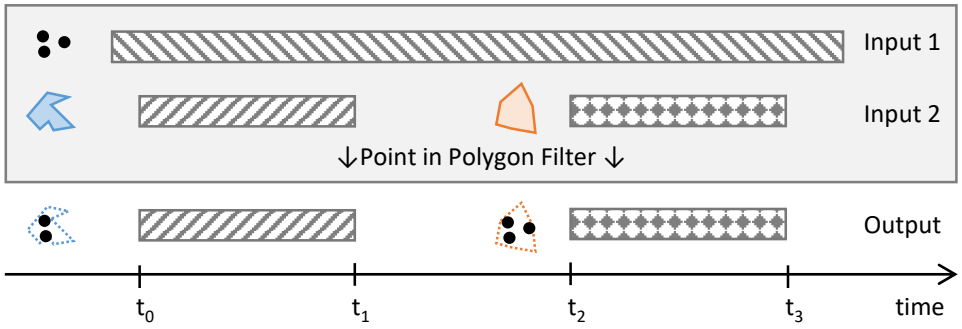

Fig. 3: Temporal point-in-polygon filter of three points and two polygons

we decided to use the points closest to the cluster center. The reason for our choice is that this information improves in accuracy when zooming in.

3.3 Support for Temporal Operations

Support for operations on time-series is a unique feature of the VAT system in comparison to other GIS-like systems. Our system supports the definition of a date and time as a global temporal reference. Recall that each query consists of a workflow with a spatio-temporal context, in particular a time interval that expresses the validity. The temporal reference slices a time series result such that it only contains elements that are valid at the given point in time.

For example, a user may add the WorldClim[10] mean annual temperature data set as a raster layer to a project. This data set is a time series that contains monthly climate variables. By means of the temporal reference the system is now able to choose the valid raster for the current month from the time series and add it to the map. Consequently, operations that include this data set also incorporate the correct raster with respect to the temporal reference. In comparison, traditional GIS oblige a user to manually add the correct raster from the data set beforehand. This is obviously a cumbersome and error-prone task.

The temporal validity of a data object is defined as an interval from a start time to an end time in which the object is incorporated into computations. When data with different validities are combined, data objects may have to be split into multiple items with different validities. Figure 3 shows an example of a point-in-polygon filter, where the output is a time series with two separate objects due to different validities of the polygons.

A problem arises when users want to combine data with non-overlapping temporal validity. One example is to compare measurements from today with measurements from the same day of the last year. In order to support this important type of temporal operations, WAVE offers a temporal shift operator to change the temporal validity of objects. A shift can either be absolute or relative. It is applied to a query first before any other processing

[10] www.worldclim.org

takes place. The result has to adapt its temporal validity to the temporal operation. In the previous example, after retrieving the measurements from last year, we have to set their validity to the current year in order to be able to compare them with each other.

When a user changes the temporal reference, WAVE triggers a re-computation of all views. Thus, there will be an update of the map, the layers, the connected table and the associated plots. The incorporation of temporal functionality in the user interface is still object of our future research. We currently only allow the specification of the temporal reference. Harmonizing the validity of different data sets is not yet possible. We will perform a user study in order to find an appropriate and intuitive way to extend our user interface for such kind of operations.

3.4 Provenance Information

Citations are very important for scientific work. They allow researchers to classify, comprehend and reproduce published results. They are also important for the publishers of those results, as they facilitate assessing the impact of their work. Aside from scientific results, also raw data has to be properly cited for the above reasons. Today, papers on topics related to data are encouraged by organizations and journals to facilitate data sharing. A recent article [BDF16] stressed the importance of citations in scientific data management.

Our system automatically keeps track of the citations of the involved data sets. We call the combination of (1) citation, (2) license and (3) a URI (e.g. link to a landing page) provenance information. All source operators are responsible for collecting the provenance information for the outputs, given a specific input. Processing operators combine the provenance information of their inputs via a duplicate eliminating union operation. This behavior can however be altered for specific operators.

The provenance information for a layer is always accessible by the user in WAVE. When the user exports a layer, a zip-archive is created. In addition to the actual data it contains two files. One file contains the workflow representing the computation of the layer, including all applied transformations. The other contains all the provenance information of the data sets involved in the computation.

4 Evaluation

We conducted a user study to gain insights about an appropriate user interface design beforehand to the product development. This included creating several use cases together with domain experts in biodiversity research to (1) create realistic scenarios and (2) cover as many concepts of WAVE as possible. We created a paper prototype to try out different interface variations. This allowed us stepping back from any implementation details and focusing on concepts on a sketch board. The advantage was that it is very inexpensive to discard doubtful concepts. And in conclusion this led to rapid concept development with domain experts.

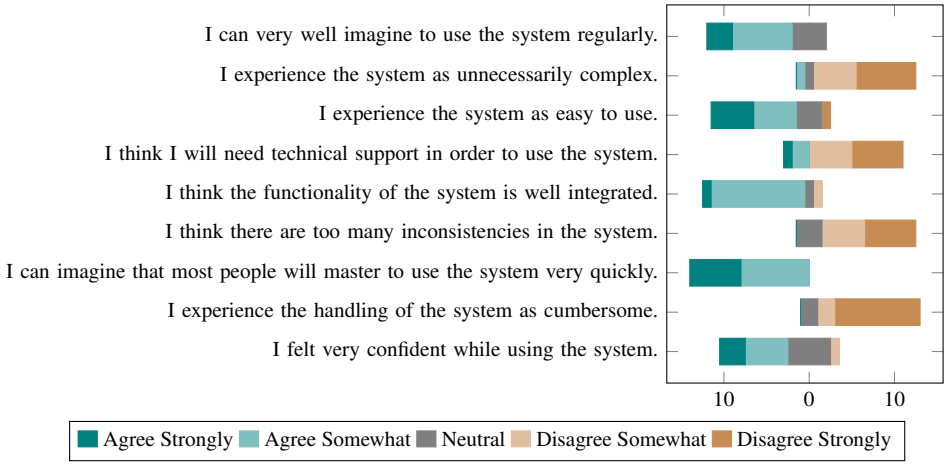

Fig. 4: Results of the use case of the user study regarding the paper prototype

The user study consisted of two parts. The first was an introduction of a use case and a 20 minute time span to solve a specific task. The users should work on the task independently without any system introduction or explanation. We observed the behavior and timed steps of certain sub tasks. The second part included a questionnaire of nine fixed questions and an additional field for free text comments. The participants had ten minutes time for this feedback. The questions aimed at different impressions about the system usage. As answers we used a symmetric typical five-level Likert scale with a neutral element.

We asked 15 users from the biodiversity domain at the Senckenberg Biodiversity and Climate Research Institute (BiK-F) in Frankfurt am Main to process our use case. Figure 4 shows the result of the study. Together with the additional comments (which we excluded here for space reasons) we did not find any reason for major changes in our design proposal. Nevertheless, we identified minor weaknesses and were able to get a better understanding of how users work with our system. One interesting fact to mention was the expectation of the users to interact with the application like in desktop GIS. This included right-clicking on elements to perform actions. This was a strong contrast to our previous experience in web application development.

5 Conclusion

We presented the design process, the fundamental concept and selected features of WAVE as the front end of the VAT system, a distributed system for supporting scientific geo applications. We conducted a user study to validate the design of the user interface. Several unique features including the support for temporal data, approaches for generalizing data, the integration of R scripts and the automatic management of workflows and citations distinguish WAVE and the VAT system from other solutions.

In our future work we will explore the possibilities of extending the system into an application builder. Here, users will be able to create custom applications using built-in functionality. For example, they create a workflow to solve a parametrized problem. Other users can then make use of this workflow in a custom user interface where they specify the required parameters without having the need to know the internals of the processing chain.

References

[A+04] Altintas, I.; Berkley, C.; Jaeger, E., et al.: Kepler: An Extensible System for Design and Execution of Scientific Workflows, In: 16th Int. Conf. on Scientific and Statistical Database Management, 2004. Proceedings. 2004, pp. 423–424.

[AAG03] Andrienko, N.; Andrienko, G.; Gatalsky, P.: Exploratory spatio-temporal visualization: an analytical review, Journal of Visual Languages & Computing 14/6, pp. 503–541, 2003.

[Au15a] Authmann, C.; Beilschmidt, C.; Drönner, J.; Mattig, M.; Seeger, B.: Rethinking Spatial Processing in Data-Intensive Science, In: Datenbanksysteme für Business, Technologie und Web (BTW) - Workshopband, 2015, pp. 161–170.

[Au15b] Authmann, C.; Beilschmidt, C.; Drönner, J.; Mattig, M.; Seeger, B.: VAT: A System for Visualizing, Analyzing and Transforming Spatial Data in Science, Datenbank-Spektrum 15/3, pp. 175–184, 2015.

[BDF16] Buneman, P.; Davidson, S.; Frew, J.: Why Data Citation is a Computational Problem, Commun. ACM 59/9, pp. 50–57, Aug. 2016, ISSN: 0001-0782.

[D+14] Diepenbroek, M.; Glöckner, F.; Grobe, P., et al.: Towards an Integrated Biodiversity and Ecological Research Data Management and Archiving Platform: The German Federation for the Curation of Biological Data (GFBio), In: GI: Informatik 2014 – Big Data Komplexität meistern, 2014.

[Jä12] Jänicke, S.; Heine, C.; Stockmann, R.; Scheuermann, G.: Comparative Visualization of Geospatial-Temporal Data. In: GRAPP/IVAPP, 2012, pp. 613–625.

[JMG12] Jetz, W.; McPherson, J. M.; Guralnick, R. P.: Integrating biodiversity distribution knowledge: toward a global map of life, Trends in Ecology & Evolution 27/3, pp. 151–159, 2012.

[M+13] McLennan, M.; Clark, S.; Deelman, E., et al.: Bringing Scientific Workflow to the Masses via Pegasus and HUBzero, In: Proceedings of the 5th International Workshop on Science Gateways, 2013, p. 14.

[Ro13] Roth, R. E.: Interactive maps: What we know and what we need to know, Journal of Spatial Information Science 2013/6, pp. 59–115, 2013.

[S+13] Steed, C. A.; Ricciuto, D. M.; Shipman, G., et al.: Big Data Visual Analytics for Exploratory Earth System Simulation Analysis, In: Computers & Geosciences, vol. 61, Elsevier, 2013, pp. 71–82.

[W+13] Wolstencroft, K.; Haines, R.; Fellows, D., et al.: The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud, In: Oxford Univ Press, 2013.