VAT: A System for Data-Driven Biodiversity Research

Christian Beilschmidt Johannes Drönner Michael Mattig Bernhard Seeger

Dept. of Mathematics and Computer Science
University of Marburg, Germany
{beilschmidt, droenner, mattig, seeger}@mathematik.uni-marburg.de

ABSTRACT

Visual analytics plays a leading role in data-driven research. This requires systems for fast and intuitive data exploration. In this paper we demonstrate VAT, a system for Visualizing, Analyzing and Transforming spatio-temporal data. The system consists of a distributed back end for low-latency processing and a web front end that allows creating workflows of computations in an exploratory fashion. A novel quality of the system is the combination of scientific processing while simultaneously tracking the provenance of the data and aggregating a list of data citations. These features make a visual analytics approach for large, heterogeneous spatio-temporal data feasible.

CCS Concepts

•Information systems → Geographic information systems; *Data analytics; Information integration;* **•Human-centered computing → Visualization;**

Keywords

Scientific Workflows, Provenance, Interactive Analysis

1. INTRODUCTION

Visual analytics plays a leading role in data-driven research. Especially in geoscience, researchers investigate spatio-temporal data by means of interactive exploration. As data sizes increase rapidly, there is a growing demand for exploratory tools with fast response time. To gain insights from the data, researchers want to examine different research ideas by expressing queries and evaluating the delivered results. However, there are multiple challenges for a scientist as detailed in the following.

Data from the geoscience domain is inherently heterogeneous. There are several data formats for vector and raster data as well as different reference systems for space and time. This makes data integration and data correlation a very cumbersome and time-consuming tasks for researchers even before tackling the actual research problem.

Scientific work itself and also the recent trend of journals to encourage data sharing make correct citations of data sources indispensable. Furthermore, it is necessary to ensure validity and reproducibility of computations. Both tasks become hard to accomplish when working in an exploratory fashion. As new ideas for additional data processing steps arise mostly when reviewing intermediate results, an upfront specification of the whole computation is not feasible. Recreating the computation steps afterwards is laborious and error-prone. To solve this, a system that offers scientific data processing should keep track of the whole path of processing steps also known as workflows. In addition, all references of incorporated source data should be aggregated as a list of citations.

To cope with these challenges, we demonstrate the VAT system in this paper, a system for Visualizing, Analyzing and Transforming spatio-temporal data in biodiversity science. It facilitates interactive data exploration and cleansing by creating and executing so-called exploratory workflows. For this, it offers processing building blocks for filtering, transforming, visualizing, and creating statistics. It enables users to join heterogeneous data and to work with time-series. They can (1) visualize data, (2) export data in consolidated formats for further analysis in custom tools, and (3) share reproducible workflows.

The VAT system is already in use in GFBio, a German national infrastructure project for managing, archiving, and providing access to biodiversity data [5]. The project aims to provide a sustainable service architecture for German research projects in biodiversity science. VAT enables researchers to identify interesting scientific topics, geographic regions and time spans by providing added value services for data visualization and analysis. It furthermore aims to facilitate reproducibility and data re-usage.

In the following, Section 2 gives an overview of the VAT system's architecture. Section 3 summarizes the functionality of the system and Section 4 describes the proposed demonstration scenario. Section 5 concludes the paper.

2. ARCHITECTURE

VAT has a client-server architecture that consists of a back end called MAPPING and a web-based front end WAVE. (c.f. Figure 1). The complete architecture is described in more detail in our previous work [1, 2].

MAPPING (Marburg's Analysis, Processing, and Provenance of Information for Networked Geographics) is a dis-

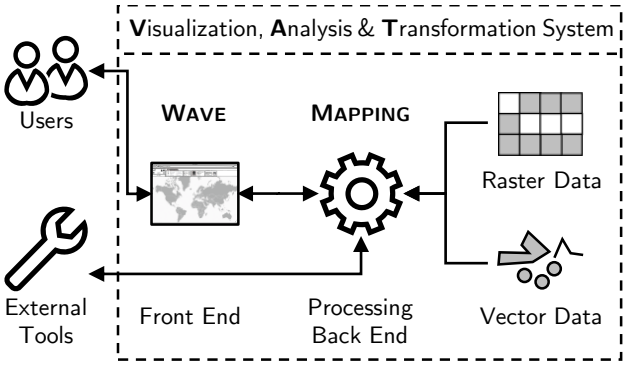

Figure 1: A condensed view on the system architecture

tributed scientific workflow processing system for low-latency processing of spatio-temporal data. It includes a workflow processing engine and various operators written in C++. For the sake of performance improvement, Mapping utilizes OpenCL to massively parallelize parts of the processing on the GPU. It also manages heterogeneous data and allows processing of raster as well as vector data. For easy access, Mapping implements important parts of the standardized OGC[1] protocols. This allows many tools to access computation results via a standard interface.

Wave (Workflow, Analysis and Visualization Editor) is an interactive web application for visual analytics and data cleansing which creates exploratory workflows [3]. It offers a reactive user interface where users apply actions on data via operators and review the results. The interface builds up on Angular 2[2] and OpenLayers 3[3], and uses an implementation of Google's Material Design components to offer appealing and touch-compatible control elements for desktop and mobile usage.

3. FUNCTIONALITY

The main scope of Vat is to support visualizing and processing collections of geo objects. These are points, lines, polygons and rasters. Examples for these data types are species occurrences for points, rivers for lines, forest regions for polygons and temperature grids for rasters. Additionally, each object has a temporal validity.

3.1 Exploring Data

Wave provides data visualization and interactive data exploration by applying operators. Operators fall into three categories: There are source operators that allow including data either from a repository of hosted environmental raster data and species related vector data, or from custom CSV files. Then, there are operators for filtering, combining and transforming data, e.g. attribute or point-in-polygon filters. Finally, there are statistics operators that allow creating figures like scatter plots and histograms. Additionally, the user can incorporate R scripts to extend the statistics functionality. The first two operator categories produce object collections that are presented in the form of layers on a map and a data table (arranged above each other, c.f. Figure 2). Wave presents the results of the latter as plots on a sidebar of the application.

For displaying large object collections, Vat offers data reduction techniques by compressing raster and vector data. It computes raster images in preview resolutions to reduce response times. As there is only a limited amount of pixels available on the user's screen, the loss of accuracy has no impact on the visualization. Vat uses a visual clustering approach (an adaptation of the method presented by Jänicke, et al. [6]) to reduce the amount of point data to be transferred and visualized. This technique facilitates recognizing the density of the data objects on the map. The data table shows aggregates of theses clusters for non-spatial attributes. Both techniques provide more accurate results when processing data of smaller areas. This means, zooming into interesting data reveals more detailed information and is therefore the intended exploration method. In the end, for scientificly valid results, Vat offers the functionality to compute the whole workflow in full resolution.

3.2 Combining Heterogeneous Data

In Vat, each object of a collection has three components: a spatial reference, a temporal reference and attributes. The spatial reference specifies a geometric object (e.g. a point) that corresponds to a coordinate reference system. The temporal reference specifies an interval from start to end time using a reference system (e.g. the Gregorian Calendar). The list of attributes contains different data types (e.g. strings or floating points). The spatial and temporal reference system is uniform for all objects within a collection. Because of the presence of temporal references in each object collection, a collection is considered as a time series. In a raster time series, each grid of cells has the same spatial and temporal reference.

To join object collections, the data needs to be in a unified reference systems. Mapping offers operators to transform data of one reference system into another. While this is usually very cumbersome for the user, Wave automatically applies these transformations whenever necessary. This makes it easy to apply operators to join initially heterogeneous object collections. Additionally, Wave suggests and restricts valid operator inputs (e.g. users can only select points and polygons for a point-in-polygon check).

Every combination operator has to consider the time series semantics. An example is the combination of a raster time series of monthly temperatures with point data (which have irregular time intervals). An input point has to be split into multiple points with different time intervals, if and only if it overlaps at least the end of one month. For instance, if an input point is valid from the first of January to the end of February, it will result in one point with temperatures from January and one from February.

Another example is to compare the temperature of the current date with the temperature of the same day of the previous year. For this, Mapping offers temporal operators for shifting the temporal context. By shifting relatively one year to the past, Vat can compute a difference expression on a single raster time series.

Wave uses a user-defined point in time to visualize the data of a time series. It uses it to select a time-slice of the series and retrieve only the data objects with matching validity. When a user changes the time, Wave triggers an update for each view, i.e. the map, the data table and the

[1]Open Geospatial Consortium, www.opengeospatial.org
[2]www.angular.io
[3]www.openlayers.org

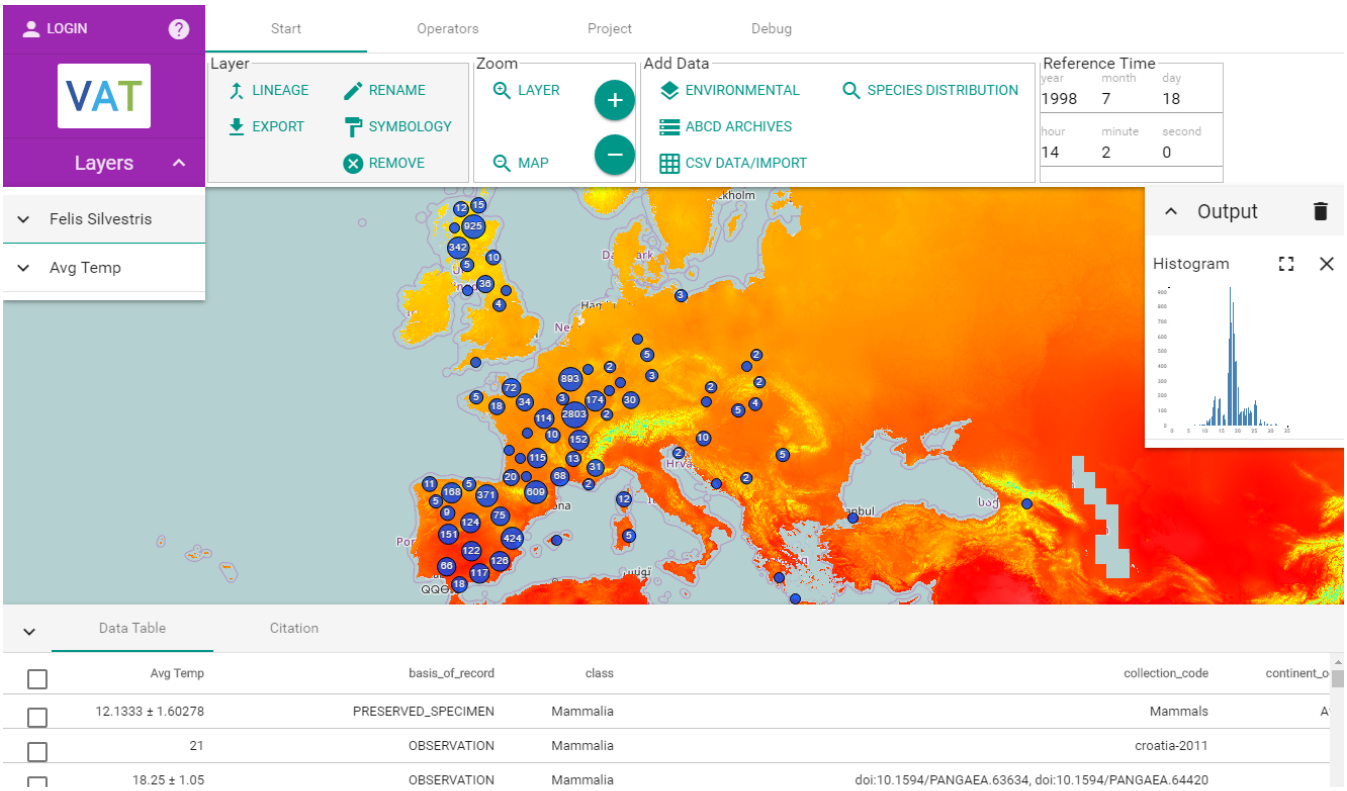

Figure 2: An overview of WAVE. The central map component shows clustered point data and a raster. Below is the data table. On top is a menu bar for data and operator selection. On the left-hand side is a list of map layers. On the right-hand side is a plot area containing a histogram.

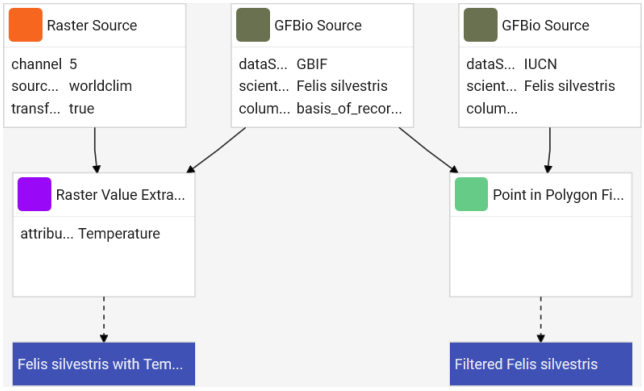

Figure 3: A lineage graph for a small workflow

plots. A video mode for uniform time steps of an interval (e.g. monthly) allows visualizing the changes over time.

3.3 Provenance Tracking

Provenance tracking is of utmost importance for the reproducibility of computations. WAVE allows data exploration by interactively applying operators on data. Users can utilize different views to evaluate results of computations. They can form new ideas and discard dead ends. We use the notion of exploratory workflows that describe the path of computations from the sources to a final result as a rooted tree. WAVE automatically updates corresponding workflows on every user action. The user is able to look up the full processing path at any time in the so-called lineage graph. Figure 3 shows an example of a series of applied operations in an exploratory workflow. The data flows from the source operators at the top to the resulting layers (blue boxes). VAT uses a workflow representation in the human-readable and interchangeable JSON data format.

The workflow data structure makes it furthermore possible to share computations and results with other researchers. This means either publishing fixed parametrized workflow results for reproducibility reasons or configurable workflows that allow other researchers to use validated workflows for their own data processing.

3.4 Collecting Citations

Correctly citing all sources of a workflow is indispensable for scientific work. While this task is complex in generic scenarios when querying database systems [4], VAT can take advantage of the custom implementation of each operator. In VAT, the tracking of citations is an inherent part of every operator's implementation.

More precisely, there is a default method that applies a duplicate eliminating union operator to the citations of the input operators. Certain operators, for instance source operators with filtering option, can change this behavior to only include citations for selected data. However, it is essential to never remove any citation of a data object that is incorporated in generating a result. An example could be a point data subtraction, where a data object is responsible

for removing another object and needs therefore to be included in the citation list. MAPPING's workflow framework guarantees this restriction.

3.5 Data Export

When exporting data sets for further usage, VAT bundles a ZIP file containing three components. The first is the result of the workflow in a raster (GeoTIFF) or vector format (CSV or GeoJSON), computed in full resolution. The second is the workflow description itself, containing the parametrization of each operator. The third one is a complete list of aggregated citations. VAT allows several metadata formats for workflow descriptions and citations like CSV or JSON.

4. DEMONSTRATION SCENARIO

In this section, we present two real-world scenarios that exemplify working with exploratory workflows in VAT. Because of the brevity of this paper we will show only the success case. Of course, one can easily imagine that the user took many wrong turns in order to achieve the result. Because of the automatic tracking of the workflow, the user can always trace back the steps.

4.1 Data Cleansing

The user is interested in the distribution of animals of the cat family, e.g. Felis silvestris (wildcat) from GBIF[4] in Europe. For this, the user adds occurrence data from the repository with the intend to cleanse it. The data occurs as a layer on the map and the user recognizes possible outliers by visually inspecting the clustering on the map.

For a first outlier removal (e.g. zoo animals), the user looks up so called expert ranges from IUCN[5] that outline the expected habitat of a species. The user filters the occurrence points by applying the point-in-polygon filter operator using the expert ranges. The result is a new layer which contains all occurrences contained by the expert ranges.

When looking at the data table, there are aggregates of default parameters from GBIF. From literature, the user knows that the species lives between sea level and a certain height. However, there is currently no elevation information present. The user adds hosted elevation raster data from WorldClim[6] as a new layer to WAVE. Then, the user attaches the raster data by applying the raster-value-extraction operator on both layers. The result is an enriched layer that serves then as an input for creating a histogram plot. The user applies a numeric range filter to remove all outlier occurrences which are not in the expected range.

The next step is exporting the resulting cleansed data and inspecting the file. It contains the data, the workflow and all citations that were included into the computation.

4.2 Statistical Analysis of Time Series

This part of the demonstration shows the impact of time series computations by observing bird movements. For this, the user starts in a clean project in WAVE and adds a layer of a migratory bird species, e.g. Sterna paradisaea (Arctic tern) occurrence points, to the map. The user first adds hosted environmental data of averaged monthly temperatures from WorldClim to the map. Then, the user inspects patterns where these birds have clusterings in the world and assumes a movement that is correlated with temperatures. As a third step, the user applies the raster-value-extraction operator to enrich the occurrence points with the corresponding temperatures. The user then applies a temporal operator to form a small temporal interval around the bird occurrences. The associated temperature values in the data table change over time when traversing in monthly intervals. To get a better understanding of the temperature attribute distribution, the user plots a histogram and observes a dense peak in the diagram.

To compare the observation, the user adds data of a resident bird, e.g. Columba oenas (Stock dove), to the map. As this is a species is stationary, there should be different results when using the same workflow. The user simply changes the source of the previous workflow and VAT computes new results. The map and the diagram show significantly different distribution patterns.

[4]Global Biodiversity Information Facility, www.gbif.org

[5]The International Union for Conservation of Nature, www.iucn.org

[6]Global Climate Data, www.worldclim.org

5. CONCLUSIONS

In this demo paper we presented VAT, a system for visualizing, analyzing and transforming spatio-temporal data while tracking citations and provenance information. We showed a brief system overview and pointed out the most important features. In our usage scenario, we presented two real-world applications of the field of biodiversity that also reveal interesting research opportunities for the database community.

6. ACKNOWLEDGMENTS

This work has been supported by the Deutsche Forschungsgemeinschaft (DFG) under grant no. SE 553/7-2.

7. REFERENCES

[1] C. Authmann, C. Beilschmidt, J. Drönner, M. Mattig, and B. Seeger. Rethinking Spatial Processing in Data-Intensive Science. In *BTW Workshops*, pages 161–170, 2015.

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

[3] C. Beilschmidt, J. Drönner, M. Mattig, M. Schmidt, C. Authmann, A. Niamir, T. Hickler, and B. Seeger. Interactive Data Exploration for Geoscience. In *BTW Workshops*, 2017.

[4] P. Buneman, S. Davidson, and J. Frew. Why Data Citation is a Computational Problem. *Communications of the ACM*, 59(9):50–57, 2016.

[5] M. Diepenbroek, F. O. Glöckner, P. Grobe, A. Güntsch, R. Huber, B. König-Ries, I. Kostadinov, J. Nieschulze, B. Seeger, R. Tolksdorf, 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-Jahrestagung*, pages 1711–1721, 2014.

[6] S. Jänicke, C. Heine, R. Stockmann, and G. Scheuermann. Comparative Visualization of Geospatial-temporal Data. In *GRAPP/IVAPP*, pages 613–625, 2012.