Using Chained Views and Follow-up Queries to Assist the Visual Exploration of the Web of Big Linked Data

Aline Menin, Minh Nhat Do, Carla Dal Sasso Freitas, Olivier Corby, Catherine Faron, Alain Giboin, Marco Winckler

Using Chained Views and Follow-up Queries to Assist the Visual Exploration of the Web of Big Linked Data

Aline Menin[1], Minh Nhat Do[1], Carla Dal Sasso Freitas[2], Olivier Corby[1], Catherine Faron[1], Alain Giboin[1], and Marco Winckler[1]

[1]Univ. Côte d'Azur, CNRS, Inria, Sophia Antipolis, France
[2]Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, Brazil

ARTICLE HISTORY
Compiled January 10, 2022

ABSTRACT
The Web of Linked Open Data (LOD) provides access to a great number of dynamic datasets containing valuable information to support decision-making processes in diverse application domains while being publicly accessible and up-to-date. While information visualization techniques are useful to explore, analyze, and explain relationships within LOD data, the existing tools are limited to visualizing a single dataset at a time and, often, use static and preprocessed data. In this paper, we leverage the linked aspect of LOD to support the dynamic integration of data into a visualization system by connecting views and distributed LOD datasets using the so-called *follow-up queries*. We demonstrate how our approach uses dynamic SPARQL queries to integrate external data into the exploration flow through visualization techniques and enrich the ongoing analysis. We ran a semi-structured interview to assess the usefulness of our approach, which results were encouraging while showing its relevance to explore big linked data.

KEYWORDS
linked data; big data; SPARQL queries; distributed datasets; follow-up queries; information visualization, chained views, exploratory process.

1. Introduction

Professionals in diverse application domains such as public health (Preim and Lawonn, 2020), social media (Chen et al., 2017), and finance (Ko et al., 2016) are confronted with the analysis of huge datasets to generate synthetic knowledge that supports decision-making processes. In the last 15 years, with the growth of the Open Data movement, a huge amount of data became available on the Web. For example, the EU portal gathers over 1,3 million datasets regarding subjects as diverse as environment, agriculture, justice, science and technology, government, economy, etc. Particularly, the Big Data era made available a great number and variety of very large datasets that are dynamic in nature (Bikakis and Sellis, 2016). These are often publicly available as Linked Open Data (LOD), i.e., structured data which is interlinked with other data, so it becomes more useful through semantic queries, as illustrated by the LOD Cloud, which, as of May 2020, gathered 1,301 datasets with 16,283 links or connections between datasets.

CONTACT Aline Menin. Email: aline.menin@inria.fr

Information visualization techniques are useful to discover patterns and causal relationships within LOD datasets (Bikakis, 2019). However, since the discovery process is often exploratory (i.e., users have no predefined goal and do not expect a particular outcome (Leng, 2011)), when users find something interesting, they should be able to (i) retrace their exploratory path to explain how results are found, and (ii) branch out the exploratory path to compare data observed in different views or found in different datasets. Indeed, as most of LOD datasets are very specialized, users often must explore multiple datasets to obtain the knowledge required to support decision-making processes. Thus, the design of visualization tools is confronted with two main challenges: the visualization system should provide multiple views to enable the exploration of different or complementary perspectives to the data; and the system should support the combination of diverse data sources during the exploration process.

Designing a single view to display all data is tempting, as displaying as much information as possible at once would minimize the need for exploration. However, the risk of engendering cognitive overload and visual clutter-related problems increases with the number of data dimensions. Moreover, it is not possible to display all types of data structures and relationships using a single view (Munzner, 2014). Multiple views can be used either through a coordinated (a change in one view will affect the other coordinated views) or chained (all views are connected but are not necessarily coordinated) layout; the latter leverage the capability of keeping the history of user interaction, which can serve for purposes of provenance analysis (North et al., 2011). Nonetheless, the integration of multiple data sources remains a challenge as most visualization systems only operate in an offline way, limiting the exploration to static and, often, small datasets of preprocessed data (Bikakis, 2019).

In a previous work (Menin et al., 2021a), we used chained views to support the exploration of LOD datasets while depicting provenance data. In this paper we introduce the concept of follow-up queries that allows users to create queries on demand during the exploratory process while connecting multiple LOD datasets with chained views. Our approach relies on a exploration process supported by the use of predefined SPARQL queries that the user can select on-the-fly to retrieve data from different SPARQL endpoints. It enables users to enrich the ongoing analysis by bringing external and complementary data to the exploration process, while also supporting the visual analysis and comparison of different subsets of data (from the same or different SPARQL endpoints) and, thus, the incremental exploration of the LOD cloud.

Contributions. This paper describes a generic visualization approach to assist the analysis of multiple LOD datasets based on the concepts of chained views and follow-up queries. We demonstrate the feasibility of our approach via four use case scenarios and a formative evaluation where we explore scholarly data described by RDF graphs publicly available through SPARQL endpoints. These scenarios demonstrate how the tool supports (i) composing, running, and visualizing the results of a query; (ii) subsetting the data and exploring it via different visualization techniques; (iii) instantiating a follow-up query to retrieve external data; and (iv) querying a different database and compare datasets. The usability and usefulness of the proposed approach is confirmed by results obtained with a series of semi-structured interviews. The results are encouraging while showing the relevance of the approach to explore big linked data.

The remainder of this paper is organized as follows. Section 2 compares related work on visual exploration of big and linked data, uses of chained views, and on-the-fly query processing for data exploration. Section 3 presents our approach and the tools support. A formative evaluation is presented at section 4. Section 5 discusses our contributions and limitations. Section 6 summarizes the conclusions and future work.

2. Related Work

In this section, we summarize and distinguish our contributions from previous ones made towards the visual exploration of big and linked data, the usage of chained views to improve the exploration process and keep provenance information, and the usage of dynamic queries within visualization systems.

2.1. Big Data Visualization

Bikakis (2019) identifies visualization methods used to handle the challenges of the Big Data era, which includes data reduction, hierarchical exploration, progressive results, incremental and adaptive processing, caching and prefetching, and user assistance.

Data reduction is used for computing abstract sets of data to enable efficient abstraction and summarizing mechanisms through approaches such as sampling and filtering (Fisher et al., 2012; Park et al., 2016) and aggregation (Bikakis et al., 2017; Elmqvist and Fekete, 2010). Hierarchical approaches allow the visual exploration of very large datasets at multiple levels, offering both an overview and an intuitive way for finding specific parts of a dataset (Bikakis et al., 2017).

For the purpose of providing real-time response while dealing with huge datasets, several systems adopt progressive techniques, where results and visual elements are computed/constructed incrementally based on user interaction or as time progresses (Park et al., 2016; Stolper et al., 2014). Another approach to reduce response time is based on caching and/or prefetching the sets of data that the user is likely to use during the exploration process (Kalinin et al., 2014). Since usually only a small fragment of the input data is accessed by the user, on-the-fly exploration techniques are used over large and dynamic datasets by incrementally processing and indexing data according to users' interactions (Alagiannis et al., 2012; Olma et al., 2017). Finally, in terms of user assistance, there are various solutions using visual recommendation to help users to choose suitable visualizations for data exploration (Key et al., 2012).

2.2. Linked Data Visualization

A thorough survey on LOD visualization is beyond the scope of this paper, so we refer the interested reader to the surveys by Antoniazzi and Viola (2018); Bikakis and Sellis (2016); Dadzie and Rowe (2011) and to the set of papers organized by Dadzie and Pietriga (2016). As suggested by the number of surveys on the subject we find in the literature, there are various visualization approaches to assist the exploration of linked data. Particularly, there are numerous tools that support the visual inspection and debugging of RDF through node-link representations (Chawuthai and Takeda, 2015; Graziosi et al., 2018), which show the relationship between subjects and objects determined by its predicates (e.g., in the triple `?parent dbo:influenced ?child`, the values of `?parent` and `?child` represent nodes and `dbo:influenced` represents the edge between those nodes). A common application of such tools would be to discover linked RDF graphs on the Web by using approaches such as revealing/hiding neighboring resources to explore and visualize relevant data of very large RDF graphs (De Vocht et al., 2015; Deligiannidis et al., 2007; Jacksi et al., 2018). A few tools support the exploration of OWL/RDF schema (Anutariya and Dangol, 2018; Kremen et al., 2018), which is of great importance to inspect the datasets to learn how to extract information from them, for example.

Although there are tools that support dynamic representation of different RDF data based on datatypes, they are mostly restricted to a single visualization technique, which provides a single perspective to the data and often does not consider the semantics behind the data (i.e., related to the application domain), resulting on a unsuitable visualization to solve domain-related tasks. Our approach supports exploratory search through various complementary visualization techniques instantiated on demand according to the task at hand, strengthening the analysis. Furthermore, regardless of the exploration goal and contrariwise to our approach, these tools often do not support the integration and simultaneous exploration of data originating from different RDF graphs, except when the data has been merged using, for instance, the SERVICE clause of SPARQL (Menin et al., 2021b).

2.3. Chained Views

Systems implementing chained views or similar concepts provide two or more visualizations in the display and connect them with visual links to represent one-to-one and one-to-many relations between data items. Connected Charts tool (Viau and McGuffin, 2012), the Domino technique (Gratzl et al., 2014), GraphTrail (Dunne et al., 2012) and SOMFlow (Sacha et al., 2017) are good examples. These tools support dynamic instantiating of multiple views during the exploration process and leverage the visual linking between the views to enable provenance tracking. In particular, GraphTrail supports the exploration of large multivariate, heterogeneous networks through drag-and-drop interactions that refine subsets of data in a new view while showing users' exploration history by lines connecting the views. In the SOMFlow tool, each view shows a cluster refinement of a dataset and links represent the analytical workflow of the exploration partition process. Although both GraphTrail and SOMFlow employ visual connections between views to show the exploration history, they are limited to exploring subsets of a unique dataset, while our approach supports the creation of views based not only on the refinement of already displayed data, but on data dynamically obtained by querying external linked datasets.

2.4. On-the-fly Query Processing

Dynamic queries are usually employed to support the exploration of datasets through the selection or filtering of data items, which resulting subsets are displayed in the same visualization or in multiple coordinated views (Shneiderman et al., 1992).

Early systems like DEVise (Livny et al., 1997) and VIKING (Visual Interactive QueryING) (Olsten et al., 1998) allow users to specify queries through graphical user interfaces. In both systems, querying and data browsing are unified into a single metaphor: the direct manipulation of visual representations. Such a visual query paradigm allows users that are not database experts to generate sophisticated SQL queries through intuitive graphical operations. In VQE (Derthick et al., 1997), the user is provided with a schema browser and a visual representation of the query, allowing the users to easily find the attributes they need. The authors distinguish between extensional and intentional queries: the former refers to selecting sets of objects via direct manipulation, while the latter has distinct declarative representation from what they evaluate on the current data, which can be reused on different data.

Heer et al. (2008) also present direct manipulation techniques that combine declarative selection queries (in a SQL-like query language) with a query relaxation engine

Table 1. Summary of related work according to: access to external data, multiple datasets exploration, adopted approach to display queries' results, the data type, the query language (**RDB:** relational database, **LD:** linked data, **N/A:** not applicable), and whether the solutions provide visual query builders (**VQB**).

Reference	External data	Multiple datasets	Results Display	Data Type	Query Language	VQB
Shneiderman et al. (1992)	×	×	update view	RDB	N/A	×
Livny et al. (1997)	✓	×	new view	RDB	SQL	✓
Olsten et al. (1998)	✓	×	replace view	RDB	SQL	✓
Derthick et al. (1997)	✓	×	new/update view	RDB	SQL	✓
Heer et al. (2008)	×	×	update view	RDB	SQL-like	×
Stolte et al. (2002)	✓	✓	new view	RDB	SQL	✓
Beyer et al. (2013)	×	×	update view	RDB	set algebra	✓
Destandau et al. (2021)	×	✓	new view	LD	SPARQL	×
Brunetti et al. (2013)	✓	✓	replace view	LD	SPARQL	×

that enables users to interactively generalize their selections. The users create selection queries through direct manipulation, and can reapply them dynamically over streaming or time-varying datasets or across different visualizations of a dataset, thereby supporting linking across views. The Polaris system (Stolte et al., 2002), precursor of Tableau, also supports the generation of relational queries through visual specifications, which allow subsetting data for analysis, filtering, sorting, and grouping the results into panels. Users could yet drill down in the visible dimensions or display different dimensions.

The ConnectomeExplorer tool (Beyer et al., 2013) uses knowledge-based query algebra to support the interactive specification of dynamically evaluated queries during the exploration process. The system is based on a coordinated views approach. It allows the user to formulate and answer domain-specific questions, either by interactively exploring the data or by posing dynamic queries through a visual query builder that translates queries into a query algebra. The results of queries are then represented as sets that can be explored in all views, used as input to more advanced queries, or stored and loaded from disk.

In the context of Linked Data, Destandau et al. (2021) propose S-Paths, a browsing tool that systematically identifies the best-rated visualization technique to represent a given resource set. By selecting different semantic paths, the tool allows users to switch to different resource sets or to get a different perspective on the same set through SPARQL query templates applied to a SPARQL endpoint that stores a set of RDF graphs. Another approach is the Linked Data Visualization Model (LDVM) (Brunetti et al., 2013), which allows connecting different datasets, data analysis, and visualizations in a dynamic way. The users can enter or select a SPARQL endpoint and choose the graphs to visualize. Many tools are available to transform the data and extract

information from specific types of LOD constructs (e.g., class hierarchy, property hierarchy, SKOS concepts hierarchy, etc.) that can be visualized through charts such as treemaps, tables, and heat maps.

Table 1 presents a summary of the above-described on-the-fly query processing solutions. We can observe that most solutions are focused on relational databases. Despite the fast growth of LOD datasets (both in terms of production and usage), there is still little work towards the dynamic integration of external data in visualization interfaces designed for exploring LOD datasets. The solutions we identified in the literature allow the visualization of data from different datasets through multiple visualization techniques. Nonetheless, these solutions provide a single visualization technique to represent the whole SPARQL result set, restraining the analysis to a single view of the data. Moreover, S-Paths (Destandau et al., 2021) is limited to the exploration of RDF graphs available via a specific SPARQL endpoint only while the analyzers provided by the LDVM (Brunetti et al., 2013) provide mainly RDF abstractions, which restrain the user from exploring domain-relevant subsets of data. In this context, our proposal differs from the existing solutions by (i) allowing the exploration of any SPARQL endpoint through SPARQL queries that can assist the exploration of RDF graphs, ontologies, or particular phenomena (e.g., bibliographic networks), and (ii) instead of replacing the existing views with a view of the new dataset, it integrates the query and the new view as chained views in the dashboard to provide exploration awareness while supporting visual comparison of different datasets.

3. Our Approach

This section presents an approach based on the concepts of chained views and follow-up queries to support the visual exploration of multiple large datasets. Particularly, we propose the visual exploration of LOD via:

- the *incremental exploration of large datasets* by pre-filtering them via SPARQL queries, which not only reduces the size of the data to be explored but also enables the analysis of more meaningful data to solve the task at hand;
- the *exploration of data from multiple perspectives* by using the chained views method, which allows users to interactively subset the data and further explore it using different visualization techniques that display the data from a different angle, while keeping the exploration path by linking the views; this allows users to focus their attention on smaller, more meaningful subsets of data;
- the *exploration of multiple datasets (from one or multiple databases)* by instantiating new queries during the exploration process (i.e., using follow-up queries) to incrementally include external data into the exploration flow, which can be analyzed and compared within a same visualization dashboard.

We propose an approach that combines the chained views concept and the visualization techniques used by MGExplorer (Menin et al., 2021a) and the query construction implemented by LDViz (Menin et al., 2021b). The resulting combination is then extended to include a visual representation of queries and enables on-the-fly query and data processing. Hereafter, we present the features that compose our approach.

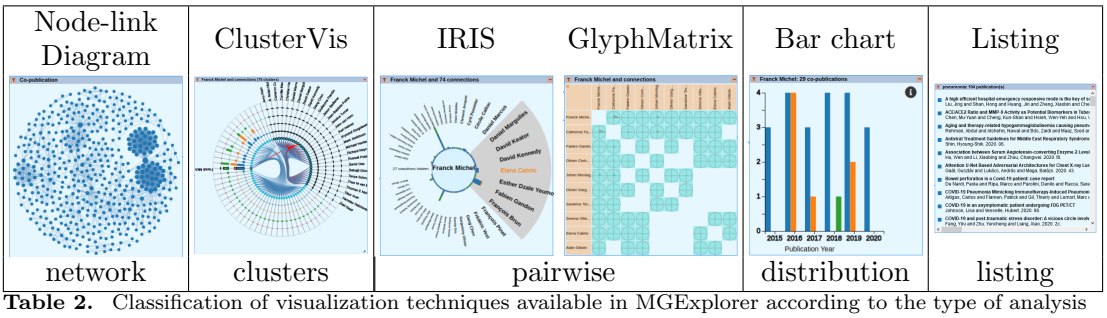

Table 2. Classification of visualization techniques available in MGExplorer according to the type of analysis they provide.

3.1. Interactive Visualization using Chained Views

We provide data visualization through an extension of the tool MGExplorer (Menin et al., 2021a), which implements chained views to assist the exploration of multidimensional and multivariate graphs. Each view in MGExplorer features a unique visualization technique. Using chained views, MGExplorer allows to compare (i) two or more different subsets of data through a particular perspective generated for each view, and (ii) multiple perspectives of the same subset of data using several views. Table 2 summarizes the set of available visualization techniques, briefly described herein.

Network visualization is provided via a **node-link diagram**, which shows items as nodes connected via line segments to represent the relationship between them. The size of nodes encodes the number of relationships of the associated item within the network. The **ClusterVis** (Cava et al., 2017) technique is used for displaying clusters obtained according to a particular relationship among data items. It has a multi-ring layout, where the innermost ring is formed by the data items (represented by circles), and the remaining rings display the data attributes (represented by rectangles). Curved lines connect the items belonging to the same cluster. Pairwise relationships between items can be visualized in two ways. The **IRIS** technique (Cava et al., 2014) isolates the item of interest (at the center) and shows the other items with which it has a relationship on a circular axis. The **GlyphMatrix** technique (Cava and Freitas, 2013) is a matrix-based visualization where rows and columns represent data items, and the intersecting cells embed a star-plot-like glyph encoding attributes that describe the relationship between the two items. Furthermore, a **bar chart** shows the distribution of values of data attributes for an item or set of items over an ordered variable (e.g., time), and a **listing** technique shows the set of items corresponding to the relationship between two or more nodes in the network. The original implementation of MGExplorer created chained views by subsetting the data obtained from a single dataset. In Subsection 4.3, we illustrate how chained views have been extended to feature follow-up queries.

3.1.1. Exploration Process and History

The dashboard of the tools is initialized with a blank query panel and a history panel; both panels are interactive and visible throughout the whole exploration process. From the query panel, users should select a starting endpoint and query. While the other views are created throughout the exploration process, the system progressively fills in the history panel with provenance information (views and dependencies between them). Each created view is connected to the previous one using line segments to represent their relationship. This approach allows a prompt recovering of the multiple analytical paths that emerge from a particular view.

The views can be moved around, allowing the user to rearrange the visualization space in meaningful ways. Further, users can hide any of the currently displayed views, which they may revisit later using the history panel, thus cleaning the display area in a way that helps them to focus on what is relevant to the task at hand.

The system supports the visualization of multiple datasets simultaneously. Each dataset can be further explored via filtering operations that allow to select an item of interest directly on a view and choose suitable visualization techniques to explore the resulting subset of data. Upon an element selection, the system filters the dataset accordingly, and the resulting subset undergoes a process that transforms and visually maps the data attributes to the chosen visualization technique. Throughout this process, the information regarding the selection operation, the dataset, the resulting subset, the chosen view, and the transformed data are recorded in the exploration history. The integration of datasets into the exploration flow is supported by dynamic follow-up queries, described in the following.

3.2. Query Representation and Processing

The visualization tool includes a querying process for retrieving datasets from multiple SPARQL endpoints and simultaneously exploring them within the same visualization dashboard. The SPARQL endpoint and query selection is supported by a query panel (Figure 1), which features the list of SPARQL endpoints available and queries supported by the endpoint. Upon the selection of a query, the system displays a set of custom parameters that allow users to filter the data (e.g. in a bibliometric network, these could be the publication period and research institution of scholarly articles) (Figure 1c). Users can choose a visualization technique among those present in Table 2. Then, upon clicking on "Run" (Figure 1g), the system will retrieve the information from the form, apply the query against the chosen endpoint, transform the resulting data, and instantiate the visualization technique to display the transformed data. For the purpose of optimizing the process, we use a cache that stores the results of queries for a certain amount of time, reducing the accesses to the data server. The user can deliberately clear the results stored in the cache by using the button "Clear cache" (Figure 1g), upon which the system will apply the query to the SPARQL endpoint at the next execution, acquiring updated data directly from the data server.

The visualization tool provides three types of queries:

- the *initial query* is the starting point of the exploration process, where the user defines the initial SPARQL endpoint and query. The initial query requires (i) a set of predefined queries (currently exported from LDViz – see Subsection 3.2.1), and (ii) at least one query that does not require an input value, for example, in Use Case Scenario III (see Subsection 4.3.3) the query requires the name of the author for which it would recover the co-authorship network.
- the *follow-up query* is used to integrate new data to the exploration flow by applying a new query to the current or a different SPARQL endpoint during the exploratory process. Further to a set of predefined queries, the follow-up query requires (i) at least one visualization displayed on the dashboard, and (ii) a selected item from a visualization to serve as input data for the new query.
- the *cloned query* allows the user to reuse the input data and query parameters from an existing query panel by creating a copy of the panel, where the user can make the necessary modifications to obtain a new visualization with different data. It requires at least one follow-up query displayed on the dashboard.

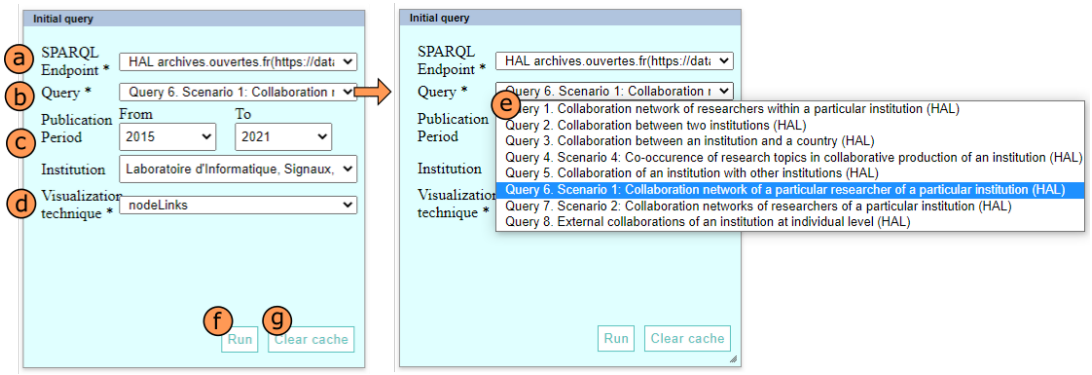

Figure 1. Overview of the query panel. (a) Select a SPARQL endpoint. (b) Select a query. (c) Custom query parameters. (d) Visualization technique to represent the results. (e) List of predefined queries for the selected SPARQL endpoint. (f) Execute the query using the selected parameters and visualize the results. (g) Clear results stored in cache for the selected query.

3.2.1. Query Specification via LDViz

To assist the use of the tool by people with little or no knowledge of SPARQL, we use predefined queries in the visualization dashboard. Then, to define new SPARQL queries, we use the LDViz (Linked Data Visualizer) tool[1]. The queries are then exported as a JSON file that describes the queries and the parameters needed to execute them (i.e., SPARQL endpoint, query type, custom variables), which is used as input data for the query panels in MGExplorer. LDViz implements a generic visualization pipeline for LOD datasets based on web technologies, i.e., JavaScript, the D3 (Data-Driven Documents) library to create visualizations, and the NodeJS library to manage the linked data access server that handles data retrieval through SPARQL queries. More information about the data model and the operating mode of LDViz can be found in Menin et al. (2021b).

3.3. Implementation Details

The extended version of MGExplorer provides a modular architecture based on web components that reduce the complexity of the interface and support the reuse of elements. It also allows the easy inclusion of new visualization techniques and functionalities. We used Stencil JS[2], a compiler that generates Web Components and builds high-performance web apps to implement the new architecture of MGExplorer. Finally, data and SPARQL queries processing is handled through a NodeJS server.

Web components are a set of web platform APIs used to create new custom, reusable, encapsulated HTML tags in web pages and web apps. They are based on four main specifications: the *custom elements* are the foundation for designing and using new types of DOM elements; the *shadow DOM* defines how to use encapsulated style and markup; the *ES modules* define the inclusion and reuse of JavaScript documents in a standards-based, modular, performing way; and the *HTML template* defines how to declare fragments of markups to be used during runtime. The Shadow DOM specifications allow components to have their own `dom` tree, which cannot be accessed accidentally from the main document. With our approach, we can provide full encapsulation,

[1] http://covid19.i3s.unice.fr:8080/ldviz
[2] https://stenciljs.com/

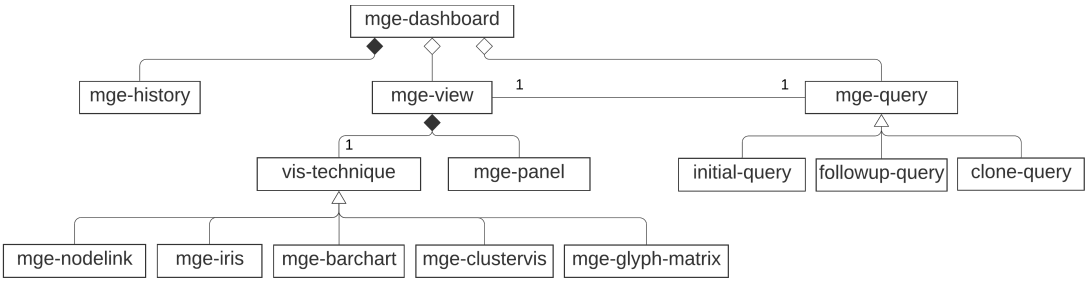

Figure 2. Overview of MGExplorer architecture and interconnected web components.

reducing the dependencies between components while preventing style specifications either to change a component from the outside or styles from inside a component to bleed out. Thus, one can properly include a new visualization technique to the tool without affecting the rest of the components. Currently, MGExplorer comprises six main components interconnected as illustrated in Figure 2 and described as follows:

- the *dashboard* component (`mge-dashboard`) stores and manages the views, the data, and the user interactions (e.g., hiding, displaying, dragging, and dropping views, etc.). This component also stores the provenance data describing the exploration process and show the exploration path by drawing the connection lines between views and updating the history tree (displayed by `mge-history`). It includes a list of views, a query panel (initial query), and the history panel.
- the *view* component (`mge-view`) comprises a visualization technique (i.e., `vis-technique`) and a settings panel that serves to customize certain aspects of the visualization (e.g., sorting items, search, etc).
- the *settings panel* component (`mge-panel`) serves to customize the visualization techniques by adjusting certain parameters. It uses a predefined template rendered as HTML inside the component according to the visualization technique.
- the *query* component (`mge-query`) comprises the query panel where users can select and customize SPARQL queries to retrieve and explore data using the visualization. This component is placed inside the `mge-view` component to reuse common features (i.e., drag-and-drop, hide, display, etc.). As mentioned earlier, it can take three different forms: initial, follow-up, or clone query.
- the *history* component (`mge-history`) displays the exploration path through a hierarchy that shows the dependencies between views and supports interaction for hiding and displaying views.
- the *visualization techniques* (`vis-technique`) encompass six components (i.e., `mge-barchart`, `mge-clusterviz`, `mge-glyph-matrix`, `mge-iris`, `mge-listing`, `mge-nodelink`), each one comprising its own properties and methods.

4. User Study

To assess our approach, we performed a user study using semi-directive questions that followed a demonstration of our visualization tool. This user study was proposed as a formative evaluation, which aims to collect observations and recommendations that can be immediately used to improve the design of the product or service, and refine the development specifications (Burmester et al., 2010). In a typical formative evaluation

we addresses questions such as: what are the usability issues in our implementation? Do users understand the underlying concepts and tasks supported by the tools? Does the system comply with recognized usability principles? The results are typically qualitative instead of summative, focusing on the needs of the design team including developers, designers, project managers, and other members. With this evaluation, we aim primary at identifying gaps between the goals of our conceptual approach and the current implementation, which is done through feedback from expert users and through the collection of specific points, where improvement is necessary before pursuing the implementation towards a professional use. For that, we designed questions for asking participants through the interview to address the following research questions:

Q1. How do users relate the content of visualizations and queries?
Q2. Are users able to distinguish subsetting operations (i.e., filtering applied to existing visualizations) from the follow-up queries (resulting in visualizations of new data)?
Q3. Would users be able to track the data provenance using chained views? Here, we are interested in understanding whether users can compare and distinguish data coming from different sources during the exploration process.

4.1. Participants

For this study we were looking for users representing the public concerned by the exploration of LOD datasets. For that, we recruited ten participants (5 female and 5 male), aged from 26 to 61 years (N=4 between 20-29, N=3 between 30-39, N=2 between 40-59, and N=1 person aged over 60 years old) in a convenience sample. We ensure that all participants were knowledgeable of Semantic Web and are representative users of LOD datasets. Half of the participants were Ph.D. students or engineers, while the other half were post-doctoral or research fellows at universities. Semantic Web and Artificial intelligence are the main research fields of all participants. Specific research topics include linked open data, natural language processing, information retrieval, web audio, and knowledge representation. Two participants also reported using visualization tools in their research.

Participants were asked what types of data they deal with on daily and how they explore or exploit that data. Most participants (n=7) work with knowledge graphs and ontologies, which they explore using SPARQL queries, dedicated software (e.g., Protégé Ontology Editor[3], Corese[4], and Virtuoso SPARQL Query Editor), or software specifically created (sometimes by themselves) to answer their needs. The other participants work with text and tabular data, which they explore using mainly Python, tabular explorers, and simple charts (e.g., scatter plots and directed graphs).

4.2. Materials and Methods

4.2.1. Protocol

In order to demonstrate the usage of the visualization tool we define a set of use case scenarios that were used to guide users during a semi-directed interview. The questions proposed to the participants might require follow-up questions according to the answers of the interviewees.

[3] `https://protege.stanford.edu/`
[4] `https://project.inria.fr/corese/`

Participants gave written consent to participate in the interviews. All the data collected was anonymized. Each interview took around 35 minutes. We used Google Forms to collect the data, which the interviewer filled out to allow the interviewee to naturally express their opinions through speaking instead of writing them down.

The interview began with a presentation of the study's goals and the visualization tool. We applied a pre-test questionnaire to collect information regarding the participants' profile (i.e., age, gender, education level) and to understand their habits and needs regarding the exploration of large datasets; for example, what types of data they work with and how they explore it, whether they need to explore multiple datasets simultaneously, and what tools they use to explore those data.

In the sequence, the interviewer presented and demonstrated our visualization tool via four use case scenarios (Sect. 4.3). After each scenario, the interviewer asked the participant to describe: (i) three things they liked in the scenario; (ii) three things they disliked in the scenario; and (iii) whether, when, and why they would use the visualization, and particularly, the features presented in the scenario under consideration. The last part of the interview was dedicated to debriefing and thanking the participants.

4.2.2. Data and Dataset

In this study, we used data describing bibliometric networks (i.e. keywords co-occurrence and co-authorship networks). The data was retrieved from two different RDF graphs storing metadata that describe scientific publications (e.g., title, publication date, authors, research topic, etc.): the HAL Open Archive[5] and the Microsoft Academic Knowledge Graph (MAKG)[6]. The HAL Open Archive is a French open archive where authors can deposit scholarly documents from all academic fields, gathering over 2,6 million scholarly articles, including over 860,000 with full-text. The MAKG collects metadata describing over 209,7 million scientific publications from across the web and partnered sources.

4.3. Use Case Scenarios

For collecting the appropriate data to answer our research questions, we designed four use case scenarios, each one presenting a different aspect of the tool, i.e. (i) composing, running, and visualizing the results of a query; (ii) subsetting the data and exploring via different visualization techniques; (iii) instantiating a follow-up query to retrieve external data; and (iv) querying a different database and compare datasets. Hence, scenarios I, III, and IV address aspects of **Q1** by showing the visual representation of three different datasets retrieved using three different queries. Scenario II presents the subsetting feature, which, together with the remaining scenarios, should allow us to collect feedback to answer **Q2**. Finally, **Q3** should be answered with feedback collected throughout all scenarios since the participants should be able to identify exploratory paths, which requires the instantiating of a certain number of visualizations and queries.

[5] `https://data.archives-ouvertes.fr/doc/schema`
[6] `https://makg.org/`

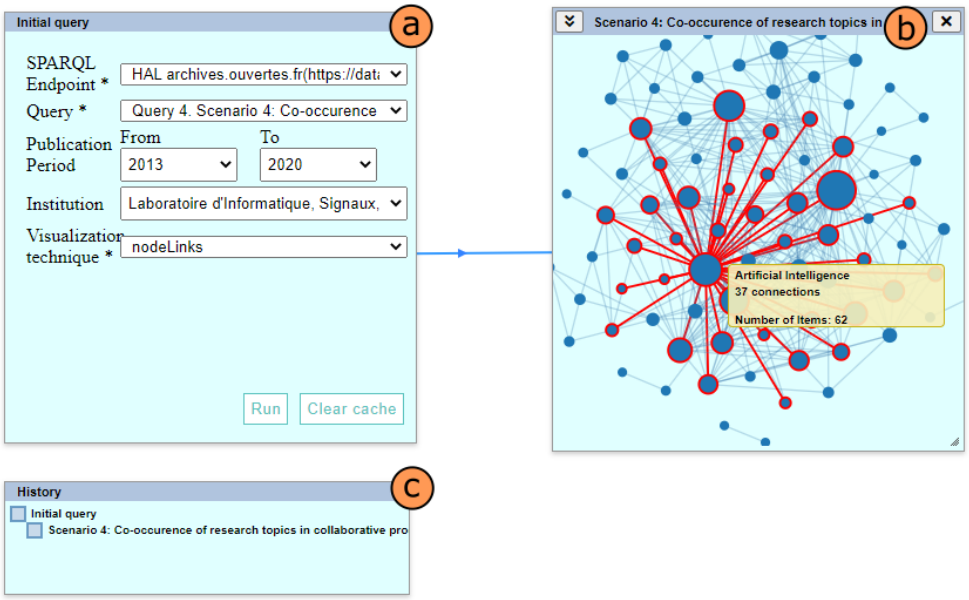

Figure 3. Exploratory path of Use Case Scenario I. (a) Initial query window. (b) Node-link diagram representing the co-occurrence network of keywords within scientific publications in HAL. (c) History tree showing the information regarding the initial query and the node-link diagram.

4.3.1. Use Case Scenario I

This scenario demonstrates how one can compose and execute a SPARQL query and visualize the results using the tool. Since this is the first scenario, we start by presenting the MGExplorer's interface, including the initial views (i.e., initial query and history) and their operating mode. Then, we select a SPARQL endpoint and a query in the query panel to begin the exploration. In this scenario, we are interested in a query to retrieve a dataset describing a keyword co-occurrence network within scientific publications of a particular research institution stored in the SPARQL Endpoint HAL. Upon choosing the query, the system displays a set of custom parameters that the user may change to filter the data by publication period or research institution. Illustrating the usage of these parameters, we set the publication period as 2013 to 2020.

Upon clicking on run, a view is created showing the results represented as a node-link diagram where keywords are represented by nodes, which size encodes the number of connections (i.e., the number of keywords with which a keyword co-occurs), and the links are defined by the publications where they jointly appear to represent the co-occurrence of both keywords in at least one publication. As the new view is displayed, the system draws a line segment with an arrow connecting the initial query to the view. The history tree is updated to include the new view, representing the dependency between the visualization and the query.

4.3.2. Use Case Scenario II

In this scenario, we demonstrate how one can use MGExplorer to further explore subsets of data via different visualization techniques to get different perspectives to the data. For that, we select an item in the node-link diagram by right-clicking on the corresponding node (e.g., the one representing the keyword "artificial intelligence"). This action will display a context menu featuring the list of available visualization

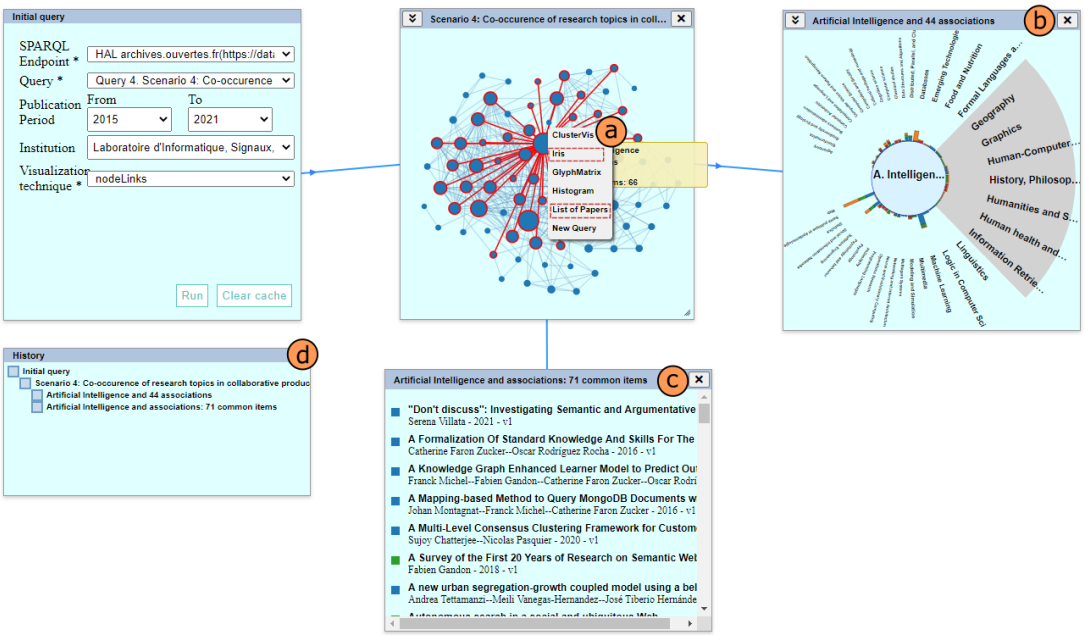

Figure 4. Exploratory path of Use Case Scenario II. Starting from the previous scenario, we have (a) a context menu triggered by right-clicking on a node of interest; (b) the IRIS technique showing the pairwise relationship between "artificial intelligence" and its associated items; (c) the list of papers featuring the scholarly articles where the keyword of interest appears together with one or more of its associated keywords; and (d) the history tree updated to show information about the newly instantiated visualizations.

techniques, from which we choose the IRIS to show the pairwise relationships between the keyword of interest and every other keyword with which it co-occurs. The IRIS illustrates the importance of each relationship via colored bars, where the color represents the type of publication and the height encodes the number of publications where both keywords co-occur. Further, we return to the node-link diagram, where we right-click again on the node of interest to launch a different visualization technique. This time, we select the List of Papers technique, which shows us a list of publications where the keyword "artificial intelligence" appears together with one or more of its associated items.

4.3.3. Use Case Scenario III

This scenario demonstrates the application of follow-up queries: extending the analysis by querying an external database using existing information in the visualization. We focus the exploration on the List of Papers. Let us consider that we are interested in the paper entitled "A Survey of the First 20 Years of Research on Semantic Web and Linked Data". We want know more about its author, "Fabien Gandon" and, particularly, about his scientific collaborations.

For that purpose, since the data being explored focus on keywords co-occurrence, we should query the HAL SPARQL endpoint to retrieve the co-authorship network of "Fabien Gandon". For that, we right-click on the item representing the publication of interest and select the option "New Query", which instantiates a query view giving the author's name as input data to be used as a parameter in the follow-up query. This view is automatically connected to the List of Papers through a line segment to represent their relationship. Furthermore, notice that the query panel's title contains

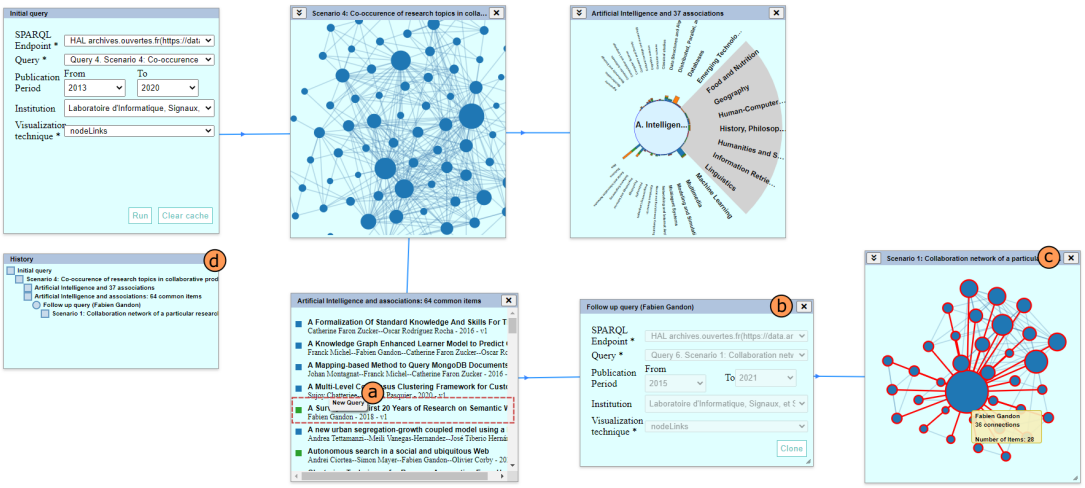

Figure 5. Exploratory path of Use Case Scenario III. Starting from the List of Papers from the previous scenario, we have (a) the context menu regarding the publication of interest; (b) the query panel featuring a form filled out with the SPARQL endpoint, the query, and query parameters; (c) the node-link diagram featuring the co-authorship network of the given author retrieved from HAL; and (d) the updated history tree representing the query panel through a circle and the new visualization technique with a square.

the name of the author, which allows knowing at all times from which item the query was derived.

In the query view, we select the HAL SPARQL endpoint and query 6 from the list, which provides the result we are looking for, i.e., the co-authorship network of "Fabien Gandon". We keep the default parameters and execute the query. The system then instantiates a new node-link diagram where nodes correspond to authors, which size encodes the number of co-authors, and links are defined by the publications they co-authored jointly, representing their co-authorship through at least one publication. The new view is automatically connected to the query view via a line segment, and the history is updated accordingly; notice that the history tree represents queries via circles and visualization techniques via squares, providing a clear indication of the start and end of workflows.

4.3.4. Use Case Scenario IV

This scenario demonstrates another application of follow-up queries: comparing data from different databases. Up to this point, we have explored data from the HAL SPARQL endpoint, which gives us the scientific collaboration network of "Fabien Gandon" from the perspective of the data stored in this database. Now, let us compare this data with the co-authorship network of this researcher retrieved from MAKG.

For that purpose, we must create a new query and a new view, which would rapidly clutter the visualization space. Thus, before continuing the exploration, we hide the views that we no longer need (i.e., the first node-link diagram and the IRIS) and rearrange the remaining views to create space for this new exploration path. As we can observe, by hiding a view, the system replaces it with a gray square that serves both to indicate the existence of previously used visualizations and as a button that, upon clicking, re-displays the corresponding view.

There are two ways for performing the necessary query: (i) we start from the node-link diagram, by right-clicking on the node that represents "Fabien Gandon", choose

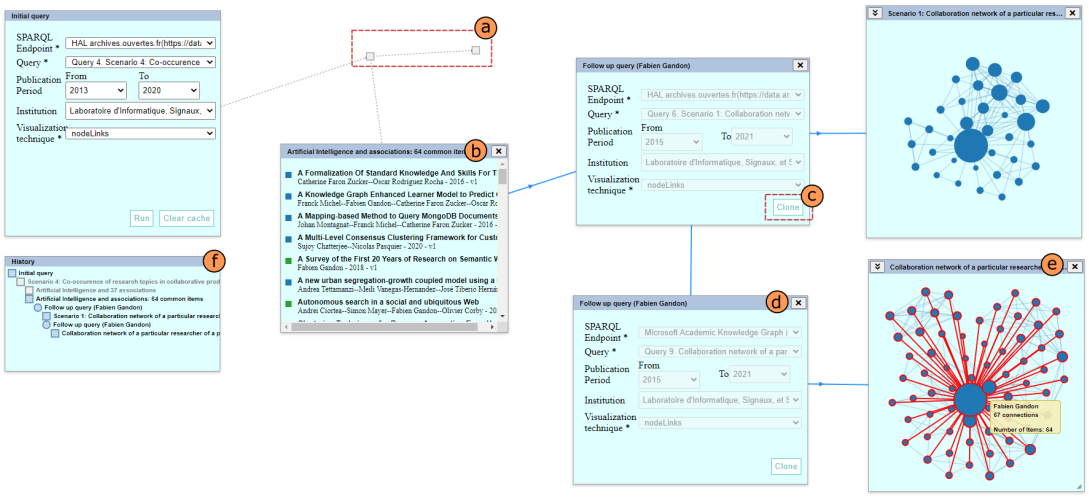

Figure 6. Exploratory path of Use Case Scenario IV. After hiding the views we no longer need, we remain with (a) two gray squares and the link between them indicating a hidden part of the exploration path and the (b) list of papers. In the query panel, we use the clone button (c) to create a cloned query that reuses the same input data and parameters of this one. We now have a (d) cloned query view, which we modify to select the necessary SPARQL endpoint and query for the new exploration; and (e) the node-link diagram featuring the co-authorship network of the given author retrieved from MAKG.

the option "New Query" in the context menu that appears, and follow the same steps as in the previous scenario; or (ii) since we are using the same input data (i.e., the author's name) that came from the List of Papers, we could also reuse the information on the previously used query view by cloning it, which has the advantage of allowing the reuse of parameters and input data.

By following the second way, we clone the query view by clicking on "Clone", which instantiates a new query view containing the same information as in the previous one. In the cloned query view, we change the SPARQL endpoint to MAKG and select query 9, which retrieves the co-authorship network of "Fabien Gandon". Upon executing the query, the system displays a node-link diagram where nodes represent the authors linked by the publications they co-authored together. We can compare both visualizations side-by-side, where we can promptly observe that the network found in MAKG is slightly larger than the one found in HAL. By hovering over the node that represents "Fabien Gandon" in both node-link diagrams, we observe that this author had 36 co-authors between 2015 and 2021 in 28 scholarly articles in the network retrieved from HAL. For the same period, the MAKG provides a network where this author had 67 co-authors through 64 scholarly articles.

4.4. Results

From the ten interviews, we could extract a total of 116 comments, classified at Figure 7 as positive (what they value and appreciate) and negative (suggestions for improvements). The positive and negative aspect of each comment was determined at the moment of the interview, when the participants were asked to provide three things they liked and disliked, respectively, about the scenario they had just watched. We observed that the content of comments reach a scope that surpass the topics addressed by our research questions. Thus, to proceed with the analysis, we categorized the comments into four dimensions, defined as follows:

Figure 7. Count of positive and negative topics addressed by participants regarding each studied dimension and the number of times they were mentioned during the interview.

- the *Data Integration* dimension gathers comments regarding the capability of the tool for integrating multiple datasets from one or multiple databases and visualizing them jointly within the same interface. The information gathered by this dimension should help us to answer **Q1** since this question seeks to understand how the users relate the content of visualization and the different queries;
- the *Exploratory Process* dimension gathers comments regarding the exploratory procedure based on chained views that are obtained through selection and filtering and exploring data from the diverse perspectives provided by different visualization techniques while keeping a trace of the exploratory path through both line segments that connect the different views and a panel featuring a history tree. The feedback gathered by this dimension should help us to answer **Q1** and **Q2**, since these questions seek to understand whether and how users perceive the different datasets and visualizations being displayed and the connection between them;
- the *Query Composition and Management* dimension gathers comments regarding the capability of the tool for querying different SPARQL endpoints in runtime by using predefined and customizable queries, as well as the ability to instantiate new queries at any moment of the exploration process and applying them to different SPARQL endpoints, at will. This dimension provides feedback that will help to improve the querying process and refers to **Q3**; and
- the *Usability* dimension gathers comments regarding overall comments about the user interface. This dimension was named usability as most of the comments refer to how users perceived the utility, performance, and overall satisfaction with the design of the user interface. This category further distinguishes comments regarding each visualization technique used during the scenarios, the history and query panels, and the visualization dashboard. This dimension provides feedback that will help improving usability and user experience with the tools presented. The comments were also analyzed with respect to problems in specific views.

Participants were mostly pleased with the visualization tool and the scenarios we presented, as most of the remarks were positive (56%). Although the number of negative comments seems large, accounting for 44% of all remarks, most of them refer to smalls suggestions for improving the usability of individual visualization techniques and do not hinder the use of chained views and follow-up queries. If we exclude these comments referring to usability issues (4^{th} dimension), the positive remarks account for 81% of the comments, suggesting that participants were overall pleased with the approach we propose to explore multiple, large datasets. Hereafter, we present and discuss these comments according to each of the above-described dimensions.

4.4.1. Data Integration

There were a total of six remarks regarding the data integration dimension, which were mentioned a total of nine times and were all regarding aspects that participants enjoyed. Although most participants reported the need for exploring multiple datasets simultaneously (9 out of 10 people), they would mostly do so by querying the different databases and comparing the results through non-visual and non-automatized ways (e.g., observing the different result sets in a table or text file) or through a script that map the differences between the datasets and provides the results as a table or text file. Thus, the possibility of integrating datasets from different databases and comparing them in a visual way within the same interface was perceived as a positive and useful functionality, even for those participants that mention using only one dataset at a time. The most frequent remarks (66% of comments on data integration) refer to the possibility of accessing diverse datasets and compare sets of data. Particularly, P9 said that "the visual comparison of information from different datasets allows to grasp the extension and differences between them", which illustrates the benefit of performing such comparisons.

4.4.2. Exploratory Process

This dimension received a total of 22 comments, which were mostly positive (86%). The 19 positive comments were mentioned 36 times, while the three negative comments received 11 mentions. The aspects that mostly pleased the participants were the creation of queries at any moment of the exploration process and the usage of data available in the visualization as input to instantiate a new query. The positive aspect is to reuse data in the display to build new queries, as illustrated by P10 who enjoyed the idea of "choosing a specific element in a visualization and instantiating a query, allowing to start a new exploration from that specific item of interest". Every participant enjoyed the possibility of including external data to the visualization by instantiating new queries on-the-fly. Only one participant (P6) mentioned that "the operating mode of follow-up queries is not intuitive", which is sensible since the concept was introduced to participants at the time of the interview, being totally new to most of them.

Furthermore, other two negative comments were raised with particular regard to the simultaneous visualization of different datasets. The usage of a non-contextualized node-link diagram to visualize different datasets was considered to hinder the comparison between them. Further, six people missed a visual representation of differences and commonalities of the different datasets under exploration, either via colors or linking and brushing approaches. P10 suggested including a customizable node between views "that would hook the results from two or more queries and perform operations on it

before visualizing, such as finding the differences between them", which is similar to what is done in Dunne et al. (2012).

Overall, participants appreciated simultaneously having on the display the visualization techniques and the queries, as well as the connection between them that illustrates the data provenance. Participants also appreciated the fact that we keep the history of the exploration via both the visual exploratory path and the interactive history panel. For instance, P5 said that "having an exploration path allows to contextualize themselves within the process" and P1 enjoyed the possibility of "seeing the history and reopen the windows that we closed".

The comment raised by P10 saying that they liked "seeing the visualizations in parallel to compare the information" illustrates how participants appreciated having side-by-side views, which remark was also raised by two other participants. A recurrent remark regards the usefulness of having different perspectives to the data either by applying new queries and retrieving new data to expand the analysis or by using multiple visualization techniques that allow exploring subsets of data from a different angle.

4.4.3. Query Composition and Management

A total of fifteen remarks were collected regarding the query composition and management dimension, which were mostly positive (66,6%), accounting for 10 over 15 comments. While the positive remarks were mentioned 26 times, the negative remarks were only mentioned once each. There were particularly two topics that stood out, which regards the possibility of customizing query parameters (in our scenarios those would be the publication period and the research institution used in the query) and the feature that allows cloning a follow up query to modify it or instantiate a new query using the same input data that originates from a particular visualization technique. As mentioned by P2, "the custom parameters allow to customize the query and to define what we want as data" and "cloning the query is useful to reuse the input data without repeating the process [of instantiating a new query]". The latter statement is completed by P6, who liked the cloning feature because "it reuses the query parameters of the follow up query". However, one participant mentioned that, although they appreciate being able to clone a query, they dislike having to clone it only to modify the parameters.

Using a list of predefined queries was overall appreciated by the participants, which according to P3 and P8 "encourage users who have no knowledge of SPARQL on using the tool". However, one participant mentioned that using predefined queries restrain the exploration and they would like to write new queries themselves, which feature is part of LDViz, as mentioned earlier.

4.4.4. Usability

The majority of remarks concern usability aspects of the tool (total of 73 comments, 30 positives). These remarks as presented hereafter in categories referring to the views.

Dashboard. This category got half of the usability remarks (37 comments – 15 positives); it refers to the composition of the user interface that host the views (visualization techniques, queries, history and the connections between them). The total of positive and negative remarks received both 30 mentions from participants. The most appreciated aspects of the dashboard were the high speed to process the queries and display the results, which we assure by using a cache that keeps the query's results in

the memory for a particular amount of time speeding up the process, and the intuitiveness of the interface, which users mentioned to be simple and easy to use. Surprisingly, only one participant said that using the interface is not intuitive, since the whole tool and the concept of chained views was new to most participants. Overall, participants also enjoyed the possibility of hiding the visualizations. P3 mentioned during the presentation of the use case scenario IV that "it helps focusing on the data resulting from the new query". In this regard, one participant (P10) also mentioned that they did not like "the fact that we cannot remove anything, only hide [views and queries]"; this participant wanted to remove the views from the dashboard when starting a new exploration flow, which is supported by refreshing the dashboard on the browser.

History Panel. Four comments regard the history layout, such as the hierarchical representation and the use of different symbols to represent views and queries, which were appreciated by the participants. As a negative aspect of the history panel, two people mentioned that they would have rather identifying the visualization technique used in each view directly in the history, e.g., by using different symbols or colors.

Query Panel. The few comments regarding the usability of the query panel (3 negative remarks) address the layout, which participants believe that could be improved. Furthermore, P3 mentioned that they would like to know directly in the interface "whether the available values for query parameters (e.g. research institutions) are exhaustive in regard to what exists in the database". Regarding the list of available queries, P10 mentioned that they would have rather "having an indication of suitable queries for the selected item by sorting the list or using colors".

Node-link diagram. Twelve comments (5 positives) acknowledge aspects such as hover over a particular node and highlight the nodes and edges connected to it, and the use of nodes' size to represent the number of connections (i.e. co-authors or associated keywords) of a particular item. Similarly, they would have liked to see the importance of such connections directly on the graph: as mentioned by P4 "it is a shame that the links are all the same, while the thickness could represent another information such as the number of publications between two nodes". A recurrent comment was regarding the non-contextualization of the node-link diagram, which hindered the understanding of the graph's nature and, therefore, the visual distinction between graphs that represent different data (e.g., keywords co-occurrence and co-authorship networks).

IRIS. Most of the comments (11 remarks; 6 positive) made about the IRIS technique regard its intuitiveness, interaction and ability to quickly identify the important connections of a particular item. P4 said that they liked "having the number of publications represented between two items" through the colored bar. Participants also enjoyed having the types of publications encoded by color in both IRIS and the List of Papers (discussed below), but disliked the lack of legend of colors that would improve the reading of these visualization techniques. Although the participants seemed to appreciate the IRIS due to the amount of information it provides, some aspects such as the focus and the size of items were troubling. Furthermore, one participant mentioned that the IRIS is a complex technique and, thus, not intuitive to use.

List of Papers. Five comments were regarding the List of Papers (2 positive). Participants appreciated having the list of publications as it gives a reference to the data source. Three people mentioned that they would like to have the possibility of sorting the list by author, publication date, title, etc. One participant even suggested allowing the user to export the list of publications as a reference file, e.g., Bib TeX.

It is worthy mentioning that the comments classified as "negative" are often suggestions for improving the tool. Most of the suggestions are quite simple and easy to solve and do not compromise the use of the tool.

4.5. Response of Research Questions

The interviews resulting on a great amount of data regarding diverse aspects of the visualization tools, reaching a scope that goes beyond of the focus of our research questions, which we try to respond hereafter.

Regarding **Q1: How do users relate the content of visualizations and queries?**, we observe that the connection between the views and the queries are considered easy to understand by the participants. Participants promptly associated that graphs displayed data in the views resulting from the queries. Many were happy to see the visual description of the queries (data source, query, and parameters of the query). Explicitly showing queries is something often missing in other tools and this functionality was positively valued by participants. The participants had no trouble in perceiving the meaning in the views and queries or the change between these visual components. A clear remarks of this point was stated by P2 who acknowledged "the consecutive search allows to go from keywords to authors".

Regarding **Q2: Are users able to distinguish subsetting operations from follow-up queries?**, we could confirm their ability of visually distinguishing these two operations. Participants understood and appreciated both features as they allow further exploration of existing data and the perspective change by bringing new data to the exploration flow. Particularly, P10 was pleased with "the possibility of starting [a new exploration flow] from a specific item displayed in a visualization technique", which illustrates their understanding and appreciation of this feature.

As for **Q3: Would users be able to track the data provenance using chained views?**, various comments illustrate how participants appreciate the visual exploratory path and the history panel, such as P8 "to see everything at once with the connections between views and the possibility of resuming the exploration flow". Many suggestions on how to represent views and queries in the history and how to encode visualization techniques, indicate that participants understood how to trace back the exploration path. In use case scenario IV, the participants were able to compare information retrieved from different databases, which confirms the need of displaying the differences and commonalities between datasets.

5. Discussion

In this paper, we presented a visualization approach based on chained views and follow-up queries to assist the exploration of multiple, large datasets. In our case study, we focused on linked open data by exploring data from two RDF graphs (HAL and MAKG) that describe scholarly articles via information such as authors, title, research topic, publication date, etc. These RDF graphs are accessible through SPARQL endpoints, which enable on-the-fly querying and data processing. We assessed our approach by the means of interview with ten expert users, where we demonstrated the usage of our visualization tool through four use case scenarios and asked for the participants' opinion on the usefulness of the presented features. The results showed a positive outlook towards exploring multiple datasets within the same visualization dashboard, allowing the instantiating of new queries during the exploration process, and keeping a visual trace of the latter. Although we applied our approach to big linked data, the concept could be easily generalized to any type of big data. Therefore, hereafter we organize the discussion according to the requirements of an efficient and effective visualization system for big data exploration as identified by Bikakis (2019).

Real-time Interaction. For the purpose of providing a pleasant user experience and keeping the system's response in the range of a few milliseconds, we use a caching approach that keeps the queries' results in memory for a certain amount of time, thus reducing the request time for follow-up queries. Since the participants of our study are used to work with large datasets in their routine and, therefore, understand the difficulties of processing such large amounts of data, the quick response of our system did not go unnoticed and has been perceived as a positive aspect of the tool.

On-the-fly Processing. To support the simultaneous exploration of multiple datasets we proposed a technique called "follow-up queries", which allows the user to instantiate and execute new queries on-the-fly, which resulting datasets are processed (e.g., filtered, transformed, and enriched) in runtime and integrated in the exploration flow. Further to enabling the comparison of data originating from one or multiple sources, this approach supports yet the incremental exploration of a large database by focusing the analysis on a different perspective at the time. Additionally, we allow the user to focus on useful datasets to the task at hand by using customizable queries, where certain parameters can be modified to meet the user's needs. Although these queries cannot be written by the user directly on the MGExplorer interface, we can eventually provide access to LDViz, where users can define meaningful queries that can later on be included into MGExplorer to enrich the exploration.

Visual Scalability. For the purpose of providing an efficient exploration of the different datasets and subsets of data, we use a customizable dashboard approach. This way, users can use drag and drop interaction to arrange the views within the dashboard in meaningful ways to the ongoing analysis, as well as hide views that they no longer use without losing the exploratory path and having the possibility of re-displaying those views, if needed. This approach allows users to explore as many datasets as they please through multiple visualization techniques without leaving the interface, which has been perceived as a positive aspect of our tool by the participants.

User Assistance. In the current state of our tool, user assistance is a major limitation, as the participants duly noticed during the interviews. In the short term, we intend to improve the tool to include information that allow users to fully leverage MGExplorer features to explore data on their own. In the long term, future work includes extending provenance features, such as improving visual representation and increasing the types and amount of data collected, as well as studying such data to, for instance, identify the most common usages of the system (standard choices of visualizations and instantiating order) according to different types of tasks, which could be used to introduce the system to new users, suggest some well-known analysis workflows, and to improve the overall user experience.

Personalization. Both the flexible dashboard and the follow-up queries approaches provide users with the possibility of personalizing their exploration according to the task at hand. According to Leng (2011), in an exploratory context, the user has no defined goal and is looking for no particular outcome. Thus, when finding something interesting, users should be able to (i) retrace their exploratory path to explain how they found the results, and (ii) branch out the exploratory path to compare data observed in different views or found in different databases. Both features are provided through our visualization approach.

The user study mainly includes expert users as part of a formative evaluation. Despite our user study being limited by the 10-participants sample, which might be considered small for raising more general conclusions, the assessment was overall positive and allowed us to validate the core concepts related to data integration, exploratory process, and query composition and management. We rely on the concept of data

saturation in qualitative studies (Guest et al., 2006), which allowed us to cover most relevant aspects during the in-depth interviews. For the *exploration process* and *query composition and management* dimensions, we could observe a consistent patterns of recurrent comments that quickly converge; for the *data integration* dimension we received very few comments, all positive, which may suggest that it does not represent a big issue for the participants. However, the results regarding the *usability* dimension are much more varied and reflect individual perceptions of the tool. Nonetheless, we believe that the usability issues do not compromise the validity of the core concepts of our approach, which we sought to validate with our research questions.

6. Conclusions and Future Work

The approach presented in this paper supports data from different sources and the exploration of multidimensional networks from multiple perspectives shown by various visualization techniques. The on-the-fly data processing detects the relationships between the dataset items and their attributes before visually integrating them in the exploration flow. Furthermore, a server supports the querying of multiple SPARQL endpoints, allowing data retrieval from different datasets at runtime. Hence, the user can explore new hypotheses on the data and bring new data to the process to expand and improve the ongoing analysis. We illustrate the use of the tool on two datasets. However, the tool is generic and can accommodate queries and visualizations to any linked open data available through a SPARQL endpoint. A video demonstrating the use of the tool can be found at `https://youtu.be/CA1AfQlagOE`.

The results of our user study are encouraging as they reveal the importance of our approach for exploring large LOD datasets, comparing data from different sources, and using multiple visualization techniques to explore the different perspectives of a dataset. This study allowed us to validate the positive perception of the core concepts of the approach. While we cannot claim to have covered all usability issues, our findings during this formative evaluation are exactly what we expected at this stage of the project. Currently we are fixing the usability issues for making the tool suitable for a quantitative study in the near future.

This work is an important step toward understanding visualization flows in linked open data. The use of chained views enables collecting and representing provenance information. Currently, the tool records information regarding the SPARQL result set, the transformed data, the chosen views and visualization techniques, the subsets of data created on the fly, and the order in which these elements were instantiated throughout the exploration process. This information could be extracted and studied to understand the exploration patterns of different users in different contexts, which could assist the development of visualization recommendation systems to assist users throughout the analysis process by recommending the most suitable visualizations for solving different tasks or exploring different datasets based on types of data (e.g. network, hierarchical) or application domain.

Future work also includes extending the proposed method to support the exploration of different types of open data accessible through querying approaches (e.g., relational databases) to enrich the analysis with complementary data. Finally, we would like to perform a usage study of chained views and follow-up queries to understand the variations of strategies that real users might develop to find information using these methods. This would help us to improve the user experience by adjusting the exploration process according to users' usage patterns and exploration needs.

Acknowledgments

We acknowledge the I3S laboratory for funding the internship of Minh Nhat Do. We gratefully acknowledge the participants of our user study for their time and valuable feedback that allowed us to strength the results of this research. This work is also partially funded by University of Côte d'Azur through its IDEX$^{\text{JEDI}}$ program (CC : C870A06232 EOTP : LINKED_OPEN_DATA DF : D103).

References

Alagiannis, I., Borovica, R., Branco, M., Idreos, S., and Ailamaki, A. (2012). Nodb: Efficient query execution on raw data files. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, page 241–252, New York, NY, USA. Association for Computing Machinery.

Antoniazzi, F. and Viola, F. (2018). RDF graph visualization tools: A survey. In *2018 23rd Conference of Open Innovations Association (FRUCT)*, pages 25–36. IEEE.

Anutariya, C. and Dangol, R. (2018). Vizlod: Schema extraction and visualization of linked open data. In *2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 1–6. IEEE.

Beyer, J., Al-Awami, A., Kasthuri, N., Lichtman, J. W., Pfister, H., and Hadwiger, M. (2013). ConnectomeExplorer: Query-Guided Visual Analysis of Large Volumetric Neuroscience Data. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2868–2877.

Bikakis, N. (2019). Big Data Visualization Tools. In Sakr, S. and Zomaya, A. Y., editors, *Encyclopedia of Big Data Technologies*, pages 336–340. Springer International Publishing, Cham.

Bikakis, N., Papastefanatos, G., Skourla, M., and Sellis, T. (2017). A hierarchical aggregation framework for efficient multilevel visual exploration and analysis. *Semantic Web*, 8(1):139–179. Publisher: IOS Press.

Bikakis, N. and Sellis, T. K. (2016). Exploration and visualization in the web of big linked data: A survey of the state of the art. *CoRR*, abs/1601.08059.

Brunetti, J. M., Auer, S., García, R., Klímek, J., and Nečaský, M. (2013). Formal linked data visualization model. In *Proceedings of International Conference on Information Integration and Web-Based Applications Services*, IIWAS '13, page 309–318, New York, NY, USA. Association for Computing Machinery.

Burmester, M., Mast, M., Jäger, K., and Homans, H. (2010). Valence method for formative evaluation of user experience. In *Proceedings of the 8th ACM conference on Designing Interactive Systems*, pages 364–367, Aarhus, Denmark. ACM.

Cava, R. and Freitas, C. D. S. (2013). Glyphs in matrix representation of graphs for displaying soccer games results. In *The 1st Workshop on Sports Data Visualization. IEEE*, volume 13, page 15.

Cava, R., Freitas, C. M., Barboni, E., Palanque, P., and Winckler, M. (2014). Inside-in search: an alternative for performing ancillary search tasks on the web. In *2014 9th Latin American Web Congress*, pages 91–99. IEEE.

Cava, R., Freitas, C. M. D. S., and Winckler, M. (2017). Clustervis: visualizing nodes attributes in multivariate graphs. In *Proceedings of the Symposium on Applied Computing*, pages 174–179.

Chawuthai, R. and Takeda, H. (2015). Rdf graph visualization by interpreting linked data as knowledge. In *Joint International Semantic Technology Conference*, pages 23–39. Springer.

Chen, S., Lin, L., and Yuan, X. (2017). Social media visual analytics. *Computer Graphics Forum*, 36(3):563–587.

Dadzie, A.-S. and Pietriga, E. (2016). Visualisation of linked data - reprise. *Semantic Web*,

8(1):1–21.

Dadzie, A.-S. and Rowe, M. (2011). Approaches to visualising linked data: A survey. *Semantic Web*, 2(2):89–124.

De Vocht, L., Dimou, A., Breuer, J., Van Compernolle, M., Verborgh, R., Mannens, E., Mechant, P., and Van de Walle, R. (2015). A visual exploration workflow as enabler for the exploitation of linked open data. In *IESD'14 Proceedings of the 3rd International Conference on Intelligent Exploration of Semantic Data*, volume 1279, pages 30–41. CER-WS. org.

Deligiannidis, L., Kochut, K. J., and Sheth, A. P. (2007). Rdf data exploration and visualization. In *Proceedings of the ACM first workshop on CyberInfrastructure: information management in eScience*, pages 39–46.

Derthick, M., Kolojejchick, J., and Roth, S. F. (1997). An Interactive Visual Query Environment for Exploring Data. In *Proceedings of the 10th Annual ACM Symposium on User Interface Software and Technology*, UIST '97, page 189–198, New York, NY, USA. Association for Computing Machinery.

Destandau, M., Appert, C., and Pietriga, E. (2021). S-Paths: Set-based visual exploration of linked data driven by semantic paths. *Semantic Web*, pages 1–18.

Dunne, C., Henry Riche, N., Lee, B., Metoyer, R., and Robertson, G. (2012). *GraphTrail: Analyzing Large Multivariate, Heterogeneous Networks While Supporting Exploration History*, page 1663–1672. Association for Computing Machinery, New York, NY, USA.

Elmqvist, N. and Fekete, J.-D. (2010). Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *IEEE Transactions on Visualization and Computer Graphics*, 16(3):439–454.

Fisher, D., Popov, I., Drucker, S., and schraefel, m. (2012). *Trust Me, i'm Partially Right: Incremental Visualization Lets Analysts Explore Large Datasets Faster*, page 1673–1682. Association for Computing Machinery, New York, NY, USA.

Gratzl, S., Gehlenborg, N., Lex, A., Pfister, H., and Streit, M. (2014). Domino: Extracting, comparing, and manipulating subsets across multiple tabular datasets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2023–2032.

Graziosi, A., Di Iorio, A., Poggi, F., Peroni, S., and Bonini, L. (2018). Customising lod views: a declarative approach. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, pages 2185–2192.

Guest, G., Bunce, A., and Johnson, L. (2006). How many interviews are enough?: An experiment with data saturation and variability. *Field Methods*, 18(1):59–82.

Heer, J., Agrawala, M., and Willett, W. (2008). Generalized selection via interactive query relaxation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, page 959–968, New York, NY, USA. Association for Computing Machinery.

Jacksi, K., Zeebaree, S. R., and Dimililer, N. (2018). Lod explorer: Presenting the web of data. *Int. J. Adv. Comput. Sci. Appl. IJACSA*, 9(1).

Kalinin, A., Cetintemel, U., and Zdonik, S. (2014). Interactive data exploration using semantic windows. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, page 505–516, New York, NY, USA. Association for Computing Machinery.

Key, A., Howe, B., Perry, D., and Aragon, C. (2012). Vizdeck: Self-organizing dashboards for visual analytics. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, page 681–684, New York, NY, USA. Association for Computing Machinery.

Ko, S., Cho, I., Afzal, S., Yau, C., Chae, J., Malik, A., Beck, K., Jang, Y., Ribarsky, W., and Ebert, D. S. (2016). A survey on visual analysis approaches for financial data. *Computer Graphics Forum*, 35(3):599–617.

Kremen, P., Saeeda, L., and Blasko, M. (2018). Dataset dashboard-a sparql endpoint explorer. In *VOILA@ ISWC*, pages 70–77.

Leng, J. (2011). *Handbook of Research on Computational Science and Engineering: Theory and Practice: Theory and Practice*, volume 2. IGI.

Livny, M., Ramakrishnan, R., Beyer, K., Chen, G., Donjerkovic, D., Lawande, S., Myllymaki, J., and Wenger, K. (1997). Devise: Integrated querying and visual exploration of large datasets. In *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, SIGMOD '97, page 301–312, New York, NY, USA. Association for Computing Machinery.

Menin, A., Cava, R., Freitas, C. M. D. S., Corby, O., and Winckler, M. (2021a). Towards a Visual Approach for Representing Analytical Provenance in Exploration Processes. In *25th International Conference Information Visualisation*.

Menin, A., Faron, C., Corby, O., Freitas, C., Gandon, F., and Winckler, M. (2021b). From Linked Data Querying to Visual Search: Towards a Visualization Pipeline for LOD Exploration. In *In Proceedings of the 17th International Conference on Web Information Systems and Technologies (WEBIST 2021)*.

Munzner, T. (2014). *Visualization Analysis and Design*. A.K. Peters visualization series. A K Peters.

North, C., Chang, R., Endert, A., Dou, W., May, R., Pike, B., and Fink, G. A. (2011). Analytic provenance: process+interaction+insight. In Tan, D. S., Amershi, S., Begole, B., Kellogg, W. A., and Tungare, M., editors, *Proceedings of the International Conference on Human Factors in Computing Systems, CHI 2011, Extended Abstracts Volume, Vancouver, BC, Canada, May 7-12, 2011*, pages 33–36. ACM.

Olma, M., Karpathiotakis, M., Alagiannis, I., Athanassoulis, M., and Ailamaki, A. (2017). Slalom: Coasting through raw data via adaptive partitioning and indexing. *Proc. VLDB Endow.*, 10(10):1106–1117.

Olsten, C., Stonebraker, M., Aiken, A., and Hellerstein, J. (1998). VIQING: visual interactive querying. In *Proceedings. 1998 IEEE Symposium on Visual Languages (Cat. No.98TB100254)*, pages 162–169.

Park, Y., Cafarella, M., and Mozafari, B. (2016). Visualization-aware sampling for very large databases. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 755–766.

Preim, B. and Lawonn, K. (2020). A survey of visual analytics for public health. *Computer Graphics Forum*, 39(1):543–580.

Sacha, D., Kraus, M., Bernard, J., Behrisch, M., Schreck, T., Asano, Y., and Keim, D. A. (2017). Somflow: Guided exploratory cluster analysis with self-organizing maps and analytic provenance. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):120–130.

Shneiderman, B., Williamson, C., and Ahlberg, C. (1992). Dynamic queries: Database searching by direct manipulation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '92, page 669–670, New York, NY, USA. Association for Computing Machinery.

Stolper, C. D., Perer, A., and Gotz, D. (2014). Progressive visual analytics: User-driven visual exploration of in-progress analytics. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1653–1662.

Stolte, C., Tang, D., and Hanrahan, P. (2002). Polaris: a system for query, analysis, and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):52–65.

Viau, C. and McGuffin, M. J. (2012). Connectedcharts: explicit visualization of relationships between data graphics. *Computer Graphics Forum*, 31(3pt4):1285–1294.

Author Biography

Aline Menin is a Postdoctoral Researcher at the Wimmics/SPARKS team of University of Côte d'Azur, CNRS, Inria in France. Her interests are Visualization, Visual analytics, Human-Computer Interaction, Geovisualization, and User-Centered Design. **Minh Nhat Do** is a second-year student of the master's program in Data Science

and Artificial Intelligence at University of Côte d'Azur. **Carla Dal Sasso Freitas** is a full professor at the Institute of Informatics, Federal University of Rio Grande do Sul. Her general research theme is interactive data visualization, mainly novel visualization techniques, evaluation of 2D and 3D interactions in the context of data visualization, and more recently on immersive analytics. **Olivier Corby** is an Inria Researcher at the Wimmics/SPARKS team of University of Côte d'Azur, CNRS, Inria in France. His research interests are Semantic Web of Linked Data & Knowledge Representation and Reasoning. He is also interested in Graph based Knowledge Representation, RDF/S, SPARQL. **Catherine Faron** is an Assistant Professor at University of Côte d'Azur and a permanent researcher at the Wimmics team of University of Côte d'Azur, CNRS, Inria in France. Her research interests are Knowledge Engineering and Modeling, Ontologies, Graph based Knowledge Representation and Reasoning, Semantic Web. **Alain Giboin** is an emeritus Inria Researcher at the Wimmics/SPARKS team of University of Côte d'Azur, CNRS, Inria in France. His research interests are ergonomics/interaction design of Human-Computer Interaction and Human-Computer-Human Interaction, User Interaction with and through the Social Semantic Web, Human-Data Interaction, and Knowledge Engineering. **Marco Winckler** is a Full Professor at University of Côte d'Azur and a permanent researcher at the Wimmics/SPARKS team of University of Côte d'Azur, CNRS, Inria in France. His research interests are Engineering of Interactive Systems, Web Engineering, and Information Visualization.