Understanding Interlinked Data

– Visualising, Exploring, and Analysing Ontologies –

Olaf Noppens and **Thorsten Liebig**
(Ulm University, Germany
olaf.noppens@uni-ulm.de, thorsten.liebig@uni-ulm.de)

Abstract Companies are faced with managing as well as integrating large collections of distributed data today. Here, the challenging task is not to store these volumes of structured and interlinked data but to understand and analyze its explicit or implicit relationships. However, up to date there is virtually no support in navigating, visualizing or even analyzing structured data sets of this size appropriately. This paper describes novel rendering techniques enabling a new level of visual analytics combined with interactive exploration principles. The underlying visualization rationale is driven by the principle of providing detail information with respect to qualitative as well as quantitative aspects on user demand while offering an overview at any time. By means of our prototypical implementation and a real-world data set we show how to answer several data specific tasks by interactive visual exploration.

Key Words: metadata, visual analytics, interactive exploration
Category: I.2.4, H.3.3, H.5.2

1 Motivation

Making value from collecting and integrating distributed and heterogeneous data is a critical factor of business success. In fact, the underlying problem is a very general kind of challenge of our information society and closely related to community efforts such as DBpedia [Auer et al., 2007] or YAGO [Suchanek et al., 2007] which have recently extracted large volumes of structured data from the Web (Wikipedia, US Census Data, WordNet, etc.). Those repositories are extreme in the sense that they are extraordinary in size and dominated by interlinked data incorporating only a small and typically lightweight schema. However, there is currently only little or no adequate support in navigating, visualizing or even analyzing large volumes of interlinked data in an reasonable way.

In this paper we present our approach which combines techniques from visual analytics and interactive exploration suitable even to grasp large volumes of heavily interrelated data sets. The approach has been implemented and integrated into our ontology authoring framework OntoTrack [Liebig and Noppens, 2005]. The following describes the various selection, exploration and analysis techniques with help of an example data set introduced in the next section.

2 The MONDIAL Database

For illustrating purposes we have chosen a simple but real-world data set, namely the Mondial Database[1] (MONDIAL). It consists of a collection of geographic information compiled from different Web data sources such as the CIA World Factbook, Global Statistics and the Terra database [May, 1999]. The core of a MONDIAL record consists of data about countries, cities as well as deserts, rivers, or ethnic groups mainly collected from the World Factbook. In addition the collection includes statistical data about populations, area, or length. Entities are typed in a lightweight manner with respect to common geographical concepts such as countries, rivers etc. The most substantial information, however, is expressed by the various relationships which relate entities among each other. For instance, the relationship `has-City` relates countries to cities, `flows-through-country` tells us through which countries a river flows to, etc.

3 Related Work

Understanding knowledge from a visual perspective is a challenge when it comes to large amounts of entities as well as interconnections. Whereas charts help to gain insights to numerical data sets they are not appropriate to depict qualitative relationships.

Current tools for the latter task typically try to combine visualization with interaction in a sense that users can chose between sets of entities they want to inspect or types of relationships they wants to investigate.

A representative of the former is the MatrixBrowser tool [Ziegler et al., 2002]. It allows to select two sets of entities in order to visualize the existing relationships between the set members with help of a 2D matrix as shown in Fig. 1. Here, the x-axis lists all countries and the y-axis all rivers of the MONDIAL data set. A relationship between a specific river and various countries is depicted by a mark in the corresponding intersecting field and connecting horizontal and vertical lines. For instance, Fig. 1 shows the relation `flows-through-country` between the river Argun and Russia. However, in case of only sparsely related individuals a lot of space is left unused. In addition, large data sets can not be displayed on one screen and hamper to grasp the 'big picture'. Furthermore, due to the nature of a two-dimensional matrix, only the relationships between two kinds of characteristics can be visualized and analyzed at once.

A different paradigm, namely a network centered approach is implemented within Welkin [Mazzocchi and Ciccarese, 2007]. Welkin as well as other RDF visualizing tools such as RDF-Gravity and ISAViz utilize either spring, circle, or tree layout techniques for rendering graphs. Typically the user selects a type

[1] http://www.dbis.informatik.uni-goettingen.de/Mondial/

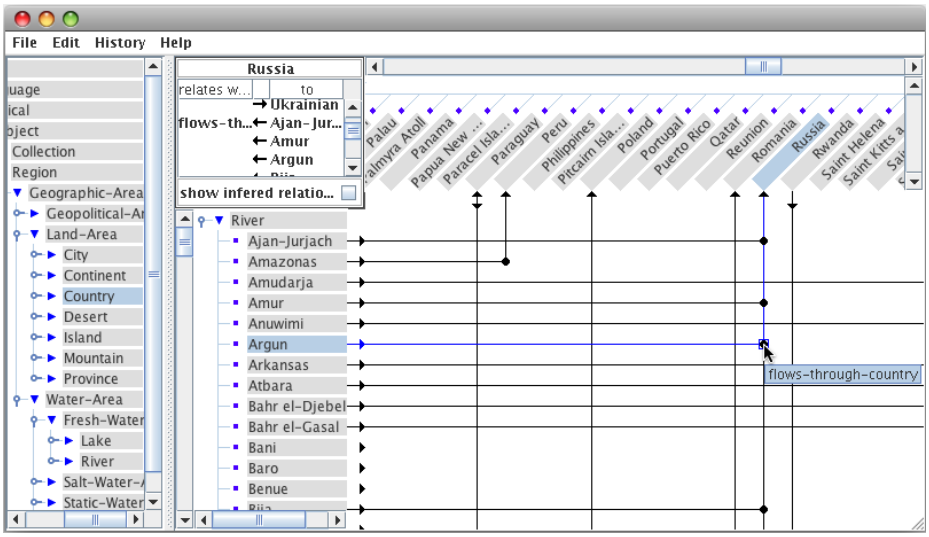

Figure 1: MatrixBrowser visualization of rivers flowing through countries.

of relationship and the corresponding network renders all existing connections of this type. For instance, Fig. 2 shows the same correlation of rivers to countries with respect to the `flows-through-country` relationship as in Fig. 1. This approach obviously is able to show the connectivity within larger data sets. However, it lacks of quantitative information, may distract in case of many interconnections, and provides no clear distinction between different types of entities.

For the purpose of comparison see the left hand side of Fig. 6 in order to get an idea how the aforementioned coherences are displayed within our approach.

4 Visual Analysis through Interactive Exploration

When visualizing ontologies one needs to take their inherent characteristics into account. Therefore our novel visualization paradigm tries to consider *qualitative* aspects such as clustering of individuals with respect to explicit/implicit concept and relationship assertions as well as *quantitative* aspects as shown in the following.

4.1 Visualizing Qualities

Several studies have shown that it is not advisable to arbitrarily visualize both all dependencies and all particulars at any time [Keim, 2001]. Therefore, in order to

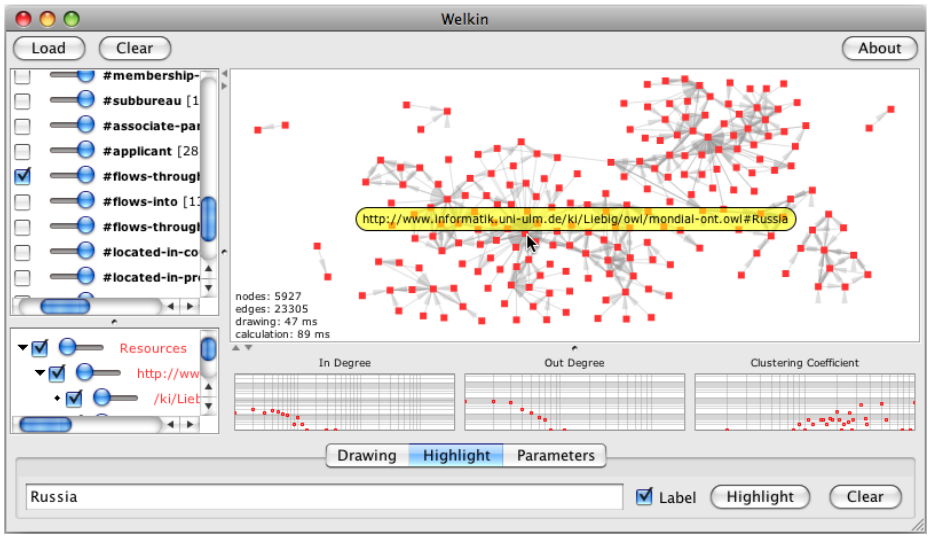

Figure 2: Welkin visualizer with network layout between rivers and countries.

prevent the user in being overwhelmed with currently non-relevant information pieces our user-directed interactive exploration strategy allows for focusing on relevant parts of an ontology, or fractions thereof which promise to unveil deeper insights. Initially, one can either start with a user selected entity (e. g. as the result of a query) or with entities belonging to the same concept such as all European capitals. As the visualization and analysis component is integrated into our OntoTrack framework the latter task can be carried out by dragging a concept from the schema representation pane on the data analysis pane.

We believe that, from the user's point of view, entities with similar characteristics should build obvious clusters from a visual perspective. These clusters are built either from concept or relationship assertions (implicit or explicit). For instance, in the MONDIAL domain all European capitals and all countries to which these capitals belong to can be pooled within a cluster as shown in Fig. 3. Here, entities are visualized as small filled circles within clusters.

A second dimension of abstraction is used to draw entities in a cluster only if the amount of entities is below a user-definable limit. To easily compare the amount in that cluster its diameter approximates the amount of entities. In addition, the number of fillers are drawn on a cluster as it can be seen in the middle cluster in Fig. 5. Additional detail information for each entity such as an image, information about the population of countries etc. in the MONDIAL domain is provided in an optional scroll-able list as shown on the left hand side

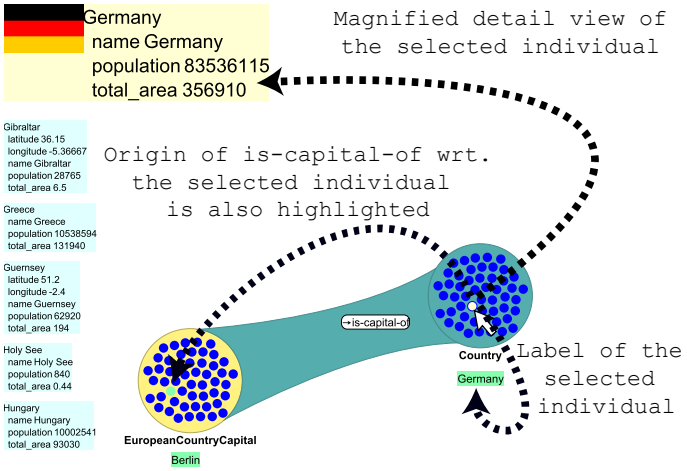

Figure 3: Clustering and club visualization.

of Fig. 3.

Instead of given each relationship between single entities a first-class visual representation and overloading the screen area, entities are grouped by their connecting relationship(s). Relationships are represented by clubs originating from the set of entities which are considered as the relationship's subject to its objects. This "club visualization" metaphor can be seen in Fig. 3. Here, the club connects the cluster containing all European capitals and all countries to which these capitals belong to, i. e. the cluster sets are connected via the `is-capital-of` relationship. In addition, not only the union of all entities in a cluster can form the origin of a club but also single entities as shown in Fig. 5.

A graphical radial preview menu of related entities grouped by their connecting relationships guides the user through the next exploration steps after clicking on the graphical representation of an entity or a cluster. In order to allow a more flexible exploration, the exploration direction is not limited to the defined direction of the relationship. For instance, the club shown in Fig. 4 represents all countries (right-hand side of the club) that are bordered to European countries (left-hand side). Here, the preview displays different relationships such as `ethnicgroups` or `encompassed` and one can easily grasp that in Armenia there are 3 ethnic groups, or in ontological terms: "Armenia" is related to 3 entities with respect to the relationship `ethnicgroups`. Note that the preview also shows the hierarchical structure of the relationships: `has-city` is a super-relationship of `has-province-capital`. The implied direction of an expansion is denoted by an arrow sign next to the relationship's label.

Each cluster as well as each (visible) entity can serve as a follow-up point for

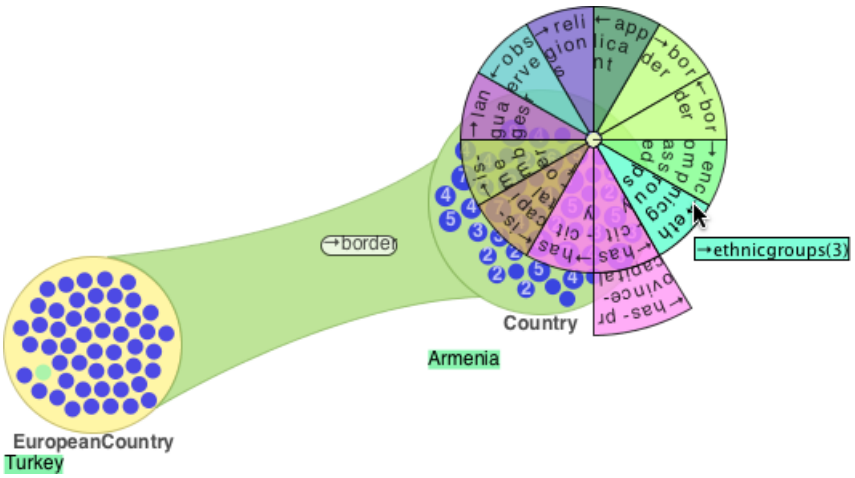

Figure 4: Previews for connected entities grouped by relationships.

further expansions to allow not only parallel expansion paths but also chains of expansions. Note that in a chain of expansion the result cluster of the previous query (i. e. expansion) is used as follow-up point for the next expansion step. For instance, Fig. 5 shows the religions worshiped in countries encompassed by Europe (expansion chain) as well as all continents those countries belong to in parallel.

To easily grasp the related entity/entities to a given one with respect to expanded relationships all entities of the origin of the club and, if further club expansion is available, of the entities of the corresponding relationship wrt. the highlighted entity are recursively highlighted, too. For instance, in Fig. 3 the mouse pointer is hovering over "Germany". As a result its label is rendered at the bottom of the cluster and because "Berlin" is the only origin of the `is-capital-of` property the corresponding graphical representation is also highlighted and its label also rendered at the bottom of the `EuropeanCountryCapital` cluster.

4.2 Visualizing Quantities

Besides qualities, the representation of quantities is another important dimension of visual analytics: we found out that quite a number of queries inherently require to take quantities into account. Even if one can easily grasp how many entities are related to a given one with respect to a specific source by manually counting them, it is more complex to visually answer how many entities in a cluster are related to a specific one within its successor cluster. It is also not enough to count them because it is more a question about the distribution of entities.

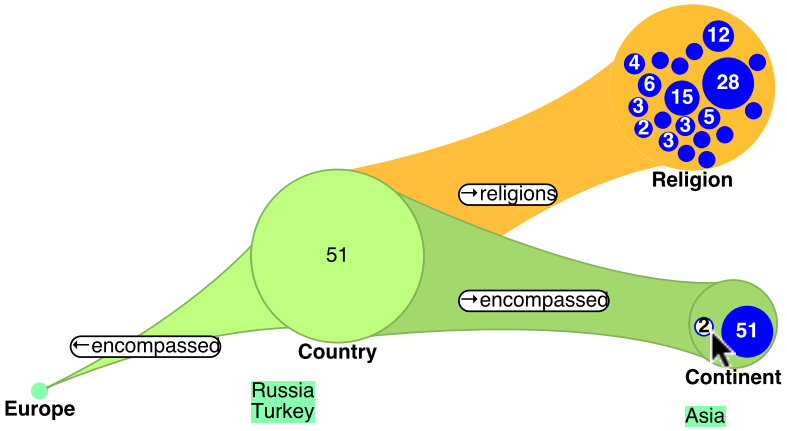

Figure 5: Multiple expansion paths.

For instance, to answer the question within the MONDIAL domain which are the countries with the most rivers flowing through, one would start with the cluster representing all rivers. After expanding the club connected via the `flows-through-country` relationship one gets a cluster consisting of all corresponding countries as shown in Fig 6. Note that one can also recognize the distribution of countries to rivers as well as interactively get the corresponding rivers. Derived from well-known methods from visual analytics we have implemented a simple but powerful solution: the diameters of each country circle scales proportionally with the number of related deserts in the predecessor club. In case that there is more than one single source entity their number is also drawn within the entity circle as shown in Fig. 6. For instance, 29 rivers flow through Russia. One can also figure out on which continents the river flows. By expanding the country cluster with respect to the `encompassed` relationship it shows a follow-up quantity, namely the distribution of those countries to encompassing continents. Consider another example shown in Fig. 5. Here one can easily grasp the distribution of European countries to continents and one can see that two countries, namely Russia and Turkey, belongs also to the Asian continent.

5 Conclusion

In this paper we introduced our "club visualization" metaphor which combines established methods from visual exploration and visual analytics in order to discover hidden connections between individuals while not disturbing the user when

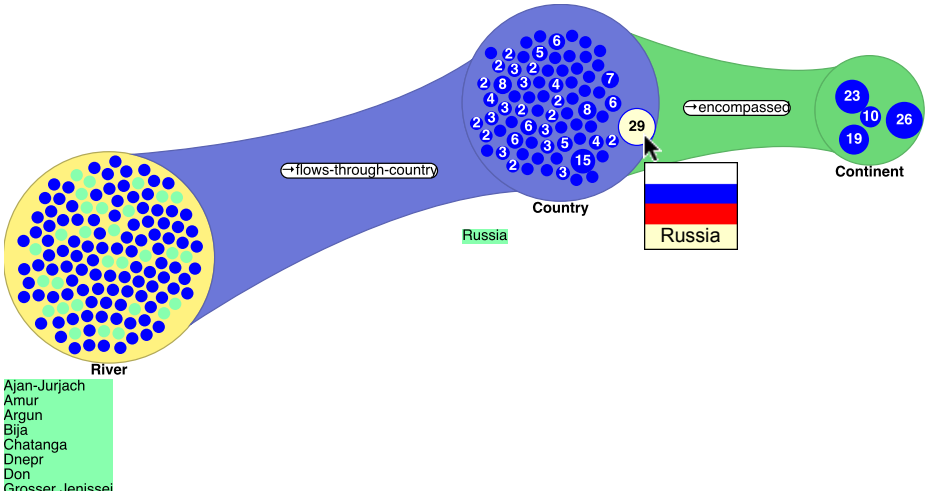

Figure 6: "Which is the country with the most rivers flowing through?"

exploring large ontologies. A visual feedback about quantities and the outlining of related individuals supplement this visualization in order to enable the user to gain deeper insights into large and heavily interrelated volumes of individuals. Our implementation adds new exploration and understanding possibilities not found in currently available tools. In order to have some fundamental qualitative results concerning average navigation and finding of not-obvious correlation we currently conduct a controlled user study.

References

[Auer et al., 2007] Auer, S., Bizer, C., Lehmann, J., Kobilarov, G., Cyganiak, R., and Ives, Z. (2007). DBpedia: A Nucleus for a Web of Open Data. In *Proc. of the 6th International Semantic Web Conference (ISWC 2007)*, pages 722–735. Springer.

[Keim, 2001] Keim, D. A. (2001). Visual exploration of large data sets. *Communications of the ACM*, 44(8):38–44.

[Liebig and Noppens, 2005] Liebig, T. and Noppens, O. (2005). OntoTrack: A semantic approach for ontology authoring. *Journal of Web Semantics*, 3(2):116 – 131.

[May, 1999] May, W. (1999). Information extraction and integration with Florid: The Mondial case study. Technical Report 131, Universität Freiburg.

[Mazzocchi and Ciccarese, 2007] Mazzocchi, S. and Ciccarese, P. (2007). Simile: Welkin. `http://simile.mit.edu/welkin/`.

[Suchanek et al., 2007] Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. In *Proc. of the WWW 2007*, pages 697–706, Banff, AL, Canada. ACM Press.

[Ziegler et al., 2002] Ziegler, J., Kunz, C., Botsch, V., and Schneeberger, J. (2002). Visualizing and exploring large networked information spaces with matrix browser. In *Proc. of the 6th Int. Conf. on Information Visualization (IV'02)*, pages 361–366.