Towards Visualization Recommendation – A Semi-Automated Domain-Specific Learning Approach

Pawandeep Kaur
Heinz-Nixdorf Chair for Distributed Information Systems
Friedrich-Schiller-Universität, Jena
pawandeep.kaur@uni-jena.de

Michael Owonibi
Heinz-Nixdorf Chair for Distributed Information Systems
Friedrich-Schiller-Universität, Jena
michael.owonibi@uni-jena.de

Birgitta Koenig-Ries
Heinz-Nixdorf Chair for Distributed Information Systems
Friedrich-Schiller-Universität, Jena
birgitta.koenig-ries@uni-jena.de

ABSTRACT

Information visualization is important in science as it helps scientists in exploring, analysing, and presenting both the obvious and less obvious features of their datasets. However, scientists are not typically visualization experts. It is therefore difficult and time-consuming for them to choose the optimal visualization to convey the desired message. To provide a solution for this problem of visualization selection, we propose a semi-automated, context aware visualization recommendation model. In the model, information will be extracted from data and metadata, the latter providing relevant context. This information will be annotated with suitable domain specific operations (like rank abundance), which will be mapped to the relevant visualizations. We also propose an interactive learning workflow for visualization recommendation that will enrich the model from the knowledge gathered from the interaction with the user. We will use biodiversity research as the application domain to guide the concrete instantiation of our approach and its evaluation.

Categories and Subject Descriptors

D.2.12 [**Data mapping**]

General Terms

Human Factors, Design

Keywords

Data Visualization, Machine Learning, Biodiversity Informatics, Text Mining, Recommender Systems

1. INTRODUCTION

The human brain can comprehend images a lot easier than words or numbers. This makes effective graphics an especially important part of academic literature [19]. Visualization that condenses large amounts of data into effective and understandable graphics is therefore an important component of the presentation and communication of scientific research [14]. Supporting scientists in choosing the appropriate visualization during the research process is very important. We believe that an optimal choice leads to more interpretable graphics which keep the reader interested in the publication, and make them understand the research work and possibly build on it. Ultimately, this results in increased citation of such publications. In addition, it aids researchers in detecting recurring patterns, formulating hypotheses and discovering new knowledge out of those patterns [24].

In this paper, we will focus on the issue of visualization selection for data presentation and will be using biodiversity research as an application domain. In the next section, we will first explain the biodiversity research domain and then analyze the challenges and requirements of researchers with respect to the visualization selection. Then, we will present the literature review of the existing solutions (Section 3). In Section 4, we will present our approach to address to the challenges that we have identified.

2. REQUIREMENTS ANALYSIS

Biodiversity research aims to understand the enormous diversity of life on earth and to identify the factors and interactions that generate and maintain this diversity [20]. Biodiversity data is the data accumulated from the research done by biologists and ecologists on different taxa and levels, land use, and ecosystem processes. For proper preservation, reusability, and sharing of such data, metadata is provided along with the data. This metadata contains vital contextual information related to the datasets like purpose of the research work, data collection method and other important keywords. In order to answer the most relevant questions of biodiversity research, synthesis of data stemming from integration of datasets from different experiments or observation series is frequently needed. Collaborative projects thus tend to enforce centralized data management. This is true, e.g., for the Biodiversity Exploratories [16], a large scale, long-term project funded by DFG. The Exploratories use the BExIS platform [15] for central data management. The instance of BExIS used within the Biodiversity Exploratiories (BE) serves as one of the primary sources for collecting requirements for this study. The large collection of data available in the BE BExIS is the result of research activities by many disciplines involved in biodiversity science over the last eight years. This data is highly complex, heterogeneous and often not easy to understand. To interpret, analyze, present, and reuse such data a system is required that can analyze and visualize these datasets effectively.

According to the survey of 57 journals conducted in [21], natural science journals use far more graphs than mathematical or social science journals [21]. The objective of any graphics in the context of scientific publications and presentations is to effectively communicate information [19]. For that, it is important to choose

27th GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), 26.05.2015 - 29.05.2015, Magdeburg, Germany.

the appropriate visualization with respect to available data and message to convey. However, the studies [23] have shown that the potential of visualization has not been fully utilized in scientific journals. In [23] Lauren et al identify two main reasons for this failure: scientists are overwhelmed by the numerous visualization techniques available and they lack expertise in designing graphs.

In general, a visualization process is considered as a 'search' process in which the user makes a decision about visualization tools and techniques at first, after which other decisions are made about different controls like layout, structure etc. until a satisfactory visualization is produced [13]. With the growing amount of data and increasing availability of different visualization techniques this 'search' space becomes wider [13]. In order to successfully execute this search process, one needs to have clear knowledge about the information contained in the data, the message that should be conveyed and the semantics of different visualizations.

We argue here that to understand this complex process and then work aptly, one needs to have some visualization expertise. However, scientists typically do not have the proficiency to manipulate the programs and design successful graphs [22]. Interactive visualization approaches make the visualization creation process more adaptive, but, due to their insufficient knowledge, scientists often have difficulties in mapping the data elements to graphical attributes [12]. The result of inappropriate mapping can impede analysis and even result in misleading conclusions [1].

Furthermore, matters related to visualization are made even more complex by human perception subjectivity [9], which means people perceive the same thing differently under different circumstances. For better understanding, readers primarily need to relate the visualization to the realm of their existing knowledge domain [2]. To ensure that the chosen visualization does indeed convey the intended message to the target readers, a model like the one proposed in [6] should be the base of visualization design.
.

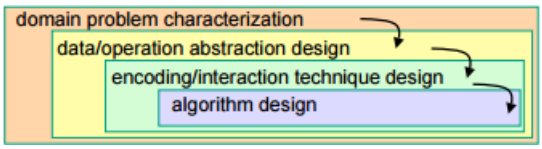

Figure 2. Nested Model for Visualization Design [6]

This model, as depicted in Figure 2, divides the design process into four levels which are: 1) characterize the tasks and data in the vocabulary of the problem domain, 2) abstract this information into visual operations and data types, 3) design visual encoding and interaction techniques, and lastly create algorithms to execute these techniques efficiently.

An approach as depicted in model above, needs to rely on the domain knowledge and visualization used in that domain. In Section 4 we will propose such an approach.

3. STATE OF THE ART

The literature on visualization recommendation can be found from the early Eighties of the last century. The earliest such work is BHARAT [25], APT [26], Vis-WIZZ [27], Vista [28] and ViA [29]. BHARAT, APT and VIA have a similar direction: They all aim at encoding the data variables to the visual clues, human perception analysis, exploit the knowledge of graphic designs and displays. Such work was independent of any domain. Vis-WIZZ and VISTA have noticed the need of knowledge accelerated visualization mapping techniques, but their research is limited to numerical or quantitative data. Casner's BOZ system [5] analyses task descriptions to generate corresponding visualizations. However, the task first needs to be fed manually to the mapping engine. Many Eyes [10] by IBM which uses the rapidly adaptive visualization engine (REVA) based on the grammar of graphics by Leland Wilkinson [11] is an example of commercial approaches in this area. Similarly, Polaris' work on Visual query language (VISQL) is used in the Show_me data module of the Tableau [17] software. Both of these approaches do not consider contextual information for recommending visualization.

PRAVDA (Perceptual Rule-Based Architecture for Visualizing Data Accurately) [4] introduced a rule based architecture for assisting the user in making choices of visualization color parameters. The appropriate visualization rule is selected based on higher-level abstractions of the data, i.e., metadata. They were the first who introduced knowledge from the metadata into the visualization process.

Current knowledge-based visualization approaches are highly interactive [3] and use semantics from different ontologies to annotate visual and data components (see, e.g., Gilson at. al [8]). They extract the semantic information from the input data and try to find the best match by mapping three different ontologies, where one is the domain ontology, another is the visualization ontology and the last one is their own ontology which is created by mapping first two.

Though knowledge-based systems reduce the burden placed upon users to acquire knowledge about complex visualization techniques, they lack expert knowledge [13]. Such solutions should be based on some ground truth collected from relevant domain experts. Additionally, we argue that limited user interaction to obtain feedback would be useful to enhance the knowledge base.

4. PROPOSED APPROACH

Based on the requirements identified in Section 2 and the shortcomings of existing approaches discussed in Section 3, we propose a visualization recommendation model which will help scientists in making appropriate choices for presenting their data.
It will be based on a knowledgebase created by reviewing the visualizations presented in biodiversity publications. Such knowledge will enrich our understanding on current trends in visualizations for representing biodiversity data. It will also enhance the system with scientific operations and concepts and variables used in the presenting those concepts. We will be extracting information from metadata (which contains a description of various characteristics of the data and the context of the data collection and usage), integrating the knowledge from the domain vocabularies, and classifying this information with respect to the visual operations performed on the dataset. The knowledge obtained in this way will serve as a key parameter in recommending visualization.

In addition, to deal with the problem of human perception subjectivity, we propose an interactive machine learning approach for visualization recommendation. We will track the input from the user at each interaction and will update that into the respective modules in mapping engine. This will make system learn from the user interaction. However, users will be only prompted to interact in case they do not get satisfactory results. Thus for a non-computer experts (biodiversity researchers in our case) the interaction would be nil, if his choice of result is present in our recommended list.

In general, our approach is made up of two main components namely: the Visualization Mapping Model (Section 4.1) and the Interactive Learning workflow (Section 4.2). The approach is explained using the metadata of a dataset from the BE BExIS as an example [30].

4.1 Visualization Mapping Model

Figure 3 provides an overview of the visualization mapping model. Each of the five phases identified and marked on the figure will be explained in detail below.

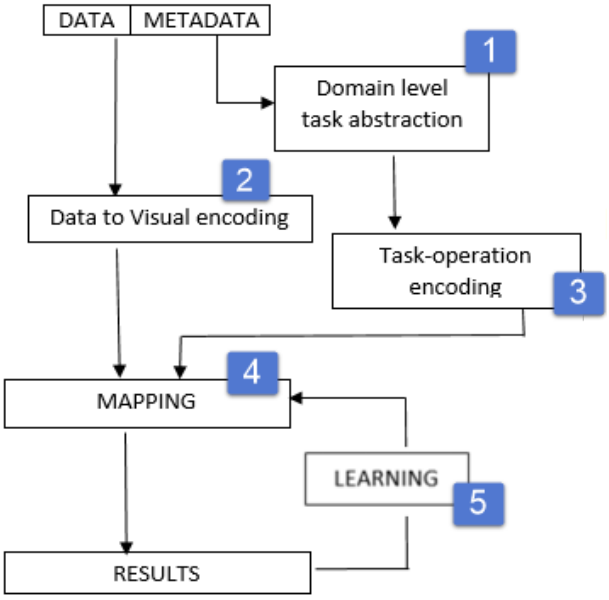

Figure 3: Visualization Mapping Model

1) **Domain level task abstraction:** Task here refers to domain specific analytic operations which are computed on several variables of the dataset in order to derive a concept. For instance, species distribution is an ecological concept which is about computation of distribution (a task) of some species over a geographical area.

To understand the domain problem well, first, we need to understand the dataset, the goals of the data collection, the analysis performed on the data and how these analytical operations can be mapped visually. Metadata provides information about the what, why, when, and who about data and context, methodology, keywords related to dataset and research. Extracting this information from the data and metadata and mapping it with the domain specific vocabularies can reveal the biodiversity related tasks that can be performed on the dataset. As an example, consider the excerpt of metadata of a specific dataset from the BE BExIS [30] shown here:

Detection of forest activities (harvesting / young stock maintenance, etc.) of the forest owner (Forest Service) on the EPs of exploratories. amount and ***spatial distribution*** *of forest harvesting measures by the Forest Service on the EPs.*

To keep it short and precise, just one keyword ("spatial distribution") has been extracted and will be analyzed and processed. By annotating it with terms from an ontology, e.g., the SWEET ontology [7], a relation such as shown in Fig. 4 has been found.

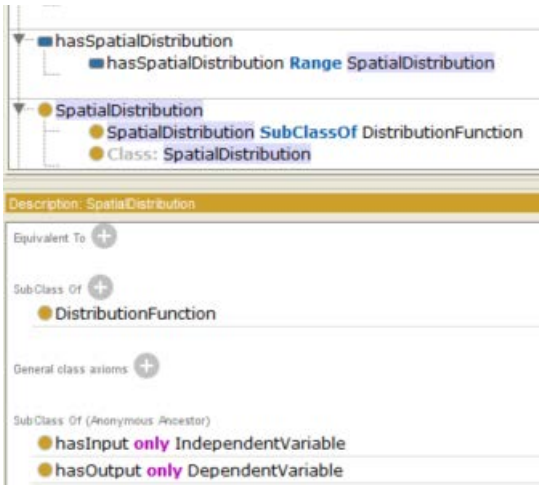

Figure 4 : Keyword Annotation

This annotation can be explained as:

Domain Specific Task: Spatial Distribution on any of the distribution functions like Probability Distribution Function or Chisquare Distribution.

Representational Task : Distribution

Representational Variables: atleast 2 (independent and dependent)

2) Data to visual encoding: Here, we will perform visual encoding on the variables and the values of the dataset. We will map the data variables to their equivalent visual marks/icons/variables (as in Figure 5 [18]) on the basis of some existing classification scheme (such as the one presented in [17]) for graphical presentation. Figure 5 shows how the relationship among various aspects of data can be represented within the visualization. For example, the variable that represent different size elements (like area or length) of some entity could be best represented via bars of different sizes in a visualization.

To give an example, we have used the same dataset as above and have extracted some variables as shown in Table 1. In the visualization creation process, first, the variables are identified with their respective datatypes (measurement units). This we have done and have appended another column as UNIT. Then, by taking a reference from Figure 5, we have transformed these variables into their respective visual icons/variables (shaded column named 'Visual Icons' in the figure). Trees species is a ordinal or categorical variable thus could be best represented as colour, shape or orientation styles. In the same way nominal variables could be used as X,Y scales in a 2-D visualization

In the next steps, we will be using these icons to represent the relations between variables in the visualization.

Table 1 : Dataset variables

name	description	Units	Visual Icons
Tree	Tree species shortcuts	Ordinal	Shape, Color, Orientation
NRderMassnahme	consecutive number of measures	Nominal	Position, Size, Value
Tree Height	Tree Height	Quantitative	Position, Size, Value
PLOT	Number of experimental plots	Nominal	Position, Size, Value

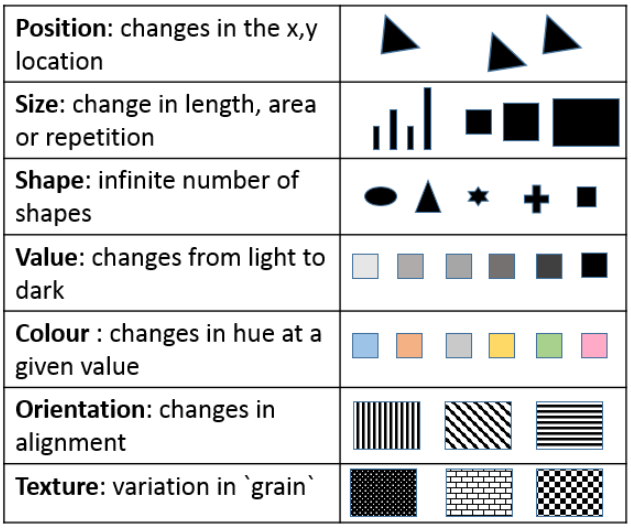

Figure 5: Bertin's Visual Variables [18]

3) Task to operation encoding: At this stage, we will combine the information from the conceptual knowledge gained from metadata and the visual representation knowledge. Visual representation knowledge will be derived by analyzing existing publications. We will be creating a domain knowledgebase about visualizations used in biodiversity publications and will ask scientists to verify it and provide their feedback. This phase is important to get the domain expertise about current visualization trends for representations of different studies. The candidate visualization will be chosen from this knowledgebase according to the domain tasks that we have extracted in Step 1.

In our preliminary work, we have tried to understand the different visualizations used in biodiversity research by reviewing the publications from the information system of the Biodiversity Exploratories. A small sample of the results is depicted in Figure 6. This figure shows what visualization has been used to represent which biodiversity study/analysis within the reviewed publications. Taking the same example as in the previous steps, with the information contained in Figure 6 we can accomplish two jobs: First, we can infer concepts that are related to the identified concept. For example, spatial distribution can be associated with other spatial analysis methods like Spatial Heterogeneity, Spatial Autocorrelation etc. Distribution itself relates to various concepts like Trajectory Distribution, Diversity Distribution, PFT Distribution. Second, we can identify related visualizations. In our example, these are Grid Heatmap, Kriged Map and Line Graph.

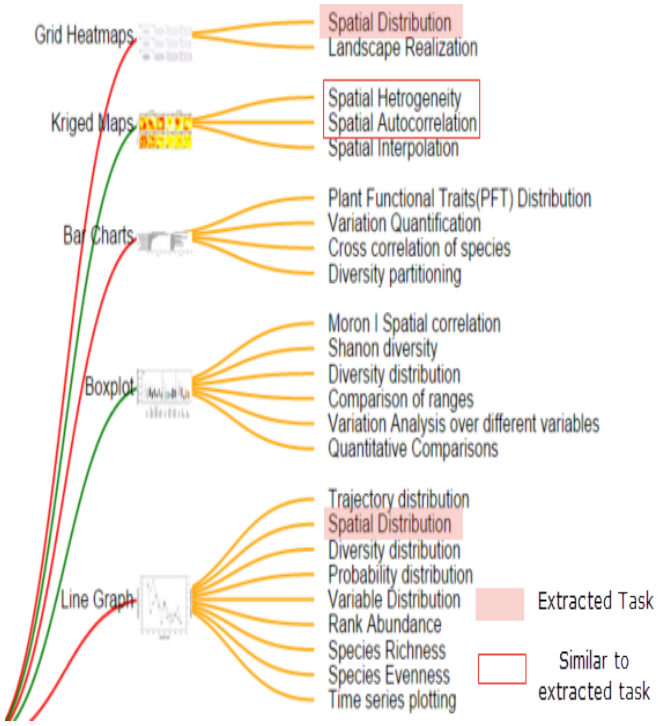

Figure 6: Sample Visualizations used in Biodiversity Research

4) Mapping: Our mapping model is an algorithm that will integrate the knowledge from the previous stages. This algorithm will generate a visualization recommendation list based on the priority of domain specific tasks and feedback from users on the results. The following tasks will be performed:

- It will use the knowledge from previous steps to understand and define the structure of the visualizations appropriate for this dataset.

  In our example, in Step 1 we have understood that the task to perform is 'Distribution' with the use of some distribution function for 'Spatial Analysis'. From Step 2, we have identified three candidate visualizations.

- It will integrate the knowledge from Step 2, to map the data attributes within the candidate visualizations.

  For example if the user selects 'heatmap', then a possible mapping of variables to visual icons are depicted in Table 3 (consider Table 1 and Figure 5 also)

- It will score/rank the candidate visualizations based on:

  Review Phase (Step 3): By choosing the candidate visualizations most popular for that study.

  System learning: Based on user's feedback as introduced in Section 4.2 below.

Table 2 : Variable attribute mapping

Visual Icons	Variables
X axis	PLOT
Y axis	NRderMassnahme
Value	Tree Height
Colour	Tree

5) Learning: If the user is not satisfied with the results of our automatic mapping process, he or she will be presented with an interactive workflow which will be explained in detail in the next section and which will improve future system suggestions.

4.2 Interactive Learning Workflow

We believe that trying to fully automate the task of visualization recommendation is an extremely difficult area. Classical machine learning approaches, in which the system can be trained on visualization mapping for different domain concepts, might be an option. However, this is an expensive process as it takes tremendous effort in gathering knowledge about the domain (especially for wide domain areas like ours) and then takes a long time to train the model on the huge database. Moreover, this approach is not user centric. Therefore, we suggest the use of interactive machine learning approaches to overcome these problems. Algorithms used in the mapping process can be continuously refined, by training them from the logs of user interaction.

Such an interactive learning workflow is presented in Figure 7.

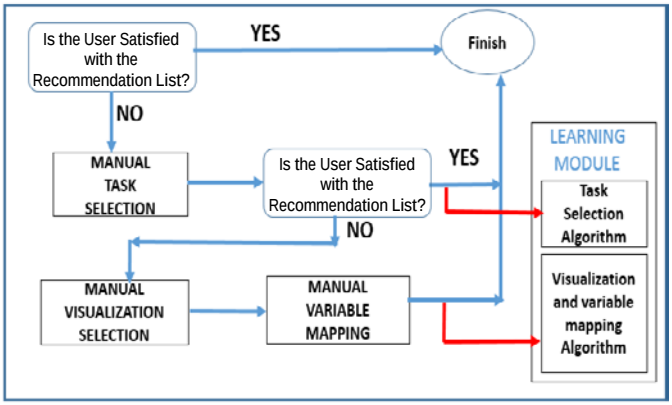

Figure 7: Interactive learning workflow

The learning aspect of the visualization will be triggered in three different cases: In the first case, if the user is satisfied with the list of recommended visualizations, the system will consider it as a *hit* case. Every *hit* case will trigger the following actions:

1) The weight parameter will be increased for that recommended list

2) Within the list, the visualization that a user selected will score higher than the other visualizations. Returning to the example from Section 4.1, consider the identified visualization list and the respective probability of the visualizations to be selected. Initially, all will have the same probability of being chosen by the user. For example, given the following visualizations:

- Grid Heatmap: 33.33 %
- Krig map: 33.33 %
- Line graph: 33.33 %

Suppose now the user has selected "Line Graph". Then it will rank higher in the list and will have a higher probability to get selected. As, it is here:

- Line graph: 66.66%
- Grid Heatmap: 16.67 %
- Krig map: 16.67 %

The second case is, if the user is not satisfied with the list of recommended visualizations. We will consider this as a *miss* case. Then, we need to know why the intended visualization (i.e., the visualization that a user wants) could not be generated. Therefore, we will ask the user to do a manual task selection. When the user has selected the task, a new visualization list will be recommended. If that is a *hit* case, then we will update our semantic algorithm (which we used in Step 3 of Section 4.1) with this task. In other words, we will make our algorithm consider this task (that the user has selected), when similar context (metadata) is encountered next time. This has been depicted via red lines in Figure 7. Now if the user is still not satisfied with the result, or it is a *miss* case again, then we know that the problem is not with the task extraction algorithm, but with something else. So, we will ask the user to select the visualization and variables. The selected visualization will be updated in the list (from the publication review phase, which we used in Step 3 of Section 4.1), with the corresponding variables that the user has selected.

5. CONCLUSION AND FUTURE WORK

In this paper, we have proposed an approach to semi-automatic visualization recommendation. It is based on understanding the problem domain and capturing knowledge from the domain vocabularies. We are certain that this will assist users (biodiversity researchers in our case) in making suitable choices of visualization from the recommended list without needing to get into any technical details. We have also presented an interactive learning workflow that will improve the system from the users' feedback in case the recommended visualizations are not suitable for them. This will make the system more human centric by inculcating knowledge from different viewpoints, which will produce more effective and interpretable graphics.

Our work is in its initial stages and we are in the process of gathering the visualization requirements from the domain experts via surveys and publications review. This knowledge will be used as a ground truth for mapping the conceptual knowledge to the visual operations.

6. REFERENCES

[1] Grammel,L., Troy,M., and Story,M. 2010. How information visualization novices construct visualizations. *IEEE transactions on visualization and computer graphics,16(6)*.943-952.DOI= 10.1109/TVCG.2010.164

[2] Amar, R. and Stasko, J. 2004. A Knowledge Task-Based Framework for Design and Evaluation of Information

Visualizations. *In Proceedings of the IEEE Symposium on Information Visualization* (INFOVIS '04). IEEE Computer Society, Washington, DC, USA, 143-150. DOI=10.1109/INFOVIS.2004.10

[3] Martig, S., Castro,S., Fillottrani, P. and Estévez, E.2003. Un Modelo Unificado de Visualización. *Proceedings, 9º Congreso Argentino de Ciencias de la Computación*. Argentina. 881-892

[4] Bergman, L.D., Rogowitz, B.E. and Treinish L.A. 1995. A rule-based tool for assisting colormap selection. *In Proceedings of the 6th conference on Visualization '95 (VIS '95)*. IEEE Computer Society, Washington, DC, USA, 118-125

[5] Stephen M. C., 1991. Task-analytic approach to the automated design of graphic presentations. *ACM Trans. Graph.* 10, 2 (April 1991), 111-151. DOI=10.1145/108360.108361

[6] Munzner T. 2009. A Nested Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (November 2009), 921-928. DOI=10.1109/TVCG.ma2009.111

[7] NASA JPL California Institute of Technology. Semantic Web for Earth and Environmental Technology (SWEET) version 2.3. Available at https://sweet.jpl.nasa.gov/download

[8] Gilson, O., Silva, N., Grant, P.W. and Chen, M. 2008. From web data to visualization via ontology mapping. *In Proceedings of the 10th Joint Eurographics / IEEE - VGTC conference on Visualization (EuroVis'08), Eurographics Association, Aire-la-Ville, Switzerland*, 959-966. DOI=10.1111/j.1467-8659.2008.01230.x

[9] Rui, Y., Huang, T.S., Ortega, M. and Mehrotra, S. 1998 Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval. *IEEE Transactions On Circuits and Systems for Video Technology*, 8(5).1-13

[10] IBM. Many Eyes. Available at http://www-969.ibm.com/software/analytics/manyeyes

[11] Wilkinson, L. *Statistics and Computing, The Grammar of Graphics*. Springer Press. Chicago, 2005

[12] Heer, J., Ham,F., Carpendale,S., Weaver,C. and Isenberg, P. 2008. Creation and Collaboration: Engaging New Audiences for Information Visualization. *In Information Visualization, Lecture Notes In Computer Science*, 4950. 92-133. DOI=10.1007/978-3-540-70956-5_5

[13] Chen, M., Ebert,D., Hagen,H., Laramee,R.S., Liere,R.V., Ma,K.L., Ribarsky,W., Scheuermann,G., and Silver,D. 2009. Data, Information, and Knowledge in Visualization. *IEEE Computer Graphics Application*, 29(1). 12-19. DOI=10.1109/MCG.2009.6

[14] Ware, C. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers, San Francisco, CA. 2000

[15] Lotz, T., Nieschulze, J., Bendix, J., Dobbermann, M. and König-Ries,.B. 2012. Diversity or uniform? Intercomparison of two major German project databases for interdisciplinary collaborative functional biodiversity research. *Ecological Informatics*, 8, 10–19 DOI=10.1016/j.ecoinf.2011.11.004

[16] Fischer, M., Boch, S.,Weisser, W.W.,Prati, D.and Schoning, I. 2010. Implementing large-scale and long-term functional biodiversity research: The Biodiversity Exploratories. *Basic and Applied Ecology*, 11(6).473-485. DOI=10.1016/j.baae.2010.07.009

[17] Tableau. Available at http://www.tableau.com/products/trial?os=windowsbertinbook

[18] Bertin, J. 1983. *Semiology of graphics*. University of Wisconsin Press, Berlin.

[19] Kelleher, C., Wagener, T. 2011. Ten guidelines for effective data visualization in scientific publications, *Environmental Modelling & Software*.1-6. DOI=10.1016/j.envsoft.2010.12.006

[20] The Biodiversity Research Centre, University of British Columbia. Available at http://www.biodiversity.ubc.ca/research/groups.html

[21] Cleveland.,W.S. 1984. Graphs in Scientific Publications. *The American Statistician*. 38(4). 261-269. DOI=10.1080/00031305.1984.10483223

[22] Schofield.,E.L. 2002. Quality of Graphs in Scientific Journals: An Exploratory Study. Science Editor 25(2). 39-41

[23] Lauren, E.F., Kevin, C.C. (2012). Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter, *the Journal of Hand Surgery*, 37(3). 591-596, DOI=10.1016/j.jhsa.2011.12.041

[24] Reda, K., Johnson, A., Mateevitsi, V., Offord, C., & Leigh, J. (2012). Scalable visual queries for data exploration on large, high-resolution 3D displays. *In High Performance Computing, Networking, Storage and Analysis (SCC). IEEE.* 196-205. DOI=10.1109/SC.Companion.2012.35

[25] Gnanamgari S. 1981. Information presentation through default displays. Ph.D. dissertation, Philadelphia, PA, USA

[26] Mackinlay J. 1986. Automating the design of graphical presentations of relational information. *ACM Transactions of Graph*, 5(2), 110-141. DOI=10.1145/22949.22950

[27] Wehrend S. and Lewis C. 1990. A problem-oriented classification of visualization techniques. *In Proceedings of the 1st conference on Visualization '90 (VIS '90)*, Arie Kaufman (Ed.). IEEE Computer Society Press, Los Alamitos, CA, USA, 139-143.

[28] Senay, H. and Ignatius, E. 1994. A Knowledge-Based System for Visualization Design. *IEEE Computer Graphics and Applications.* 14(6), 36-47. DOI=10.1109/38.329093

[29] Healey C. G., Amant R. S., and Elhaddad M. S.. 1999. Via: A perceptual visualization assistant, *In 28th Workshop on Advanced Imagery Pattern Recognition (AIPR-99)*, pp.2–11.

[30] Biodiversity Exploratories Information System (BExIS). Available at https://www.bexis.uni-jena.de/Data/ShowXml.aspx?DatasetId=4020. Accessed on 07/05/2015