Neirynck, Thomas and Börner, Katy. (2007) Representing, Analyzing, and Visualizing Scholarly Data in Support of Research Management.
In Conference Proceedings of 11th Annual Information Visualization International Conference (IV 2007),
Zurich, Switzerland, July 4-6, pp. 124-129, IEEE Computer Society Conference Publishing Services.

Representing, Analyzing, and Visualizing Scholarly Data in Support of Research Management

Thomas Neirynck & Katy Börner
School of Library and Information Science, Indiana University
10th Street & Jordan Avenue, Bloomington, IN 47405
{tneirync@indiana.edu, katy@indiana.edu}

Abstract

This paper describes a data management system that serves leaders and members of scholarly units (e.g., research groups, laboratories or centers) to organize, store, and access information relevant to their work and to the operation of their unit. In particular, we discuss the representation of major data types such as people, projects, teaching, publications, presentations, calls & events, datasets, software, hardware, and funding as well as their interlinkages. Based on this representation, different analyses and visualizations become possible. Each visualization is presented in detail and its usage and relevance for the management and presentation of a unit's operation is discussed. We conclude with an outlook and discussion.

Keywords: data management, visualization, network analysis, network visualization

1. Introduction

This paper presents a novel approach and tool for the tracking, management, and sense making of data relevant for the daily operation of a scholarly unit such as a research groups, laboratory, or center. Small research groups with about five students might not require the proposed tool as all data can be easily managed by one person. Huge groups, e.g., entire departments or schools, with more than 50 different people typically have full-time paid assistants and staff to take care of finances, hiring and firing, information technology, travel arrangements, etc.. However, medium size research units do not enjoy the luxury of paid assistants. They are often lead by one person with part time help from students or an hourly staff person. The task of keeping track of up to 50 people (mostly students that stay a rather short time), up to 30 research projects, teaching up to 10 different classes, up to 30 publications a year, about 50 presentations a year, about 10 co-organized calls & events, a growing number of used and developed datasets, different software packages and their versions and dependencies, owned and licensed hardware, as well as associated funding is easily overwhelming.

There are tools like Excel and Access that support spreadsheet design and documentation. However, to our knowledge, there exists no system that is designed to support the management of scholarly data that is relevant for the operation of a research unit and the communication of this data to different stakeholders such as group lead(s) and members, information technology staff, funding agencies, or other scholars.

Subsequently, we discuss the needs of different stakeholders. Next, we review related work. Then, we present data representations, analyses, and visualizations that are tailored to the identified stakeholders and their needs. The paper concludes with a summary and discussion of future work.

The proposed data representation, analysis, and visualization approach has been fully implemented in the InfoVis Lab management system, or IVL for short. All examples in this paper are drawn from this prototypical implementation.

2. Different stakeholders and their needs

As mentioned before, the presented tool aims to support knowledge management for medium size research units. Subsequently, we discuss the information needs of different stakeholders.

Leaders of medium size units frequently need to provide references for current or previous members. Receiving a phone call about a student which worked in the unit four years ago as a summer intern, they quickly need to retrieve data about the project(s) the student was involved in; software, papers, presentations s/he produced and their quality; and any other, e.g., teaching, contributions the students might have made. Leaders also need to report the progress of different projects to different funding agencies in different time intervals, e.g., quarterly or annually. A particular challenge is caused by the fact that almost every project starts on a different day of the year and has a different funding duration. Whenever a report is due, one needs to retrieve details about all people, educational/teaching activities, publications, presentations, calls & events, datasets, software, and hardware associated with a particular project for the reporting duration, e.g., the last three or 12 months.

Group members might include Ph.D. students that need to report their progress on an annual basis. That is, they need to retrieve all projects they contributed to, teaching activities, publications, presentations, events they organized, service activities, and funding acquired

Neirynck, Thomas and Börner, Katy. (2007) Representing, Analyzing, and Visualizing Scholarly Data in Support of Research Management.
In Conference Proceedings of 11th Annual Information Visualization International Conference (IV 2007),
Zurich, Switzerland, July 4-6, pp. 124-129, IEEE Computer Society Conference Publishing Services.

and/or received. Some members are project team leads themselves and are in need to manage the project team and associated publication plans, presentation travels, organized events, as well as the datasets, software, and hardware associated with their project(s).

Information technology support staff needs to know what projects and people are affected if certain hardware is replaced or a piece of software is updated. Ideally, they even know what publication or software release deadlines the different projects are aiming for in order to cause the least interruptions of scholarly activity while building and maintaining a highly reliable infrastructure.

Funding agencies are interested to see the impact of the work they fund. Impact might be measured by the number and quality of produced ‘artifacts’ such as Ph.D. students, publications, teaching material, datasets, or software. Quality might be calculated based on the ‘consumption’ of artifacts by others, i.e., the citation of papers, the hiring of students, the number of invited talks, and the geospatial (national or international) and topical (within domain of research or interdisciplinary) spread of scholarly artifacts.

Scholars might like to gain a quick overview about the expertise and productivity of a certain unit. They might be interested to know when Ph.D. students become available for hire, if an expert has ‘cycles’ left to serve as a collaborator on a new funding proposal, if a unit member is available to give a talk at a certain date or is already committed to other travels, or how often unit members will travel to Tokyo, Japan this year and might be able to give a talk without requiring additional flight costs.

3. Related work

There exists a rich diversity of knowledge management books that report innovative approaches and best practices of how individuals or organizations can and should manage their knowledge [5]. However, most of this work is rather qualitative than quantitative.

Recent work on mapping knowledge domains [3, 7] aims to represent, analyze, and map scholarly activity on a global scale. This work is of great utility if general trends, patterns, outliers have to be identified to support decision making. However, as we have seen in section 2, most actions and decisions happen on a much more local scale. To our knowledge there is a desperate need for more sophisticated scholarly knowledge management tools.

4. Scholarly knowledge organization

It is non-trivial to decide what general scholarly record types are needed, what attribute values different record types should have, and how the records should be interlinked. These decisions need to be made based on the information needs that the knowledge management system is supposed to support, see section 2.

Semantic Association Networks (SANs) have been introduced in [2] as a novel means of using semantic web technology to tag and interlink scientific datasets, services (e.g., algorithms, techniques, or approaches), publications (e.g., papers, patents, grants), and expertise (i.e., author and user information) to improve scholarly knowledge and expertise management. Building on this work, we propose a knowledge organization that interlinks main scholarly data entities such as people, projects, teaching, publications, presentations, calls & events, datasets, software, hardware, and funding, see Figure 1. Note that people and projects are the two central nodes in this network.

Using a visual metaphor, any node can be ‘pulled’ out of the network and other nodes linked to it will follow. For example, a specific person can be selected and all scholarly records associated with it will be retrieved.

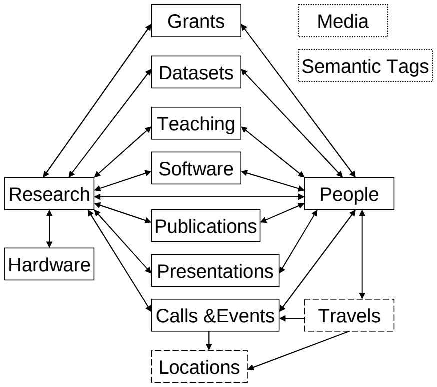


Figure 1. Simplified representation of the IVL database schema showing the main associations between the main data entities. Single and double arrows respectively indicate one-to-many and many-to-many relationships. All entity types can be linked to media items or semantic tags.

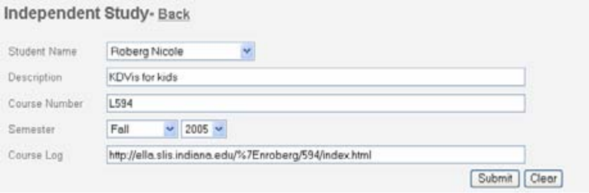
The database schema, represented in a simplified manner in Figure 1, was implemented in a relational PostgreSQL database for the IVL. In general, each scholarly data type corresponds to a table in the database. Research projects are stored in *tblResearch*, publications in *tblPublications*, and so on. However, some entity types are not coerced into a single data model and are represented by multiple tables. An example is teaching, which is used to represent everything from semester-long classes to half-day tutorials. Therefore, each teaching type (courses, tutorials, educational materials) has its own table. Other peripheral tables store the different people types like students that perform independent studies, research center members, and others. These are not mutually exclusive, as people can both be doctoral students and enrolled in an independent study.

In addition to the main scholarly data entities and their subtypes, the IVL also monitors the movement of people in geo-space from one event to another, adding a spatio-temporal dimension to scholarly activity. Travels are stored in a separate table, each record specifying a displacement from one location to another. To simplify data entry, every time a new presentation or meeting is added, an automatic function adds a corresponding record in the travel table. In the IVL, the default start location is the default location of the person and the end location being whatever the location of the related event is. The default travel date is the first date of the event or meeting. Multi-legged travels can be recorded. Unique latitude and longitude are stored in a separate location table, with corresponding address information.

There are two tables that have linkages to all other tables: the media table and the semantic tag table. In Figure 1, these tables are shown in the top right corner, but to keep the diagram easy to understand, the links are not drawn. The media table is used to store files associated with different data entity types, from photos of people and events, to text files of publications. The semantic tag table stores a list of unique terms that can be associated with any of the main data type records. For example, tags are used to represent concepts such as “Cyberinfrastructure”, “Knowledge Management” or “Information Visualization” that are too specific or too general to correspond to a specific project. Hence, tags serve as an additional means to organize and retrieve records.

So-called bridge tables store the relationships between different data records, e.g. *brdgAuthorsEditors* links *tblPeople* and *tblPublications* or *brdgTeamCollabs* links *tblPeople* and *tblProjects.* In most cases, bridge tables also store additional attributes. For example, *brdgTeamCollabs* stores the time during which a link was active, e.g., how long a person was associated with a certain project and in what capacity (team lead, team member, external collaborator). To ensure readability of Figure 1, the bridge tables are not shown.

The database was prototypically implemented in the IVL, with the information displayed at http://ivl.slis.indiana.edu. Data entry proceeded in a variety of ways. Initially, spreadsheets were used to enter the records of main entities. Later, web based input forms were developed and used to interlink entities, see sample forms in Figure 2. Throughout, a database administration tool, PgAdminIII, was used to make corrections as necessary.



Independent Study- Back

Student Name	Roberg Nicole
Description	KDVis for kids
Course Number	L594
Semester	Fall 2005
Course Log	http://ella.slis.indiana.edu/%7Enroberg/594/index.html

Figure 2. Data input form for independent studies.

In the near future, the system will support data entry via text files (.txt, .cvs formats) or XML files to populate the tables of the main data types. For the web based forms, fielded forms, combo-boxes, and automated data checking are employed widely to enforce data quality and integrity.

5. Data analysis and visualization

Given the data representation introduced in the previous section, high quality data about scholarly entities can be captured for subsequent analysis and visualization in support of needs outlined in section 2. Firstly, the tool supports the generation of progress reports for funding agencies. Secondly, the data can be analyzed to gain a further insight into the areas of expertise within a research lab. If one wants to gain a deeper understanding of all the work and resources that have been invested in research on “Information Visualization” (IV), one can pull out all the projects that are tagged with this term. All related data records, e.g., of type software, datasets, publications, are retrieved as well. Such an approach goes beyond the mere reporting of a lab’s activities. Rather, these rich semantic networks represent the core of a research unit’s activity, which can be navigated and explored. Returning to the IV example; a colleague might have attended a presentation of a Ph.D. student on an IV project. When visiting the research unit’s web pages, the researcher not only finds a link to the abstract and slides of the presentation in the Presentations section, he/she can also find the datasets for the related IV project and the publications written on the topic. Hence, s/he can learn about collaborators who are active on other IV projects, and in turn study their publications and presentations.

Several graphical displays have been implemented to aid in the understanding and communication of this data. Subsequently, we discuss tabular formats, graphs and networks, and geospatial maps that can be generated by querying the database.

As for now, all displays are generated by querying the database manually and use other software packages or tools, such as Excel or Pajek [6], to generate the visualizations. In the near future, these steps will be automatized using CIShell and the Network Workbench technology [4].

5.1 Tabular displays

Sortable tables provide a global view of people, projects, publications, presentations, developed software, datasets, etc. and their attributes. They are useful for

- Retrieving data for writing project specific progress and final reports.
- Quickly gaining access to complete, up to date information about a person for writing letters of reference or for giving phone references.
- Having an up to date list of all publications, activities, etc.
- Identifying who has the most publications, projects, etc.

5.2 Graph displays

Graph displays such as scatter plots and timelines help understand correlations and changes over time. They can be used to correlate dollars spent to the number of papers published or to check the number of publications, presentations, people, etc. associated with a certain person/project in a given time frame.

Results are displayed as simple graphs that represent the number of entities of the selected type (y-axis) for a specified time frame (x-axis), see example in Figure 3.

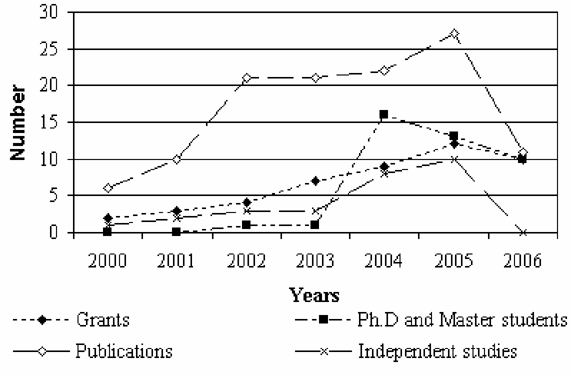


Figure 3. Number of students and publications in the IVLab between 2000 – 2006.

5.3 Network displays

Almost every project is a team effort that consumes resources, e.g., people, funding, hardware, software, and produces artifacts such as expertise, publications, datasets, software, etc. Any team and project can only be as good as the resource networks that support it and the artifact networks it produces.

The explicit interlinkage of entity types presented in section 4 supports the extraction, analysis, and display of networks. Examples of networks that interlink entities of only one type are co-author, co-PI, or paper-citation networks. Figure 5 is an example of a coauthorship network from 2005 – 2006. It represents all the collaborations on papers produced by staff and students active in the IV lab in that timeframe. The nodes are color coded according to number of network components he/she is connected too (when this number is 2 or higher, this means that these people are articulation points. Take them away and the network would fall apart). This is useful to identify which people have ties to other research groups. Node size represents the number of papers written, while edge size represents the number of times to authors collaborated. The border of the nodes is color-coded too; red for Ph.D. students that have worked at the lab at one point, blue for other team members, and black for outside collaborators.

Other networks of interest connect entities of two types, e.g., people and papers, people that are principal investigators and projects, people and projects, projects and funding, etc.

An example network of *People* and *Funding* records for 2001 to 2006 is given in Figure 6. Circular nodes represent awards, their color indicating the year it was awarded. Squares (with or without the images as available in the IVL database) represent the investigators. Networks are displayed as a static image or dynamic network visualization that evolves over time. Networks can be downloaded as .net file for further analysis and exploration.

5.4 Geospatial map displays

Space matters – even in the Internet age [1]. Given that addresses are available for most entities of type *People*, *Presentations* and *Events* the joint production of results documented in papers, presentations, meetings, travels etc. can be overlaid on a geospatial map.

Figure 4 shows an exemplary overlay of *Travels* in 2003 and 2004 using the Google Maps API.

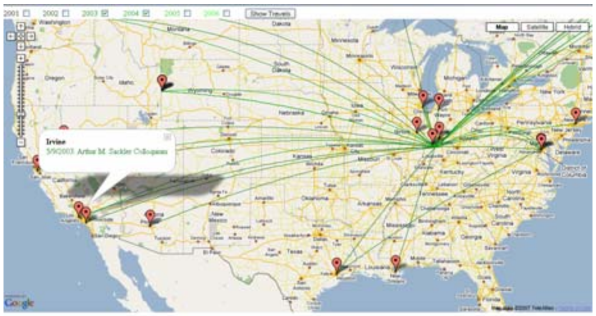


Figure 4. Interactive Google Maps interface that shows travels of IVL members in 2003 and 2004.

6. Implementation and usage

The scholarly data organization and representation discussed in this paper was prototypically implemented in the data management system of our Information Visualization Laboratory, online at http://ivl.slis.indiana.edu. Selection of any of the main entity types brings up dynamically generated pages with details about that record type, and other associated information (team members, related publications, links to software downloads, etc.). Such a representation supports the needs of the different stakeholders as outlined in section 2.

Given the easy to use interface and the up-to-date status of all relevant data, the IVL management system also serves another interest of scholars: to diffuse their own data, software, resources, and expertise in exchange for reputation and funding.

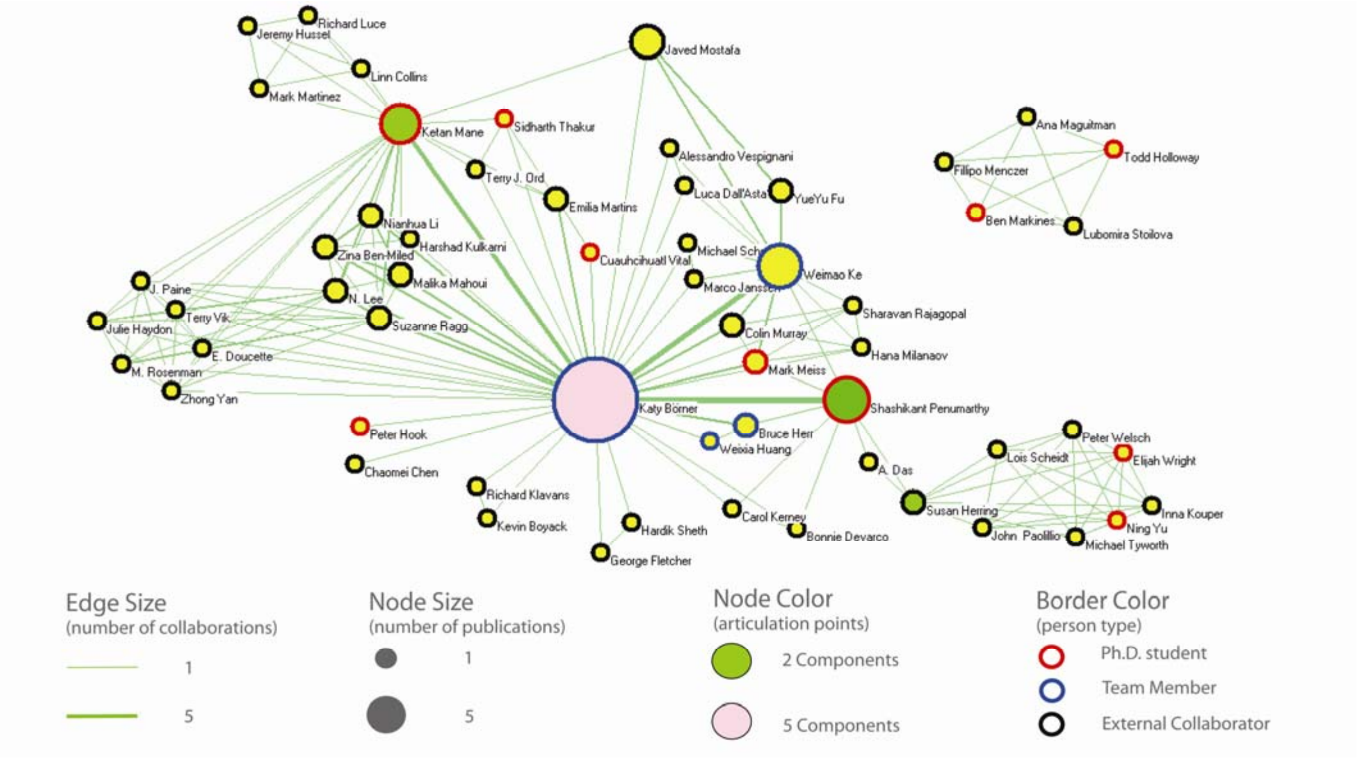


Figure 5: Co-authorship network for all publications of InfoVis lab group members in 2005 to 2006

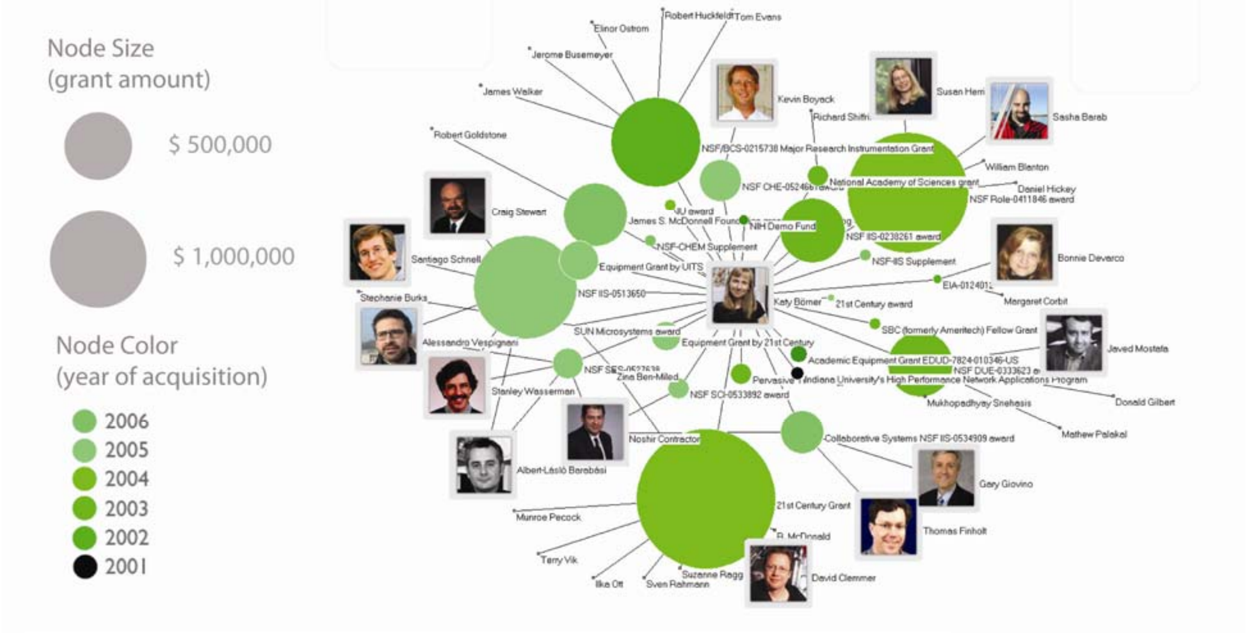


Figure 6: Awards at the InfoVis lab and their investigators from 2001 to 2006

7. Discussion

If every research group would keep track of scholarly data using an IVL-like setup then access to high quality papers, data, software, resources, and expertise would be eased considerably. Any scholar could access or disseminate data, software, resources, expertise. For each expert/author/awardee in the system, a list of associated datasets, software, resources, papers and their 'quality' could be retrieved. A person's affiliation, country, and geo-location as well as his/her area of expertise could be determined based on his/her datasets, software, resources, papers, etc. Hence, the research trajectory of any person could be analyzed and mapped.

Given the rich interlinkage of the different scholarly data entities, new ways of searching for/sharing of people/datasets/software become possible. For example, all people that have worked with a certain dataset can be retrieved. Analogously, all software that was applied to study a dataset could be viewed.

Adding semantic tags and/or comments to datasets, software, resources, papers, and other entities makes it possible to retrieve entities based on word matches.

Ratings of datasets, software, resources, papers, experts (manually done or automatically based on #downloads, #citations, #comments) can be used to determine what datasets, software, resources, papers are most valuable. A true marketplace of scholarly records can be created. Plus, whoever utilizes this new way to interlink and manage scientific data will be able to generate high resolution inserts that can be embedded in a global map of science.

8. Future work

While the IVL is operational, we plan a number of extensions in the near future. Among them are:

- Web based search interfaces to easily build detailed queries, and generate various visualizations.
- Add applets to the WebPages in support for the dynamic network visualizations.
- Upload papers automatically from Endnote or BibTEX files.
- Export data in different file formats for easy data exchange, such as .kml for the spatial visualizations, and GraphML or .nwb for the network visualizations.

The IVL software and database will also be used to serve the web pages for the newly funded Cyberinfrastructure for Network Science (CNS) center at Indiana University. While the IVL serves data relevant for the Information Visualization Lab, the CNS site will serve data relevant to network science research and cyberinfrastructure development in accordance with the center's mission.

Acknowledgements

We would like to thank Sumeet Ambre, Gavin LaRowe, and Russel Duhon for setting up the IVL PHP code base and their expert advice for the presented work. Elisha Hardy designed the Web interface. This work is supported by the National Science Foundation under IIS-0513650 and IIS-0534909. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

References

1. Batty, M. *The Geography of Scientific Citation*, Environment and Planning A, 35, 2003, 761-765.
2. Börner, K. Semantic Association Networks: Using Semantic Web Technology to Improve Scholarly Knowledge and Expertise Management. in Geroimenko, V. and Chen, C. eds. *Visualizing the Semantic Web*, Springer Verlag, 2006, 183-198.
3. Börner, K., Chen, C. and Boyack, K. Visualizing Knowledge Domains. in Cronin, B. ed. *Annual Review of Information Science & Technology*, Information Today, Inc./American Society for Information Science and Technology, Medford, NJ, 2003, 179-255.
4. Herr, B.W., Huang, W., Penumarthy, S. and Börner, K. Designing Highly Flexible and Usable Cyberinfrastructures for Convergence. in Bainbridge, W.S. and Roco, M.C. eds. *Progress in Convergence – Technologies for Human Wellbeing*, Annals of the New York Academy of Sciences, Boston, MA, 2007, 161-179.
5. Montano, B. (ed.), *Innovations of Knowledge Management*. IRM Press, 2005.
6. Nooy, W.d., Mrvar, A. and Batagelj, V. *Exploratory Social Network Analysis with Pajek*. Cambridge University Press, 2005.
7. Shiffrin, R.M. and Börner, K. (eds.). *Mapping Knowledge Domains*. PNAS, 101, 2004.