Bull. Nov. Comp. Center, Comp. Science, 31 (2010), 17–28

An extension of a visualization component of ontology based portals with visual analytics facilities[*]

Z.V. Apanovich, P.S. Vinokurov

Abstract. The process of development of an ontology-based knowledge portal and creation of its content is time-consuming and laborious. The lifetime of such portals is sufficiently long and they collect a great volume of valuable information. This information can be analyzed from various points of view. This paper describes an extension of the visualization subsystem, developed the A.P. Ershov Institute of Informatics Systems, with new tools for visual analysis. An example of extracting information about scientific cooperation from the content of a knowledge portal and two ways of its visualization are demonstrated.

Key words: hierarchical edge bundles, knowledge portal, modularity, ontology.

1. Introduction

A generally accepted way to facilitate understanding of large and complex data sets is to use graph visualization methods. Any ontology can be represented as a graph, where the graph vertices correspond to the ontology entities, such as classes and instances, and the graph edges correspond to the ontology relationships. By browsing drawings of various sub-graphs, the portal developer can detect the errors caused by manual input of information, as well as the design errors [1]. During its life time, information portals accumulate more and more heterogeneous data and become a valuable source of information suitable for various forms of analysis. In the case of an information portal devoted to some scientific field, the problem of the scientometric analysis of its content is very important. In particular, we are interested in extraction and visualization of information about cooperation between various scientific communities. This paper describes new facilities of a visualization component developed at the A.P. Ershov Institute of Informatics Systems [2, 3]. These facilities focus on a deeper analysis of the knowledge portals content. The User Interface of the visualization component is shown in Figure 1. The state of the art in the scientific cooperation research is shortly discussed in Section II. The first newly developed facility is related to the "Clustering" menu tab. It allows for extracting the co-authorship networks from the content of ontology based portals, their

[*] Partially supported by RFBR under grant 09-07-00400) and RAS Project 2/12.

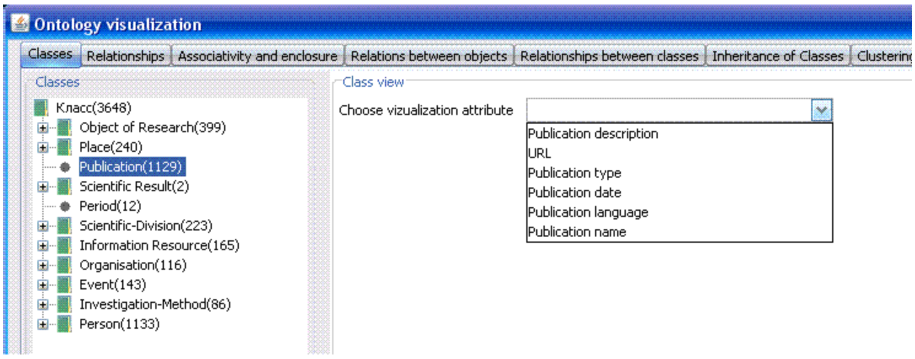


Figure 1. User interface of the visualization component

division into scientific communities and visualization. This facility will be discussed in Section 3.

The second new facility is related to the “Associativity and enclosure” menu tab. It allows for combined visualization of the co-authorship networks and partonomy relationships of the portal. This visualization mode will be demonstrated in Section IV. Finally, Section V presents conclusion and perspectives for further work.

2. Ontology based portals and scientific cooperation networks

One of the considered portals is an archaeological portal that contains a great volume of information about scientific publications in various fields of archeology. This information evolves over time. In addition to the fact that this information is intended for professionals working in the field of archeology, it can be used in all kinds of scientometric experiments. One common method of bibliographic information analysis is the analysis of various networks, such as co-authorship networks, citation and co-citation networks, etc.

Investigation of co-authorship networks is getting a very popular scientific challenge, since joint research becomes the dominant and most promising mode of production of high-quality scientific results. Collaborative multidisciplinary research projects and co-authored publications point to the mode of today’s knowledge production [6, 7]. Co-authorship networks are studied extensively from various points of view, such as the degree distribution analysis [8], social community extraction [9], etc.

We construct the co-authorship networks by using the content of knowledge portals. To this end, we generate a graph, where vertices correspond to researchers, and edges represent co-authorship relations between them. If a publication has n authors, a clique connecting the publication’s authors is created. Since authors can have several joint publications, the weight of the edge connecting these researchers is equal to the number of joint publications.

Note that the co-authorship relationship did not explicitly exist in our test ontology. At the same time, this relationship can be described, like many others, as a superposition of already available relationships. For example, in the test ontology there exists the relationship "*Author-Of-Publication*" linking the classes "**Researcher**" and "**Publication**". It is clear that the co-authorship relationship can be described as a superposition of the relationship "*Author-Of-Publication*" and its inversion. A facility to describe this kind of superposition was incorporated into the internal language of our visualization component.

3. Clustering for analysis and visualization of scientific cooperation

When the co-authorship network is created, we start investigating it so as to extract and visualize the scientific communities. The standard force-directed algorithms [10, 11] are not quite suitable for visualization of scientific communities, because they seek to place all vertices of the graph at the same "ideal" distance. We need an algorithm that would visually separate groups of highly connected researchers.

A standard way of extraction of scientific communities uses various clustering methods. A clustering algorithm based on the *modularity* measure has been selected and implemented in our visualization component. Modularity [9] is a property of a specific division of a network into communities. It determines if the division is a good one. (The division into communities is considered to be good if there are many edges within communities and only a few between them).

Let us define a $k \times k$ symmetric matrix $\mathbf{e}$ whose element e_{ij} is a fraction of all edges in the network that link vertices in the community i to vertices in the community j. Then we can define the row (or column) sums $a_i = \Sigma_j e_{ij}$, which represent the fraction of edges that connect vertices inside the community i.

The modularity is expressed through a_i and e_{ij}:

$$Q = \sum_i (e_{ii} - a_i).$$

An example of modularity calculation for two communities C_1 and C_2 is shown in Figure 2. It is known from experiments [9], that a value greater than 0.3 is a good indicator of a significant community structure in a network. Our algorithm of communities extraction consists in removing the

$e_{12} = 1/10$,
$e_{11} = 6/10$,
$e_{22} = 3/10$,
$a_1 = 7/10$,
$a_2 = 4/10$
$Q = 41/100$

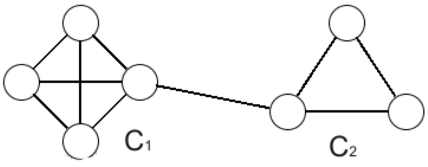


Figure 2. The modularity calculation for two communities C_1 and C_2

edges with the highest *edge betweenness* [9]. To calculate the edge betweenness, we have to find the shortest paths between all pairs of vertices and calculate how many paths run along each edge. The edge with the greatest value of betweenness is removed from the graph. If the edge removal increases the number of components, the modularity for a newly obtained partition is calculated. If the newly found modularity value is greater than the former one, this value is stored, and the edge removal process continues until the difference between the current modularity value and the best value is greater than a *stop_parameter*. At this point, the clustering process terminates and the components corresponding to the best found value of the modularity are used as a result of clustering.

The next step consists in visualization of the results of scientific communities extraction. We want to make the found communities easily visible, as well as the relationships between them.

To generate this kind of visualization, a three-step placement algorithm is used. First, we construct a global placement of the graph, whose vertices are the found communities. During this step, the ideal length of the edge connecting the components c_i and c_j, is considered to be proportional to the value e_{ij}, a fraction of edges that join vertices in the community i to vertices in the community j.

Then another force-directed algorithm creates a detailed placement of each community. At this point, all vertices of one group are placed at the approximately equal distances from each other. Finally, all the detailed placements of components are substituted in the global component placement and the inter-component edges are rendered.

Figure 3 shows an example of placement of the greatest connected component extracted from the co-authorship network of the archeological portal. The network consists of 2090 researchers. The greatest connected component has 370 vertices and 1690 edges. This component is laid out with the standard Fruchterman–Reingold algorithm [11].

Figure 4 shows the same component clustered and laid out with the algorithm implemented in our system. As a result of this procedure, 35 communities were identified. The greatest community comprises 50 authors. At the same time, there are communities comprising 2-3 researchers. Typically

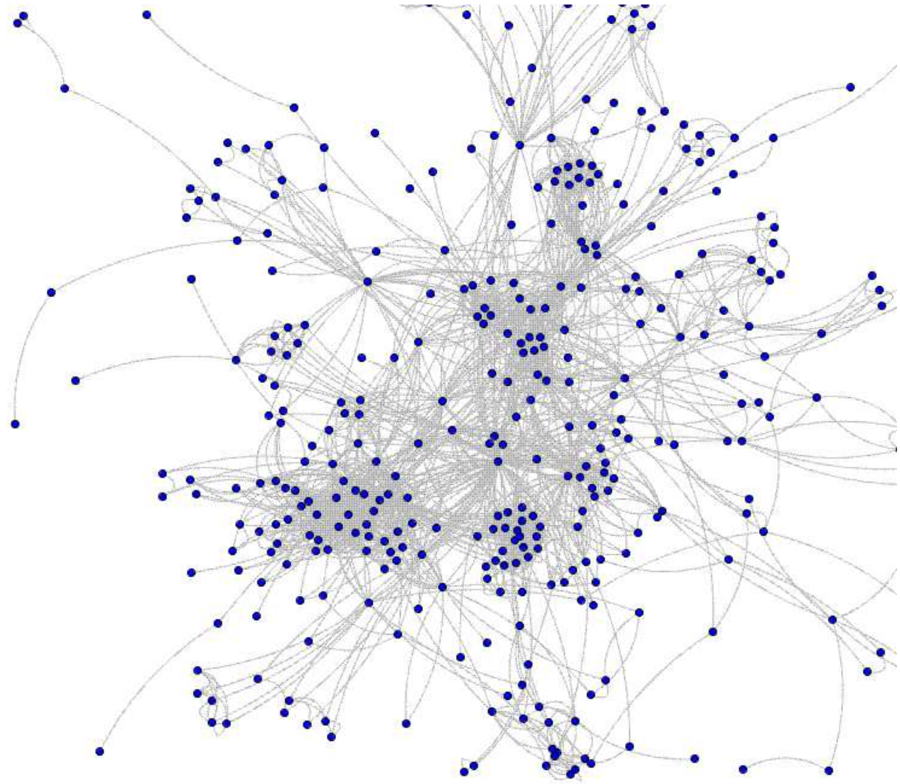


Figure 3. The greatest connected component placement using the standard Fruchterman–Reingold algorithm

each community is formed around an author with the maximum number of publications in the community. To make the found components easily recognizable for a user, the intra-community edges are shown in black, and the inter-community edges are shown in gray.

From our point of view, the main disadvantage of such visualization is its incompleteness. Specialized portals contain, in addition to bibliography, a huge amount of extra information: about scientific organizations and branches of science, details of research and scientific activities, such as scientific expeditions, etc.

All this information is organized by the portals' ontology. We believe that juxtaposition of cooperation relations with other elements of the portal content can give us better understanding of these relations. Figure 5 shows all relations of the test ontology for the **Person** and **Researcher** classes. The dashed edge corresponds to the inheritance relationship and all other edges represent other associative relationships.

It is possible to see that the class **Person** is connected with other classes via the following associative relationships: "*Acquaintances*" (the

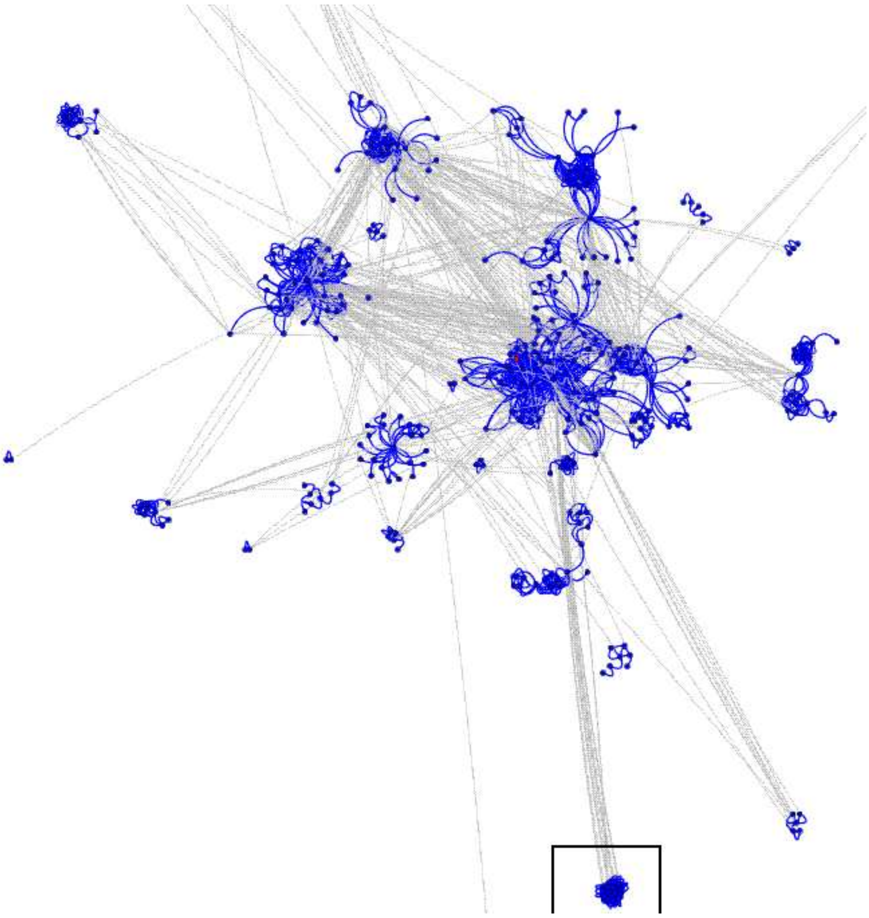


Figure 4. The greatest component visualization with the algorithm implemented in our system. The intra-community edges are shown in black and the inter-community edges are gray

class "**Person**"), "*Event-Participants*" (the class "**Event**"), "*Resource-of-a-person*" (the class "**Information-Resource**"), "*Applies-method*" (the class "**Investigation-Method**"), and "*Student*" (the class "**Researcher**").

The subclass **Researcher** inherits relationships from the class **Person** in addition to its own relationships. It is connected with:

- the class **Scientific Division** via the "*Investigation-Direction*" relationship,
- the class **Publication** via the "*Author-Of*" relationship,
- the class **Period** via the "*Investigates-Period*" relationship,
- the class **Project** via the "*Project-Participant*" relationship,
- the class **Person** via the "*Student-Of*" relationship.

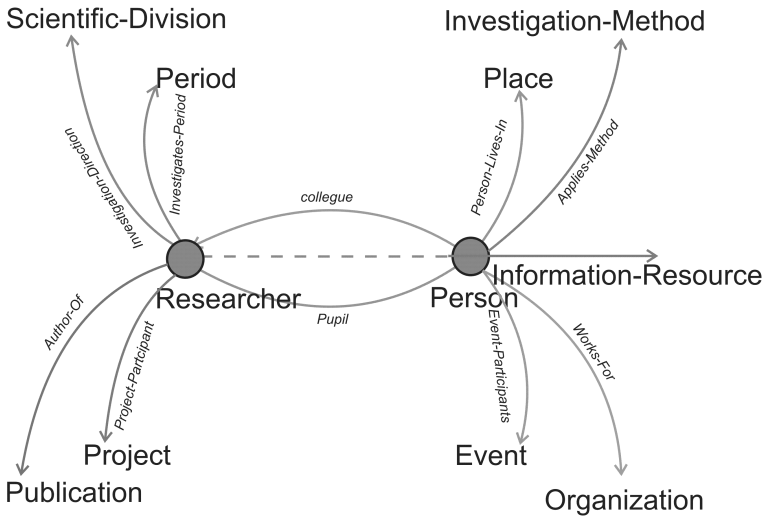
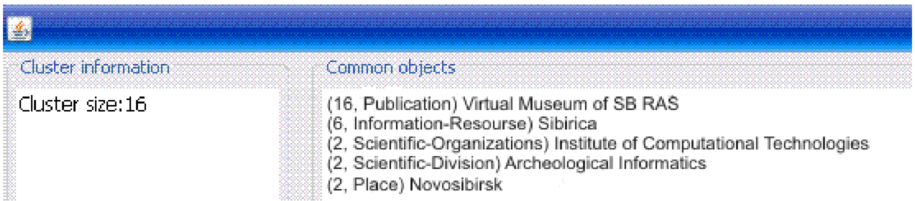


Figure 5. Ontology's relations for the **Person** class and the **Researcher** subclass



Cluster information

Cluster size:16

Common objects

(16, Publication) Virtual Museum of SB RAS
(6, Information-Resourse) Sibirica
(2, Scientific-Organizations) Institute of Computational Technologies
(2, Scientific-Division) Archeological Informatics
(2, Place) Novosibirsk

Figure 6. The common properties of members of the selected community

Therefore, it would be of interest to study the dependencies of partnership relations on scientific organizations, scientific directions, geographical location and other parameters described in the content of the information portal. One possible way to get this information is implemented as follows. When a user selects some vertex in the drawing of the network, the visualization program identifies a community that includes the selected node. The program looks through the links of the vertices-researchers belonging to the selected community and searches for the attributes that are common to them. The common characteristics are arranged in a descending order.

For example, Figure 6 shows the common properties of researchers in the community shown in Figure 4. This community consists of 16 researchers. It is possible to see that the main common property of researchers in this community is their participation in the electronic publication "Virtual museum of SB RAS". Because of this property, a clique connecting all these researchers was generated in the co-authorship network. It should be noted that the human observer, familiar with the field, would immediately have noticed that all members of this group work for the same Institute. But the message from Figure 6 says that only two members of the selected community work for the same organization and stay in the same geographical location. When checking the "**Scientific-Organization**" and the "**Scientific-Division**" attributes of the community members, the user can discover that some information is missing in the content. This kind of unfinished job is quite usual during the manual data input. It makes difficult any search for correlations in the portal content and indicates clearly that automatic methods for portals population should be used.

4. Using the hierarchical edge bundles method for visualization and analysis of scientific cooperation

Another method of co-authorship network visualization takes into consideration more data contained in the information portal knowledge base and uses hierarchical edge bundles for their visualization [15]. The method of hierarchical edge bundles [13] allows users to combine a drawing of co-authorship networks with drawings of other elements of the portal content. This method is implemented in our visualization component as one of several options of the rendering mode "**Associativity and nesting**". We have noticed that a considerable quantity of instances of knowledge portals is organized into hierarchies induced by the partonomy relationships. These relationships are "*Method-of-Research-Includes*", "*Place-Includes*", "*Organization-includes*", etc.

The mode "**Associativity and nesting**" serves to construct a combined image of some partonomy relationship and any other relationships chosen by users. The same visualization strategy was applied to render co-authorship networks. It is implemented as follows. The chosen partonomy relationship is drawn as a tree. The internal nodes of the tree are instances connected by the partonomy relation. The leaves of the tree represent researchers. The tree layout is created either with radial or with circular placement algorithm. Then each edge of the co-authorship network is laid out using the nodes of the tree as cubic B-spline control points. Each edge of the network is modeled as a single B-spline using the control points along the shortest path in the tree layout from one leaf point to another. For example, Figure 7 shows the co-authorship relationship between the research staff working in different geographical areas.

The basis of this drawing is a balloon layout corresponding to the relation "*Place-includes*". Small black circles represent geographic objects, such as country, city, village, etc., straight edges correspond to the relation "*Place-includes*". That is, the edge connecting the objects "Russia" and "Irkutsk" states that Irkutsk is located in Russia. Little gray circles depict individual researchers and the straight edges connecting researchers with localities represent the researcher's residence in the specified locality. The bright

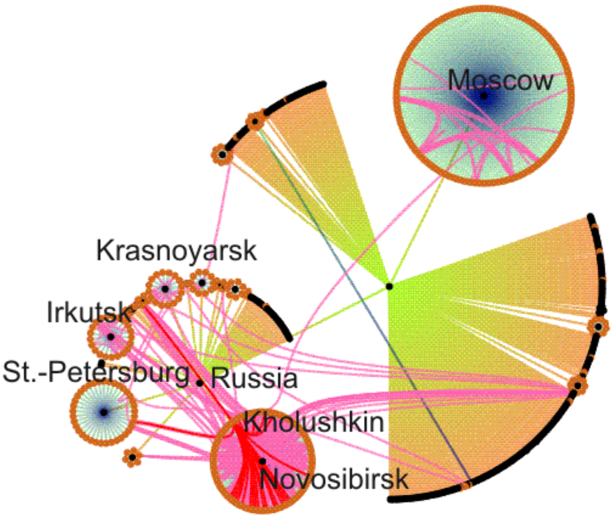


Figure 7. The co-authorship relations between the research staff working in different geographical locations

bundles depict the co-authorship relationship between researchers from different cities, and more subtle and darker bundles show the co-authorship relations of one selected researcher. It is possible to see that this researcher has collaborators in Novosibirsk, St. Petersburg, Krasnoyarsk, etc., but no collaborators in Moscow. The same drawing can serve as an example of visualization of defects in the input data. For example, all Russian cities are laid out along the circumference centered in a vertex “Russia”. But the city of Moscow is located on the perimeter of another circle. Apparently, the information that Moscow is located in Russia is missing in the test data.

Figure 8 shows the hierarchy of scientific divisions and the research staff working in different scientific divisions depicted by the radial placement algorithm. It is possible to see the whole structure of co-authorships between various scientific divisions, as well as the co-authorship relations of academician A.P. Derevianko. This figure demonstrates one more interesting effect. It is possible to see that the co-authorship relationship is rather wide and connects academician Derevianko A.P. with researchers working in many different scientific divisions. Each scientific division is shown as a circle with a small black circle in the center. The small gray circles situated on the perimeter of each black circle correspond to researchers working in the depicted scientific division. It is worthy to note that in the content of the

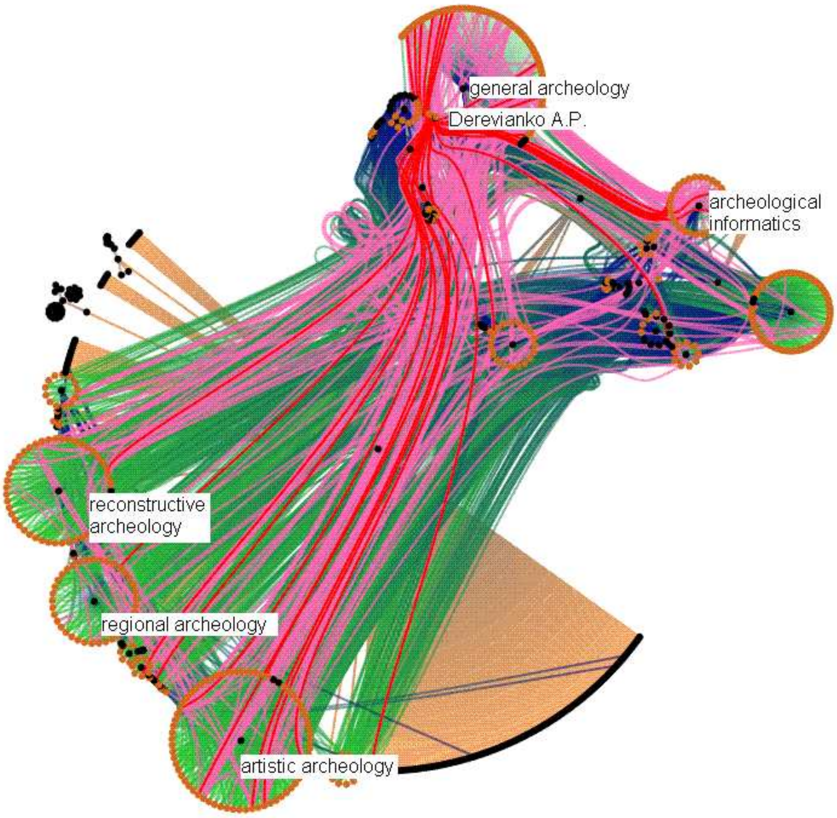


Figure 8. The co-authorship relations between the research staff working in different scientific divisions

archeological portal there is no explicit information about the scientific division where academician Derevianko A.P. works. This information is visible only through the co-authorship relations with other researchers having a well-described scientific division attribute.

5. Conclusion

We have presented two different methods for visualization of co-authorship networks extracted from the content of ontology-based portals. It is evident that the same visualization methods can be applied to many other kinds of networks. From our point of view, the main advantage of this kind of visualizations is their ability to juxtapose different aspects of data. These methods were tested on the data that describe the content of portals on archeology and computer linguistics, as well as the site of the Programming Department of the Mathematical Faculty of the Novosibirsk State University. They can be used at the stage of an ontology-based portal development to detect some errors or unfinished job of manual data input. Also, they can be used for visual analysis of the portal content during its entire life cycle.

In the near future, we plan to develop our visualization component in several directions. First, we plan to extend the input language of our subsystem towards the whole OWL language. It is also expected to fill up our library of algorithms with several new ones. In particular, it is planned to implement an algorithm showing the evolution of the portal content over time. When developing our subsystem, we used a free Java class library called JUNG [12].

Acknowledgement. The authors are very grateful to Yu.A. Zagorulko and S.V. Bulgakov for the test data in the xml-format.

References

[1] Katifori A., Halatsis C., Lepouras G., Vassilakis C., Giannopoulou E. Ontology visualization methods – a survey // ACM Comput. Surv. – 2007. – Vol. 39, No. 4.

[2] Apanovich Z. V., Vinokurov P. S., Elagin V. . An approach to visualization of knowledge portal content // Bull. Novosibirsk Comp. Center. Ser. Computer Science. – Novosibirsk, 2009. – IIS Special Iss. 29. – P. 17–32.

[3] Apanovich Z.V. Methods of navigation for graph visualization // Vestnik NGU. – 2008. – Vol 6, Iss. 3. – P. 35–47 (In Russian).

[4] Zagorulko Yu.A., Borovikova O.I., Kholushkin Yu.P. Construction of ontology for archeological knowledge portal // Information Technologies in Humanitarian Research. – Novosibirsk, 2006. – Iss. 10 (In Russian).

[5] Kholushkin Yu.P., Grazhdannikov E.D. System Classification of Archeological Science (Elementary Introduction in Science of Science). – Novosibirsk, 2000. – 58 p. (In Russian).

[6] Wuchty, S. Jones, B. Uzzi, B. The Increasing Dominance of Teams in Production of Knowledge // Science Express. – 2007. – Vol. 316, No. 5827. – P. 1036–1039.

[7] Jones, B. F. Wuchty, S, Uzzi, B. Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science // Science 322. – 2008. – No. 5905. – P. 1259.

[8] Barabasi, A.-L. The origin of bursts and heavy tails in human dynamics // Nature. – 2005. – No. 435. – P. 207–211.

[9] Newman M. E. J., Girvan M. Finding and evaluating community structure in networks // Physical Review E. – 2004. – No. 69. – P. 26113.

[10] Huang J. et al. Collaboration over time: characterizing and modeling network evolution // Proc. of the Internat. Conf. on Web Search and Web Data Mining. – 2008. – P. 107–116.

[11] Fruchterman T. M. J., Reingold E. M. Graph Drawing by Force-Directed Placement // Software -Practice and Experience. – 1991. – Vol. 21, No. 11. – P. 1129–1164.

[12] Kamada, T., Kawai, S. An algorithm for drawing general undirected graphs // Information Processing Letters. – 1989. – Vol. 31. – P. 7–15.

[13] Holten D. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data // Transactions on Visualization and Computer Graphics. – 2006. – Vol. 12, No. 5. – P. 741–748.

[14] Madahain O., Fisher D., Smyth P., White S., Boey Y. Analysis and visualization of network data using jung // J. of Statistical Software. – VV(II).

[15] Apanovich Z.V., Kislitsyna T.A. Vinokurov P. S. Flexible component for content visualization of ontology based portals during their lifetime // XII All-Russian Research Conf. RCDL-2010, Kazan, Russia, 2010. – P. 265–272 (in Russian).