Towards creating an evolvable semantic platform for formation of research teams

Eduard Babkin, Nikolay Karpov, Oleg Kozyrev

National Research University Higher School of Economics,
25/12 Bolshaja Pecherskaja Ulitsa, Nizhny Novgorod 603155, Russia ,

{eababkin, nkarpov, okozyrev}@hse.ru

Abstract. In this paper authors wish to present an approach to information modeling and software design suitable for developing evolvable semantic applications in the domain of research team formation. The novel proposals include specialization of generic paradigm of ontological engineering, specific types of machine-readable RDF ontologies and application of temporal look at information relevant for team formation. In order to validate theoretical approach a software prototype of the evolvable web-based semantic platform InfoPort was developed.

Keywords: teams formation, ontologies, service-oriented design

1 Introduction

Modern competitive research projects mostly require formation of strong and consistent inter-disciplinary teams [1][2]. In this case the cohesive team of scientific experts shall be combined taking into account competences of each specific performer. The team as a whole shall provide possibility of carrying out the research.

In universities such teams are frequently composed on demand. The issue of team formation, if it is understood as "personalization of the individual's allocation to a group"[3], becomes quite challenging in presence of time pressure, dynamically changed organizational structures and evolving employees skills. To our best knowledge most of university team leaders and grant writers try to solve such issue by *ad hoc* manner, mostly relying upon their intuition and informal inter-personal communications. At the same time the team leaders pessimistically estimate attempts to automate the team formation using information and communication technologies (ICT).

There are several important reasons for such pessimistic attitude. The team formation methods include both formal and human-oriented tasks to accomplish. For example, obviously semantic interoperability should be maintained among the members of the team who have overlapping scientific interests and research qualification. Also personal interconnections and social aspects have great impact, stimulating or declining willingness of team members to be vulnerable to the actions of others on the basis of the expectation that the other members will perform needed actions. The later issue usually is identified as trust management. In [4] Germain mentions that ability, benevolence, and integrity can parsimoniously encompass the concept of trustworthiness, the immediate precursor to trust.

Although several research initiatives and software tools exist for formation of teams in such contexts as manufacturing, collaborative learning, multi-disciplinary research [2][5][6] , the problem of proper conceptualization of the cross-disciplinary domain of team formation and effective information modeling of such complex matters as trustworthiness or competencies still reduces potential benefits of application of information technologies. Most of available software tools exploit traditional database approach and techniques of Human Resource Management (HR). Such approach does not fit well the situation. In presence of dramatic variations in the target domain of research and expected characteristics of the research team end-users wish to get access to evolvable applications. In that case the developers cannot rely upon stable database scheme, predetermined queries and stable application designs.

Because the task of information modeling has great significance for design of evolvable and dynamic applications in general, first of all we suggest to consider the problem of team formation from the information modeling point of view. In that context we applied principles of Semantic web and ontological engineering to provide adequate information models for an evolvable semantic platform in the domain of research team formations with particular interest to the trust management. In order to achieve semantic interoperability relevant well-known ontologies are used to link together different chunks of data. In our research the ontology engineering approach was similar to well-known METHONTOLOGY methodology [7]. Basic RDF inference techniques help to deduce entailments of organizational structure and domains of research interests.

Our principal novel proposal includes application of temporal look at information relevant for team formation. In parallel to traditional semantic repositories which contain only the latest snapshot of information we implemented a particular scheme for historic repository. This feature allows analyzing personal and collective evolution over the time. We found such capability very important for making strategic decisions about feasible directions of advances.

To validate our theoretical approach we designed a prototype of the evolvable web-based semantic platform InfoPort which supports formation of university research teams. The prototype combines four major own modules (data extractor/crawler, import, knowledge engine and web-frontend) and third-party RDF repositories. Paradigm of linked data, service-oriented design and wide application of open standards allow easy extending InfoPort platform and building full-fledged evolvable web-applications for automated or semi-automated formation of research teams.

The rest of the article is structured as follows. Section 2 describes design of principal data models, including overview of historical repository for temporal analysis of research interests. In Section 3 we present design of InfoPort platform and technical information about particular methods and algorithms of data crawling applied in our prototype. In the conclusion we compare our approach to similar researches and define further steps needed to mature the proposed models and software design.

2 Information modeling of the discourse domain

Information modeling and proper world representation play a crucial role in the task of teams formation. Until now in that domain the consensus is not achieved yet in precise definition and conceptualization of the core concepts, including the notion of teams. For example, in [1] J.R. Holenbeck et al. emphasize, that "any researcher doing a meta-analysis might find it difficult to explore the moderating influence of tasks, samples, and contexts because of the lack of consensus on classification systems." In [6] Hodik et al. also support the same statement in terms of lacking in adequate methods of competence management in distributed systems: "there is no common terminology of competency management; the terms 'competency', 'competency class', 'competency instance' and 'profile' are used in several slightly different meanings." For our purposes the following definition of competency [8] seems to be the most adequate: "The notion of competency provides an observable account of concrete human capacities under specific work conditions."

The situation of unstable terminology and confronting view points on team formation processes becomes the key obstacle on the way of exploiting information systems in general, and traditional database technologies with predetermined data schema in particular.

To get rid of repeating reengineering of DB schema in our approach to information modeling we offer to accept the hypothesis of open world and to implement the world representation using the generic concept of "linked data" and technologies of Semantic Web. In this case all information may be represented in the machine-readable form of RDF[1] triples. In the considered scenario RDF allows for extending the informational contents by demand supporting evolvability of our solution.

The second advantage of the proposed approach consists of strong semantic interoperability. Indeed, in the global network large amount of relevant ontologies and vocabularies exist in the form of RDF. Unified representation of heterogeneous information in the form of RDF and ability of SPARQL[2] query language to process federated queries give bold foundations for establishing tight relations between different relevant sources of information and maintain consistence and up-to-date state of information easily.

Selection of the basic semantic technologies determines a methodology for data modeling in our approach. To structure information about researchers, their skills, competences and other relevant facts we apply the method of ontological engineering. From our point of view it is similar to well-known METHONTILOGY methodology [7] which has been recommended by FIPA for the ontology construction task. In our research we also have passed through such development activities as specification, conceptualization, formalization and implementation. On the conceptual level we distinguish four types of ontologies which comprise the information model in the semantic platform, and are described below:

- organizational ontology;

[1] http://www.w3.org/RDF/
[2] http://www.w3.org/TR/rdf-sparql-query/

- ontology of scientific areas;
- personal ontology;
- historical ontology.

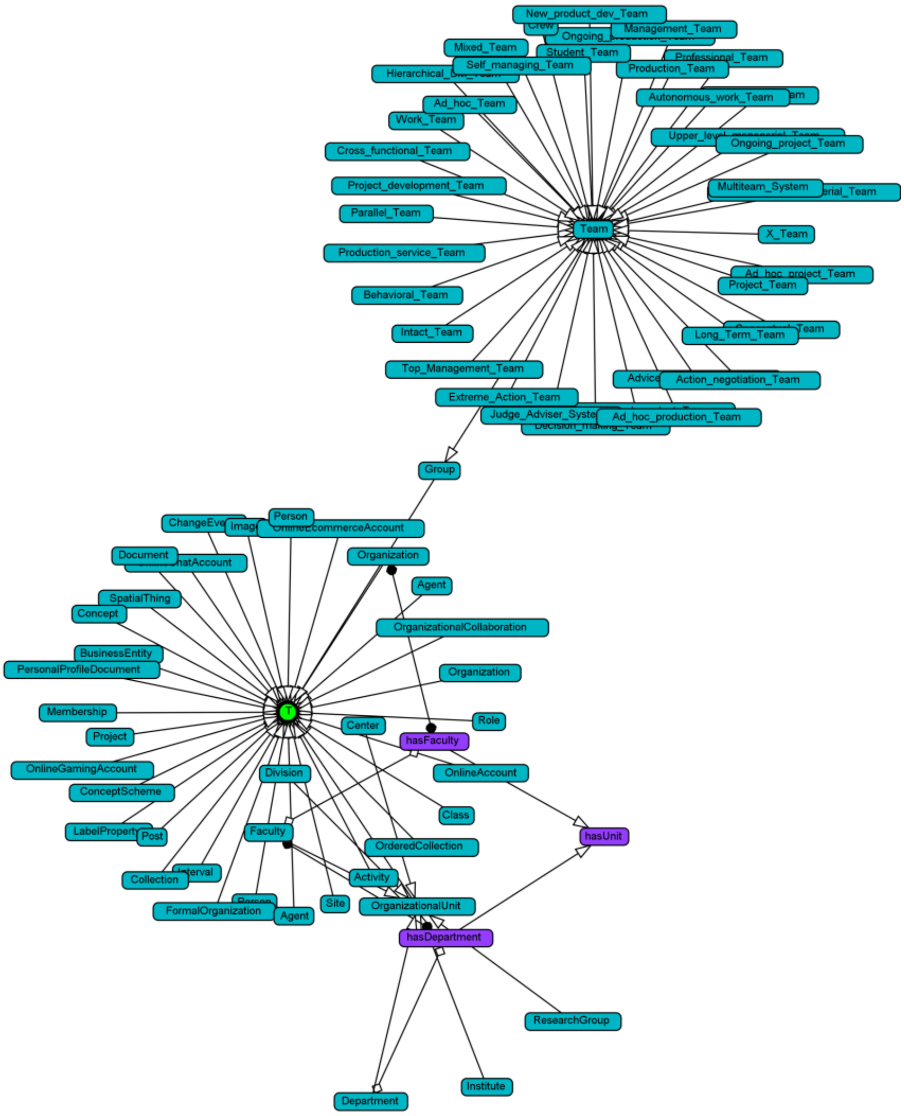


Fig. 1. Organizational ontology

Organizational ontology (Fig.1) contains classes and instances describing an organizational structure of particular research institution. Instances of the ontology represent both stable organizational units like departments or chairs, and *ad hoc* virtual structures like project teams. In order to maintain semantic interoperability we propose to build the organizational ontology on the basis of W3C organization ontology[3]. This ontology is aimed at supporting linked data publishing of organizational information across a number of domains. In order to present more elaborated foundations for teams formation we extended the original W3C organizational ontology introducing the typology of teams as it was proposed in [1].

The second ontology represents scientific areas and brings considerable contribution to the unification and semantic integration of researchers' characteristics. That's a matter of fact that multiple grant application systems, HR components and other ICT-systems require specification of research interests in terms of the limited set of taxonomies which describe scientific areas. This circumstance provides us for a practical ability to perform the first step of the semantic search among corresponding team's candidates without application of sophisticated natural language processing algorithms.

In our research several science taxonomies were considered. Currently two taxonomies are used together as a machine-readable ontology of scientific areas in our project. The first part of this ontology is international UNESCO nomenclature for fields of science and technology[4]. This nomenclature has such attractive features as deep hierarchical structure of concepts, good coverage of scientific domains, several national translations and wide acceptance in international science as an international standard. Also this nomenclature has more affordable structure in comparison with Dewey decimal system classification[5]. The second part of the ontology was developed by our research team on the basis of the national scientific classification called e-library, which is widely used in the national practice of scientific research, multiple grant application systems, etc. Exploiting of common parent concepts in both parts of the ontology may leverage their simultaneous using and interconnecting with other search engines and techniques. That's why for developing the part of national scientific classification we propose to use the concepts of knowledge structuring developed in W3C technology "Simple Knowledge Organization System" (SCOS)[6]. In such case the developed ontology becomes a kind of instance-based ontology where taxonomy topics are represented as instances of skos:Concept class. Relations are established by link types skos:narrower and skos:broader. Predicate links skos:prefLabel and skos:notation are used to define human-readable properties of the instances. In the result the principles of mapping between existing taxonomy and derived ontology of scientific areas may be represented as Fig.2 demonstrates.

---

[3] http://www.w3.org/TR/2012/WD-vocab-org-20121023
[4] http://skos.um.es/unesco6/
[5] http://dewey.info/
[6] http://www.w3.org/2004/02/skos/

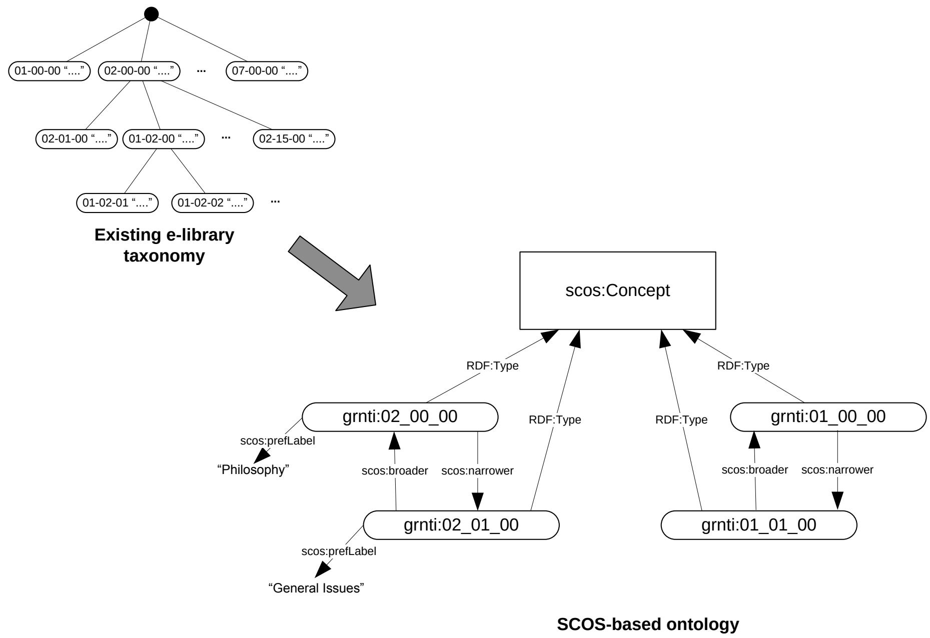


Fig. 2. Principles of mapping between taxonomy and the ontology of scientific areas

The third personal ontology represents a significant part of professional knowledge needed for team formation. This ontology represents an actual snapshot of present person-oriented information available in the result of periodical web crawling through the university portal, results of the articles processing, and results of logical inference which provide for description of competences and skills. In our research we classify information about researchers according the following principal aspects:

a. Researcher as a person.
b. Researcher as a skillful agent.
c. Researcher as a team member.

At some extent the specified aspects reflect the Gero's Function-Behavior-Structure (FBS) Ontology [5] and leverage application of the paradigm of smart design of research community in presence of situatedness. The first aspect of personal ontology describes such traditional and more less stable characteristics as last and first names, education, etc. The second aspect is tightly related to the notion of skills in terms of relevant topics of scientific classification taxonomies and keywords describing interests and skills of the person in free form. The last aspect contains information about participating the person in different kinds of projects or teams, connects personal ontology with the organizational ontology.

In order to consistently represent all three aspects of the person in terms of the same ontology we offer to use major concepts of W3C FOAF (Friend of Friend) vocabulary specification[7]. This specification is aimed at machine –readable representation of reusable information about people, groups, companies and other entities acting as agents in different activities. In this case their relationships become explicitly represented by different types of FOAF concepts from such sections as FOAF Basic (familyName, givenName, firstName), FOAF Personal Info (interest, currentProject, pastProject, etc), FOAF Projects and Groups, and FOAF Documents and Images (image).

In addition to the standardized FOAF classes our personal ontology also contains several new link classes for semantic integration with other aspects. For example, our personal ontology provides class infoport:classify for linking with the ontology of scientific areas and class infoport:freeKeywords for specification of skills in terms of free text keywords. At the same time W3C FOAF and W3C organization ontology have connection via a shared concept foaf:Agent. An example of consistent representation of different aspects is depicted on Fig.3.

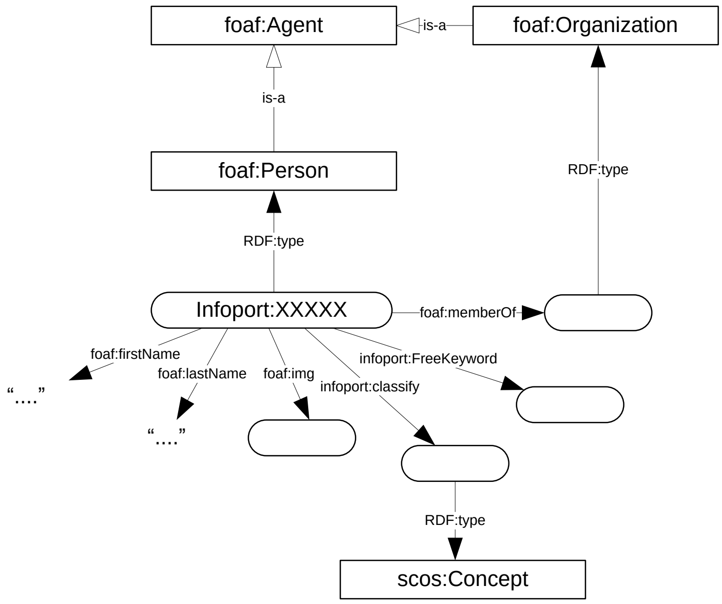


Fig. 3. Classes of personal ontology

The introduction of the fourth element in our approach, the historical ontology, strongly differentiates our work from other known ontology-based approaches to the teams formations. We believe that decision support of such complex task as formation of effective research teams should take into account not only snapshot of present skills and competences, but instead it should be based on the temporal representation and the historic, evolutionary view point.

[7] http://xmlns.com/foaf/spec/20070524.html

Temporal representation may be achieved using the paradigm of system's life-cycle. More specifically, in our approach we use the concepts of BORO [9] and principles of temporal modeling proposed in ISO 19526[8]. Temporal aspects of personal information may be represented in the historical ontology using reification of RDF statements as follows (Fig.4).

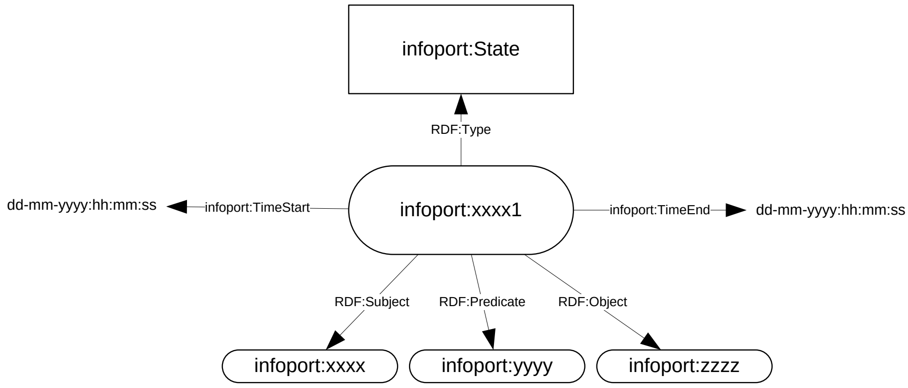


Fig. 4. Structure of the historical repository

Each modification in personal ontology is reflected in the historical ontology via modification of corresponding instances of infoport:State class. Modifications include creating new instances with corresponding values of infoport:TimeStart attribute and attaching attribute infoport:TimeEnd to existing instances. Delete operation is never applied to the historical ontology.

Given complete temporal information from the historical ontology InfoPort platform may provide a team builder with intuitively clear representation of personal evolution in terms of time line view (Fig.5).

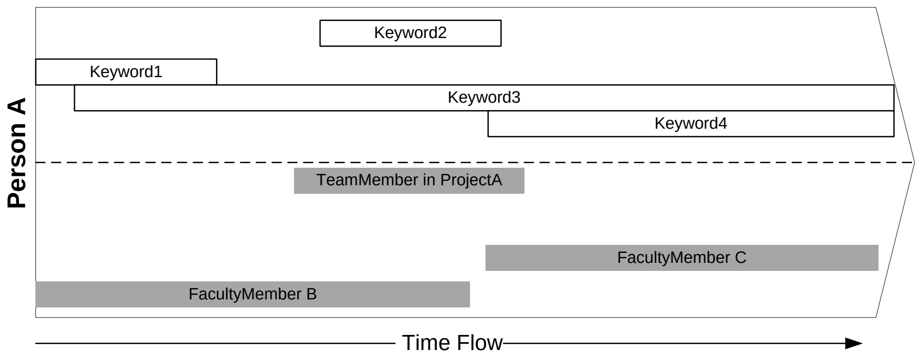


Fig. 5. The timeline of the person

[8] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=29556

This view combines segments of keywords or topics of science taxonomy together with temporal information about participating in particular research projects or stable organizational structures. Such combination allows for deep analysis of transient or sustainable scientific interests of the person as well as providing the basis for the trust management. Indeed, the temporal view point on the person's research interests has direct connection with the concept of cognition-based trust. In [4] Germain notes that cognition-based trust occurs due to perceptions of competence, reliability, and dependability It is based on reasoning about others' reliability and dependability. The competence, integrity, ability, and past record of the person being trusted form the rational basis for withholding trust.

Combining all aspects together our approach to information modeling facilitates effective design of evolvable semantic platforms for support of team formation providing:

1) foundations for semantic interoperability;
2) extensibility of information model;
3) temporal representation of information.

Taking a pragmatic point of view we may note that explicit separation of four ontologies allows for flexible implementation strategy. It means that given specifics and update policies for each of the four ontologies the developers can easily choose the most appropriate implementation of RDF repository in terms of performance and representation power of the query language. The available solutions may vary from simple RDF triple stores to quite sophisticated OWL inference engines. Example of such ontology-related implementation strategy is given in the next section.

3 Design of the platform

Proposed methods of the information modeling were practically applied during design and implementation of the evolvable semantic platform for research team formation called Information Port (InfoPort). High level design of the InfoPort platform is based on principles of Service-Oriented Architecture. The following services are distinguished (Fig.6):

- Crawler Service.
- Import Service.
- Web Backend Service.
- RDF Store Service.

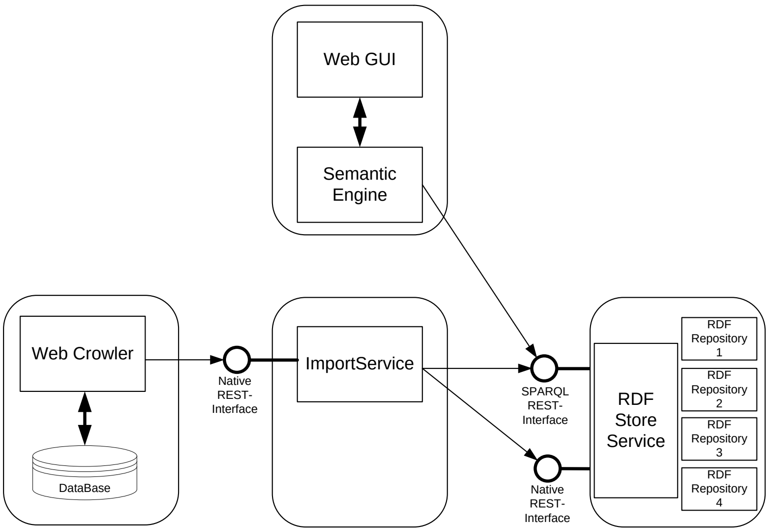


Fig. 6. InfoPort Conceptual Design

Implementation of Crawler Service uses Python[9] programming language and Lxml[10] library for processing HTML. The service is monitoring the information in the website using Hash function within a week period. After changing of information new structured data crawls and extracts from their pages using XML parser. The data represents as RDF triples and stores to the MySQL[11] database. Triples include hierarchical information as it originally is in the source. The first level is an alphabetical ordered list of group of scientist, second is a scientist with his personal interests and papers, and third is papers with its features. This data compose JSON request for native REST interface of ImportService

Implementation of Import and Web Backend services is based on JRuby programming environment and RubyOnRails Framework. During implementation several Ruby libraries for working with RDF were tested like RDF.rb[12], Sparql-client library[13], as well as RDF mappers SPIRA[14], ActiveRdf[15]. Due to performance issues after several experimental mock-ups the final design decision was to refuse using high-level object-mapping libraries (i.e. SPIRA). Instead, the implementation uses low-level SPARQL queries and RDF manipulation methods via native REST-based interface provided by RDF.rb library.

---

9 http://www.python.org
10 http://lxml.de
11 http://www.mysql.com
12 http://rdf.rubyforge.org/
13 https://github.com/ruby-rdf/sparql
14 https://github.com/datagraph/spira
15 http://activerdf.org/

For RDF repository Services two alternatives were tested in our project: OpenRDF Sesame[16] and free version of Allegro Graph 4.0 [17]. Four different kinds of RDF repositories were created for each of four ontologies:

1) Persistent repository for personal ontology with capability of RDFS inference.
2) In-memory repository for e-lib taxonomy with capability of RDFS inference.
3) In-memory repository for organizational ontology with capability of RDFS inference.
4) Persistent plain RDF repository for historical ontology.

InfoPort platform implements communication between services using REST-interfaces and JSON-based encoding. For example, the table below describes the REST interface of the import service.

Table 1. Interface specification for InfoPort Import Service

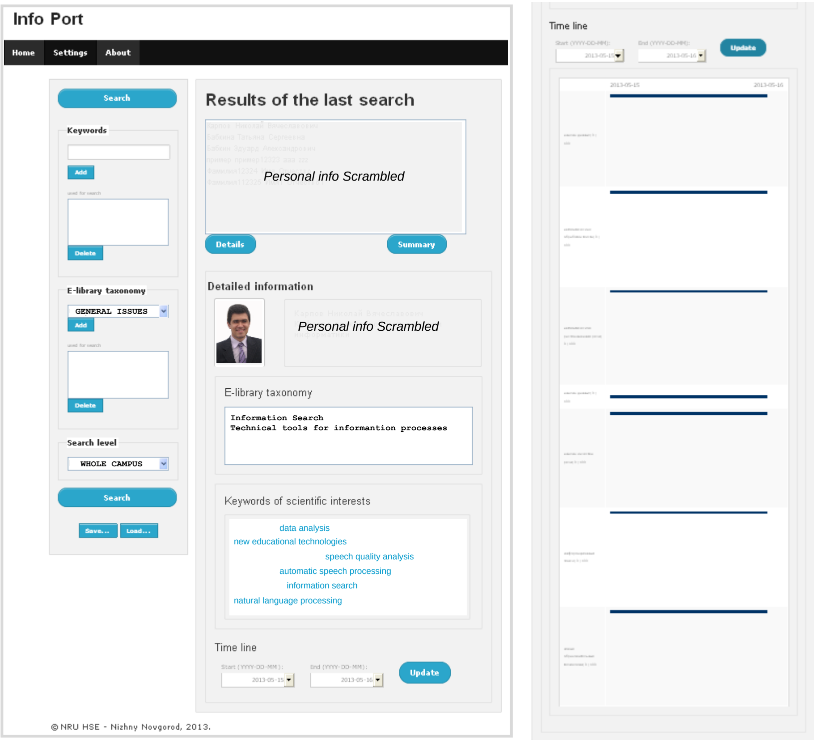
Method	HTTP	URL	Data Encoding
Create	POST	http://xx.yy.xx.ww/infoport/entities	JSON
Read	GET	http://xx.yy.xx.ww/infoport/entities/:id.json	JSON
Update	PUT	http:// xx.yy.xx.ww/infoport/entities/:id	JSON
Delete	DELETE	http:// xx.yy.xx.ww/infoport/entities/:id	

Currently the InfoPort production environment uses deployment of all components to the same Tomcat server, however distributed configuration is also available due to wide using federated SPARQL queries.

Experimental setup of InfoPort contains factual information about more than three hundred employees of HSE NRU branch at Nizhny Novgorod. InfoPort updates information every month and temporal part of ontology is modified accordingly.

The multi-service graphical user interface is available for team builders (Fig.7). This interface provides such functions as multi-criterion search, visualization of present keywords in the form of tag cloud, visualization of temporal characteristics of the person in the form of time line.

[16] http://www.openrdf.org/
[17] http://www.franz.com/agraph/allegrograph/



a) **b)**

Fig. 7. InfoPort User Interface:

a) – front page; b) – enlarged view of personal time line.

4 Discussion and conclusion

In this article we presented a new approach to support of research team formation focusing in information modeling and ICT-support via the evolvable semantic platform. The proposed methods of modeling use the concepts of "linked data" and ontological engineering, supporting the following statement of Christakis and Fowler: "To know who we are, we must understand how we are connected"[10].

Several other approaches also use semantic modeling for teams or group formation and offer using of FOAF vocabulary for building virtual teams and request- based virtual organizations in manufacturing [6][11], Communities of Practice (CoP) oriented towards learning activities [12][13][3]. However those proposals mainly target the domain of collaborative learning and do not take into account intrinsic features of the research teams formation (like specific team types mentioned in [2]).

On the level of information modeling the principal distinction between our approach and other mentioned works mostly deals with modeling the temporal aspect.

Such research works in social networking as [14] introduce a similar concept of dynamic co-authorship maps but only partially investigate the temporal aspects of connections between persons and skills.

Machine-readable structured ontologies developed according to our approach laid in the foundation of the semantic service-oriented platform InfoPort. The purpose and functionality of that platform may be compared with e-Cat architecture [6][11] and such systems as ECOLEAD and PANDA which represent prototype implementations of the e-Cat architecture. The e-Cat is an agent-based architecture for the facilitation of members’ profiles and competencies in alliances of small and medium enterprises. E-Cat architecture also uses the concept of life cycle, but that concept is applied to modeling the whole virtual organization instead of modeling individual evolution of skills and competences of a researcher as it is done in our approach. Also e-Cat architecture precisely defines the notion of competency and uses graph-oriented structures for their representation which are close to ontologies.

The developed ontologies and software implementation of the InfoPort platform open opportunities for further developing methods and tools for decision support during research teams formation. Several principal research interests form the roadmap for the next steps in our research. Because e-Cat architecture does not provides complete solutions for the problems of automated deducing competencies and ensuring consistency among the shared definitions of competencies we wish to include these issues to the roadmap of our further research. The next target is to develop a comprehensive formal model and practical algorithms for automated team formation based on gathered ontological data. In [15] Ounnass et al propose to use the technique of constraints satisfaction to form the learning groups. In [16] Vokřínek et al. propose to use distributed multi-agent methods of formation of virtual organizations based on contract net and auctions. Their proposals give bold foundations for effective application Multi-agent technologies and multi-agent communication protocols. In our case application of soft–computing paradigm of Kohonen’s self-organizing Maps (SOM) allow formation of hierarchical teams.

A promising direction of the research includes integrating own personal ontologies and existing individual-oriented vocabularies and ontologies for competency specification. The later include General User Model Ontology (GUMO)[17] and the former include the Reusable Definition of Competency or Educational Objective (RDCEO) or generic competency schema (GCS)[8]. Such integration will facilitate advanced search capabilities augmented by psychological testing, thorough description of researchers behavior and analysis of competencies gaps. This integration will be a solid basis to cope the problems of strong personality in research teams overviewed in [2]. Inclusion into consideration such aspect will bring us closer to the solution of trust management in the research teams and practical application of the concept of templates for successful virtual teams, as it was described [4].

References

1. Hollenbeck, J.R, Beersma, B., Schouten, M.E., "Beyond team types and taxonomies: A dimensional scaling conceptualization for team description," *Academy of Management Review*, Том 37, Выпуск 1, January 2012. — с. 82-106.
2. Dodson, M.V., Guan, L.L., Fernyhough, M.E., Mir, P.S., Bucci, L., McFarland, D.C., Novakofski, J. et al., "Perspectives on the formation of an interdisciplinary research team (short survey)," *Biochemical and Biophysical Research Communications*, Том 391, Выпуск 2, January 2010. — с. 1155-1157.
3. Ounnas, A., Davis, H. C., Millard, D. E. Semantic Modeling for Group Formation // Procs. of Workshop on Personalisation in E-Learning Environments at Individual and Group Level (PING) at the 11th International Conference on User Modeling UM2007. — Corfu, Greece, 2007.
4. Germain, M.-L., "Developing trust in virtual teams," *Performance Improvement Quarterly*, Том 24, Выпуск 3, 2011. — с. 29-54.
5. Gero, J.S., Kannengiesser, U., "An ontology of situated design teams," *Artificial Intelligence for Engineering Design, Analysis and Manufacturing: AIEDAM*, Том 21, Выпуск 3, June 2007. — с. 295-308.
6. Hodík, J., Vokřínek, J., Bečvář, P., "Support for virtual organisation creation - Partners' profiles and competency management," *International Journal of Agent-Oriented Software Engineering*, Том 3, Выпуск 2-3, 2009. — с. 230-251.
7. Corcho, O., Fernández-López, M., Gómez-Pérez, A., "Methodologies, tools and languages for building ontologies. where is their meeting point? ," *Data and Knowledge Engineering*, Том 46, Jan. 2003. — с. 41-64.
8. García-Barriocanal E., Sicilia M.-A. Sánchez-Alonso, S., "Computing with competencies: Modelling organizational capacities," *Expert Systems with Applications*, Том 39, Выпуск 16, 2012. — с. 12310-12318.
9. Business Objects: Re-engineering for Re-use, 2nd ed. / Partridge, C. — The BORO Centre, 2005.
10. Connected: the surprising power of our social networks and how they shape our lives / Christakis, N. A., Fowler, J. — New YorkLittle, Brown and Co, 2009.
11. Hodík, J., Vokřínek, J., Bíba, J., Bečvář, P, "Competencies and profiles management for virtual organizations creation ," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Том 4996, 2007. — с. 93-102.
12. Da Silva, J. L.T., Ribeiro, A. M., Primo, T.T., Viccari, R.M., "A reference profile ontology for communities of practice," *International Journal of*

Metadata, Semantics and Ontologies, Том 7, Выпуск 3, November 2012. — с. 185-196.

13. Isotani, S., Inaba, A., Ikeda, M., Mizoguchi, R., "An ontology engineering approach to the realization of theory-driven group formation," *International Journal of Computer-Supported Collaborative Learning*, Том 4, Выпуск 4, 2009. — с. 445-478.

14. Steffen-Fluhr, N., Collins, R., Passerini, K., Wu, B., Gruzd, A., Zhu, M., Hiltz, R. Leveraging Social Network Data to Support Faculty Mentoring: Best Practices // Women in Engineering Program Advocates Network (WEPAN) National Conference. — Columbus, OH., USA, 2012.

15. Ounnas, A., H., D., Millard, D. A Framework for Semantic Group Formation // In Procs. of The 8th IEEE International Conference on Advanced Learning Technologies (ICALT 2008). — Santander, Cantabria, Spain, 2008. — с. 34-38.

16. Vokřínek, J., Bíba, J., Hodik, J., Vybíhal, J., "The RBVO formation protocol," *International Journal of Agent-Oriented Software Engineering*, Том 3, Выпуск 2-3, 2009. — с. 135-162.

17. Heckmann, D., Schwartz, T., Brandherm, B., Schmitz, M., Von Wilamowite-Moellendorff, M., "GUMO - The general user model ontology," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* , Том 3538, 2005. — с. 428-432.