Semantic Fragment of a Research e-Infrastructure: necessary information objects, tools and services

Sergey Parinov

Abstract

Basically semantic linkage technique is used to specify in computer readable form known facts or relationships that definitely exist between information objects like people, organizations, research results, etc. Recently developers also started using it to visualize scientists' opinions or scientific hypotheses (e.g. inference/deduction, impact/usage, theoretical hierarchy, etc.) about relationships between research objects. Based on CERIF Link entity and the Semantic Layer and assuming that scientists typically re-use research objects by making relation-ships between them we propose a sketch of a research e-infrastructure semantic fragment, which allow scientists unlimited re-use of research information systems (RIS) content. After some development a semantic linkage technique provides scientists with tools and services for semantic linking of any pair of research objects, which metadata are available within content of any RIS. This application also allows scientists a decentralized development of semantic vocabularies that guarantee a covering by this technique any new types of relationships. In the paper we discuss information objects, tools and services which are necessary for proper functioning of proposed fragment of the research e-infrastructure. We also discuss a "quality control" topic which in this context is very important.

Keywords: ESRI ArcGIS for INSPIRE, Estonia, Spatial Data Infrastructure, automatic data update mechanisms.

1 Introduction

Basically a semantic linkage technique is used to visualize known facts or obvious relationships that exist between information objects representing people, organiza-tions, research results, scientific assertions, etc. At abstract level this technique is specified in RDF (http://www.w3.org/RDF/). It is also developed in CERIF as a spe-cific data model with a focus on research (Jorg et al. 2012a, 2012b). Some examples of practical implementation of initial RDF Semantics specification (http://www.w3.org/TR/2004/REC-rdf-mt-20040210/) can be found e.g. in Nano-publication approach (Groth et al. 2010). Implementation of the CERIF approach occurs in many projects; e.g. in the Semantic Linkages Open Repository (SLOR) project (Parinov 2012).

Information objects which might be semantically linked are produced by research information systems (RIS), which include a whole spectrum from Institutional Repos-itories (IR) to full-functional CRIS. We assume more or less homogeneous structure of information objects (e.g. CERIF compatible) and each of them has a unique identifier. We also assume that all information objects matched these requirements belong to research Data and Information Space (DIS) and we can operate with them in some standardized way.

For example, a fact that a person "A" works for organization "B" can be visualized by linking the information objects of the person and the organization existed in DIS. Person's current position "C" can be expressed as a semantic assigned to the linkage. Typically such semantic linkages are created by developers of RIS when they are designing data storage of the system. And current semantic linkage technique assumes that a linkage exists as some additional fields to metadata of one (or both) of linked objects. See examples at (Jorg et al. 2012c).

Recently appeared "nano-publication" approach (Groth et al. 2010) and "object-for-reuse" concept (Parinov 2010) demonstrate an advanced semantic linkage tech-nique that allows a creation of semantic linkages also by scientists when they deposit their assertions/research artifacts at some repository (e.g. at ConceptWiki.org). Na-nopub.org explains how scientists can create nano-publications: "Specific relations between entities … can be thought of as specific scientific assertions. When converted to RDF, these assertions can be represented as a collection of semantic triples."

In this paper we discuss the next step: how the semantic linkage technique should be developed to visualize not only scientific assertions but also professional opinions or scientific hypotheses about relationships between research objects. This approach is definitely demanded by the community since typically scientists produce not only texts (papers, articles, books, etc) but scientific relationships between existed research objects as well. And existed technique to visualize such relationships – e.g. a mecha-nism of citations – still is not upgraded to abilities of modern ICT.

Some new types of relationships between research outputs that scientists may want to express are: inference/deduction, impact/usage, theoretical hierarchy, and so on (Parinov 2012; Parinov and Kogalovsky 2011). In this case the semantic linkages can carry information that, e.g. a research output "D" is produced by a person "A" (in a role "author") and with a financial support of an organization "B" (in a role "fund-ing") as the scientist's research result produced by a logical inference based on anoth-er research output "E". And at the same time, say, the research output "E" provides a broader theoretical concept then the "D".

In this example the first type of semantic relationships between "D" and "E" is the "scientific inference" with the meaning "used as a base", and the second type – the "theoretical hierarchy" with the value "broader".

So now there are two cases of using semantic linkage technique: a) to visualize re-lationships between known facts and/or research assertions, and b) to visualize opin-ions and/or hypothesis. Important differences between "a" and "b":

(1) typically linkages are created: for "a" at the same time when linking infor-mation objects are created; but for "b" – when linking information objects already exist and available for scientists;

(2) a necessity to have for the case "b" something like a scientific journal's per-reviewing to provide a "quality control" over publishing semantic linkages.

The main objective of this paper is to propose an approach which responds on:

- what information objects, tools and services are needed to give scientists a sim-ple and reliable way to visualize their opinions and hypothesis about relation-ships between information objects over all available content of RIS; and

- what tools and services are necessary for the scientific community to provide some kind of “quality control” over individual opinions/hypothesis and to keep the right balance of a freedom for scientific opinions/hypnotizes clashing and a respect of scientific ethic and norms.

Proposed in this paper approach should works as a universal solution covers the both cases of using semantic linkage technique. A set of tools implemented this tech-nique has be supplemented by some services which allow scientific community to provide some quality control over published semantic linkages. Altogether these tools and services are forming a semantic fragment of a research e-infrastructure.

In the next section we discuss a concept of the semantic linkage technique devel-opment according requirements listed above.

A revision of a set and templates of information objects is necessary for proper representation of new research relationship types. We propose two interrelated types of semantic objects: (1) a semantic linkage and (2) a semantic meaning. In the section “Information Objects” we discuss its templates based on CERIF and other relevant topics.

In the section “Tools and Services” we give a description of necessary instruments. One part of the instruments must be designed inside local RIS environment, and an-other one - outside it. This second part of tools and services can be called as a seman-tic fragment of research e-infrastructure.

Technical realization of discussed tools, services and the whole concept are cur-rently implemented as a part of the Socionet system (socionet.ru). But a discussion of the realization topics comes outside of this paper.

At the conclusion we summarize benefits of the proposed application.

2 A concept of the semantic linkage technique development

When semantic linkages are carrying not only known facts but also hypotheses ex-pressed as a personal opinion of some scientists the implementation of this technique becomes more complex. E.g. in the mentioned above example of relationships be-tween “D” and “E” the specified semantic values have a sense if only the linkages are oriented and the orientation is: for the first - “D” -> “E”; for the second - “E” -> “D”.

According our objective the semantic linkage technique has to satisfy following important requirements:

- linkages with assigned semantic meaning should be created not only by RIS de-velopers, but also directly by scientists or their assistants with explicit indication of who is an author of the linkage and responsible for semantically expressed professional opinions or scientific hypothesis;
- a technique of semantic linkages should work at standalone mode, i.e. inde-pendently of linked objects metadata, since in many cases the semantic link-age’s attributes cannot be directly included into linked objects metadata;
- semantic linkages should be deposited by their authors into DIS as a public in-formation resource;
- since linkages are created in decentralized mode there should exist a submission procedure, which implies a moderation and validation of semantic linkages by the community before they will be publically available;
- since a set of relationship types used for semantic linkage creation cannot be completely predefined, there should exist an ability for scientists and developers to expand in some controlled way semantic vocabularies associated with types of relationship;
- the scientific community should be able to make selection by quality, impact evaluation and multiple re-use over all created semantic linkages.

Ideally any scientist should be able to establish any number of consistent and rele-vant research relationships in visual and computer readable form between a pair of any available research objects. And any scientist should have an opportunity to pro-pose new types of research relationships for covering by this technique.

But at the same time the scientific community should have some kind of quality control over new semantic linkages and new types of relationships submitted by sci-entists for a public use. And the community should be able to evaluate impact of submitted semantic objects and re-use it in multiple forms.

Technically it means that scientists should have a personalized tool to cre-ate/manage the semantic linkages and, as well, vocabularies of semantic meanings. Such personalized tool can be a part of a Content Management System (CMS), which exist in many RIS, including repositories.

A scientist can use a CMS of some local RIS "A" for creation semantic linkages only. It means the linking objects belong, e.g. to RIS "B" and "C". In this case the linkage's attributes created by the scientist in RIS "A" cannot be directly embedded into metadata of linking objects, as it is proposed e.g. in CERIF 1.4 (Jorg et al. 2012c), because owners of the metadata are RIS "B" and "C".

To be universal the semantic linkage technique should produce linkages having a status of regular DIS information objects and so they should exist separately from the metadata of linked information objects. In such situation we have to specify a new data type "semantic linkage" and design its template which can be used by CMS of local RIS to create semantic linkages as autonomous objects.

All semantic linkages created in any RIS by this decentralized way should have unique identifiers. In the Section 4.3 we discuss how to perform this requirement.

Each semantic linkage has as an attribute a semantic meaning that belongs to some semantic vocabulary. A set of research relationship types is open for decentral-ized extension and development by scientists. So we should have a template to create semantic vocabulary as a collection of standardized objects of the “semantic mean-ing” data type, which also can be done by using CMS of local RIS.

Initial set of rendered scientific relationships built from different already existed ontologies (Parinov 2012; Parinov and Kogalovsky 2011) includes: (1) relationships between various research outputs like inference, usage, impact, comparison, etc; (2) relationships between elements of the set {scientists, organizations}; (3) relationships between research outputs on the one hand and elements of the set {scientists, organi-zations} on the other.

A flow of semantic linkages and/or semantic vocabularies submitted by scientists for a public use is moderated as it is typically organized for the process of research papers depositing and self-archiving.

For better utilization of created linkages and its multiple re-uses by the community all semantic objects (linkages and vocabularies) created at local RIS are aggregated at some central data storage. In more details it is discussed in the Section 4.3.

To perform necessary utilization and re-use of accumulated at the central data storage semantic objects we have to design some basic services, which is forming a research e-infrastructure semantic fragment:

1. An aggregation, storage and sharing of semantic linkages and semantic vocabular-ies at the central data storage;

2. Maintaining of requests from local RIS to the central data storage about (a) existed ingoing and outgoing semantic linkages for specified information object and (b) available semantic vocabularies and updates of its content;

3. Interpretation of semantics, which in the general case are necessary for visualiza-tion of linkages at central data storage as multilayer networks of scientific relation-ships different types;

4. Navigation and searching over the objects of both types at the central data storage;

5. Monitoring service, which is tracing changes in linked objects and linkages itself. It runs a notification service and collects statistics;

6. Notifications for authors of linked objects, for authors of linkages, for readers of linked objects;

7. Scientometrics to collect quantitative data (e.g. numbers of linkages) and qualita-tive data about relationship types and semantic meanings.

These listed services give the community some additional capabilities that can be characterized as better “information metabolism”. It is discussed in more details in the section "Tools and services".

3 Information objects

In this section we discuss templates of two important data types of information objects: (1) a semantic linkage and (2) a semantic meaning. A semantic linkage includes a semantic meaning as an attribute, to characterize a type of relationship between linked information objects.

To make this discussion compact we use by default CERIF terminology and speci-fications wherever possible, primarily the Link Entity and the Semantic Layer (Jorg et al. 2012a, 2012b).

We assume that templates proposed in this section should be implemented at local RIS: 1) to create standardized objects of both types which can be easily re-used out-side local RIS; and 2) to harvest in proper way standardized semantic vocabularies from other local RIS or from the central data storage and to use it within a tool to create/edit semantic linkages.

Since we have to operate with created semantic linkages as regular information objects of DIS, which should be convenient for storing and processing, displaying for navigation across all linkages and/or across linked objects, delivering, harvesting, indexing for searching by keywords, and so on, in some cases we have to expand the initial CERIF notation of the both semantic objects. This extension of initial CERIF set of fields can look as redundancy, but it is justified in our practical implementation of this concept.

3.1 Semantic linkage

As an initial model of information objects with the "semantic linkage" type we took the specification of CERIF Link Entity (Jorg et al. 2012a, p. 33; Jorg et al. 2012b, p. 13).

To make the semantic linkage template self-contained we add to the initial set of fields:

- more data about pair of linked objects, including clear specification of the link-age orientation (which object is a source the target one), unique IDs of linked objects, its data types and titles;
- ID and a name of selected semantic meaning and also URI and a name of parent semantic vocabulary;
- a field for comments that allows scientists to provide explanations and com-ments about specified semantic linkage parameters;
- a group of fields with personal, organizational data about an author of the se-mantic linkage and about a provider of the service;
- a title and unique ID of the linkage itself, creation and revision dates and other extra attributes if necessary.

Personal data about the linkage's creator including his/her e-mail address is used to notify the creator about a need to revise the linkage’s correctness because of changes in linked objects.

A title of the semantic linkage is needed to build a table of contents and for navigation across the whole set of created semantic linkages.

We assume that semantic linkage parameters are changeable and, in principle, have a status similar to electronic publication. So it explains why we require a revision date.

Some additional details about semantic linkage specification can be found in (Pa-rinov 2012; Parinov and Kogalovsky 2011). Below we provide specification details.

In XML notation the main attributes of the link entities are (CERIF 1.3 XML, p. 13):

```
<cfPers_OrgUnit>
   <cfPersId>ID</cfPersId>
   <cfOrgUnitId>ID</cfOrgUnitId>
   <cfClassId>ID</cfClassId>
   <cfClassSchemeId>ID</cfClassSchemeId>
   <cfFraction>Float</cfFraction>
   <cfStartDate>Timestamp</cfStartDate>
   <cfEndDate>Timestamp</cfEndDate>
</cfPers_OrgUnit>
```

This CERIF 1.3 example of a semantic linkage metadata provides the minimum required attributes. It is not enough for our application, where the

semantic linkages are created by scientists as their intellectual products in decentralized mode by tools of different RIS and the linkages should be visual at research DIS not only as connections between objects, but also as regular information objects.

CERIF gives a recommendation: “The physical name of link entities is composed of the names of the two linked entities, including the CERIF prefix as follows: cfEnti-ty1Name_Entity2Name.” (CERIF 1.3 FDM, p. 33). It is restrictive for us, since we don't know in advance what new types of entities will be served this application in the future. Information about data type of linked objects should be specified in appropri-ate fields of the link entity metadata, but not as its physical name.

Since we have to operate with the semantic linkage as regular information objects of DIS, which should be convenient for storing and processing, displaying for naviga-tion across all linkages and/or across linked objects, delivering, harvesting, indexing for searching by keywords, and so on, we have to modify initial CERIF notation of the Link Entity. To make the semantic linkage template as universal as possible, we designed a template consisted of three groups of XML tags:

- the <cfLinkage> group is an expansion of four initial tags <cfPersId>, <cfOrgUnitId>, <cfClassId> and <cfClassSchemeId>; now it contains more data about pair of linked objects, where one is a source object and the another - a target one;
- we added the <cfAuthorship> group of tags with personal, organizational data about authors of the semantic linkage and provider of the service;
- other tags from initial CERIF template moved to <cfGeneral> group, where there are a data type code, title and unique ID of the linkage, dates and more attributes if necessary.

The first group of XML tags as an example describes that the source information object with the type “person” has an outgoing linkage to the target one of the same type “person” with the semantic “Supervisor”:

```
<cfLinkage>
   <cfObjects>
      <cfFrom>   data about the source object
         <cfDataType>person</cfDataType>
         <cfPersID>ID</cfPersID>
      </cfFrom>
      <cfTo>   data about the target object
         <cfDataType>person</cfDataType>
         <cfPersID>ID</cfPersID>
      </cfTo>
   </cfObjects>
   <cfData>
      <cfClassification>   data about semantic
         <cfClassSchemeId>ID</cfClassSchemeId>   vocabulary
         <cfName>CERIF Semantics for Person</cfName>
         <cfURI>.../oai.cgi?verb=ListRec&set=wlhurq</cfURI>
         <cfClass>   data about semantic meaning
            <cfClassID>ID</cfClassID>
            <cfClassMeaning cfLangCode="en">Supervisor</cfClassMeaning>
         </cfClass>
      </cfClassification>
      <cfNote cfLangCode="en">comments to the semantic link-age</cfNote>
   </cfData>
</cfLinkage>
```

Notes for tags of the first group:

A: Inside the nested tags <cfClassification><cfClassSchemeId> with parameters of used semantic vocabulary there are tags <cfName> and <cfURI> with name and URI of the vocabulary, which is also embedded into linkage metadata to make it more autonomous;

B: Inside the nested tags <cfClassification><cfClassScheme> with parameters of selected semantic meaning additionally to ID of the selected meaning we embedded its name in tag <cfValue>;

C: <cfNote> allows scientists to provide explanations and comments to parameters specified by them for the semantic linkage.

The second group of XML tags:

```
<cfAuthorship>
   <cfProvider>
      <cfName cfLangCode="en">Socionet</cfName>
      <cfOrgUnitID>ID</cfOrgUnitID>
   </cfProvider>
   <cfOrganization>
      <cfName cfLangCode="en">CEMI RAS</cfName>
      <cfOrgUnitID>ID</cfOrgUnitID>
   </cfOrganization>
   <cfAuthor>
      <cfEmail>sparinov@gmail.com</cfEmail>
      <cfName cfLangCode="ru">Паринов Сергей Иванович</cfName>
      <cfName cfLangCode="en">Sergey Parinov</cfName>
      <cfPersID>ID</cfPersID>
   </cfAuthor>
```

```
</cfAuthorship>
```

This group of tags provides personal data about the linkage's creator, including e-mail address which will be used to notify the creator about a need to revise the link-age's correctness because of changes in linked objects. It also provides information about organizations involved with some roles into this semantic linkage creation.

The third group of XML tags:

```
<cfGeneral>
   <cfTemplateType>linkage</cfTemplateType>
   <cfTitle>source object ID + target object ID + semantic meaning</cfTitle>
   <cfID>semantic linkage ID</cfID>
   <cfDate>
      <cfStartDate>Timestamp</cfStartDate>
      <cfEndDate>Timestamp</cfEndDate>
      <cfRevision>Timestamp</cfRevision>
   </cfDate>
   <cfFraction>Float</cfFraction>
</cfGeneral>
```

The last group of tags provides information which typically used to build a table of contents and navigation across the whole created semantic linkages, and for some other processing of this data.

About data mandatory CERIF requires: *"Whereas the classification and classification scheme references are mandatory, the fraction attribute is not. Besides, each linking record requires a startdate and enddate."* (CERIF 1.3 FDM, p. 33). It should be appended by: information about authorship and data from General group of tags are also mandatory.

In CERIF: *"... the inherited identifiers and the date attributes build the primary key of link entities"* (Jorg et al. 2012a, p. 33). In our application this model of primary key building will not guarantee the unique identification of semantic linkages. For the same pair of objects scientists can create at the same time many semantic linkages with different authorship and/or with different types of relationship (take semantic meaning from different semantic vocabularies).

To have uniqueness of identifiers we propose to use the RePEc model of building object's ID (it is called here "Handle"). RePEc model is very useful when objects are created in decentralized mode by many people: *"The Handle: field content starts with the name of the authority (organization), for example RePEc. The next element is the code of the archive, then follows the code of the series and finally the number of the paper within the series. All these parts are separated by the colon character, i.e. :. Note that this field may not contain whitespace."* (Krichel 1997).

According this model the ID of a semantic linkage looks as a text string merged of 4 domains: "orgunit_code:archive_code:collection_code:object_code".

Additionally in RePEc for three domains "orgunit_code", "archive_code", and "collection_code" exist a template, which provides information about creators, edi-tors, etc. It is a convenient way to specify complex dependences between involved in this activity organizations, people and resources.

3.2 Semantic meaning

According CERIF the Semantic Layer - "*supplies the means for maintaining the CERIF Semantics: types, roles, terminology, subject classifiers, or mappings. It stores the semantic values that are carried by or referred to from the link entities via the cfClassSchemId attribute references, and it assigns each semantic value to a particu-lar classification scheme.*" (Jorg et al. 2012a, p. 33).

In our application an information object with the type "semantic meaning" exactly corresponds to cfClass. As well, a semantic vocabulary as a collection of semantic meanings representing different aspects of a specific type of research relationships corresponds with cfClassScheme. (Jorg et al. 2012b, p. 14; Jorg et al. 2012a, p. 37)

In XML notation the main attributes of these cfClass and cfClassScheme are de-fined as follow (CERIF 1.3 XML, p. 14):

```
<cfClass>
   <cfClassId>ID</cfClassId>
   <cfClassSchemeId>ID</cfClassSchemeId>
   <cfStartDate>Timestamp</cfStartDate>
   <cfEndDate>Timestamp</cfEndDate>
   <cfURI>String</cfURI>
</cfClass>
```

```
<cfClassScheme>
   <cfClassSchemeId>class-scheme-cerif-publication-types</cfClassSchemeId>
   <cfURI>http://www.eurocris.org/fileadmin/cerif-2008/CERIF2008_1.0_Semantics.pdf</cfURI>
</cfClassScheme>
```

Additionally, CERIF Semantic Layer allows for a representation of multilingual terms (cfClassTerm) and term descriptions (cfClassDescr). The two

class-type entities (cfClass, cfClassScheme) are interconnected with two recursive entities (cfClass_Class, cfClassScheme_ClassScheme) to allow for the representation of structures and for the mappings between classifications or classification schemes. The recursive entities of the CERIF Semantic Layer consistently support fractional values for classification references. (CERIF 1.3 FDM, p. 37)

CERIF 1.4 also provides a definition of the classification term (cfClassDef) and examples for the classification term (cfClassEx). These elements represent the lan-guage-dependent tables in the data model.

Following fragment of XML notation of a semantic vocabulary provided by CERIF Task Group can be used for our application almost without modifications:

```
<cfClassScheme>
   <cfClassSchemeId>307d6abf-13c2-40d7-ae1b-72609fde3145</cfClassSchemeId>
   <cfURI>CERIF-2008-1.2-cfOrganisationTypes</cfURI>
   <cfName cfLangCode="en" cfTrans="o">CERIF Organisation Types in Release 2008-1.2</cfName>
   <cfDescr cfLangCode="en" cfTrans="o">The CERIF Organisation Types - 2008-1.2 scheme provides the list of organisation types
according to the specification in the CERIF Semantics document. The organisation types are referred to from within the CERIF Link Entity
cfOrgUnit_Class.</cfDescr>
<!--Tags above give a description of a semantic vocabulary-->
<!--Tags below give a description of semantic meanings as elements of the vocabulary-->
   <cfClass>
      <cfClassId>229d5b5c-841d-4a5e-870c-c6c809f3a81d</cfClassId>
      <cfStartDate>2010-11-02T00:00:00</cfStartDate>
      <cfURI>CERIF-2008-1.2-cfOrganisationTypes#Academic_Institution</cfURI>
      <cfTerm cfLangCode="en" cfTrans="o">Academic Institution</cfTerm>
      <cfDescr cfLangCode="en" cfTrans="o">Academic institution is an educational institution dedicated to education and research, which
grants academic degrees.</cfDescr>
   </cfClass>
   <cfClass>
    ...
  </cfClass>
 ...
</cfClassScheme>
```

Information objects of this data type and a whole semantic vocabulary are created in our application by scientists or developers in decentralized mode. So all created objects of this type have to include a group of fields with "Authorship", which should be exactly the same as it described above for the semantic linkage data type.

And we also use RePEc model (Krichel 1997) to build ID of information objects instead of proposed in CERIF UUID model (Jorg et al. 2012a, p. 13 footnote).

4 Tools and Services

In this section we discuss two main groups of tools and services divided by its al-location: inside local RIS and outside it in a semantic fragment of a research e-infrastructure.

Following tools and services are working inside of local RIS: 1) a creation, editing and managing of both types of objects and their collections; 2) a submission of creat-ed objects to make them publicly available and moderation of submitted objects; 3) an output gateway to serve requests from an aggregator of central data storage; 4) API to send requests to the central data storage to get back existed ingoing and outgoing semantic linkages for specified information objects of local RIS; 5) API to synchronize a local set of semantic vocabularies with it current content at central data storage.

Outside of local RIS should work following services: 1) an aggregator to synchro-nize central data sets of semantic objects with content of local RIS-provides; 2) an output gateway to give away specified objects and/or any part of central data storage on requests from local RIS; 3) a navigation and searching tools over full content of central data storage; 4) a semantic interpretation service to build proper visualization for multilayer networks of relationships; 5) a monitoring service to trace changes in the central data sets and linked objects to build data for a notification service; 6) e-mail notifications for author of linkages and linked objects; 7) a statistic service to process central data sets and build various scientometric indicators.

4.1 RIS tools and services to deposit semantic objects

If RIS has functionality for scientists to deposit electronic publications, the same can be used for semantic objects. Typically CMS, as a tool for depositing papers at local RIS, can be configured to use additional templates for creating new types of information objects. The templates discussed above in the section "Information ob-jects" can be used by this way.

The deposited semantic objects will be available for public utilization only if it passed through some usual quality control routine. Collections of these objects after moderation should be available at the RIS output gateway for harvesting by a central aggregation service using one of popular protocols, e.g. OAI-PMH (OAI-PMH 2008).

Only after a RIS manager will register the output gateway at some open list of pro-viders (see the next section) and this information passed validation, the central aggre-gator starts everyday synchronization of the source with the central data storage.

4.2 A registry of semantic objects collections

A registry of semantic objects collections is an open catalogue of output gateways provided open access to collections of semantic linkages and semantic vocabularies. The registry can be organized by the same way as e.g. the Registry of Open Access Repositories (ROAR,

http://roar.eprinst.org/). A provider of semantic objects collec-tions (typically it is a research organization) fills in at the register an online form with parameters of the output gateways (gateway's description, URL and its protocols). This information is validated by the central data storage manager and if positive it is used by the central aggregation service to regularly synchronize a content of this pro-vider at central data storage with its sources at local RIS.

Each semantic objects provider receives at the registry a unique ID, which can be used to build unique identifiers for all semantic objects harvested to the central data storage.

4.3 Aggregation of semantic objects at central data storage

An aggregation service of the central data storage collects at one dataset all semantic linkages from diverse output gateways registered at the registry of semantic objects collections. Semantic vocabularies are also aggregated and stored at the cen-tral data storage to be available for using them by RIS tools for semantic linkages creation.

The central aggregator takes data from output gateways using all the most popular protocols. At least it should work as OAI-PMH harvester. Collections of semantic objects of both types at central data storage should be regularly (everyday) synchro-nized with its local sources.

For a situation when semantic objects IDs become not unique inside the central data storage, the provider's unique ID at the registry can be used to correct objects' ID. This correction can be made on a "fly" when semantic objects are recorded into the storage.

Central data storage also has an output gateway for giving away requested data using all popular protocols (at least OAI-PMH).

Altogether it should work just as a simple information hub. And so the central data storage has no special requirements for an authority, security, privacy and other non-functional properties.

Since in general any central data storage can be a bottleneck in a functioning sys-tem we suppose that this problem will be solved by mirroring or replication of its con-tent, e.g. as it was implemented in RePEc system (repec.org).

4.4 Data sharing and embedded software

Any local RIS is able to check presence at the central data storage already existed linkages for the local information objects. If positive, the linkages' data is transferred from the central data storage to the RIS. By this way the local RIS can visualize a network of linkages composed of articles and other information objects belong to this RIS.

Proposed application will provide some API for using within local RIS. This addi-tional software can be integrated into RIS: (1) to visualize already accumulated at the central data storage outgoing/ingoing semantic linkages for information objects of local RIS when a user is browsing over them; and 2) to update/synchronize local set of semantic vocabularies with it current content at the central data storage.

4.5 Interpretation, visualization and utilization

As an end-user interface to the central data storage we propose a service of semantic meanings interpretation. It can make a specific rearranging and visualizing of linked information objects by processing of its semantic. For example, information objects connected by semantic linkages with a meaning "components of a scientific composition" can be visualized as a networked document, or as a collection of scien-tific artifacts, or a table of contents.

This interpretation service provides data for specific visualization of multilayer networks of various types of relationships over integrated content of research DIS.

Additionally the central data storage should have a typical navigation and search-ing tools over full content accumulated semantic objects.

4.6 Monitoring of changes and notifications of users

Semantic linkages between a pair of research objects (e.g. between two articles) may lose their consistency if one or both of linked objects are revised by their au-thors. E.g. a meaning of the text fragment cited by a semantic linkage may be changed by an author of linked article, or this text fragment may disappear or move to another part of the linked article. In all such cases the author of the article that cited changeable text fragments must be informed to make reconsideration of related semantic linkages.

Scientists also can change already established semantic linkages, including: (a) a complete deletion of a linkage; (b) a redirecting of the linkage on another target ob-ject, since the new target object is better, e.g. it gives better illustration or evidence for a scientist's research output; (c) a changing of the current semantic meaning since the scientist changed his/her opinion on it.

The monitoring service has to register all such events. It stores necessary data to provide it for other services (a notification and scientometic services).

Initially designed monitoring service is processing only a flow of current changes in semantic linkages. But it can be developed to record a history of the evolution of thinking about the hypothetical relationship between DIS information objects. In this case the service also should store all previous states of semantic linkages.

A notification service uses data about changes in semantic linkages generated by the monitoring. Different types of notifications produced by this service support a scientific circulation/communication by distributing signals about semantic linkages creation/revision. To keep consistence of research DIS this service should notify:

1. the authors of objects linked by created or revised semantic linkage, just to inform them about this event, let them know about specified semantic and give them an ability to react on this event (e.g. to protest against specified semantic);

2. the author who is changing his/her object (e.g. article), if the object has linked (cit-ed) in other objects (articles), that by this action she/he can violate have estab-lished linkages and/or its semantic;

3. the authors of semantic linkages, if there were changes in objects specified as a source and a target of the linkages, so they should reconsider their linkages and, if it necessary, correct it;

4. the users of research DIS while they are viewing some DIS object (e.g. the readers of electronic articles) that certain semantic linkages made for the displaying source object (e.g. citations in reading text) can be violated because of the target object (e.g. cited articles) was changed, and an author of the linkages has not updated suspicious linkages (e.g. citations).

If the first three types of notifications in the list above can be made by e-mail only, the last one should work as warning, that displayed on the screen when it necessary.

Thus the notification service improves scientific circulation because it immediately informs scientists about using their research outputs. And it improves research com-munication because authors of semantic linkages can receive a feedback on their actions from authors of linked research objects.

4.7 Collecting statistics and producing scientometrics

Traditional statistical representations of changes in DIS scope and structure are well known and have had examples of good implementations (e.g. LogEc, MESUR, Socionet Stats, and other). If scientists start an intensive building of semantic linkage multilayer networks over research DIS objects it opens a new space for statistics development.

The monitoring service associated with the central data storage collects all availa-ble statistics about semantic linkages. It allows us to form a scientometric database both quantitative (number of linkages, etc.) and qualitative (semantic meanings) characteristics of scientific relationships.

Quantitative data about all accumulated semantic linkages at the central data storage includes different types of its structuring and aggregation. E.g. numbers of linkages (total and by types of scientific relationships) for selected objects, aggregated numbers of linkages for all objects of one author (total, by relationship types, by values of semantic vocabularies, etc.), and many others.

Qualitative data about relationship types accumulated at the central data storage is structured by semantic vocabularies (layers of semantic network) and then by meanings (a distribution by types). It also includes graphs of linkages with semantic meanings assigned to each edge of the graph, and so on.

This new scientometric data will give the community useful additional information for better research assessment of individual scientists and research organizations as well.

Some additional statistics about accumulated semantic linkages can be also pro-duced by adopting to this specific case the well-known PageRank (Google page rank) algorithm.

5 Conclusion

Proposed application of the enhanced semantic linkage technique provides to the scientific community obviously benefits. It is a new type of semantic interoperability: any scientists can “interact” with any available information object of research DIS by expressing in computer readable form his/her opinion on research relationships and so re-use DIS content in some new forms. In fact, it gives scientists a new dimension for personal scientific creativity. For the community it gives better visualization of research outputs usage, improves scientific life-cycle and research communication, and supports measurements and assessment of research activity.

On one side, new tools implemented semantic linkage technique give scientists a freedom to express their professional opinions and research hypotheses. On the other side, a set of services from semantic fragment of research e-infrastructure provides the scientific community with some new abilities.

The community is immediately notified about all new opinions/hypothesis ex-pressed by individual scientist. The community can react on individual activities by expressing opinions on its correctness (from positive to negative, including a block-ing). Each individual activity gets publicly available statistical portrait which accumu-lates both: (1) data about linkages and expressed opinions/hypotheses made by a scientist, and (2) data about reactions made by the community on activities of the scientist.

As a unified system it gives the community an ability to select opinions/hypothesis (semantic linkages) by their quality, to evaluate its impact and to reuse it in multiple forms.

6 References

1. Jorg B, Jeffery KG, Dvorak J, Houssos N, Asserson A, van Grootel G, Gartner R, Cox M, Rasmussen H, Vestdam T, Strijbosch, L, Clements A, Brasse V, Zendulkova D, Hollrigl, T, Valkovic L, Engfer, A, Jagerhorn M, Mahey M, Brennan N, Sicilia M-A, Ruiz-Rube I, Baker D, Evans K, Price A, Zielinski M (2012a): CERIF 1.3 Full Data Model (FDM): In-troduction and Specification. euroCRIS, 2012. http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.3/Specifications/CERIF1.3_FDM.pdf

2. Jorg, B.; Dvorak, J.; Vestdam T.; van Grootel, G.; Jeffery, K.G.; Clements, A.; (2012b): CERIF – 1.3 XML: Data Exchange Format Specification. euroCRIS, January 2012. http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.3/Specifications/CERIF1.3_XML.pdf

3. CERIF 1.3 Semantics: Research Vocabulary. CERIF Task Group, euroCRIS, 2012, http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.3/Specifications/CERIF1.3_Semantics.pdf

4. CERIF 1.3 Vocabulary. CERIF Task Group, euroCRIS, 2012, http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.3/Semantics/CERIF1.3_Vocabulary.xls

5. Groth P.; Gibson A.; Velterop J. (2010): The Anatomy of a Nano-publication. Information Services and Use, Volume: 30, Number: 1/2, 2010, http://iospress.metapress.com/content/ftkh21q50t521wm2/

6. Parinov S. (2010): The electronic library: using technology to measure and support Open Science. In proceedings of the World Library and

Information Congress: 76th IFLA General Conference and Assembly, 10-15 August 2010, Gothenburg, Sweden http://www.ifla.org/files/hq/papers/ifla76/155-parinov-en.pdf

7. Parinov S. (2012): Open Repository of Semantic Linkages. In proceedings of 11th Interna-tional Conference on Current Research Information Systems "e-Infrastructure for Research and Innovations" (CRIS-2012), Prague 6-9 2012, http://socionet.ru/publication.xml?h=repec:rus:mqijxk:29

8. Parinov, S.; Kogalovsky M. (2011): A technology for semantic structuring of scientific digital library content. In Proc. of the XIIIth All-Russian Scientific Conference RCDL'2011 "Digital libraries: Advanced Methods and Technologies, Digital Collections", Voronezh State University, October 19–22, 2011 pp. 94-103. (In Russian - http://socionet.ru/publication.xml?h=repec:rus:mqijxk:28)

9. Jorg, B.; Dvorak J.; Vestdam T. (2012c): Streamlining the CERIF XML Data Exchange Formats Towards CERIF 2.0. In proceedings of 11th International Conference on Current Research Information Systems "e-Infrastructure for Research and Innovations" (CRIS-2012), Prague 6-9 2012

10. Krichel, T. ReDIF version 1, working paper, 1997, http://socionet.ru/publication.xml?h=repec:rpc:rdfdoc:redif

11. OAI-PMH: The Open Archives Initiative Protocol for Metadata Harvesting (2008), http://www.openarchives.org/OAI/openarchivesprotocol.html

12. LogEc - Access Statistics for Participating RePEc Services, http://logec.repec.org/

13. MESUR: MEtrics from Scholarly Usage of Resources, http://www.mesur.org/MESUR.html

14. Socionet Stats (in Russian), http://www.socionet.ru/stats.xml

---

About authors

Sergey Parinov – Central Economics and Mathematics Institute of RAS e-mail: sparinov@gmail.com