Semantic Linkages in Research Information Systems as a New Data Source for Scientometric Studies

Sergey Parinov[1] and Mikhail Kogalovsky[2]

[1] *sparinov@gmail.com*
Central Economics and Mathematics Institute,
Russian Academy of Sciences
Nakhimovsky pr. 47, Moscow, 117418 (Russia)

[2] *kogalov@gmail.com*
Market Economy Institute,
Russian Academy of Sciences,
Nakhimovsky pr. 47, Moscow, 117418 (Russia)

Abstract

A growing number of research information systems use a semantic linkage technique to represent in explicit mode information about relationships between elements of its content. This practice is coming nowadays to a maturity when already existed data on semantically linked research objects and expressed by this scientific relationships can be recognized as a new data source for scientometric studies. Recent activities to provide scientists with tools for expressing in a form of semantic linkages their knowledge, hypotheses and opinions about relationships between available information objects also support this trend. The paper presents one of such activities performed within the Socionet research information system with a special focus on (a) taxonomy of scientific relationships, which can exist between research objects, especially between research outputs; and (b) a semantic segment of a research e-infrastructure that includes a semantic interoperability support, a monitoring of changes in linkages and linked objects, notifications and a new model of scientific communication, and at last - scientometric indicators built by processing of semantic linkages data. Based on knowledge what is a semantic linkage data and how it is stored in a research information system we propose an abstract computing model of a new data source. This model helps with better understanding what new indicators can be designed for scientometric studies. Using current semantic linkages data collected in Socionet we present some statistical experiments, including examples of indicators based on two data sets: (a) what objects are linked and (b) what scientific relationships (semantics) are expressed by the linkages.

JEL codes: C6, C8

1. Introduction

A growing number of modern research information systems (RIS) are using a semantic linkage technique to visualize for its users some information about relationships between entities form its content. This practice is coming nowadays to a maturity when a pool of already existed semantic linkages provided information on what research objects are semantically linked and what kind of relationships were expressed by this can be recognized as a new data source for scientometric studies.

Basically the semantic linkage technique is used to explicitly express all kinds of relationships that exist between information objects representing in RIS content people, organizations, research results, scientific assertions, etc. At abstract level this technique is specified in RDF (http://www.w3.org/RDF/) and used as a base by different teams of developers like Linking Open Data (LOD)[1], Open Annotation[2], and many others. Specifically for RIS the semantic linkage technique is also developed by the CERIF task group (Jörg et al. 2012a). Some of these applications have a convergence tendency, e.g. between CERIF and LOD by exposing CERIF-driven relational data as linked data (Joerg et al. 2012b).

[1] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
[2] http://www.w3.org/community/openannotation/

At conceptual level the RDF Semantics specification[3] underlies for well-known nano-publication approach (Groth et al. 2010), as well the CERIF semantic model is used in many projects; e.g. to build the Semantic Linkages Open Repository (Parinov 2012a).

There are several public RIS where the semantic linkage technique is available for scientists, e.g. at NanoPub.org (Groth et al. 2010), SiteULike.org (Shotton 2010), Socionet.ru (Parinov and Kogalovsky 2011; Parinov 2012a, 2012b) and some others.

One of these RIS - Socionet.ru – is designed to improve standard semantic web techniques (Parinov 2012a, 2012b) and to be compatible with the CERIF Semantics specification (Jörg et al. 2012a). Socionet serves the research community and gives scientists an ability to express and share with the community some additional research information by linking semantically pairs of information objects from content of available RIS.

For general purposes the similar idea is developed by Open Annotation Community Group as a common, RDF-based, specification for annotating digital resources[4].

Semantic linkages created at Socionet match well with the open annotation model and can be recognized as a specific case of the digital resources annotating. But the Socionet approach is more focused on developing of: (a) taxonomy of scientific relationships, which can exist between research objects, especially between research outputs; and (b) a semantic segment of a research e-infrastructure that includes a semantic interoperability support, a monitoring of changes in linkages and linked objects, notifications and a new model of communication for scientists, and at last – scientometric indicators built by processing of semantic linkages data.

Implementation of a semantic linkage technique in a form of a researcher’s personal tool within the Socionet RIS demonstrates also strong similarities with collaborative tagging services which are very popular among users of many social networks like Twitter, Flickr, etc. The collaborative tagging (also known as a folksonomy) is defined as a “method of allowing anyone (users) to link keywords or tags to content, at pleasure” (Dix et al. 2006). The semantic linkage technique can be defined as a method of allowing anyone (e.g. scientists as users of RIS) to link any pair of available research objects from RIS content. A semantic meaning of the linkage expresses scientist’s knowledge, hypothesis or an opinion about a scientific relationship between the linked objects. And scientists do it at pleasure, as their regular scientific creativity (Parinov 2012a, 2012b).

In contrast with typical collaborative tagging at Socionet the semantic linkage technique uses available taxonomy of scientific relationships that can exist between different types of research information objects (between person and organization profiles, projects, research outputs, etc.). Within the Socionet the taxonomy is represented as hierarchical structure of controlled semantic vocabularies.

Currently in the research community there are several initiatives provided proper ontologies and semantic vocabularies for scientific relationships classification. One of them is the Semantic Publishing and Referencing (SPAR) Ontologies (Shotton 2010b). Another one – CERIF Semantics (CERIF 1.3 Vocabulary, 2012) developed by euroCRIS[5] in collaboration with CASRAI[6] and other organizations. Ontologies like SWAN, SKOS and some elements of the Annotation Ontology[7] are also can be used to build a classification of scientific relationships.

While a substance of scientific relationships between a person and an organization are well known and more or less have been classified, the taxonomy of relationships between research

[3] http://www.w3.org/TR/2004/REC-rdf-mt-20040210/

[4] http://www.openannotation.org/

[5] http://www.eurocris.org/

[6] http://casrai.org/

[7] http://code.google.com/p/annotation-ontology/

outputs is still under development. SPAR ontologies include very useful the Citation Typing Ontology (CiTO) (Shotton 2010c), which allows us to build initial semantic vocabularies.

In addition some available studies give us empirically based corrections and improvements for initial classification of relationships between research outputs. It is a long-standing study of the rhetorical and argumentative characteristics of scientific discourse (Shum et al. 2010) and recent research on analysis of sentiments in the text surrounding references in scientific papers that conducted in a framework of the Maps of Science building (Small 2011).

The study of scientific discourse provided us with some classification of typical motivations for making citations (Shum et al. 2010). Information about the same motivations can be also extracted from the citation context by using special software that discussed in Shum et al. (2010).

The sentiments' extraction from the citation contexts about the cited research papers can be done by existed sentiment analysis tools which have been developed to build taxonomy over social networks data (Galassini et al. 2011).

Since the taxonomy should also be "live", at Socionet we provide scientists with a tool to update and develop scientific relationship classifications (semantic vocabularies) in decentralized, but to some extent controlled way (Parinov, 2012b).

Semantic linkages created by scientists between research objects of RIS content can be organized either as a specialized information system, e.g. an Open Repository of Semantic Linkages (Parinov 2012a) or this data can be stored just within a publication's metadata (Nanopub, SiteULike). In any way the data of semantic linkages accumulated and stored in RIS can be processed to provide for the scientific community some additional useful information e.g. like the "Semantic Halo" (Dix et al. 2006) initially designed for traditional social networks.

By processing semantic linkages data collected in RIS one can build at least two types of new scientometric indicators: 1) statistical aggregators of research objects properties (e.g. download statistics, number of citations , etc.) according information what objects are linked; 2) statistical distributions for scientific relationship classes associated with research objects from RIS content.

Using these two types of indicators at least following new scientometric studies can be done:

- quantitative studies of all accumulated semantic linkages including different sort of its structuring and aggregation, e.g. numbers of linkages (total and by scientific relationship classes) for specific objects (authors, organizations, etc.); aggregated numbers of linkages for all objects belonged to one author (total, by relationship classes, by values from semantic vocabularies, etc.); and many others;
- qualitative studies of relationships specified by scientists over all available research objects, including statistical distribution of expressed relationships for specific objects, graphs of linkages with semantic values assigned to each edge of the graph, and so on.

These new scientometric indicators can significantly improve a visualization of research outputs usage and impact. It gives the community useful additional information for better research assessment and evaluation of individual scientists and research organizations as well.

Scientomentics studies based on proposed approach, if the collected data is statistically significant, can e.g. answer: a) who used what research results as a basis for creating a new scientific knowledge (e.g. a "use method from" of a semantic linkage meaning that should have the highest positive assessment); b) whose research results prove/repeat or are proved by other results that can be assessed as an indicator of credibility; c) whose results are mentioned just as illustrations that should be assessed as a weak usage; d) whose results are criticized or disapproved that means assessment of suspicious research results with currently not clear research impact; and so on.

In the second section of this paper we provide basic information about a semantic linkage technique and its possible application within RIS. We discuss requirements to proper implementation of this technique at RIS which should allow a collecting of new data for scientometric studies. We also present in this section our approach to provide scientists with the semantics which necessary for expressing of scientific relationships over available research objects. It includes examples of basic semantic vocabularies that setting up taxonomy used also when we build the data sets for scientometric studies.

In the third section we propose an abstract model of the data source used to aggregate and process the semantic linkage data. We discuss possible scientometric indicators which can be built according this model.

In the fourth section we discuss ideas and provide some examples of scientometric indicators based on two data sets: (a) what objects are linked and (b) what scientific relationships (semantics) are expressed by the linkages. Most of these indicators have been generated at the moment by Socionet services and are available for users. Statistical significance of the indicators will be improved with a growth of users and amount of semantic linkages accumulated at Socionet.

In conclusion we summarize benefits of this new data source for scientometrics studies.

2. Semantic Linkages in Research Information Systems

Presently the semantic linkage technique is implemented in several RIS. In some way it works at NanoPub.org, in another - at CiteULike.org. At VIVO (vivoweb.org) it is implemented as semantic web platform that reveals research and scholarship through linked profiles of people and other research-related information. VOA3R (voa3r.eu) uses the semantics technology to deploy an advanced, community-focused integrated service for the retrieval of relevant open content and data. Below we describe how this technique is designed within Socionet RIS to make possible using of accumulated semantic linkages data for scientometric studies.

Semantic linkage data

A template to create semantic linkages in our RIS was designed as compatible to ~~a~~ specification of CERIF Link Entity (Jörg et al. 2012a, p. 33; Jörg et al. 2012b, p. 13). Making semantic linkages self-contained within RIS content we partly upgrade initial CERIF specification by adding more characteristics. So when a scientist creates a semantic linkage we get in RIS following data:

1. data about pair of linked objects, including a specification of the linkage orientation (which object is a source of the linkage and what is a target one), unique IDs of linked objects, its data types, titles and authors;
2. ID and a name of selected semantic meaning and also URI and a name of the parent semantic vocabulary;
3. comments that allows scientists to provide explanations and comments about specified semantic linkage parameters;
4. personal, organizational data about an author of the semantic linkage and about a provider of the service;
5. a title and unique ID of the linkage itself, creation and revision dates.

A title of a 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 attributes are changeable (e.g. items 2-3 above) and, in principle, have a status similar to electronic publication. So it explains why we require a revision date. In a case to have a history of changes the system can store dates and details of all revisions. The linkage's last revision date in combination with revision dates of linked objects give us important information about synchronization of changes in the triple: source object – linkage – target object.

Some additional details about semantic linkage specification can be found in (Parinov 2012a; Parinov and Kogalovsky 2011).

To be universal the semantic linkage technique should produce linkages having a status of regular information objects from RIS content and so they should exist separately from the metadata of linked information objects.

Semantic linkages in research information systems

The semantic linkages can be used as a data source for scientomentic studies if its implementation in RIS satisfies several important requirements.

Linkages with assigned semantic meaning should be created not only by RIS developers, but also directly by scientists or their assistants with explicit indication of who is an author of the linkage and responsible for semantically expressed knowledge, professional opinions or scientific hypothesis.

A technique of semantic linkages should work at standalone mode, i.e. independently of linked objects metadata, since in many cases the semantic linkage's attributes cannot be directly included into linked objects metadata.

Created semantic linkages should be deposited by their authors into RIS as a public information resource.

Semantic linkages are created in decentralized mode and its semantic can break some ethical norms (e.g. wrong accusation in plagiarism, etc.). Nevertheless any created semantic linkage includes data about its author, who cannot escape the responsibility, and there should exist a submission procedure, which implies moderation and some quality evaluation of semantic linkages before it will be publically available.

Since a set of relationship classes used for semantic linkage creation cannot be completely predefined, scientists should be able to expand in some controlled way semantic vocabularies of relationships.

Ideally any scientist should be able to establish any number of consistent and relevant scientific relationships in visual and computer readable form between a pair of any available research objects. And any scientist should have an opportunity to propose new classes of scientific relationships for covering by this technique.

But at the same time the scientific community should have some kind of quality control over submitted for a public use new semantic linkages and new classes of relationships.

Socionet.ru almost completely satisfies these requirements. Below in the section "Examples of Scientometric Indicators" we provide some details about this RIS.

Semantic vocabularies

A template for creating semantic meanings as units of a semantic vocabulary in our RIS based on cfClass specification from the CERIF Semantic Layer (Jörg et al. 2012a). As well, a semantic vocabulary as a collection of semantic meanings representing different aspects of a specific class of research relationships corresponds with cfClassScheme (Jörg et al. 2012b, p. 14; Jörg et al. 2012a, p. 37).

Initial set of rendered scientific relationship classes has been built from different already existed ontologies (Parinov and Kogalovsky, 2011; Parinov 2012a) includes: (1) relationships between research outputs like inference, usage, impact, comparison, evaluation, 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, organizations} on the other.

Since a semantic linkage expresses a relationship between *two* objects, we should determine which scientific relationship classes (semantic vocabularies) applicable for each combination of pairs from a list research objects' types: a *source* object type {"person", "organization", "research output", "project", etc.} -> a *target* object type {"person", "organization", "research output", "project", etc.}.

In (Parinov and Kogalovsky 2011) we proposed initial classes of scientific relationships and a set of semantic vocabularies. For the pair of object types "research output" -> "research output" following classes of scientific relationships and associated semantic vocabularies were specified (ontologies used as a source for semantic vocabularies are mentioned below in brackets):

- Type "Inference", initial semantic vocabulary (CiTO): "obtain background from", "updates", "used as evidence", "confirms", "qualifies", etc.;
- Type "Impact/usage", initial semantic vocabulary (CiTO): "contains assertion from", "uses data from", "uses method from", "corrects", "refutes", etc.;
- Type "Hierarchical and associative relationships", initial semantic vocabulary (SKOS, SWAN): "broader", "narrower", "related", "alternative to", etc.;
- Type "Components of scientific composition", initial semantic vocabulary (DoCo): "duplicate", "revised", etc.

Additionally we made up a relationship class "Usage proposal" which is also valid for pair of data types "research output" -> "research output" and has initial semantic vocabulary: "can improve", "can illustrate", "can replace", etc. Using it scientists can share with the community their ideas on what research outputs can be used to improve/develop some other research outputs.

For the pair of types "person" -> "research output" there is a class "Professional opinions" with initial semantic vocabulary (SWAN): "responds negatively to", "responds positively to", "responds neutrally to", etc. Using this class of semantic linkages a scientist can, e.g. protest (the value "responds negatively to") against wrong opinions expressed by other scientists with their semantic linkages.

And there are also relationship classes and semantic vocabularies for some other pairs of objects:

- "person" -> "organization", a class "Person-Organization" relationships, initial semantic vocabulary (CERIF semantic vocabulary): "employee", "head", "member", "director", etc.;
- "person" -> "person", a class "Person-Person", initial semantic vocabulary (CERIF semantic vocabulary): "manager", "supervisor", "mentor", etc.;
- "person" -> "research output", a class "Person-Research Output", initial semantic vocabulary (CERIF semantic vocabulary): "author", "editor", "reviewer", "translator", etc.;
- "organization" -> "research output", a class "Organization-Research Output", initial semantic vocabulary (CERIF semantic vocabulary): "intellectual property rights claim", "publisher", "organizational author", etc.

These classes of scientific relationships and initial collections of associated semantic vocabularies can be used as taxonomy for scientometric studies if scientific community recognized them as proper characteristics. It can be achieved if the relationship classes and semantic vocabularies are opened for decentralized development, i.e. scientists can propose new relationship classes and/or semantic vocabularies for public use and there is a competition between different semantic vocabularies to be used by the community for making semantic linkages (Parinov, 2012a, 2012b).

Scientific relationships taxonomy

Scientific relationships taxonomy defines classes and subclasses of relationships which can exist over a set of research objects available for scientists in RIS content. Research objects from RIS content belong to one of scientific entity types: "person", "organization", "research output", "project", etc. (see a more complete list of types, e.g. in CERIF Semantic vocabulary). All research objects of the same type have identical classes of possible scientific relationships with other research objects.

Since a semantic linkage supports a *binary* relationship between two research objects we have to regularize classes of relationships according their applicability to possible variation in pairs of research objects types: a type of source object – a type of target object. We present this regularization as a two-dimensional matrix. Columns and rows of the matrix correspond to a complete list of scientific entity types. A cell of this matrix contains names one or more semantic vocabularies, which classify possible scientific relationships for the pair of research object types from the column and the row.

Each cell of the matrix contains at least one semantic vocabulary. So the proposed scientific relationships taxonomy covers all existed research objects and whole RIS content.

Some cells contain more than one semantic vocabulary. E.g. for the pair of types "research output" -> "research output" currently there are 5 classes of possible scientific relationships. It means that for the same pair of research outputs a scientist can create up to 5 linkages with semantic meanings from different classes.

3. An Abstract Computing Model of a New Data Source

Based on introduced above entities like a semantic linkage, a semantic meaning, a semantic vocabulary and a scientific relationships taxonomy we can define at more or less abstract level what data (statistics) can be extracted from semantic linkages accumulated in RIS. By this we set up an abstract model of this new data source. The model can be useful for understanding in general what scientometric indicators can be designed using this data source.

For any research object *A* taken from RIS content the scientific relationships taxonomy allows defining of two sets of research objects types: 1) what objects can be the target for outgoing linkages of the object *A*; and 2) what objects can be the source for ingoing linkages to the object *A*.

So for any research object from RIS content there are sets of types of the *target* and the *source* objects. For any object type from these sets the matrix of applicable relationship classes (defined above) provides relevant semantic vocabularies, where each vocabulary contains a set of values (semantic meanings) specified possible diversity of a corresponded scientific relationship.

A scientific relationship presented in RIS by some semantic vocabulary has following characteristics: 1) a class defined by a title of corresponded semantic vocabulary and optionally by a description of the vocabulary; and 2) a specific meaning as an item from the vocabulary with optional additional description of the item. These characteristics are stored within a description of semantic linkage data together with URIs of both the source and the target objects.

By processing semantic linkages accumulated in RIS we can compute for a selected research object *A*: 1) a set of the *target* research objects distinguished by scientific relationship classes and then by specific meanings (relationship subclasses) of corresponded semantic vocabularies; 2) a set of the *source* research objects distinguished by the same as in the first case. In some cases these sets can be empty.

Each research object from the *target* or *source* set can have own sets the target and the source objects, and so on. It is a consequence from a statement that all research objects can have

semantic linkages. It means that for the object *A* we can compute scientometric indicators by processing data of semantic linkages for any number of remote levels relatively the object *A*.
According this abstract computing model we can generate for any research object *A* at least two types of characteristics: 1) indicators based on information about linked objects and 2) indicators based on information about expressed semantic meanings and associated relationship classes.

Indicators computed by using information about linked research objects

If the object *A* linked by some research relationships with some set of other research objects, then some properties of the linked objects can be transferred to the object *A*. The class of research relationships in this case determines an interpretation of properties attached to the object *A*. The properties' transfer should be made according linkages' direction and for any necessary number of remote levels.
For instance, research object *A* is scientist's personal information (profile). It has semantic linkages with research objects like papers, articles and other forms of his/her research output. A semantic meaning of each such linkage specifies an authorship of this scientist for these linked research objects. Research objects with the type "research output" typically have in RIS a characteristic like number of views/downloads, number of citations and so on. These statistical data can be aggregated over a set of objects linked to the same author's profile and attached to other properties of this author. Interpretation of attached properties is obvious in this particular case: aggregated number of views/downloads characterize a total level of the community demand to the author's research outputs; and aggregated number of citations (excluding self-citations) is an indicator of the community response on author's research outputs.
As an illustration of a properties' transfer for more than one remote level the described above aggregated properties of personal profiles can be attached to properties of an organization profile which presents a workplace of these authors. In this case we have two levels of aggregation of properties: 1) according semantic linkages between research outputs and personal profiles of its authors; and then 2) according semantic linkages (with the meaning – "workplace") between personal profiles and the organization profile. As a result an organization profile gets characteristics like a total community demand for research outputs (views/downloads aggregated statistics) produced by scientists from staff of this organization, and total indicator of community respond (total number of citations) on research outputs produced in this organization.
In the next section there are examples of such indicators based on real data.

Indicators computed by using semantic meanings and research relationship classes

Sets of relationships classes and/or semantic meanings (subclasses) of all collected in RIS semantic linkages can be ordered by its frequency in a form of a statistical distribution. Each single semantic meaning in this statistical distribution has following characteristics: 1) a corresponding semantic vocabulary (and associated class of a scientific relationship); 2) a pair of linked research objects (the *source* and the *target*), including data about its authors, scientific area, discipline, and other typical attributes if the linked objects belong to "research output" type; 3) a scientist who created the linkage and responsible for its semantic meaning; and some other. These characteristics grouped for the *source* or the *target* linked research objects and then can be processed to build different scientometric indicators.
Statistical distributions built by this way can have interpretation, e.g. as characteristics of a popularity of some scientific relationship classes and specific meanings expressed by research community over RIS content. If this statistical distribution is built e.g. for a research article it

can characterize what class of scientific relationships between this article with the others currently dominates.

Such statistical distributions computed for a research object *A* (e.g. it is a research output) can be associated in some way with the object's properties to be transferred according semantic linkages of the object *A* with other objects and can be attached to properties of linked objects. For example, a distribution of scientific relationship classes and subclasses expressed for some set of papers, articles, etc. can be attached to the personal profile of these research outputs author.

The next section provides some illustrations of such statistical distributions.

4. Examples of Scientometric Indicators

Socionet as RIS produced, harvested and accumulated semantic linkages

The Socionet system development was started in 1997 as a project to design a Russian Virtual Laboratory for Economists and Sociologists. At the beginning it provided a mirror of RePEc.org data and functionality. It also included the first in Russia scientific open archive to submit research papers in Social Sciences for its online presenting, and some simple tools of virtual workspace (Krichel and Parinov 2002). In 2000 this information system got its current name "Socionet" (socionet.ru). Since from that time it has own harvester, which federates more research collections and archives, than RePEc provided (Parinov et al. 2003). It allowed a building and, from that time, an everyday updating the Russian research data and information space (DIS) initially for Social Sciences only.

In 2002 a Socionet Personal Zone service was created as add-in online workbench. It allowed a management of information object's collections for 9 object types ("person", "institution", "paper", "article", "chapter", "book", etc.). The Personal Zone service also included software of the "personal information robot" to trace new additions/changes within DIS according personal research interests of users and notify them about relevant findings (Parinov and Krichel 2004).

In 2004 Socionet users got some new tools to create and manage semantic linkages between information objects of DIS (Parinov and Krichel, 2004).

In 2007 we ran a service to monitor permanently all significant changes in DIS information objects including semantic linkages. It allowed a collecting of different statistics about DIS current state and changes. The Socionet scientometric database started accumulating of its data from 2007.01.01 (Kogalovsky and Parinov 2008, 2009).

Currently the Socionet is working as a public multidiscipline RIS based on Open Science ideas (Parinov 2009, 2010b). Socionet tends to be a full-functional modern CRIS driven by the community of Russian writing scientists (Parinov 2010a).

On July of 2013 the Socionet system federates more than 4000 collections with scientific materials organized by RePEc.org initiative and about 500 collections mostly from Russian research organizations and academic publishers. In total it is about 2M research objects and with every day average surplus of 300 new objects and 1-2 new collections per week. It covers 15 scientific disciplines organized by 16 object types sections.

In addition there are about 6M semantic linkages over research objects of Socionet DIS, which came to the Socionet by different ways. The biggest part of them - about 5M – is harvested with regular updates from the CitEc.repec.org system (Barrueco and Krichel 2005). All of these linkages have a semantic meaning just as the "citation", but authors of these citation linkages can semantically enrich them by using Socionet tools and semantic vocabularies presented above. Another part - about 700 thousands of semantic linkages - Socionet received with RePEc collections of research objects. Linkages here have obvious semantics associated with personal and organization profiles. And the last part of semantic linkages was created by scientists inside the Socionet Personal Zone.

If a publication metadata includes a code(s) of some scientific classification we also recognize it as a semantic linkage(s) between the publication and the scientific classification system. This class of semantic linkages means a relation of a publication to certain scientific area(s), mapped by a scientific classification system.

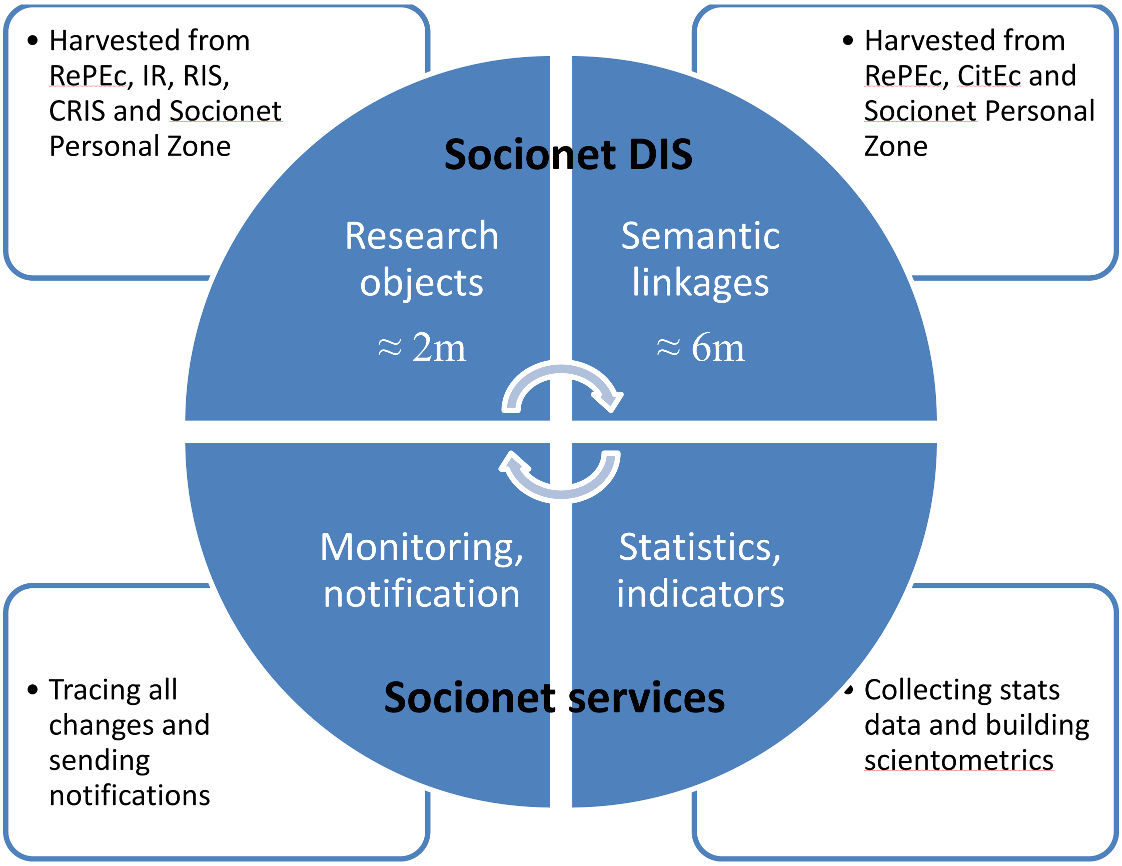
Harvested from RePEc, IR, RIS, CRIS and Socionet Personal Zone
Harvested from RePEc, CitEc and Socionet Personal Zone
Socionet DIS
Research objects
≈ 2m
Semantic linkages
≈ 6m
Monitoring, notification
Statistics, indicators
Socionet services
Tracing all changes and sending notifications
Collecting stats data and building scientometrics

Figure 1: Socionet main subsystems

Indicators based on information about linked objects

The first part of the abstract computing model defined above in the Section 3 is in operation as a Socionet Statistics subsystem from 2007. Using information on what research objects are linked we built aggregated scientometric indicators for personal profiles, organization profiles, scientific classification systems, and some others. Initially aggregated data contained only views/downloads statistics processed according LogEc specifications (Karlsson 2011). Some other aggregated properties (number of citations, statistical distributions of semantic meanings and classes, etc.) were added recently and under testing now.

(1) *For all personal profiles* in Socionet DIS (research objects with type "person", total number is about 50 thousands profiles) we compute statistics of: (a) views/downloads aggregated over a set of research outputs (paper, article, book, chapter, etc.) semantically linked to the personal profiles of its author; and (b) total amounts of outgoing/ingoing semantic linkages existed for the same set of research outputs. In the table 1 there is an example of such aggregators for the top three authors with the highest number of ingoing linkages.

Table 1: An example of aggregated data for personal profiles

Objects	**Linkages * (outgoing/ingoing)**	**Statistics ** (downloads/views)**
Shleifer, Andrei	3534 / 35102	10543 / 42776
Barro, Robert J	1497 / 24016	9732 / 30724
Heckman, James J	4339 / 23017	10874 / 35878

*) number of linkages counted by Socionet services including CitEc.repec.org data on July 2013
**) number of downloads/views counted by LogEc.repec.org on July 2013 over the past 12 months

Specifically in this example the most of ingoing linkages are citations of materials authored by people listed in the table. Socionet statistic subsystem allows viewing of a structure for all aggregators. E.g. the aggregator of linkages for Andrei Shleifer personal profile has following structure (counted by Socionet services using full data set of linkages on 2013-07-17):

1. Outgoing linkages - 3534 (9% of total number of linkages for the personal profile)
    1. 1. direct outgoing linkages from the personal profile - 377 (11% of total outgoing linkages)
        1.1.1. from the personal profile to organization’s profiles – 2 (1% of total direct linkages)
        1.1.2. from the personal profile to materials - 375 (99% of total direct linkages)
    1.2. secondary outgoing linkages from the objects linked with the personal profile - 3208 (89% of total outgoing linkages)
        1.2.1. from materials authored by the person (number of citations made by the person) - 3157 (100% of total secondary outgoing linkages)
2. Ingoing linkages – 35102, (91% of total number of linkages for the personal profile)
    2.1. direct ingoing linkages - 0 (0%)
    2.2. secondary ingoing linkages to the objects linked with the personal profile - 35102 (100% of total ingoing linkages)
        2.2.1. to materials authored by the person (number of citations of the person’s materials) - 35102 (100% of total secondary ingoing linkages)

For statistics of downloads/views there are also a lot of options to change a time interval, to switch to a structural or dynamic statistical representation, to see detailed statistics for each single author including a statistical distribution of author’s research outputs by a frequency of its views/downloads, etc.

(2) *For all organization profiles* in Socionet DIS (research objects of the type “institution”, total number is about 12500 profiles) the system computes the same two types of aggregators: outgoing/ingoing linkages and downloads/views. In the table 2 listed the top three organizations with the highest number of ingoing linkages.

Table 2: An example of aggregated data for profiles of organizations

Objects	**Linkages *** **(outgoing/ingoing)**	**Statistics **** **(downloads/views)**
National Bureau of Economic Research (NBER)	366960 / 960708	7386 / 26593
Centre for Economic Policy Research (CEPR)	271126 / 374956	6384 / 18848
Institute for the Study of Labor (IZA)	373266 / 332458	7287 / 28091

*) number of linkages counted by Socionet services on July 2013

**) number of downloads/views counted for 2007.01.01 – 2013.07.15, but only at one RePEc node Socionet

The statistics of linkages and views/downloads are aggregated independently by two classes of relationships: a) over personal profiles linked with the semantic "workplace" to the organization profiles; and b) over collections belonged to this organization. The same as in previous case there are a lot of additional options to change parameters and have different details.

(3) *For all codes of scientific classification systems* (research objects with type "scheme", total number is about 2700 codes) included into metadata of Socionet DIS research objects the system computes statistics of views/downloads aggregated over research objects marked by the same classification codes. Additionally a parent code property aggregates statistics of all child codes with lower hierarchy. In the table 3 there is an example of daily aggregated number of downloads and views for objects marked by codes JEL:g including codes with lower hierarchy.

Table 3: An example of aggregated data for some codes of JEL classification system

Objects	**Downloads***	**Views***
JEL:g - Financial Economics	38	116
JEL:g2 - Financial Institutions and Services	12	25
JEL:g28 - Government Policy and Regulation	5	6
JEL:g21 - Banks; Depository Institutions; Micro Finance Institutions; Mortgages	3	9
JEL:g3 - Corporate Finance and Governance	11	48
JEL:g32 - Financing Policy; Financial Risk and Risk Management; Capital and Ownership Structure; Value of Firms; Goodwill	2	14
JEL:g38 - Government Policy and Regulation	2	3

*) number of downloads/views counted only at one RePEc node - Socionet for one day 2013.07.15

One can see daily list of classification codes (http://socionet.ru/stat4.xml?rubric=jel&obj=class&l=en) ordered by aggregated number of downloads with the same additional options as described for the first indicator.

Indicators based on semantic meanings data

In contrast with the previous sort of indicators in this section we discuss mostly the design of new indicators, since at the moment (July 2013) we still have no in Socionet statistically significant amount of semantic data. All examples of scientometric indicators on Figures 2-4 below are for illustration purpose only. It uses real taxonomy of scientific relationships, but all numbers are not real.

According the second part of our data source abstract computing model we count a frequency of individual classes and subclasses of scientific relationships taxonomy in a form of statistical distributions. Since our taxonomy has two hierarchical levels: 1) classes of scientific relationship (semantic vocabularies); and 2) its subclasses (semantic meanings) the statistical distributions can be built for only one or for both levels of the taxonomy and for different subsets of semantic linkages data.
The Figure 2 gives an example of overall statistical distribution of scientific relationship classes expressed by scientists in making semantic linkages between research outputs. It is a macro picture of dominated scientific relationships ("research usage" on the top) over whole Socionet DIS. This type of indicators has some similarities with the "Semantic Halo" made for traditional social networks (Dix et al. 2006).
Such "semantic cloud" in combination with data about groups of research objects linked by certain scientific relationships makes possible a multilayer stratification of Socionet DIS. By this way one can map scientific areas, disciplines, specific objects, groups of authors, etc., who are "producers" of certain scientific relationships, or which the specified relationships are established with.

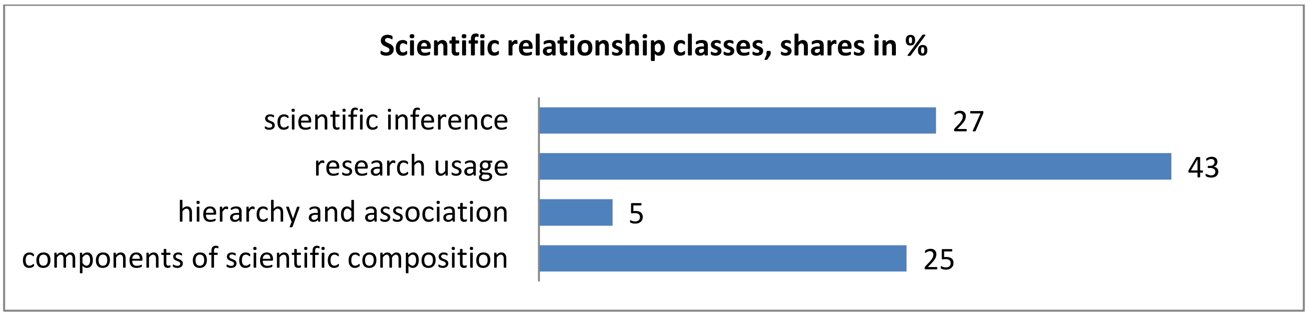

Figure 2: An example of overall statistical distribution for all expressed relationship classes

For any specific class of scientific relationships we can count a frequency of using its subclasses. The Figure 3 illustrates this for the "scientific inference" relationship class, where the most popular expressed by scientists a type of the inference is the "updates".

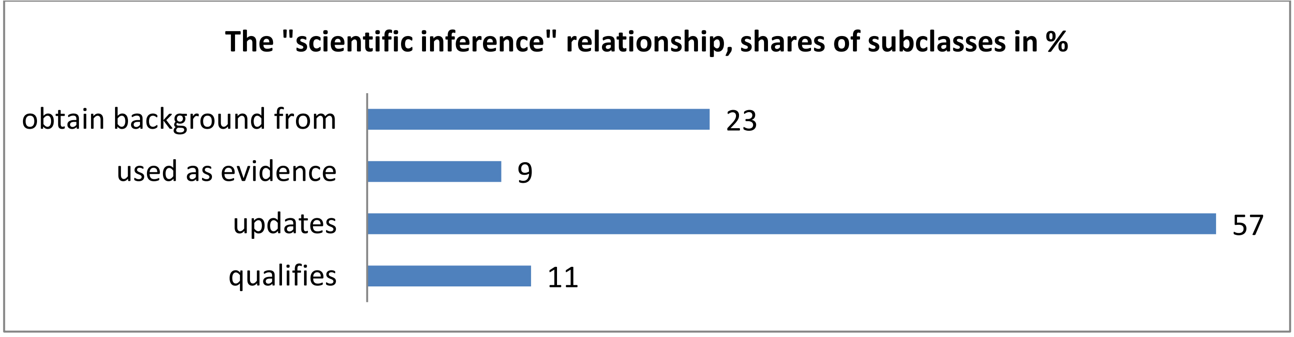

Figure 3: An example of a statistical distribution within the "scientific inference" class

Making semantic linkages scientists can also express their professional opinions about available research outputs. In this case they create linkages between their personal profiles and research outputs. The Figure 4 illustrates a distribution of expressed professional opinions, where the most popular sentiment is the "responds neutrally to".

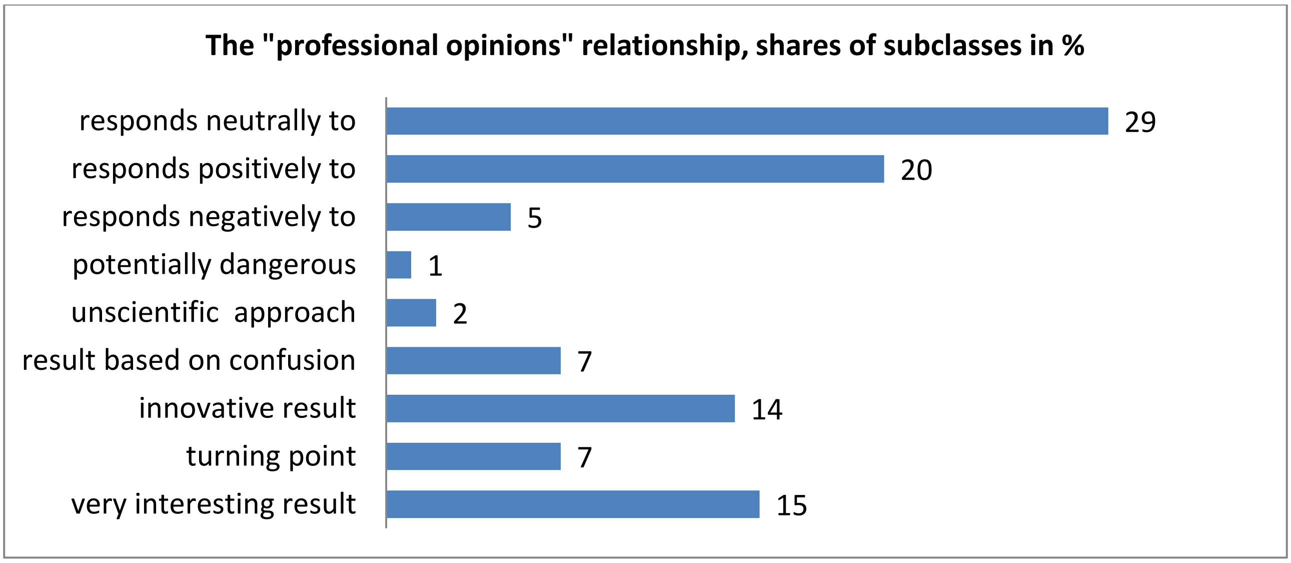

Figure 4: An example of sentiment distribution by the "professional opinions" class

As described in previous section the semantic data can be also aggregated by using information about linked objects. Such aggregators can characterize different objects with variation in selected relationship classes or subclasses. It can present e.g. accumulated opinions about research outputs of one author, or a distribution of sentiments expressed by one scientist, or the same aggregators for a research organization, a scientific journal, an academic publisher, and so on.

Handling some specific classes of relationships we can make studies for selected groups of research outputs, authors, or scientific disciplines: which research outputs is a background for scientific inference of another research result, what results are claimed to be a theoretical generalisation of another, and many other according our taxonomy of relationships.

5. Conclusion

A semantic linkage technique implemented in RIS opens for scientists a new dimension for scientific creativity. Monitoring and processing of created semantic linkages within RIS content establish a data source for new scientometric studies, which gives the research community different benefits and new opportunities.

Additionally to already existed scientometric indicators the proposed data source allows a deep inspection of impact/usage characteristics for a scientist and a research organization in both static and dynamic aspects.

Obtained in this way, assessments of research productivity and performance will be substantially more informative than the currently used citation indexes and characteristics of publication activity. Thus, using this new opportunities in procedures of professional certification and in research funding schemas will create more effective incentives and motivations for scientists.

Acknowledgments

We are grateful to Keith Jeffery and anonym reviewers for useful comments. The paper presents some results of a research project "A technology of decentralized interactive semantic structuring of digital libraries contents" funded by Russian Foundation for Basic Research, grant 12-07-00518-A.

References

Barrueco, J. M., Krichel, T. (2005). Building an autonomous citation index for GL: RePEc, the Economics working papers case. *The Grey Journal*, 1(2), 91-97.

Dix, A., Levialdi, S., & Malizia, A. (2006). Semantic halo for collaboration tagging systems. *In the Social Navigation and Community-Based Adaptation Technologies* Workshop-June 20th

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

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

Galassini, C., Malizia, A., & Bellucci, A. (2011). An approach for developing intelligent systems for sentiment analysis over social networks. *Intelligent Systems and Control / 742: Computational Bioscience*, J.F. Whidborne, P. Willis, G. Montana, Eds. Cambridge, United Kingdom, July 11 – 13, 2011.

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

Jörg, B., Jeffery, K.G., 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., Höllrigl, T., Valkovic, L., Engfer, A., Jägerhorn, 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): Introduction and Specification. *euroCRIS*, http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.3/Specifications/CERIF1.3_FDM.pdf

Joerg, B., Ruiz-Rube I., Sicilia M., DVOŘÁK J., Jeffery K., Hoellrigl T., Rasmussen H. S., Engfer A., Vestdam T., and Barriocanal E.G. (2012b). Connecting Closed World Research Information Systems through the Linked Open Data Web. *International Journal of Software Engineering and Knowledge Engineering* 22, no. 03 (2012): 345-364.

Karlsson, S. (2011). About LogEc, http://logec.repec.org/about.htm

Kogalovsky, M.; Parinov, S. (2008). Metrics of online information spaces. *In Economics and Mathematical Methods*, v.44, no. 2, 2008 (in Russian), authors' version - http://socionet.ru/publication.xml?h=repec:rus:mqijxk:17

Kogalovsky, M.; Parinov, S. (2009). Scientometrics by using a citation type of linkages in Socionet system. *Deposited by authors at Socionet* (in Russian), http://socionet.ru/publication.xml?h=repec:rus:rssalc:web-32

Krichel, T.; Parinov, S. (2002). The RePEc database and its Russian partner Socionet. *In Russian Digital Libraries Journal* vol. 5, no. 2, 2002, http://www.elbib.ru/index.phtml?page=elbib/eng/journal/2002/part2/KP

Parinov, S.; Lyapunov V.; Puzyrev R. (2003). Socionet system as a platform for developing information resources and online services for researchers. *In Russian Digital Libraries Journal* vol. 1, no. 6, 2003 (in Russian) http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2003/part1/PLP

Parinov, S.; Krichel, T. (2004). RePEc and Socionet as partners in a changing digital library environment, 1997 to 2004 and beyond. *In Russian Conference on Digital Libraries*, Pushchino, Russia, http://eprints.rclis.org/archive/00001830/01/bonn.pdf

Parinov, S. (2006). Information Hubs. *Deposited by author at Socionet* (in Russian), http://socionet.ru/publication.xml?h=repec:rus:mqijxk:9

Parinov S. (2007). e-Science: Online Future of the Science. *In Information Technology*, No. 9, 2007 (in Russian), author's version - http://socionet.ru/publication.xml?h=repec:rus:mqijxk:19

Parinov, S. (2009). Electronic libraries development is a way to Open Science. *In Proceedings of the XI All-Russian Research Conference "RCDL2009"*, Petrozavodsk, Russia (in Russian), author's version - http://socionet.ru/publication.xml?h=repec:rus:mqijxk:21

Parinov, S. (2010a). CRIS driven by research community: benefits and perspectives. *In proceedings of the 10th International Conference on Current Research Information Systems*. Aalborg University, Denmark, June 2-5, 2010, pp. 119-130. http://socionet.ru/publication.xml?h=repec:rus:mqijxk:23

Parinov, S. (2010b). 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*, Gothenburg, Sweden, August 10-15 (2010), http://www.ifla.org/files/hq/papers/ifla76/155-parinov-en.pdf

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*, October 19-22, pp. 94–103. Voronezh State University (2011) (in Russian), http://socionet.ru/publication.xml?h=repec:rus:mqijxk:28

Parinov, S. (2012a). Open Repository of Semantic Linkages. *In: Proceedings of 11th International Conference on Current Research Information Systems e-Infrastructure for Research and Innovations (CRIS 2012)*, Prague (2012), http://socionet.ru/publication.xml?h=repec:rus:mqijxk:29

Parinov, S. (2012b). Towards a Semantic Segment of a Research e-Infrastructure: necessary information objects, tools and services. *Metadata and Semantics Research, Communications in Computer and Information Science*. J. M. Dodero, M. Palomo-Duarte, P. Karampiperis, Eds. *Springer*, vol. 343, pp. 133-145. http://socionet.ru/pub.xml?h=RePEc:rus:mqijxk:30

Shotton, D. (2010a). Use of CiTO in CiteULike, http://opencitations.wordpress.com/2010/10/21/use-of-cito-in-citeulike/

Shotton, D. (2010b). Introduction the Semantic Publishing and Referencing (SPAR) Ontologies. October 14, 2010. http://opencitations.wordpress.com/2010/10/14/introducing-the-semantic-publishing-and-referencing-spar-ontologies/

Shotton, D. (2010c). CiTO, the Citation Typing Ontology. *Journal of Biomedical Semantics* 2010, 1(Suppl 1): S6. http://www.jbiomedsem.com/content/1/S1/S6

Shotton, D., Peroni, S. (2011). DoCO, the Document Components Ontology. 17/02/2011. http://purl.org/spar/doco/

SKOS - Simple Knowledge Organization System. http://www.w3.org/TR/skos-reference/

Small H. (2011). Interpreting maps of science using citation context sentiments: a preliminary investigation. *Scientometrics, Springer Netherlands*, Volume 87, Issue 2, pp 373-388

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

Shum, S.B., Clark T., de Waard A., Groza T., Handschuh S., Sandor A. (2010). Scientific Discourse on the Semantic Web: A Survey of Models and Enabling Technologies. *Semantic Web Journal: Interoperability, Usability, Applicability*. Special Issue on Survey Articles, Ed. Pascal Hitzler, http://www.semantic-web-journal.net/content/scientific-discourse-semantic-web-survey-models-and-enabling-technologies

SWAN (Semantic Web Applications in Neuromedicine) - Scientific Discourse Relationships Ontology Specification. http://swan.mindinformatics.org/spec/1.2/discourserelationships.html