Open Repository of Semantic Linkages

Sergey Parinov

Central Economics and Mathematics Institute, Russian Academy of Sciences, Russia

Summary

The growing use of the CRIS-CERIF concept in combination with the natural development of research information systems de facto leads to a certain standardization of the systems, e.g. concerning content, a set and a structure of information objects that they are operating with. Modern research information systems operate with virtually the same set of objects: "person," "organization," "project," "research output/outcome", "event" and some others. This allows us: (a) to consider the public information objects maintained within independent local CRIS, as objects of a common research Data and Information Space (DIS), and (b) to create tools and services for the re-use of these information objects in some new ways. Proposed open repository of semantic linkages enables scientists to express in a computer-readable form their opinions about multiple scientific relationships that can exist between research objects. Technically it means an establishing of a multilayer network of semantic linkages over all DIS content. Local CRIS can use some API associated with the repository to visualize semantic linkages for the objects from their content. Semantic vocabularies, used when scientists are creating linkages, are open to the completion and development by the community. The repository is supplemented by some services like monitoring, notifications and scientometrics. Using this semantic linkages open repository (SLOR) the scientific community can get multiple benefits and makes one more step to a universal interoperability on the CRIS-CERIF model base.

1 Introduction

Modern research information systems operate with virtually the same set of objects: "persons", "organizations", "projects", "research outputs/outcomes", "events" and some other. This allows us to consider the information objects that exist within a content of separate local CRIS, as having some homogeneous and belonging to common research Data and Information Space (DIS). Based on this we can build a shared and collaborative repository infrastructure and services for the re-use of these information objects in some new ways.

One of opportunities to operate with information objects of such virtual research DIS is a public tool for researchers to visualize multiple scientific relationships between DIS objects in a computer-readable form of semantic linkages.

It is a well-known fact about scientific work that researchers create different types of relationship between research outputs, outcomes and other objects of scientific DIS. Some of these relationships are visible (e.g. citations), most of them are not observable and exist in a mental form only. Research community will benefit, if these relationships will be visible including its specific characteristics, e.g. like a scientific inference dependence and concrete values of impact/usage, that can exist for research outputs.

If scientists have an ability to establish and visualize scientific relationships in computer-readable form over different types of information objects owned by local CRIS, it opens for the research community a new dimension for scientific creativity.

Scientists will be able virtually operating with all available research objects as with a "building kit" to express their opinions on existed research relationships between DIS objects or/and about a new scientific knowledge emerges from a combination of existed research outputs. Making semantic linkages between research objects they can design and share with the community logical/conceptual models of new research ideas or actual problems. Scientists can also create their own research outputs in a style of some semantic networks, e.g. by semantically linking separate research objects, "units of thought" or other types of research outputs, which can belong to different authors.

Thus networks of semantic linkages with declared authorship can be submitted by an author for its public exposition at DIS as new intellectual product and again can be re-used and developed by the community.

Technically the proposed tool can be realized as a specific open repository that delivers a non-traditional content for the community: a set of semantic linkages. We called it SLOR (Semantic Linkages Open Repository). The repository will allow any researcher to create semantic linkages caring by the semantic a data about relationship between DIS objects. It will store all semantic linkages metadata created by users and provide facilities for the community to manage and accumulate semantic linkages, to navigate and search over a set of accumulated linkages. The repository should also have some basic services like users' notifications about linking their objects and a new scientometrics produced on a base of accumulated semantic values.

The proposed SLOR should include some API to connect it with local CRIS. This API will provide data on request to external CRIS to display existed in/out semantic linkages for the viewed information objects.

A combination of SLOR, as a tool to link semantically any two objects or research DIS, and associated API, as added software for local CRIS to visualize created linkages, is producing an effect of universal interoperability. Based on this any CRIS can operate to some extent with content of other CRIS.

In designing SLOR there is a challenge to provide the complete and proper semantic values for covering all types of relationships that scientists can wish to express for linking objects of research DIS. A background for creation necessary semantics vocabularies includes: a semantic section of CERIF (Common European Research Information Format), recommendations of W3 - SKOS (Simple Knowledge Organization System), SWAN (Semantic Web Applications in Neuromedicine), SPAR (Semantic Publishing and Referencing Ontologies) and especially its parts CiTO (Citation Typing Ontology) and DoCo (Document Components Ontology).

We propose the "linkage" data type, which is based on CERIF "Link" entity specification (CERIF 2011). It is designed to carry out the subjective opinions of scientists about the relationship that can exist between any pairs of DIS objects, including "person", "organization", "project", "research result", "event" and some others. Rendered scientific relationships include: (1) relationships between the 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, organizations} on the other.

An openness for SLOR means: a) it is free to use, i.e. any scientist can use it to create semantic linkages between any available objects of research DIS (such proposals are moderated to be public); b) all public semantic linkages in the repository are open for harvesting and external using by other research information systems; c) openness of multiple semantic vocabularies for replenishing and development (proposals are moderated); d) DIS data types, which objects can be linked in SLOR can be expanded.

SLOR should notify scientists about linking/using their research outputs, as well about changes in research objects that they linked with own outputs. It improves research communication and increase efficiency of research work at large.

SLOR can provide new scientometrics based on quantitative and qualitative data about scientific relationships, like impact, usage and others. It can help with research assessment and evaluation and it improves the professional signaling system of the scientific community.

At the next section we are discussing technical details of the proposed repository. At third section we describe some supplementary services to increase benefits of the repository. The forth section presents a current state of our understanding on research relationship types structuring and initial content of semantic vocabularies associated with them. The last section concludes the topic.

This article develops an approach proposed by (Parinov & Kogalovsky 2011) for semantic structuring of digital libraries' contents, which resulted in explicit implementation of semantic linkages among information objects, opens new opportunities for scientific creativity and increase a quality of the libraries' contents. In combination with categorizing of semantic linkages it is establishing multilayer semantic networks over information objects that allow new qualitative scientometrics measurements and investigation on structuring properties of the corpus of science represented by digital libraries contents.

2 SLOR technique

It this section we are discussing technical details that necessary for SLOR proper functioning, including: (a) requirements to information objects owned by local CRIS to be linked with other objects; (b) a design of a semantic linkage itself; (c) requirements for repository of semantic linkages.

2.1 Objects for linking

When a scientist selects an object to link it with another one the SLOR should have following metadata of the selected object:

- object's URI to provide a valid hyperlink to the object web page at the web page of created semantic linkage;
- object's data type to determine subset of applicable semantic vocabularies;
- object's title for navigation and searching within SLOR over all stored linkages;
- object's author and his/her contact details (e.g. e-mail address) to determine who should be notified when a linkage will be created for the object;
- and the object's current modification date should be available to monitor if the object is changed and the authors of already created linkages with this changed object should be notified about a necessity to revise correctness of their semantic linkages.

There are at least three ways how thus metadata can be provided to SLOR: 1) it can be specified manually by a scientist who is creating a semantic linkage; 2) it can be harvested by SLOR once the scientist specified URI of the object from a parent research information system to which the object belongs; and 3) it can be received by a SLOR's request from some services of research e-infrastructure.

SLOR can harvest the metadata from the parent research information system if the object's metadata is available through a harvesting gateway (e.g. by OAI-PMH or RSS protocol) and in some popular format (CERIF is preferable).

There is an alternative opportunity for SLOR to harvest necessary metadata from the parent system if the metadata is integrated into HTML markup, e.g. by SCHEMA.ORG rules.

If parent information system does not provide a necessary metadata of the object through the listed above options, the SLOR in some cases can receive the metadata from research e-infrastructure. It can be done if a parent information system shares its metadata with some content integrator services, with information hubs and so on.

2.2 A semantic linkage design

A semantic linkage is an object of research DIS which includes URIs of a source and a target objects. Thus, all semantic linkages are oriented from a source object to a target one. A pair of objects URI is complemented by semantic value, which characterizes a type of relationship between the objects considering the orientation. Variety of semantic values is organized as a set of semantic vocabularies according different types of scientific relationships.

For practical using in SLOR a semantic linkage template also includes additional fields:

- URI, which can be generated by software or specified by the linkage creator manually, since the linkage exists as a research DIS object and has to have unique ID;
- a linkage title, which can be build by software, for presenting the linkage within SLOR table of contents;
- a comment where the linkage's creator (author) can explain specified semantic value;
- 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 linkage correctness because of changes in linked objects;
- dates of the linkage creation and revision, which is necessary for making decisions about sending notifications to authors of linked objects and/or to the linkage creator.

By creating semantic linkages the scientists is expressing their individual opinions about relationships between research DIS objects. So the same pair of the objects can have many different semantic linkages between them of both orientations. Each object from this pair can be the source or the target for the relationship orientation.

Semantic values of linkages between the same pair of objects are the intellectual products of scientists and so, in principle, they can contradict each other. To exclude possible inconsistency SLOR forbids a creator to make more than one linkage between the same pair of DIS objects with a semantic value from the same vocabulary. But it is feasible if a scientist creates linkages between the same pair of objects but with semantic values from different vocabularies for diverse types of relationships.

We defined a semantic linkage as an object of research DIS. It means this type of objects can be semantically linked with other DIS objects.

2.3 A repository

A repository of semantic linkages should have usual features. It should provide scientists with some personalized facilities to create semantic linkages, to store and manage them at some personal area, to submit the linkages for publication at open area of the repository. The submitted linkages are available for public utilization only if it passed through some usual control routine.

On the other side the repository has to provide some API for external usage. This software can be integrated into CRIS: (1) to make easier for users a semantic linking of information objects from local CRIS and storing the linkages into SLOR; and (2) to visualize already accumulated at SLOR outgoing/ingoing semantic linkages for information objects of local CRIS when a user is browsing over them.

Any external CRIS can in an automated mode check presence in SLOR of linkages for own information objects. If positive, the linkages' data can be harvested from SLOR to the CRIS using API. So external CRIS can visualize a network of linkages composed of articles and other information objects belong to this CRIS.

3 Supplementary services

Many SLOR benefits for the community of scientists will be produced by some supplementary services, which can be built either as a part of SLOR or outside, e.g. within a research e-infrastructure services. Useful supplementary services are: (a) monitoring of semantic linkages creation/change; (b) notifications of scientists about important events within DIS, e.g. about linking their research objects and so other; and (c) processing semantic values of accumulated linkages at SLOR to produce a new class of scientometric data.

3.1 Monitoring

A fundamental feature of the modern DIS is a changeable status of all information objects including linkages and their semantic. Time to time scientists can revise their paper, articles and other information objects deposited at DIS, e.g. to improve their research outputs, their personal records and so on. This phenomenon was called "liquid publication" (Casati et al. 2007) or "living document" (Parinov 2010).

If some semantic linkages were previously established by some scientists between a pair of research objects (e.g. between two articles), the created linkages may lose their consistency if one of the objects or both are revised by their authors. 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 object, 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 value since the scientist changed his/her opinion on it.

The monitoring service has to register all such events and then this service should send notifications and form/collect statistics.

3.2 Notifications

To fulfill the scientific circulation/communication on semantic linkages creation/revision and to support consistence of research DIS the monitoring services 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 (cited) in other objects (articles), that by this action she/he can violate have established 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 and, if it necessary, correct their linkages;
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 notification service improves scientific circulation because it immediately informs scientists about using their research outputs. And it improves research communication because authors of semantic linkages can receive a feedback on their actions from authors of linked research objects. As a result Science System gets more efficient information metabolism.

3.3 Scientometrics

Whereas some kinds of monitoring (e.g. changes in DIS scope and structure, viewing and download statistics and so on) are well known and have examples of good implementations (e.g. LogEc, MESUR, Socionet Stats, and other), there will be a new area of monitoring, created by possible intensive development of semantic linking over research DIS objects.

The monitoring service associated with SLOR should collect all available data about semantic linkages. It allows us to form a scientometric database both quantitative (number of linkages, etc.) and qualitative (semantic values) characteristics of scientific relationships:

- quantitative data about all accumulated semantic linkages at SLOR including 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 and semantic values specified by authors of semantic linkages accumulated at SLOR, including graphs of linkages with semantic values 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.

4 Scientific relationships and semantic vocabularies

We assume there are many types of initially hidden scientific relationships between objects of research DIS, which can be visualized for the research community within DIS and registered in some computer-readable form. Each type has a set of values, which characterize different possible states of the relationship. Technically thus set of values can be realized as a semantic vocabulary, associated with the relationship.

The repository of semantic linkages is a tool to create and make visual at the end the scientific relationships between linked objects of research DIS. Since the repository allows making semantic linkages between any two objects of research DIS, we should determine the scientific relationship types for each combination of pairs from a list of DIS objects data types: a source object type {"person", "organization", "research output", "project", etc.} → a target object type {"person", "organization", "research output", "project", etc.}.

In (Parinov & Kogalovsky 2011) we proposed 6 types of scientific relationships and 9 initial semantic vocabularies. For the pair "research output" → "research output" following types of scientific relationships and associated semantic vocabularies were specified (a source of initial semantic values is in brackets):

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

There is a situation when a scientist would like to recommend some research output to an author of some other research output, since the recommended one can improve the target output. E.g. an author of a new research output would like to recommend it for possible future using by other researchers. For such situations we made the relationship type “Usage proposal” which is valid for pair of data types "research output" → "research output” and has initial semantic vocabulary (source: none): “*can improve*”, “*can illustrate*”, “*can replace*”, etc.

For the pair of types "person" → "research output" there is a type "Professional opinions" with initial semantic vocabulary (SWAN): "*responds negatively to*", "*responds positively to*", "*responds neutrally to*". Using this type 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 relationship types and semantic vocabularies for some other pairs of data types:

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

Specified above types of scientific relationships and initial collections of associated semantic vocabularies should be recognized by the research community as proper and real characteristics. It can be achieved if relationship types and semantic vocabularies are opened for development and its new versions submission, and if users have a right of choice, i.e. there is a competition between different semantic vocabularies to be used for making semantic linkages.

Technically it means that scientists should be able to create a semantic vocabulary and/or to propose changes or new values for existing vocabularies. Than, such vocabularies as intellectual products with an explicit authorship should be submitted for public use as a part of SLOR procedure of making semantic linkages.

A semantic vocabulary belongs to the "classification" entity (data type) of CERIF. A value of the semantic vocabulary should be specified by a following template: "short name", "description", "usage recommendations", "authorship" data (personal/organizational), "creation/revision dates", and a "reference" to a source, if any.

To include into SLOR a semantic vocabulary as a set of values according such template the vocabulary's author should also specify a pair of data types and name of scientific relationship which the vocabulary was made for.

5 Conclusion

The growing use of the CRIS-CERIF concept in combination with the natural development of research information systems de facto leads to a certain standardization of the systems, e.g. concerning a substance, a set and a structure of information objects that they are operating with. Modern research information systems operate with virtually the same set of objects: "person," "organization," "project," "research output/outcome", "event" and some others. This allows us to consider the information objects that exist in the separate CRIS, as objects of a common research DIS and create tools and services for the use of these information objects in some new ways.

We propose a concept of specific open repository called SLOR, which will allow any researcher to create semantic linkages caring by semantic a relationship data between DIS objects, to store, manage and accumulate semantic linkages, to provide navigation and searching tools over a set of accumulated linkages. This repository should have some API, e.g. to provide data about linkages on request to external CRIS for visualization of a network of linkages composed of articles and other information objects belonging to this CRIS. And the repository of course should have as supplementary services a monitoring, notifications and scientometrics that processed accumulated semantic linkages data and related events.

With SLOR the researchers have a new tool and a new professional dimension for scientific creativity as well. Additionally to traditional way of scientific work now they can express a new scientific knowledge about relationships between separate research results by building multilayer networks of semantic linkages.

With SLOR scientists can create their research publications in a style of semantic networks. E.g. by linking separate research objects, "unit of thought" or other types of research outputs, which can belong to different authors. As a result, SLOR improves a scientific circulation mechanism and research outputs (nodes of semantic networks) can be easily reused by the research community.

SLOR notifies scientists about linking/using their research outputs, as well about changes in research objects that they linked with own outputs. It improves research communication and increase efficiency of research work at large.

SLOR provides new scientometrics based on semantic characteristics of research impact, usage and others. It can help with research assessment and evaluation and it improves the professional signaling system of the scientific community.

References

CERIF–1.3 (2011), euroCRIS. http://www.eurocris.org/Index.php?page=CERIF-1.3&t=1

Shotton, D. (2010a) 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. (2010b): CiTO, the Citation Typing Ontology. *J. 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/

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

Casati, F.; Giunchiglia, F.; Marchese, M. (2007): Publish and perish: why the current publication and review model is killing research and wasting your money, In *ACM Ubiquity* 8 (3), Feb 2007. http://www.acm.org/ubiquity/views/v8i03_fabio.html

Parinov, S. (2010): The electronic library: using technology to measure and support Open Science. In Proc. of the *World Library and Information Congress: 76th IFLA General Conference and Assembly*. 10-15 August 2010, Gothenburg, Sweden. pp. 1-13. http://socionet.ru/publication.xml?h=repec:rus:mqijxk:25

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://ceur-ws.org/Vol-803/paper13.pdf)

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

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

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

Contact Information

Sergey Parinov

Central Economics and Mathematics Institute (CEMI)
Russian Academy of Sciences (RAS)
Nakhimovsky pr. 47
Moscow
Russia
117418

sparinov@gmail.com