

Институту магистратуры СПбГУТ 5 лет

УДК 004.822

Визуализация структур данных в результатах запросов к RDF-хранилищам

Губин А. Н., Литвинов В. Л.✉, Филиппов Ф. В., Кравец А. Ю.

Санкт-Петербургский государственный университет телекоммуникаций им. проф. М. А. Бонч-Бруевича
Санкт-Петербург, 193232, Российская Федерация

Постановка задачи: современные поисковые системы должны использовать онтологические технологии для максимального удовлетворения информационных потребностей пользователя. Онтологии позволяют искать информацию не по конкретному слову, а по предметной области и связанным с ней объектам с помощью языка запросов SPARQL. При этом остается актуальной задача использования различных визуальных элементов для отображения сложных масштабных или структурных элементов. **Целью работы** является разработка методики визуализации результатов глобальных запросов в виде направленного графа, вершины которого отображают значения свойств сущностей, информация о которых извлекается из хранилища в результате работы (реализации) запроса. **Используемые методы:** решение задачи визуализации результатов глобальных запросов основано на применении онтологического подхода к построения семантических глобальных поисковых систем. **Новизна:** элементами новизны представленного решения является решение задачи отображения взаимосвязей данных, полученных в результате выполнения запросов к RDF-хранилищам данных. Результат визуализации представляет собой направленный граф, вершины которого отображают значения свойств сущностей, информация о которых извлекается из хранилища по окончании работы (реализации) SPARQL-запроса. **Результат:** расширение функциональности веб-сервиса для анализа и визуализации RDF-данных «RDF Grapher». Поддерживаемые форматы входных данных (сериализации RDF): Turtle, RDF/XML, RDF/JSON, N-Triples, TriG и N-Quads. Поддерживаемые форматы выходного изображения: PNG, SVG, PDF, PS, EPS, GIF и JPG. **Теоретическая/Практическая значимость:** представленная методика может быть реализована как совокупность программных библиотек в задачах проектирования глобальных информационных поисковых систем.

Ключевые слова: глобальные информационные системы, семантический поиск, онтологии, когнитивный поиск, SPARQL, RDF

Актуальность

Современные технологии интеграции данных позволяют эффективно взаимодействовать в одном пространстве сразу с несколькими информационными системами. Наиболее подходящим решением при проектировании таких информационно-справочных сервисов является семантическая сервисно-ориентированная архитектура SOA (аббр. от англ. Service Oriented Architecture) [1].

Фактически, глобальные информационные системы строятся в веб-пространстве, и реализация концепции управления ими как открытыми системами означает переход к технологиям Linked Open Data

Библиографическая ссылка на статью:

Губин А. Н., Литвинов В. Л., Филиппов Ф. В., Кравец А. Ю. Визуализация структур данных в результатах запросов к RDF-хранилищам // Вестник СПбГУТ. 2024. Т. 2. № 1. С. 1. EDN: HTOURZ

Reference for citation:

Gubin A., Litvinov V., Filippov F., Kravets A. Visualizing Data Structures in Query Results to RDF Storages // Herald of SPbSUT. 2024. Vol. 2. Iss. 1. P. 1. EDN: HTOURZ

(*аббр. от англ.* LOD) и предоставлению потребителю релевантной информации, полученной из достоверных источников [2]. Управление глобальной информационной системой, функционирующей в открытой среде, должно предусматривать решение не только внутренних проблем, но и всю совокупность проблем глобального информационного ресурса, связанных с внешней средой [3]. При этом современные поисковые системы должны использовать онтологические технологии для максимального удовлетворения пользователя. Онтологии помогают искать информацию не по конкретному слову, а по предметной области и связанным с ней объектам с помощью языка запросов SPARQL (*сокр. от* SPARQL Protocol and RDF Query Language).

SPARQL позволяет пользователям запрашивать информацию из баз данных, которые могут быть сопоставлены с RDF (*аббр. от англ.* Resource Description Framework). RDF – это абстрактная модель представления знаний, которая не отличает данные от метаданных, чтобы избежать расширения существующей модели с помощью аннотаций метаданных на уровне оператора, таких как показатели достоверности, веса, временные ограничения и информация о происхождении; например, если бы это была аннотация, измененная вручную. RDF использует интернационализированные идентификаторы ресурсов (IRI, *аббр. от англ.* Internationalized Resource Identifier) для определения сущностей, свойств и значений, чтобы обеспечить однозначное представление и интеграцию данных из разных источников. Можно сказать, что RDF – это стандартизированная графовая модель данных семантической паутины, а SPARQL – ее язык запросов и обновлений.

Для обеспечения понимания этих стремительно расширяющихся данных нужны соответствующие инструменты визуализации. Одним из таких общепризнанных инструментов является визуализация информации с применением графовых моделей [4]. Следует заметить, что осмысленные множества данных имеют разную структуру и требуют существенно различных стратегий при визуализации.

Постановка задачи

Визуализация данных предполагает использование таблиц, диаграмм или графиков. Обработка информации больших массивов данных, переведенных в визуальное представление, происходит гораздо легче [5]. В работе рассматривается решение задачи отображения взаимосвязей данных, полученных в результате выполнения запросов к RDF-хранилищам данных. Результат визуализации представляет собой направленный граф, вершины которого отображают значения свойств сущностей, информация о которых извлекается из хранилища в по окончании работы (реализации) SPARQL-запроса. Для решения поставленной задачи используем веб-сервис для анализа и визуализации RDF-данных «RDF Grapher» [6]. Поддерживаемые форматы входных данных (сериализации RDF): Turtle, RDF/XML, RDF/JSON, N-Triples, TriG и N-Quads. Поддерживаемые форматы выходного изображения: PNG, SVG, PDF, PS, EPS, GIF и JPG. Web-интерфейс сервиса представлен на рис. 1.

Методика решения

Рассмотрим предлагаемую методику на конкретном примере. Поставлена задача определения типов самолетов, выпущенных объединением Микояна, с указанием наименования самолета, даты первого полета, скорости полета, количества двигателей, типа двигателя, наименования предприятия, выпускающего двигателя. Используем ресурсы сервиса «SPARQL Query Editor» [6] (<http://dbpedia.org>). Текст запроса представлен на рисунке 2. Значение параметра LIMIT в запросе ограничим величиной 5 для обеспечения читабельности результатов запроса. Результаты выполнения запроса представим в виде таблицы 1.

Для обеспечения функционирования сервиса «RDF Grapher» необходимо преобразовать результаты работы запроса в один из поддерживаемых им формат данных. Рассмотрим работу сервиса с данными, представляющими собой совокупность RDF-триплетов (формат Turtle). Преобразуем исходный запрос в запрос типа «CONSTRUCT», содержание блока «where» оставим без изменений (как показано на рисунке 3), в этом случае «RDF Grapher» отобразит все связи, указанные в блоке «where».

Исполнение запроса генерирует сериализацию RDF, ссылка на которую представлена на рисунке 4. Загрузив ссылку в окно «RDF data or URI» и активизировав режим «Visualize», получим требуемый результат, представленный на рисунке 5.



Рис. 1. Web-интерфейс сервиса «RDF Grapher»

```
Select distinct ?airplane ?name ?date ?speedKmh
?numberEngine ?name1 ?name2
where
{
?airplane a dbo:Aircraft.
?airplane dbo:manufacturer dbr:Mikoyan.
?airplane rdfs:label ?name.
?airplane dbp:firstFlight ?date.
?airplane dbp:maxSpeedKmh ?speedKmh.
?airplane dbp:eng1Number ?numberEngine.
?airplane dbp:eng1Name ?engine.
?engine rdfs:label ?name1.
?engine dbp:manufacturer ?works.
?works rdfs:label ?name2.
filter (lang (?name )='en').
filter (lang (?name1 )='en').
filter (lang (?name2 )='en').
}
LIMIT 5
```

Рис. 2. Текст запроса типа Select

```
CONSTRUCT
where
{
?airplane a dbo:Aircraft.
?airplane dbo:manufacturer dbr:Mikoyan.
?airplane rdfs:label ?name.
?airplane dbp:firstFlight ?date.
?airplane dbp:maxSpeedKmh ?speedKmh.
?airplane dbp:eng1Number ?numberEngine.
?airplane dbp:eng1Name ?engine.
?engine rdfs:label ?name1.
?engine dbp:manufacturer ?works.
?works rdfs:label ?name2.
filter (lang (?name )='en').
filter (lang (?name1 )='en').
filter (lang (?name2 )='en').
}
LIMIT 5
```

Рис. 3. Текст запроса типа CONSTRUCT

Таблица 1. Результаты выполнения запроса типа Select

airplane	name	date	speedKmh	number Engine	name1	name2
http://dbpedia.org/resource/Mikoyan-Gurevich_DIS	"Mikoyan-Gurevich DIS"@en	1941	610	2	"Mikulin AM-37"@en	"Mikulin OKB"@en
http://dbpedia.org/resource/Mikoyan-Gurevich_MiG-15	"Mikoyan-Gurevich MiG-15"@en	1947	1076	1	"Klimov VK-1"@en	"Klimov"@en
http://dbpedia.org/resource/Mikoyan-Gurevich_I-320	"Mikoyan-Gurevich I-320"@en	1949	1090	2	"Klimov VK-1"@en	"Klimov"@en
http://dbpedia.org/resource/Mikoyan-Gurevich_Ye-8	"Mikoyan-Gurevich Ye-8"@en	1962	1386	1	"Tumansky R-21"@en	"Tumansky"@en

```

https://dbpedia.org/sparql?default-graph-uri=http%3A%2F%2Fdbpedia.org&query=%23select+distinct+%3Fairplane+%3Fname+%3Fdate+%3FspeedKmh%0D%0A+++++++%23+++%3FnumberEngine+%3Fname1+%3Fname2%0D%0A%0D%0A%7B%0D%0A%3Fairplane+a+dbo%3AAircraft.%0D%0A%3Fairplane+dbo%3Amanufacturer+dbr%3AMikoyan.%0D%0A%3Fairplane+rdfs%3Alabel+%3Fname.%0D%0A%3Fairplane+dbp%3AfirstFlight+%3Fdate.%0D%0A%3Fairplane+dbp%3AmaxSpeedKmh+%3FspeedKmh.%0D%0A%3Fairplane+dbp%3Aeng1Number+%3FnumberEngine.%0D%0A%3Fairplane+dbp%3Aeng1Name+%3Fengine.%0D%0A+%3Fengine+rdfs%3Alabel+%3Fname1.%0D%0A+%3Fengine+dbp%3Amanufacturer+%3Fworks.%0D%0A+%3Fworks+rdfs%3Alabel+%3Fname2.%0D%0A%0D%0A%28%28lang+%28%3Fname+%29%3D%27en%27%29.%0D%0A%0D%0A%28%28lang+%28%3Fname1+%29%3D%27en%27%29.%0D%0A%0D%0A%28%28lang+%28%3Fname2+%29%3D%27en%27%29.%0D%0A%7D%0D%0A+LIMIT+5%0D%0A%0D%0A&format=text%2Fturtle&timeout=10000&signal_void=on&signal_unconnected=on
  
```

Рис. 4. Ссылка на результаты работы запроса типа CONSTRUCT

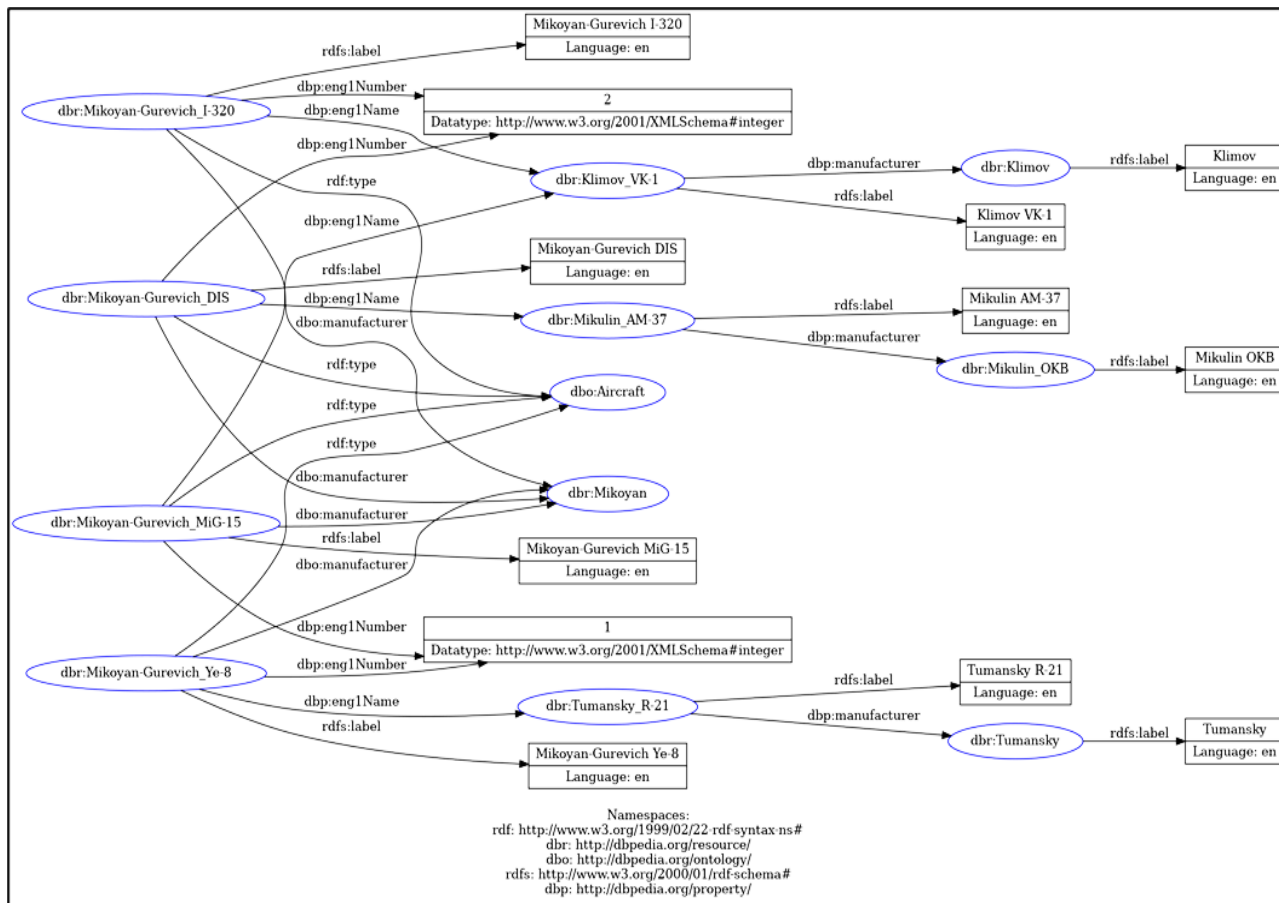


Рис. 5. Результат визуализации структуры данных запроса к RDF-данным

Если требуется отобразить только часть связей, указанных в блоке «WHERE», то блок «CONSTRUCT» используется с параметрами, в которых необходимо указать набор требуемых триплетов. Например, при необходимости отобразить структуру данных, относящихся только к двигателям искомых самолетов, в блок «CONSTRUCT» запроса необходимо добавить следующие данные, показанные на рисунке 6. Результат визуализации структуры данных, полученных при выполнении последнего запроса, представлен на рисунке 7.

```

CONSTRUCT
{
?airplane a dbo:Aircraft.
?airplane dbo:manufacturer dbr:Mikoyan.
?airplane rdfs:label ?name.
?airplane dbp:eng1Number ?numberEngine.
?airplane dbp:eng1Name ?engine.
?engine rdfs:label ?name1.
?engine dbp:manufacturer ?works.
?works rdfs:label ?name2.
}
where
{
?airplane a dbo:Aircraft.
?airplane dbo:manufacturer dbr:Mikoyan.
?airplane rdfs:label ?name.
?airplane dbp:firstFlight ?date.
?airplane dbp:maxSpeedKmh ?speedKmh.
?airplane dbp:eng1Number ?numberEngine.
?airplane dbp:eng1Name ?engine.
?engine rdfs:label ?name1.
?engine dbp:manufacturer ?works.
?works rdfs:label ?name2.
filter (lang (?name)='en').
filter (lang (?name1)='en').
filter (lang (?name2)='en').
}
LIMIT 5
    
```

Рис. 6. Запрос получения RDF-триплетов блока CONSTRUCT с параметрами

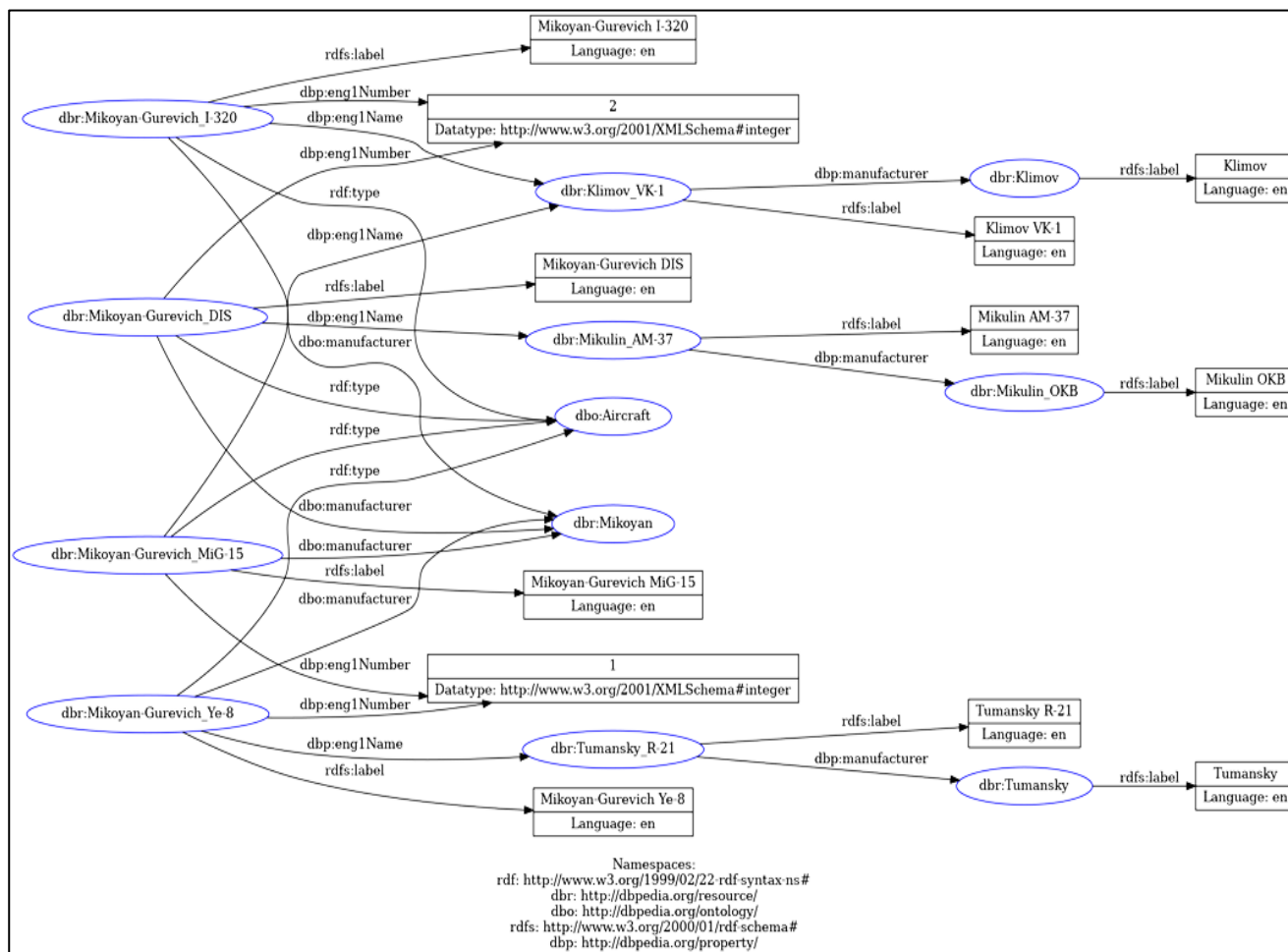


Рис. 7. Результат визуализации структуры данных запроса к RDF-данным с учетом ограничений

Заключение

Таким образом, предложенная методика визуализации структуры RDF-данных позволяет получить отображения взаимосвязей данных и представляет собой направленный граф, вершины которого отображают значения свойств сущностей, информация о которых извлекается из хранилища в результате работы (реализации) запроса. С его помощью можно визуализировать результаты запросов к большому количеству неоднородных данных в различных форматах и источниках, что позволит более эффективно извлекать скрытую в них информацию. К недостаткам предложенной методики можно отнести необходимость трансформации SPARQL-запросов типа SELECT в запросы типа CONSTRUCT. Представляет интерес в дальнейшем рассмотреть возможность реализации вывода результатов запросов типа SELECT в формате ориентированного графа.

Литература

1. Кашалкин Д. Ю., Курчидис В. А. Принципы построения семантической сервис-ориентированной архитектуры // Моделирование и анализ информационных систем. 2007. Т. 14. № 1. С. 48–53. EDN: LAOFQD
2. Vandenbussche P.-Y., Vatant B. Linked Open Vocabularies // ERCIM News 96. 2014. PP. 21–22. URL: <https://ercim-news.ercim.eu/en96/special/linked-open-vocabularies> (дата обращения 03.03.2024)
3. Litvinov V. L., Filippov F. V. Paradigm of controls concept for global information systems // Proceedings of the 3rd International Conference on Control in Technical Systems (CTS). 2019. PP. 228–230. DOI: 10.1109/CTS48763.2019.8973235. EDN: MQYXFT
4. Апанович З. В., Винокуров П. С., Кислицина Т. А. Средства визуального анализа информационного наполнения порталов, входящих в облако Linked Open Data // Труды 13-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2011. Воронеж: Воронежский государственный университет, 2011. С. 216–223. EDN: SCQDRD
5. Базалева О. И. Мастерство визуализации данных. М.: Изд-во Диалектика, 2018. 192 с.
6. RDF Grapher. URL: <https://www.ldf.fi/service/rdf-grapher> (дата обращения 03.03.2024)
7. Global and Unified Access to Knowledge Graphs // DBpedia. URL: <http://dbpedia.org> (дата обращения 03.03.2024)

Статья поступила 04 марта 2024 г.

Одобрена после рецензирования 15 марта 2024 г.

Принята к публикации 16 марта 2024 г.

Информация об авторах

Губин Александр Николаевич – кандидат технических наук, доцент, доцент кафедры информационных управляющих систем Санкт-Петербургского государственного университета телекоммуникаций имени проф. М. А. Бонч-Бруевича. E-mail: gan50_60@mail.ru

Литвинов Владислав Леонидович – кандидат технических наук, доцент, доцент кафедры информационных управляющих систем Санкт-Петербургского государственного университета телекоммуникаций имени проф. М. А. Бонч-Бруевича. E-mail: vlitvinov@sut.ru

Филиппов Феликс Васильевич – кандидат технических наук, старший научный сотрудник, доцент кафедры информационных управляющих систем Санкт-Петербургского государственного университета телекоммуникаций имени проф. М. А. Бонч-Бруевича. E-mail: 9000096@mail.ru

Кравец Алексей Юрьевич – магистрант 2 курса (направление 09.04.02 Информационные системы и технологии) Санкт-Петербургского государственного университета телекоммуникаций имени проф. М. А. Бонч-Бруевича. E-mail: kravets.au@gmail.com

Visualizing Data Structures in Query Results to RDF Storages

A. Gubin, V. Litvinov, F. Filippov, A. Kravets

The Bonch-Bruevich Saint Petersburg State University of Telecommunications,
St. Petersburg, 193232, Russian Federation

Purpose. Modern search systems must use ontological technologies to maximally satisfy the user's information needs. Ontologies allow you to search for information not by a specific word, but by a subject area and related objects using the SPARQL query language. At the same time, the task of using various visual elements to display complex large-scale or structural elements remains relevant. The goal of the work is to develop a methodology for visualizing the results of global queries in the form of a directed graph, the vertices of which display the values of the properties of entities, information about which is retrieved from the storage as a result of the operation (implementation) of the query. **Methods.** The solution to the problem of visualizing the results of global queries is based on the use of an ontological approach to the construction of semantic global search systems. **Novelty.** The novelty of the presented solution is the solution to the problem of displaying the relationships of data obtained as a result of executing queries to RDF data stores. The visualization result is a directed graph, the vertices of which display the values of the properties of entities, information about which is retrieved from the storage as a result of the operation (implementation) of the request. **Results.** Expanding the functionality of the web service for analysis and visualization of RDF data "RDF Grapher". Supported input data formats (RDF serialization): Turtle, RDF/XML, RDF/JSON, N-Triples, TriG and N-Quads. Supported output image formats: PNG, SVG, PDF, PS, EPS, GIF and JPG. **Practical relevance.** The presented methodology can be implemented as a set of software libraries in problems of designing global information search systems.

Keywords: global information systems, semantic search, ontologies, cognitive search, SPARQL, RDF

Information about Authors

Alexandr Gubin – Ph.D. of Engineering Sciences, Associate Professor, Associate Professor at the Department of Information Control Systems (The Bonch-Bruevich Saint Petersburg State University of Telecommunications). E-mail: gan50_60@mail.ru

Vladislav Litvinov – Ph.D. of Engineering Sciences, Associate Professor, Associate Professor at the Department of Information Control Systems (The Bonch-Bruevich Saint Petersburg State University of Telecommunications). E-mail: vlitvinov@sut.ru

Felix Filippov – Ph.D. of Engineering Sciences, Associate Professor, Associate Professor at the Department of Information Control Systems (The Bonch-Bruevich Saint Petersburg State University of Telecommunications). E-mail: 9000096@mail.ru

Alexey Kravets – a 2nd year master's student (The Bonch-Bruevich Saint Petersburg State University of Telecommunications). E-mail: alexeipyp@mail.ru