ИМПОРТ РАЗНОРОДНЫХ МЕТАДАННЫХ НАУЧНОГО ЭКСПЕРИМЕНТА В ЦЕНТРАЛИЗОВАННОЕ ОНТОЛОГИЧЕСКОЕ ХРАНИЛИЩЕ

А.Ю. Кайда, М.Ю. Губин
Томский политехнический университет
anastasiakaida@gmail.com

Введение

На сегодняшний день проведение научных экспериментов сопряжено с необходимостью агрегирования больших объемов разнородных метаданных. Более того, требуется обращение к полученным ранее данным и поиск необходимой информации, сопряженных с ней файлов различных типов. Решить эту задачу помогает единое централизованное онтологическое хранилище метаданных. Однако для наполнения такой базы требуются программные компоненты, отвечающие за импорт метаданных из разнообразных источников данных. Существенной проблемой является отсутствие на рынке программных продуктов, соответствующих требованиям разработчиков базы знаний научного эксперимента и способных решить поставленную задачу, в связи с чем возникает необходимость разработки соответствующего инструмента.

Модуль импорта метаданных

Хранимые в репозиториях гетерогенные данные, как правило, не имеют между собой семантической связи, что затрудняет поиск необходимой информации в больших научных экспериментах. Data Knowledge Base – централизованное онтологическое хранилище агрегированных из структурированных и документальных источников метаданных о научных исследованиях – разрабатывается для решения данной задачи. Хранилище состоит из двух уровней: агрегатора метаданных и самого онтологического хранилища, в основе которого лежит онтологическая модель научного исследования с параметрическим описанием экспериментов и сопроводительных документов. Для загрузки в RDF-хранилище Virtuoso Universal Server требуется определенное представление данных, поступающих на обработку в виде TXT и JSON-файлов, содержащих необходимые метаданные. Извлеченные данные должны быть представлены некоторым единым образом и загружены в хранилище. Такое представление реализуется в синтаксисе Turtle, поддерживаемом Virtuoso Universal Server.

Модель данных представляет собой ориентированный граф (граф, ребрам которого присвоено направление) и предназначена для представления интегрированной гетерогенной информации, взятой из множества разнородных источников. Согласно онтологической модели данных, метаданные представляются в виде триплетов – троек «субъект-предикат-объект».

Синтаксис Turtle является упрощенным вариантом записи RDF-триплетов в виде групп URI-ресурсов. URI-идентификатор ресурса ссылается на абстрактный или физический ресурс, причем идентификатор может ссылаться на любой тип ресурса. Субъект представляет собой унифицированный идентификатор ресурса URI, ссылающийся на конкретный описываемый элемент. Объект может быть представлен как в виде литерала, так и в виде ссылки URI. В свою очередь, предикат, соединяющий субъект и объект, является своеобразным индикатором, указывающим на то, как связаны объект и субъект в модели. Каждое свойство может обладать рядом характеристик, таких как транзитивность, рефлексивность, иррефлексивность, симметричность, инверсия и т.д.

Задача модуля состоит в том, чтобы извлечь данные из файлов и представить в виде триплетов, избегая формирования неполных троек или дублирования записей. Метаданные, представленные в таком виде, готовы для загрузки в систему. Обработка данных в системе происходит в потоковом режиме. В каждом сообщении передаются данные из файлов, предназначенные для обработки. На выход поступает сообщение с набором триплетов и передается для записи в хранилище данных согласно общей логике работы системы. Роль модуля определена на уровне подготовки данных согласно концепции связанных данных и представлена на рисунке 1.

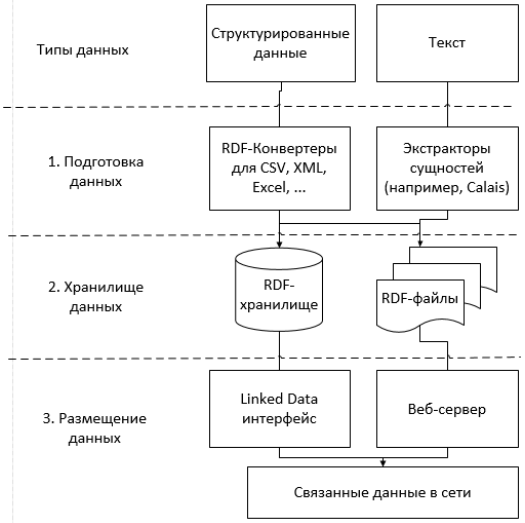

Рис. 1. Общая архитектура системы в концепции связанных данных

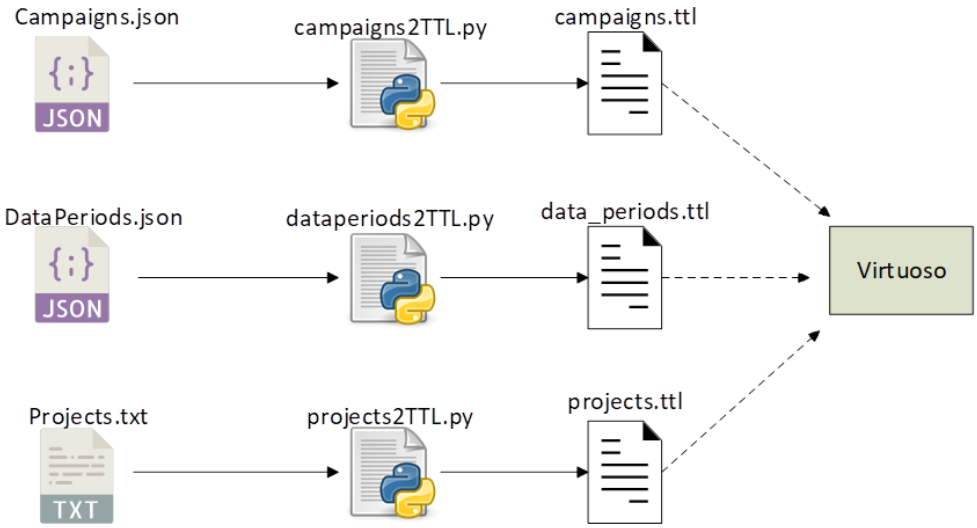

Рис. 2. Принцип работы модуля импорта

Заключение

В результате был разработан прототип модуля импорта метаданных, представляющий собой набор скриптов и обеспечивающий получение информации из файлов TXT и JSON, содержащих метаинформацию о научных экспериментах, и обеспечивающий корректное добавление информации в единую базу знаний научного эксперимента. Сама база знаний должна ускорить и упростить поиск необходимых метаданных и соответствующих документов о научном эксперименте. Для разработки прототипа модуля был выбран высокоуровневый язык программирования Python.

Стандарты W3C, относящиеся к Semantic Web, на протяжении лет не находили отклика у широкой аудитории, однако являются эффективным набором инструментов в решении задач по созданию баз знаний и агрегированию гетерогенных данных в единую систему. Тем не менее, логика работы модуля строго зависит от заданной структуры входных данных и построенной модели данных.

Полученные результаты могут быть использованы в узкоспециализированных базах знаний научного эксперимента для автоматизации процесса хранения, накопления и обработки метаданных.

Список использованных источников

1. Soldatova L., King R. An Ontology of Scientific Experiments // Journal of the Royal Society Interface. ─ 2006. ─ Issue 11. ─P.795-804, DOI: 10.1098/rsif2006.0134
2. Allemang D., Hendler J. Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL // Elsevier. — 2011, ISBN 978-0-12-385965-5
3. Tim Berners-Lee et al. Tabulator: Exploring and analyzing linked data on the semantic web. In Proceedings of the 3rd International Semantic Web User Interaction Workshop, 2006. [Электронный ресурс]. – URL: http://swui.semanticweb.org/swui06/papers/BernersLee/Berners-Lee.pdf
4. Powers S. Practical RDF – O'Reilly Media, 2003. – 352 p.
5. Semantic Web Standarts. RDF – Resource Description Framework. [Электронный ресурс]. – URL: https://www.w3.org/RDF/ (дата обращения 20.08.2017)
6. Hitzler P., Krötzch M., Rudolph S., Foundations of Semantic Web Technologies – FL.: Chapman & Hall/CRC, 2009. – 455 p.
7. H. Wache, T. Voegele, T. Visser, H. Stuckenschmidt, H. Schuster, G. Neumann, and S. Huebner. IJCAI-01 Workshop: Ontologies and Information, page 108--117. (2001)
8. Nasser Alalwan, Hussein Zedan, François Siewe, «Generating OWL Ontology for Database Integration», Third International Conference on Advances in Semantic Processing, 2009
9. RIF Overview (Second Edition) [Электронный ресурс]. - Режим доступа: http://www.w3.org/TR/rif-overview/, свободный
10. Terse RDF Triple Language [Электронный ресурс]. - Режим доступа: http://www.w3.org/TR/turtle/, свободный
11. Python 2.7.14rc1 documentation [Электронный ресурс]. - Режим доступа: https://docs.python.org/2/, свободный