Применение семантических технологий для обработки связанных данных в геоинформационных системах

И. О. Сычёв, Ю. А. Кораблёв
СПБГЭТУ «ЛЭТИ»
Санкт-Петербург
orestes2358@gmail.com, juri.korablev@gmail.com

Л. С. Звягин
Финансовый университет при Правительстве Российской Федерации (Финуниверситет),
Financial University
lszvyagin@fa.ru

***Аннотация*. Рассматриваются методы построения хранилищ данных для геоинформационных систем. Предлагается архитектура семантического хранилища данных, обеспечивающая возможности распределённой обработки информации.**

Ключевые слова: система поддержки принятия решений; онтология; база знаний; RDF; ГИС

I. Введение

В современных географических информационных системах (ГИС) различного назначения важной проблемой является организация распределённого процесса принятия решений в условиях неполноты и нечёткости исходной информации. Лицам, принимающим решения (ЛПР), необходимы развитые инструменты для обработки больших массивов входных данных, которые могут повысить качество принимаемых решений. Онтология является подходящим инструментом для формализации знаний о предметной области и формирования модели данных, удобной для машинной обработки. Интеграция онтологии в ГИС является сложным процессом, сопряжённым с необходимостью слияния онтологической модели со схемами и структурами данных, используемых в существующих ГИС. В статье рассматриваются общие принципы применения онтологий в морских ГИС.

II. Общие принципы функционирования морской ГИС

ГИС – система сбора, хранения, анализа и графической визуализации пространственных данных и связанной с ними информации. Морские ГИС широко применяются для решения задач в различных судовых и корабельных информационно-управляющих системах. Современная ГИС следующие задачи [1]:

- сбор, обработка и хранение информации, получаемой от внешних источников;
- представление на экране карты морской обстановки;
- прогнозирование местоположения подвижных объектов на заданный момент времени;
- ввод, хранение и отображение информации об объектах, состоянии объектов и технических средствах;
- обеспечение навигационного ориентирования судов;
- обеспечение возможности поиска кораблей и судов.

В ГИС требуется обработка неформализованной информации (текстовые сообщения, фото) и формализованной информации (данные от сенсоров, картографические данные).

Для выработки решений ЛПР необходимо обрабатывать информацию от внешних источников о текущей морской обстановке. Располагая развитым программным комплексом, ЛПР имеет возможность оценить возможные угрозы и принять соответствующе решение. При принятии решении оператор учитывает как данные, поступающие в реальном времени (пеленг, курс, скорость, метеорологические условия и т.д.), так и данные, относящиеся к справочной информации (характеристики судов, технических средств, местоположения географических объектов). В случае отсутствия функции выработки рекомендаций в ГИС, решение принимается на основании личного опыта ЛПР и его знаний.

Формализация знаний экспертов и хранение нормативно-справочной информации (НСИ) в виде, удобном для машинной обработки является ключевым аспектом для обеспечения функции поддержки принятия решений в морской ГИС.

III. Архитектура информационной подсистемы ГИС

В ГИС целесообразно разделение информации на два типа (рис. 1).

- оперативные данные, поступающие от внешних источников;
- база знаний, содержащая выявленные знания о предметной области.

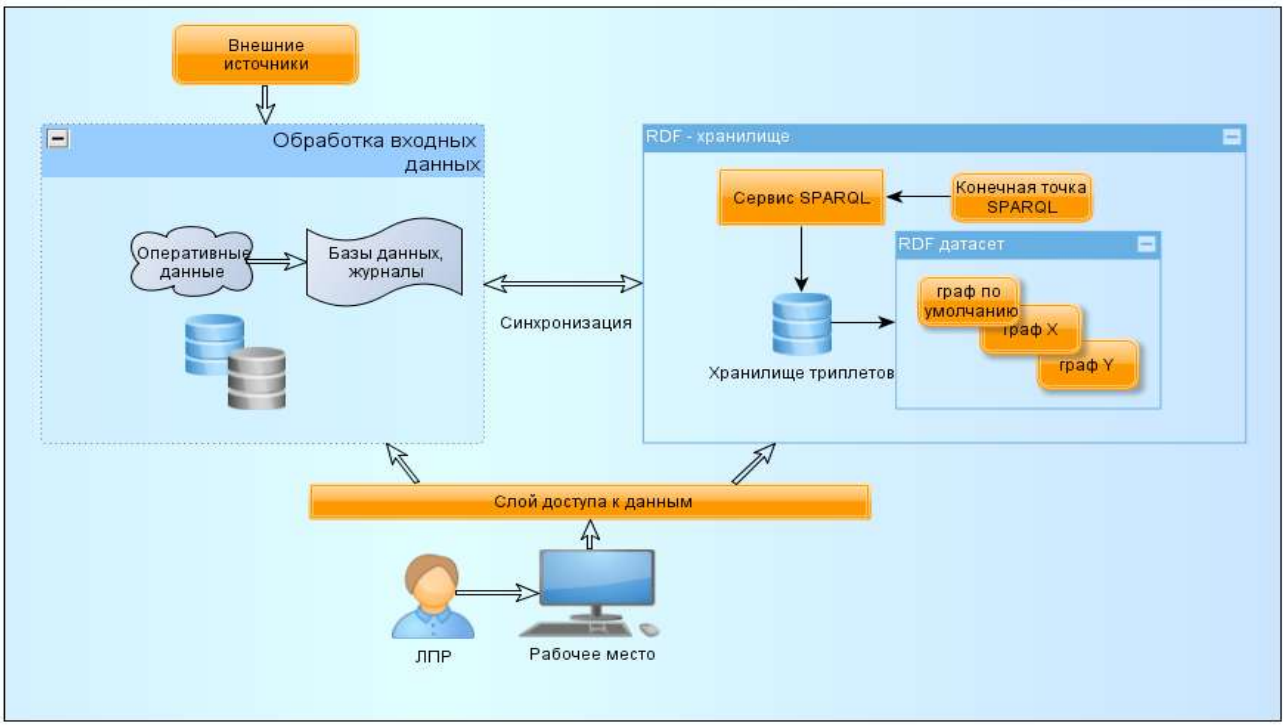

Рис. 1. Хранение данных в ГИС

Модуль обработки входных данных выполняет приём и обработку информации от внешних источников и взаимодействующих систем. Запись информации может производится в базы данных (БД) различного вида: реляционные БД (PostgreSQL, MySql), встраиваемые БД типа «ключ-значение» (mdbx), документоориентированные БД (MongoDb). Оперативная информация позволяет оператору проводить оценку текущей обстановке на карте.

База знаний реализуется средствами linked data: RDF – среда описания ресурса, модель представления данных; OWL – язык описания онтологий; SPARQL – набор языков и протоколов для извлечения и модификации данных в RDF хранилище.

Структура RDF – это коллекция триплетов вида «субъект-предикат-объект», образующих RDF граф. Каждый триплет описывает связь между двумя элементами (субъект и объект) через дугу. Дуга всегда направлена от субъекта к объекту.

Исходная онтология создаётся в редакторе Protégé. Чтобы обеспечить совместный доступ нескольких операторов к данным, онтология сохраняется в хранилище триплетов – базе данных, предназначенной для хранения и извлечения триплетов с помощью семантических запросов.

Взаимодействие с хранилищем осуществляется по протоколам Sparql 1.1 Protocol и Graph Store HTTP Protocol [2]. Клиентские программы (SPARQL клиенты) выполняют запросы к хранилищу. Запросы отправляются по протоколу HTTP на конечную точку SPARQL – URL, по которому сервис принимает запросы. Сервис SPARQL выполняет SPARQL-запросы и возвращает результаты на клиентов. Информация хранятся в наборе данных RDF – коллекции, включающей один или более графов.

В качестве хранилища триплетов выбрана свободная платформа Apache Jena. К её преимуществам относятся полная поддержка стандарта SPARQL 1.1, наличие программного интерфейса для работы с OWL и поддержка логического вывода. На разработанном наборе данных Apache Jena обеспечила высокую скорость выполнения запросов, не уступающую реляционным БД.

В рамках работы были разработаны [4]:

- онтология для морской ГИС, включающая информацию следующих видов: объекты информационного поля (корабли, суда); технические средства судов, технические характеристики средств и др. Базовые идеи построения онтологий для схожих направлений хорошо описаны в [3];
- интерфейс для просмотра и изменения данных в онтологии.

Пример триплетов представлен на рис. 2. В процессе разработки онтологии не всегда очевидно, к какому типу данных относится заданное понятие предметной области. Так, например, классификаторы судов разумно вынести на уровень индивидов через задание отношения «тип-подтип», поскольку нельзя выстроить одно дерево классов судов для разных стран. Технические характеристики судов вынесены в отдельные наборы свойств: такой подход позволяет задавать одни и те же наборы свойств нескольким судам.

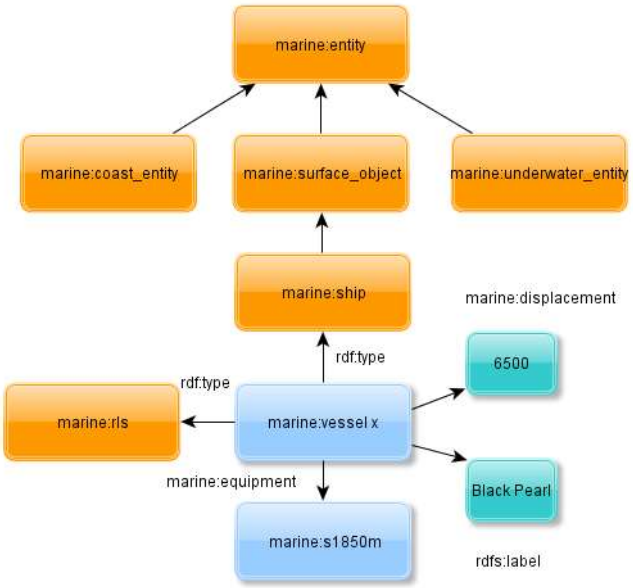

Рис. 2. Пример RDF триплетов

Оператор может сам добавлять новые свойства объектов в онтологию через интерфейс изменения данных, расширяя словарь. Данная функция повышает гибкость системы в случае, если на этапе разработки отсутствуют полные требования к модели данных.

Для обеспечения согласованности данных в онтологии и во внешних источниках данных используются уникальные URI. URI используются для идентификации объектов в других базах данных, содержащих информацию, необходимую для выполнения задач в реальном времени.

IV. Поиск и извлечение данных

Использование онтологии в ГИС даёт ряд преимуществ по сравнению с БД других типов. Реляционные СУБД эффективны для обработки транзакций в реальном времени, но плохо подходят для формализации знаний экспертов и хранения справочной информации. Упрощается процесс написания запросов для поиска нужной информации, исключается необходимость формирования сложных запросов с использованием соединений таблиц.

Структура RDF графа позволяет сформировать словарь предметной области и выстроить сложные зависимости между объектами. Наборы классов и свойств формируют словарь, позволяющий создавать индивидов и описывать характеристики и отношения между ними. Для поиска нужной информации пользователю онтологии (которым может быть как оператор, так и программный агент) нужно знать используемый словарь, а не фиксированную схему данных как в реляционной БД. Например, пользователь может сформировать запрос для выборки «всех объектов, имеющих водоизмещение 35000 тонн» без указания класса объекта:

```
select ?entity ?label ?type ?value
where {
 ?entity rdf:type ?type.
 ?entity nsi:has_displacement_full ?value
 filter (?value = "35000"^^xsd:integer)}
```

Для поиска связи между двумя элементами графа может быть использован синтаксис property path (путь свойства). Property path – это возможный маршрут в графе между двумя узлами. Тривиальный случай использования property path – поиск индивида, который является экземпляром непрямого потомка заданного класса:

```
select ?entity ?label
where {
 ?entity rdf:type ?type.
 ?entity rdfs:label ?label.
 ?type rdfs:subClassOf* nsi:surface_object
 }
```

Отметим, что процесс проектирования онтологии не является тривиальной задачей, поскольку структура онтологии должна быть адаптирована под алгоритмы, используемые в системе. При разработке онтологии должен быть учтён весь спектр задач для определения, того к какому типу отнести выбранное понятие предметной области. Не всегда естественное разделение по классам и индивидам является оптимальным для алгоритмической обработки.

V. Механизмы распределённого взаимодействия

Важным элементом хранения данных в морской ГИС является возможность распределённого взаимодействия нескольких операторов в рамках единого информационного поля (рис. 3). ЛПР могут работать с собственными хранилищами данных, содержащими результаты выполнения расчётных задач в локальном пространстве пользователя. Такая технология обеспечивает возможность децентрализованного процесса принятия решений как в условиях отсутствия центрального сервера БД, так и в случае необходимости организации локального хранения данных в пространстве оператора.

Онтологии изначально разрабатывались для работы со связанными данными (linked data). Linked data определяет набор практик для размещения данных во всемирной паутине [5]. В основе концепции лежат понятия: URI, HTTP, RDF, SPAQL. Информация об объекте должна содержать ссылки на другие URI, чтобы пользователи могли получать большее количество информации. Использование принципов linked data может помочь в формализации большого объёма информации, накопленного в морских ГИС и формировании единого общего морского словаря для работы с данными.

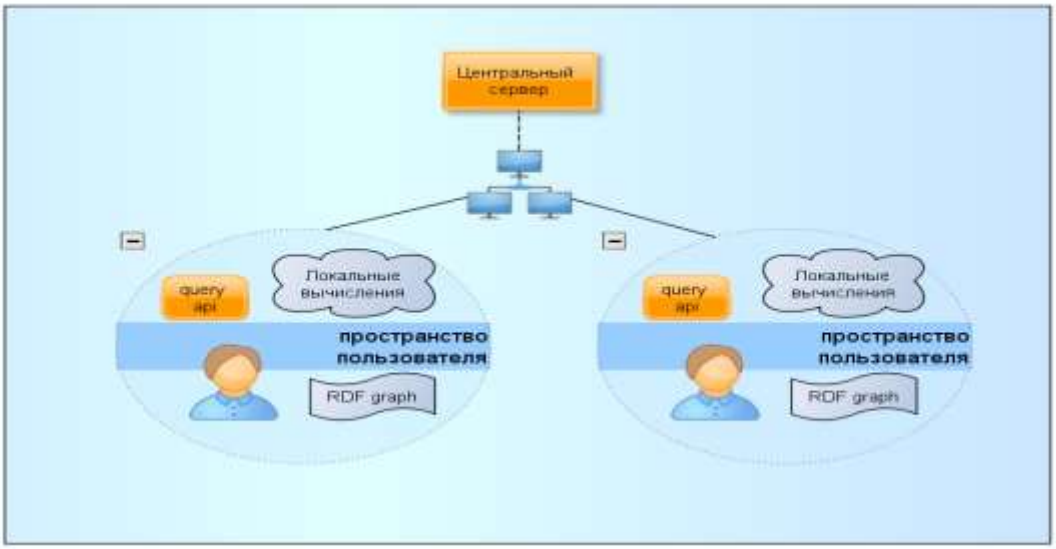

Рис. 3. Распределение информации

Ключевой элемент linked data – возможность хранения данных на разных узлах сети. Расширение SPARQL 1.1 Federated Query позволяет выполнять запросы над несколькими конечными точками SPARQL. Таким образом, оператор получает возможность получить доступ к данным по объектам, отсутствующим в его собственном локальном пространстве.

Распределение вычислений в сети позволяет обеспечить распределённую работу операторов: обработку документов, выполнение задач и т.д. Такая архитектура повышает отказоустойчивость и гибкость системы, но требует развитых алгоритмов для синхронизации информации при её изменении.

VI. Заключение и направление дальнейших исследований

В статье рассмотрены методы разработки подсистемы хранения и обработки информации для морской ГИС. Представлен подход к хранению информации с использованием технологического стека linked data, как наиболее подходящего инструмента для хранения справочной информации и формализации знаний экспертов. Широкое применение linked data в морских ГИС позволит сформировать единые словари для обработки данных.

Дальнейшим направлением работы является поиск и применение алгоритмов, обеспечивающих возможность динамической генерации запросов к графу. Актуальность данной проблемы обусловлена тем, что для формирования запроса к графам знаний пользователю необходимо знать внутреннюю структуру и словарь, используемый в онтологии.

Другое направление исследования – аналитическая обработка данных в графе:

- предсказание связи – предсказание существования заданных рёбер в графе (триплетов);
- разрешение сущностей – определение, какие объекты относятся к одним и тем же сущностям.

Алгоритмы аналитической обработки данных в графе позволят находить скрытые закономерности в данных и дополнять граф в условиях неполноты исходной информации.

Список литературы

[1] Интеллектуальные географические информационные системы для мониторинга морской обстановки / под редакцией В.В. Поповича, М.: Наука 2013. 283 с.

[2] «Обзор SPARQL 1.1» Электронный ресурс, url: www.w3.org/TR/sparql11-overview

[3] Metadata and sematic research 5th international conference: «Serhan Kars An ontology for a naval wargame conceptual model»

[4] Кондратьев С.А., Сычёв И.О. Применение семантических технологий для хранения и обработки данных в корабельных информационно-управляющих системах // Морская радиоэлектроника № 2(68) июнь 2019, с.38–41.

[5] Linked Data - The Story So Far - Linked Data - The Story So FarChristian Bizer, Freie Universität Berlin, GermanyTom Heath,Talis Information Ltd, United KingdomTim Berners-Lee, Massachusetts Institute of Technology, USA.

[6] Chandrasekaran B., Josephson J.R., Benjamins V.R. “What are Ontologies, and Why Do We Need Them?”, IEEE Intelligent S ystems and their Applications, vol. 14, pp. 20-26, January/February 1999.