К ВОПРОСУ СОЗДАНИЯ ИНФОРМАЦИОННО-АНАЛИТИЧЕСКОЙ СИСТЕМЫ МЕСТОРОЖДЕНИЙ РЕДКИХ МЕТАЛЛОВ

В.В.Костин, А.Н.Бездушный
В.В.Костин
Вычислительный Центр им. А.А. Дородницына РАН
kosvic11@mail.ru
А.Н.Бездушный
Вычислительный Центр им. А.А. Дородницына РАН
anb@ccas.ru

Аннотация

В работе рассматривается инициатива по реализации информационно-аналитической системы месторождений редких металлов. С этой целью совместно с прикладными специалистами были проанализированы сущности, их взаимосвязи, сопутствующие понятию месторождения. Отдельно рассмотрены специализации месторождений для редких металлов. Предложена онтологическая (owl) спецификация понятий, их свойств и связей предметной области. Реализованы хранилище данных (база данных добываемых природных ресурсов), среда визуального анализа сведений о месторождениях, позволяющая проводить OLAP анализ. По онотологии на работу с этими данными была настроена машина вывода Pellet, заданы таблицы фактов и сформулированы логические правила, что позволило пополнить массив хранимых данных вычисленными истинными фактами (по системе условий целостности). По результатам визуального OLAP анализа получен ряд достоверных фактов о соотношениях сущностей системы, сформулирован и проверен ряд гипотез о свойствах месторождений редких металлов.

Введение

На сегодняшний день правительством РФ перед министерством природных ресурсов и экологии РФ поставлена задача сбора информации по всем природным ресурсам страны в единое хранилище. Российским фондом фундаментальных исследований был проведен конкурс, один из пунктов которого нацелен на разработку элементов информационной системы, способной накапливать и анализировать соответствующие данные.

Многие прикладные заинтересованы в создании подобной системы. Каждый в своем аспекте. Специалисты по редкоземельным металлам сектора Т.Ю.Усовой [16] Института минералогии, геохимии и кристаллохимии редких элементов, сформулировавшие исходную постановку для нашей работы, наиболее актуальным вопросом считают анализ экономической эффективности добычи редких металлов (в частности, в отвалах месторождений полезных ископаемых), обусловленный их рассеянностью в земной коре, трудностями извлечения из сырья для выделения в чистом виде [16], насущными потребностями рынка[17]. Большую заинтересованность к работе проявили и оказали существенную поддержку сотрудники отдела В.М. Ряховского Музея Истории Земли им. В.И. Вернадского.

Постановка задачи

На момент постановки перед нами задачи ее инициатор Татьяна Юрьевна Усова находилась в такой ситуации, что :

1. В распоряжении сектора экономических исследований ФГУП «ИМГРЭ» сформирован определенный объем данных по редким металлам.
2. Каждое из месторождений редких металлов описывается доброй сотней разнотипных параметров.
3. Перед коллективом стоит задача поиска скрытых закономерностей между параметрами месторождений, полезными ископаемыми, на них добываемыми, и размером их запасов.
4. Анализ проводится вручную путем сопоставления таблиц Excel с данными. Попытка своими силами автоматизировать деятельность не увенчалась успехом.

Поэтому первоочередной целью стали:

1. Создание системы, позволившей ускорить процесс сравнения данных и упростить поиск закономерностей специалистом. Для решения данной задачи применена OLAP технология визуализации данных.
2. Поиск скрытых закономерностей и проверка гипотез, проведение семантического анализа и получения новых достоверных данных. Для этого использован метод создания OWL онтологии и машины логического вывода.

Данные

Данные по месторождениям редких металлов собираются разными структурами – научно-исследовательскими институтами для фундаментальных исследований и коммерческими структурами для внутренних нужд на закрытой территории. Большой объем информации уже собран государственными структурами ранее и хранится в соответствующих архивах.

В ходе работы получены данные из двух источников – от В.М. Ряховского[1] и от Т.Ю. Усовой[2]. Создана единая БД и выработана универсальная структура данных. Данные представлены на разных языках (английском и русском). База (а в последствие и хранилище) данных видоизменена таким образом, чтобы имелась возможность хранения и задания соответствия между одними и теми же данными на разных языках.

Ключевой проблемой сбора информации является отсутствие открытого доступа к информации, собранной государством. Доступ к ней закрыт, она засекречена.

OWL онтология

Оптимальным средством передачи данных в ходе работы *выбрана* OWL онтология. Совместно с прикладными специалистами на основе работ [1] и [2] проанализированы сущности, определена четкая система понятий. Выделены ключевые с точки зрения анализа сущности. Создана система свойств и подсвойств, определяющих взаимосвязи сущностей. (Рис.1) Формализованы все понятия, с целью избежания неоднозначной трактовки различных терминов. Составлены полные словари экземпляров сущностей. Отдельно рассмотрены специализации месторождений для редких металлов, выделены географическая, геологическая и геоэкономическая составляющие. Предложена онтологическая (owl) спецификация понятий, их свойств и связей предметной области. Особенность структуры онтологии – система свойств, большинство из которых соединяет определенную сущность с сущностью «месторождение», что указывает на явное выделение

[1] Владимир Михайлович Ряховский, зав. отделом информатизации Музея Истории Земли им. В.И. Вернадского.
[2] Усова Татьяна Юрьевна, зав. сектором экономических исследований Института минералогии, геохимии и кристаллохимии редких элементов.

схемы «звезда». Заданы ограничения целостности сущностей. Определены и заданы мощности связей.

Пополнение и обмен информацией

Оптимальным вариантом работы системы представляется следующая организация – центральный, ключевой, сервер с хранилищем данных и системой оперативного анализа. К данной инициативе может присоединиться специалист любой области, который, с одной стороны, может передавать новые данные, а с другой – получить доступ к уже имеющимся и результатам их анализа. Специалисты разных областей формируют разносторонний взгляд на одни и те же объекты и процессы. Такие специалисты (или группы специалистов) формируют удаленные элементы системы из имеющейся у них информации. На первоначальном этапе эта информация может храниться в любом формате – на бумажных носителях, в виде списка, в виде таблицы excel, в виде реляционной базы данных, в любом другом наиболее удобном для специалистов формате. Актуализируется вопрос о создании универсального формата хранения информации, оптимальной средой для обмена информации представляется сетевая онтология, позволяющая ссылаться на сущности из других онтологий при описании своих данных, что упрощает взаимоинтеграцию.

Каждый из специалистов проекта имеет собственное видение необходимых сущностей для решения стоящей перед ним задачи. Многие понятия трактуются по-разному различными специалистами, что требует совместное формирование системы понятий сетевой онтологии главного сервера специалистами, участвующими в проекте. Учитывая онтологическую спецификацию данных, полученная схема может при необходимости легко редактироваться.

Информация у каждого из специалистов онтологии может храниться в наиболее удобном для него формате. Для передачи ее на центральный сервер необходимо создать «адаптер» – инструмент, позволяющий преобразовать информацию из имеющегося у специалиста формата в формат сетевой онтологии центрального сервера. Схема сущностей онтологии, получившаяся при этом процессе, будет частью схемы онтологии центрального сервера.

Визуализация онтологии

Для оценки наиболее подходящих средств визуализации осуществлен анализ имеющихся на сегодняшний день решений. Проведен обзор литературы по сравнению решений*[5], [6], [7]*. На его основе выделены необходимые критерии оптимальной системы отображения данных:

1. Возможность отображать одновременно сущности, их экземпляры и свойства для возможности одновременно рассматривать онтологию целиком;
2. Возможность создавать и хранить срезы онтологии;
3. Возможность объединять онтологии;
4. Возможность одновременно отображать несколько онтологий;
5. Возможность экспортировать графическую структуру в другие форматы;
6. Возможность редактировать онтологию во время просмотра;
7. Возможность наглядно отображать связи класс-подкласс и свойство-подсвойство.

Наиболее привлекательным средством представляется надстройка Ontograf системы работы с онтологиями Protégé 4.1. Она использована в работе (Рис.1.1). Не обладая рядом возможностей – наглядно указывать связи «класс» - «подкласс» и «свойство» -

«подсвойство», удобной системой автоматического упорядочивания элементов и возможностью на графике вносить изменения в онтологию, она отвечает всем остальным требованиям. Значительную функциональность продемонстрировала написанная на языке java система RDF Gravity. Основным недостатком этой системы является неудобная система автоматического отображения структуры сущностей и экземпляров. Другие системы – OntoVisT, OWLViz, NavigOWL – не продемонстрировали столь высокой эффективности с точки зрения решения поставленных нами задач.

Хранилище данных

Используя схему сетевой онтологии центрального сервера, реализовано хранилище данных (база данных добываемых природных ресурсов). Отличительной особенностью хранилища данных является отсутствие возможности изменять данные и увеличенная скорость чтения и анализа имеющейся информации. В виду необходимости унификации и нормализации данных, задана нормативно-справочная информация – сформированы библиотеки экземпляров соответствующих сущностей.

Структурный анализ

Создан дружественный интерфейс работы со средой. Созданная система позволяет проводить визуальный анализ. Имеется возможность рассматривать загруженную в систему информацию в виде двумерных и трехмерных графиков. Задана возможность рассматривать имеющуюся информацию в виде сводных таблиц. Имеется возможность рассматривать иерархическую структуру (иерархия географии – «континент» – «государство» – «область» представлена на рис.2). Система позволяет рассматривать распределение информации в сводной таблице более чем по двум параметрам (рис.3).

Имеется возможность проведения OLAP анализа месторождений. Выделены закономерности в зависимостях расположения крупных месторождений лития и геологических характеристик. Имеется возможность графически рассматривать 1-, 2- и 3-мерные срезы куба и задавать фильтры на другие измерения. Отдельно проведен анализ в геоэкономическом аспекте, проанализирована экономическая целесообразность добычи редких элементов в отвалах других месторождений.

Также возможно применение других типов анализа – «что-если» анализ, кластерный анализ, статистический анализ и другие.

Семантический анализ

Данные преобразуются в вид «субъект – предикат – объект» для семантического анализа. Концепция представления данных в таком виде называется Resource Description Framework (RDF). Множество RDF-утверждений образует ориентированный граф, в котором вершинами являются субъекты и объекты, а рёбра помечены предикатами. Данные хранятся на языке web ontology language (OWL). Организация данных таким образом позволяет задавать логические правила связи для семантического анализа. Для проведения анализа используется машина вывода.

В систему интегрирован инструмент, позволяющий преобразовывать информацию из таблиц хранилища данных в тройки «субъект – предикат – объект» сетевой онтологии и обратно.

Опираясь на работу [3], проведен анализ имеющихся машин вывода. В онтологии для работы с этими данными была интегрирована и настроена машина вывода Pellet [4]. Заданы таблицы фактов и сформулированы логические правила. Заданы ограничения целостности.

Получены достоверные данные, основанные на ограничениях целостности. Проведен логический анализ, получены новые истинные данные (по системе условий целостности). На основе этих данных проведен визуальный OLAP анализ и получен ряд достоверных фактов о соотношениях сущностей системы.

Проведен анализ предоставленных специалистами гипотез. Гипотеза «На месторождениях с рудой гранитных пегматитов должен добываться литий» приведена к следующему логическому виду: «Если на месторождении добывается руда гранитных пегматитов, то на месторождении содержится элемент литий». В качестве результатом запуска машины вывода получена информация: «месторождение Кингс-Маунтин содержит элемент литий». Это достоверная информация, которой не было в системе до запуска машины вывода. Информация, введенная в систему, была крайне небольшого объема, поэтому погрешность ее работы крайне высока.

На основе описываемых элементов построено html приложение, позволяющее ознакомиться с возможностями системы удаленно.

Интеллектуальный анализ

Проведен анализ современных методов интеллектуального анализа данных и обзор соответствующей литературы ([9], [10], [11]):

1. Нейронные сети[14]. Этот метод заключается в создании большого числа «нейронов» - неких элементов, связывающих входные данные и результат. На основе обучающей выборки система подбирает оптимальные значения коэффициентов. Метод достаточно точно работает даже при наличии большого количества шумов. Для успешного использования метода необходим большой объем «обучающей» выборки. *Кроме того*, не представляется возможным проанализировать процессы, проходящие внутри нейронной сети. Нейронные сети позволяют с высокой точностью инициировать изображение.
2. Метод рассуждений на основе аналогичных случаев. В этой системе для прогноза будущего случая система ищет похожие случаи в уже имеющейся подборке. Недостатками такого метода является отсутствие формулировки четких правил и неоднозначность определения схожести случаев. Этот подход демонстрирует хорошие результаты при анализе большого числа повторяющихся событий. [8]
3. Метод построение дерева решений. Он использует логические переходы наподобие «если - то». Однако данный процесс не позволяет найти «лучшие» правила в данных, а многие закономерности не отслеживаются, что приводит к ненадежному анализу.

Описанные выше методы не позволяют получить максимального эффекта для решения поставленной задачи.

4. Алгоритм ограниченного перебора[13]. Он основан на подходе «что-если». Однако, максимальная длина комбинации «что-если» в правиле в системе, как правило, равна 6, и с самого начала работы алгоритма производится эвристический поиск простых логических событий, на которых потом строится весь дальнейший анализ.

Это направление выглядит перспективно с точки зрения анализа геологической информации.

5. Метод визуального анализа позволяет при помощи многомерных графиков каким-либо образом визуализировать информацию.

Данный подход реализован в рассматриваемой системе.

Реализация

На данный момент система реализована следующим образом. (Рис.4) Схема данных реализована семантической онтологией на приложении Protégé 4.1[12]. В онтологии заданы все необходимые библиотеки классификаторов. Схема данных транслирована в РСУБД MS SQL Server. Данные хранятся в базе данных Microsoft SQL Server и импортированы в хранилище данных. Хранилище реализовано при помощи MS Analysis Services. На основе этих сервисов реализован OLAP анализ. Для визуализации анализа использована программа MS Excel. Для проведения семантического анализа использована машина вывода Pellet. Онтология визуализируется надстройкой Ontograf. Создана система передачи данных из Protégé 4.1 на MS SQL Server и обратно.

Заключение

Система, созданная в ходе данной работы, показала высокую вариативность. Проведенные испытания продемонстрировали эффективность при поиске скрытых взаимосвязей характеристик месторождения и при оценке экономической эффективности добычи. В случае использования на более крупных массивах данных система позволит выделить скрытые взаимосвязи между характеристиками месторождений и спрогнозировать экономическую оправданность ведения добычи как первичных, так и вторичных полезных ископаемых.

ЛИТЕРАТУРА

[1] Т.Ю. Усова, Н.А. Архипова, Е.А. Калиш, М.Ф. Комин, Д.С. Ключарев. Редкие металлы – сырьевое обеспечение инновационных технологий// ФГУП «ИМГРЭ», Москва.

[2] Вершинин А.В., Дьяконов И.А., Ряховский В.М., Шкотин А.В. Архитектура распределенной геоинформационной среды на основе формальных онтологии пространственных данных и сервисов. «Геоинформатика», №2, 2008.

[3] J Bock, P Haase,Q Ji,Volz. Benchmarking OWL Reasoners. // In: Proc. of the ARea2008 Workshop, 2008.

[4] E. Sirin, B. Parsia, B. Cuenca-Grau, A. Kalyanpur, Y. Katz. Pellet: A practical OWL-DL reasoner. // Journal of Web Semantics, 2007.

[5] http://rcdl.ru/doc/2010/265-272.pdf

[6] Злобин А.Н. Обзор методов визуализации онтологий. // VI Всероссийская межвузовая конференция молодых ученых, 2009.

[7] http://shcherbak.net/razrabotka-vysokoeffektivnyx-sredstv-sozdaniya-i-obrabotki-ontologicheskix-baz-znanij/

[8] Айвазян С. А., Бухштабер В. М., Юнюков И. С., Мешалкин Л. Д. Прикладная статистика: Классификация и снижение размерности. — М.: Финансы и статистика, 1989.

[9] Knowledge Discovery Through Data Mining: What Is Knowledge Discovery? — Tandem Computers Inc., 1996.

[10] http://www.inftech.webservis.ru/it/database/datamining/ar2.html

[11] Кречетов Н.. Продукты для интеллектуального анализа данных. — Рынок программных средств, № 14–15, 1997, с. 32–39.

[12] http://protege.stanford.edu/
[13] http://www.intuit.ru/department/database/datamining/
[14] А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод. Технологии анализа данных. Data Mining, Visual Mining, Text Mining, OLAP. – СПб: БХВ-Петербург, 2007.
[15] Т.Ю. Усова. Редкие металлы и их месторождения. / Соросовский образовательный журнал, 2001, №11, с. 79-85.
http://window.edu.ru/window_catalog/redir?id=20916&file=0111_079.pdf
[16] А.В. Наумов Обзор мирового рынка редкоземельных металлов. Металлургия редких и благородных металлов. / 2008 г. http://www.kvarltd.ru/getfile/15.pdf
[17] Рынок редких и редкоземельных металлов 2011. Аналитический обзор. / РБК.research. Москва, 2011. http://marketing.rbc.ru/research/562949980390944.shtml

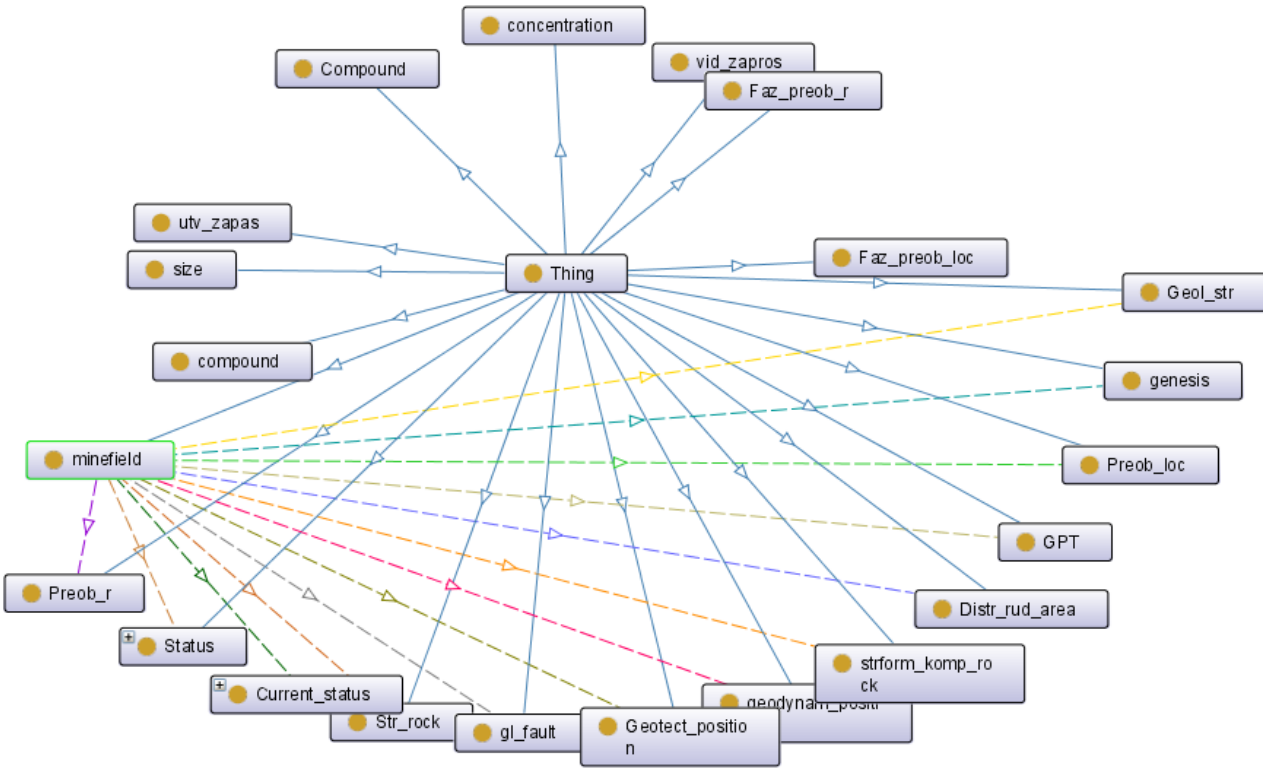


Рис.1. Схема соотношения классов онтологии

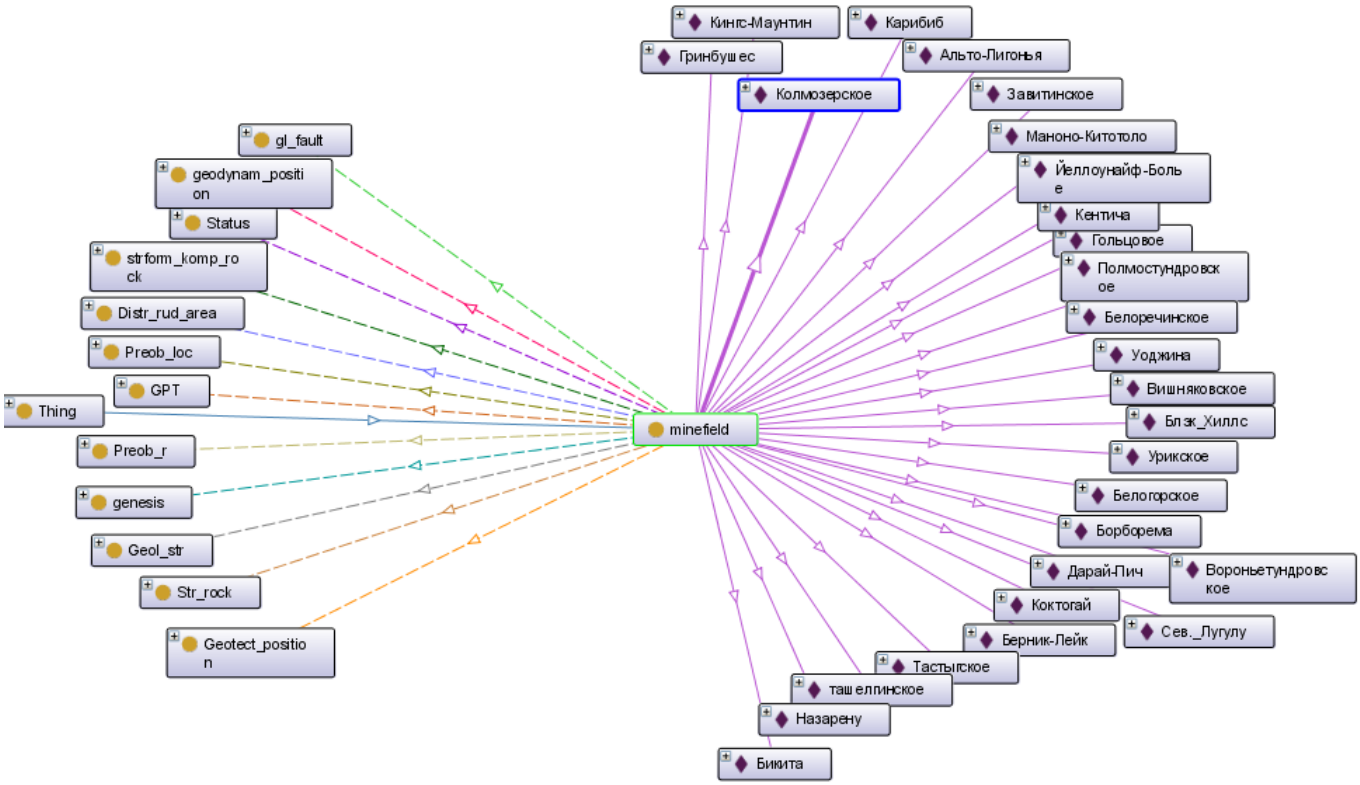
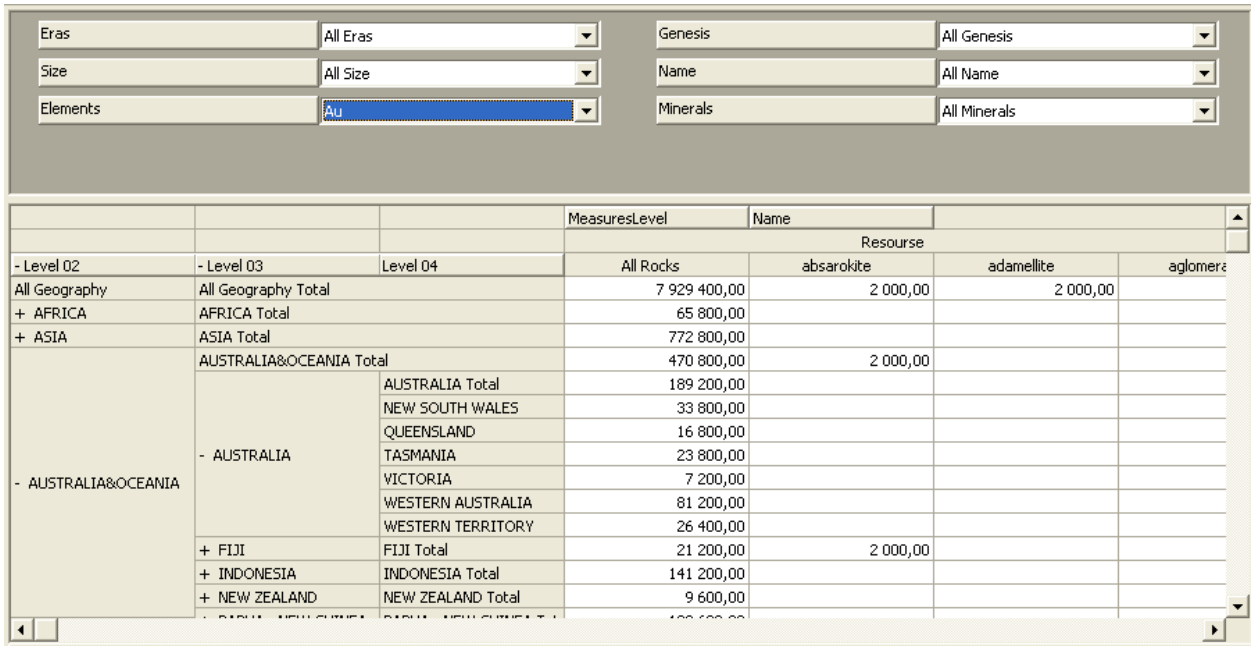


Рис.1.1. Связи сущности «месторождение», программа визуализации онтологий Ontograf.



Eras	All Eras	Genesis	All Genesis
Size	All Size	Name	All Name
Elements	Au	Minerals	All Minerals

			MeasuresLevel	Name		
			Resourse			
- Level 02	- Level 03	Level 04	All Rocks	absarokite	adamellite	aglomera
All Geography	All Geography Total		7 929 400,00	2 000,00	2 000,00	
+ AFRICA	AFRICA Total		65 800,00			
+ ASIA	ASIA Total		772 800,00			
- AUSTRALIA&OCEANIA	AUSTRALIA&OCEANIA Total		470 800,00	2 000,00		
	- AUSTRALIA	AUSTRALIA Total	189 200,00			
		NEW SOUTH WALES	33 800,00			
		QUEENSLAND	16 800,00			
		TASMANIA	23 800,00			
		VICTORIA	7 200,00			
		WESTERN AUSTRALIA	81 200,00			
		WESTERN TERRITORY	26 400,00			
	+ FIJI	FIJI Total	21 200,00	2 000,00		
	+ INDONESIA	INDONESIA Total	141 200,00			
	+ NEW ZEALAND	NEW ZEALAND Total	9 600,00			

Рис.2. Двумерный срез 8-мерного куба по измерениям «география» и «породы». Измерение «география» представлено в виде иерархической структуры.

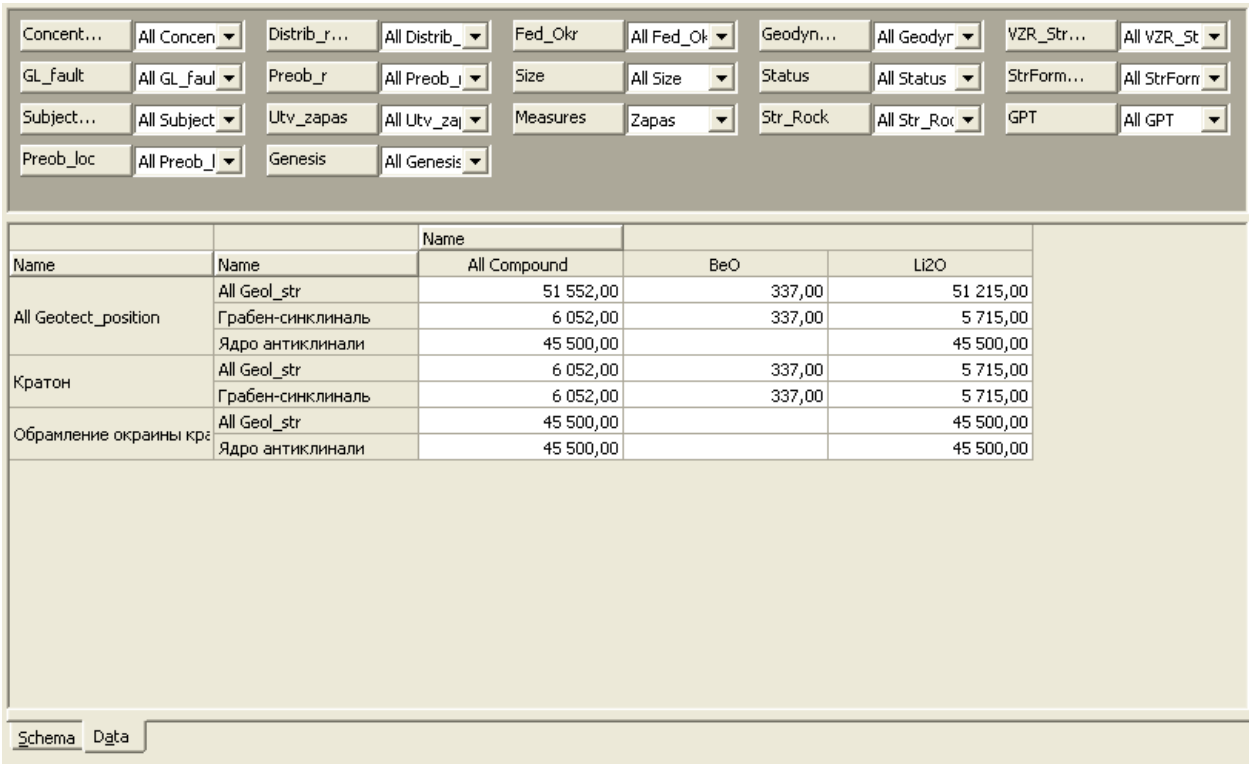


Рис.3. Трехмерный срез 20-мерного куба по измерениям «геотектоническая позиция», «геологическая структура» и «соединения».

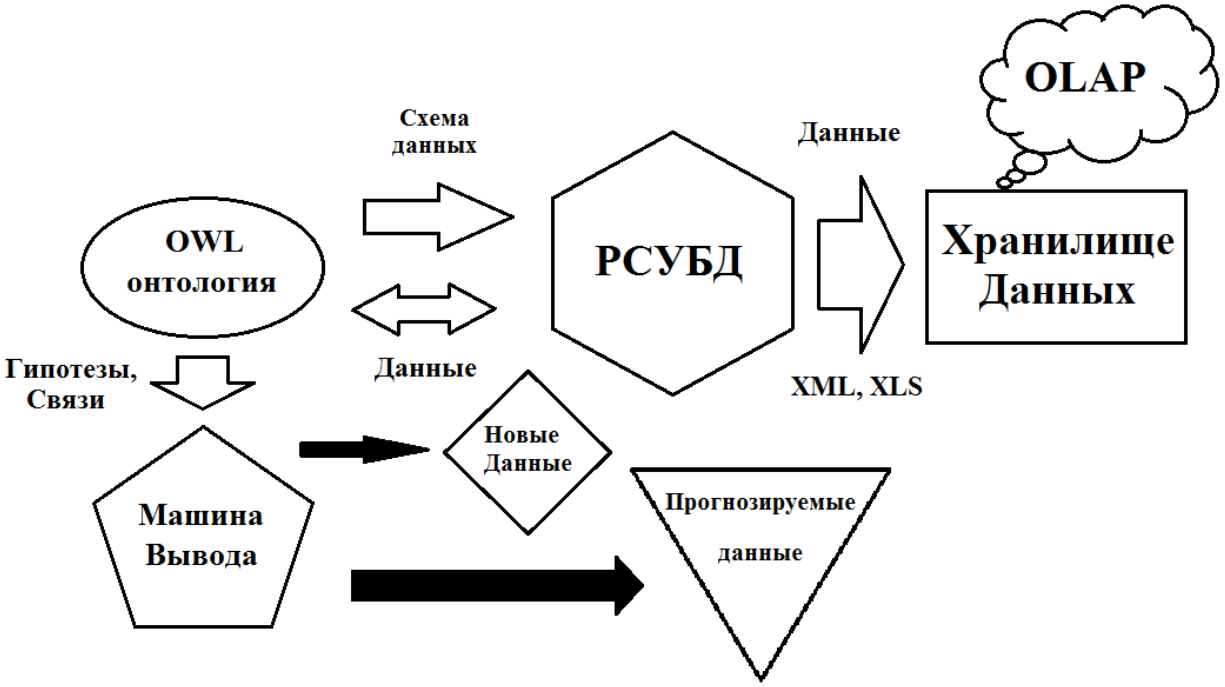


Рис.4. Структура реализованной на данный момент системы.