GEOLOGICAL ONTOLOGIES

FOR DESCRIPTION AND CLASSIFIACTION OF OBJECTS, STRUCTURES

AND MAPPED ROCK UNITS

IN GEOLOGIC AND CARTOGRAPHIC DATABASES

Vicror Snezhko,

A.P. Karpinsky Russian Geological Research Institute (VSEGEI), St. Petersburg, Russian Federation

Email Viktor_Snezhko@vsegei.ru

Ontologies in geology are based on formal descriptions made in a controlled natural language. They are used for unification of description of geological objects, structures and mapped rock units, enhancement of search engines in thematic databases and correlation of geological map legends across large regions on the national and international scale. For the latter, the interoperability of ontologies is the key principle. On the national scale in Russia, one of the most eminent results is the correlation between the sheets of the State Geological Map scaled 1:1000000, and the main international dimension of work is the interoperability of the domestic ontologies with the GeoSciML (GeoScience Markup Language) standards.

Назначением системы онтологий является обеспечение поиска геологической и картографической информации в базе данных при помощи запросов, построенных на использовании формализованных описаний геологических подразделений и их свойств. В результате запроса из БД может быть получен перечень подразделений, удовлетворяющий условиям, определенным в запросе (возраст, таксон районирования, литологический состав, генезис и т.д.). Полученные результаты могут быть непосредственно отражены на карте, а могут использоваться для дальнейшего анализа, обобщения или пере интерпретации геологической информации - проведение межсерийной корреляции, построение обобщенных легенд, построение легенд по другому основанию классификации (легенда формационной схемы, схемы геодинамических комплексов и т.п.).

В предложенном выше определении необходимо четко понимать смысл, вкладываемый нами понятия - «формализованные описания», «эффективный поиск» и «онтологии».

Формализованные описания – это описания какого-либо геологического подразделения (толщи, свиты, комплекса, например), вещества его слагающего (известняк мраморизованный, алевролит, гранодиорит), свойств (генетический тип, степень консолидации и т.п.) в терминах (или наборе терминов) имеющихся в наборе словарей, используемых для построения информационной системы. Эти термины являются неотъемлемой частью системы, имеют однозначное написание, словарное определение и обязательную ссылку на источник заимствования определения.

Таким образом, формализованные описания переходят в ранг «контролируемых понятий», т.е. понятий которые используются только в том значении, которое формально определено для них в создаваемой системе или в стандарте обмена данными (в виде текстового описания из источника заимствования и формализованного описания в определенном формате), а уникальность таких понятий самостоятельно отслеживается системой либо определяется стандартом обмена (т.е. не может быть двух терминов «песчаник» с различным определением из различных источников). Важно отметить, что при таком подходе и само геологическое подразделение тоже становится термином - «контролируемым понятием», имеющим однозначное определение в виде текстового описания и ссылку на источник этого описания (серийная легенда, например). Поэтому геологическое подразделение (как «контролируемое понятие») может быть охарактеризовано определенным набором терминов (других «контролируемых понятий», используемых для описания его ранга, возраста, мощности, слагающих его пород и т.д.). Эти термины, в свою очередь так же могут быть описаны в виде набора терминов, характеризующих те или иные их свойства. Например, термин «андезит» может быть охарактеризован терминами, отражающими его генезис - «вулканический генезис», его химический состав – «силикат средний» и «силикат нормальнощелочной» и т.д. (согласно Петрографическому кодексу как источнику заимствования определения термина).

Контролируемые понятия могут быть связаны между собой различными геологическими отношениями (эквивалент, частичный эквивалент, вмещает, прорывает и т.п.). При построении запроса, можно указывать на тип обрабатываемой связи (например, учитывать или не учитывать подразделения, связанные по типу «частичный эквивалент»), в то же время система различает все эти понятия (как имеющие различный ранг – свита, толща и серия в авторских определениях).

Эффективный поиск информации. Создание информационной системы «национального» уровня, содержащей десятки-сотни тысяч описаний геологических подразделений, должно обеспечиваться адекватными механизмами поиска информации. Традиционные средства контекстного поиска, основанные на поиске по вхождению слова в текстовое описание, при таком объеме информации становятся не эффективными и зачастую не обеспечивают адекватного выбора информации по запросу пользователя. Одно из перспективных направлений развития информационно-поисковых систем - построение моделей «семантического», т.е. «смыслового» поиска - поиска ресурсов, наиболее релевантных запросу, а не просто содержащих слова из запроса. Попытки реализации такого подхода предпринимаются с конца 20 века. За прошедшее время было предложено несколько методик, наиболее эффективной в области описания геологической информации является модель, основанная на онтологии предметной области [Соловьев В.Д, 2006].

Онтологии включают доступные для компьютерной обработки определения основных понятий и объектов предметной области, а так же формальные описание их иерархии, свойств и связей между ними [Соловьев В.Д, 2006]. Не углубляясь далее в развитие темы онтологий, отметим, что создание онтологии – это одна из наиболее наукоемких и трудозатратных задач в создании поисковой системы. Как правило, создание онтологий выполняется экспертами в данной предметной области на основе нормативно-методических и справочных документов, содержащих как точное определение термина, так и определяющие правила, обеспечивающие возможность его группировки с другими терминами, а в конечном счете - определяющими его место в принятой иерархии.

В качестве оценки эффективности подходов, основанных на различных способах поиска информации, можно привести пример запроса к базе данных, содержащей геологические карты масштаба 1:1 000 000 по 27 номенклатурным листам, с целью выбора подразделений, содержащих породы плутонического генезиса. В случае построения такого запроса по методу «прямого» контекстного поиска – в самом запросе необходимо будет вручную перечислить 537 названий пород (согласно перечню пород плутонического генезиса, содержащихся в описаниях геологических подразделений на 27 номенклатурных листах) – и нет никакой уверенности, что геолог сможет вспомнить все эти литологические разности; в случае же поиска по методу «онтологий» ему необходимо будет выбрать в описании подразделения поле «характеристика породы по генезису», а в нем установить значение «плутонический генезис». Преимущество метода «онтологий» не вызывает сомнения.

Таким образом, создание полноценного набора терминов (в виде «контролируемых понятий») составляющих словари геологических понятий, увязка их между собой и формирование из этих терминов наборов онтологий является важнейшей задачей при создании любой, достаточно большой базы данных геологических карт, поскольку создает единую, унифицированную систему описания геологической информации.

Принцип интероперабельности онтологий (interoperability). Создание интероперабельной системы онтологий и терминологической базы в виде «контролируемых понятий» открывает путь к возможности международного обмена геологической информацией не только на уровне физической сопоставимости объектов (точка, линия, полигон), достигаемой через унификацию форматов цифровых материалов, но и на содержательном уровне – через сопоставление используемых онтологий и корреляции терминов на уровне «контролируемых понятий».

Одной из наиболее интересных задач сегодняшнего дня является реализация такого сопоставления и корреляции с терминологической базой и структурой международного языка GeoSciML (GeoScience Markup Language). Проект GeoSciML развивается с 2003 года под эгидой Commission for the Management and Application of Geoscience Information, которая входит в International Union of Geological Sciences. В настоящий момент для GeoSciML реализована уже 3 версия, которая включает более 30 словарей по различным аспектам геолого-картографической терминологии. Спецификации GeoSciML используются многими геологическими службами, в том числе Австралии, США, Канады, Франции, Италии, Словении. Рассматривается вопрос об использовании GeoSciML как унифицированной основы цифровых геолого-картографических материалов в рамках пан-Европейской инициативы INSPIRE и для создаваемой в настоящее время пан-Африканской инфраструктуры «AEGOS Project». Таким образом, корреляция российской терминологической основы (онтологии и «контролируемые понятия») с терминологической базой GeoSciMLоткрывает возможность для проведения содержательной интеграции геолого-картографической информации по огромным территориям.