Построение системы онтологий для предметной области “Петрология”

(Формальное представление знаний в петрологии.v2)

Александр Шкотин, В. Ряховский, Д. Кудрявцев

Отдел ГИС
Государственный Геологический Музей им. В.И. Вернадского РАН
www.sgm.ru

ashkotin@acm.org

Формальная онтология это

Представление знаний текстом

- в узком смысле - на OWL 2,
- в широком смысле - на одном из языков математической логики.

Содержание

- Введение
- Формализация фактов
- Формализация словаря
- Элементы формальной теории
- Выводы и дальнейшие планы
- Благодарности
- Ссылки

Введение

Петрология - наука, изучающая горные породы и условия их образования.

Играет важную роль при описании строения частей земной коры и при выявлении закономерностей в расположении полезных ископаемых.

В настоящее время накоплен большой объем петрологической информации, которая нуждается в систематизации, интеграции и поддержке в согласованном состоянии.

Эти задачи могут быть решены путем формализации знаний.

Формализация знаний

Конечная цель: создании формальной теории, объединяющей ключевые понятия петрологии и отношения между ними.
Основополагающую роль в процессе создания играют онтологии - знания, организованные на основе математической логики.
Определения играют решающую роль в изложении теории, т.к. задают именно те понятия, свойства которых будут использоваться и изучаться.

Формальная теория. Назначение

Построение формальной теории области естественных знаний по образцу математической является уточнением формы знаний до математического уровня точности.
Это (как всегда) приводит к выявлению неточностей, устранению неясностей и скрытых противоречий неформальных знаний.

Формализация даёт возможность многие свойства определений терминов (например, противоречивость или то что два термина взаимно исключающи) проверять универсальными алгоритмами.

Формальная теория. Построение

Область «пробега» переменных - любое твёрдое тело, в котором может быть жидкая и газовая фазы.

Область значений некоторых функций - десятичные числа или строки.

Построение:

- Выявление предикатов и функций.

- Поиск определений одних и «назначение» первичными (не определяемыми в данной теории) других.

- Выявление аксиом: правил, свойств, законов, закономерностей предметной области.

Правила вывода такие же как в математической логике.

Формализация фактов

Важнейшие термины предметной области - термины применяющиеся при записи фактов.
В настоящий момент научные факты сосредоточены в основном в базах данных.
Задача: Необходимо уметь извлекать знания из баз данных.
Результат: БД Proba, хранящая сведения об образцах горных пород, преобразована в OWL-онтологию фактов.
Был выявлен состав терминов необходимых для записи фактов.

Онтология словаря

Определения терминов накоплены в специализированных словарях. Различные научные школы, направления могут иметь отличающиеся определения.

Задача: выявление состава и определений терминов, а также не определяемых в данной предметной области, первичных терминов.

Результат: «Словарь терминов изверженных горных пород» преобразован в OWL-онтологию. Начато формальное описание отношений между терминами (например, синонимия).

Концентратор определений

Цели:

- собрать в одном месте разные определения,

- дать возможность экспертам решить какое из них надо формализовать.

Нужна веб система класса wiki.
Система webProtege позволяет хранить и обрабатывать формальные определения, которые являются нашей целью.

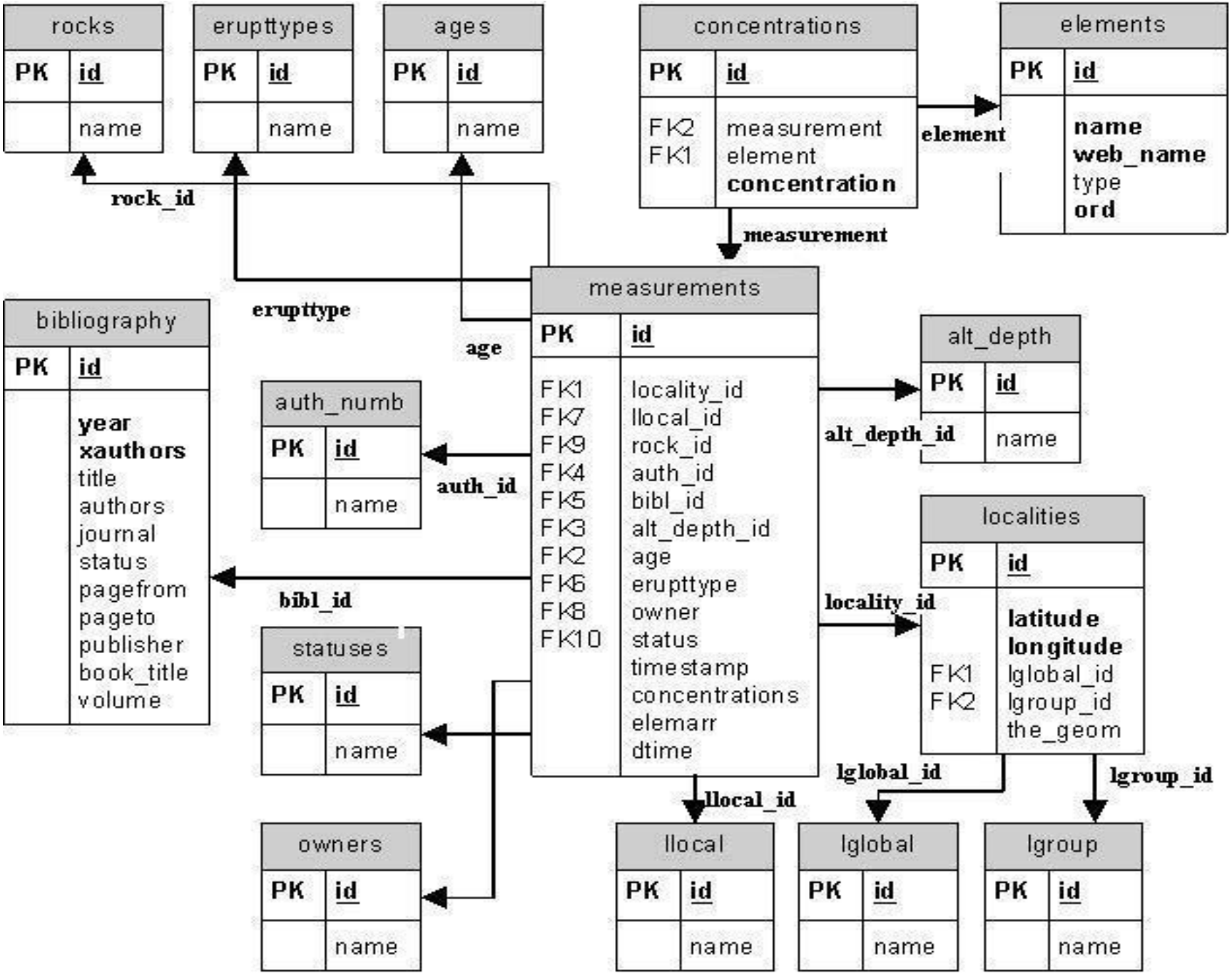
Формальные определения

Для получения точных определений и их формализации необходимо использовать правила классификации, описываемые в методиках, рекомендациях, стандартах.
Были взяты «Igneous Rocks: A Classification and Glossary of Terms» - рекомендации IUGS по классификации образцов горных пород.
Правила классификации были уточнены до алгоритма.
Из алгоритма были получены формальные определения.

Формализация фактов

Подход

Базы данных не представляют собой знаний.
Они требуют существенной и кропотливой переработки, чтобы получить знания.
Прямой путь получения знаний из данных: преобразование БД в традиционную форму знаний - знания на естественном языке.
Естественный язык ограничивается до ОЕЯ.
ОЕЯ есть универсальное средство представления формальных знаний.
Об ОЕЯ см. http://ru.wikipedia.org/wiki/CNLs


rocks
PK id
name
erupttypes
PK id
name
ages
PK id
name
concentrations
PK id
FK2 measurement
FK1 element
concentration
element
elements
PK id
name
web_name
type
ord
rock_id
measurement
erupttype
age
bibliography
PK id
year
xauthors
title
authors
journal
status
pagefrom
pageto
publisher
book_title
volume
measurements
PK id
FK1 locality_id
FK7 llocal_id
FK9 rock_id
FK4 auth_id
FK5 bibl_id
FK3 alt_depth_id
FK2 age
FK6 erupttype
FK8 owner
FK10 status
timestamp
concentrations
elemarr
dtime
alt_depth
PK id
name
alt_depth_id
auth_numb
PK id
name
auth_id
localities
PK id
latitude
longitude
FK1 lglobal_id
FK2 lgroup_id
the_geom
locality_id
bibl_id
statuses
PK id
name
lglobal_id
lgroup_id
llocal_id
owners
PK id
name
llocal
PK id
name
lglobal
PK id
name
lgroup
PK id
name

Описание БД Proba

БД Proba содержит сведения из 1174 научных статей (таблица bibliography) о 49285 образцах магматических пород (таблица measurements). Образцы собраны по всему земному шару, что отражено в таблицах localities, llocal, lglobal, lgroup. Образцам приписаны: вид породы (таблица rocks), вид происхождения (errupttypes), возраст (ages) и главное: весовое процентное содержание (concentrations) химических веществ и изотопов (список в таблице elements).

Проблемы и задачи

Идентификаторы таблиц и колонок лишь приблизительно соответствуют терминам употребляемым петрологами при обмене информацией об образцах.

Задача: превратить данные накопленные в РБД в знания в форме непосредственно понятной специалистам предметной области.

ОЕЯ предложения. Подход

Создать шаблоны всех видов ОЕЯ предложений чтобы представить все факты, содержащиеся в БД Проба.

Использовать локальные («внутренние») имена собственные.

Слова в сложных терминах соединить буквой «_».

В тексте могут встретиться и глобальные общепринятые имена собственные, например: Iceland, Atlantic_Ocean.

В остальном мы имеем простые понятные предложения английского языка.

OEЯ предложения. Пример

PUB5633 is a publication.
A title of PUB5633 is "A CONTRIBUTION TO THE GEOLOGY OF THE K...".
SAM32994 is a sample. SAM32994 is a rhyolite.
PUB5633 describes SAM32994.
PLC32994 is a place. PLC32994 is a part of Iceland.
A gathering_place of SAM32994 is PLC32994.
SUB469812 is a substance. SAM32994 includes SUB469812.
WPC469812 is a weight_percent. A value of WPC469812 is 73.95.
A component of WPC469812 is SUB469812.

Промежуточные итоги

Для записи фактов, содержащихся в БД нужен очень ограниченный естественный язык.

Получилось так потому, что БД нормализована.

Но не везде. Доведение нормализации до конца есть одна из задач большой и кропотливой работы по приведению БД к состоянию в котором автоматическое преобразование в знания возможно.

Были разработаны правила отображения содержимого БД в ОЕЯ.

Эти правила являются спецификацией для SQL-скриптов выгружающих БД в ОЕЯ текст [otch08].

OWL онтология

Все порождаемые предложения являются высказываниями ОЕЯ - ACE.
http://en.wikipedia.org/wiki/Attempto_Controlled_English
Предложения таковы, что транслятор APE транслирует их в OWL.
БД преобразуется в 1174 онтологии.
Значения колонок формируют в основном значения атрибутов, но некоторые - названия классов (rhyolite, harzburgite) и предметов (Iceland).

OWL 2. -1-

Язык формальной записи знаний.
OWL_Working_Group

Основные элементы языка:

- литералы - текстовые константы различной структуры,
- индивиды - предметные константы и переменные,
- классы - унарные предикаты,
- объектные свойства и свойства данных - бинарные предикаты и функции.

OWL 2. -2-

Основные конструкции языка:

- операторные выражения дающие свойство, класс, «диапазон данных»;

- аксиомы - **высказывания** об отношениях классов, свойств, индивидов.

OWL 2. Пример

http://attempto.ifi.uzh.ch/ape/

Prefix(:=<http://attempto.ifi.uzh.ch/ontologies/owlswrl/test#>)

Ontology(<http://attempto.ifi.uzh.ch/ontologies/owlswrl/test>

ClassAssertion(:publication :PUB5633)

DataPropertyAssertion(:title :PUB5633 "A CONTRIBUTION TO THE GEOLOGY OF THE K..."^^<http://www.w3.org/2001/XMLSchema#string>)

ClassAssertion(:sample :SAM32994)

ClassAssertion(:rhyolite :SAM32994)

ObjectPropertyAssertion(:describe :PUB5633 :SAM32994)

ClassAssertion(:place :PLC32994)

OWL 2. Пример. окончание

ObjectPropertyAssertion(:part :Iceland :PLC32994)

ObjectPropertyAssertion(:gathering_place :SAM32994 :PLC32994)

ClassAssertion(:substance :SUB469812)

ObjectPropertyAssertion(:include :SAM32994 :SUB469812)

ClassAssertion(:weight_percent :WPC469812)

DataPropertyAssertion(:value :WPC469812 "73.95"^^<http://www.w3.org/2001/XMLSchema#double>)

ObjectPropertyAssertion(:component :WPC469812 :SUB469812)

)

OWL-онтология № 5633

- Classes:

place, publication, rhyolite, sample, substance, weight_percent.

- Data properties: authorial_number, chemical_formula, first_page, latitude, longitude, reference, title, value, year.

- Object properties:

component, describes, gathering_place, includes, mixture, part.

- Individuails: Iceland, PLC32994..., PUB5633, SAM32994..., SUB469812..., WPC469812...

Усмотрение

Все использованные термины, кроме rhyolite, относятся к контекстам смежным с петрологией и даже геологией. Таковы контексты: географии (place...), научных публикаций (publication...), твёрдого тела (sample, substance, weight_percent...), химии (chemical_formula). В дальнейшем мы сосредоточимся на получении определений для видов горных пород, в том числе для rhyolite.

OWL-онтология словаря
и
Концентратор определений

Толковый словарь

- важный и специфический вид знаний.
Он содержит номенклатуру терминов предметной области и неформальные определения этих терминов. Неформальные определения даются экспертами, которые обычно принадлежат к одной из научных школ.
Задачи:
- преобразовать толковый словарь в формальные знания,
- собрать вместе определения различных школ.

Пример статьи словаря

HARZBURGITE. An ultramafic plutonic rock composed essentially of olivine and orthopyroxene. Now defined modally in the ultramafic rock classification (Fig. 2.9, p.28). (Rosenbusch, 1887, p.269; Harzburg, Harz Mts, Lower Saxony, Germany; Tröger 732; Johannsen v.4, p.438; Tomkeieff p.247)

[IRCGT], p.88

От текста словаря к онтологии

Словарь

"*Словарь терминов изверженных горных пород*". 1567 статей, подавляющее большинство которых является наименованиями горных пород.

Владелец

Межведомственный петрографический комитет при ОНЗ РАН.

Текст

http://www.igem.ru/site/petrokomitet/slovar.htm

Преобразования текста словаря в текст OWL-онтологии

Лексика

«_», «pele_s_hair»

Заголовок статьи

Термин → класс

Синонимия (3179 классов и 1659 аксиом эквивалентности классов)

Текст статьи

определение термина, комментарий, список ссылок на литературу, описание происхождения термина.

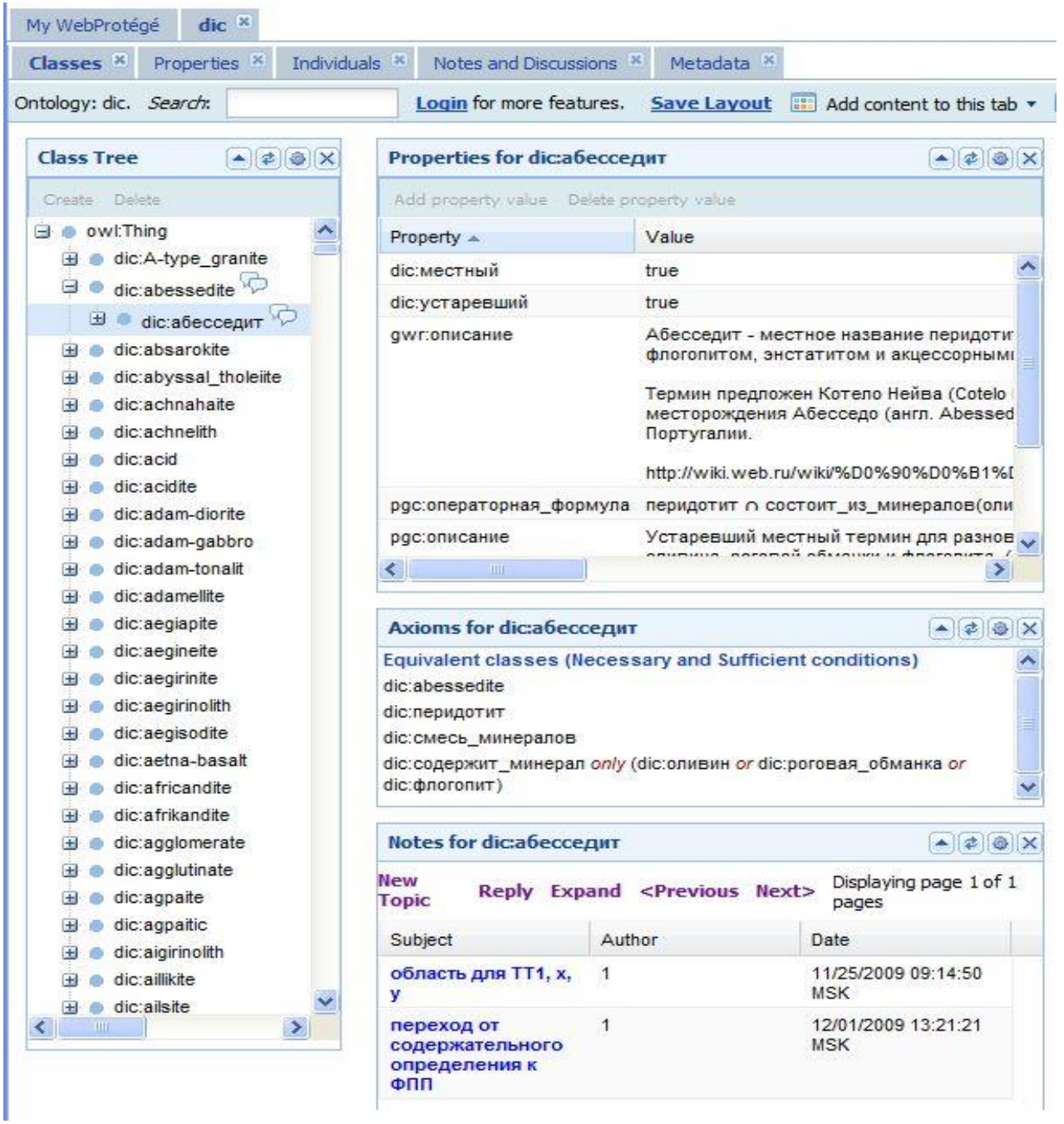
Доступ к OWL-онтологии словаря

Основные способы использования:
- программами, например для импорта её в онтологию фактов. А также для запросов, например, о наличии и характеристиках того или иного термина, его определении;
- людьми, для просмотра и обсуждения определений;
- специалистам для редактирования онтологии.

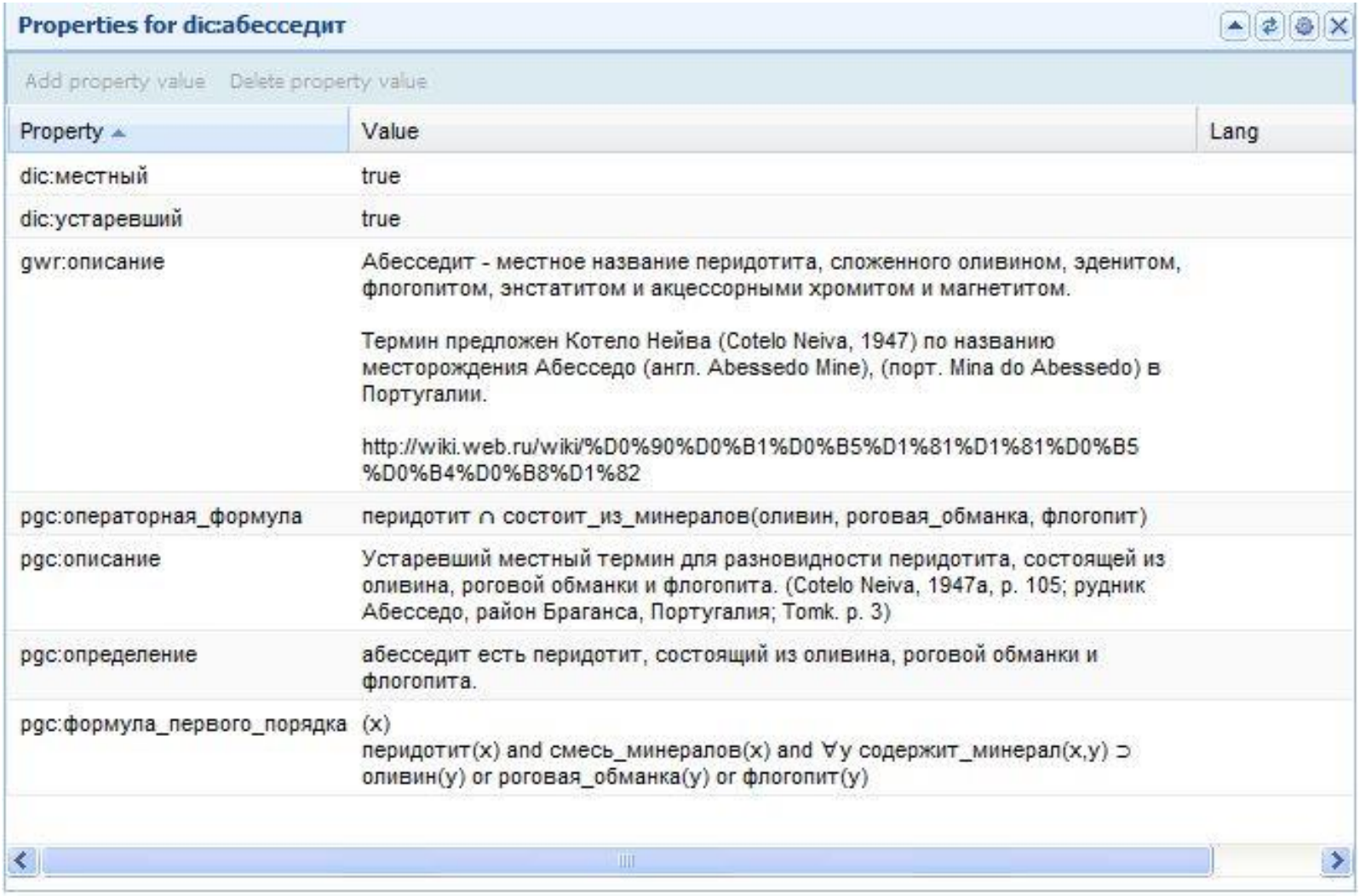
Адрес: http://earth.jscc.ru/ontologies/dic.owl

Концентратор определений

Цель - коллективное ведение определений научных терминов, включая формальные определения.
OWL-онтология словаря изверженных горных пород заложена в webProtege, имя dic.
Некоторые термины пополнены определениями из других словарей.
Адрес на портале Геология:
http://earth.jscc.ru/webprotege/


My WebProtégé
dic
Classes
Properties
Individuals
Notes and Discussions
Metadata
Ontology: dic. Search:
Login for more features.
Save Layout
Add content to this tab
Class Tree
Create
Delete
owl:Thing
dic:A-type_granite
dic:abessedite
dic:абесседит
dic:absarokite
dic:abyssal_tholeiite
dic:achnahaite
dic:achnelith
dic:acid
dic:acidite
dic:adam-diorite
dic:adam-gabbro
dic:adam-tonalit
dic:adamellite
dic:aegiapite
dic:aegineite
dic:aegirinite
dic:aegirinolith
dic:aegisodite
dic:aetna-basalt
dic:africandite
dic:afrikandite
dic:agglomerate
dic:agglutinate
dic:agpaite
dic:agpaitic
dic:aigirinolith
dic:aillikite
dic:ailsite
Properties for dic:абесседит
Add property value
Delete property value
Property
Value
dic:местный
true
dic:устаревший
true
gwr:описание
Абесседит - местное название перидоти
флогопитом, энстатитом и акцессорными
Термин предложен Котело Нейва (Cotelo
месторождения Абесседо (англ. Abessed
Португалии.
http://wiki.web.ru/wiki/%D0%90%D0%B1%D
pgc:операторная_формула
перидотит ∩ состоит_из_минералов(оли
pgc:описание
Устаревший местный термин для разнов
Axioms for dic:абесседит
Equivalent classes (Necessary and Sufficient conditions)
dic:abessedite
dic:перидотит
dic:смесь_минералов
dic:содержит_минерал only (dic:оливин or dic:роговая_обманка or dic:флогопит)
Notes for dic:абесседит
New Topic
Reply
Expand
<Previous
Next>
Displaying page 1 of 1 pages
Subject
Author
Date
область для TT1, x, y
1
11/25/2009 09:14:50 MSK
переход от содержательного определения к ФПП
1
12/01/2009 13:21:21 MSK

Свойства термина



Properties for dic:абесседит

Add property value Delete property value

Property	Value	Lang
dic:местный	true	
dic:устаревший	true	
gwr:описание	Абесседит - местное название перидотита, сложенного оливином, эденитом, флогопитом, энстатитом и акцессорными хромитом и магнетитом. Термин предложен Котело Нейва (Cotelo Neiva, 1947) по названию месторождения Абесседо (англ. Abessedo Mine), (порт. Mina do Abessedo) в Португалии. http://wiki.web.ru/wiki/%D0%90%D0%B1%D0%B5%D1%81%D1%81%D0%B5%D0%B4%D0%B8%D1%82	
pgc:операторная_формула	перидотит ∩ состоит_из_минералов(оливин, роговая_обманка, флогопит)	
pgc:описание	Устаревший местный термин для разновидности перидотита, состоящей из оливина, роговой обманки и флогопита. (Cotelo Neiva, 1947a, p. 105; рудник Абесседо, район Браганса, Португалия; Tomk. p. 3)	
pgc:определение	абесседит есть перидотит, состоящий из оливина, роговой обманки и флогопита.	
pgc:формула_первого_порядка	(x) перидотит(x) and смесь_минералов(x) and ∀y содержит_минерал(x,y) ⊃ оливин(y) or роговая_обманка(y) or флогопит(y)	

Пространства имён

Для терминов нашего словаря, терминов самой онтологии, терминов Геовеб портала МГУ и терминов Петрографического кодекса России соответственно:

prefix pgc: <http://www.igem.ru/site/petrokomitet/slovar#>
prefix dic: <http://earth.jscc.ru/ontologies/dic.owl#>
prefix gwr: <http://wiki.web.ru/wiki#>
prefix pgcc: <http://www.igem.ru/site/petrokomitet/code#>

Заключение-1

Петрографический кодекс России (задачи):

- "...создание формализованной терминологической базы для широкого обмена информацией." Статья XI.1.
- "...в широком масштабе организовать формализацию петрографической информации" Статья XI.5.

Отчёт [otch09] содержит подробную информацию о проделанной работе.

Элементы формальной теории

Первоисточники

Рекомендации IUGS [IRCGT] и уточняющий отчёт BGS [BGSRCS] описывают:

- правила начальной классификации,
- правила дальнейшей классификации в рамках выявленных свойств,
- диаграммы окончательной классификации по процентному содержанию существенных минералов

т.е. содержат алгоритм классификации.

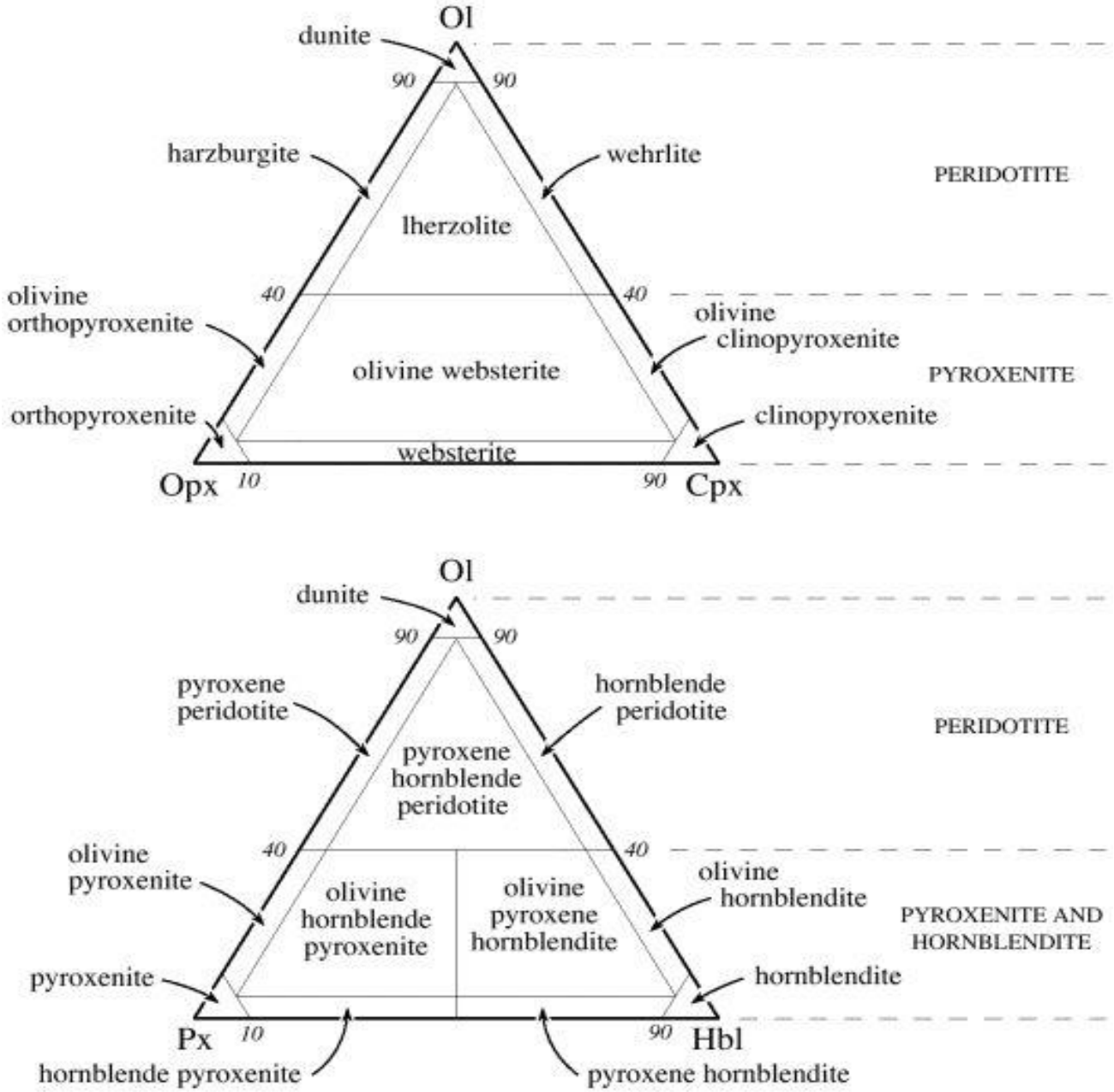


Fig. 2.9. Modal classification of ultramafic rocks based on the proportions of olivine (Ol), orthopyroxene (Opx), clinopyroxene (Cpx), pyroxene (Px) and hornblende (Hbl) (after Streckeisen, 1973, Figs. 2a and 2b).

Требования к системе определений

- Определения видов горных пород взаимоисключающи.
- Совокупность определений полна: всякий объект предметной области подпадает под какое-то из определений.

«Первичные» предикаты и функции

Предикаты (унарные кроме part_of)

part_of; clast, clastic, volcanic_eruption_result;

kimberlite, lamproite, lamprophyre, charnockite, plutonic, volcanic;

melilite, kalsilite, leucite, Ol, hornblende, garnet, spinel, biotite;

Opx, Cpx;

Q, A, P, F.

VPC означает Volume Percentage Content - объёмное процентное содержание, обычно минерала в образце, известное также как модальное содержание.

Формально VPC() – оператор: на входе предикат вещества, на выходе числовая функция.

Определения предикатов и функций

Предикаты

Группировки минералов диаграмм
Px(x) = Opx(x) ∨ Cpx(x)
OOC(x) = Ol(x) ∨ Px(x)
OPH(x) = Ol(x) ∨ Px(x) ∨ hornblende(x)

pyroclastic(x) = clastic(x) ∧
(∀y **part_of**(y,x) ∧ clast(y) → volcanic_eruption_result(y))

Функции

Показатель мафичности
M(x) = 100 - (VPC(Q)(x)+VPC(A)(x)+VPC(P)(x)+VPC(F)(x))

Определение предиката вида горной породы harzburgite

harzburgite(x) = plutonic(x) ∧
¬(pyroclastic(x) ∨ kimberlite(x) ∨ lamproite(x) ∨
lamprophyre(x) ∨ charnockite(x))

∧ VPC(carbonates)(x)≤50 ∧ VPC(melilite)(x)≤10 ∧
VPC(M)(x) ≥ 90 ∧ VPC(kalsilite)(x)=0 ∧
VPC(leucite)(x)=0 ∧ VPC(hornblende)(x)=0 ∧

0.4*VPC(OOC)(x)≤VPC(Ol)(x)≤0.9*VPC(OOC)(x)
∧ VPC(Cpx)(x)<0.05*VPC(OOC)(x)

Обсуждение формулы

Формальное определение вида магматической горной породы harzburgite состоит из трёх частей:

1. качественные характеристики,
2. абсолютные ограничения на минеральный состав,
3. относительные ограничения на минеральный состав.

Формальное определение ничего не подразумевает и не содержит ссылок на диаграмму.
Оно содержит в себе нужную часть диаграммы.

Свойства системы предикатов

Каждый предикат содержат совокупность условий - систему линейных неравенств (СЛН). Система предикатов в целом обладает важными математическими свойствами:

- каждые две СЛН несовместны, т.к. образец не должен одновременно «принадлежать» двум предикатам.
- объединение всех условий вместе даёт «опорную формулу» без линейных неравенств.

Опорная формула для ультрамафических пород

$$UM(x) =$$

$$plutonic(x) \wedge$$

$$\neg(pyroclastic(x) \vee kimberlite(x) \vee lamproite(x) \vee lamprophyre(x) \vee charnockite(x))$$

$$\wedge\ VPC(carbonates)(x) \leq 50 \wedge VPC(melilite)(x) \leq 10 \wedge VPC(M)(x) \geq 90 \wedge VPC(kalsilite)(x) = 0 \wedge VPC(leucite)(x) = 0$$

Применение машин вывода

Описанные свойства могут быть проверены автоматически, если определения загрузить в машину вывода (МВ) - reasoner, работающую с линейными неравенствами.

И такие МВ есть (например, Racer), т.к. запись линейных неравенств возможна используя расширение OWL 2 [OWL2LE].

Powerful Reasoning over OWL ontologies

- **Consistency:** determines whether the ontology contains contradictions**.**
- **Satisfiability**: determines whether classes can have instances.
- **Subsumption**: is class C1 implicitly a subclass of C2?
- **Classification**: repetitive application of subsumption to discover implicit subclass links between named classes
- **Realization**: find the most specific class that an individual belongs to.

http://www.slideshare.net/micheldumontier/formal-representation-of-models-in-systems-biology

Эквивалентность классов -1-

гарцбургит =

ОПС_клинопироксен some decimal [>0,<10]

and ОПС_оливин some decimal [>40,<90]

and ОПС_ортопироксен some decimal [>10,<60]

and ОПС_хромшпинелид some decimal [>0,<5]

Эквивалентность классов -2-

harzburgite =

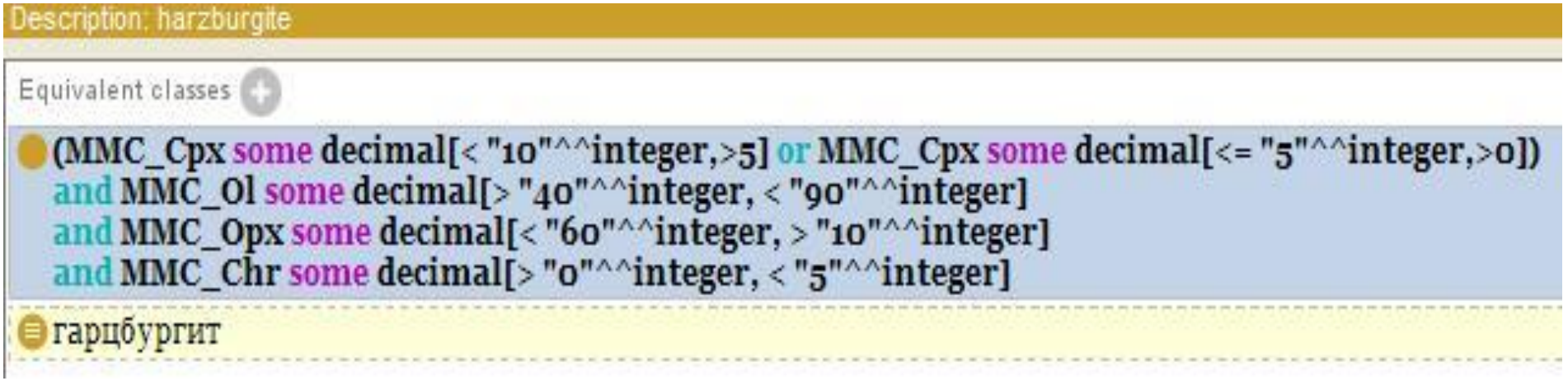
(MMC_Cpx some decimal [>5,<10] or MMC_Cpx some decimal[>0,<= 5])

and MMC_Ol some decimal [>40,<90]

and MMC_Opx some decimal[>10,<60]

and MMC_Chr some decimal[>0,<5]

Эквивалентность классов -3-



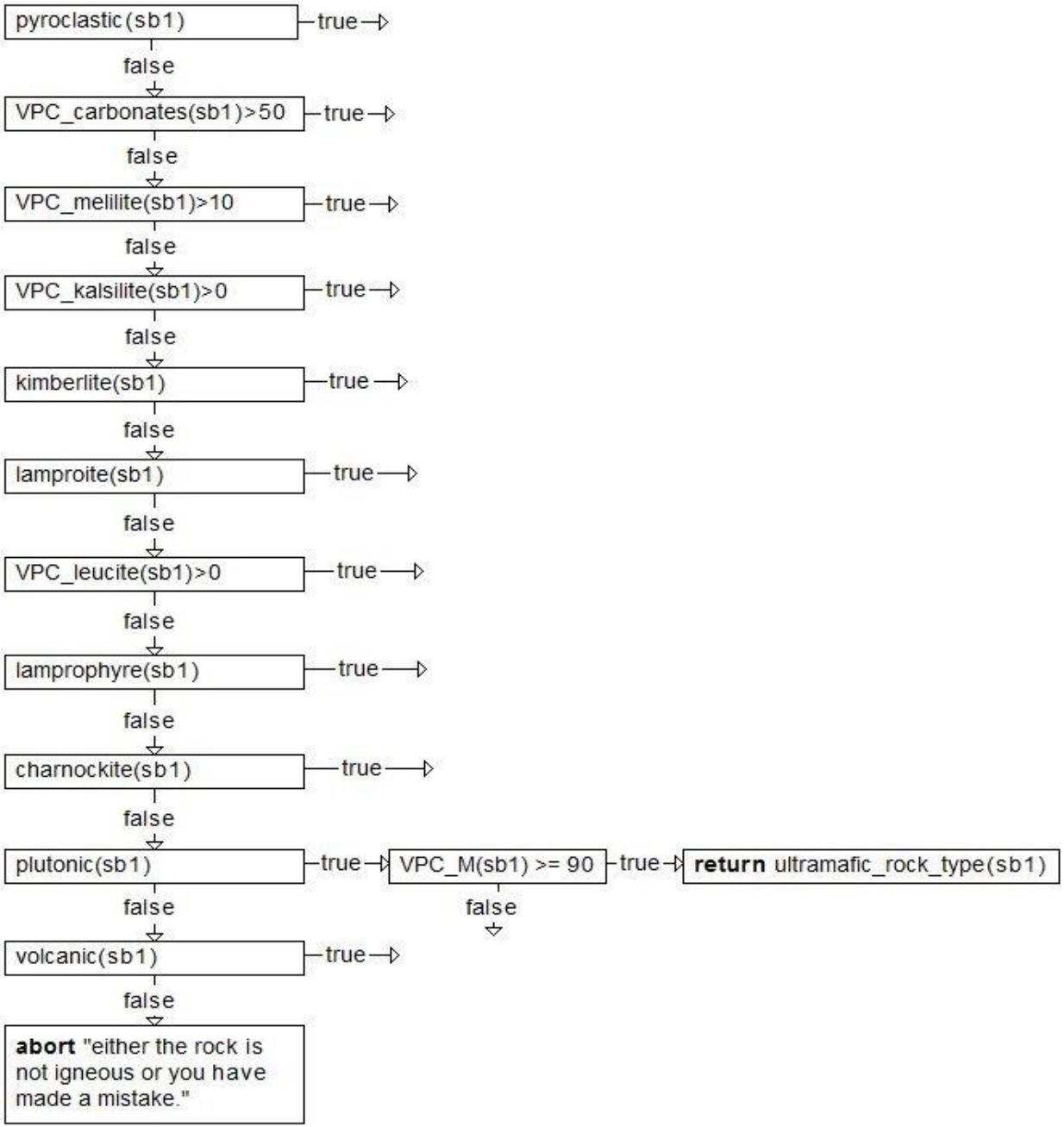
https://sites.google.com/site/alex0shkotin/formal-geology/ovgpmms

Задачка Эйнштейна: http://test.feofan.com/

Алгоритм классификации

Вход: сведения об образце.
Выход: термин - тип горной породы образца.
Алгоритм задаётся как совокупность функций для каждой из которых дана блок-схема, понятная петрологу.
Алгоритм использует числовые функции и предикаты.
При этом функции и предикаты мыслятся применяемыми к какому-то конкретному твёрдому телу (измерения).

Definition
rock_type(sb1) =



Извлечение предиката вида породы

Алгоритм классификации:

- содержит в неявном виде определения всех видов магматических горных пород, т.е.
- задаёт систему предикатов видов горных пород.

Получены формулы для предикатов harzburgite и dunite, которые оказались формулами с одной свободной переменной исчисления предикатов первого порядка с числами.

Ещё пример формального определение вида горной породы

абесседит есть

перидотит and смесь_минералов and содержит_минерал only (оливин or роговая_обманка or флогопит)

OWL синтаксис - Manchester.

Важно что её могут прочитать петрологи. Получение формальных (математических) определений, особенно в понятной экспертам форме - важнейшая цель проекта.

Выводы и дальнейшие планы

- Определение возможно
- Начато построение формальной теории
- Опробованы средства ведения формальных знаний
- Создан опытный образец концентратора определений

Возможный проект:

- Концентратор определений
- Формализация рекомендаций IUGS
- Формализация законов предметной области
- КРЯ (контролируемый русский язык)

Ссылки

[IRCGT] Le Maitre, L.E., ed. 2002. Igneous Rocks: A Classification and Glossary of Terms 2nd edition, Cambridge. url

[BGSRCS] Gillespie, M R, and Styles, M T. 1999. BGS Rock Classification Scheme, Volume 1, Classification of igneous rocks. British Geological Survey Research Report, (2nd edition), RR 99–06. url

[OWL2LE] OWL 2 Web Ontology Language. Data Range Extension: Linear Equations. url

[otch08] "БД Проба. онтология. промежуточный отчёт. Осень 2008", ГГМ РАН. url

[otch09] "Онтология словаря научных терминов. промежуточный отчёт. Осень 2009", ГГМ РАН. url

[otch10] «Алгоритм классификации магматических горных пород и формальное определение вида горной породы. Исследовательский отчёт. Осень 2010», ГГМ РАН. url

Acknowledgments

We would like to thank

- Dr. Stephen M. Richard from Arizona Geological Survey for comments on the report [otch10], helpful discussion and reference to [BGSRCS]
- Pavel Klinov from University of Manchester for numerous invaluable comments
- Dr. Kaarel Kaljurand from Attempto group for idea to use proper names

Спасибо за внимание:-)

ashkotin@acm.org