Метаданные о научных методах для обеспечения их повторного использования и воспроизводимости результатов

Н.А.Скворцов, Д.О.Брюхов, Л.А.Калиниченко, Д.Ю.Ковалёв, С.А.Ступников

nskv@ipi.ac.ru

RCDL-2013, Ярославль

План

- Науки с интенсивным использованием данных
- Проблема повторного использования научных методов и воспроизводимости результатов
- Пример среды исследований MyExperiment
- Структура потоков работ
- Набор дополнительных метаданных для решения обсуждаемой проблемы
- Реализация прототипа
- Запросы к метаданным
- Выводы

Науки с интенсивным использованием данных

- Науки с интенсивным использованием данных (data-intensive sciences)
  - Новые требования к объёмам и скорости обработки данных
- Изменение парадигмы исследований
  - От поиска данных для решения задачи к анализу больших объёмов данных для нахождения новых знаний
  - The Fourth Paradigm: Data-Intensive Scientific Discovery. T. Hey, et al (Eds). – Microsoft Research. – Redmond, 2009.
  - Jim Gray & Alex Szalay
- Сбор данных от инструментов наблюдения или моделирования
- Курирование данных и организация долгосрочного хранения
  - Семантические подходы к представлению данных
  - Эффективность представления и доступа
  - Обеспечение надёжности данных
- Анализ данных
  - Доступность методов, алгоритмов и инструментов обработки данных
  - Простота использования методов
  - Постоянная и всесторонняя обработка больших объёмов данных

Использование новой парадигмы в исследованиях

- Постоянное автоматическое применение широкого ассортимента известных методов
  - Подготовка сырых данных к анализу
  - Нахождение существенных свойств и параметров объектов
  - Классификация объектов
  - Выявление особых объектов, ошибок
  - Проверка научных гипотез
  - Подтверждение или опровержение экспериментальных моделей
  - Применение научных методов над всеми доступными данными
  - Агрегация
- Предоставление исследователям богатого набора методов анализа и среды исследования для анализа больших объёмов данных
  - Потоки работ для применения методов и поэтапной обработки данных
  - Специфические методы и законы предметной области
  - Аналитические методы: статистика, машинное обучение и др.
- Использование производной информации и методов в последующих исследованиях
  - Воспроизведение тех же результатов другими группами исследователей
  - Применение тех же методов над другими наборами данных
  - Результаты применения научных методов сохраняются и становятся источником данных для работы других методов в данной области и сопряжённых проблемных областях

Повторное использование методов и воспроизводимость результатов

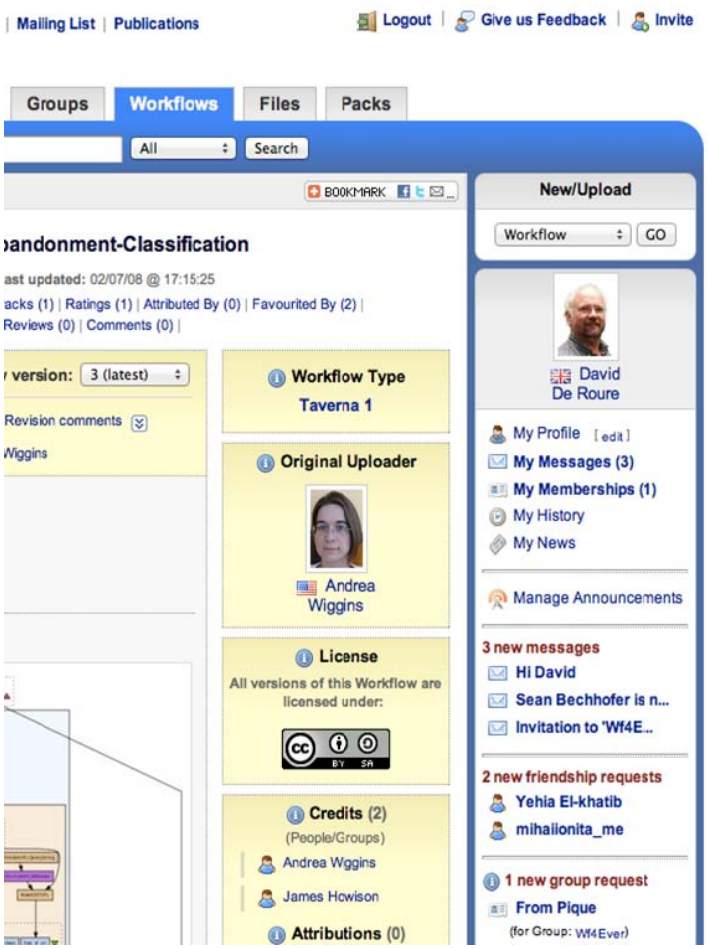
- Помимо накопления научных данных необходимо накапливать реализации научных методов
  - Доступность методов в сообществе
  - Возможность совмещения применения методов
  - Сервисы и потоки работ
- Возможность выбора, совмещения источников данных
- Независимость реализаций методов от источников информации
  - Возможность применения над произвольными данными
- Возможность поиска накопленных методов по различным критериям и их применения
  - Семантический поиск доступных методов в предметной области научного сообщества
  - Семантическое описание входных/выходных параметров, их этапов в потоках работ, данных, передаваемых между этапами
  - Систематизация накопленных данных и методов
  - Развитие спецификаций предметных областей внутри сообществ исследователей
  - Условия и среды воспроизведения
- Надёжность данных и результатов
  - Обеспечение необходимого качества данных
  - Информация о точности и полноте открытых данных, точности и полноте результатов
  - Данные о происхождении исходных данных, методах получения производных данных
  - Сохранение результатов
  - Тестовые наборы исходных данных и результатов для проверки различных ситуаций в работе методов

Примеры проектов

- Visier
  - Накапливает всевозможные каталоги астрономических данных
  - Организует их поиск и поиск в них
  - Предоставляет набор наиболее востребованых сервисов, расчёт производных некоторых параметров над конкретными каталогами
- Astrogrid
  - Реестры каталогов
  - Удалённый доступ к данным, к сервисам различного назначения
  - Рабочая область
- MyExperiment
  - Обеспечение взаимодействия пользователей
  - Накопление методов: тысячи потоков работ
  - Десятки проектов
- Wf4ever
  - Набор сервисов для поддержки повторного использования потоков работ
  - Проверка работоспособности методов и выяснение причин недоступности
  - Спецификации происхождения результатов
  - Публикация объектов исследования

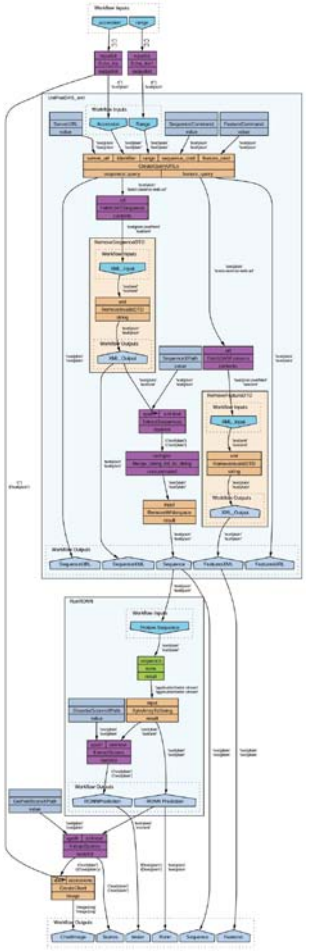
MyExperiment

- Сервер MyExperiment
- Социальная сеть поддержки научных экспериментов
- Коллекция объектов исследования
  - Файлы
  - Ссылки
  - Научные потоки работ
- Структура данных (онтология)
  - Пользователи
  - Группы (пользователи, доступ к объектам исследования)
  - Пакеты (объекты исследования)
    - Файлы (данные, документы, …)
    - Потоки работ (Taverna, …)
    - Ссылки на внешние ресурсы
  - Аннотации (теги, рейтинги, …)



Taverna

- Потоки работ - магистраль анализа данных
- Управление вызовом сервисов и информационных ресурсов, направление данных между ними
- Описание инструкций обработки данных повторяющихся научных экспериментах
- Плагины
  - Различные сервисы, вызываемые из потоков работ
  - Происхождение
  - Специализированные астрономические сервисы



Структура метаинформации о потоках работ MyExperiment

- Потоки работ
  - Поток работ как объект (Workflow)
  - Поток работ как набор компонентов (Dataflow)
    - Свойство has-component
- Узлы потоков работ
  - WorkflowComponent
    - Свойство belongs-to-workflow
  - Разновидности узлов (суперпонятие NodeComponent)
    - Source – входные узлы потоков работ
    - Processor – узлы обработки
      - Разновидности процессоров ConstantProcessor, WSDLProcessor, DataflowProcessor
      - Свойство processor-uri
    - Sink – выходные улзы
- Соединения компонентов
  - Разновидности компонентов соединения (суперпонятие IOComponent)
    - Input – данные входа узла
    - Output – данные на выходе узла
    - Link – данные, передаваемые между узлами
      - Свойства from-output, to-input

Интерфейсы (API) MyExperiment

- http-запросы (REST)
  - GET: http://www.myexperiment.org/workflows/16
    Accept: application/rdf+xml
  - …
    <Workflow rdf:about="3565">
    <content-url rdf:resource="wf.t2flow">
    </Workflow>
    …
- MyJPI - Java API (реппер над REST)
- Точка доступа Sparql
  - http://rdf.myexperiment.org/sparql?query=...&formatting=XML&reasoning=1
  - PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> SELECT DISTINCT ?type WHERE { <http://www.myexperiment.org/workflows/16> rdf:type ?type }
  - …
    <binding name="type">
    <uri>http://rdf.myexperiment.org/ontologies/contributions/Workflow</uri>
    </binding>
    …
- Нет описания некоторых метаобъектов
  - компоненты потоков работ

Анализа среды MyExperiment

- Реализованного аннотирования тегами недостаточно
  - Аннотируются только целые потоки работ и файлы, но не компоненты и интерфейсы потоков работ
  - Не обеспечивают семантического подхода
- Нет требования независимости потоков работ от источников данных
  - Многие потоки работ состоят из сервисов доступа к определённым базам данных
- Для обеспечения требований повторного использования методов и вопроизводимости результатов необходим набор дополнительных метаданных
  - На основе доступных API: средствами онтологий и RDF
  - Аннотирование компонентов потоков работ
  - Возможность задания запросов одновременно к метаданным и структуре потоков работ

Требования к дополнительным метаданным

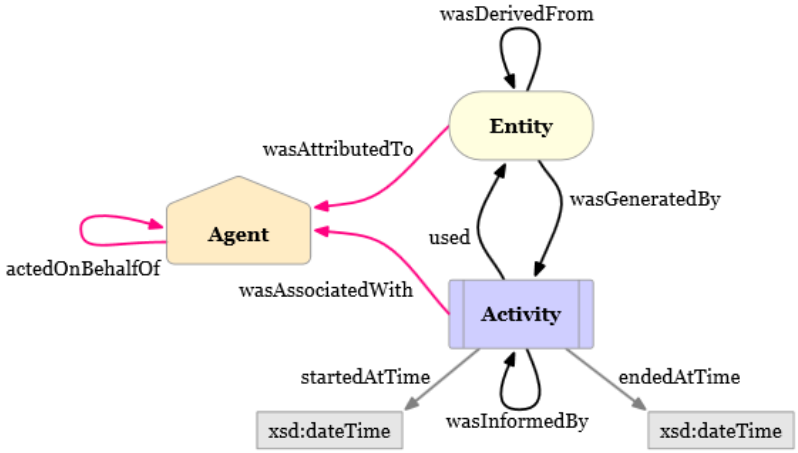
- Семантический подход
  - С описанием предметной области научного метода
  - Понимание машиной и человеком
- Доступность методов
  - Возможность поиска
  - Независимость от источников
- Информация о происхождении и качестве данных и методов
  - Источники
  - Надёжность
  - Точность
  - И другие
- Описание требований на уровне компонентов
  - Входных и выходных данных
  - Сервисов
  - Используемых информационных ресурсов
  - Данных, тестов
- Включение информации о среде воспроизведения

Онтология предметной области

- На примере астрономической области
- Модули разделов астрономии
  - Астрометрия, фотометрия, спектрометрия, астрофизика, астрономические объекты, кратные системы, затменные двойные, и др.
- Решаемые проблемы
  - Обеспечение семантического поиска в предметной области
  - Аннотация объектов исследования в терминах понятий предм. области
    - Семантика реализуемых методов в целом
    - входы/выходы, пред- и постусловия (по примеру OWL-S)
    - Семантика данных и результатов
    - Семантика ситуаций в наборах тестов
  - Систематизация объектов исследования
- Астрометрия
  - Coordinate
  - CoordinateSystem
  - EquatorialCoordinateSystem
  - CoordinateSystemComponent
  - Epoch
  - RightAscension
  - Declination
- Астрономические объекты
  - AstrObject
  - StellarObject
  - CompoundObject
  - Star
  - MultipleStar

Онтология происхождения данных

- Онтология PROV-O
  - Агент
    - человек, организация, программа
  - Сущность
    - описываемая сущность, план, множество, комплект
  - Деятельность
  - Связи
    - Взаимодействие сущностей, агентов и деятельностей
- Рекомендация W3C
- Решаемые проблемы
  - Трекинг применяемых компонентов потоков работ
  - Систематизация версий
  - Контроль источников ошибок
  - Спецификация источников данных и методов
  - Трекинг в результатах тестов
  - Обеспечение достоверности данных и реализаций методов
  - Сравнение работы разных реализаций
  - И другие



Пример метаданных происхождения

Метод resolve_coordinates, возвращающий координаты по имени астрономического объекта

```
wf3514:resolve_coordinates
    rdf:type     prov:SoftwareAgent .
wf3514:resolve_coordinates_outputTable
    rdf:type     prov:Entity;
    prov:wasAttributedTo    wf3514:resolve_coordinates;
    prov:wasGeneratedBy    wf3514: .
```

Онтология качества данных и методов

- Измерения
  - Полнота, Точность, Объем, Возраст данных, Целостность, Надёжность
- Метрики (примеры сервисов)
  - Полнота – относительное количество непустых значений
  - Точность – рассчитанная или взятая из данных точность
  - Объем – количество кортежей
  - Возраст данных – разница даты создания и текущей даты
  - Целостность – соответствие набору определённых правил
  - Надёжность – рассчитанное на основе значений других метрик

Среда вопроизведения

- Параметры
  - Среда, система, сервисы
  - Модели, методы, алгоритмы, интерфейсы, стандарты
  - Исходный код и средства разработки
  - Документирование
  - Данные и входные параметры
  - Цели и результаты
- Многое из этого выразимо в онтологии происхождения
  - частично дублирует онтологию происхождения
  - Возможно сделать над ней

Результирующая структура необходимых метаданных

- Онтология предметной области исследования
  - Понятия предметной области
  - Связи и знания
- Онтология происхождения данных и методов
  - Агенты
  - Деятельности
  - Сущности
- Онтология качества данных
  - Измерения (dimension)
  - Метрики
- Онтология сред воспроизведения

Реализация

- На ontology.ipi.ac.ru
  - Модули онтологии предметной области
    - http://ontology.ipi.ac.ru/ontologies/astront/
  - Онтология качества
    - http://ontology.ipi.ac.ru/ontologies/quality.owl
  - Точка доступа SPARQL (Jena)
    - http://ontology.ipi.ac.ru:3030/
- Онтология PROV
  - http://www.w3.org/ns/prov
- MyExperiment
  - Точка доступа http://rdf.myexperiment.org/sparql
  - Онтологии MyExperiment http://rdf.myexperiment.org/ontologies/

Пример совместного запроса к метаданным

```
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix mecomp: <http://rdf.myexperiment.org/ontologies/components/>
prefix astrobjects: <http://ontology.ipi.ac.ru/ontologies/astrobjects.owl>
prefix astrometry: <http://ontology.ipi.ac.ru/ontologies/astrometry.owl>
prefix prov: <http://www.w3c.org/ns/prov#>
SELECT ?workflow ?output WHERE
{
  ?input rdf:type astrobjects:AstrObject .
  ?output rdf:type astrometry:Coordinate .
  ?output prov:wasGeneratedBy ?workflow .
  ?output prov:wasAttributedTo ?service .
  SERVICE <http://rdf.myexperiment.org/sparql>
  {
    ?input rdf:type mecomp:Source .
    ?output rdf:type mecomp:Sink .
    ?input mecomp:belongs-to-workflow ?workflow .
    ?output mecomp:belongs-to-workflow ?workflow .
  }
}
```

Результат запроса

```
<?xml version="1.0"?>
<sparql xmlns="http://www.w3.org/2005/sparql-results#">
  <head>
    <variable name="workflow"/>
    <variable name="output"/>
  </head>
  <results>
    <result>
      <binding name="workflow">
        <uri>http://www.myexperiment.org/workflows/3514/versions/2</uri>
      </binding>
      <binding name="output">
        <uri>http://www.myexperiment.org/workflows/3514/versions/2
              #dataflows/1/components/2</uri>
      </binding>
    </result>
</results>
</sparql>
```

Выводы

- Предложен состав метаданных, необходимых для реализации сервисов обеспечения повторного использования научных методов и воспроизводимости результатов
- Выполнены требования парадигмы исследований наук интенсивным использованием данных
  - Семантизация поиска и спецификаций методов, потоков работ и данных
  - Независимость реализаций методов от источников информации
  - Возможность применения над произвольными данными
  - Обеспечение надёжности и качества данных и результатов исследований
- Результаты могут быть использованы для создания сред поддержки исследований в условиях роста объёма данных и объёма необходимых исследований над ними