Повторное использование данных

Курс: Концептуальное моделирование предметных областей

План

- Сопровождение (курирование) данных
- Принципы обеспечения повторного использования данных (FAIR)
- Различные интерпретации FAIR
- Инфраструктуры данных
- Перспективы концептуального моделирования для FAIR

Сопровождение данных

- Курирование данных (Data Curation) – сопровождение данных, обеспечивающее их оптимальное долгосрочное использование
  - Способы длительного хранения (long-term preservation)
  - Очистка, структурирование и стандартизация данных
  - Обеспечение доступа
  - Описание метаданными и обеспечение поиска данных
  - Метаданные о происхождении данных
  - Связывание данных со средствами их обработки
- MyExperiment, WF4Ever – воспроизводимость результатов
  - Сообщества исследователей, проекты
  - Research Objects (RO) – капсулы с данными, потоками работ для их обработки, документацией
  - Публикация и поиск потоков работ
- FORCE11 – сообщество по электронным публикациям на основе семантических технологий

Принципы FAIR-данных

Findable – обнаруживаемые

- Постоянные уникальные идентификаторы
- Обширное описание метаданными
- Метаданные включают идентификаторы данных
- Регистрация, индексирование для возможности поиска

Accessible – доступные

- Стандартный протокол
- Извлечение по идентификаторам
- Возможна авторизация и аутентификация
- Метаданные доступны, даже когда данные больше не доступны

Interoperable – интероперабельные

- Использование формального языка представления знаний
- Интерпретация машиной
- Использование словарей, отвечающих FAIR
- Ссылки на другие данные

Reusable – повторно используемые

- Определение множества атрибутов применимости (нефункциональные)
- Лицензирование
- Детальное писание происхождения
- Поддержка стандартов предметных сообществ

Интерпретации принципов FAIR-данных

- Дальнейшая детализация требований к данным
  - Появление новых аббревиатур, уточняющих принципы для различных областей и целей
- Оценка данных с точки зрения FAIR
  - Метрики оценки, сертификация центров данных и архивов
- Подбор технологий, обеспечивающих реализацию принципов
  - Основанные на RDF и LOD: контейнеры данных (LDP), отображение данных (RML), извлечение данных (TPF)
  - Конкретные реализации принципами не продиктованы
- Применение к другим областям, кроме архивов данных
  - Инфраструктуры исследовательских данных
  - Цитирование данных, планы управления данными,
  - Исследование принципов автоматизации (IFDS)
  - Потоки работ

Охватываемые цели FAIR

- Формальность описаний данных
  - Возможность вывода
- Интерпретируемость машиной
  - От понимания определённого набора метаданных к пониманию данных, с которыми машина ранее не работала
  - Данные не только машинно-читаемые (readable), но и машинно-управляемые (actionable)
- Ресурсы, связанные с данными, также являются данными
  - Метаданные, словари и онтологии, потоки работ, программы, форматы, интерфейсы, протоколы, цитирования, документы и другие
  - Они должны следовать принципам FAIR, чтобы сами данные отвечали принципам FAIR-данными
- Принципы признаются рассчитанными на перспективу
  - Нет проектов, полностью выполняющих или не выполняющих принципы
  - Есть направление развития

Реализации разных принципов FAIR

- Снабжение данных метаданными для семантически значимого поиска
  - По текстовым описаниям, по словарям
  - Организация реестров данных
- Унификация интерфейсов доступа к данным
  - Унификация протоколов доступа, (данные могут оставаться неоднородными)
  - Стандартизация словаря атрибутов в определённой предметной области
- Использование стандартных представлений данных
  - Представление в примитивных форматах без семантики данных
  - Сохранение семантики данных в форматах определённых предметных областей
- Глобальная идентификация ресурсов
  - Идентификация наборов данных в целом, доступ к классам данным по идентификаторам
  - Идентификация по срезам данных, атрибутам, типам
  - Идентификация объектов
- Создание контейнеров
  - Агрегация данных с необходимыми ресурсами для их обработки
  - Подробное документирование
- Подходы к автоматизации обработки данных
  - Планы управления данными
  - Предопределённые потоки работ
  - Действия на основе событий

Состояние проблемы

- Разрыв между подходами к доступу и повторному использованию данных, превалирующими в конкретных дисциплинах
- Семантика данных используется слабо
- Решение проблем неоднородности данных в каждой задаче (они занимают больше половины времени в исследованиях)
- Обработка данных, управляемая машиной, встречается редко

Инфраструктура EOSC

- Заблаговременно разработанные планы управления данными (DMP) в течение проекта
- Цифровые объекты (DO)
  - Данные
  - Постоянные уникальные идентификаторы (PID)
  - Стандарты и форматы
  - Метаданные, происхождение
  - Код
  - Лицензии
- Реестры этих видов данных
- Протокол управления цифровыми объектами (DOIP)
- Разделение на сервисы для обеспечения
  - Сервисы для клиентов-людей
  - Сервисы для машин
- Признаётся фактическое отсутствие подходов к управлению данными машиной (machine-actionable)

Потенциал концептуального подхода

- Необходимость формального концептуального описания знаний предметной области
  - Онтологии предметной области
  - Основа для возможности вывода
  - Предложение общих схем данных для предметных областей
  - Сообщество заинтересовано в описании своей предметной области
- Подробное описание ресурсов данных, сервисов в терминах предметной области
  - Возможность точной классификации данных
- Реестры ресурсов
  - Реестры, основанные на онтологиях предметных областей
  - Хорошо классифицированные коллекции ресурсов, используемых сообществом
  - Коллекции данных, метаданных, методов
  - Ресурсы, интегрированные в соответствии со знаниями предметной области
- Сервисы поддержки взаимодействия в сообществах
  - Семантический поиск ресурсов на основе ограничения требований
  - Средства интеграции данных и методов (соответствие семантики данных, интерфейсы, идентификация объектов)
  - Подробная публикация средствами формальных онтологических описаний
  - Поддержка жизненного цикла решения задач над данными сообщества

Принципы цитирования данных

- Данные являются продуктом исследований, которые могут цитироваться наряду с другими видами исследовательских объектов, в частности, публикациями
- Цитирование должно перечислять исследователей (авторов), причастных к данным
- Любое использование данных в исследованиях необходимо цитировать
- Цитирование использует постоянные идентификаторы
- Цитирование данных должно обеспечивать доступ к данным и необходимым для их использования метаданным, к документации, коду
- Метаданные и идентификаторы данных, используемые в цитировании, должны быть постоянными, даже если данные претерпевают изменения или заканчивают жизненный цикл
- Цитирование включает метаданные происхождения, позволяющие проверить версию, временной интервал, цитируемую часть
- Цитирование должно учитывать специфику сообществ, однако это не должно быть препятствием для цитирования данных вне сообществ

Резюме

- Возможность и целесообразность использования формальных описаний данных и автоматизации на основе концептуальных подходов в сообществах декларируются и исследуются, однако редко используются в достаточной мере
- Концептуальное моделирование позволяет семантически описывать, классифицировать и накапливать используемые в сообществах данные, избавляться от неоднородности данных в сообществе
- Подход позволяет обоснованно повторно использовать интегрированные релевантные задачам данные, существенно повышать масштабируемость и автоматизацию при работе к неоднородными данными