Новый источник данных для наукометрических исследований

М.Р. Когаловский, ИПР РАН

С.И. Паринов, ЦЭМИ РАН

Работа поддержана РФФИ, проект 12-07-00518-а и РГНФ, проект 11-02-12026-в

Контекст работы

- Ранее на конференции RCDL была представлена наша работа о семантическом структурировании контента научных ЭБ на основе онтологии связей принадлежащих ему информационных объектов
- Семантическая структура контента формируется путем декларации связей между его информационными объектами с явным образом декларированной семантикой (семантических связей)
- Семантика связей определяется на основе специальной онтологии
- В нашем проекте в формировании семантической структуры участвуют зарегистрированные пользователи ЭБ в онлайновом режиме
- ЭБ с такими возможностями обеспечивают новые возможности для наукометрии, для доступа пользователей к ресурсам ЭБ, создают условия для новых форм научной деятельности
- В докладе кратко представлены основные принципы предлагаемого подхода, развиваемого в рамках проекта Соционет, рассматриваются результаты его реализации, относящиеся к наукометрии
- Реализация предлагаемого подхода осуществляется в среде системы Соционет. Ключевые элементы предлагаемого подхода реализованы.

О практике наукометрических исследований

- В сложившейся практике оценки продуктивности научной деятельности исследовательских организаций, групп ученых и отдельных ученых основной критерий индексы цитирования
- При оценке научных публикаций обращается внимание на импактфактор журналов, в которых публикуются работы
- Однако зарубежный опыт показывает, что такой критерий в ряде стран используется либо с осторожностью и обязательно дополняется экспертной оценкой, либо вообще не используется (см. презентацию доклада авторов на заседании Президиума РАН)
- Более того, целесообразность использования некоторых элементов такой системы оценки подвергается в последнее время критике: категорическое неприятие импакт-фактора (см. San Francisco Declaration on Research Assessment San Francisco DORA), неудовлетворение использованием индекса Хирша.

Особенности данного проекта в наукометрии

- Слабая сторона индексов цитирования они основаны на «немых» ссылках (связях) цитирования, которые не несут какой-либо информации о мотивах цитирования
- Парадоксы «немых» связей высокое цитирование претенциозных абсурдных публикаций как ответная реакция научного сообщества
- Однако мотивы цитирования часто обозначены в контексте ссылки либо могут быть установлены экспертом в предметной области публикации
- В предлагаемом подходе связи цитирования обогащаются явно определенной семантикой, характеризующей мотивы цитирования
- Вместе с тем, предоставляются средства для указания с помощью семантических связей также и других факторов влияния одних публикаций на другие, которые не обозначены связями цитирования
- Учитываются факторы влияния не только публикаций, но и научных информационных объектов другого рода: средств программного обеспечения, наборов данных и др.
- Такой подход позволяет проводить более содержательные наукометрические исследования и обеспечивает другие полезные возможности, которые рассматриваются в докладе.

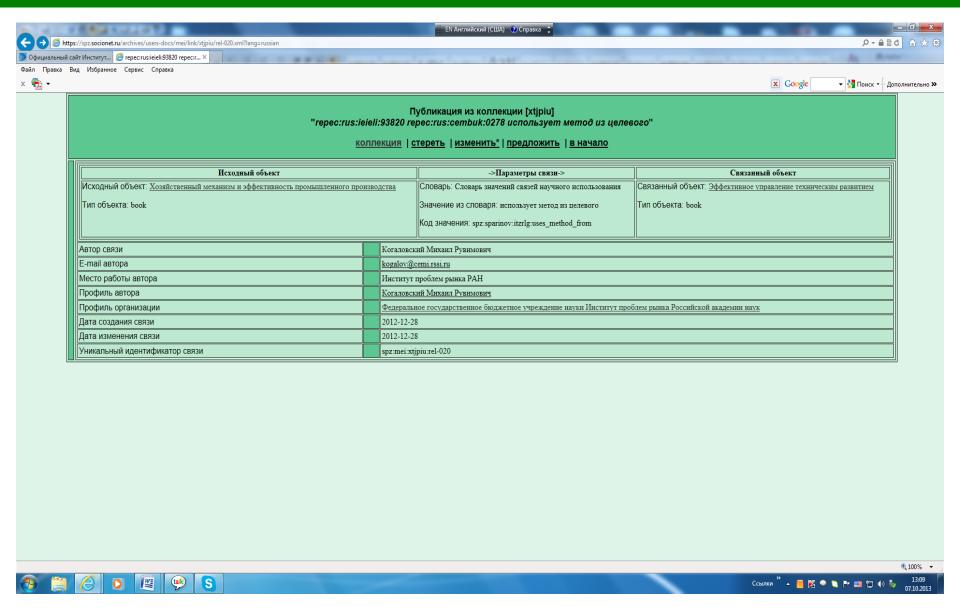
Используемые виды семантических связей

- Между информационными объектами контента ЭБ поддерживаются различные отношения, представляемые семантическими связями, в том числе, связи цитирования
- Рассматриваются бинарные ориентированные связи
- Участники связей информационные объекты разных типов, в частности: публикации (монографии, статьи, диссертации и авторефераты, научные отчеты, научные данные...), персоны, организации и др.
- В нашем проекте мы не ограничиваемся связями цитирования
- Семантические связи в нашем подходе отражают отношения:
 - научного характера (публикация-публикация или персона-публикация)
 - авторства (организация—публикация, персона—публикация),
 - *административные* (организация–персона, персона-персона)
 - *ролевые отношения* между публикациями и/или их фрагментами (публикация—аннотация, публикация—библиография ...) и др.
- Семантика связей определяется на основе онтологии связей.

Встроенные и автономные декларации связей

- В известных авторам проектах, использующих семантические связи, их декларации включаются в метаданные исходных информационных объектов связей мы называем их встроенными декларациями
- Встроенная декларация, т.о., является частью метаданных исходного объекта связи и создается на стадии подготовки этих метаданных
- Декларации другого вида автономные декларации могут создаваться только для уже существующих информационных объектов и представляются в нашем проекте как самостоятельные информационные объекты специального типа (linkage)
- В автономной декларации указываются: уникальный ID связи в системе, ее семантика, уникальные ID объектов-участников, автор связи, дата создания, комментарий
- Автономные связи организуются *в коллекции*, как и информационные объекты других типов
- Встроенные декларации могут быть преобразованы в автономные.

Пример представления автономной декларации



Известные разработки онтологий связей (1)

- Первые существенные результаты в разработках онтологий связей между научными публикациями были опубликованы в 2010 г.
- Наиболее активно этим занимаются специалисты в области наук о жизни, прежде всего в биомедицине
- Однако полученные результаты имеют общий характер и могут применяться в других предметных областях
- Разработаны, например: онтологии CiTO (Citation Typing Ontology), DoCo (Document Components Ontology) и др., модульный комплекс онтологий SWAN (Semantic Web Applications in Neuromedicine) и ряд др.
- Эти частные онтологии были сведены в единую систему онтологий, названную SPAR (Semantic Publishing and Referencing Ontologies)
- В ней наряду с семантикой цитирования акцентируется использование в издательском деле: включена, в частности, онтология ролей в издательском деле PRO (Publishing Role Ontology), онтология состояний издания PSO (Publishing Status Ontology) ...

Известные разработки онтологий связей (2)

- Другая известная разработка рекомендация SKOS (Simple Knowledge Organization System) консорциума W3C
- По существу, SKOS это онтология связывания научных данных; она более агрегирована и более ограничена по сравнению с онтологиями, входящими в состав SWAN
- Наконец, нужно упомянуть европейский проект CERIF (Common European Research Information Format), цель которого разработка обобщенной концептуальной модели научных данных
- В рамках этого проекта также предложен документ, в котором предлагается определение стандартизованной семантики отношений между объектами научных информационных систем CRIS (Common Research Information System).

Семантические связи и онтология данного проекта

- В нашем проекте допускаются оба вида деклараций семантических связей:
 - *Встроенные, используемые, в частности* для описания связей цитирования публикаций из списка литературы
 - *автономные* для уже существующих информационных объектов участников связи
- В проекте используется *«синтетическая»* онтология, включающая некоторые фрагменты рассмотренных онтологий, а также ряд дополнений
- На ее основе разработана для реализации таксономия семантических связей
- Наша таксономия представляет собой двухуровневую иерархию классов семантических связей (верхний уровень классы связей по функциональности, нижний их подклассы)
- Таксономия семантических связей реализуется в виде совокупности контролируемых словарей, каждый из которых представляет класс верхнего уровня.

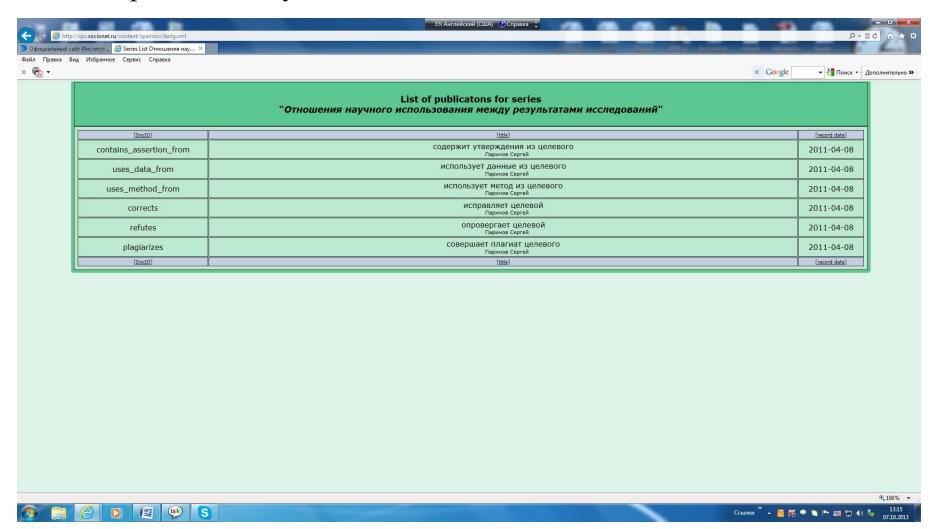
Примеры словарей семантических связей (1)

• Словарь мнений и оценок



Примеры словарей семантических связей (2)

• Словарь связей научного использования



Области действия словарей связей

- Контент ЭБ содержит информационные объекты различных типов: Организации, Персоны, Публикации, Связи этих объектов и др.
- В зависимости от типов объектов—участников, связи могут отражать различные отношения между ними, например: персона-публикация (авторство, мнение), публикация-публикация (оценка, характер научного использования)...
- Допустимые отношения для конкретной выбранной пары типов связываемых объектов определяются матрицей допустимости
- Строки и столбцы матрицы соответствуют типам объектов, а в ее клетках указаны допустимые для этих типов виды отношений со ссылками на соответствующие словари классов связей
- Обращаясь к ней в процессе описания связи, пользователь выбирает вид отношения, получает доступ к соответствующему контролируемому словарю классов связей, выбирает требуемый класс, который и указывается в декларации создаваемой связи.

Матрица применимости словарей связей

• Фрагмент матрицы применимости словарей



Кто может создавать семантические связи?

- Создавать связи могут в онлайновом режиме зарегистрированные в системе (ЭБ):
 - авторы информационных объектов (публикаций, связей и др.)
 - пользователи системы, выступающие в роли экспертов
- Для зарегистрированных в системе авторов информационных объектов и ее пользователей поддерживаются их профили
- Автор информационного объекта (или его представитель), в отличие от пользователя системы, может создавать связи со встроенной декларацией, прежде всего, связи цитирования
- В профилях авторов указываются адреса электронной почты для направления им сообщений системы оповещений об изменениях в составе и/или в свойствах связей, участниками которых являются их информационные объекты; то же делается для объектов-связей
- Два информационных объекта могут быть участниками нескольких связей одного и того же класса
- Множество таких связей, учрежденных одним лицом, может быть семантически противоречивым.

Как используются связи для наукометрии?

- Семантические связи, представляющие отношения научного характера (оценочные, научного использования, научного вывода) используются собственно для вычисления показателей, характеризующих оценку и использование публикаций как результатов научной деятельности:
 - Связи цитирования с встроенной декларацией (характеризуют мотив цитирования от имени авторов исходных публикаций этих связей)
 - Связи публикация-публикация с автономной декларацией (то же, но от имени авторов связей)
 - Связи персона–публикация с автономной декларацией (то же от имени персон – экспертов, являющихся авторами связей – дополнение к традиционному рецензированию)
- Семантические связи, представляющие отношения авторства (персона/организация-публикация), административные отношения (организация персона, персона—персона) используются для идентификации и агрегирования указанных выше статистических показателей.

Другие использования семантических связей (1)

- Наряду с использованием семантических связей в системе для формирования наукометрической статистики, предусматривается их использование и для других целей:
 - Доступ пользователей к информационным объектам контента ЭБ путем семантической навигации в его многослойной семантической структуре
 - Исследование топологии отдельных слоев графа связей
 - Использование как участников связей объектов, не принадлежащих контенту библиотеки, но доступных в Веб (это могут быть, например, различные ресурсы – страницы Веба, файлы ftp-архивов, веб-сервисы)
- Открытый доступ к ресурсам ЭБ, интерактивный режим доступа и возможность автономной декларации семантических связей, оповещение авторов информационных объектов, а также возможность обратной связи, обеспечивают среду для нового вида научной деятельности.

Другие использования семантических связей (2)

- В этой среде пользователи могут выражать мнение о представленных в ней ресурсах с помощью семантических связей вида персонапубликация (используя семантику связи + комментарий), а также о мнениях других пользователей с помощью связей персона-связь
- Тем самым создается виртуальная онлайновая социальная среда для совместной деятельности пользователей-ученых в качестве экспертов, добровольно высказывающих свои мнения о представленных в ЭБ (или в Вебе) ресурсах, а также о высказанных мнениях других экспертов о них
- Деятельность в таком виртуальном дискуссионном пространстве дополняет традиционную практику анонимного рецензирования научных публикаций, но в оперативном режиме, с доступным авторам содержанием высказанных мнений, с возможностью их ответной реакции
- Открытость высказанных оценок для научного сообщества позволяет ответным образом реагировать на них, способствует более высокой ответственности и объективности авторов этих оценок.

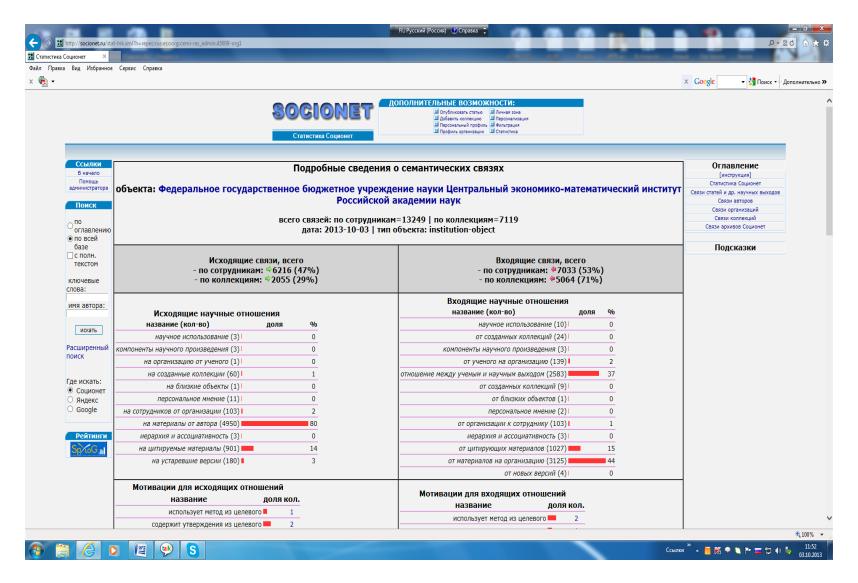
Состояние реализации в среде системы Соционет

- В настоящее в системе Соционет реализован набор основных сервисов для создания и использования семантических связей, выполняющих следующие функции:
 - Создание, обновление словарей связей
 - Создание и поддержка матрицы применимости словарей
 - Создание, обновление, удаление встроенных деклараций связей
 - Создание, обновление, удаление автономных деклараций связей
 - Поддержка коллекций автономных деклараций связей
 - Просмотр связей данного объекта и навигация по семантической структуре контента ЭБ
 - Генерация статистических данных по структуре семантических связей для заданного информационного объекта в целом, а также:
 - ✓ по входящим и исходящим связям,
 - ✓ по типам целевых объектов,
 - ✓ по классам связей
- Развитие комплекса системных сервисов продолжается.

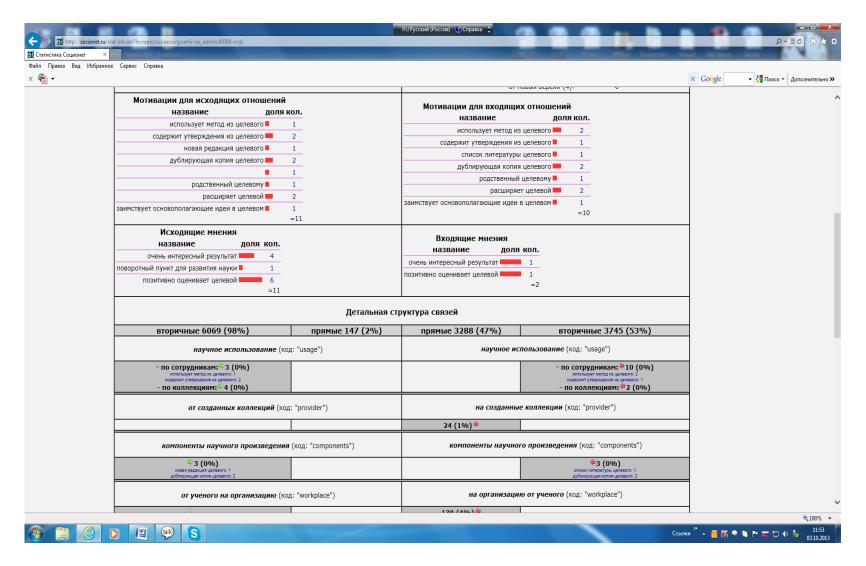
Наукометрическая статистика о связях

- Созданные сервисы системы Соционет генерируют статистику о связях заданного информационного объекта:
 - Для организации
 - Для автора
 - Для информационного объекта-публикации
 - Для связей персона-публикация
- Генерируется интегральная статистика по всем классам связей и дифференцированная по отдельным их классам
- При этом для некоторых классов связей учитываются транзитивные связи
- Далее приведено несколько примеров сгенерированных статистических данных
- Выводимые формы включают только поля, для которых существуют значения.

Пример статистики по организации

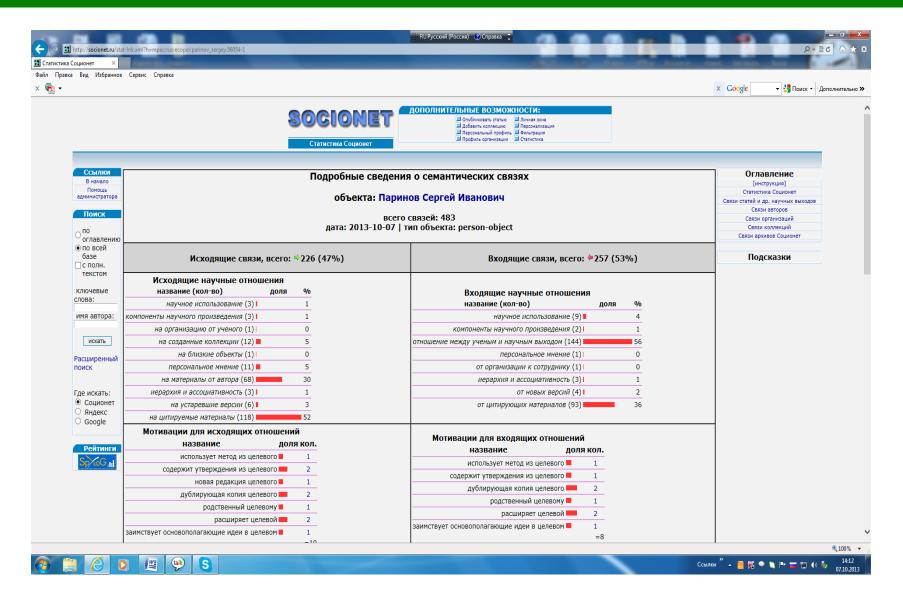


Пример статистики по организации (окончание)

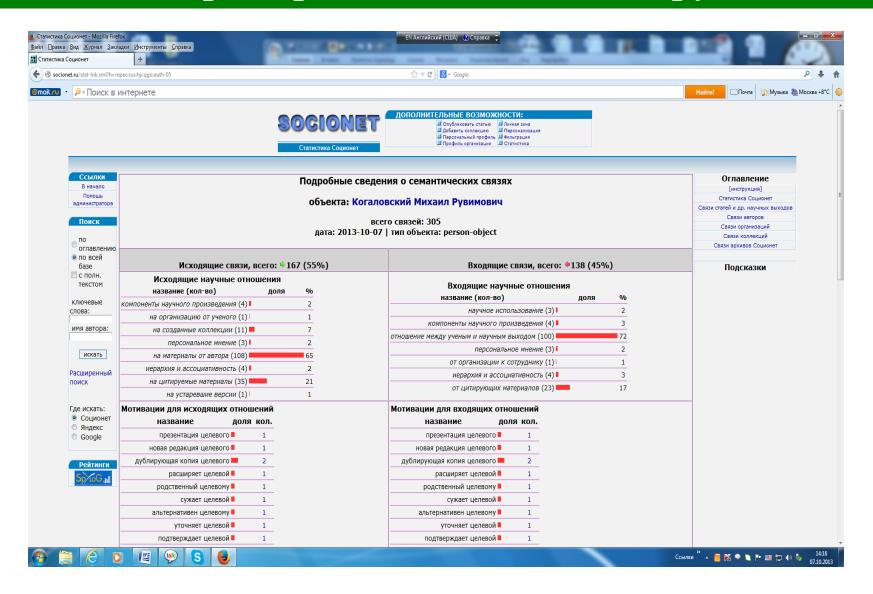


XV Всероссийская научная конференция RCDL-2013, Ярославль, 14-17 октября 2013 г.

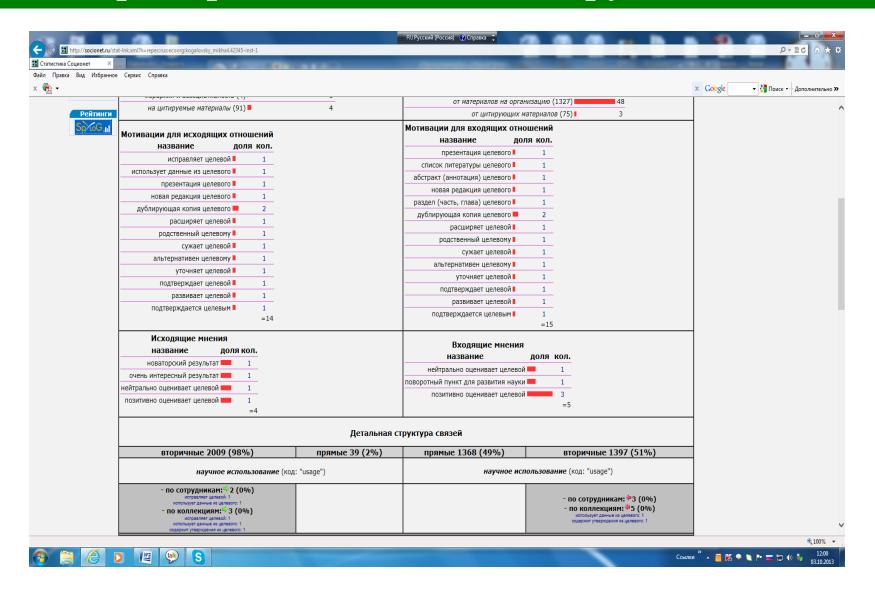
Пример статистики по автору



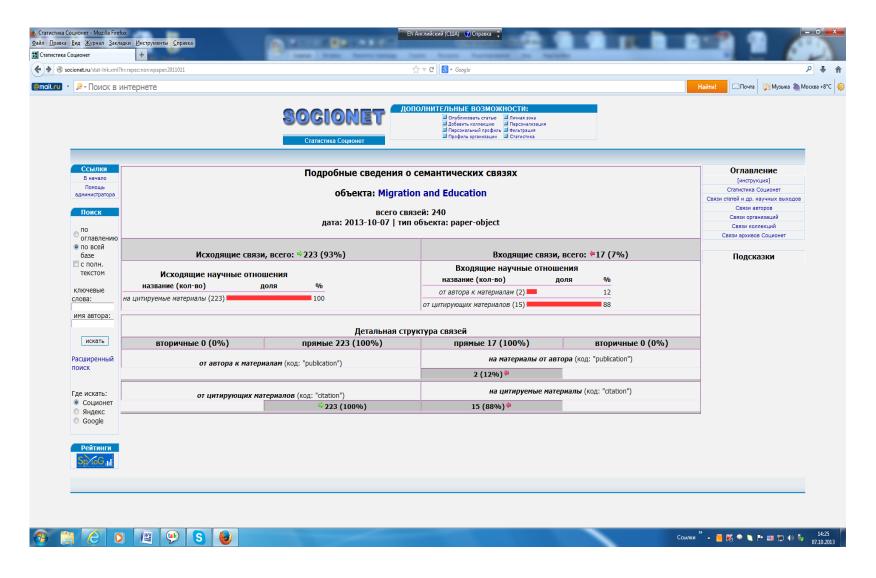
Пример статистики по автору



Пример статистики по автору (окончание)



Пример статистики по публикации



Репозитории семантических связей

- Семантические связи организуются в коллекции
- Из них можно сформировать репозиторий семантических связей (связей с автономной декларацией)
- Такой репозиторий является своего рода «семантическим ореолом» контента ЭБ (Semantic Halo Термин позаимствован из Dix A., Levialdi S. & Malizia A. Semantic halo for collaboration tagging systems. In: Social Navigation and Community-Based Adaptation Technologies Workshop June 20th, 2006)
- Возможна интеграция репозиториев семантических связей, построенных над различными библиотеками (при этом необходимо решить проблемы уникальности идентификации экземпляров связей и информационных объектов их участников в интегрированном репозитории)
- Если репозитории связей построены на основе технологии открытых архивов, эта задача решается относительно легко.

Итоги

- Предложенный в работе подход обеспечивает создание и использование семантического ореола (Semantic Halo) контента электронной библиотеки
- Важной его составляющей являются семантические связи, отображающие *научные отношения* между информационными объектами контента библиотеки (и это не только связи цитирования)
- Коллекции таких связей образуют новый источник данных для наукометрических исследований
- Система, поддерживающая рассмотренную функциональность, образует среду для нового вида научной деятельности
- Возможно *повторное использование* ресурсов семантического ореола в других репозиториях научной информации
- Обеспечивается семантическая навигация по структуре связей эффективный способ доступа пользователей к ресурсам библиотеки
- Семантический ореол контента электронной библиотеки может включать связи, участники которых *внешние* (не принадлежащие контенту, но доступные в Вебе) информационные объекты.

Конец

Благодарю за внимание