Модель семантического управления личной информацией

© А. А. Бездушный © А. Н. Бездушный © В. А. Серебряков

МФТИ ВЦ им. А.А. Дородницына РАН
andrey.bezdushny@gmail.com anb@ccas.ru serebr@ccas.ru

Аннотация

Целью данной работы является рассмотрение основных подходов к управлению информацией и знаниями, а также прототипирование системы предоставляющей человеку возможность организации личного информационного пространства в соответствии со стандартами Semantic Web и инициативой Linked Open Data. Структурированное представление данных позволяет проводить автоматизированный анализ сведений, с которыми ежедневно сталкивается человек, а использование стандартов Semantic Web предоставляет гибкие возможности для интеграции с репозиториями Linked Open Data. Предлагаемая методика развивает идею подхода Semantic Desktop, введенного Leo Sauermann, способа организации данных на персональном компьютере, в котором любой объект на компьютере – файл, e-mail или событие календаря, рассматривается как RDF ресурс (объект с уникальным идентификатором – URI).

1 Введение

Решением вопросов эффективной организации и работы с информацией и знаниями, занимаются системы управления личной информацией (Personal Information Management Systems). Одним из первых свое видение подобной системы, в 1945 г., высказал Вэнивар Буш в эссе «Как мы можем мыслить» [2]. В нем Буш описывает устройство под названием Мемекс (Memex), в котором люди могут хранить всю свою личную информацию – мысли, записи, книги, и которое может выдавать нужную информацию с достаточной скоростью и гибкостью. В основу работы Мемекс Буш закладывал механизмы ассоциативных ссылок и примечаний. Устройство, по его задумке, должно было точно имитировать ассоциативные процессы человеческого мышления, исключая присущие человеку недостатки, такие как забывание информации. Одной из технологий, необходимых для реализации своего устройства, Буш считал возможность организации хранилища, содержащего практически неограниченное количество информации, такое что «даже если бы пользователь вставлял в него по 5000 страниц сведений в день, ему бы потребовались сотни лет чтобы заполнить свое хранилище». Сейчас стоимость жестких дисков мала настолько, что человек может хранить всю изученную им информацию в течение неограниченного количества времени, при необходимости просто увеличивая объем хранилища, путем добавления нового жесткого диска. Таким образом, на пути создания Мемекса, в настоящее время лежит лишь проблема проектирования гибкой системы, способной помогать человеку при выполнении ежедневных задач, дополняя и структурируя его мыслительные процессы.

2 Управление информацией

С каждым годом количество информации, с которой ежедневно сталкивается человек, растет, и все больший ее объем переходит в электронный формат – публикуется в сети или сохраняется на персональных компьютерах. Эта информация распределяется между различными источниками, в которых хранится в разнородных форматах. Часть сведений может храниться в виде документов, другая – в виде ссылок или заметок, третья – в контексте не связанных между собой информационных систем. Такая организация данных приводит к *фрагментации информации* – в рамках различных источников, человеку приходится поддерживать различные, зачастую не связанные между собой, но обладающие общей структурой, организационные схемы. Разнородность форматов хранения данных затрудняет процесс задания зависимостей между этими схемами, в результате, сведения о взаимосвязях фиксируются только в памяти человека. Со временем, эти сведения неизбежно забываются, затрудняя процесс воссоздания контекста работы и поиска ресурсов, работа с которыми велась ранее.

Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.

Системы управления личной информацией автоматизируют процессы ведения и работы с *информационным пространством* – совокупностью всех сведений, с которыми человек работает сейчас или работал ранее. Рассмотрим основные функции, которые, с нашей точки зрения, должны выполнять системы управления личной информацией:

1. Ведение информационного пространства и структуризация *информационных ресурсов*, находящихся в нем. Под информационными ресурсами, будем понимать любые данные, имеющие важность для человека, и выделяемые им в отдельную сущность – это могут быть файлы, заметки, посещенные веб страницы, письма и пр. Системой управления личной информацией, должны обеспечиваться процессы, способствующие формализации информационных ресурсов, а также операций, выполняемых над ними.

2. Поиск по информационному пространству. Часто человек сталкивается с задачей повторного поиска информации, с которой он работал ранее. В таком случае, при поиске, он обычно обладает большим количеством косвенной информации, касающейся искомого ресурса. Эта информация чаще имеет ассоциативный формат, например, с какими еще ресурсами велась работа одновременно с искомым, в какой период времени это выполнялось, какая последовательность действий была совершена при первичном нахождении ресурса. Системы управления личной информацией должны поддерживать ведение такого рода метаданных и предоставлять возможность поиска по ним.

3. Автоматический анализ информационного пространства. Поскольку информационные ресурсы хранятся в структурированном виде, становится возможным проведение их автоматической обработки и анализа. Можно выделить несколько аспектов анализа. Автоматическое пополнение метаданных об информационных ресурсах, сведениями, найденными в сети. Другим аспектом является поиск схожих или связанных ресурсов, как в рамках информационного пространства пользователя, так во внешней сети.

4. Категоризация ресурсов информационного пространства. Часть ресурсов может быть категоризирована системой автоматически, например, научные публикации, музыка, видео и прочие ресурсы, категории для которых определены в сети. Другая часть – с участием человека, в таком случае частичное распределение по категориям производится пользователем, а системой, на основании этого распределения, предлагаются категории для новых ресурсов.

5. Возможность совместной работы. При ведении общего проекта, люди часто сталкиваются с необходимостью совместной работы над частью информации. Для такого случая, системы управления личной информацией должны поддерживать обмен информационными ресурсами, их метаданными или частичное объединение информационных пространств разных пользователей системы, а также предоставлять возможности для комментирования и обсуждения информационных ресурсов.

Таким образом, систему управления личной информацией можно рассматривать, как своеобразного интеллектуального цифрового помощника, сопровождающего и помогающего пользователю вести его информационное пространство. Дополняя сырые данные структурой и семантикой, пользователь получает возможность автоматизации, выполняемых им, интеллектуальных процессов.

3 Существующие подходы к управлению информацией

Несмотря на существование большого количество работ в области управления личной информацией, в настоящее время распространены лишь так называемые «персональные органайзеры», рассматривающие самые простые задачи, такие как планирование событий, установка напоминаний, ведение заметок и контактов, работа с электронной почтой. Наиболее популярными примерами таких органайзеров являются Microsoft Outlook, Mozilla Thunderbird и др. Предоставляя возможности для хранения ограниченных типов информации, эти средства, тем не менее, опускают вопросы управления накопленными сведениями – формирования взаимосвязей между данными, организации составных структур, совместной работы.

Рассмотрим основные направления работ, опубликованных в последние годы, по теме управления личной информацией. В работе «The Gnowsis Semantic Desktop for Information Integration» [16], описывается концепция *Semantic Desktop* - подхода к организации данных на персональном компьютере, в соответствии с которым любая информация, используемая пользователем, - файл, e-mail или событие календаря, рассматривается как RDF ресурс с собственным уникальным идентификатором. В этой работе вводится модель личной информации (Personal Information Model - PIMO) – формализующая ментальную модель информационного пространства, составленную пользователем. Основная задача PIMO – предоставить общую модель данных, с которой смогут работать различные приложения, используемые пользователем. Единое представление данных предоставит больше возможностей для организации более гибкой интеграции между приложений. К недостаткам Gnowsis можно отнести то, что основной акцент делается на организации модели данных и ее совместном использовании различными приложениями, в то время как вопросы управления накопленными данными почти не рассматриваются. На этих вопросах концентрируются работы SemEx

[7], IRIS [5], Haystack [13], MyLifeBits [9], DeepaMehta [15]. В SemEx, IRIS и Haystack данные представляются в иерархическом виде, в основе иерархии лежат наиболее распространенные типы данных, такие как email, контакты, проекты. В IRIS и Haystack для каждого типа данных определен набор интерфейсов, предоставляющих базовые операции, такие как возможность ответа или пересылки письма, создания события или напоминания. В Haystack, дополнительно предоставляется возможность настройки стандартных и создания собственных визуальных представлений данных, а также определяются программные интерфейсы для создания дополнительных операций над данными. В работах MyLifeBits и DeepaMehta, рассматриваются альтернативы к иерархическому подходу представления данных. В MyLifeBits, предлагаются интерфейсы отображения ресурсов с использованием временной шкалы. По мнению авторов, введение временной шкалы позволяет более наглядно отобразить ресурсы, а также увеличить количество одновременно выводимых ресурсов. В работе DeepaMehta, данные предоставляются пользователю в форме тематической карты (topic map) – ориентированного графа, узлами которого являются ресурсы, определенные пользователем.

В работе Beagle++ [6] подробно рассматриваются вопросы ранжирования ресурсов, полученных в результате поискового запроса. Ранжирование производится на основании объединения результатов, полученных с помощью алгоритмов ObjectRank и TF/IDF.

Работы iMecho [4] и Fledspar [3], рассматривают вопрос использования ассоциаций при поиске ресурсов. В iMecho, предлагается формировать журнал работы пользователя с ресурсами, который в дальнейшем анализировать для выделения зависимостей между ресурсами. В Fledspar, предоставляется удобный интерфейс для ассоциативной навигации по ресурсам, а также реализуется возможность осуществлять поиск ресурсов на основании информации связанной с ними. В работе Desktop Gateway [12], помимо интеграции между приложениями, также рассматриваются вопросы использования данных полученных из сети.

Перечисленные работы делают больший упор на формирование информационного пространства и работу с ним, в меньшей мере затрагивая вопросы анализа данных и автоматизации действий, выполняемых пользователем. В них слабо освещены вопросы объединения информационных пространств различных пользователей и совместной работы с ними.

Среди коммерческих систем можно выделить Google Now [10] и Dropbox Datastore [8]. Основной задачей Google Now является отображение нужной информации в нужный момент. Основываясь на истории совершенных ранее действий, а также на текущем местоположении и моменте времени, Google Now предоставляет пользователю релевантную информацию, такую как прогноз погоды, пробки и др. Dropbox Datastore обеспечивает возможность хранения структурированной информации в «облаке» Dropbox. Основной структурой в Dropbox Datastore являются таблицы, для чтения и записи в них предоставляется программные интерфейсы.

4 Предлагаемое решение

В данном разделе приводится описание архитектуры предлагаемого решения, а также проводится сравнение двух схем, соответствующих разным моделям работы пользователей с данными, одна из которых не использует систему управления информацией (рис. 1), а другая – использует (рис. 2). На схемах, рассматривается работа двух пользователей, чьи информационные пространства частично пересекаются. В процессе работы каждый пользователь пополняет собственное информационное пространство, состоящее из разнородных данных – документов, событий, писем. Работа с ресурсами ведется по средствам различных приложений и информационных систем.

Без использования системы, информационные пространства пользователей нигде формально не определены, только сам пользователь может определить, какие данные связаны между собой и как именно. Вследствие этого, все дальнейшие взаимодействия с данными – поиск, совместная работа, формирование иерархической структуры, могут производиться лишь в рамках того приложения, которое отвечает за конкретный тип ресурсов. Поскольку информация делится между различными приложениями, большое количество метаданных о ресурсах, таких как иерархическая организация, связи и зависимости между ресурсами, могут дублироваться в каждом из них. Системы управления личной информацией вводят дополнительный уровень организации данных, позволяют пользователю явно определить своей информационное пространство и предоставляют интерфейсы для работы с ним. Сведения хранятся в системе в соответствии с форматами, определенными в OWL онтологии, за счет чего, к слабоструктурированным данным добавляется семантика, а также появляется возможность производить их автоматический анализ, категоризацию и индексацию. Также важным моментом является то, что пользователи могут осуществлять работу с данными с помощью привычных для них приложений, т.к. по средствам адаптеров и агентов информация из внешних источников может быть автоматически выгружена в систему. Дополнительно, поскольку система по своей сути является многопользовательским приложением, в рамках нее возможна совместная работа различных пользователей с общим информационным пространством.

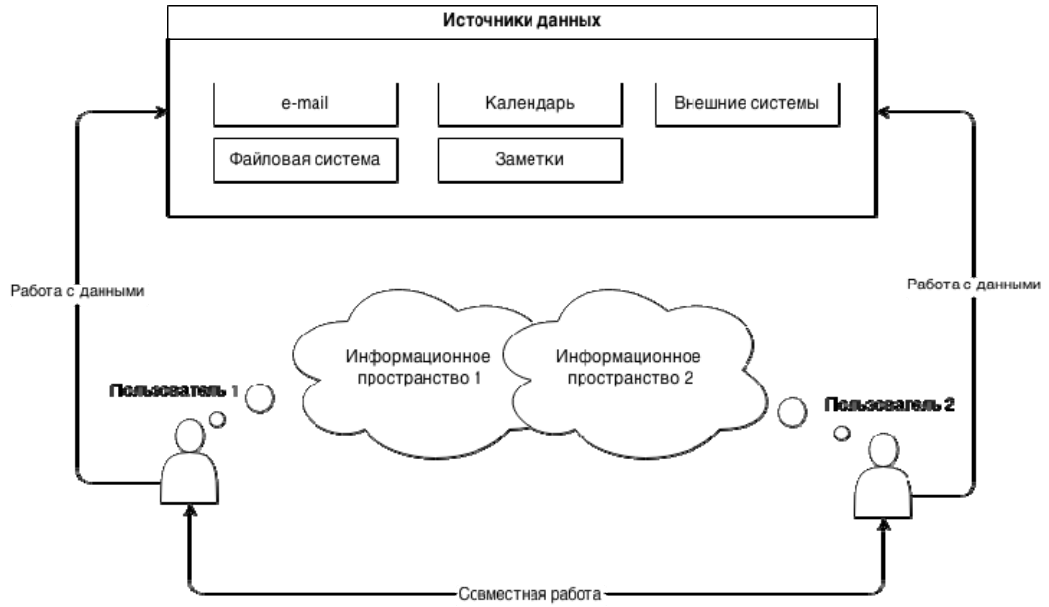

Рис. 1. Работа пользователей без использования системы

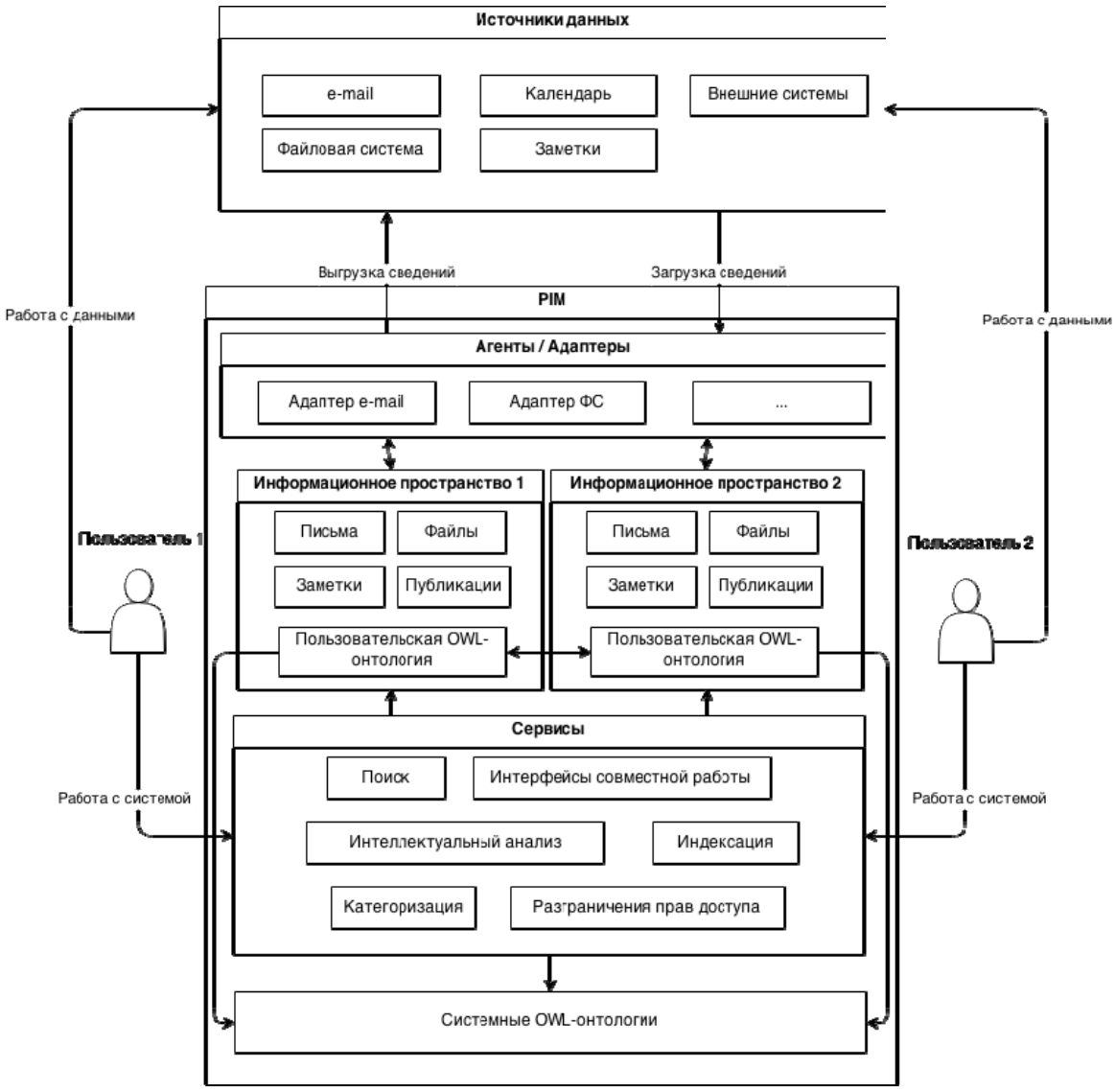

Рис. 2. Работа пользователей с использованием системы

5 Информационное пространство пользователя

Информационное пространство – это совокупность всех сведений, с которыми человек работает сейчас или работал ранее. Любые данные, имеющие важность для человека, и выделяемые им в отдельную сущность рассматриваются как элементы информационного пространства – информационные ресурсы. Информационными ресурсами могут быть файлы, заметки, посещенные веб страницы, письма и пр.

В рамках системы, в качестве информационных ресурсов, рассматриваются RDF ресурсы. Таким образом, каждый пользователь в рамках системы ведет работу с собственным RDF репозиторием данных, представляющим его информационное пространство. Структура информационного пространства описывается с помощью OWL онтологии пользователя. По умолчанию, предоставляется системная онтология, которую, при необходимости, пользователь может изменять, добавляя новые классы и свойства, а также расширяя уже имеющиеся. Наполнение репозитория происходит либо по средствам автоматического импорта информации, либо вручную, через пользовательский интерфейс системы.

6 Категоризация

Деление файлов на различные иерархические категории, является одним из основополагающих процессов, используемых пользователями при работе с информацией на персональных компьютерах. Тем не менее, исследования [18] показывают, что, несмотря на понимание того, что категоризация в дальнейшем может существенно облегчить поиск, многие пользователи, по различным причинам, игнорируют эту возможность. В качестве объяснения такого поведения пользователи обычно ссылаются на сложности при принятии решения, в какую из категорий отнести файл, проблемы при формировании подкатегорий, таких, чтобы их содержимое не пересекалось, а также на нехватку времени. Поэтому возможность автоматической или полуавтоматической категоризации сведений попадающих в систему, является крайне важной.

Необходимым элементом, для проведения категоризации, является выбор категорий, на которые будут делиться ресурсы. В ряде случаев возможно выбрать их полностью автоматически – это относится к ресурсам, набор категорий для которых, может быть получен из внешних источников. Например, научные статьи делятся на категории на основании тематики работы, музыка и фильмы на основании жанров и направлений.

Поскольку в общем случае, автоматическое выделение категорий не возможно, формировать требуемые классы можно по мере работы пользователя с системой. В таком случае категоризация может проводиться по двум направлениям:

- выделение в общие категории ресурсов, находящихся в общих разделах иерархической структуры, организованной пользователем;
- выделение в общие категории на основании добавленных пользователем метаданных, таких как «теги».

За счет такой категоризации, при добавлении нового ресурса в систему, на основании уже внесенных пользователем сведений, пользователю будет предложено возможное расположение нового ресурса в иерархической структуре, а также метаданные, которые могут быть к нему добавлены.

7 Интеллектуальный анализ данных

Одной из наиболее важных функций в системах управления личной информацией, является возможность хранить метаданные о созданных в системе ресурсах. Хорошо известно, что обычно, люди забывают или затрудняются заносить метаданные вручную, поэтому важно, чтобы система могла сформировать максимальное количество метаданных в автоматическом режиме. Как было описано выше, часть метаданных формируется адаптерами к источникам данных, на основании содержимого импортируемого ресурса. Кроме того, в ряде случаев, можно получить дополнительную информацию из глобальной сети, для этого системой предоставляется ряд *адаптеров к внешним репозиторям*. Адаптеры осуществляют поиск «аналогов», для имеющихся в системе ресурсов, в различных репозиториях глобальной сети (в частности, в репозиториях Linked Open Data), и, в случае успеха, переносят соответствующую информацию из найденных ресурсов в репозиторий пользователя. Каждый адаптер отвечает за поиск «аналогов» одного или нескольких классов OWL-онтологии пользователя. Другим аспектом анализа данных является поиск похожих или связанных ресурсов. Для каждого ресурса, находящегося в информационном пространстве пользователя, осуществляется поиск связанных с ним ресурсов, как внутри информационного пространства, так и вне его – в глобальной сети. Алгоритмы поиска схожих ресурсов могут сильно отличаться в зависимости от класса искомого ресурса. Поэтому, за поиск схожих ресурсов для разных классов отвечают различные компоненты.

8 Реализация

В рамках данной статьи реализован прототип системы, поддерживающий хранение и анализ научных публикаций. Прототип соответствует описанной выше архитектуре. Уровень адаптеров данных представляет адаптер файловой системы. На уровне сервисов реализован модуль анализа публикаций, осуществляющий пополнение репозитория данными из Академии Google (Google Scholar), а также выполняющий поиск новых публикаций, схожих с загруженными ранее.

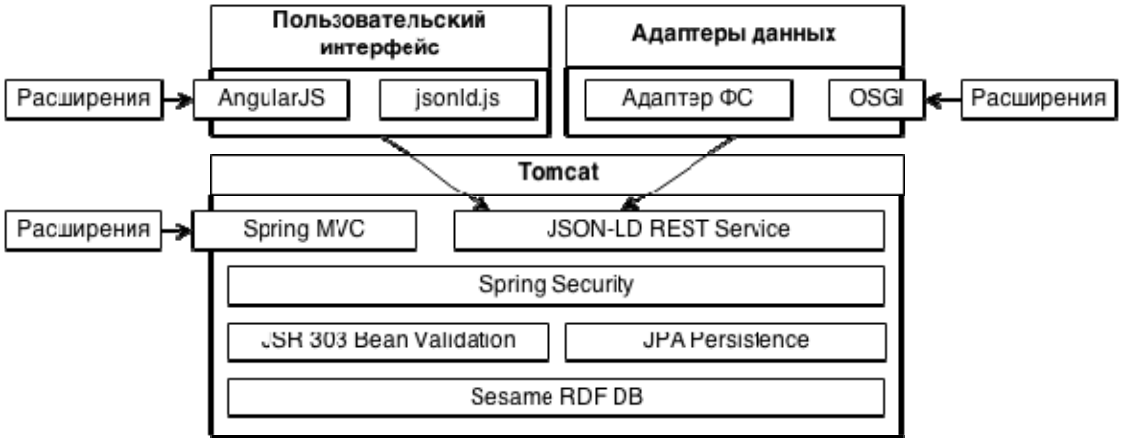

Рис. 3. Архитектура системы

На рис. 3 представлена модель реализованного приложения. Большой упор при реализации делался на расширяемость системы. Серверная часть системы выполнена на языке Java и представляет собой веб приложение, реализованное с использованием библиотеки Spring Framework [17]. В качестве хранилища данных используется RDF-база данных Sesame. Для получения и записи данных в хранилище системой предоставляется REST веб-сервис, использующий формат JSON-LD [11] для представления RDF-данных. Расширение системы возможно как на клиентской, так и на серверной стороне. Использование библиотеки AngularJS [2] предоставляет гибкие возможности для расширения пользовательского интерфейса системы, добавления новых визуальных представлений данных и изменения системных интерфейсов. Для расширения серверной части приложения используется стандарт OSGI [14].

Адаптер файловой системы представляет собой приложение, запущенное на компьютере пользователя, которое отвечает за передачу скачанных пользователем публикаций на удаленный сервер. Приложение работает в фоновом режиме и, при изменениях содержимого папок, передает информацию об этом на сервер. На стороне сервера, на основании текстового содержимого выгруженных файлов, осуществляется поиск статей в Академии Google. В случае успешного поиска, в систему заносятся метаданные, такие как название, год, авторы, а также ссылки на схожие статьи.

Работа пользователя с информационным пространством осуществляется через веб-интерфейс. На рисунке 4 представлен скриншот пользовательского интерфейса системы. Верхнее меню отвечает за навигацию по типам ресурсов, а также предоставляет возможность поиска ресурсов по системе. Рабочая область портала поделена на три блока. В левой части выводится навигационное меню, отображающее папки, синхронизованные с системой. В центральном блоке выводится список публикаций, находящихся в выбранной папке. Публикации, находящиеся в центральном блоке, могут быть отсортированы по средствам интерфейса Drag&Drop. В правой части рабочей области выводится информация, о выбранной статье – схожие статьи и детальная информация. В панели схожих статей, выводятся ссылки на статьи, схожие с выбранной публикацией. На панели детальной информации, выводятся метаданные о выбранной публикации. Все поля, выводимые на панели детальной информации, при необходимости, могут быть отредактированы пользователем.

9 Заключение

В данной статье были рассмотрены задачи управления личными знаниями и информацией, описана архитектура системы, автоматизирующая основные процессы, возникающие в ходе выполнения этих задач, представлен прототип, соответствующий описанной архитектуре. Основные направления, по которым проводится автоматизация это: структурирование информации, поиск информации с которой ранее велась работа, категоризация информации, пополнение сведений метаданными, полученными из внешних источников, поддержка совместной работы. В качестве реализации был представлен прототип системы, поддерживающий работу с научными публикациями. В рамках прототипа реализованы модули каждого из описанных уровней архитектуры – адаптер к файловой системе, сервис анализа публикаций, сервис поиска. За хранение ресурсов, загруженных пользователями, отвечает RDF-база данных Sesame.

В дальнейшем, большее внимание, планируется уделить созданию формальной модели системы, явно описывающей основные модули и процессы, выполняемые в рамках системы. Другим направлением работ, является введение элементов логического вывода, в рамках анализа информационного пространства. Поскольку информация хранится в формате RDF и описывается с помощью языка OWL, принципиальных ограничений в этом вопросе нет. Также, отдельным вопросом, заслуживающим изучения, является анализ процесса работы пользователя с информацией. Анализируя выполняемые

пользователем действия, можно выявлять скрытые зависимости между ресурсами, а также формировать метаданные, которые в дальнейшем могут быть использованы пользователем, например при поиске. Исследования по данному направлению обычно исследуются в рамках работ Task Mining.

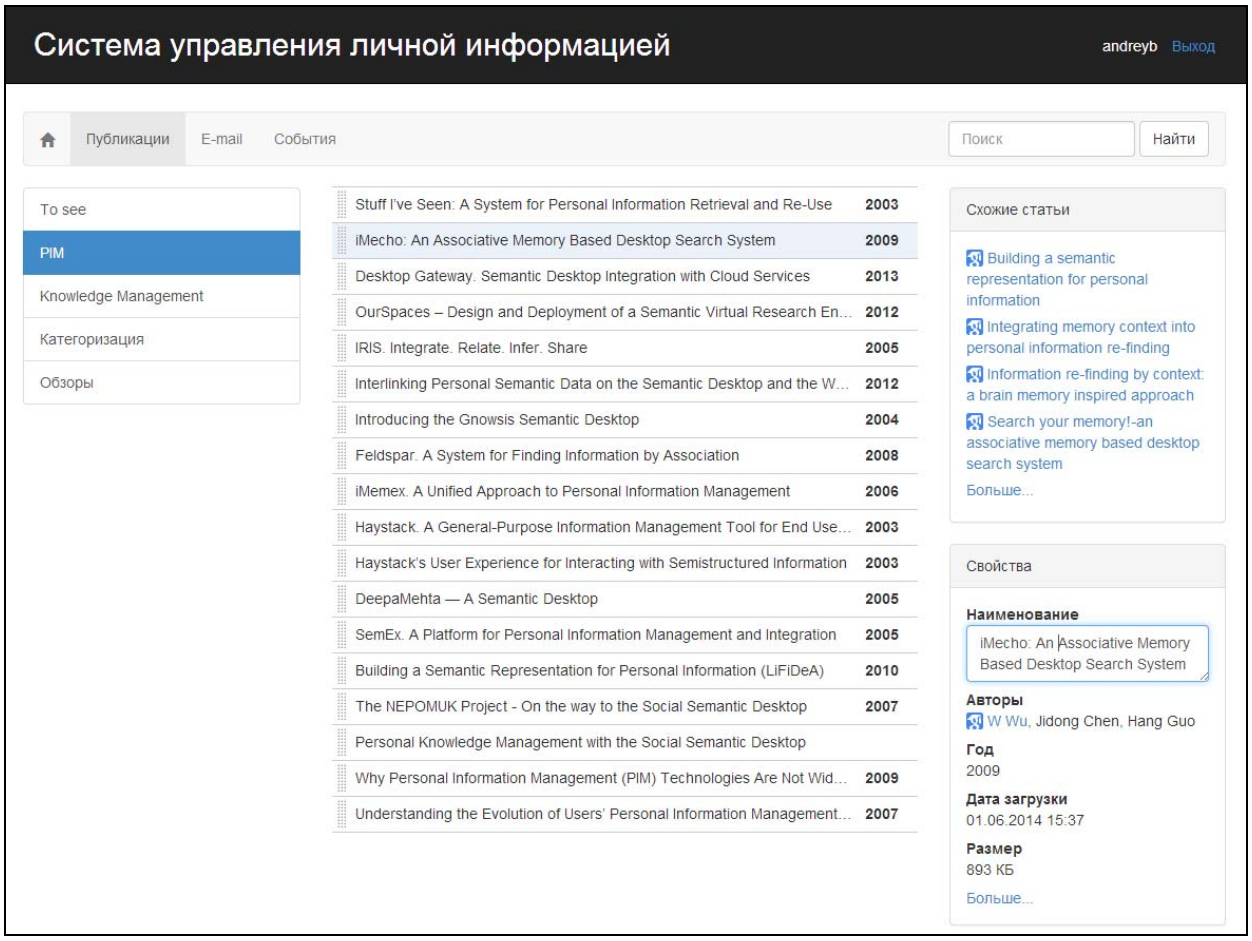

Рис. 4. Пользовательский интерфейс системы

Литература

[1] AngularJS. https://angularjs.org/

[2] Bush V. As We May Think // The Atlantic Monthly. – Atlantic Media Company, Washington, DC, USA 1945. – V. 176 – P. 101–108.

[3] Chau D. H., Myers B., Faulring A. What to Do when Search Fails: Finding Information by Association // Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Florence, Italy, April 5–10, 2008 – / ACM – New York, NY, USA, 2008. – P. 999–1008.

[4] Chen J., Guo H., Wu W., Wang W. iMecho: an associative memory based desktop search system // CIKM '09: Proceedings of the 18th ACM Conference on Information and Knowledge Management, Hong Kong, China, November 02–06, 2009. – / ACM – New York, NY, USA, 2009. – P. 731–740.

[5] Cheyer A., Park J., Giuli R. IRIS: Integrate, Relate. Infer. Share // Proceedings of the Semantic Desktop Workshop at ISWC Galway, Ireland, November 6, 2005. – 2005. – ISSN 1613-0073. http://ceur-ws.org/Vol-175/17_park_iris_final.pdf

[6] Chirita P. A., Costache S., Nejdl W. et al. Beagle++: Semantically enhanced searching and ranking on the desktop // The Semantic Web: Research and Applications. – Springer Berlin Heidelberg, 2006. – P. 348–362.

[7] Dong X. L., Halevy A. A platform for personal information management and integration // In Proceedings of CIDR 2005, Asilomar, CA, USA, January 4-7, 2005. – P. 119–130. http://www.cidrdb.org/cidr2005/papers/P10.pdf

[8] Dropbox Datastore. https://www.dropbox.com/developers/datastore

[9] Gemmell J., Bell G., Lueder R. et al. MyLifeBits: fulfilling the Memex vision // Proceedings of the tenth ACM international conference on Multimedia, Juan les Pins, France, December 1–6, 2002 /Association for Computing Machinery – New York, NY, USA, 2002. – P. 235–238.

[10] Google Now. http://www.google.com/landing/now/

[11] JSON-LD. A JSON-based Serialization for Linked Data. http://www.w3.org/TR/json-ld/

[12] Kareski A., Jovanovik M., Trajanov D. Desktop Gateway: Semantic Desktop Integration with Cloud Services // BCI'13: Proceedings of the 6th Balkan Conference in Informatics, Thessaloniki, Greece, September 19–21, 2013 / ACM – New York, NY, USA, 2013. – P. 162–168.

[13] Karger D. R., Bakshi K., Huynh D. et al. Haystack: A General-Purpose Information Management Tool for End Users Based on Semistructured Data // Proceedings of CIDR 2005, Asilomar, CA, USA, January 4–7, 2005. – 2005. – P. 13–26. http://www.cidrdb.org/cidr2005/papers/P02.pdf

[14] OSGI. Open Service Gateway Initiative. http://www.osgi.org/Specifications/HomePage

[15] Richter J., Völkel M., Haller H. DeepaMehta – A Semantic Desktop // In Proceedings of the Semantic Desktop Workshop at ISWC Galway, Ireland, November 6, 2005. – 2005. – ISSN 1613-0073. http://ceur-ws.org/Vol-175/30_dm_poster.pdf

[16] Sauermann L., Grimnes G.A., Kiesel M. et al. Semantic desktop 2.0: The gnowsis experience // The Semantic Web – ISWC 2006. – Springer Berlin Heidelberg, 2006. – P. 887–900.

[17] Spring Framework. http://docs.spring.io/spring/docs/4.1.0.RC1/spring-framework-reference/htmlsingle/

[18] Voit K., Andrews K., Slany W. Why personal information management pim technologies are not widespread // ASIS&T 2009 Workshop on Personal Information Management, November 7–8, 2009, Vancouver, BC, Canada – 2009. http://pimworkshop.org/2009/papers/voit-pim2009.pdf

Model of Semantic Personal Information Management System

Andrey A. Bezdushny, Anatoly N. Bezdushny, Vladimir A. Serebryakov

This paper considers the problem of personal information management by using semantic technologies, proposes the architecture of semantic personal information management systems and presents the prototype of the system implemented in accordance with this architecture. Proposed method develops the idea of the Semantic Desktop – an approach to the personal information space organization in accordance with the Semantic Web and Linked Open Data.