2004 - Том 7 - Выпуск 1

# О реализации веб-системы математической информации

***Аджиев А.С.***
**ЦНТК РАН**

***Бездушный А.Н., Серебряков В.А.***
**ВЦ РАН**

На основе проведенного ранее анализа российских математических электронных ресурсов, а так же опыта зарубежных математических информационных систем описан проект создаваемой математической информационной системы Math-Net.RU. Базовой платформой системы Math-Net.RU является универсальная информационная система ИСИР.

Проект описан в терминах перечня требований и условий, которым должна удовлетворять создаваемая система. Рассмотрены и проанализированы альтернативные варианты реализации различных компонент системы, а также пути решения возникающих при этом проблем. Очерчены категории хранимой информации, целевой круг пользователей системы и требуемая функциональность. Описана общая архитектура, схема данных, пользовательские интерфейсы, а также способы наполнения системы информацией, актуализации и синхронизации данных из других информационных систем и баз данных. Рассмотрены проблемы представления математических текстов и формул в информационных системах, дан сравнительный анализ существующих форматов хранения. Очерчены так же перспективы участия системы Math-Net.RU в создаваемой Всемирной математической информационной системе Math-Net, а также требования к системе-участнику.

## *Постановка задачи*

### Предпосылки

Проведенная исследовательская работа показала, что существующие российские электронные информационные ресурсы в области математики слабо систематизированы и разрознены, как логически, так и физически. Математическая информация в целом слабо представлена для доступа в Сети. Хотя в ряде организаций и проводится работа по публикации данных в сети Internet, но эти представления информации преимущественно статические, плохо структурированы, обладают разнородными интерфейсами и форматами, и не всегда имеют средства поиска. В России отсутствует, как таковая, единая специализированная общенациональная система дистанционного поиска научной

математической информации и доступа к ней [12].

Таким образом, основными задачами, определяющими дальнейшее развитие этой информационной инфраструктуры, является не только ее информационное наполнение, но и интеграция существующих и вновь создаваемых математических ресурсов в единую интегрированную информационную систему.

## Требования к системе

Для удовлетворения информационных потребностей российских математиков необходимо создание общедоступного через Internet общероссийского математического информационного портала, удовлетворяющего весь спектр информационных и коммуникационных потребностей российских математиков, а также лиц, обучающихся математике и интересующихся математикой.

Создаваемая система будет включать в себя различные компоненты, удовлетворяющие различным потребностям российских математиков, и охватывающих в совокупности всю российскую математическую жизнь. На первом этапе предполагается создать систему, решающую более узкий круг задач, а именно, удовлетворяющую основные потребности математиков в получении существующей в настоящий момент традиционной информации научного характера, то есть, информации о математических публикациях, организациях, персонах, проектах, грантах, конференциях, семинарах, программном обеспечении и web-ресурсах.

Создаваемая информационная система будет основана на технологиях и принципах построения информационной системы ИСИР [9]. В связи с этим, ниже в этой статье не рассматриваются теоретические аспекты и принципы функционирования информационной системы как таковой, а внимание уделено приложению этих принципов для создания российской математической информационной системы.

Ниже рассмотрены требования, определяющие структуру создаваемой информационной системы.

### Пользователи системы

Предполагается, что пользователями системы будут российские и иностранные математики, а также аспиранты и студенты, выбравшие научную работу в области математики в качестве своей будущей профессии. Это означает, что, по крайней мере, часть функций создаваемой системы должна быть доступна широкому кругу любых заинтересованных пользователей, считающих себя в той или иной степени математиками.

Пользователями системы также будут административные работники Отделения математики РАН.

### Информация, хранимая и доступная в системе

В системе будет храниться любая информация, необходимая для обеспечения

научной работы или обучения в области математики. Основными типы хранимых информационных ресурсов являются:

1. *Персона.* Описывает человека, как ученого, или административного работника.
2. *Публикация.* Описывает произвольный носитель математической научной информации в виде структурированного математического текста, предназначенного для передачи математических знаний читателям. Например, статья в журнале, журнал, книга, компакт-диск, электронная публикация. В эту группу не включаются другие математические web-ресурсы (кроме электронной публикации).
3. *Организация или подразделение.* Описывает научную математическую, организацию или подразделение, а также любую другую организацию или подразделение, деятельность которой связана с научной работой ученых-математиков. Например, издательства, или административные структуры ОМ РАН.
4. *Проект или грант.* Описывает любые проекты или гранты, в рамках которых осуществляется научная работа в области математики.
5. *Конференция или семинар.* Описывает любую официальную регулярную или нерегулярную встречу ученых математиков с целью обмена научной информацией, например, конференцию или семинар.
6. *WEB-ресурс*. Описывает любой web-ресурс, полезный для ученого-математика в его научной работе (кроме электронных публикаций, поскольку их целесообразнее регистрировать как ресурсы типа публикация).
7. *Программное обеспечение.* Научное программное обеспечение, пакеты программ и библиотеки.

**Функции и возможности системы**

Система должна обеспечивать:

1. Поиск ресурсов всех перечисленных выше типов по ключевым словам в значениях их атрибутов, регулярным выражениям и сложным поисковым запросам.
2. Навигацию в пространстве связанных ресурсов по имеющимся связям между ресурсами, а также по рубрикам иерархических тематических рубрикаторов.
3. Разграничение прав доступа к информации между разными категориями пользователей.
4. Возможность пользователям системы самим предоставлять информацию для опубликования в системе или корректировки имеющейся информации. При этом необходимо разграничение прав доступа, а также возможность эффективной обработки вводимой информации редакторами.
5. Пакетный ввод информации разного уровня структурированности из электронных источников, таких как:
    1. Базы данных.
    2. Структурированный текст.
    3. Web-сайты.

6. Актуализация информации из других баз данных.
7. Участие во всемирной математической информационной системе Math-Net в качестве российского узла.

# *Реализация системы*

## Архитектура

Поскольку система должна интегрировать в себе информацию из различных источников, то для обеспечения эффективной поддержки системы, в частности, решения задач актуализации данных, архитектура должна быть распределенной. Это означает прежде всего распределенность хранения информации, т.е. должна допускаться возможность физического хранения информации в разных географически удаленных базах данных, имеющих разную структуру и поддерживаемых разными командами независимо друг от друга. Такие базы данных должны обеспечивать поддержку единых открытых интерфейсов поисковых запросов системы, чтобы пользователь мог осуществлять поиск и навигацию по всем базам данных системы одновременно.

Этим требованиям вполне соответствует разрабатываемая в настоящее время распределенная архитектура ИСИР, которая и будет использована при реализации математической информационной системы.

## Модель данных

Схема данных была выработана на основе результатов анализа российских электронных математических ресурсов, потребностей российских математиков и на основе обобщения опыта европейских и американских математических информационных систем [12].

Модель данных была разработана на основе и в терминах объектной модели данных OWL/RDF [14], то есть представляет собою фиксированный набор типов (классов) ресурсов. Каждый класс характеризуется набором характеризующих его атрибутов и допустимых связей с другими ресурсами.

Как уже упоминалось, предполагается хранить в системе 7 типов информационных ресурсов, а также информацию о связях между ресурсами.

Многие ресурсы могут быть отражены несколькими способами в этой модели данных. Например, web-сайт организации может быть представлен, как атрибут соответствующей организации, или быть включенным в систему как самостоятельный ресурс типа “web-ресурс”, если он интересен как информационный ресурс безотносительно к организации, которой он принадлежит.

### Уровни детализации описаний ресурсов в схеме данных

При разработке модели данных необходимо, прежде всего, установить уровень детализации при описании каждого типа ресурса. В простейшем случае любой

ресурс можно описать одним текстовым атрибутом, в значении которого будет изложена в свободной форме вся информация о ресурсе. В более сложных случаях одному описываемому объекту может соответствовать в схеме данных целая совокупность ресурсов разных типов, связанных между собою разными связями. Примером может служить приведенный ниже эскиз модели описания конференции в системе поддержки конференций, разрабатываемой в рамках проекта ИСИР.


Высокая детализация описания ресурса обладает следующими преимуществами:

1. Высокие возможности поиска и навигации в пространстве ресурсов (например, в приведенном выше примере можно осуществить поиск всех конференций, в которых определенное лицо входило в оргкомитет).
2. Высокие возможности по преобразованию информации к произвольному формату представления или к произвольной модели данных. Это особенно важно при обмене данными с другими системами.
3. Возможность реализации гибкой развитой системы поддержки типа ресурса, обеспечивающей ввод, манипулирование данными на основе потока работ.

Недостатками высокой детализации описания являются:

1. Сложность и высокая трудоемкость преобразования загружаемой в систему информации к детализированной схеме данных.
2. Неизбежные ошибки и потери информации при преобразовании к сложной, детализированной схеме данных. Например, приводя описания конференций в виде текста в свободной форме к описанной выше модели, вряд ли удастся обойтись без большого объема ручного труда и без ошибок.

   Пожалуй, наиболее общими и неизбежными ошибками будут ошибки идентификации описанных ресурсов. Например, если в двух конференциях участвует некто Иванов И. И., и нам ничего о нем не известно, кроме

фамилии и инициалов, в приведенной выше модели невозможно достоверно определить, стоит ли привязывать 2 доклада с этой фамилией к одной и той же персоне, или к разным.

3. Сложность логики и реализации такой системы.

Таким образом, уровень детализации и конкретный набор атрибутов и связей для ресурсов в системе должен быть разумным компромиссом между:

1. Запросами пользователей системы (детализация должна позволять выполнять наиболее востребованные пользователями запросы к системе).
2. Возможностью преобразования данных к схемам данных для обмена с другими системами.
3. Уровнем детализации загружаемых в систему данных.
4. Требованиями к допустимому проценту ошибок в данных системы.
5. Возможностями персонала поддержки системы по детализации и интеграции и контролю ошибок в загружаемых в систему данных.
6. Возможностью бесшовного (seamless) повышения уровня детализации, т.е. с обеспечением обратной совместимости по данным.

Очевидно, что нет смысла делать детализацию описания ресурсов ниже, чем детализация в большинстве источников данных системы, поскольку потеря детализации – это потеря данных и потеря возможностей поиска и трансформации.

В то же самое время излишняя детализация, мало востребованная пользователями в поисковых запросах, будет лишь порождать ошибки в данных при идентификации ресурсов.

Оптимальным, по-видимому, будет подход, при котором на первом этапе реализуется минимальная детализация, определяемая уровнем детализации источников информации, а также требованием обеспечения обработки основных пользовательских поисковых запросов, но обеспечивающая бесшовное (seamless) повышение уровня детализации.

Впоследствии, по мере развития средств детализации загружаемой и уже загруженной в систему информации, детализацию описания ресурсов можно увеличивать, расширяя, таким образом, возможности поиска, навигации и обмена с другими системами.

**Тематическая классификация ресурсов**

Как отмечалось выше, в российской математике для тематической классификации ресурсов традиционно используется международная система тематической классификации публикаций УДК (UDC) [4], а также математический рубрикатор MSC [3], созданный Американским математическим обществом (AMS), и получившим широкое всемирное признание.

Исследования показали, что российские математики в целом неплохо знакомы с

обеими этими системами тематической классификации, и знают коды MSC и основной таблицы УДК, соответствующие тематике их работы.

Для тематического обозначения специальности персон в России используется также рубрикатор ВАК. Все проекты, поддерживаемые Российским фондом фундаментальных исследований (РФФИ) также классифицируются тематическим рубрикатором РФФИ.

Рубрикатор MSC имеет древовидную иерархическую структуру, а также содержит некоторые горизонтальные связи, характерные для одноязычного тезауруса.

Код УДК является сложной синтаксической конструкцией, которая описывает коды из основной таблицы УДК, определители, и также соотношения между ними, наиболее полно отражающие тематику ресурса. Например, код

621.923.014.5-185.4:[621.922.023:621.921.34](597+598)"18" обозначает тематику «Высокоскоростное шлифование алмазными брусками в Лаосе и Вьетнаме в 19 веке».

Основная таблица, а также таблицы определителей, имеют, также как и MSC, иерархическую древовидную структуру с горизонтальными связями одноязычного тезауруса.

Задача реализация полнофункциональной поддержки УДК, включая поиск по основному коду и определителям, является самостоятельной сложной задачей, и не будет решаться на первом этапе создания системы.

В то же время, основная таблица УДК сама может являться тематическим рубрикатором, который можно использовать для классификации ресурсов всех типов. Преимуществом такого рубрикатора, определяющим необходимость его использования, является знание большинством математиков кодов основной таблицы УДК, соответствующих их специальности.

Таким образом, в качестве основных тематических рубрикаторов в системе будет использован рубрикатор MSC, и основная таблица УДК, реализованные как тезаурусы-классификаторы [13]. В перспективе будет реализована полная поддержка системы классификации УДК.

Кроме этого, в системе для тематической классификации специальностей персон будет использован рубрикатор ВАК, а для проектов – рубрикатор РФФИ.

**Принципы построения схемы данных системы**

Как уже отмечалось выше, степень детализации описаний ресурсов, а именно набор атрибутов и связи между ресурсами системы, выбирались исходя из выполнимости необходимых пользователю поисковых запросов, а также исходя из характера доступной в перспективе для загрузки в систему информации.

Анализ того, какие именно поисковые запросы и навигационные возможности в

действительности необходимы пользователям, затруднительно провести непосредственно. Потому за основу была взята совокупность поисковых и навигационных возможностей, реализованных в известных зарубежных математических информационных системах, имеющих большой опыт работы и обратной связи с пользователями. Логично предположить, что реализованные в них поисковые возможности перекрывают, как минимум, наиболее насущные потребности математиков.

Таким образом, в отдельные атрибуты и связи в схеме данных были выделены именно те атрибуты ресурсов, по которым, как можно ожидать, исходя из указанных выше соображений, в наибольшей степени будет востребован атрибутный поиск и агрегирование информации.

Кроме того, при выборе набора атрибутов была учтена также детализация описаний ресурсов в схемах данных систем, с которыми в перспективе может быть налажен обмен данными. Прежде всего, схема данных создающейся международной системы Math-Net, а также детализация описания ресурсов в стандартных международных форматах, таких как, например, DublinCore и VCard, с целью обеспечения по возможности легкого и обратимого преобразования данных к этим форматам.

Приведенные в модели списки атрибутов ресурсов и свойств связей являются предварительными и могут незначительно меняться.

**Персона**

Атрибуты:

1. *Полное имя по-русски.*
2. *Полное имя по-английски* (возможно, несколько вариантов написания).
3. *Дата рождения* (возможно, также и место рождения и дата смерти).
4. *Ученая степень + год присуждения* (возможно, дата присуждения). Несколько значений.
5. *Ученое и академическое звания* (возможно, также дата присуждения). Несколько значений.
6. *Ключевые слова, характеризующие направления деятельности.* Несколько значений на разных языках.
7. *Почтовый адрес.* Несколько значений, возможно, на разных языках.
8. *Телефон.* Несколько значений.
9. *Факс.* Несколько значений.
10. *e-mail.* Несколько значений.
11. *Персональная страница.*
12. *Направления деятельности.* Текстовое описание.
13. *Биография.* Текст.
14. *Прочая информация.* Текст.
15. *Фотография.*

Связи с рубрикаторами и тезаурусами.

1. *Коды MSC.* Несколько значений.
2. *Специальность ВАК.* Несколько значений.
3. *Код основной таблицы УДК.* Несколько значений.

Связи с другими ресурсами

1. *Публикации.* Является автором, редактором.
2. *Проекты и гранты.* Является руководителем или участником.
3. *Организации и подразделения.* Занимает определенную должность. Связь имеет атрибут *название должности*.

**Публикация**

Атрибуты:

1. *Название.* Несколько значений на разных языках.
2. *Тип публикации.* Например, монография, журнал, выпуск журнала, статья, труды конференции, доклад на конференции или семинаре, электронная публикация.
3. *Библиографическое описание.* В соответствии с ГОСТ.
4. *Язык.*
5. *Полный код УДК.*
6. *Идентификаторы* (ISBN, ISSN, DOI). Возможно несколько значений.
7. *Специальные идентификаторы.* Например, идентификаторы этой публикации в системах ZentralblattMATH [6] и MathSciNet [5]. Такие идентификаторы могут быть использованы авторами других публикаций для ссылок на данную публикацию, а также для доступа к рефератам на эту публикацию, сделанными сотрудниками этих систем.
8. *Дата опубликования.*
9. *Авторская аннотация.* Несколько значений на разных языках.
10. *Оглавление.* Несколько значений на разных языках.
11. *Ключевые слова, характеризующие содержание публикации.* Несколько значений на разных языках.
12. *Реферат.* Несколько значений, возможно, на разных языках. Кроме текста, необходимо поддерживать хранение в виде файлов в специальных форматах (подробнее об этом будет написано ниже).
13. *Полный текст.* Несколько значений. Может быть представлен файлом или URL.
14. *Информация о включении.* Информация, как именно соотносится публикация с вышестоящей по иерархии. Если публикация – статья в журнале, вышестоящая публикация – выпуск этого журнала, содержащий эту статью, то данное поле может содержать номера страниц, где она расположена.

Связи с рубрикаторами и тезаурусами:

1. *Коды MSC.* Несколько значений.
2. *Коды таблиц УДК.* Проставляются в результате синтаксического анализа полного кода УДК.

Связи с другими ресурсами:

1. *Публикация.* Возможны следующие типы связей:
    1. *Составная часть.* Например, статья является составной частью выпуска журнала или трудов конференции.
    2. *Ссылка одной публикации на другую.*
2. *Персона.* Является автором или редактором этой публикации.
3. *Организация.* Возможны следующие типы связей:
    1. *Издала публикацию.*
    2. *Является коллективным автором этой публикации.* Например, труды организации.
4. *Проект или грант.* Публикация выполнена в рамках данного проекта.
5. *Конференция или семинар.* Публикация является трудами конференции или сборником семинара.
6. *Программное обеспечение.* В публикации описано данное программное обеспечение.

**Организация или подразделение**

Организация и подразделения имеют следующие атрибуты:

1. *Название.*
2. *Сокращенное название.*
3. *Почтовый адрес.* Несколько значений, возможно. На разных языках.
4. *Телефон.* Несколько значений.
5. *Факс.* Несколько значений.
6. *e-mail.* Несколько значений.
7. *Историческая справка.* Несколько значений на разных языках.
8. *Направления деятельности.* Несколько значений на разных языках.
9. *Ключевые слова, характеризующие тематику и направления деятельности орг. единицы.* Несколько значений на разных языках.
10. *WWW-страница.*
11. *Вторичная страница Math-Net* [7]. Содержит URL вторичной страницы организации или подразделения, сделанной в соответствии со стандартами международной системы Math-Net. Именно эта страница должна выдаваться пользователями Math-Net. Если это поле не заполнено, при запросе из Math-Net выдается URL на скрипт, динамически генерирующий вторичную страницу для этой организации.
12. *Логотип* (или фотография здания, если логотип отсутствует).

Связи с рубрикаторами и тезаурусами:

1. *Коды MSC.* Несколько значений.
2. *Код основной таблицы УДК.* Несколько значений.

Связи с другими ресурсами:

1. *Персона.* Занимает определенную должность. Связь имеет атрибут *название должности*, а также, возможно, контактную информацию,

характеризующую скорее рабочее место, чем саму персону (рабочий телефон, факс, номер офиса, и т.д.).
2. *Публикация.* Возможны следующие типы связей:
    1. *Издала публикацию.*
    2. *Является коллективным автором публикации.*
3. *Организация или подразделение.* Организация или подразделение административно подчинены или входят в состав организации или подразделения.
4. *Проект или грант.* Организация или подразделения участвует в этом проекте или гранте.
5. *Конференция или семинар.* Эта организация или подразделение организует или проводит данную конференцию или семинар.

**Проект или грант**

Проект или грант имеет следующие атрибуты:

1. *Название.* Несколько значений на разных языках.
2. *Даты начала и окончания.*
3. *Описание.* Несколько значений на разных языках.
4. *Условия участия в гранте и контактная информация.* Несколько значений на разных языках.
5. *Описание основных результатов.* Несколько значений на разных языках.Для завершившихся проектов.
6. *WEB-сайт проекта.*
7. *Ключевые слова.* Несколько значений на разных языках.

Связи с рубрикаторами и тезаурусами.

1. *Коды MSC.* Несколько значений.
2. *Код рубрикатора РФФИ.* Несколько значений. Для проектов РФФИ.
3. *Код основной таблицы УДК.* Несколько значений.

Связи с другими ресурсами.

1. *Персона.* Руководитель или участник проекта или гранта.
2. *Организация.* Участник проекта ли гранта.
3. *Публикация.* Публикация по проекту или гранту.
4. *Программное обеспечение.* Написано в рамках этого проекта или гранта.

**Конференция или семинар**

Конференция или семинар имеют следующие атрибуты:

1. *Тип.* Регулярное событие или разовое.
2. *Название.* Несколько значений на разных языках.
3. *Статус конференции или семинара.* Локальный, всероссийский, международный.
4. *Даты начала и окончания.*

5. *Дата окончания срока подачи заявки или тезисов* (deadline).
6. *Ключевые слова.* Несколько значений на разных языках.
7. *Описание.* Несколько значений на разных языках. Включает описание условий участия.
8. *Место проведения.* Несколько значений на разных языках.
9. *Web-сайт конференции или семинара.*

Связи с рубрикаторами и тезаурусами:

1. *Коды MSC.* Несколько значений.
2. *Код основной таблицы УДК.* Несколько значений.

Связи с другими ресурсами.

1. *Публикация.* Возможны следующие типы связей:
    1. *Труды конференции.*
    2. *Доклад на семинаре или конференции.*
2. *Организация или подразделение.* Организует конференцию или семинар.

**WEB-сайт**

Имеет следующие атрибуты:

1. *Название.* Несколько значений на разных языках.
2. *URL.*
3. *Тип.* Например, статический ресурс, электронная TELNET-библиотека, образовательный ресурс и т.д. Несколько значений. Справочник возможных типов или видов web-сайтов пока не разработан. Возможно, этот атрибут будет разделен на несколько атрибутов. Классифицировать можно, например, по виду реализации: (статический сайт, БД, поисковая система, z39.50-система, TELNET-библиотека ... ),
   по виду представленной информации (статьи, экспериментальные данные, тезаурус, образовательные материалы…). Некоторые существующие типы ресурсов были описаны в [12].
4. *Язык.* Несколько значений.
5. *Описание.* Несколько значений на разных языках.
6. *Ключевые слова.*

Связи с рубрикаторами и тезаурусами:

1. *Коды MSC.* Несколько значений.
2. *Код основной таблицы УДК.* Несколько значений.

Связи с другими ресурсами:

1. *Организация* или *персона.* Является владельцем и поддерживает этот ресурс.

**Программное обеспечение**

Имеет следующие атрибуты:

1. *Название.* Несколько значений на разных языках.
2. *Тип.* Научное, учебное, промышленное, коммерческое, некоммерческое.
3. *Краткое описание и условия лицензирования или приобретения.* Несколько значений на разных языках.
4. *Язык интерфейсов.* Несколько значений.
5. *Программный код.* Файл или URL.
6. *Документация.* Файл или URL.

Связи с рубрикаторами и тезаурусами:

1. *Код MSC.* Несколько значений.
2. *Код основной таблицы УДК.* Несколько значений.

Связи с другими ресурсами:

1. *Организация* или *персона.* Автор, владелец или разработчик.
2. *Публикация.* Публикация об этом программном обеспечении.
3. *Проект или грант.* Программное обеспечение стало результатом проекта или гранта.

## Защита данных и разграничение прав доступа

Как уже упоминалось, в системе Math-Net.RU предусмотрено разграничение доступа к ресурсам, как для чтения, так и для редактирования (см. ниже о пользователях системы и их правах). Для реализации разграничения доступа будут использованы технологии системы ИСИР, предусматривающие индивидуальное определение прав на чтение, модификацию и удаление каждого хранимого объекта в репозитории системы.

Помимо разграничения доступа к ресурсам в целом, будет ограничен доступ на чтение к некоторым атрибутам некоторых ресурсов. Например, при вводе информации о себе некоторые персоны могут пожелать, чтобы их домашние адреса и телефоны были доступны только сотрудникам системы и Отделения Математики для контактов с ними, но не показывались в интерфейсах системы всем желающим. Публичный доступ к полным текстам некоторых публикаций также может быть ограничен, в то время как остальные их атрибуты и связи будут общедоступны.

В соответствии с моделью разграничения прав доступа ИСИР, пользователи системы могут проходить аутентификацию двумя способами: вводя логин и пароль зарегистрированного в системе пользователя, либо осуществляя запрос с зарегистрированных IP-адресов. Пользователь, не прошедший аутентификацию, имеет только права на чтение общедоступных ресурсов и их атрибутов. Права доступа прошедшего аутентификацию пользователя могут быть специфицированы индивидуально, либо в соответствии с правами группы, к которой принадлежит этот пользователь.

## Пользовательские интерфейсы

Система должна иметь пользовательские WEB-интерфейсы, чтобы быть доступной как можно большему количеству пользователей. Пользователей системы можно условно разделить на 3 группы:

1. *Администраторы данных.* Осуществляют поддержку адекватности и целостности данных в системе. На этих пользователях лежит ответственность за адекватность набора ресурсов, значений их атрибутов и связей между ними. В эту группу включаются также операторы загрузки и интеграции данных.
   Администраторы данных могут добавлять, удалять и редактировать ресурсы и связи между ними непосредственно в базе данных системы в пределах зоны своей ответственности.
   Администраторы данных должны быть достаточно компетентны в предметной области математики, а также в технических аспектах функционирования системы Math-Net.RU.
2. *Пользователи, предоставляющие информацию.* Это достаточно компетентные в своей области люди, сотрудничающие с системой. Они несут, или не несут ответственность за предоставляемую информацию. Предоставляемая этими людьми информация проходит через операторов загрузки и интеграции данных, которые осуществляют добавление предоставленной информации и, при необходимости, контроль ее адекватности и редактирование.
3. *Обычные пользователи.* Люди, осуществляющие поиск информации в системе.

### Интерфейсы обычных пользователей

Эти интерфейсы должны быть на русском и английском языках, для обеспечения возможности доступа к ресурсам в качестве обычных пользователей большей части математиков мира.

Кроме того, среди российских математиков, особенно старшего поколения, велик процент людей, слабо владеющих компьютером. Часто даже обычные широко известные общеупотребительные термины, пришедшие в русский язык вместе с компьютерами и Internet, не понятны таким людям. Потому интерфейсы по возможности должны быть рассчитаны именно на такой круг людей.

Интерфейсы обычных пользователей должны обеспечивать следующие способы доступа к информации:

#### Навигация по древовидной структуре рубрикатора или тезауруса

При навигации пользователь переходит между узлами, соответствующими разным рубрикам рубрикатора. При просмотре каждого узла он должен получать список ресурсов указанного типа, соответствующих текущей рубрике, а также список рубрик, для которых текущая рубрика является родительской.

#### Навигация по связям между ресурсами

На странице просмотра каждого ресурса должны быть соответствующие гипертекстовые ссылки на страницы просмотра связанных с ним ресурсов, по которым пользователь может осуществить переход к просмотру соответствующих ресурсов.

**Поиск ресурсов по значениям их атрибутов и их просмотр (атрибутный и полнотекстовый поиск)**

Такие интерфейсы должны быть для всех типов ресурсов. Помимо собственных атрибутов ресурсов, поиск также может быть осуществлен по рубрикам рубрикаторов.

Поисковые запросы должны вводиться в WEB-формы.

Ограничения на значения атрибутов должны иметь вид “в значении атрибута встречаются слова, удовлетворяющие данному простому регулярному выражению”.

Интерфейсы должны позволять использование в запросах логических операций между ограничениями на значения атрибутов.

Ниже приведен предварительный список атрибутов, по которым должны быть возможны поиск для каждого ресурса:

**Персона:**

- *Полное имя*
- *Почтовый Адрес*
- *E-mail*
- *Ученая степень*
- *Ученое и академическое звание*
- *Специальность ВАК* (поиск и навигация)
- *Код MSC*(поиск и навигация)
- *Код основной таблицы УДК* (поиск и навигация)
- *Ключевые слова*
- *Должность*

**Публикация:**

- *Название*
- *Тип*
- *Биб. описание (в формате ГОСТ)*
- *Идентификаторы*
- *Специальные идентификаторы*
- *Дата публикации*
- *Язык*
- *Аннотация*
- *Реферат*
- *Код MSC* (поиск и навигация)
- *Компоненты кода УДК* (поиск и навигация)
- *Ключевые слова*

- *Наименование Организации-издательства*
- *Имя автора (наименование коллективного автора)*

**Организация или подразделение**

- *Название и сокращенное название*
- *Почтовый адрес*
- *E-mail*
- *Код MSC*(поиск и навигация)
- *Код основной таблицы УДК* (поиск и навигация)
- *Ключевые слова*

**Проект или грант**

- *Название*
- *Диапазон дат начала и окончания*
- *Код MSC*(поиск и навигация)
- *Код основной таблицы УДК* (поиск и навигация)
- *Код рубрикатора РФФИ* (поиск и навигация)
- *Ключевые слова*

**Конференция или семинар**

- *Тип*
- *Название*
- *Статус*
- *Диапазон дат проведения*
- *Диапазон дат последнего срока подачи заявок*
- *Описание*
- *Код MSC*(поиск и навигация)
- *Код основной таблицы УДК* (поиск и навигация)
- *Ключевые слова*
- *Наименование Организации или подразделения* (которое организует конференцию, или является учредителем)

**WEB-сайт**

- *Название*
- *Тип/вид*
- *Язык*
- *Код MSC*(поиск и навигация)
- *Код основной таблицы УДК* (поиск и навигация)
- *Ключевые слова*

**Программное обеспечение**

- *Название*
- *Тип*
- *Краткое описание и условия приобретения*
- *Язык интерфейсов*

- *Код MSC*(поиск и навигация)
- *Код основной таблицы УДК* (поиск и навигация)
- *Ключевые слова*

Помимо атрибутного поиска и навигации для каждого типа ресурсов будет возможен также полнотекстовый поиск по значениям ряда атрибутов (для некоторых публикаций и по полным текстам публикаций), а также исходных текстов, файлов кода и документации программного обеспечения.

**Интерфейсы администраторов данных**

Делятся на *интерфейсы пакетной загрузки и интеграции* и *интерфейсы редактирования данных.*

Интерфейсы загрузки данных работают с еще не загруженными в базу данных системы данными. Их задачи:

1. Нормализация вводимых данных.
2. Контроль адекватности вводимых данных.
3. Интеграция загружаемых данных в систему.

Подробнее о загрузке, нормализации и интеграции данных будет сказано ниже.

Интерфейсы редактирования данных работают с уже загруженными в систему данными. Они должны обеспечивать ввод, удаление и модификацию атрибутов ресурсов и связей между ними.

Все интерфейсы администраторов данных будут рассчитаны только на достаточно компетентных в техническом плане пользователей.

**Интерфейсы пользователей, предоставляющих информацию**

Предполагается, что такими пользователями будут российские математики или достаточно компетентные в математике и связанные с математикой по роду занятий люди, пожелавшие на каких-либо условиях сотрудничать с системой. Например, независимые авторы рефератов. По этой причине эти интерфейсы также как и интерфейсы обычных пользователей, должны быть рассчитаны на некомпетентных в техническом плане людей, и иметь простую структуру, не привязанную к схеме данных системы.

Интерфейсы должны обеспечивать возможность ввода информации, как через web-форму, так и другими способами. Например, такие пользователи могут присылать письма в установленном текстовом формате по электронной или обычной почте (текстовые формы). Такой формат должен быть опубликован и доступен всем заинтересованным лицам. Кроме того, должны быть обеспечены механизмы эффективной обработки операторами загрузки данных текстов в этом формате.

## Наполнение системы информацией

Как уже упоминалось, предполагаются следующие источники наполнения системы

информацией:

1. Загрузка и актуализация информации из других баз данных.
2. Пакетная загрузка информации из структурированного текста.
3. Предоставление информации пользователями системы.
4. Харвестинг информации из доступных в Internet источников.
5. Ввод данных оператором данных.

В первых четырех случаях вводимая информация должна быть обработана подсистемой загрузки и интеграции данных. Эта подсистема производит преобразование информации к схеме данных системы, а также контроль адекватности информации и ошибок преобразования информации (см. об этом также выше, в описании схемы данных).

При этом, как уже упоминалось,  оператор может осуществлять контроль адекватности, а также принимать участие при необходимости в нормализации и интеграции данных с другими ресурсами.

Под *нормализацией* далее будем подразумевать приведение значений атрибутов ресурсов и связей в соответствие с областями значений этих атрибутов. Например, если фамилия персоны написана с маленькой буквы, а также содержит случайно попавшие туда посторонние символы (например, скобки), первая буква фамилии должна быть заменена на заглавную, а посторонние символы должны быть удалены. Нормализацией также будет, например, приведение дат к установленному формату или устранение опечаток.

Под *интеграцией* далее будем подразумевать процессы *идентификации* ресурса, т.е. определения, существует ли уже такой ресурс в системе, идентификации связанных с ним ресурсов, создание при необходимости нужных ресурсов и объединение загружаемых ресурсов с уже существующими ресурсами, описывающими те же объекты реального мира.

Ниже источники наполнения системы данными рассмотрены подробнее.

**Ввод данных оператором данных**

Это наиболее простой способ ввода ресурсов. Однако на практике он должен применяться относительно нечасто. Например, при исправлении ошибок. При вводе данных оператором данных используются интерфейсы редактирования данных системы, которые работают непосредственно с базой данных системы.

**Пакетная загрузка информации из структурированного текста**

Пакетная загрузка должна осуществляться из структурированных текстовых файлов, т.е. файлов, написанных в соответствии с некоторым форматом или синтаксисом. Такой формат должен позволять описывать ресурсы разных типов. В качестве такого формата предполагается использовать формат XML. В процессе загрузки на первой после нормализации значений атомарных атрибутов данные преобразуются к формату RDF, структура которого соответствует OWL-схеме репозитория системы, после чего происходит собственно загрузка данных в

репозиторий и интеграция.

Предусматривается, также поддержка форматов данных некоторых других международных информационных систем (например, SOIF [11]). Данные в таких форматах могут быть сначала преобразованы к XML классическими методами синтаксического анализа, поскольку такие методы на выходе представляют информацию входного файла в виде дерева, которое становится структурой входного XML-файла.

При пакетной загрузке предполагается, что загружаемая информация адекватна по своему содержанию. В этом случае часто удается полностью или частично автоматизировать загрузку при помощи соответствующих алгоритмов. При этом конкретные случаи, в которых работа таких алгоритмов может дать сбой, должны быть распознаны этими алгоритмами и обработаны оператором вручную с помощью соответствующих интерфейсов.

Алгоритмы нормализации и, в особенности, идентификации и интеграции не могут быть достаточно универсальны для всех ресурсов и всех источников данных. Эффективный алгоритм, обрабатывающий высокий процент ресурсов без вмешательства оператора, и допускающий при этом минимум ошибок, может быть создан только с учетом специфических особенностей конкретного источника и характерных ошибок в получаемой из него информации.

Например: В системе – источнике данных имена персон хранятся в виде текстовых полей, в каждом из которых последовательно идут разделенные пробелом фамилия, имя и отчество персоны. Иногда после фамилии в скобках для женщин указывается девичья фамилия. При выгрузки данных для Math-Net.RU скрипт первое слово прописывает в качестве значения атрибута «фамилия», второе – «имя», все остальное – «отчество». Если встречается строка с девичьей фамилией, то именем становится, соответственно, Открывающая скобка, а отчеством – девичья фамилия + закрывающая скобка + имя + отчество.

Зная такую характерную ошибку источника данных, легко написать алгоритм, исправляющий все такие случаи без вмешательства оператора, и не допускающий при этом ошибок. Написать же универсальный алгоритм, учитывающий ошибки такого рода практически невозможно.

Для разрешения этого противоречия подсистем нормализации и интеграции должна быть построена по модульному принципу, позволяющему подключать разные алгоритмические модули для разных источников данных. Модули должны обмениваться информацией и взаимодействовать между собою на основе стандартных интерфейсов. Каждый модуль должен реализовывать один из вариантов алгоритмов решения данной подзадачи, и может быть использован для одного, или нескольких ресурсов и источников данных, специально под которые он и был создан.

Процессы нормализации, идентификации и интеграции идут в автоматическом режиме, для тех ресурсов, для которых вероятность ошибки, исходя из конкретной ситуации, минимальна. Если алгоритмы подсистемы не могут

обеспечить достаточную безошибочность некоторой операции для некоторых конкретных ресурсов (например, не могут однозначно идентифицировать какой-либо ресурс), необходимо вмешательство оператора загрузки данных.

Например: фамилия персоны начинается со строчной буквы, либо в теле фамилии обнаружены заглавные буквы. Перед фамилией стоят пробелы. В этом случае подсистема удаляет лишние пробелы и приводит буквы к нужному регистру. Если в фамилии обнаружены посторонние символы – требуется вмешательство оператора.

Еще пример: при загрузке информации о персоне в базе данных системы не находится ни одной персоны с такой фамилией и инициалами, подсистема загрузки без участия оператора создает такую персону. Если же обнаружены 2 персоны с теми же инициалами, подсистема, в зависимости от реализованных в ней алгоритмов, может либо начать анализ других атрибутов, либо вызвать оператора, который должен принять решение по идентификации загружаемой персоны.

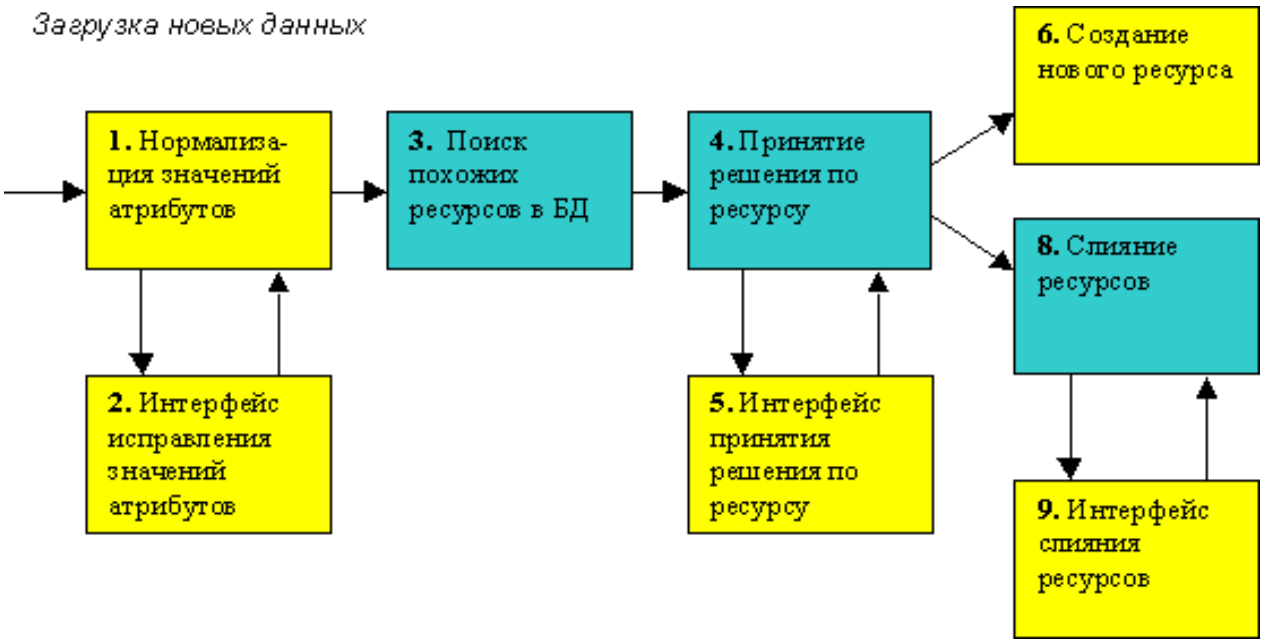
При пакетной загрузке существует 2 способа взаимодействия оператора с системой в сомнительных случаях:

1. Подсистема загрузки в сомнительном случае (когда она не может гарантированность безошибочность принятого решения) все-таки принимает решение, как с изменением данных, так и без оного, однако пишет сообщение с описанием сомнительной операции в лог загрузки данных.
   Впоследствии оператор просматривает лог, проверяет каждый сомнительный случай и при необходимости исправляет информацию в базе данных системы с помощью интерфейсов редактирования.
2. Подсистема загрузки в сомнительном случае приостанавливает загрузку и ожидает принятия решения оператором через специальные интерфейсы загрузки.

Практика работы прототипа системы Math-Net.RU (см. о прототипе ниже) показала, что первый вариант предпочтительнее, когда процент сомнительных случаев невелик (около 1 – 3 % загружаемых ресурсов). При большем проценте предпочтителен второй вариант.

Требуемые алгоритмы компоненты и технологии для реализации нормализации и интеграции, удовлетворяющие описанному модульному принципу, в настоящий момент разрабатываются в рамках проекта ИСИР.

Предварительный вариант разбиения на модули, а также взаимодействия между модулями подсистемы нормализации и интеграции ресурсов для пакетной загрузки представлен на рисунке.




**Предоставление информации пользователем системы**

Пользователь системы предоставляет данные через рассмотренные выше интерфейсы пользователей, предоставляющих информацию. Если априори известно, что предоставляемая некоторым пользователем информация не требует контроля адекватности ее содержания (получена из надежного источника), загрузка сразу проводится по технологии пакетной загрузки. В противном случае все загружаемые данные проверяются предварительно на адекватность содержания вручную (ответственность за содержание ложится на оператора).

**Загрузка информации из других баз данных. Актуализация информации из разных источников**

Для работы механизма загрузки данных из другой БД необходимо осуществить отображение  онтологии источника на онтологию нашей информационной системы.

При этом возникает задача актуализации информации, включая как добавление в систему ресурсов, добавленных в базу-источник, обновление ресурсов, обновленных в базе - источнике, и удаление ресурсов, удаленных в базе-источнике. Следует также учесть, что информация об одном и том же ресурсе может быть получена из разных источников. В этом случае, если его описание было удалено только в базе данных, являющейся одним из таких источников, вряд ли стоит удалять ресурс из системы. Конфликты также могут возникнуть при изменении описания ресурса в одном из источников.

В соответствии с технологией ИСИР каждый ресурс в системе может иметь любое количество URI разных типов. В метаданных из базы данных - источника для каждого ресурса необходимо выделить свой уникальный идентификатор, который и будет храниться в базе данных системы в форме соответствующего типа URI – *URI источника происхождения данных*. Тип этого URI должен идентифицировать базу данных - источник происхождения данного ресурса, а его значение – быть уникальным идентификатором ресурса в базе данных источнике.

Такие URI присваиваются также ресурсам, полученным из других источников. В случае, если информация об дном и том же ресурсе получается из разных источников, ресурс получает несколько URI источников происхождения. Причем, если источник не является другой базой данных, из которой система получает информацию в режиме актуализации (например, информация вводится пользователем или оператором), такой URI получает формальное значение, не идентифицирующее фактически ресурс. Имеет смысл только его тип, идентифицирующий источник его происхождения.

Сначала из базы источника осуществляется выгрузка данных в установленном формате в соответствии со схемой отображения схем данных, после чего данные загружаются  в систему по технологии пакетной загрузки. Идентификация ресурсов при этом проводится по вышеописанному URI. В режиме актуализации ресурсы, отсутствующие в базе-источнике, но имеющие в системе URI источника происхождения, тип которого соответствует этой базе-источнику данных, лишаются этого URI, поскольку они были удалены из базы – источника, а значит эта база уже не может считаться источником происхождения данных ресурсов. Если такой ресурс при этом не имеет других URI источника происхождения, он удаляется из базы данных системы.

**Харвестинг информации из доступных в Internet источников**

Под харвестингом в данном случае понимается сбор информации, не предназначенной изначально для загрузки в систему из источников в сети Internet. Такими источниками могут быть, например, математические web-сайты. Имеет смысл осуществлять харвестинг только из тех источников, адекватность информации которых не вызывает сомнений.

Компонента харвестинга должна в соответствии со структурой источника извлекать из него информацию (метаданные ресурсов), и выдавать структурированный текст описаний ресурсов в доступном для пакетной загрузки формате, после чего данные загружаются в систему по технологии пакетной загрузки.

Такая компонента харвестинга также должна быть уникальна для каждого источника, и строиться по модульному принципу.

## Форматы хранения текста. Проблема математических формул

Математические тексты в электронном виде могут храниться в разных форматах. Такие форматы должны удовлетворять по возможности следующим свойствам:

1. Быть достаточно общеупотребительными, чтобы у подавляющего большинства читателей существовало программное обеспечение для просмотра математического текста в этих форматах. Программное обеспечение для чтения этих форматов должно быть по возможности бесплатным.
2. К этому формату должны легко преобразовываться тексты в других форматах, которые используют математики для создания текстов.
3. Быть достаточно компактным. Это требование необходимо для доступа к

текстам из медленных сетей.
4. Позволять включать в текст любые математические формулы, а также информацию форматирования (заголовки, разные шрифты и т.д.).
5. Позволять легко извлекать из текста отдельные слова и фразы в формате ASCII-текста с целью индексации для обеспечения эффективного поиска. Желательна также возможность индексации формул.
6. Удовлетворять традициям создания и обмена текстами в математическом мире.

Ниже перечислены распространенные универсальные, а также специализированные математические форматы для математических текстов, а также оценка их применимости.

**ASCII-текст**

Хорошо удовлетворяет пунктам 1-3 и 5-6, однако совершенно не имеет средств представления математических формул, и вообще какого-либо форматирования текста.

**Форматированный текст (RTF) и другие форматы MSOffice**

Ограничено удовлетворяет пунктам 1, 4 и 6. Совершенно не удовлетворяет пунктам 2, 3 и 5.

**HTML**

Хорошо удовлетворяет пунктам 1-3 и 5-6. Однако возможности HTML по записи математических формул весьма ограничены. В разных браузерах формулы в HTML могут выглядеть по-разному.

**HTML с формулами в виде картинок**

Этот формат удовлетворяет пунктам 1-2 и 4 и отчасти 6. В настоящее время практически для всех используемых форматов существует программное обеспечение для приведения текстов с формулами к такому виду. Однако, такое представление получается довольно громоздким, и совершенно не позволит осуществлять индексацию и поиск по математическим формулам. Потому пункт 7 не удовлетворяется, а пункт 5 удовлетворяется частично.

**PDF**

Этот формат широко распространен и общепринят для обмена печатными текстами (удовлетворяет пунктам 1, 6). Он так же хорошо удовлетворяет пунктам 2 и 4, и отчасти пунктам 3 и 5. Однако формат довольно сложен, и позволяет один и тот же текст представить многими разными способами (например, текст может быть упакован разными способами, или вообще представлен как картинка). В связи с этим индексация текста в PDF довольно затруднительна. Математические формулы в PDF также обычно представлены графически, а потому их индексация невозможна.

**TeX**

Этот формат очень хорошо удовлетворяет всем вышеперечисленным требованиям,

за исключением пункта 2, поскольку этот формат описывает в большей степени структуру самой формулы (безотносительно к математической семантике), а не только ее отображение. Потому конверторов из каких-либо форматов в TeX практически не существует. Еще одним недостатком формата TeX является необходимость наличия громоздкого компилятора для просмотра формул не знакомыми с этим форматом людьми. Однако очень многие математики знают TeX и создают свои статьи в формате TeX. Кроме того, TeX имеет несколько версий, и каждая из них требует свой компилятор.

**DVI**

Этот формат, как и PDF, по сути является графическим, и возник, как производный от формата TeX. Он отчасти удовлетворяет требованиям 4 и 6 и не удовлетворяет остальным.

**PostScript**

Как и PDF, в этом формате в ряде случаев текст может быть представлен разными способами, в том числе и как графический объект, а математические формулы всегда будут представлены как графические объекты. Он удовлетворяет требованиям 1, 2, 4, 6, частично 5 и совсем не удовлетворяет требованию 3, поскольку очень некомпактен.

**MathML и OpenMath**

Эти форматы появились относительно недавно и основаны на формате XML. MathML, так же как и TeX, определяет структуру отображения математической формулы, в то время как OpenMath определяет семантику формулы. Запись формулы в MathML может содержать ссылки на объекты, семантика которых определяется средствами OpenMath. Структура описания математических объектов в OpenMath позволяет использовать этот формат в системах формальной логики и в системах поиска доказательств.

В настоящий момент ведется создание единой базы данных математических объектов в OpenMath, что можно считать первым шагом на пути создания единой математической базы знаний в формате, пригодном для машинной обработки.

В целом MathML и OpenMath, также как и TeX, удовлетворяет всем вышеперечисленным требованиям к форматам представления математических текстов, за исключением пункта 2. Однако в настоящий момент эти форматы и весьма перспективные технологии, основанные на них, пока не получили широкого распространения.

В качестве дальнейшего развития MathML и OpenMath в настоящее время создается также стандарт “OpenMathematicalDocuments” (OMDoc)

Максимально удовлетворяет этим требованиям формат TeX. Он полностью удовлетворяет требованиям 3, 4, 5, и в значительной степени требованиям 1 и 6. В меньшей степени удовлетворяют этим требованиям форматы DVI, PDF, PostScript, HTML, ASCII-текст и RTF. Именно эти форматы традиционно используются также в существующих математических информационных системах для хранения полных

текстов публикаций и их аннотаций. Новый перспективный формат MathML/OpenMath также хорошо удовлетворяет требованиям 3, 4 и 5, однако он не успел получить пока широкого распространения, и под него пока не создан весь спектр необходимого программного обеспечения.

Для заголовков и других строковых атрибутов математических публикаций к вышеперечисленным требованиям добавляется еще одно: тексты должны быть встраиваемы в HTML-файлы в виде ASCII-текста и хорошо читаемых формул. Этому требованию полностью удовлетворяет только формат TeX, и, частично, HTML и ASCII-текст (с неполной поддержкой формул). В перспективе этому требованию может удовлетворять также MathML/OpenMath.

В существующих математических информационных системах эта проблема решается использованием заголовков в формате TeX. Преимуществом такого подхода является возможность индексировать формулы в атрибутах так же, как и слова. Кроме того, пользователи, не знакомые с TeX, также могут читать атрибуты без формул и искать ресурсы по словам в этих атрибутах.

В системе Math-Net.RU также предполагается хранение формул в текстовых атрибутах ресурсов в форматах TeX или ASCII-текст. Кроме того, в качестве допустимых форматов для хранения полных текстов могут быть использованы форматы DVI, PDF, PostScript, HTML, ASCII-текст, RTF. По мере развития и распространения перспективных форматов MathML/OpenMath (OMDoc), их поддержка также может быть обеспечена в Math-Net.RU.

## Участие во всемирной математической информационной системе Math-Net в качестве российского узла

В настоящий момент проект всемирной системы Math-Net весьма расплывчиват и далек от стадии технического задания. Существует очень немного требований, которые в настоящее время предъявляются к информационной системе, чтобы быть узлом всемирной Math-Net. Ниже перечислены эти требования, а также описано, каким образом система Math-Net.RU будет удовлетворять этим требованиям.

### Англоязычный интерфейс

Как было указано выше, реализация системы Math-Net.RU предусматривает англоязычный интерфейс поиска информации.

### Обмен метаданными в стандартных форматах

В рамках проекта ИСИР разрабатываются технологии, обеспечивающие выгрузку и загрузку информационных ресурсов в виде наборов метаданных в стандартных форматах, таких, как DublinCore, Vcard, и RDF. Эти технологии также будут использованы и в системе Math-Net.RU.

### Вторичные страницы организаций и подразделений

Для интеграции в систему Math-NetIMU рекомендовал математическим

организациям создавать *вторичные страницы организаций*.

Вторичная страница организации или подразделения представляет собою HTML-документ установленного формата и дизайна, содержащий название и логотип организации или подразделения, а так же расположенные в установленных местах ссылки на страницы, содержащие основные сведения об организациях.

Существуют инструментальные средства, генерирующие вторичную страницу организации на основе вводимой в диалоговый интерфейс необходимой информации об организации и необходимых WEB-ссылок.

Система Math-Net.RU предусматривает выдачу по специальному запросу URL вторичной страницы любой организации. Если организация не имеет своей вторичной страницы, система выдаст URL специального вида. Ответ на запрос по такому URL система выдаст динамически сгенерированную “суррогатную” вторичную страницу на основе атрибутов организации, содержащихся в базе данных Math-Net.RU. Таким образом, каждая организация, представленная в Math-Net.RU, автоматически будет представлена и во всемирной Math-Net.

**Классификация ресурсов рубрикатором MSC**

Этот рубрикатор фактически уже стал стандартом в мировой математики для тематической классификации ресурсов любых типов. Поддержка классификации ресурсов всех типов этим рубрикатором будет реализован также и в Math-Net.RU.

**Поддержка электронных публикаций**

В настоящий момент электронная публикация рассматривается IMU и CEIC как наиболее перспективное средство обмена научной информацией. Поддержка публикаций такого рода также предусмотрена в Math-Net.RU.

**Поддержка функций узла распределенной базы данных**

Здесь под функциями узла распределенной базы данных подразумевается способность системы обрабатывать поисковые запросы в соответствии с определенным протоколом, утвержденным в качестве протокола общения между узлами системы. К настоящему моменту такой обязательный для всех участников протокол в Math-Net утвержден не был, а разные потенциальные участники Math-Net пользуются разными протоколами.

Таким образом, имеет смысл говорить не о протоколе, а о требованиях, которым он должен удовлетворять. Требования к возможностям обработки поисковых запросов для Math-Net.RU, сформулированные выше, вполне соответствуют возможностям существующих информационных систем, таких, как, например, сервисы AMS, ZentralblattMATH и немецкая система Math-Net.

## *Текущее состояние*

В настоящее время реализован прототип портала Math-Net.RU, доступный в Web по адресу http://www.math-net.ru/ . Прототип был реализован на основе уже

реализованных технологий ИСИР.

Прототип поддерживает хранение, поиск, выдачу и сопровождение (ввод и редактирование) следующих ресурсов

- *Организация*
- *Подразделение*
- *Персона*
- *Публикация*
- *Проект*

Прототип поддерживает представление информации на русском и английском языках в интерфейсах поиска и навигации.

Прототип поддерживает также интерфейсы оператора данных на двух языках (ввод любых ресурсов) и пользователя, предоставляющего информацию (ввод персон, интерфейс на русском языке). Частично реализован также интерфейс оператора, загрузки данных.

Кроме того, имеется возможность пакетной загрузки информации из простого текстового и структурированного (XML) формата.

Используя эти средства в рамках проекта Math-Net.RU, был создан Директорий российских математиков, ставший частью всемирного Директория математиков 2002 года.

В настоящий момент база данных Math-Net.RU включает достоверную информацию о более чем 4000 математиков, а также полную базу данных журналов ОМ РАН.

При создании директория были реализованы средства для рассылки математикам писем, с приглашением ввести информацию о себе для Директория российских математиков и паролей для ввода и система учета активности математиков, которым разосланы письма и даны права на ввод информации.

## Литература


1. http://www.openmath.org/ OpenMath web site.
2. http://www.w3.org/Math/ W3C Math Home Page.
3. http://www.ams.org/msc/ 2000 Mathematics Subject Classification (MSC).
4. http://www.udcc.org/ Universal Decimal Classification (UDC) Consortium.
5. http://www.ams.org/mathscinet/ Mathematical reviews on the web (MathSciNet).
6. http://www.emis.de/ZMATH/ Zentralblatt MATH abstracting and reviewing service in pure and applied mathematics.
7. http://www.math-net.org/ an Internetional Information and Communication System.
8. Ю.И.Кузякин, Г.Ф. Масич, А.В.Созыкин. Распределенная информационная система для проведения конференций. Сборник технического университета, 2002.
9. http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2003/part4/BNSBS

Бездушный А.А., Нестеренко А.К., Сысоев Т.М., Бездушный А.Н., Серебряков В.А., Архитектура и технологии RDFS-среды разработки цифровых библиотек и Web-порталов, Электронные библиотеки, Том 6, выпуск 4, 2003 г.
10. А.Клецель, Форматы графических файлов, TriArtGraphicsStudio, Тель-Авив, 1999
11. http://www.tardis.ed.ac.uk/harvest/docs/user/ Harvest Summary Object Interchange Format (SOIF).
12. А. С. Аджиев. WEB-ресурсы для российских математиков, "Информационные ресурсы России", #6, 2003
13. Аджиев А.С. Нгуен М.Х. Подходы к описанию и использованию тезаурусов в информационных системах. Труды конференции RCDL2003, Петербург, Россия, 2003, стр. 191-200.
14. http://www.w3.org/2001/sw/WebOnt/ OWL Web Ontology Language, W3C Recommendation 10 February 2004.

---

### *Об авторах*

**Аджиев Алим Сапарович** – младший научный сотрудник Центра научных телекоммуникаций и информационных технологий РАН. Сфера деятельности: программирование, Java, RDFS, Semantic Web, Web-порталы, базы данных, системное программирование, цифровые библиотеки, информационное обеспечение научной деятельности.
Тел. +7 095 938 37 09
E-mail: ajiev@ccas.ru

**Бездушный Анатолий Николаевич** - кандидат физико-математических наук, с.н.с. Вычислительного Центра имени А. А. Дородницына РАН. Сфера деятельности: системное программирование, параллельные вычисления, сети, базы данных,распределенные системы, информационно-поисковые технологии, цифровые библиотеки, Web-порталы.
Тел. (095)1355471(*4216)
E-mail: bezdushn@ccas.ru

**Серебряков Владимир Алексеевич** - доктор физико-математических наук, с.н.с., зав.отделом Вычислительного Центра имени А. А. Дородницына РАН. Сфера деятельности: системное программирование, параллельные вычисления, сети, базы данных, распределенные системы, информационно-поисковые технологии, цифровые библиотеки, Web-порталы.
Тел. (095)1355471(*4220)
E-mail: serebr@ccas.ru

---