Общероссийский WEB-портал математических ресурсов

© А. С. Аджиев ВЦ РАН ajiev@ccas.ru
© А. Н. Бездушный ВЦ РАН bezdushn@ccas.ru
© С. П. Коновалов МИ РАН serk@mi.ras.ru
© В. А. Серебряков ВЦ РАН serebr@ccas.ru

Аннотация

На основе проведенного ранее анализа российских математических электронных ресурсов, а также опыта зарубежных математических информационных систем описан проект общероссийского математического портала. Более детально описан проект создаваемой математической информационной системы Math-Net.RU как первого этапа реализации Общероссийского портала. 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, способы наполнения системы информацией и интеграции 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 также перспективы участия системы Math-Net.RU в создаваемой Всемирной математической информационной системе Math-Net, а также требования к системе-участнику.

Цель проекта

Для удовлетворения информационных потребностей российских математиков необходимо создание общедоступного через Internet общероссийского математического информационного портала, удовлетворяющего всему спектру информационных и коммуникационных потребностей российских математиков, а также лиц, обучающихся математике и интересующихся математикой, и отражающего в совокупности всю математическую деятельность в России.

Компоненты создаваемой системы

Общие свойства компонент

Все компоненты должны удовлетворять следующим требованиям:

1. Быть максимально интегрированными между собою. То есть поддерживать необходимые перекрестные связи между ресурсами разных компонент в соответствии с общей открытой схемой данных, а также обеспечивать в интерфейсах возможность навигации между ресурсами из разных компонент системы.
2. Иметь развитые интерфейсы для поиска и навигации в пространстве ресурсов на русском и английском языках, а также средства администрирования и поддержки баз данных компонент.
3. Иметь средства разграничения доступа к разной информации для разного круга лиц.
4. Поддерживать общепринятые средства тематической классификации ресурсов (рубрикаторы, тезаурусы, другие системы классификации). Обеспечивать поиск ресурсов по этим средствам классификации.
5. Иметь средства обмена информацией с другими информационными системами в открытых стандартных форматах и по открытым стандартным протоколам.

Компоненты системы:

Директорий российских математиков

Информация обо всех российских математиках, достигших определенных результатов в науке.

Российские математические журналы

База данных полных текстов российских реферируемых математических журналов.

Математические публикации

Информация о разных публикациях (книгах, сборниках, трудах, журналах, отдельных статьях, диссертациях, электронных публикациях и других текстовых изданиях).

Российский сервер препринтов

Средства для поиска и публикации препринтов (в "западном" смысле) и электронных публикаций по математике.

Система реферирования математических ресурсов России

Рефераты на российские математические публикации, сделанные независимыми экспертами - вне-

Труды 5ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL2003, Санкт-Петербург, Россия, 2003.

штатными сотрудниками команды поддержки этой компоненты.

Российские математические организации

Информация об организациях, включая организационную структуру, направления деятельности, научные результаты, контактную информацию, должности сотрудников, а также информационные ресурсы организации.

Математические проекты и гранты

Описания проектов и грантов, а также информацию о том, как принять участие в каком-либо конкретном гранте.

Математические конференции и семинары

События и научные встречи математиков (включая конференции и семинары), информация о том, как принять участие в конференции или семинаре. А также организация и проведение средствами этой компоненты математических телеконференций в Internet как в режиме реального времени, так и в асинхронном режиме.

Каталоги электронных библиотек

Средства поиска публикаций в электронных каталогах всех математических библиотек России, участвующих в проекте. Взаимодействие между этой компонентой и электронными каталогами библиотек на основе стандартных открытых протоколов.

Библиотека математического программного обеспечения

Информация о существующем математическом программном обеспечении, исследовательском, коммерческом или образовательном, включая документацию, лицензионные соглашения и исходники (если доступны).

Каталог Российских математических WEB-ресурсов

Полезные и серьезные математические web-ресурсы в России или на русском языке.

База данных вакансий для математиков

Позволяет работодателям публиковать объявления о вакансиях в математических организациях России на исследовательские и преподавательские позиции, а также осуществлять поиск среди резюме математиков. Позволяет математикам осуществлять поиск вакансий и публиковать свои резюме.

Система Math-Net.RU. Постановка задачи

Приведенный выше круг задач весьма широк и является формулировкой конечной цели создания полной информационной системы. 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.

Создаваемая информационная система будет основана на технологиях и принципах построения информационной системы ИСИР [8]. Эта система обеспечивает эффективную работу с хранящимися и поддерживаемыми распределенно онтологиями сложной структуры, включая мощные средства атрибутного поиска, разграничения прав доступа, загрузки и интеграции данных с другими системами на основе открытых стандартов.

В связи с этим, ниже в статье не рассматриваются 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

Пользователями системы будут российские и иностранные математики, а также аспиранты и студенты, выбравшие научную работу в области математики в качестве своей будущей профессии.

Пользователями также будут административные работники Отделения математики РАН.

Информация, хранимая и доступная в системе

В системе будет храниться информация, необходимая для обеспечения научной работы или обучения в области математики. Основными типами 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.
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 или учебное 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 атрибу-

тов, регулярным выражениям и сложным поисковым запросам.

2. Навигацию в пространстве 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 и интеграцию информации разного уровня структурированности из электронных источников таких, как базы данных, структурированный текст, Web-сайты.

6. Распределенное хранение и поддержку данных.

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

Реализация системы

Архитектура

Для обеспечения возможности распределенной поддержки и хранения данных система должна иметь распределенную архитектуру, 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.

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.

Тематическая классификация ресурсов

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.

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

Код УДК является сложной синтаксической конструкцией, которая включает рубрики основной таблицы УДК, определители, а 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".

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 с горизонтальными связями одноязычного тезауруса, который можно использовать для классификации ресурсов всех типов.

Таким образом, в качестве основного тематического рубрикатора в системе будет использован рубрикатор MSC, а также, возможно, основная таблица УДК. На первом этапе достаточно будет реализовать их как простые иерархические рубрикаторы. В перспективе будет реализована поддержка тезаурусов, а также полная поддержка системы классификации УДК.

Кроме этого, в системе для тематической классификации специальностей персон будет использован традиционный для России рубрикатор ВАК, а для проектов - рубрикатор РФФИ, используемый для всех проектов, поддерживаемых Российским фондом фундаментальных исследований.

Принципы построения схемы данных системы, уровень детализации описания ресурсов

При разработке модели данных необходимо, 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.

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.

На первом этапе необходимо реализовать минимальную детализацию, определяемую уровнем детализации источников информации, а также требованием обеспечения обработки основных пользовательских поисковых запросов, но обеспечивающую бесшовное 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.

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, Dublin Core и VCard, с целью обеспечения по возможности легкого и обратимого преобразования данных к этим форматам.

Некоторые общие свойства и атрибуты ресурсов

В системе Math-Net.RU предусмотрено разграничение доступа к ресурсам, как для чтения, так и для редактирования. Права доступа к ресурсу могут быть указаны через специальные атрибуты ресурса, или вычислены на основе связей этого ресурса с другими ресурсами.

Права доступа могут быть ограничены также 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 всем желающим.

Схема данных

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

Текущее RDFS-описание схемы данных Math-Net.RU находится по URL http://mathnet.ru/project/rdfs/schema.rdfs.

Пользовательские интерфейсы

Система должна иметь пользовательские WEB-интерфейсы, чтобы быть доступной широкому кругу математиков. Пользователей системы можно условно разделить на 3 группы:

1. *Администраторы данных*. Осуществляют поддержку адекватности и целостности данных в системе. 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*. 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.

Они несут, или не несут ответственность за предоставляемую информацию. Предоставляемая этими людьми информация проходит через операторов загрузки и интеграции данных, которые осуществляют загрузку 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 поиска в системе 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:

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

Должен быть возможен поиск всех типов ресурсов. Помимо значений атрибутов, поиск ресурсов 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.

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.

Навигация по древовидной структуре рубрикатора или тезауруса

При навигации пользователь переходит между узлами, соответствующими разным рубрикам рубрикатора. При просмотре каждой рубрики он должен получать ее описание, 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 гипертекстовые ссылки на страницы просмотра связанных с ним ресурсов, по которым пользователь может осуществить переход к просмотру соответствующих ресурсов.

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 ресурсов и связей между ними.

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

Интерфейсы пользователей, предоставляющих информацию

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 нормализации, загрузки и интеграции данных.

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 связей между ними, 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, которые работают непосредственно с базой данных системы.

Пакетная загрузка информации из структурированного текста

Загрузка метаданных осуществляется из структурированных текстовых файлов, т.е. файлов, написанных в соответствии с некоторым форматом или синтаксисом. В качестве таких форматов предполагается использовать принятые в ИСИР форматы на базе XML и RDF, а также форматы данных некоторых других международных информационных систем (например, SOIF [10]), поддержка которых будет осуществлена с помощью компиляторов форматов.

При пакетной загрузке необходимо преобразование схем данных, то есть решение задач интегра-

ции и нормализации. Часто удается реализовать алгоритмы, корректно решающие эти задачи для большинства ресурсов. При этом конкретные случаи, в которых такие алгоритмы могут ошибиться, должны быть распознаны этими алгоритмами и обработаны оператором вручную с помощью соответствующих интерфейсов.

Существует 2 способа взаимодействия оператора с системой в сомнительных случаях:

1. Подсистема загрузки в сомнительном случае все-таки принимает решение, как с изменением данных, так и без оного, однако пишет сообщение с описанием сомнительной операции в журнал загрузки данных. Впоследствии оператор просматривает журнал, 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 показала, что первый вариант предпочтительнее, когда процент сомнительных случаев невелик (около 1 - 3 % загружаемых ресурсов). При большем проценте предпочтителен второй вариант.

Предоставление информации пользователем системы

Пользователь системы предоставляет данные через рассмотренные выше интерфейсы. 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 проверяется предварительно адекватность загружаемых данных вручную (ответственность за содержание ложится на оператора загрузки).

Загрузка информации из других баз данных. интеграция информации из разных источников

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

При этом возникает также задача интергации информации, а именно, добавление в систему ресурсов, добавленных в базу-источник, обновление ресурсов, обновленных в базе-источнике, и удаление ресурсов, удаленных в базе-источнике. При этом информация об одном и том же ресурсе может быть получена из разных источников, что требует средств решения конфликтов интеграции по разным источникам.

Харвестинг информации из доступных в Internet источников

Под харвестингом в данном случае понимается сбор информации, не предназначенной изначально ее владельцем 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 загрузки формате, после чего данные загружаются в систему по технологии пакетной загрузки.

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 пользователей 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 и специализированные математические форматы, а также оценка их применимости.

ASCII-текст

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

Форматированный текст (RTF) и другие форматы MS Office

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

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 матема-

тическим формулам. Потому пункт 3 не удовлетворяется, а пункт 5 удовлетворяется частично.

PDF

Этот формат широко распространен и общепринят для обмена печатными текстами (удовлетворяет пунктам 1, 6). Он также хорошо удовлетворяет пунктам 2 и 4, и отчасти 3. 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 иногда довольно затруднительна. Математические формулы в PDF также обычно представлены графически, а потому их индексация невозможна [9].

TeX

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

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 в PDF, в этом формате 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, и совсем не удовлетворяет требованию 3, поскольку очень некомпактен [9].

MathML [2] и OpenMath [1]

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, так же как и TeX, удовлетворяет всем вышеперечисленным требованиям к форматам представления математических текстов, за исключением пункта 2. Однако пока эти форматы и весьма перспективные технологии, основанные на них, не получили широкого распространения.

В качестве дальнейшего развития MathML и OpenMath в настоящее время создается стандарт "Open Mathematical Documents" (OMDoc)

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 (по мере развития браузеров).

В существующих математических информационных системах проблема представления формул в заголовках решается использованием формата TeX. Это позволяет 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 них, в качестве допустимых для 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 расплывчат и далек от стадии технического задания. Существует немного требований к информационной системе, чтобы быть узлом всемирной Math-Net. Ниже перечислены эти требования, а также описано, каким образом система Math-Net.RU будет удовлетворять им.

Англоязычный интерфейс

Система Math-Net.RU предусматривает англоязычный интерфейс для поиска информации.

Обмен метаданными в стандартных форматах

В рамках проекта ИСИР разрабатываются средства выгрузки и загрузки информационных ресурсов в стандартных форматах, таких, как RDF, Dublin Core, Vcard. Эти средства будут использованы и в системе Math-Net.RU.

Вторичные страницы организаций и подразделений

Для интеграции в систему Math-Net IMU рекомендовал математическим организациям создавать свои вторичные страницы.

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

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

Система Math-Net.RU предусматривает хранение URL вторичной страницы организации. Если организация не имеет своей вторичной страницы, система выдаст URL специального вида. В ответ на запрос по такому URL будет выдана динамически сгенерированная "суррогатная" вторичная страница на основе атрибутов, содержащихся в базе данных Math-Net.RU. Таким образом, каждая организация, представленная в системе Math-Net.RU, будет представлена и во всемирной Math-Net.

Классификация ресурсов рубрикатором MSC

Этот рубрикатор фактически уже стал стандартом в мировой математике 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 электронных публикаций также предусмотрена в Math-Net.RU.

Поддержка функций узла распределенной базы данных Math-Net

Здесь под функциями узла распределенной базы данных подразумевается способность системы обрабатывать поисковые запросы в соответствии с определенным протоколом, утвержденным в качестве стандарта общения между узлами системы. К настоящему моменту такой протокол в Math-Net не утвержден, 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 [5], Zentralblatt MATH [6] и немецкая система Math-Net [7].

Текущее состояние

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

Прототип поддерживает хранение, поиск, выдачу и сопровождение (ввод и редактирование) ресурсов типов *организация, подразделение, персона, публикация* и *проект*.

Прототип поддерживает представление информации на русском и английском языках в интерфейсах поиска и навигации, а также интерфейсы оператора данных (для всех типов ресурсов). Частично реализованы также интерфейсы пользователя, предоставляющего информацию, и оператора загрузки данных (ввод данных персон).

Имеются средства пакетной загрузки информации из структурированного текстового и XML форматов.

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 информацию о 4000 математиках, а также базу данных журналов ОМ РАН.

При создании директория были реализованы средства работы с пользователями, предоставляющими информацию (рассылки писем, с приглашением ввести информацию о себе для Директория, паролей для ввода, система учета активности пользователей).

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/ American Mathematical Society.

[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] Бездушный А. А., Бездушный А. Н., Нестеренко А. К., Серебряков В. А., Сысоев Т. М. Архитектура RDFS-системы. Практика использования открытых стандартов и технологий в системе ИСИР.

[9] А. Клецель, Форматы графических файлов, TriArt Graphics Studio, Тель-Авив, 1999

[10] http://www.tardis.ed.ac.uk/harvest/docs/user/ Harvest Summary Object Interchange Format (SOIF).

Russian Mathematical WEB portal

A.S. Ajiev, A.N. Bezdushny, S.P. Konovalov,
V.A. Serebriacov

Basing on analysis of Russian mathematical electronic resources and foreign mathematical information systems the Russian Mathematical Portal project was described. Russian Mathematical Information System Math-Net.RU project as a first stage of the Portal project was described in detail. Math-Net.RU is based on the ISIR universal information system core. The project is described in terms of requirements and conditions system must meet. Alternative variants of various system components realization are considered and different ways to resolve the problems arisen is analyzed. Stored data types, user circles and functionality required are outlined. The description was done for the system framework, data scheme, user interfaces and the ways of information import and replication from another information systems and databases. Consideration was given for the mathematical formula storing and representation problem in mathematical texts. Comparative analysis was given for existing mathematical formula storing and representation formats. Consideration was given also for the World Math-Net project participation for the Math-Net.RU system and the World Math-Net requirements to the system-participant.