Министерство образования Российской Федерации
ГОУ ВПО «Удмуртский университет»
Математический факультет
Кафедра математического обеспечения ЭВМ

**ГРЕБНЕВ Андрей Николаевич**

# Информационные системы научных коммуникаций

выпускная квалификационная работа

Ижевск 2003

# Аннотация

В данной работе обоснована необходимость и изложены теоретические принципы построения информационной системы научной коммуникации (ИСНК), главным образом в аспекте научных публикаций; рассмотрено множество стандартов, технологий и инструментов, практически используемых для построения ИСНК; описана архитектура проектируемой системы; приведены результаты практической апробации некоторых модулей ИСНК.

# Оглавление

# Введение

*Актуальность* исследования обусловлена потребностью научного сообщества в решении проблемы дороговизны и медлительности научной коммуникации, формальную основу которой составляет научная публикация, на базе современных и наиболее прогрессивных информационных технологий.

*Научная новизна* исследования состоит в: анализе различных инструментов и технологий, стандартов и методологий, возможных для использования в процессе научного информационного обмена; выявлении их роли и взаимосвязи; формировании собственных новых, альтернативных взглядов на некоторые частные вопросы в рамках исследуемой проблемы; адаптации уже существующих решений.

*Практическая значимость* работы обусловлена: проектируемой информационной системой научных коммуникаций; реализованным ядром системы и несколькими ее модулями; внедренным и практически опробованным основным модулем системы, а именно модулем преобразования форматов научных публикаций и отображения математических формул. Практическая значимость работы заключается в реальной возможности использования проектируемой информационной системы в центрах научной коммуникации (университеты, институты, академии), что позволит создать единую среду научного информационного обмена.

*Объект* исследования – информатизация процессов научной публикации, составляющую основа научных коммуникаций.

*Предмет* исследования – инструменты и технологии, стандарты и методологии, в применении к единой информационной системе научных коммуникаций.

*Гипотеза* – научная коммуникация будет наиболее эффективной если:

1) будут использоваться современные и наиболее прогрессивные информационные технологии;
2) будет минимизирована стоимость информационной системы научных коммуникаций и стоимость ее обслуживания;
3) будет максимально упрощен процесс использования информационной системы конечными потребителями;
4) будет решена проблема защиты прав автора на техническом уровне;
5) будет максимально упрощен процесс закладывания, преобразования и извлечения знаний в информационной системе.

*Цель* работы – провести анализ, спроектировать и реализовать информационную систему научных коммуникаций, обеспечивающую более быстрый и качественный информационный обмен в научном сообществе.

*Задачи* работы:

1. Проанализировать инструменты и технологии, стандарты и методологии, возможные для использования в процессе построения информационной системы научных коммуникаций.
2. Минимизировать себестоимость системы за счет использования свободно распространяемых бесплатных решений.
3. Минимизировать стоимость обслуживания системы за счет автоматизации большинства процессов.
4. Максимально упростить процесс использования системы за счет доступного пользовательского интерфейса, отсутствия необходимости привлечения дополнительных инструментов при работе с системой.
5. Выявить пути решения проблем защиты прав автора и простоты извлечения знаний.

На защиту выносятся следующие *основные положения*:

1. Система проанализированных инструментов и технологий, стандартов и методологий.
2. Проектируемая система научных публикаций, как формальная основа научных коммуникаций.
3. Реализованное ядро системы и некоторые модули.
4. Внедренный и практически опробованный основной модуль системы по преобразованию форматов публикаций и отображения математических формул.

*Методологическую основу* исследования составляют материалы европейских (ECDL), американских (JCDL) и российских (RCDL) конференций по тематике электронных библиотек, а так же научные журналы и статьи по данной и смежной тематике.

*Методы исследования*:

- ✓ Анализ литературы по исследуемой проблеме.
- ✓ Анализ существующих решений в области научных публикаций и научных коммуникаций.
- ✓ Анализ, построение взаимосвязей и адаптация инструментов и технологий, стандартов и методологий необходимых для реализации системы.
- ✓ Проектирование системы
- ✓ Реализация ядра системы и некоторых модулей
- ✓ Апробация основного модуля системы в реальных условиях.

*Материалами для исследования* являются спецификации, стандарты, технологии, инструменты, методологии ведущих международных организаций по стандартизации W3C, IEEE, ISO и т.д. и крупнейших корпораций, лидеров на рынке по исследуемой области, Microsoft, Apache, Oracle, IBM, Adobe и т.д.

*Апробация работы*. Основные результаты работы докладывались и обсуждались на: XL Международной научной студенческой конференции

«Студент и научно технический прогресс» Новосибирск 2002, по итогам которой представленный доклад отмечен дипломом второй степени; на семинарах и заседаниях кафедры Математического обеспечения ЭВМ ГОУ ВПО «Удмуртский университет». По исследуемой теме опубликовано четыре работы.

*Структура работы*. Работа состоит из введения, трех глав, заключения и списка использованной литературы. Общий объем работы при сквозной нумерации составляет 58 страниц, рисунков 13, таблиц 5, библиография содержит 28 названий.

*Достоверность* полученных результатов исследования и вытекающих из них выводов обеспечивается методологической обоснованностью исходных параметров исследования, связанных с системным подходом; последовательностью в изучении предмета исследования; применением методов адекватных задачам и логике исследования; позитивным результатам экспериментальной работы.

# Глава 1. Научные коммуникации сегодня

Стремительно развивающаяся научно-техническая революция стала основой глобального процесса информатизации всех сфер жизни общества. От уровня информационно-технологического развития и его темпов зависят состояние экономики, качество жизни людей, национальная безопасность, роль в мировом сообществе.

Уровень развития современных технологий в стране зависит, в первую очередь, от интеллектуального потенциала общества и, следовательно, от уровня развития науки и образования в стране. Важнейшую роль играют качество и темпы информационного обмена в научном сообществе.

Научная публикация относится к важнейшим первичным средствам формальной научной коммуникации. «Публикация выступает как первичный источник сведений о научном знании, отношениях между учеными, строении и динамике научных объединений и т.п. Для науковеда, философа, логика, методолога, специалиста по информатике, социолога науки той конечной реальностью, из которой исследователь черпает свои представления о науке, выступают публикации. Отличающиеся друг от друга изображения науки в различных исследовательских традициях ... становятся объектами изучения лишь постольку, поскольку сведения о них имеются в научной публикации» [1].

Однако, в научную коммуникацию, помимо публикаций, входят и другие составляющие, наиболее важным элементом является обратная связь читателя с автором, представляемая различными конференциями, дискуссиями, форумам и т.д.

## 1.1 Бумажные публикации

Научные бумажные публикации переживают не лучшие времена. Не секрет, что в большинстве случаев автор научной публикации вообще ничего не получает, иногда получает копейки, нередко платит сам. Скорость выхода

публикации тоже оставляет желать лучшего, очень часто информация статьи успевает «устареть» еще до выхода бумажной версии в свет. Тираж большинства не крупных журналов позволяет обеспечить бумажной версией, как правило, только самих авторов статей и библиотечные фонды местных научных учреждений. Тем самым статья остается не доступной широкому кругу лиц читателей.

Информатизация библиотечного дела позволяет несколько улучшить положение. На сей день, уже многие библиотеки имеют электронный каталог собранных у них публикаций, тем самым, позволяя читателю установить хотя бы наличие нужной статьи. Однако подобные каталоги имеют поиск лишь по названию сборника, названию статьи и автору. Этого не достаточно чтобы обеспечить быстрый и качественный поиск нужной информации. Перевод статей в полнотекстовую форму, как правило, не осуществляется, а значит, информация является недоступной для читателя, не имеющего возможности лично придти в подобную библиотеку.

Обычные бумажные публикации (издания), будь то журналы, сборники статей, не успевают за темпами научно-технического прогресса [2]. Они уже не могут обеспечить требуемого качества и скорости информационного обмена. В обществе наметились тенденции частичного перехода (а в будущем возможно и полного) с бумажных публикаций на электронные.

## 1.2 Электронные публикации

Научные электронные публикации получают все большее и большее распространение. Их возможности в значительной мере превосходят бумажные. Это и быстрота выхода публикации, ее дешевизна, широта охватываемой аудитории. К сожалению, на данный момент, потенциальные возможности электронных публикаций задействованы лишь на очень малую долю. Бумажная публикация приобрела лишь электронную форму отображения.

К настоящему моменту можно выделить несколько основных мест сосредоточения научных публикаций в российском Интернет.

### 1.2.1 Персональные (домашние) страницы-сайты

Статьи на них размещаются самими автором работ. Форматы размещения статей различны. Как правило, публикации выкладываются в исходном формате (в котором она была набрана). В подавляющем большинстве это Microsoft Word или LaTeX. Публикация материалов в более удобных для чтения и печати форматах, таких как HTML (Hyper Text Markup Language) и PDF (Portable Document Format), требует специальных знаний в области информационных технологий и наличие под рукой необходимого инструментария для осуществления трансформации. На сегодня финансовое положение научных работников не позволяет воспользоваться удобными и легкими в использовании системами управления содержимым сайта (CMS – Content Management System). Им приходится использовать лишь стандартные средства переноса данных, предоставляемые провайдером Интернет услуг (ISP Internet Service Provider), таких как FTP (File Transfer Protocol), использование которых тоже требует наличия определенных знаний. Итак, получается что, например работники гуманитарных наук, обладающие лишь минимальными знаниями в области компьютеропользования, вообще не имеют возможности опубликовать свои работы на домашних страничках, не говоря уже о преобразовании форматов в более удобный вид и обеспечении правильного функционирования сайта, необходимого для качественной работы поисковых систем (Google, Yandex, и т.д.).

### 1.2.2 Сайты электронных научных журналов и конференций

Их можно развить на две основные категории *коммерческие* (на которых за чтение или публикацию материалов нужно платить) и *финансируемые* (например, с помощью грантов). Коммерческие сайты, ввиду своей платности, имеют ограниченный круг читателей, что в условиях

российской действительности не приводит к качественному и количественному улучшению научного информационного обмена. Финансируемые сайты имеют ограниченный бюджет, которого не хватает на разработку (покупку) и внедрение качественных систем электронных публикаций. Все публикации обрабатываются вручную или в частично автоматизированном режиме, что требует значительных затрат на оплату труда. Доступными форматами являются только HTML для просмотра и в редких случаях версия для печати в PDF. Поиск среди материалов представленных на сайте зачастую тоже отсутствует. В пример можно привести сайты ежегодной Всероссийской научной конференции RCDL «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», напрямую занимающейся тематикой научных публикаций. Организаторами конференции являются такие организации российского масштаба как Российский Фонд Фундаментальных Исследований (РФФИ), Объединенный Институт Ядерных Исследований, Московская секция ACM SIGMOD, Институт ЮНЕСКО по информационным технологиям в образовании. Однако даже на ней отсутствует поиск по материалам конференций, все тезисы выложены только в одном формате, только 22 апреля 2003 года появилась система подачи тезисов через Web-интерфейс. Еще одним недостатком подобных сайтов является их периодичность, например конференции проводятся, как правило, один раз в год, в итоге новые прогрессивные идеи авторов более года не доступны научному сообществу.

### 1.2.3 Электронные научные библиотеки

Электронные научные онлайновые полнотекстовые библиотеки на данный момент предоставляют наиболее качественный и быстрый способ публикации научных материалов. Среди подобных бесплатных библиотек можно привести два наиболее крупных в мире проекта – это arXiv.org

(физика, математика, информатика) и RePEc.org [3] (экономика). Существует российский проект socionet.ru [4] (экономика, социология, политика, закон, психология) основанный на технологии RePEc. Эти проекты используют два различных подхода к организации коллекций научных публикаций.

ArXiv.org – централизованная система научных публикаций. Статьи предоставляются авторами в формате LaTeX и его производных (PDF, PostScript, HTML+GIF). Читателю, по его запросу, может быть автоматически сгенерирована и выдана публикация в необходимом формате (PDF, PS, HTML+GIF). Метаинформация о статье довольно скудная, всего несколько полей. Возможность прямой публикации из форматов Microsoft Word или RTF (Rich Text Format) отсутствует.

Socionet.ru система, в основе которой лежат более прогрессивные технологии. Она имеет большой собственный набор атрибутов метаданных ReDIF. Имеется рейтинг публикаций, основанный на количестве прочтений. В сравнении с arXiv.org система Socionet.ru имеет распределенную децентрализованную структуру. Вообще говоря, это каталог ссылок на публикации, однако в ней есть возможность бесплатно создавать свой сайт в домене socionet.ru и публиковать материалы на нем, ссылки на публикации с личного сайта можно заносить в единый каталог socionet.ru. Все материалы размещает сам автор. Ограничение на формат публикации нет. Однако такой подход, порождает все проблемы, уже рассмотренных, персональных страничек-сайтов.

### 1.2.4 Единая среда распределенных ресурсов

Существует проект GRID [5] – единая глобальная среда распределенных ресурсов, надо сказать планетарного масштаба. Основу планируемой системы составляют такие понятия как: кластеры понятийных сетей, информационное обеспечение принятия решений и т.п.

Однако, на сегодняшний момент научный работник (исследователь) среднестатистического российского регионального университета (института), не смотря на бурное развитие информационных технологий, все еще публикует электронные версии своих статей (тезисов) на домашней страничке, используя при этом подручные, зачастую не адекватные инструменты. Авторы статей, не обладающие необходимыми знаниями, публикуют свои статьи только в медлительном бумажном виде.

Назрела необходимость создания бесплатной, дешевой и простой в использовании системы, решающей проблемы не просто публикации материалов в электронной форме, но проблемы прав автора, скорости извлечения знаний, качественного рейтинга статей, создание обратной связи автора с читателем и т.д. Назрела необходимость создания единой *информационной системы научных коммуникаций* (ИСНК) [6,7,8,9], распределенной по научным центрам (университет, институт, академия).

# Глава 2. Система и ее модули

Информационная система научных коммуникаций (ИСНК) – это подкласс электронной библиотеки, это распределенная интероперабельная среда, объединяющая коллекции, сервисы и людей для поддержки полного жизненного цикла создания, распространения, использования и сохранения полнотекстовых научных публикаций, представленных в слабоструктурированной гетерогенной форме[1].

## 2.1 Предназначение системы

ИСНК предназначена для организации научного информационного обмена как внутри научных центров (университет, институт, академия) так и между ними, посредствам сети Интернет.

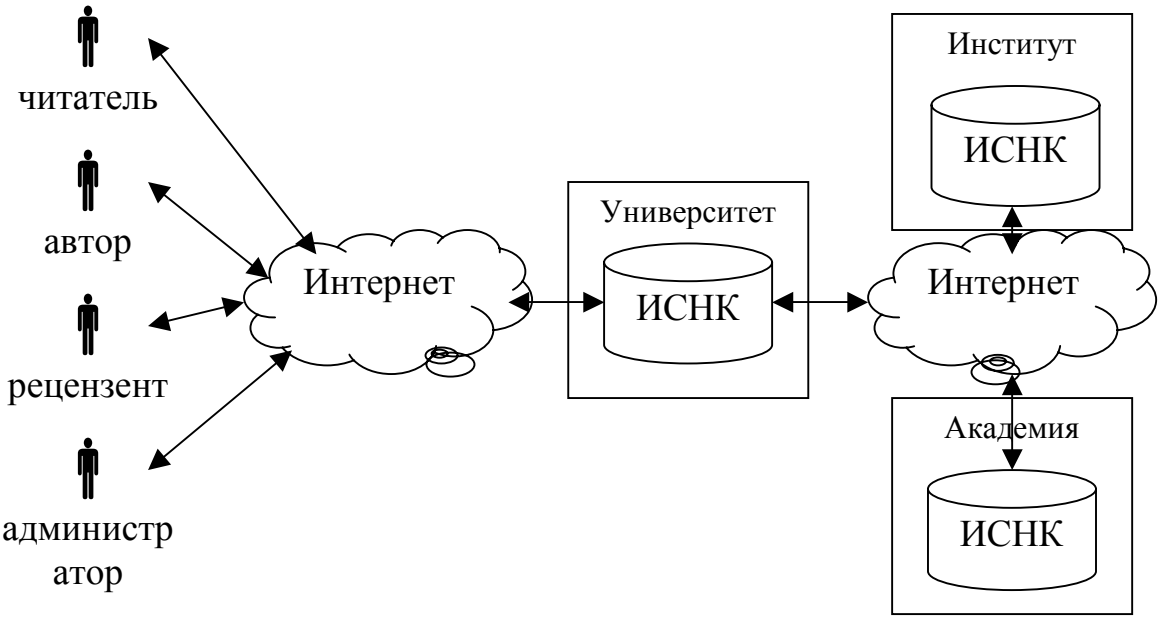


**Рисунок 1 Схема потоков информации при взаимодействии с ИСНК**

Распределение именно по центрам научной коммуникации обусловлено многими причинами. Рассмотрим два наиболее крайних случая построения системы.

[1] Далее по тексту будут расшифрованы все основные части данного определения.

Организация качественной полностью *централизованной* одиночной системы невозможна по следующим причинам:

1. Необходимы огромные финансовые вливания для построения и поддержания работоспособности подобной системы, российские бюджетные организации на данный момент не готовы к столь масштабному проекту.
2. Российские коммуникационные сети, особенно в научных центрах, не способны предоставить качественный доступ к одиночному серверу, обеспечивающему работу системы.
3. Менталитет российских ученых, консерватизм и недоверие к электронным публикациям сдерживает выход статей (тезисов) за границы родного университета (института).

Реализация качественной системы по принципу полной *децентрализации* (один центр со ссылками на публикации, размещенных на домашних страничках авторов) невозможна, т.к.:

1. Отсутствие руководящей организации, рецензирующих комитетов и т.п. приведет к потере научности системы, а в будущем и к недоверию авторов к системе.
2. Разнородность подходов (формат, стиль) к публикациям приведет к увеличению сложности взаимодействия пользователей с системой.

Реализация системы распределенной по центрам научной коммуникации будет наиболее эффективной.

1. Как правило, подобные учреждения уже имеют налаженную инфраструктуру публикации бумажных сборников статей (тезисов), научных журналов; проведения научных конференций; обеспечения рецензирования работ, которую можно

адаптировать для работы с ИСНК. Автоматизация этих процессов при внедрении ИСНК позволит освободить ресурсы учреждения необходимые для обеспечения работоспособности самой ИСНК.

2. Большинство наиболее ресурсоемких операций (публикация, рецензирование, администрирование), с точки зрения потребления каналов связи, будут происходит в рамках самого учреждения, что лишь незначительно увеличит загрузку каналов связи.

Итак, распределение ИСНК по центрам научной коммуникации позволит без значительных финансовых вливаний (что особенно важно для сегодняшних российских научных учреждений):

1. Автоматизировать существующие процессы научного обмена (в том числе и бумажные).
2. Перейти на более прогрессивную электронную форму обмена научной информацией.

## 2.2 Цель и задачи системы

Цель построения системы – ускорить темпы и улучшить качество информационного обмена в научном сообществе.

Задачи системы:

1. Автоматизировать существующие процессы информационного обмена в центрах научной коммуникации.
2. Внедрить новые прогрессивные технологии обмена научной информацией, в основе которой будет лежать электронная публикация.

3. Обеспечить обмен научной информацией между научными центрами.
4. Максимально расширить аудиторию авторов и читателей научных публикаций.
5. Привлечь к научному информационному обмену молодое перспективное поколение.

## 2.3 Требования к системе

Выделим основные требования к проектируемой системе, с точки зрения учреждения обеспечивающего работоспособность ИСНК.

*Дешевизна системы* – полное отсутствие необходимости покупать дополнительное программное обеспечение для установки и использования системы. Система должна быть полностью построена на бесплатных решениях.

*Кроссплатформенность системы* обеспечивает легкость ее установки на любую операционную систему, что позволяет без особых усилий интегрировать ее в уже существующую инфраструктуру организации.

*Дешевизна использования* – это минимизация расходов на обслуживание (администрирование, редакторская деятельность над публикациями и т.п.) системы, достигается за счет максимальной автоматизации всех процессов.

*Легкость обслуживания* системы, позволяет обеспечить быстрое внедрение системы и сэкономить средства на обучении обслуживающего персонала.

Обозначим требования к системе с точки зрения ее пользователей.

*Дешевизна использования* – это отсутствие необходимости покупать дополнительное программное обеспечение (ПО) для работы с системой,

достигается за счет использования либо стандартного широкого используемого ПО или за счет использования бесплатного легкодоступного ПО.

*Простота использования* достигается за счет интеграции с общеизвестными распространенными приложениями, отсутствие необходимости сложной и длительной установки и настройки приложений.

## 2.4 Функции системы

Основная функция ИСНК – это автоматизация всех этапов жизненного цикла научной публикации.

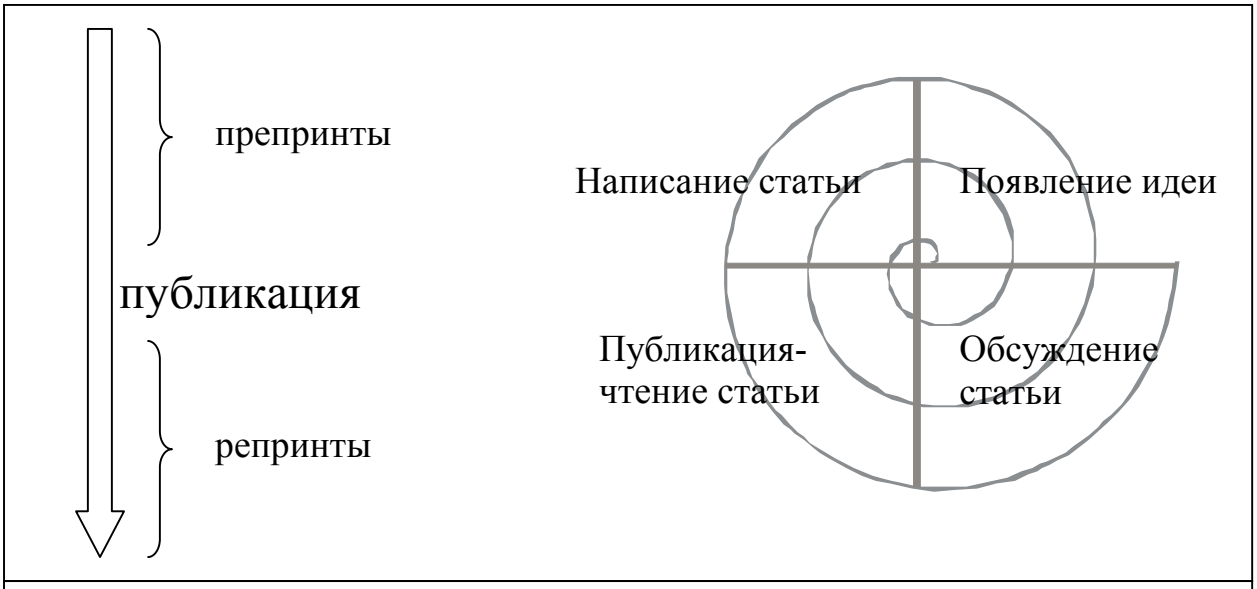


**Рисунок 2 Модели жизненного цикла научной публикации**

Рассмотрим основные этапы жизненного цикла научной публикации (например, статьи). На основе изучения опыта предыдущих исследований (как правило, изложенных в научных статьях), у автора будущей публикации *появляется идея*. Наступает длительный этап *написания статьи* автором, в процессе него публикация может переживать значительные изменения, корректировки, дополнения, в этот момент она уже может быть доступна некоторому кругу читателей в виде так называемых *препринтов*. По

завершению «оттачивания» содержимого, статья появляется в свет в виде законченной *публикации*, наступает этап *публикации статьи и прочтения* ее широким кругом читателей. Чтение статьи равномерно сочетается с процессом ее *обсуждения* между читателями и между читателями и автором. Процесс обсуждения неминуемо наводит автора на новые корректировки исходной идеи, начинают появляться *репринты* статьи. В конченом итоге обсуждение статьи приводит к *появлению конструктивно новой идеи* (хотя и не обязательно у автора исходной статьи).

Вообще говоря, понятие препринт означает допечатную электронную версию публикации, а репринт – послепечатную. Они отличаются от оригинальной бумажной статьи только форматом представления, содержимое публикации остается неизменным. Однако в эти понятия предлагается вложить дополнительный смысл – содержимое репринтов и препринтов может отличаться как от самой публикации, так и между собой. Публикация непосредственно соотносится с ее бумажной версией, если же публикация чисто электронная, то понятия препринтов и репринтов можно объединить и назвать просто версиями публикации. Это позволяет решить вопрос с так называемой «авторской волей». «Автору никто не мешает вносить в свой электронный текст любые изменения, поскольку любое изменение может быть датировано и авторизовано. Какая-то из версий текста в каждый данный момент может быть избрана истинно авторской, что не противоречит хранению всей предыдущей истории теста и демонстрации ее пытливому читателю. В известной степени электронная публикация может стать действенным ограничителем излишнего текстопорождения. Желая ответить на вопросы и возражения читателей, автор может не писать отдельных ответов на критики, а дополнять и править свой исходный текст, вставлять в него комментарии, сохраняя при этом (для себя и для читателей) всю историю этих преобразований. Наверное, большинство пишущих людей согласится с тем, что вместо одного задуманного текста на одну тему со

временем произрастает целый куст прямо или косвенно связанных с этой темой опубликованных текстов. К тому же всегда остается целый ряд не отвеченных вопросов, неопубликованных возражений; первоначальный замысел был строен и красив, результат аморфен и размазан по публикациям случайным образом» [10].


**Рисунок 3 Функциональная схема информационной системы научных коммуникаций**

Теперь попробуем определить основные *функции* ИСНК, используя уже знакомую устоявшуюся терминологию. Итак, ИСНК – это:

1. электронная библиотека (ЭБ)[2];
2. электронное издательство (ЭИ);
3. электронная дискуссия (ЭД).

ЭБ является основой ИСНК. Термин «электронная библиотека» (ЭБ) возник в 90-е годы на основе англоязычного термина «цифровые

[2] В данном случае и в дальнейшем, определяя термины, будет подразумеваться научный контекст. Например, под электронной библиотекой подразумевается электронная научная библиотека.

библиотеки» (digital library [11]). ЭБ – область исследований и разработок, направленных на развитие теории и практики обработки, распространения, хранения, поиска и анализа цифровых данных различной природы. На данный момент уже проведено большое количество исследований в области ЭБ [12,13]. Понятие ЭБ возникло вследствие эволюции обычных библиотек, сначала простой автоматизации библиотечных процессов компьютерными технологиями, а затем вследствие постепенного перехода на работу с электронными документами. Электронный документ – документ, носителем которого является электронная среда. Электронная библиотека выполняет следующие *функции*: сбора, хранения, классификации, поиска, выдачи информации и метаинформации. Заметим что сбор и выдачу уже готовой информации.

По аналогии с электронной библиотекой существует понятие электронное издательство (ЭИ) (E-Publishing House) [14]. Как и в случае с библиотекой, это понятие появилось в процессе эволюции издательств. Если обычное издательство выполняло только одну главную функцию – подготовку печатного издания, то в идеологии ИСНК ЭИ выполняет *функцию* подготовки документа для помещения ее в репозиторий библиотеки, а так же извлечения ее из репозитория и представление в требуемой форме.

Мы подошли к более или менее формальному определению ключевого понятия ИСНК – электронной научной публикации (ЭНП) (или просто ЭП). ЭНП – это электронный документ, смыслом которого является научная публикация, подготовленный электронным издательством и размещенный в электронной библиотеке. Теперь мы можем определить основной *объект* работы ИСНК – это электронная научная публикация.

Электронная дискуссия (E-Discussion) (ЭД) – в рамках идеологии ИСНК, осуществляет *функции* обратной связи читателя с автором, чего так

не хватает в бумажных публикациях. Главным элементом ЭД является форум для общения. Известно, что обсуждение идей часто приводит к рождению новых. Наведение конструктивной критики, получение откликов и рецензий обеспечивает значимое улучшение результатов работ.

Для лучшего понимания функциональной идеологии ИСНК, понятия ЭБ и ЭИ поставлены на один уровень. Однако тема электронных библиотек получила гораздо большую проработку с *теоретической* точки зрения, как в России, так и за рубежом. Поэтому в соответствии с нашим определением, ИСНК – это подкласс ЭБ, а ЭИ и ЭД являются лишь ее сервисами. Однако *практическая* проработка технологий все же позволяет рассматривать их как равнозначные.

Вообще говоря, только в комплексе все эти три отдельно развивающиеся технологии (ЭБ, ЭИ, ЭД) позволят создать ИСНК, обеспечивающую основу жизненного цикла электронной научной публикации.

## 2.1 Входные и выходные форматы

Все электронные публикации это некоторые файлы, представленные в необходимом формате. Интероперабельность данных – наиболее важная часть ИСНК.

С точки зрения использования системы, форматы подразделяются на *входные* (авторские) и *выходные* (читательские). Входные форматы являются либо *общеиспользуемыми* (Word, TeX, HTML), либо *открытыми* (DocBook, TEI), основанными на XML. Читатель будет иметь возможность получить публикации в зависимости от его требований HTML, DjVu, LuraDocument *для ознакомления*; PDF, PS *для печати*; WML *для беспроводных устройств*; eBook (OpenBook) *для чтения в электронных книгах*; подмножество XML *для обмена*; либо исходные Word, TeX *для дальнейшей работы*. При этом получаемый документ, при необходимости, может быть адаптирован под

физические возможности читателя или программные и аппаратные возможности его системы.

Существует два основных подхода к хранению публикаций. Первый (А) – *структурированный* и *гомогенный*, все публикации хранятся в едином структурированном формате XML (например, DocBook). Второй (Б) – *слабоструктурированный* и *гетерогенный*, публикации хранятся в различных исходных авторских вариантах, подчас имеющих не до конца проработанную структуру.

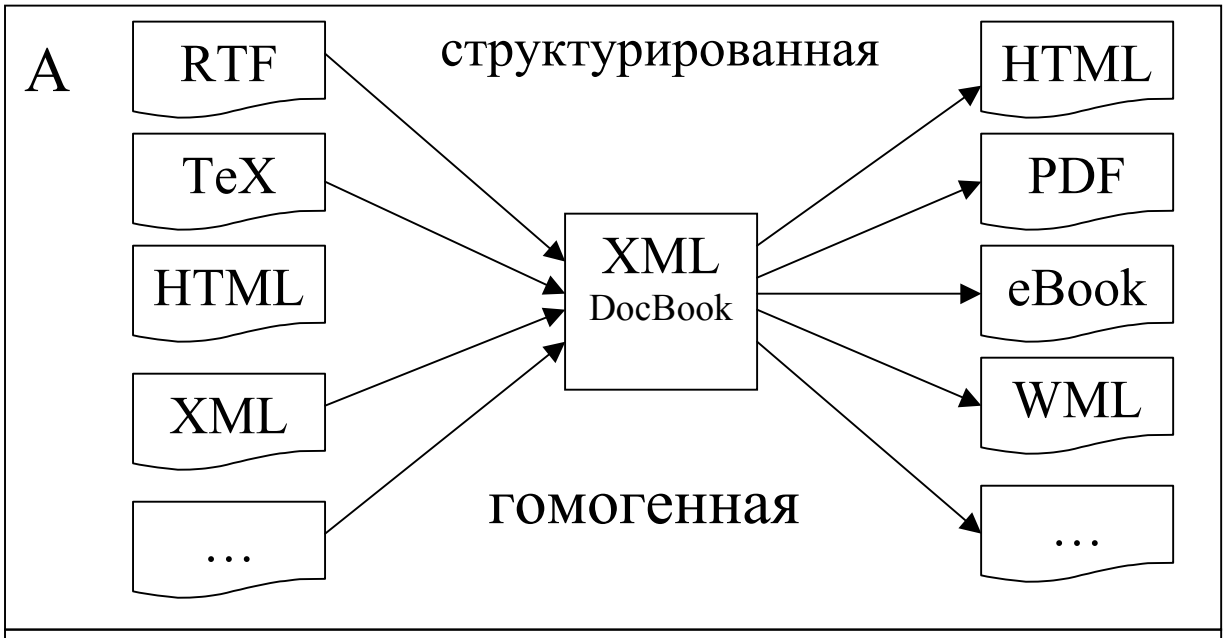


**Рисунок 4 Структурированная гомогенная среда**

Подход (А), преимущества:

1. Структурированный формат XML позволяет неограниченно, без особых усилий, расширить кол-во выходных форматов, с помощью использования технологии XSLT.

Подход (А), недостатки:

1. На сегодняшний момент, не все конвертеры (преобразователи) из входных форматов в XML, позволяют достичь желаемого

качества трансформации. Однако, прямые преобразователи, в некоторых случаях, из входных форматов в выходные позволяют этого достичь.

2. Для обеспечения правильной качественной структуры XML, необходимо чтобы входной документ был набран, используя соответствующую технику. Достичь этого на практике позволяет лишь использование стилей. Однако, не все авторы публикаций, в силу своих привычек или не знания соответствующей технологии готовы набирать материалы надлежащим образом. Второй путь решения проблемы это использование специальных визуальных текстовых редакторов XML документов.
3. Не все из входных форматов, в силу технического ограничения самого формата, позволяют обеспечить структуризацию документа в полном объеме.

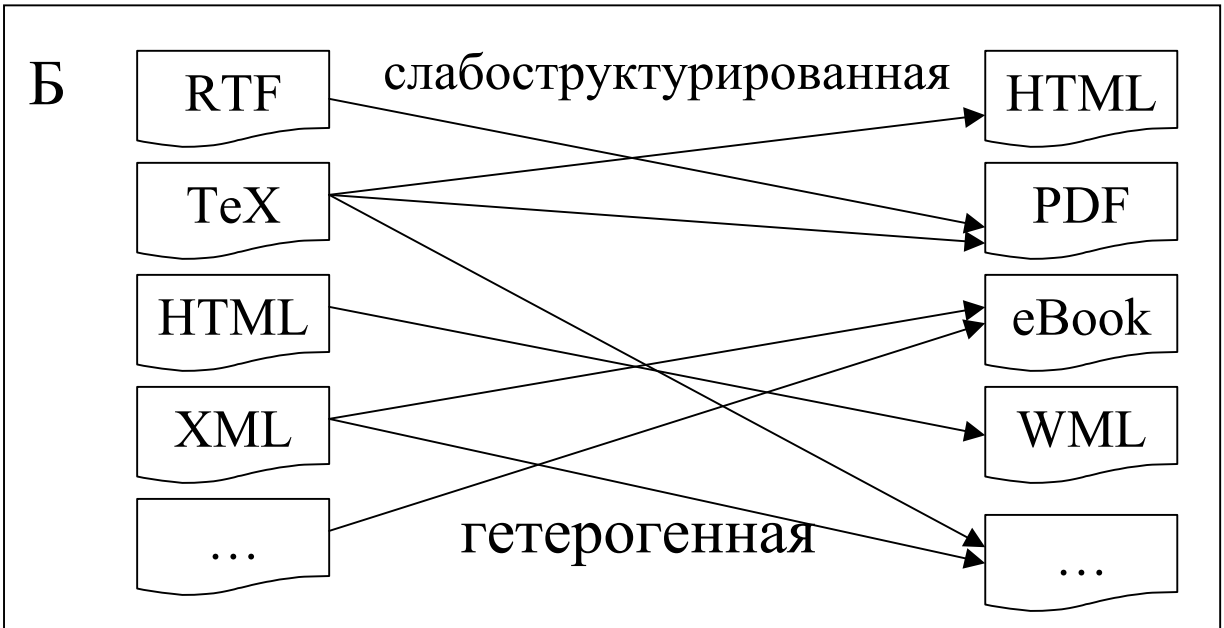


**Рисунок 5 Слабоструктурированная гетерогенная среда**

Подход (Б), преимущества:

1. Отсутствие необходимости автора работы подстраиваться под систему (использовать специальные стили).
2. Уже существует множество готовых бесплатных конвертеров, обеспечивающих приемлемое качество преобразования.

Подход (Б), недостатки:

1. Для случая использования форматов основанных на XML, качество преобразований может значительно ухудшится.

### 2.1.1 Подмножества XML для хранения

Для набора текстовой информации (в том числе и научной) в XML существует множество XML-приложений. Рассмотрим два наиболее распространенных.

*DocBook* (http://docbook.org/) – это наиболее популярный набор тегов для описания книг, статей и других документов, особенно технической документации. DocBook определен используя DTD синтаксис SGML и XML. Проект DocBook был начат в 1991 году фирмами HAL Computer Systems и O'Reilly. В связи с бурным ростом популярности языка, был организован Технический Комитет (Technical Committee TC) Организации по Продвижению Стандартов Структурированной Информации (Organization for the Advancement of Structured Information Standards OASIS http://www.oasis-open.org/).

Давайте, для общего обзора, вкратце перечислим основные элементы языка:

- главы, разделы, секции;
- оглавление, список таблиц, фигур, примеров;
- предметный указатель, глоссарий, библиография;
- мета информация: автор, заголовок, издатель и т.д.;

- списки: нумерованные, ненумерованные и т.д.;
- выделения: адрес, синопсис, эпиграф и т.д.;
- теги для описания пользовательского интерфейса: кнопки, меню, списки и т.д.;
- теги для описания конструкций языков программирования: классы, константы, функции, интерфейсы и т.д.;
- различные типы ссылок.

Существует большое кол-во приложений работающих с этим стандартом.

*TEI (Text Encoding Initiative)* (http://www.tei-c.org/) – это международный междисциплинарный стандарт, который помогает библиотекам, музеям, издательствам и отдельным ученым представлять все типы литературных и лингвистических текстов для онлайновой разработки и обучения, используя схему кодирования, которая максимально выразительна и минимально подвержена устареванию.

Проект TEI был основан в 1987 году для разработки руководящих принципов для кодирования машиночитаемых текстов в области гуманитарных и социальных наук. Руководящие принципы названные «P3» были разработаны в 1994 году и стали стандартом де-факто для кодирования литературы, лингвистических текстов и т.д. В июне 2001 был выпущена версия «P4». На данный момент о TEI можно говорить как о стандартном наборе тегов для SGML и XML.

Ввиду большой сложности TEI существует его подмножество называемое TEILite.

### 2.1.2 Визуальные текстовые редакторы XML

Для набора структурированных текстов в XML существуют специализированные редакторы. Рассмотрим два наиболее распространенных.

*Epic Editor* (ArborText, http://www.arbortext.com/) – редактор текстов внутреннее представленных в XML. Выделим основные свойства, отличающие его от обычного редактора текстов. Главной особенностью является формат внутреннего представления, это XML, в связи с этим появляется структуризация текста по существующему DTD. XML также позволяет легко трансформировать текст в нужный выходной формат, как с помощью XSLT, так и с помощью XSL-FO. В его поставку входит более 100 уже готовых представлений текста, в зависимости от нужного выходного формата, HTML для просмотра в браузере, WML для средств беспроводной связи, PDF и PS для печати, eBook и т.д. Существует импорт из Microsoft Word, Adobe FrameMaker, Interleaf, импорт таблиц из Microsoft Excel. Имеется API для многих языков (Java, JavaScript, C++, Perl, TCL), а также встроенный язык ACL (Arbortext Command Language), с помощью них можно легко сконфигурировать и интегрировать систему в уже существующие решения. Epic Editor поддерживает совместную работу с различными репозиториями Epic E-Content Engine (ArborText), Documentum 4i, Interwoven TeamXML, Oracle iFS, что позволяет организовать централизованное хранение документов, контроль версий, совместную работу с пользователями и т.д.

Подводя итог можно сказать, что профессионализм выполнения и проработки вполне оправдывает его стоимость.

*XMetal* (ранее SoftQuad, а сейчас Corel). Помимо стандартных функций редактирования XML-документов (дерево, исходный код с подсветкой) по DTD. В данном редакторе имеется множество интересных возможностей:

- Написание макросов на языках JavaScript или VBScript, средства для их отладки и редактор форм. Например, можно определить форму появляющуюся при вставке конкретного элемента.
- Возможность подключения DLL, COM, ActiveX.
- Проверка орфографии и тезаурус.

Возможность создания под каждое конкретное XML-приложение элементов редактирования. В качестве примера приведены элементы редактирования XML-документа типа «статья». Обеспечена возможность WYSIWYG редактирования.

**Таблица 1 Сравнительные характеристики структурных редакторов XML документов**

| Свойства | EpicEditor | XMetal |
|---|---|---|
| Форматы | XML, SGML | XML, SGML |
| Виды редактирования | WYSIWYG, source code | WYSIWYG, source code |
| Отслеживание изменений (рецензирование) | ДА | ДА |
| Поддержка таблиц | CALS, HTML 4.0 | CALS, HTML 4.0 |
| Возможность подстройки текста под различные аудитории (profiling) | ДА | НЕТ |
| Редактирование стилей для отображения тегов | ДА | ДА |
| Интеграция с репозиториями | Documentum 4i, Interwoven TeamXML, Oracle iFS, Epic E-Content Engine, etc | НЕТ |
| Поддержка DOM (Document Object Model) | ДА | ДА |
| Поддержка COM (Component Object Model) | ДА | ДА |
| API | C/C++, VB, Java, JavaScript, TCL, Perl | C++, VB, Java |

| | | |
|---|---|---|
| Макро языки | ACL (Arbortext Command Language) | JavaScript, VBScript |
| Встроенный XSL преобразователь (XSLT, XSLFO) | ДА | НЕТ |
| Проверка орфографии | 18 языков | Английский (возможно подключение) |
| Тезаурус | 13 языков | Английский (возможно подключение) |
| Встроенный редактор формул | ДА | НЕТ |
| Импорт в не производные от XML форматы | Microsoft Word, Adobe FrameMaker, Interleaf | Microsoft Word |
| Экспорт из не производных от XML форматов | Microsoft Word, Adobe FrameMaker | НЕТ |
| Платформы | Windows 95, 98, 2000, NT 4.0, Sun Solaris 7, 8 | Windows NT4.0, 2000 |

Из распространенных бесплатных решений хочется упомянуть визуальные текстовые редакторы, с возможностью сохранения документов в XML:

1. AbiWord (http://www.abisource.com)
2. Kword (http://www.koffice.org/kword/)
3. Текстовый редактор из OpenOffice (http://www.openoffice.org/)

### 2.1.3 Графические форматы отображения

Основными и всем известными представителями этих форматов являются PostScript и PDF от Adobe. О преимуществах и недостатках этих форматов друг перед другом можно говорить очень долго, что из них лучше, что хуже. Но как бы там ни было на данный момент PDF стал фактически стандартом в области электронной информации и Web-публикации, в то

время как PostScript не теряет свое лидерство в допечатных технологиях. В связи с этим мы не будем рассматривать PostScript, да и ко всему прочему plugin'ы для просмотра PostScript в браузере не так распространены как для PDF.

PDF стандарт де-факто и ничего с этим не поделаешь. Но в последнее время на его место начинают претендовать графические форматы основанные на волновых алгоритмах. Это DjVu от LizardTech (http://www.lizardtech.com) и LuraDocument от Algo Vision LuraTech (http://www.luratech.com). Оба эти формата специально предназначены для хранения и отображения текстовой информации. Очень высокая степень сжатия и возможность хранения документов в отсканированном, но не распознанном виде, дают большие преимущества этих форматов перед PDF.

**Таблица 2 Сравнение форматов PDF, DjVu, LuraDocument**

| Атрибут | PDF | DjVu | LuraDocument |
|---|---|---|---|
| Спрятанный текстовый слой | Да | Да | Нет |
| Встроенный поисковый индекс | Да | Нет | Нет |
| Подсветка терминов | Да | Нет | Нет |
| Постраничная навигация | Да | Нет | Нет |
| Поиск | Да | Да | Нет |
| Копирование и вставка текста | Да | Нет | Нет |
| Постраничная загрузка из Web | Да | Да | Нет |
| Спрятанный текстовый слой | Да | Да | Нет |
| Встроенное предпросмотровое изображение (Thumbnails) | Да | Да | Да |
| Закладки | Да | Нет | Нет |
| Гиперссылки | Да | Да | Нет |
| Навигационное дерево | Да | Нет | Нет |
| Дерево закладок | Да | Нет | Нет |
| Переход на страницу (GoTo) | Да | Нет | Да |
| Аннотации | Да | Нет | Да |
| Примечания | Да | Нет | Нет |
| Средства безопасности | Да | Нет | Нет |
| Поддержка средств безопасности третьих фирм | Да | Нет | Нет |
| Поддержка плугинов третьих фирм | Да | Нет | Нет |

| | | | |
|---|---|---|---|
| Размер тестового файла | 70628 | 8453 | 6796 |
| Размер плугина к браузеру | 10433Кб | 1961Кб | 3341Кб |

## 2.2 Набор и отображение мат. формул

В настоящее время в научных публикациях актуальна проблема набора математических формул и их отображение в браузерах. Наряду с классическим форматом TeX, все большее и большее распространение приобретает XML-формат MathML (Mathematical Markup Language). В качестве примера можно привести тот факт, что уже все основные математические системы, такие как Mathematica, Maple, MathCAD имеют возможность экспорта в формате MathML.

### 2.2.1 MathML

*MathML*[15] – это XML приложение для описания структуры и содержания математической нотации. Цель MathML – создать возможность для размещения и обработки математических документов в World Wide Web подобно тому, как HTML открыл такие возможности для текста.

MathML предназначен для облегчения использования и повторного использования математического и научного наполнения Сети, а также для различных приложений типа компьютерных алгебраических систем, типографского набора и голосового синтеза. MathML может использоваться с целью кодирования и представления математического содержания для последующей высококачественной визуальной интерпретации для приложений, в которых основную роль играет семантика.

Хотя MathML – документы можно создавать вручную, ожидается, что всегда, кроме самых простых случаев, будут использоваться редакторы формул, программы преобразований и другие специализированные программные средства для работы с MathML. Уже существует несколько

версий таких программ, и разрабатываются еще как свободно распространяемые, так и коммерческие продукты.

Чтобы представить, как с помощью MathML кодируются математические выражения, рассмотрим следующий пример $\sum_{x=1}^{n} 2x$ описанный с помощью презентационных тегов (см. Код 1) и с помощью семантических тегов (см. Код 2). Для сравнения можно рассмотреть это же выражения в системе TeX (см. Код 3).

**Код 1**

```
<mrow>
  <munderover>
    <mo>&sum;</mo>
    <mrow><mn>x</mn><mo>=</mo><mn>1</mn></mrow>
    <mi>n</mi>
  </munderover>
  <mn>2</mn><mo>&InvisibleTimes;</mo><mi>x</mi>
</mrow>
```

**Код 2**

```
<apply>
  <sum/>
  <bvar><ci>x</ci></bvar>
  <interval>
    <cn>1</cn><cn>n</cn>
  </interval>
  <apply>
      <times/>
    <ci>2</ci>
    <cn>x</cn>
  </apply>
</apply>
```

**Код 3**

```
\sum\limits_{x=1}^n2x
```

### 2.2.2 Набор математических формул

Для набора математических формул в двух наиболее распространенных форматах LaTeX и MathML существует большое кол-во приложений, рассмотрим два наиболее весомых.

*MathType* (Design Science, http://www.dessci.com/) – профессиональный редактор формул, который является ActiveX объектом и следовательно с легкостью может встраиваться во многие приложения в том числе и Microsoft Word. В сравнении с редактором формул входящим в поставку Microsoft Office он имеет множество преимуществ:

- большее кол-во символов и шаблонов;
- выделение цветом;
- автоматическая нумерация формул и простановка ссылок;
- экспорт в форматах TeX, LaTeX, AMS-TeX, AMS-LaTeX, MathML1.0, MathML2.0
- и т.д.

Интересной возможностью является сохранения документа Word в HTML с формулами в виде:

- картинок в форматах GIP, JPEG;
- MathML для просмотра с помощью MathPlayer;
- MathML для просмотра с помощью WebEQ Viewer;
- MathML для просмотра с помощью IBM Techexplorer;

MathML для просмотра с помощью браузеров поддерживающих MathML Mozilla или Amaya.

*Workplace, Word, Notebook* (MacKichan Software, Inc, http://www.mackichan.com/) – это семейство продуктов, которые используются для создания и редактирования научных и математических документов. Говоря проще это текстовый редактор, сочетающий в себе легкость набора Microsoft Word и мощь, функциональность и качество печати LaTeX (в систему LaTeX встроен). Математические выражения можно набирать как с

помощью WYSIWYG редактора формул, так непосредственно в LaTeX. В системе предусмотрено около 150 стилей (журнал, статья, книга …).

В приложение встроены две Компьютерные алгебраические системы, собственная разработка компании MuPAD и система от Maple 5.1. MuPAD является довольно мощной системой, достаточно сказать лишь то, что эта система основана на объектно-ориентированном языке, она имеет отладчик, профайлер и т.д.

Система имеет возможность сохранять результат работы в формате LaTeX. Не плохо реализовано публикация документа в HTML (или XHTML) с возможностью сохранения формул в картинках различных форматов или в MathML (предпочтительно для IBM Techexplorer). Сохранение в формате PDF возможно лишь при наличии продуктов от Adobe.

Из полезных свойств можно отметить проверку орфографии на 19 языках.

Сочетание мощи TeX и простоты Word делает этот продукт очень привлекательным для набора научных документов.

### 2.2.3 Отображение математических формул

Существующие на данный момент способы отображения мат. формул можно подразделить на два основных класса: *генерация представления на стороне сервера* и *на стороне клиента*. Рассмотрим их по порядку. Огромное распространение имеет представление мат. формул в виде изображений в формате GIF и PNG. Оригинальным решением является отображение с помощью комбинации стандартных шрифтов HTML, однако оно не всегда дает приемлемое качество. Завоевывает все большее и большее распространение способ отображения с помощью графических форматов высокого сжатия основанных на волновых алгоритмах, это DjVu от LizardTech и LuraDocument от Algo Vision LuraTech, для их просмотра требуется установка специального плугина. Высокое качество отображения

дает векторный формат SVG (Scalable Vector Graphics) [16], просматриваемый с помощью стандартного плугина SVGView от Adobe.

Теперь перейдем к представлениям, генерируемым на стороне клиента.

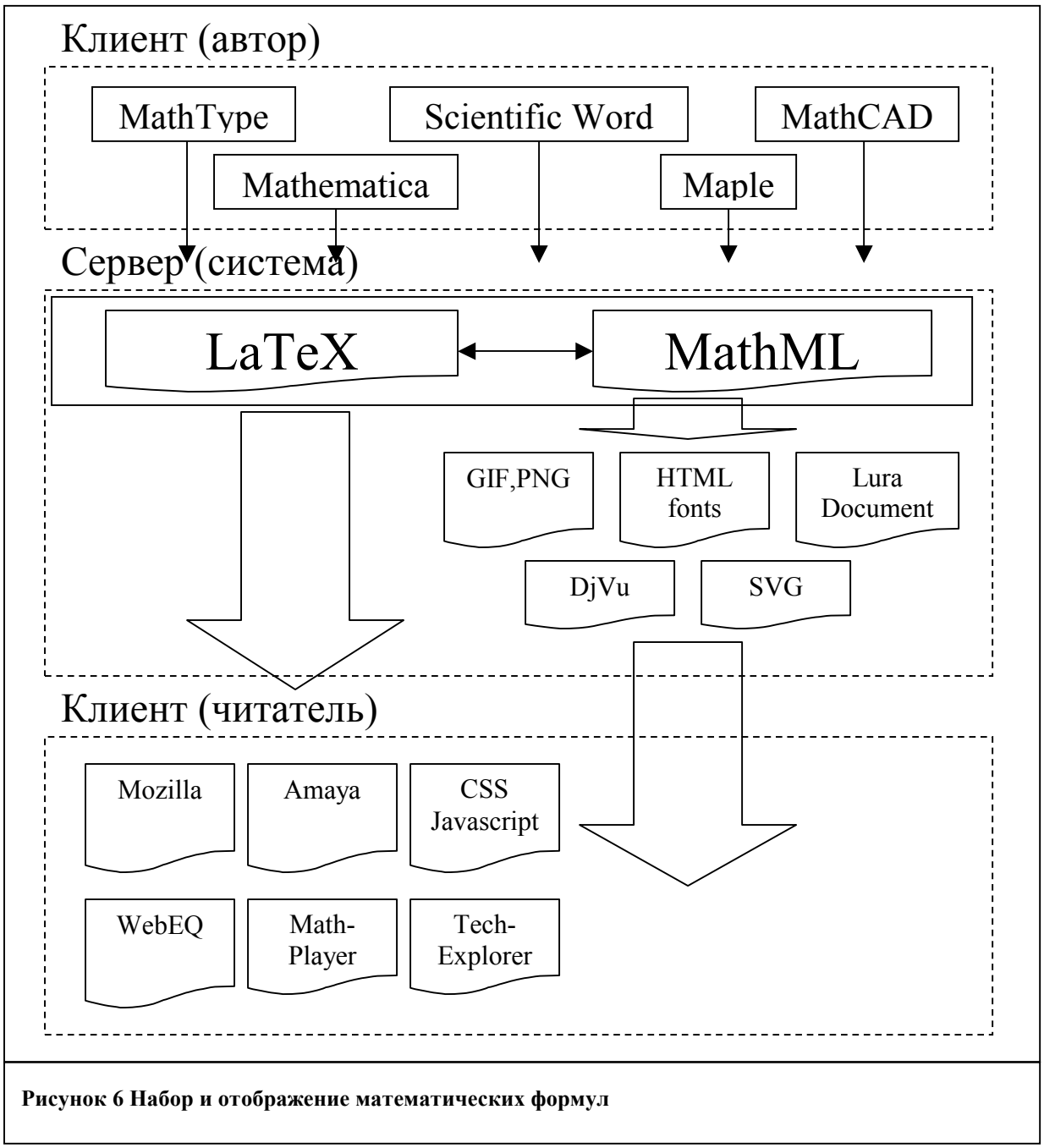


**Рисунок 6 Набор и отображение математических формул**

В настоящее время некоторые браузеры уже самостоятельно поддерживают отображение формул в MathML, в пример можно привести бурно развивающийся браузер Mozilla и браузер Amaya. Фирма Design Science предлагает решение WebEQ (http://www.dessci.com/webmath/webeq/),

отображение с помощью Java аплета. Она же предлагает ActiveX объект к браузеру, MathPlayer (http://www.dessci.com/en/products/mathplayer/welcome.asp). Конкурентом является фирма IBM предлагающая сходное, но уже коммерческое, решение, TechExplorer (http://www.software.ibm.com/techexplorer/).

Из широкого разнообразия, реализованных в системе форматов, читатель имеет возможность выбрать подходящий для него метод просмотра математических формул.

## 2.3 Систематизация

Подобная система немыслима без строгой систематизации, классификации статей. Существует несколько классификаций: Универсальная десятичная классификация, Десятичная классификация Дьюи (http://www.oclc.org/dewey/), Классификация и Предметные рубрики Библиотеки Конгресса и т.д.

Давайте рассмотрим, применяемую в системе, Универсальную Десятичную Классификацию (http://users.kpi.kharkov.ua/library/oglavlen.Htm). По этой классификации в России в 1962 году введено обязательное

| | | | |
|---|---|---|---|
| 600 | Техника (Прикладные науки) | | |
| | 630 | Сельское хозяйство и родственные отрасли | |
| | | 636 | Животноводство |
| | | | 637.7 Собаки |
| | | | 636.8 Кошки |

**Рисунок 7 Пример иерархии Универсальной Десятичной Классификации**

индексирование всех научных публикаций.

Система построена на содержательном принципе, что делает ее идеальной для организации базовых знаний: нотация записывается в везде признанных Арабских цифрах, хорошо определенны категории, хорошо разработана иерархия, существует богатая сеть взаимоотношений между темами. Базовый уровень делится на десять основных классов, которые покрывают всю область знаний. Основной класс далее делится на десять разделов, который в свою очередь, делится на десять секций (не все разделы и секции могут быть использованы). Первая цифра в каждом трех числовом номере представляет основной класс. Например, 500 представляет естественные науки и математику. Вторая цифра указывает раздел. Например 500 используется для базовых работ в науке, 510 для математики, 520 для астрономии, 530 для физики. Третья цифра обозначает секцию. Таким образом, 530 используется базовых работ в физике, 531 для классической механики , 532 для гидромеханики, 533 для газомеханики.

Арабские цифры используются для представления каждого класса в УДК. Десятичная точка следует за тремя цифрами номера класса, после которой деление на десять продолжается на уровни по мере необходимости. Тема может появляться более чем в одной дисциплине. Например «одежда» может рассматриваться в различных аспектах и участвовать в различных дисциплинах. Психологическое воздействие одежды принадлежит 155.95 как часть дисциплины психология; таможня, связанная с одеждой, находится в 391 как часть дисциплины о таможне; одежда в смысле моды лежит в 746.92 как часть дисциплины искусства.

Более чем 65000 тысяч взаимосвязанных понятий позволяют организовать кроме обычной выборки статей по категориям, качественный смысловой поиск.

## 2.4 Метаинформация

Является очевидным, что каждая статья есть ни что иное, как информационный ресурс, а значит, она имеет стандартный набор атрибутов. Для описания ресурсов можно использовать формат MARC (используемый Библиотекой Конгресса США и рядом других библиотек), тем более что разработано специальное расширение MARC для электронных документов. Этот формат позволяет очень детально каталогизировать электронный документ аналогично традиционной книге. Однако подобная детализация затрудняет использование MARC без соответствующего обучения и недоступно широкому кругу пользователей, создающих информационные ресурсы в Интернет. Простым способом отображения метаинформации является набор полей «Дублинское ядро».

Дублинское ядро это стандарт на представление метаинформации о ресурсе. Это словарь для стандарта на описание ресурса RDF (Resource Data Framework) [17] от W3C.

Инициатива "Дублинское ядро" (Dublin Core – DC) (http://purl.org/dc/) – это международная и междисциплинарная попытка определить основной набор элементов описания информационных ресурсов. Для согласования основного набора элементов с целью эффективного поиска ресурсов в Интернет проводились семинары библиотекарей, специалистов по информационным технологиям и телекоммуникациям, экспертов, специалистов по музейной информации. Первый Дублинский семинар, прошел в марте 1995 г.

В результате проведения семинаров были выработаны рекомендации по набору полей и методам каталогизации, доступным без предварительного обучения. В этот набор входит 15 полей, в которых описываются основные характеристики информационного ресурса.

1. Заглавие

2. Автор или создатель
3. Предмет и ключевые слова
4. Описание
5. Издатель
6. Сведения об ответственности
7. Дата создания или модификации
8. Тип ресурса
9. Формат данных
10. Идентификатор ресурса
11. Источник информации
12. Язык
13. Отношения с другими ресурсами
14. Зона действия (охват)
15. Правовые аспекты использования ресурса

Поля могут повторятся и кроме этого поле может разбиваться на подполя. В настоящее время стандартизован набор полей DC. Ряд полей имеет подполя, перечень которых еще не полностью определен. Этим занимается специальная рабочая группа, организованная на пятом семинаре DC в Хельсинки.

При описании поля вводятся понятия схемы и подполей. *Схема* – это наименование правил, в соответствии с которыми приводится содержание данного поля. Так например, для поля «Предмет» указывается какая система классификации используется, для поля «Дата» указывается какой стандарт представления даты и т.д. *Подполе* – это информация, утоняющая содержание поля. В каждом поле, имеющем подполя, выделятся одно

подполе, называемое «подполе по умолчанию». Если информация не разделена на подполя, считается, что она занесена в подполе по умолчанию. Например, поле «Отношения» имеет подполя «Первоисточник», «Оригинал», «Составная часть» и т.д.

Сопоставление каждой статье набора атрибутов Дублинского ядра позволяет, реализовать качественный унифицированный поиск.

## 2.5 Защита прав автора

Соблюдение прав автора самый краеугольный камень электронных публикаций. Однако уже все механизмы и инструменты их обеспечения уже разработаны. Это *трастовые системы управления цифровыми правами* [18] обеспечивающие следование, пользователями системы, некоторых правил. *Расширяемый язык разметки прав* XrML (eXtensible Rights Markup Language) используемый для описания прав трастовых систем. *Идентификатор цифрового объекта* DOI (Digital Object Identifier) – уникальный не изменяющийся идентификатор произведения, обеспечивающий связь между пользователем произведения и обладателем прав на это произведение, присвоенный специально созданным международным фондом DOI. *Само-защищающийся документ* SPD (Self-Protecting Document) – активный документ, сохраняющий свою конфиденциальность и целостность, принуждающий пользователей следовать правам, связанным с этим документом. В данный момент многие компании предлагают различные решения, реализующие эти технологии, например: Adobe, Glassbook, ContentGuard, InterTrust Technologies, SoftLock и т.д.

Адаптация всех этих механизмов к системе позволит решить техническую сторону проблемы защиты прав автора.

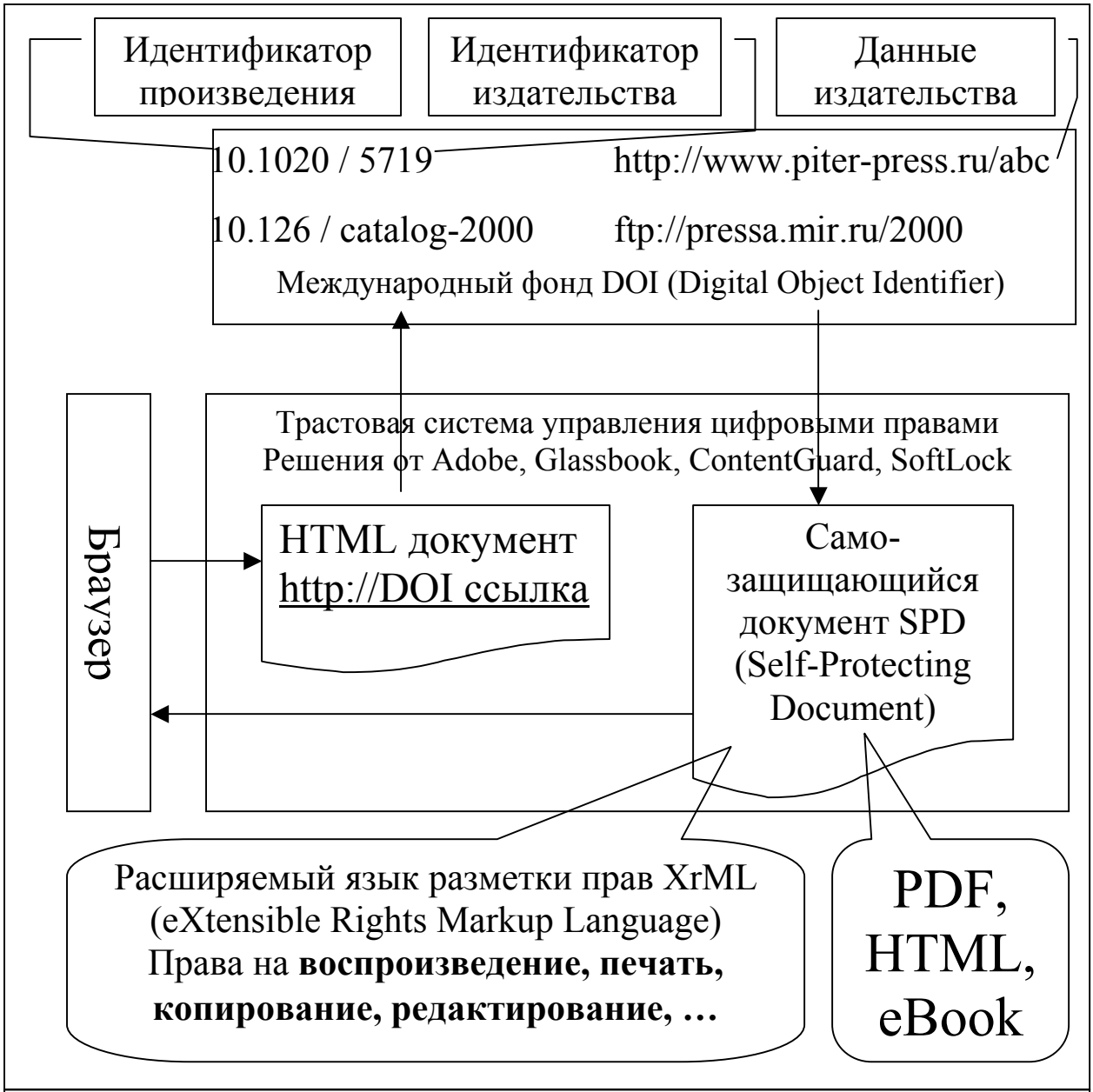


**Рисунок 8 Трастовая система управления цифровыми правами**

## 2.6 Извлечение знаний

Объем накопленных человечеством знаний огромен. Нередко время для изучения всех накопленных материалов, прочтения всех существующих книг и статей, по интересующей тематике, ограниченно временем человеческой жизни. Чтобы получить ответ на интересующий нас вопрос, как правило, нам приходится прочитать многие и многие тысячи строк текста, изучить не один десяток страниц. Однако процесс поиска и

извлечения знаний можно автоматизировать. Для этого машину нужно заставить «понимать» смысл текста.

Существуют реализованные разработки ассоциативно-статистического подхода анализа текста [19], основанного на результатах нейропсихологических исследований [20], которые установили, что анализ печатного текста, опираясь на зрительное пространственное (а не на линейное слуховое) восприятие, реализуется преимущественно правым полушарием мозга, использующим ассоциативную статистическую модель. В основе подхода лежит интегральное представление смысла текста в форме ассоциативной семантической сети, в качестве критерия построения которой используется частота совместной встречаемости понятий в предложениях текста. Читатель, исследуя такую сеть, представленную в виде тематического дерева (дерево ключевых и связанных с ними понятий текста), может в значительной степени ускорить процесс исследования текста и выявления требуемой информации. Семантическая сеть так же используется для решения таких задач, как автоматическое реферирование, тематическая классификация и кластеризация текстов, смысловой поиск, автоматическое

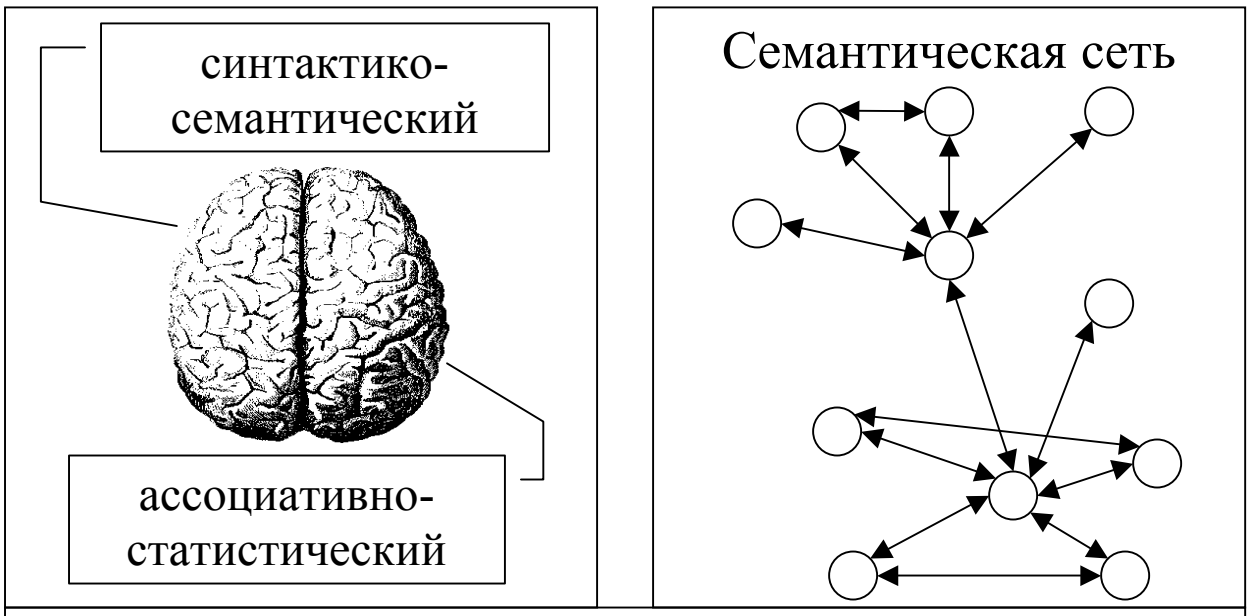


**Рисунок 9 Семантическая сеть в ассоциативно-статистическом подходе**

построение гипертекста и т.д.

### 2.6.1 Решения на ассоциативно-статистическом подходе

*InterMedia Text* – картридж для Oracle8i, основной задачей которого является задача поиска документов по их содержанию, при этом используются следующие свойства:

- ✓ расширение слов запроса всеми морфологическими формами, что реализуется привлечением знаний о морфологии языка.
- ✓ interMedia Text допускает расширение слов запроса близкими по смыслу словами за счет подключения тезауруса – семантического словаря.
- ✓ расширение запроса словами, близкими по написанию и по звучанию - нечеткий поиск и поиск созвучных слов.

interMedia Text пользуется тезаурусом – семантическим словарем, содержащим около полумиллиона слов, которые классифицированы по тематическим категориям (рубрикам) и синонимическим рядам: для каждого слова установлены его синонимы, более общие и более частные понятия, а также "родственные" слова, часто имеющие с ним смысловую связь в тексте.

Так, если контекстный поиск находит все документы, содержащие заданные слова, то тематический поиск возвращает лишь те документы, в которых словам запроса соответствует одна из ключевых тем. Кроме того, он позволяет найти документы, вовсе не содержащие слов из названия заданной темы, однако имеющие к ней отношение.

Существует русская локализованная версия interMedia Text – Russian Context Optimizer, разработанная компанией Гарант-Парк-Интернет. К сожалению, тезауруса в ней пока нет.

*TextAnalyst* (Microsystems Ltd, http://www.analyst.ru/) – система автоматического анализа текста. TextAnalyst разработан в качестве инструмента для анализа содержания текстов, смыслового поиска информации, формирования электронных архивов, и предоставляет пользователю следующие основные возможности:

- ✓ анализа содержания текста с автоматическим формированием семантической сети с гиперссылками – получения смыслового портрета текста в терминах основных понятий и их смысловых связей;
- ✓ анализа содержания текста с автоматическим формированием тематического древа с гиперссылками – выявления семантической структуры текста в виде иерархии тем и подтем;
- ✓ смыслового поиска с учетом скрытых смысловых связей слов запроса со словами текста;
- ✓ автоматического реферирования текста – формирования его смыслового портрета в терминах наиболее информативных фраз;
- ✓ кластеризации информации – анализа распределения материала текстов по тематическим классам;
- ✓ автоматической индексации текста с преобразованием в гипертекст;
- ✓ ранжирования всех видов информации о семантике текста по «степени значимости» с возможностью варьирования детальности ее исследования;
- ✓ автоматического/автоматизированного формирования полнотекстовой базы знаний с гипертекстовой структурой и возможностями ассоциативного доступа к информации.

## 2.7 Система рейтингов

Существует понятие рейтинга журнала, так называемый Impact Factor [21], основанный на частоте цитирования его статей. В рамках разрабатываемой системы, предлагается новый альтернативный метод подсчета рейтинга, основанный не на трудоемком (как правило, ручном) подсчете ссылок на статью, а на взаимосвязи трех основных понятий: *рейтинга автора* (РА), *рейтинга статьи* (РС) и *рейтинга журнала* (РЖ). РЖ оценивается как суммарный рейтинг всех его статей. РС зависит от суммарного рейтинга всех ее авторов, а так же кол-ва реплик в форуме ассоциированных с данной статьей и рейтинга авторов сделавших эти реплики. РА оценивается как суммарный рейтинг всех его статей и рейтинг журналов, в которых они размещены.

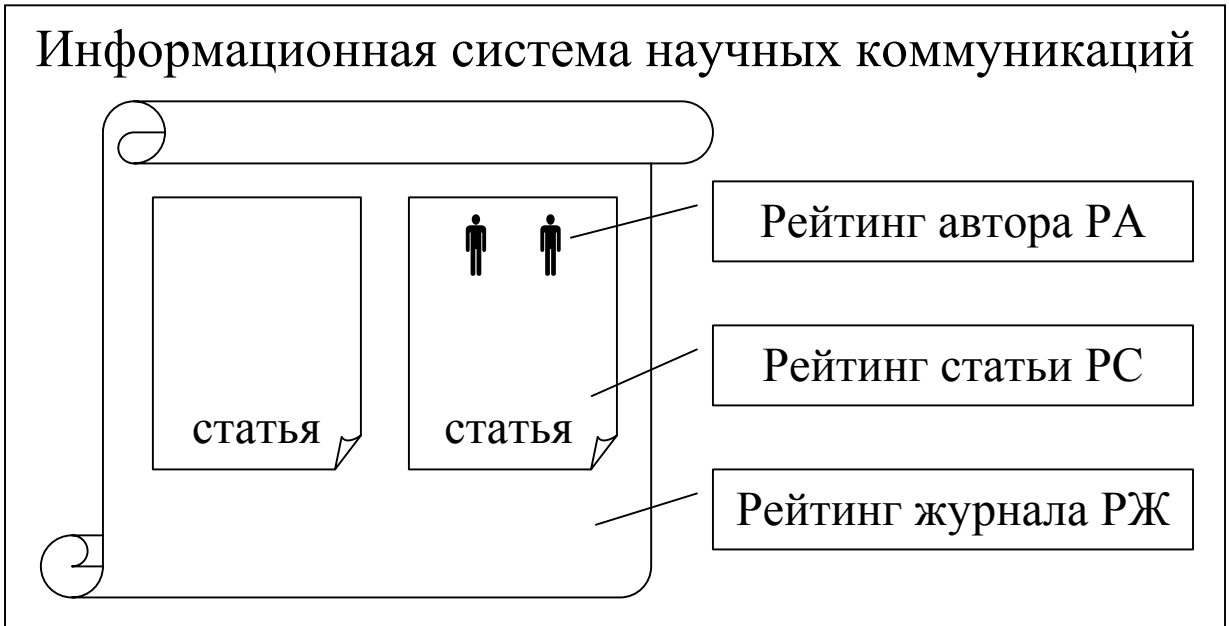


**Рисунок 10 Основные элементы системы рейтингов**

Проектируемая система подсчета, основанная на множестве этих и других различных факторов, позволит в автоматическом режиме и более качественно оценить  важность и достоверность научной публикации, тем самым, предоставляя читателю возможность выборочного чтения наиболее рейтинговой информации.

# Глава 3. Архитектура системы

К моменту написания данной работы, процесс разработки ИСНК не завершен. Различные части системы находятся на различных стадиях. Однако основа системы уже заложена и будет развиваться в дальнейшем.

**Таблица 3 Состояние разработки ИСНК на момент написания данной работы**

| Наименование | Анализ | Проектирование | Реализация | Апробация |
|---|---|---|---|---|
| Ядро системы | Да | Да | Да | Нет |
| Web уровень | Да | Да | Да | Нет |
| Уровень хранения данных | Да | Да | Да | Нет |
| Хранение статей | Да | Да | Да | Нет |
| Контроль версий и совместная авторская разработка | Да | Да | Да | Нет |
| Преобразование входных форматов в выходные | Да | Да | Да | Да |
| Отображение математических формул | Да | Да | Да | Да |
| Систематизация (УДК) | Да | Да | Нет | Нет |
| Метаинформация | Да | Да | Да | Нет |
| Защита прав автора | Да | Нет | Нет | Нет |
| Извлечение знаний | Да | Нет | Нет | Нет |
| Система рейтингов | Да | Да | Нет | Нет |
| Электронная дискуссия | Да | Да | Нет | Нет |

В соответствии с предъявляемыми требованиями, система полностью строится на кроссплатформенном языке Java. Внешние, подключаемые конвертеры, являются зависимыми от операционной системы, однако их исходный код свободно доступен и написан на платформо-переносимых языках C, C++, Perl, что в конечном итоге все же позволяет обеспечить платформонезависимость системы.

## 3.1 Входные, выходные форматы и отображение мат. формул

**Таблица 4 Преобразование форматов**

| | Входные (авторские) | | |
|---|---|---|---|
| Выходные (читательские) | | RTF | LaTeX |
| | **Просмотр** | | |
| | HTML | rtf2latex | tex4ht |
| | DjVu | rtf2latex | djvulibre |
| | **Печать** | | |
| | PDF | rtf2latex | pdflatex |
| | PostScript | rtf2latex | ghostscript |
| | **Редактирование** | | |
| | RTF | --- | latex2rtf |
| | LaTeX | rtf2latex | --- |
| | **Обмен (XML)** | | |
| | XHTML | rtf2latex | tex4ht |
| | TEI | rtf2latex | tex4ht |
| | DocBook | rtf2latex | tex4ht |
| | **Специальные** | | |
| | WML | rtf2latex | XSLT+XSLFO |
| | eBook (OpenBook) | rtf2latex | XSLT+XSLFO |

Это наиболее проработанная часть ИСНК. Для ее апробации было сделано отдельное Web приложение – Онлайновые текстовый конвертер (ОТК). На момент написания работы он доступен по адресу http://otc.udsu.ru ОТК уже практически используется студентами и преподавателями Удмуртского Государственного Университета для подготовки материалов при публикации в Web. На данный момент ОТК находится на стадии тестирования и уже обеспечивает:

1. Преобразование документов из двух наиболее распространенных форматов RTF и LaTeX.
2. Корректную работу с русским и английским языком.
3. Правильное преобразование математических формул как из формата RTF (подготовленного с помощью Microsoft Word), так и из формата LaTeX.

Преобразование форматов и отображение математических формул реализовано исключительно на основе бесплатных инструментов.

1. rtf2latex – http://members.home.net/setlur/rtf2latex2e/
2. tex4ht – http://www.cis.ohio-state.edu/~gurari/TeX4ht/
3. djvulibre – http://djvu.sourceforge.net/
4. latex2rtf – http://sourceforge.net/projects/latex2rtf/
5. pdflatex – входит в большинство поставок LaTeX
6. ghostscript – http://www.ghostscript.com/

Процесс апробации ОТК показал достаточную готовность всех рассмотренных цепочек преобразования форматов и инструментов для отображения математических формул.

## 3.1 Совместная авторская разработка и контроль версий

Совместная авторская разработка и контроль версий в Web (Web Distributed Authoring and Versioning) [22] – так называется стандарт, расширяющий функции HTTP 1.1, если говорить коротко, возможностями

сохранения информации на сервере. В настоящее время данный стандарт бурно развивается, в качестве примера можно привести то, что его поддерживают такие продукты как: Microsoft Internet Explorer 5.0 и выше, Windows 2000 и выше, Office 2000 и выше; Adobe Photoshop 6.0 и выше; Atlovas XML Spy; и т.д. В WebDAV все файлы хранятся в абстрактной файловой системе, которая легко может распределена. Физически же файлы могут находиться в обычной файловой системе, СУБД или другом хранилище информации. Каждый ресурс может иметь неограниченный набор атрибутов, в нашем случае набор атрибутов Дублинское ядро. С помощью расширения DASL (DAV Searching & Locating) [24] стандартизован поиск по этим атрибутам. Для организации совместной авторской разработки реализуется механизм блокировок. Расширение DeltaV позволяет хранить несколько версий файла, разветвлять дерево ревизий и ставить метки. С помощью спецификации ACL (Access Control Extension) [23] возможно с легкостью реализовать гибкую систему управления правами доступа.

WebDAV реализуется в ИСНК с помощью библиотеки Slide из проекта Apache Jakarta (http://jakarta.apache.org/slide/), именно она обеспечивает

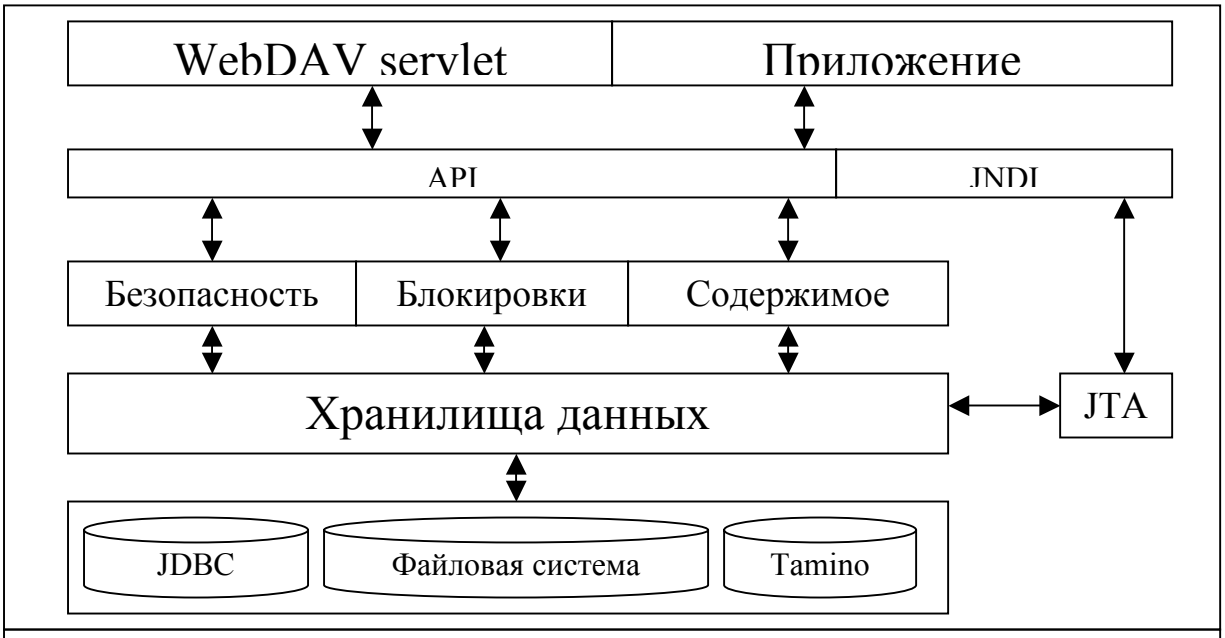


**Рисунок 11 Совместная авторская разработка и контроль версий (WebDAV)**

хранение публикаций.

Реализация протокола WebDAV во многих распространенных программных продуктах, позволяет с легкостью взаимодействовать с системой даже начинающим пользователям.

## 3.2 Отображение данных в Web и хранение данных

Web уровень спроектирован с использованием паттерна MVC (Model

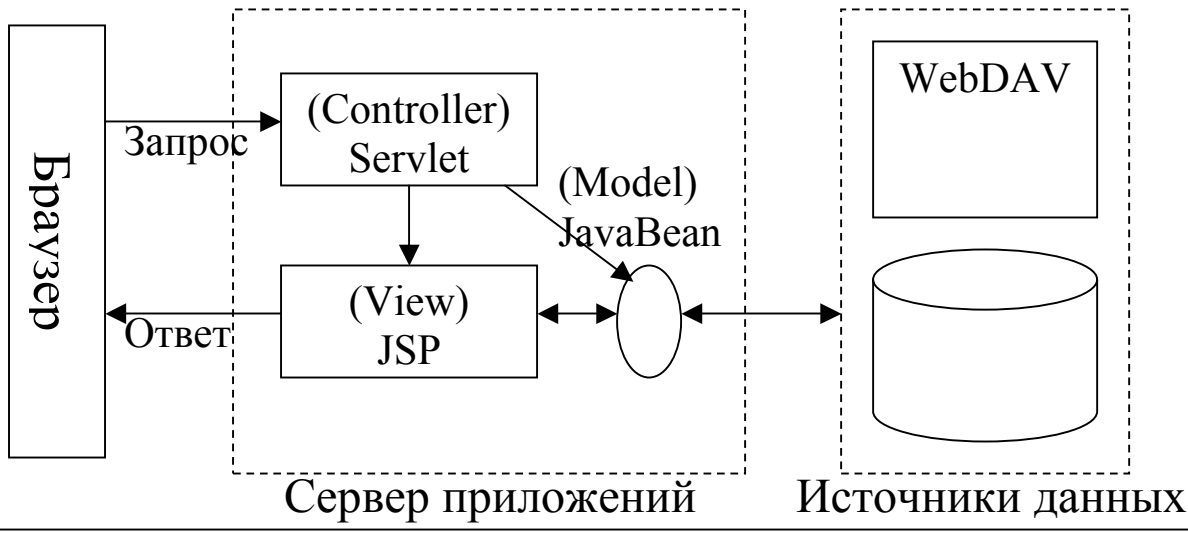


**Рисунок 12 Модель отображения данных в Web (MVC)**

View Controller) [25,26]. Эта модель (MVC) является наиболее прогрессивной на сегодняшний момент, и позволяет обеспечить максимальную масштабируемость и расширяемость системы.

Web уровень реализуется с помощью библиотеки Struts из проекта Apache Jakarta (http://jakarta.apache.org/struts/). Платформой является контейнер Tomcat из проекта Apache Jakarta (http://jakarta.apache.org/tomcat/).

Для взаимодействия JavaBean'ов приложения с реляционной СУБД используется Object Relational Bridge (OJB) с построением запросов на JDO (Java Data Object) (http://java.sun.com/products/jdo/) [27] через Object Transaction Manager (OTM).

**Таблица 5 Сравнение технологий хранения постоянных объектов**

| | Сериализация | JDBC | ODBMS | EJB | JDO |
|---|---|---|---|---|---|
| Транзакции | Нет | Да | Да | Да | Да |
| Возможность запросов | Нет | Да | Да | Да | Да |
| Стандартный API | Да java.io | Да JDBC | Нет ODMG[3] | Да EJB | Да JDO |
| Стандартный язык запросов | Нет | Нет SQL[4] | Нет OQL | Да EJBQL | Да JDOQL |
| Поддерживаемые парадигмы хранения данных | Файловая система | RDBMS | ODBMS | RDBMS, EAI | RDBMS, ODBMS, EAI, Файловая система, и др. |
| Прозрачное закрытие постоянных объектов | Нет | Нет | Да | Нет | Да |
| Прозрачная модель доменов | Нет | Нет | Да | Нет | Да |
| Истинно объектная база данных | Нет | Нет | Да | Нет | Нет[5] |
| Поддержка существующей табличной структуры | Нет | Да | Нет | Нет | Да[6] |

По мнению многих аналитиков в ближайшем будущем технология хранения постоянных объектов JDO примет наибольшее распространение в

[3] Этот стандарт еще качественно не реализован
[4] Зависит от конкретной реализации производителя
[5] Имеет объектно-ориентированный API
[6] Это не стандартизовано в JDO, однако все реализации поддерживают эту возможность.

сравнении с уже используемыми технологиями, таких как ODMG (Object Data Management Group) и EJB (Enterprise Java Beans).

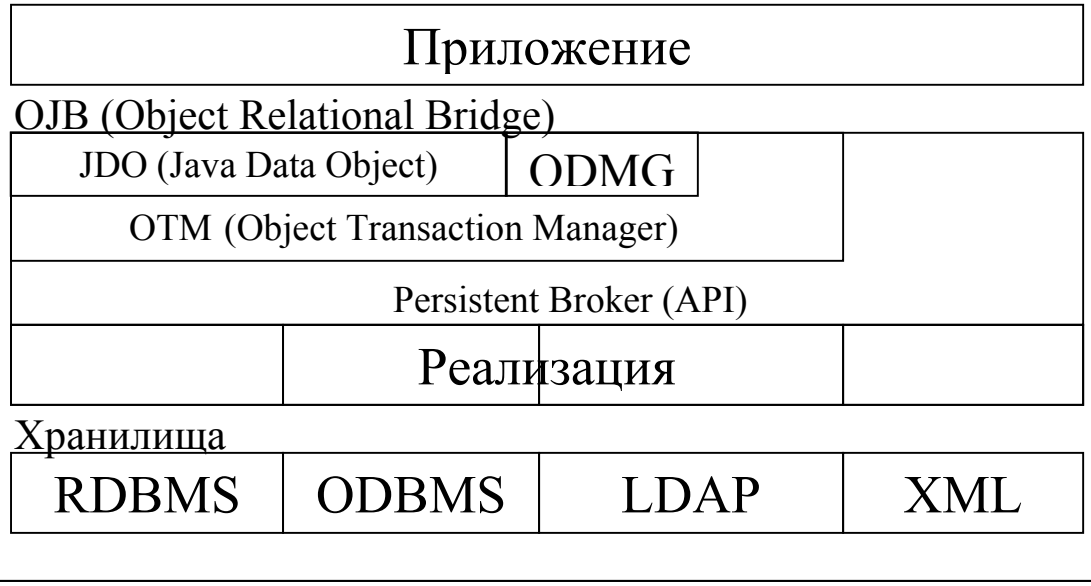


**Рисунок 13 Модель хранения данных (JDO)**

Для реализации хранения данных используется библиотека OJB из проекта Apache Jakarta (http://jakarta.apache.org/ojb/).

Претендентами на роль СУБД в системе, были MySQL и PostgreSQL. Выбор был сделан в пользу PostgreSQL, возможностей которой достаточно для обработки больших объемов данных. MySQL для данной задачи не подходит, в силу отсутствия необходимого быстродействия.

Использование с одной стороны наиболее прогрессивных методологий, технологий и бесплатных решений с другой, позволяет добиться необходимой дешевизны, гибкости, масштабируемости и функциональности системы.

# Заключение

Существующие на данный момент методы публикации научных материалов уже не могут обеспечить требуемое качество и скорость. Громоздкие и дорогие системы оказались не доступными российским центрам научной коммуникации (университеты, институты, академии).

Осознание необходимости построения системы, подобной ИСНК, пришло давно. До сего времени шел процесс накопления и оттачивания знаний. Наступил момент нового витка в эволюции процессов информационного обмена в научном сообществе.

Не так давно разработанные стандарты платформы XML [28] и реализованные на их основе инструменты, подвели мощнейшую базу для построения интероперабельных открытых систем. Только в 2001-2002 годах, были разработаны реально применимые инструменты для извлечения знаний. К этому же времени были практически отработаны механизмы защиты интеллектуальной собственности, являющимися ключевыми для построения реально используемой ИСНК. Только 10 января 2002 года был принят федеральный закон РФ «Об электронной цифровой подписи», являющийся основополагающим в механизмах защиты авторского права.

К настоящему моменту существуют и отработаны все компоненты необходимые для построения качественно новой информационной системы научных коммуникаций.

Исследования показали, что развитие большого количества технологий и инструментов, разрабатываемых под лицензиями GNU, GPL, Apache, позволяют создать недорогую, но в тоже время достаточно мощную систему, обеспечивающую необходимое качество и скорость обмена информацией в научном сообществе.

Результаты практической апробации системы преобразования форматов и отображения математических формул показали достаточную готовность подобранного инструментария. Однако в процессе его тестирования были выявлены и исправлены уже более 10 ошибок и недочетов. Некоторые элементы системы преобразования форматов и отображения математических формул все еще требуют доработки.

Многие из проанализированных частей ИСНК не достигли даже стадии проектирования, некоторые спроектированные модули требуют реализации. Однако даже рассмотренные в данной работе модули системы автоматизируют не все этапы жизненного цикла публикации. Требуются еще дополнительные разработки в области автоматизации внесения знаний в систему, отображения химических формул и т.д.

Продолжение работ в данной области необходимо для поддержания российской науки и образования на надлежащем уровне, а значит и улучшения состояния экономики, качества жизни людей, национальной безопасности и повышения роли в мировом сообществе.

# Список использованной литературы


1. Э.М. Мирский. Массив публикаций и система научной дисциплины. "Системные исследования", 1977 г.
2. С.Г. Маслов, С.В. Моченов. Методология создания электронных учебных изданий //Материалы XXIII научно-метод. конф. ИжГТУ (19-22 февраля).- Ижевск: Изд-во ИжГТУ, 2001.- С. 29-31.
3. Thomas Krichel, С.И. Паринов. База данных RePEc и ее российский партнер система Соционет //Электронные библиотеки, 2002, Том 5, Выпуск 2.- 2002 (http://www.elbib.ru/journal/2002/200202/KP/KP.ru.html)
4. С.И. Паринов, В.М. Ляпунов, Р.Л. Пузырев. Система Соционет как платформа для разработки научных информационных ресурсов и онлайновых сервисов //Электронные библиотеки, 2003, Том 6, Выпуск 1.- 2003 (http://www.elbib.ru/journal/2003/200301/PLP/PLP.ru.html)
5. А.В. Жучков, С.А. Арнаутов. Единая среда распределенных ресурсов (GRID) и цифровые библиотеки. Институт химической физики им. Н.Н. Семенова РАН. (http://rcdl2001.krc.karelia.ru/papers/papers/zhuchkov_arnautov/arnautov_paper.rtf)
6. А.Н. Гребнев. Научные информационные системы // Вестник УдГУ: Математика.- Ижевск: Изд-во УдГУ, 2003. С.99-106.
7. А.Н. Гребнев. Система электронных публикаций научных статей //Материалы XL Международной научной студенческой конференции «Студент и научно-технический прогресс»:

Информационные технологии / Новосиб. гос. ун-т. Новосибирск, 2002. С. 32-33.

8. А.Н. Гребнев. Необходимость информационной системы научной коммуникации. //Сборник работ членов Студенческого научного общества и Научного общества молодых ученых и аспирантов УдГУ 2002-03 гг.- Ижевск: Изд-во УдГУ, 2003. (в печати).
9. А.Н. Гребнев. Информационная система научных коммуникаций //Материалы XLI Международной научной студенческой конференции «Студент и научно-технический прогресс»: Информационные технологии / Новосиб. гос. ун-т. Новосибирск, 2003 (в печати).
10. О.В. Кукушкина, А.А Смирнов, А.А. Соколов. Текстология и гипертекстология сети. «Текстология», 2002. (http://www.textology.ru/public/textnet.html)
11. S.M. Griffin NSF/DARPA/NASA Digital Libraries Initiative. D-Lib Magazine, July/August 1998.
12. М.Р. Когаловский. Научные коллекции информационных ресурсов в электронных библиотеках. Институт проблем рынка РАН. 1999. (http://www.dl99.nw.ru/PDF/02.pdf)
13. С.А. Арнаутов. Роль и место виртуальных цифровых библиотек в Интернете. Институт химической физики им. Н.Н. Семенова РАН.2001. (http://rcdl2001.krc.karelia.ru/papers/papers/arnautov/arnautov_paper.rtf)
14. П.А. Дмитриев. Проектирование комплексных систем поддержки Электронных Изданий. Вычислительный центр РАН. Москва. 2001.

(http://rcdl2001.krc.karelia.ru/papers/papers/dmitriev/complete_text.doc)

15. *Mathematical Markup Language (MathML) Version 2.0*, W3C Recommendation, 21 February 2001. (http://www.w3c.org/TR/MathML2)
16. *Scalable Vector Graphics (SVG) Version 1.0*, W3C Recommendation, 04 September 2001. (http://www.w3.org/TR/SVG/)
17. *Resource Description Framework (RDF)Model and Syntax Specification*, W3C Recommendation, 22 February 1999. (http://www.w3c.org/TR/1999/REC-rdf-syntax/)
18. Правовая охрана и защита цифровой интеллектуальной собственности. Институт Открытое общество. (http://www.intellect.vsu.ru)
19. А.Е. Ермаков, В.В. Плешко. Ассоциативная модель смысла текста в прикладных задачах компьютерного анализа полнотекстовых докуметов. // Русский язык: исторические судьбы и современность. Международный конгресс. Труды и материалы.- Москва, МГУ, 2001. (http://research.metric.ru/art_ling.asp)
20. Н.Н. Брагина, Т.А. Доброхотова. Функциональные ассиметрии человека.- М.: Медицина, 1981.- 287 с.
21. М.В. Алфимов, А.Н. Либкинд, И.А. Либкинд, В.А. Минин. Информационные потоки в РФФИ: Новый подход к цитированию. (http://intra.rfbr.ru/pub/vestnik/V4_0 1 /1 _1 .htm)
22. HTTP Extensions for Distributed Authoring WEBDAV RFC 2518 (http://asg.web.cmu.edu/rfc/rfc2518.html)

23. Versioning Extensions to WebDAV (Web Distributed Authoring and Versioning) RFC 3253 (http://www.faqs.org/rfcs/rfc3253.html)

24. DAV Searching and Locating Protocol (http://www.webdav.org/dasl/protocol/draft-davis-dasl-protocol-00.html)

25. Understanding JavaServer Pages Model 2 architecture. JavaWorld magazine. 1999. (http://www.javaworld.com/javaworld/jw-12-1999/jw-12-ssj-jspmvc_p.html)

26. Model-View-Controller. Java Blue Prints. Sun Microsystems. 2002. (http://java.sun.com/blueprints/patterns/MVC-detailed.html)

27. Robin M. Roos. Java Data Objects. Addison-Wesley. Pearson Education. 2003.

28. М.Р Когаловский. Стандарты платформы XML и баз данных. Институт проблем рынка РАН. 2001. (http://rcdl2001.krc.karelia.ru/papers/papers/kogalovsky/kogalovsky_paper.rtf)