

Сибирский государственный аэрокосмический университет
имени академика М.Ф. Решетнёва

Цифровое библиотечное хранилище на базе CMS Битрикс

Дамов Михаил Витальевич
Зотин Александр Геннадьевич

Исследование выполнено при поддержке Министерства образования и науки Российской Федерации, соглашение 14.В37.21.1526

Красноярск – 2013

Цели и задачи ЭБС

Разрабатываемая электронно-библиотечная система должна обеспечивать выполнение следующих целей и задач:

1. Функционирование библиотеки в соответствии с законодательством РФ и отраслевыми библиотечными и Интернет- стандартами;
2. Интеллектуальный поиск по библиотечному каталогу и полному тексту единиц хранения;
3. Доступ с любого Интернет-узла в соответствии с ролями «студент», «преподаватель», «автор», «библиотекарь», «администратор»;
4. Наличие расширенной статистики по посетителям электронной библиотеки и цифровому книжному фонду;
5. Возможность интеграции в существующие системы автоматизации библиотеки;
6. Возможность размещения мультимедийных материалов, например видео- и аудиофайлов;
7. Предоставление изданий с сохранением вида страниц.

Постановка задачи

Целью настоящей работы является модификация формата хранения мультимедийных электронных книг для обеспечения поддержки внедренных видео- и аудио-материалов и возможности сохранения оригинального вида страниц и верстки книг

Формат должен обеспечивать выполнение следующих задач:

1. Простота хранения и доступа к электронной книге, используя различные виды устройств (РС, планшет, смартфон, мобильный телефон) и ПО (браузер, специализированное ПО);
2. Возможности поиска по тексту и метаданным;
3. Расширенные возможности навигации по текстовому и мультимедиа-содержанию;
4. Возможность отображения электронной книги в модифицированном формате на устройствах без поддержки расширений формата;
5. Простота дальнейшего расширения формата;
6. Возможность интеграции в существующие системы автоматизации библиотеки.

Стандарты и форматы

Стандарты электронных каталогов

UNIMARC

MARC21

RUSMARC

Стандарты доступа

OPAC

Z39.50

Стандарты метаданных

MARC21/MARCXML

MODS

MADS

EAD

Стандарты электронных книг

Word Document
(DOC)

Microsoft Help
(CHM)

Hypertext
(HTML)

Portable Document
Format (PDF)

DJVU

Electronic Publication
(EPUB)

Формат EPUB

Достоинства

- Формат на основе HTML5 и XML;
- Широкий набор ПО для создания, редактирования и просмотра документов
- Поддержка документов с внедренными объектами формул и мультимедиа
- Формат открыт для модификации

Недостатки

- Недостаточная поддержка возможностей для образовательных и научных целей
- Последние версии базовых форматов не являются стандартом и находятся в стадии утверждения
- Отсутствие возможностей для сохранения оригинальной верстки книг

Сохранение оригинального вида и верстки книги

<PAGE size="210mm 297mm" margin="20mm 10mm 10mm 10mm">page content </PAGE>

```
<?xml version="1.0" encoding="UTF-8"?>
<html>
<body>
<page>
<p>Lorem ipsum dolor sit amet, consectetur adipiscing elit. Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies, purus lectus malesuada libero, sit amet commodo magna eros quis urna. Nunc viverra imperdiet enim. Fusce est.</p>
<p>Vivamus a tellus. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Proin pharetra nonummy pede. Mauris et orci. Aenean nec lorem.</p>
</page>
<page>
<p>In porttitor. Donec laoreet nonummy augue. Suspendisse dui purus, scelerisque at, vulputate vitae, pretium mattis, nunc. Mauris eget neque at sem venenatis eleifend. Ut nonummy.</p>
<p>Fusce aliquet pede non pede. Suspendisse dapibus lorem pellentesque magna. Integer nulla. Donec blandit feugiat ligula. Donec hendrerit, felis et imperdiet euismod, purus ipsum pretium metus, in lacinia nulla nisl eget sapien.</p>
</page>
</body>
</html>
```

Рис. 1. Пример XML-разметки книги

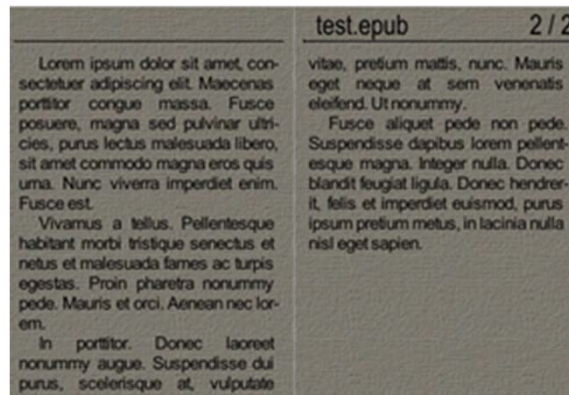


Рис. 3. Изображение текста книги в CoolReader 3.0

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies, purus lectus malesuada libero, sit amet commodo magna eros quis urna. Nunc viverra imperdiet enim. Fusce est.

Vivamus a tellus. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Proin pharetra nonummy pede. Mauris et orci. Aenean nec lorem.

In porttitor. Donec laoreet nonummy augue. Suspendisse dui purus, scelerisque at, vulputate vitae, pretium mattis, nunc. Mauris eget neque at sem venenatis eleifend. Ut nonummy.

Fusce aliquet pede non pede. Suspendisse dapibus lorem pellentesque magna. Integer nulla. Donec blandit feugiat ligula. Donec hendrerit, felis et imperdiet euismod, purus ipsum pretium metus, in lacinia nulla nisl eget sapien.

Рис. 2. Изображение текста книги в браузере

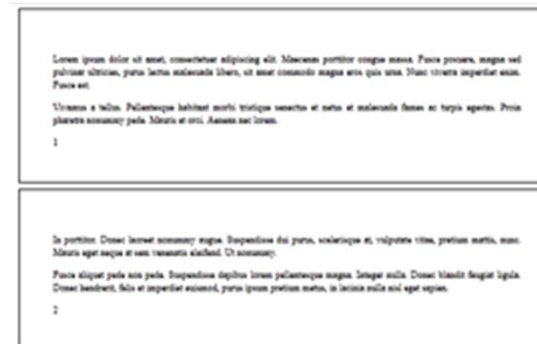
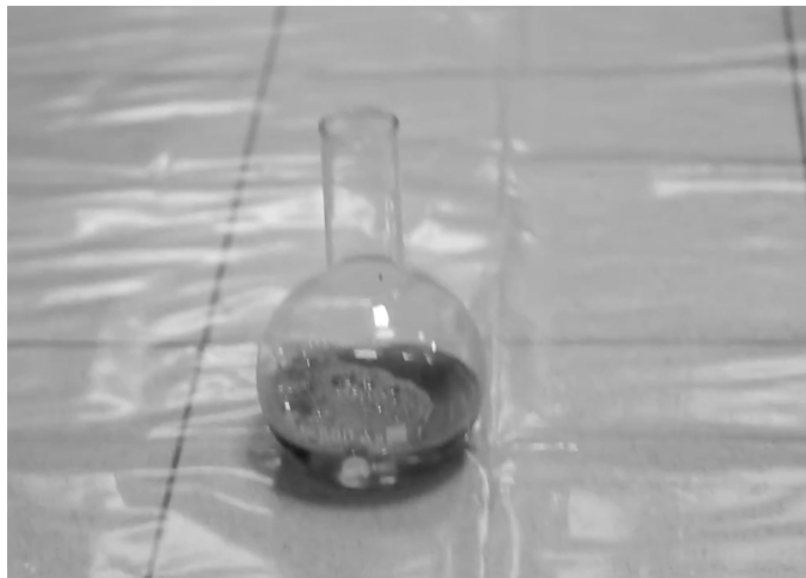


Рис. 4. Изображение текста книги в браузере сохранением оригинального вида страниц

Поддержка внедренного мультимедиа-содержимого

`<VIDEO src="test.mp4" poster="test.jpg">`This is fallback content to display if the browser does not support the video element.

`<SCENE start="01:04" name="Scene 01" poster="poster.jpg">` `<p>`Description`</p>`
`</SCENE></VIDEO>`



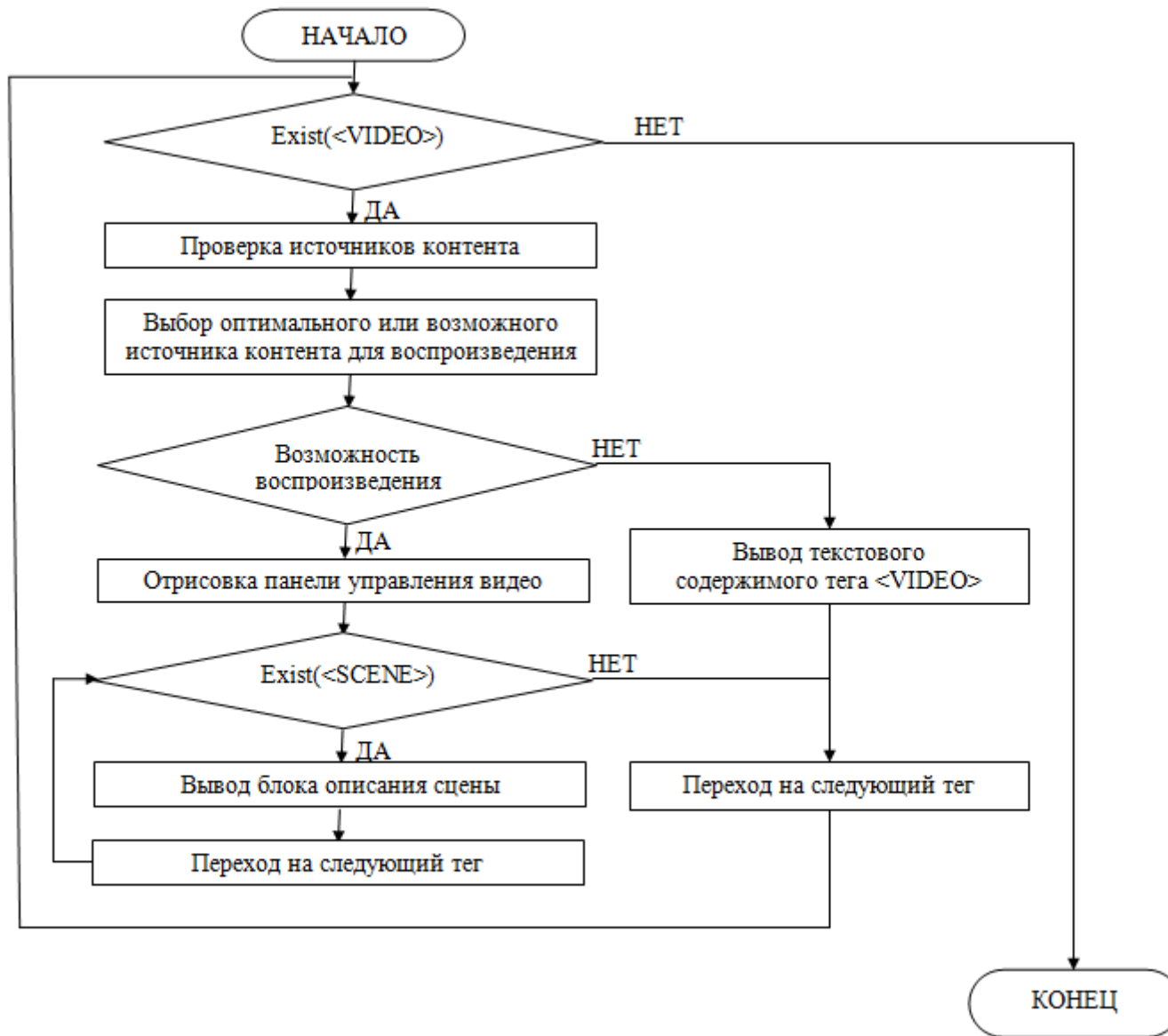
[01:04](#) Scene 01

In porttitor. Donec laoreet nonummy augue. Suspendisse dui purus, scelerisque at, vulputate vitae, pretium mattis, nunc. Mauris eget neque at sem venenatis eleifend. Ut nonummy.

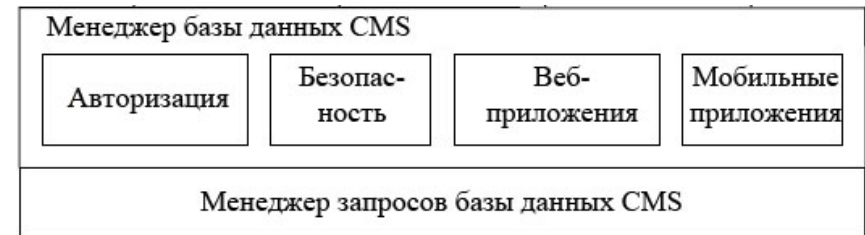
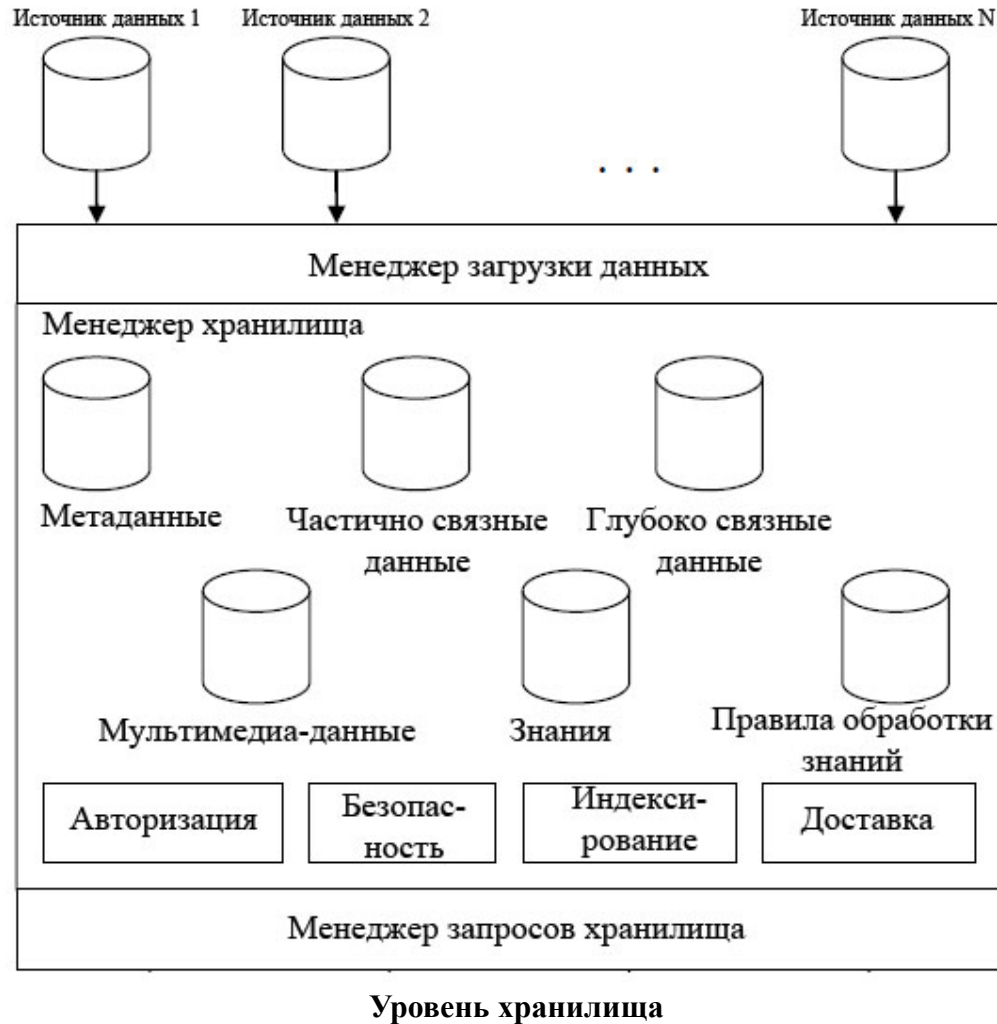
[05:10](#) Scene 02

In porttitor. Donec laoreet nonummy augue. Suspendisse dui purus, scelerisque at, vulputate vitae, pretium mattis, nunc. Mauris eget neque at sem venenatis eleifend. Ut nonummy.

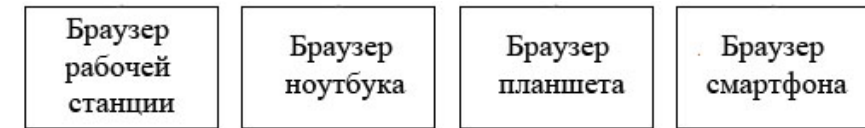
Алгоритм обработки мультимедиа-тегов



Планируемая архитектура хранилища



Уровень базы данных CMS



Стационарные рабочие места

Мобильные рабочие места

Уровень конечного пользователя

Требования к CMS

1. Работа под управлением LAMP-сервера (Linux, Apache, Mysql, PHP)
2. Широкий набор стандартных функций
3. Удобство пользования стандартными функциями
4. API с широкой функциональностью
5. Многофункциональная система управления правами пользователей
6. Встроенные средства управления мультимедиа-содержимым
7. Встроенные средства безопасности, противодействия атакам и вирусам
8. Встроенные средства диагностики компонентов продукта
9. Встроенные средства резервного копирования
10. Встроенные средства масштабирования продукта
11. Доступность документации
12. Возможность обращения в техническую поддержку продукта
13. Возможность обучения пользователей online и создание учебных курсов.
14. Отсутствие критичных уязвимостей в актуальных версиях продукта

Безопасность проекта

Количество уязвимостей общее

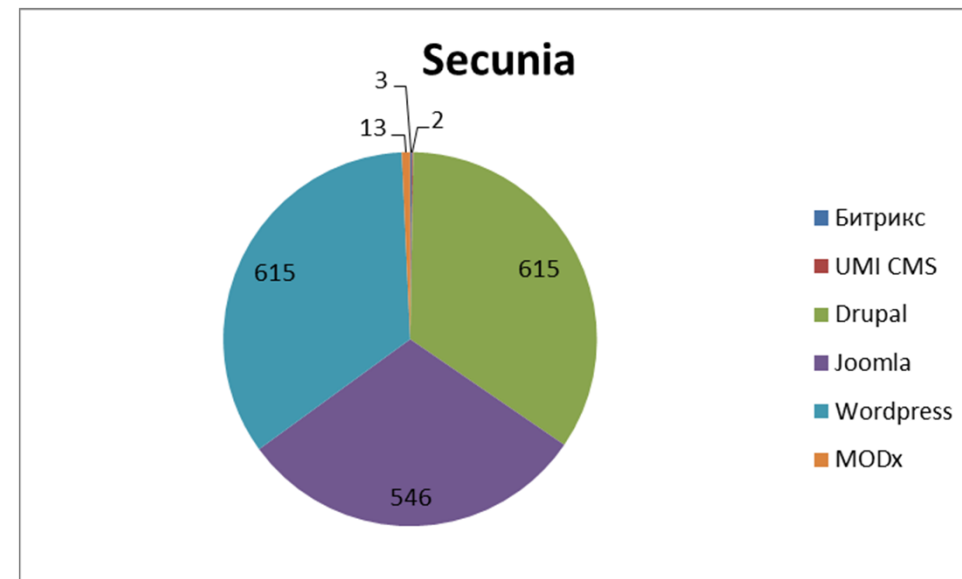
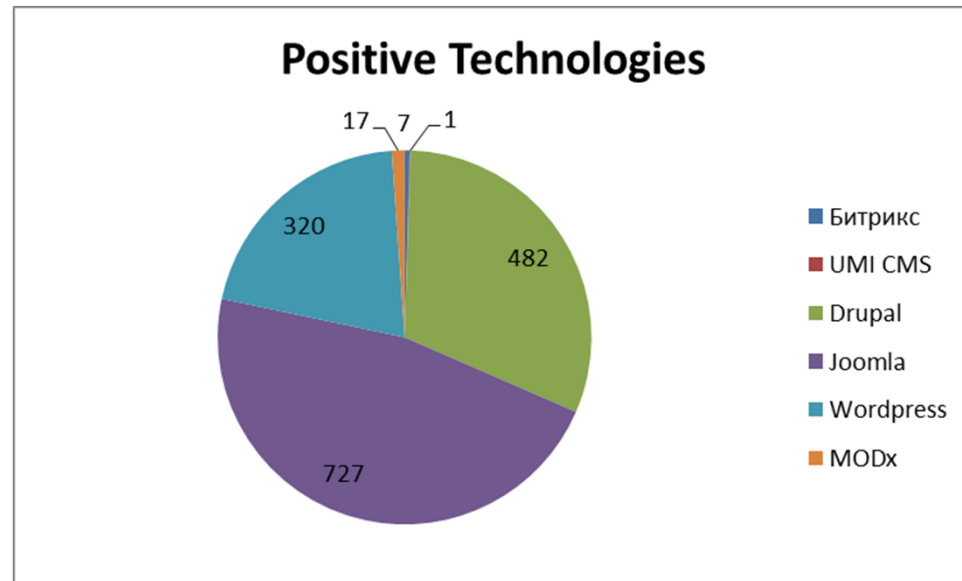
Название	Positive Technologies (Россия)	Secunia (Европа)
Битрикс	7	3
UMI CMS	1	2
Drupal	482	615
Joomla	727	546
Wordpress	320	615
MODx	17	13

Количество уязвимостей по типам

Название	High	Low	Medium
Битрикс	1	4	2
UMI CMS	0	0	1
Drupal	71	180	231
Joomla	424	75	228
Wordpress	92	57	171
MODx	2	1	14

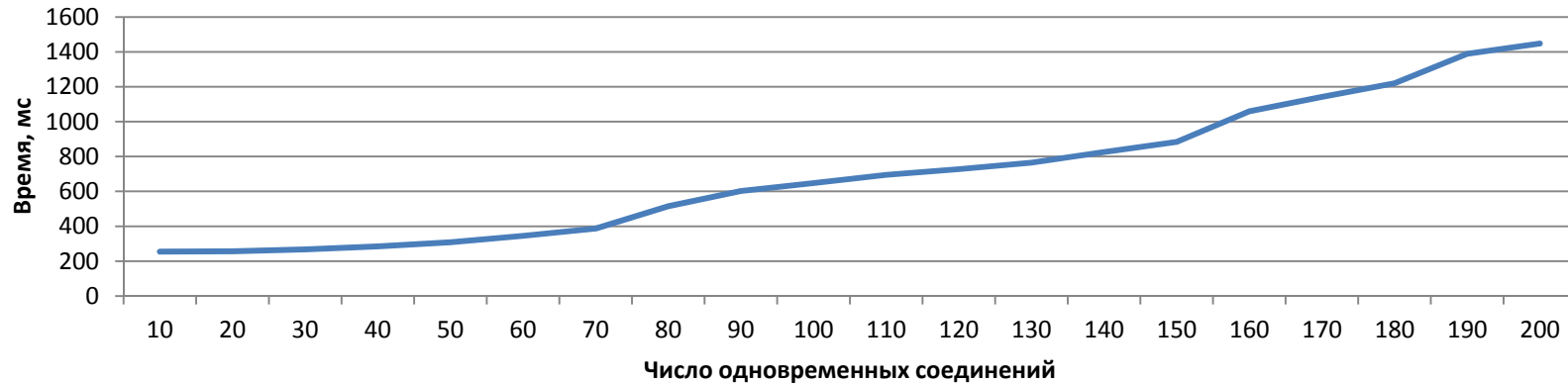
Количество уязвимостей по типам в 2012 г.

Название	High	Low	Medium
Битрикс	0	0	0
UMI CMS	0	0	0
Drupal	4	>20	>20
Joomla	12	4	>20
Wordpress	>20	10	>20
MODx	0	0	1



Производительность проекта

Количество времени, потраченное на выполнение одного запроса

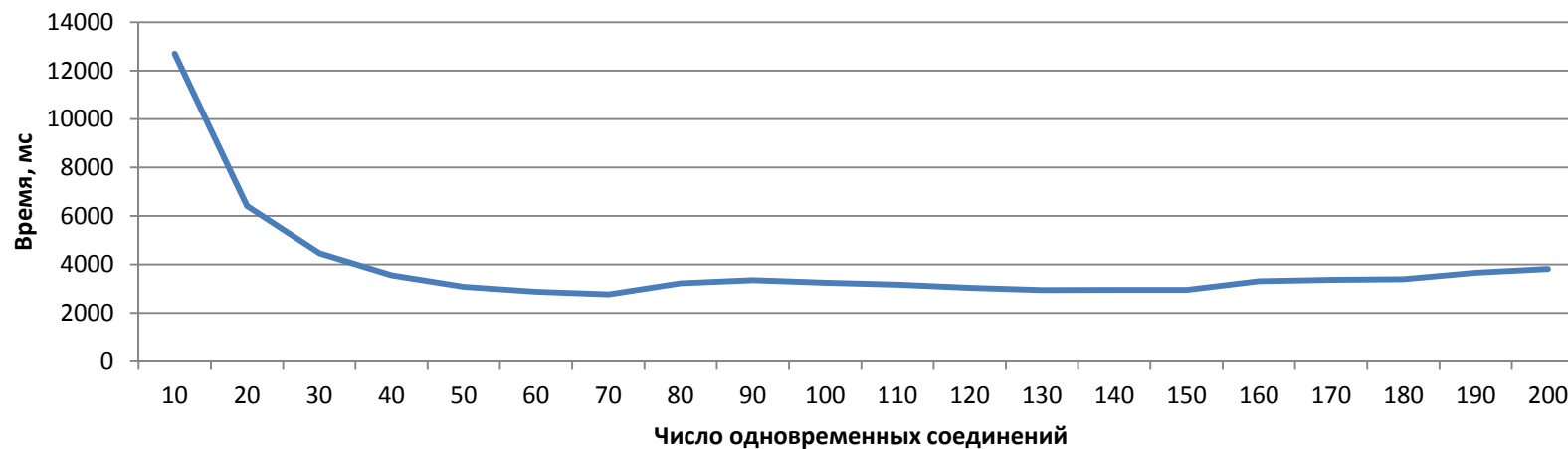


Количество запросов, выполненных в единицу времени



Производительность проекта

Общее время выполнения теста



Оценка производительности проекта Битрикс

Критерий	Результат	Эталон
Среднее время отклика, с	0,0185	0,033
Производительность процессора, млн оп./с, 1 ядро	14,2	9
Производительность жесткого диска и файловой системы, оп/с	7281	10000
MySQL, запись, запрос/с	2381	5600
MySQL, изменение, запрос/с	3250	5800
MySQL, изменение, чтение/с	9120	7800
Общая оценка, производительности, баллы Битрикс	54,15	30

Результаты представленной работы

- Проведено исследование форматов электронных книг и форматов данных, используемых в библиотечном деле;
- Предложены и разработаны модификации формата EPUB для обеспечения поддержки внедренных видео- и аудио-материалов, и возможности сохранения оригинального вида страниц и верстки книг;
- Проведена модификация алгоритма обработки мультимедиа-тегов в тексте электронной книги при ее отображении на устройстве вывода;
- Предложена архитектура цифрового библиотечного хранилища;
- Проведена оценка обеспечения безопасности популярных систем управления контентом на основе которых выбрана система для реализации уровня базы данных CMS хранилища;
- Проведены исследования производительности проекта.

[Слайд 1]

Уважаемые коллеги! Представляю вам свою работу по созданию электронной библиотеки вуза на базе существующей системы управления сайтами Битрикс, которая включает создание системы хранения мультимедийных электронных книг и модификацию формата для хранения электронных книг. В связи с развитием дистанционного образования широко востребовано применения различных новых образовательных технологий, в том числе внутривузовских и межвузовских коллекций литературы в электронном виде, а также различного рода каталогов. Работа по созданию электронной библиотеки вуза поддержана грантом Министерства образования в рамках ФЦП «Научные и научно-педагогические кадры инновационной России».

[Слайд 2]

Рассмотрим цели и задачи, стоящие перед электронно-библиотечной системой вуза. Основные из них перечислены на слайде. Цели и задачи сформулированы, во-первых, из нормативных актов, таких как современный государственный стандарт, соответствующие приказы Рособнадзора, лицензионные требования, а во-вторых, из технических требований Научной библиотеки СибГАУ, в сотрудничестве с которой идет работа над грантом.

[Слайд 3]

Исходя из целей и задач электронно-библиотечной системы, можно сформулировать цель представленной работы, состоящую в выборе и последующей модификации формата хранения мультимедийных электронных книг для обеспечения поддержки внедренных видео- и аудио-материалов и возможности сохранения оригинального вида страниц и верстки книг в рамках стоящих перед форматом задач.

[Слайд 4]

Создание электронной библиотеки предполагает следование определенным отраслевым стандартам. Все стандарты каталогов и метаданных разработаны на основе MARC-форматов (машиночитаемая каталогизация). Разработка таких форматов началась еще в 70 годах прошлого века. В 1993 году был создан формат UNIMARC, как попытка создать наднациональный стандарт. RUSMARC – официальная российская версия формата UNIMARC. Формат предполагает как поля стандартного библиографического описания, так и пользовательские поля.

Z39/50 – протокол обмена информацией в распределенной неоднородной среде с использованием клиент-серверной архитектуры. Данный протокол используется для организации шлюзов между различными библиотечными системами и получения информации о книжном фонде этих сторонних библиотек.

OPAC – стандарт обмена библиотечной информацией ориентированный на Web.

Стандарты метаданных основаны на стандартах электронных каталогов с использованием XML. Предназначены как правило для поисковой выдачи.

Что касается форматов электронных книг, как правило, такие книги специально не готовятся к публикации, и публикуются в том виде, как обрабатываются в типографии. В основном с этим связано такое разнообразие форматов книг. Также существует еще ряд экзотических форматов, разработанных и применяемых вместе с устройствами для чтения. Сами форматы, думаю, не имеет смысла описывать, все с ними работали.

[Слайд 5]

После анализа форматов было принято решение выбрать для дальнейшей работы формат EPUB.

К достоинствам формата относится широкий набор программного обеспечения для просмотра и создания документов, ограниченная поддержка мультимедиа-объектов. В

качестве недостатков можно выделить ограниченную поддержку возможностей для образовательных и научных целей, а также отсутствие возможности сохранения оригинальной верстки бумажного издания книги. Однако преимуществом формата является открытость и расширяемость, поэтому можно предложить и обеспечить поддержку собственных необходимых *XML*-элементов разметки. В связи с этим, формат *EPUB* является наиболее пригодным для создания, хранения и доставки электронных документов в современной научно-образовательной библиотеке. В тоже время в целях совместимости следует обеспечить поддержку других форматов, таких как *PDF* и *DJVU*.

[Слайд 6]

Для реализации возможности сохранения оригинальной верстки книги предлагается дополнение стандарта тегом `<Page>` с атрибутами для описания размеров страницы и ее полей. Теги могут быть вложенными. В таком случае они описывают колоночную структуру документа.

На слайде вы можете видеть пример *XML*-кода, изображение текста книги в браузере и программе для чтения книг без поддержки расширения стандарта, и страницу в браузере с поддержкой предложенных тегов. На данный момент поддержка в браузере реализована с помощью таблиц *CSS* и преобразования *XML* в *DHTML*.

[Слайд 7]

Для поддержки навигации по внедренному мультимедиа-содержимому предлагается дополнение стандарта тегом `<SCENE>` внутри тега `<VIDEO>` с атрибутами для заголовка сцены, времени начала сцены и изображения. Внутри тега может быть текстовое описание сцены.

На слайде вы можете видеть пример встроенного видео с навигацией, заголовками и описанием сцен. Следует отметить, что даже стандартные возможности *HTML5* и соответственно *EPUB* по встраиванию мультимедиа реализованы только в последних версиях браузеров, и не реализованы в программах для создания и чтения книг, таких как *Adobe Digital Editions 2.0*, например.

[Слайд 8]

На слайде представлен алгоритм обработки мультимедиа-тегов, включая предлагаемые дополнения к стандарту.

Для каждого встречаемого на странице мультимедиа-тега, проверяются указанные в атрибутах *SRC* источники контента, выбирается оптимальный источник, и при возможности его воспроизведения отрисовывается элемент управления. Если внутри имеются теги описания сцены, для каждого выводится блок описания сцены из заголовка, времени начала, постера и текстового описания. В случае невозможности воспроизведения мультимедиа-содержимого выводится текстовое содержимое мультимедиа-тега.

[Слайд 9]

На слайде изображена предлагаемая планируемая архитектура цифрового библиотечного хранилища. Архитектура состоит из трех уровней – хранилища, базы данных системы управления контентом и конечных пользователей. Источники данных предоставляют информацию из различных мест, например, интернета, интранета, других цифровых библиотек. Хранилище может содержать различные типы данных. Например, частично связанные данные включают в себя текстовые файлы, глубоко связанные данные используются в сложных запросах, метаданные содержат описания хранимых данных для системы управления контентом *web*-сайта, база знаний и правил их обработки необходима для интеллектуального индексирования и поиска данных. Также в хранилище встроены функции авторизации и управления безопасностью.

База данных CMS имеет собственные функции авторизации, безопасности, работы с веб-приложениями и мобильными приложениями.

Уровень конечных пользователей может быть представлен широким спектром приложений, обеспечивающих необходимые роли.

[Слайд 10]

В рамках подготовки заявки на грант было принято решение о создании мультимедийной библиотечной системы ориентированной на функционирование в среде Internet. Для минимизации работ было принято решение о создании библиотеки на основе существующей системы или фреймворка управления контентом. Таким образом, стандартные функции, такие как управление пользователями, файлами, базой данных будут возложены на CMS. Были подготовлены требования, которые вы можете видеть на слайде. В результате анализа этих требований в качестве системы управления контентом была выбрана 1С-Битрикс. Следует отметить, что все крупные системы обладают примерно похожим функционалом, поэтому критичным для выбора оказался вопрос безопасности проекта.

[Слайд 11]

На слайде представлена статистика по обнаруженным уязвимостям, собранная российской компанией Positive Technologies и европейской Secunia. Представлена общая статистика за время жизни продукта по уязвимостям в виде таблицы и диаграммы по каждой лаборатории отдельно. Также статистика по видам уязвимостей. Так же по видам уязвимостей за последний год. Последние данные только от компании Positive Technologies. Следует отметить, что коммерческие системы наиболее безопасны, очевидно, потому что имеют возможность проводить единую политику создания и поддержки продукта и осуществлять аудит безопасности сторонними экспертами.

[Слайд 12]

На слайде приведены результаты замера производительности сервера и системы с помощью утилиты Apache Bench. Для тестов использовался сервер на базе i7 975, 12 Гб ОЗУ, файловая система представлена тремя 1,5Тб дисками, объединенными в программный RAID-5.

Тесты проводились организацией нескольких одновременных соединений с последующим повышением их количества. Общее количество соединений 2000.

[Слайд 13]

При выполнении тестов проводился замер параметров, указанных на слайдах. Исходя из представленных графиков, можно сделать вывод, что система способна обрабатывать без срывов производительности до 180 запросов в секунду при 70 одновременных соединениях, что дает теоретическую производительность до 15 млн. запросов в сутки.

Также на слайде представлена таблица результатов тестов системы, выполненная встроенным автоматическим мастером диагностики системы Битрикс. Результаты тестов показывают достаточную производительность разработанной системы.

[Слайд 14]

На слайде представлены основные результаты представленной работы