Проектирование комплексных систем поддержки Электронных Изданий

Дмитриев[1] П.А.

[1] Вычислительный центр РАН, Москва

Введение.

Под *Электронным Изданием* (ЭИ) в нынешнем IT сообществе обычно понимается организованная коллекция структурированных полнотекстовых документов, а под *функционированием ЭИ* подразумевается комплекс сервисов, позволяющих работать с ЭИ как с электронным аналогом обычной книги, оснащённой, помимо стандартных для книги возможностей использования, сложными сервисами, такими как поиск фрагментов по лексемам, гиперпереходы между фрагментами (однозначные и многозначные) и т.д.

Под *Системой Поддержки Функционирования ЭИ* понимается комплекс программных и/или аппаратных средств, обеспечивающих

- функционирование коллекции ЭИ;
- комплекс действий по подготовке, хранению и администрированию этой коллекции.

Система Поддержки Функционирования ЭИ состоит из следующих подсистем:

- Подсистема Подготовки Документов
- Подсистема Загрузки Информации
- Подсистема Показа Информации
- Подсистема Администрирования

Функции этих подсистем будут рассмотрены ниже.

Жизненный цикл ЭИ состоит из следующих стадий:

- Стадия подготовки информации.

  На этой стадии происходит:

  1) сканирование и распознавание печатного оригинала;
  2) проверка, редактирование и форматирование текста;
  3) разметка структуры, оформления, гипертекстовых связей;
  4) формирования массивов данных и метаданных.

  Ответственной за выполнение этих работ является Подсистема Подготовки Инфориации.

- Стадия загрузки информации в систему.

  Поддержка этой стадии осуществляется Системой Загрузки Информации, которая производит преобразование входного массива данных и метаданных в некоторый внутренний для системы формат, предназначенный для обеспечения реализации сервисов, предоставляемых Системой Показа Информации .

- Стадия работы пользователя с ЭИ.

Пользователь получает доступ к коллекции ЭИ через интерфейс Подсистемы Показа Информации, обеспечивающей функционирование ЭИ.

Типичная Подсистема Показа Информации предоставляет пользователю следующие сервисы:

- Навигация по структуре документов (иногда поддерживается несколько альтернативных структур);
- Поросмотр текста документа;
- Возможность переходов по гиперссылкам между документами;
- Возможность просмотра дополнительной и специальной информации о документе (метаинформация, сноски);
- Контекстный поиск;
- Атрибутивный поиск;
- Возможность распечатки текста документа.

- Стадия модификации ЭИ – добавление новых документов в систему и/или удаление из системы существующих.

  Реализуется через интерфейс Подсистемы Администрирования.

  Основная задача Подсистемы Администрирования заключается в том, чтобы обеспечить возможность модификации коллекции ЭИ без её полной перегрузки.

Направления развития и примеры систем.

Работа по созданию систем поддержки Электронных изданий в настоящее время ведётся многими группами разработчиков разного масштаба по следующим основным направлениям:

- Разработка портативных электронных устройств eBook (см. [1]). Таковые являются оригинальными устройствами (мини ПК), предназначенными для работы с электронными версиями книг (сейчас в различных кустарных форматах. В настоящее время стандартном для подготовки данных к публикации через eBook'и становится формат OEB (см. [2]). Данное направление активно развивается сейчас на Западе.

- Разработка Издательских систем, поддерживающих как процесс специальной подготовки данных (после получения их от авторов), так и сервисы по их бумажному, электронному тиражированию и поддержки издательской номенклатуры, обеспечивающей возможность профессиональной издательской деятельности. Такие системы имеются в основном за рубежом, и они сделаны под конкретное издательство, т.е. изначально ориентированы на *фиксированный технологический процесс*, а значит, трудноприменимы в рамках всей отрасли. Например, такие системы используются в издательских домах Elsevier (см. [3]) и Springer Verlag (см. [4]).

  Например издательский дом Elsevier предлагает авторам опубликовать свои статьи в из журналов, с которыми он сотрудничает. Автор вместе с бумажной версией предоставляет электронный вариант статьи в одном из предопределённых

форматов (LaTex, Word, …). Формат определяется требованиями конкретного журнала. Электронная версия статьи получается путём несложного преобразования авторских текстов в HTML. Затем ЭИ помещается в соответствующий раздел Web-сайта.

Рассмотрим, каким образом поддерживаются различные стадии жизненного цикла ЭИ в этой системе:

- Стадия подготовки информации.
  Все функции целиком возлагаются на автора статьи.
- Стадия загрузки информации в систему.
  Для каждого входного формата отдельный технологический процесс. В данной системе эта фаза не сопряжена с большими трудностями из-за отсутствия требований реализации сложных сервисов функционирования ЭИ.
- Стадия работы пользователя с ЭИ.
  Пользователю предоставляются следующие сервисы:
  - Навигация по структуре изданий.
  - Просмотр текста статьи.
  - Переходы по гиперссылкам (локальные – в пределах статьи + на содержание номера журнала).
  - Контекстный поиск.
- Стадия модификации ЭИ.
  Реализуется полной перезагрузкой ЭИ.

Как мы видим, в данной системе не реазованы многие сервисы функционирования ЭИ (просмотр специальной информации, атрибутивный поиск, ...). Это объясняется тем, что система изначально предназначена для публикации отдельных статей (несвязанных между собой ЭИ сравнительно небольшого размера), а не крупных многотомных произведений и собраний сочинений. Для такой разновидности коллекции ЭИ отсутствие, например, функции изменения ЭИ без его полной перезагрузки не очень критично. Эта система практически не адаптируема к применению в более широкой области. Например, для добавления сервиса атрибутивного поиска необходимо, в первую очередь, изменение входного формата, что требует пересмотра налаженной схемы взаимодействия с авторами.

К этому же типу систем можно отнести системы формирования и управления содержанием Web-библиотеки. Такие системы предназначены для решения задач преобразования электронной версии документа (файл MS Word) в HTML-формат, публикации его на сайте и управления содержанием этого сайта. В качестве примера можно привести такие крупные Web-библиотеки, как ASM[17], IEEE[19] или более мелкие – УИС “Россия”[18].

УИС “Россия” объединяет более 20 самостоятельных электронных изданий. Система предоставляет следующие сервисы:

- преобразование электронной версии публикации в HTML-формат (разбиение на части с сохранением логической структуры, обработка

оглавления, сносок, примечаний, расстановка ссылок на связанные разделы внутри текста);
- интеграция различных публикаций (единообразное оформление, создание сводного оглавления, поддержка ретроспективы);
- навигация по сайту и возможность сквозного поиска;
- ведение сайта (удаление, изменение, добавление документов, поддержка целостности).

Рассмотрим подробнее работу этой системы:

- Стадия подготовки информации.

  Входными данными для Подсистемы Подготовки Информации являются один или несколько файлов в форматах ASCII, RTF, HTML, Word, Excel, а также графические файлы в форматах GIF и JPG. На выходе имеется набор HTML-файлов, предназначенных для публикации на сайте. Конвертация осуществляется в несколько этапов:

  - сохранение публикации в формате HTML с использованием стандартных средств (например, функция «Сохранить как Web-страницу» в Word или Excel);
  - разметка логической структуры полученного HTML-файла с помощью некоторого набора правил-эвристик. В файле документа расставляются специальные пометки, определяющие гиперссылки, выделяются самостоятельные смысловые блоки (сноски, примечания, таблицы, графики и т.п.), формируются заголовки, как отдельных блоков, так и сборника в целом;
  - проверка и корректировка автоматической разметки вручную при помощи тексового редактора;
  - преобразование размеченного макета в HTML-страницы, предоставляемые конечному пользователю. При этом происходит «разрезание» текста на отдельные компоненты в соответствии с предложенной разметкой и построение связей между ними по заданным правилам. Производится также оформление заголовков и вставка таблиц стилей.

- Стадия загрузки информации в систему.

  Для обеспечения возможности реализации сервисов функционирования ЭИ в системе поддерживается один или несколько файлов карты сервера (в формате XML), которые представляют собой централизованное хранилище всех существующих связей между документами. В этих файлах для каждого документа указываются:

  - разделы оглавлений, к которым относится данный документ;
  - документы, являющиеся следующими и предыдущими для полиграфического и сводного оглавления;
  - ссылки, которые необходимо соотнести этому документу;

- информацию о том, как и где эти ссылки отображать.

- Стадия работы пользователя с ЭИ.

  Пользователю предоставляются следующие сервисы:

  - Навигация по структуре изданий. Поддерживаются несколько альтернативных структур доступа: обычная (ссылки на следующий и предыдущий документы, переход к вышестоящему разделу оглавления) и три сводных тематических оглавления, включающих более 3000 пунктов.
  - Просмотр текста документа.
  - Переходы по гиперссылкам.
  - Контекстный поиск (по данной статье и сквозной по всей коллекции).
  - Возможность дублирования отдельных частей коллекции на CD-ROM.

- Стадия модификации ЭИ.
  Благодаря наличию файлов карты сервера модификация содержания документа и модификация его связей с другими документами могут производиться независимо друг от друга, что очень удобно для добавления/удаления документов.

Как мы видим, данная система представляет собой хорошее решение для публикации в Интернете небольших коллекций документов с не очень сложными связями между ними. В случае больших коллекций со сложной структурой документов неизбежно возникнут проблемы с подготовкой документов (на этапе разметки потребуется большой объём ркчной работы), а также со скоростью работы системы (файлы карты сервера будут гигантских размеров).

- Разработка дистрибутивных (здесь – развёртываемых с «нуля») систем, обеспечивающих полный цикл создания ЭИ (подготовка, создание и электронная публикация) в «настольных» условиях. Таковые нацелены на самые разные слои рынка, а использование открытых стандартов и внедрение передовых технологий позволяет им быть востребованными на протяжении срока, превыщающего срок устаревания среднего IT ноу-хау (1-2 года).

  Наиболее (с нашей точки зрения) удачными примерами таковых являются такие системы как Orthogon Publisher[5], HyperMethod ePublisher [20] и разрабатываемая в ВЦ РАН система «Elysium», являющаяся частью проекта «ИСИР» общей информационной системы РАН (см. [6]).

  Рассмотрим систему ePublisher компании HyperMethod.
  - Стадия подготовки информации.
    Программа ePublisher предназначена для авторов (издательств), желающих создать электронные версии своих книг, справочников, документации, каталогов продукции. Предполагается, что книга уже

написана и свёрстана. Входными данными для программы является вёрстка в одном из форматов PDF либо RTF, графические файлы (BMP, GIF, JPG, WMF, EMF), звуковые файлы (WAV, MIDI, MP3), видео-файлы (AVI, MPEG). Программа автоматически определяет структуру текста (главы-подглавы-параграфы) и предлагает варианты схем публикации и стилей оформления. Имеется около 50 предопределённых шаблонов, кроме того в дистрибутивный комплект входит редактор, позволяющий создавать новые шаблоны.

- Стадия загрузки информации в систему.

  EPublisher автоматически производит разбивку (нарезку) текста в соответствии с выбранной схемой и устанавливает гипертекстовые ссылки(к содержанию, к главам, к словарю). Программа включает компонент, позволяющий определить дополнительные правила для автоматической расстановки гиперссылок.

  EPublisher позволяет сохранить публикацию в одном из трёх видов:

  - При распространении публикации на CD-ROM, дискетах, DVD создаётся exe-файл и соответствующий стандартам дистрибутив.
  - При распространении публикации через Internet создаётся либо запакованный exe-файл для "скачивания", либо набор HTML-страниц для размещения на Web-сайте.
  - При распространении публикации в виде Windows Help создаётся набор hlp-файлов.

- Стадия работы пользователя с ЭИ.

  Пользователю предоставляются следующие сервисы:

  - Навигация по структуре документов.
  - Просмотр(прослушивание) документов.
  - Переходы по гиперссылкам (сгенерированным на предыдущем шаге).
  - Контекстный поиск (по всему тексту или только в заголовках).
  - Импорт документов в PDF или RTF формат.

- Стадия модификации ЭИ.

  Реализуется импортом имеющегося издания в редактор, добавлением/удалением/модификацией документов и экспортом нового ЭИ обратно.

- Разовые, или «штучные» системы, разрабатываемые под **конкретный** набор данных (например, одну книжку), **конкретный** заказ на пользовательский интерфейс и **конкретный «**запрос» информационного рынка и не претендующие на масштабируемость или повторное использование внедрённых в них технологических решений. Такие системы отличаются проработанным интерфейсом и обычно являются частью коммерческой деятельности конкретной компании.

  Спектр таких систем также необычайно широк, что обуславливается их массовостью, но с нашей точки зрения стоит отметить лишь некоторые из них, отличительной чертой которых является использования *структурированного документа* как базового элемента оперирования в системе:

- Например, широкий класс таких систем составляют т.н. *электронные энциклопедии*, которые бывают как ON-LINE'овые (см. «Кирилл и Мефодий» [8]) , так и выпускаемые на CD – Britannica, Microsoft Encarta, Ibm World Book Deluxe и т.д. (см. [9], [10], [11]).

- Другим ярко выраженным представителем являются справочно-правовые системы по нормативным документам , такие как СПС «Гарант» ([12]), «Консультант+» ([13]) или «Кодекс» ([14]).
  Например, СПС «Гарант», например, предоставляет пользователю следующие сервисы:
  - Навигация по структуре документов.
  - Поросмотр текста документа;
  - Возможность переходов по гиперссылкам между документами;
  - Возможность просмотра справки о документе;
  - Контекстный поиск;
  - Атрибутивный поиск (по реквизитам, по ситуации, по источнику опубликования, по словарю терминов);
  - Возможность распечатки текста документа.

- Есть также подкласс «штучных» систем, ориентированных на создание электронных версий конкретных книг, обеспечивающих как привычные по бумажным оригиналам способы использования (работа с содержанием, страничное разбиение, просмотр текста, ...) так и более сложные, но привычные в Internet среде сервисы: гиперссылки, поиски с морфологической нормализацией, индексы терминов и т.д. В наше время множество авторов выпускают такой диск одновременно с твёрдой копией материала. Никакой стандартизации интерфейса или хотя-бы общего прообраза таковой системы не имеется и все разработки в данном направлении являются сугубо оригинальными, что не умаляет их практическую ценность. Обычным средством распространения для них являются CD-R носители с записанной на них оболочкой и данными, которые образуют замкнутую и цельную информационную систему.

  Примерами удачных решений в этой области можно считать те-же [9] или CD-версию "Энциклопедического словаря" Брокгауза и Ефрона, выполненную при содействии проекта "Яndex" фирмы Комптек (см. [15],[16]).

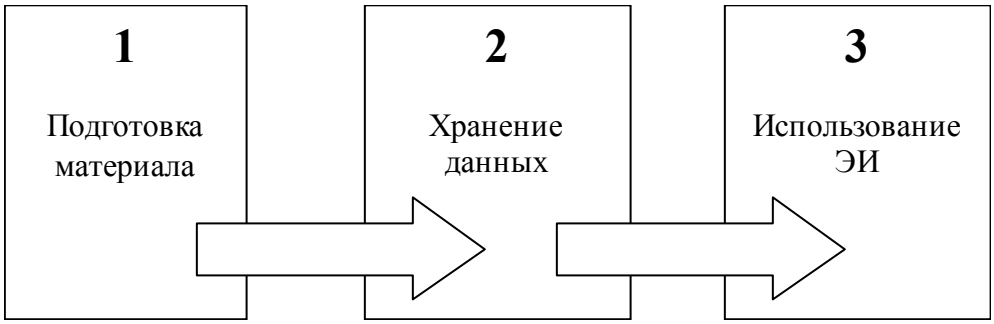
Направление текущей разработки:

В ближайшем будущем будет изложена архитектуру, стандарты и технологии, положенные в основу системы «Elysium», в состав группы разработчиков которой входит автор. Эта система относится к разряду *комплексных дистрибутивных систем,* т.е. предназначена для поддержки всего жизненного цикла существования ЭИ и способна развёртываться в готовое к работе состояние на конкретной платформе из инсталляционного комплекта. Фактически, такая система планируется как Издательская Система для **подготовки** и **изготовления** Электронных Изданий.

Система состоит из:

- подсистемы подготовки и загрузки материалов,
- сервера данных и
- сервера приложений.

В простом (без возможности редактирования данных, а лишь с полной перезагрузкой) жизненном цикле ЭИ эти три компоненты используются последовательно, отражая следующие этапы работы с ЭИ:



Подсистема подготовки и загрузки материалов

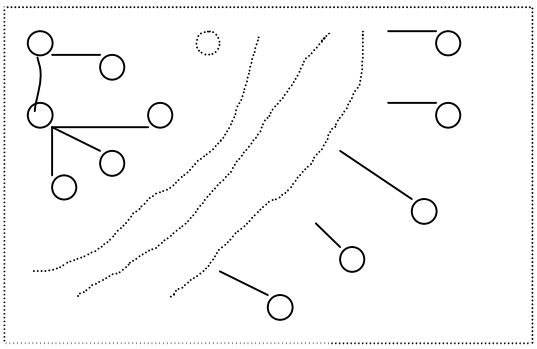
Целью работы такой подсистемы является формирование входного массива документов, содержащих в себе данные ЭИ в формате OEB' (оригинальный DTD), который в отличии от OEB (см. [2]) позволяет хранить структурные элементы, более высокого уровня, чем отдельные Произведения или Издания – т.н. именованные Контейнеры, которые включают в себя Контейнеры, Издания, Произведения как структурные подэлементы.

Процесс поддерживается любым HTML и XML редактором, поскольку все DTD предоставляются явно.

Сервер данных

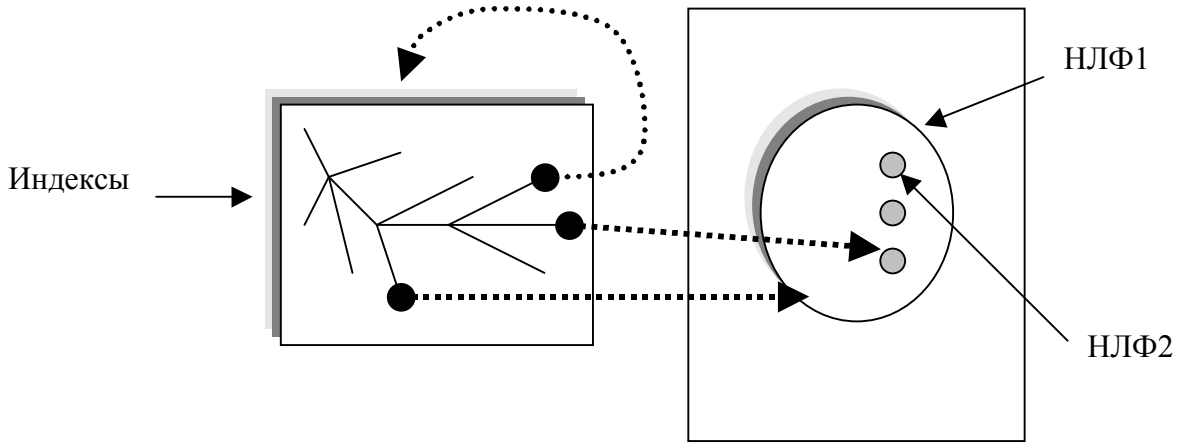
Хотя система имеет дело с несколькими древовидными структурами над одним и тем же массивом фрагментов, одну из них (соответствующую древовидной структуре входного XML файла с учётом специальных меток, сигнализирующих о специальном типе фрагментов данных) можно считать первичной, грузимой вместе с данными; а остальные – вторичными, т.е. генерируемыми искусственно при функционировании пост-загрузчика, работающего на уже созданных элементах на сервере приложений. Хотя, тем не менее, в конечном виде реализация этих иерархий одинакова с точки зрения системы, т.е. реализуется одним и тем же набором классов и методов.

Хотя всю совокупность информационных фрагментов, хранимых репозиторием, логично представлять как одно огромное дерево: от «Литературных корпусов» до «Параграфов»,



репозиторий уже предусматривает разделение этого массива на 2 части:

Хранение первой из них, структуры, возлагается на объекты типа «Индекс», позволяющими оперативно строить дерево потомков N-ого уровня для каждого из НЛФ, а второй – на объекты типа «НЛФ1» ,соответствующие «крупным» фрагментам, и на объекты «НЛФ2», соответствующие остальным.



Смысл деления заключается в том, что «индексы» отвечают за организацию *элементов* с точки зрения структуры входных данных (а, значит, и логики их вывода), а «НЛФ» - за организацию элементов данных в приемлемом для системы виде.

В связи с этим возникают и три вопроса:

1. Что такое «крупный»?
2. Как соотносятся НЛФ1 и НЛФ2?
3. Как организованы индексы?

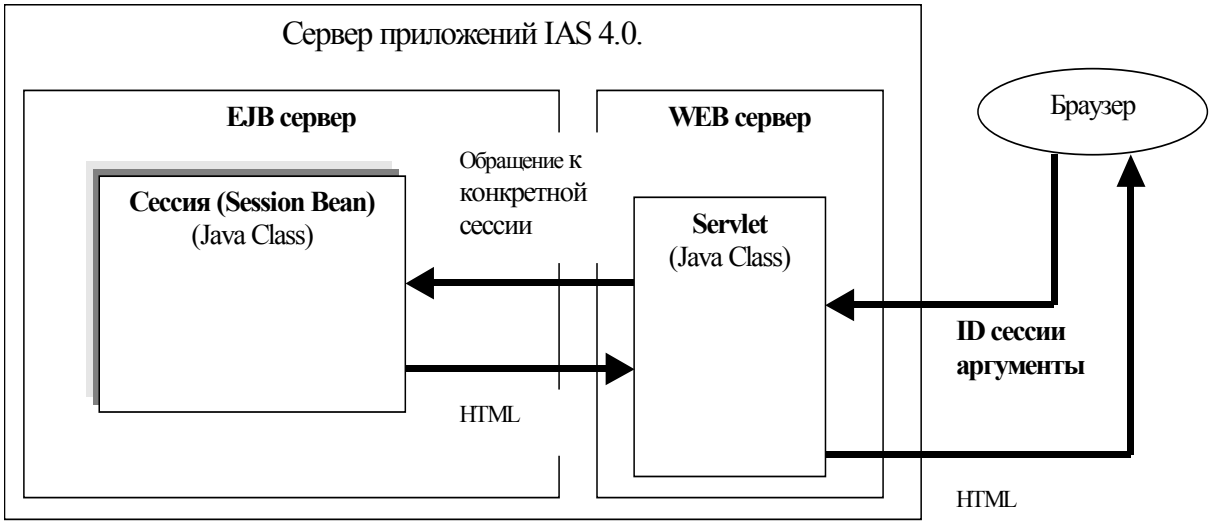
В статье формально решаются все эти три вопроса и приводится объектная модель классов сервера, ориентированная на реализацию на платформе J2EE (см. [28]).

Сервер приложений

Суть функционирования сервера приложений определяется как выполнение следующих функций:

- • получение событий и аргументов форм от браузера (отождествляется с одной пользовательской сессией), содержащих запросы пользователя к ЭИ,
- • обработка полученных событий и аргументов форм с целью преобразования их в последовательность обращений к методам сервера данных («декодирование пользовательского ответа»),
- • формирование блока данных для следующей HTML страницы в браузере («ответ пользователю») соответственно с логикой сборки структурных фрагментов, хранящихся на сервере,
- • фрагментация выводимых пользователю данных для ускорения получения им ответа (загрузки HTML страницы), позволяющей избежать передачи на клиента (через http) блоков данных большого размера, а ограничиться только минимальными визуально-оправданными («страницами»).

Выполнение этих задач возлагается на 3 основных класса , функционирующих на сервере приложений и на Web-сервере:



Серверным образом конкретного браузера является класс «Сессия» (HttpClientSession), который хранит *полный* контекст этого браузера, т.е. другими словами каждому отображённому объекту в браузере соответствует элемент данных в неком объекте «Сессия». В связи с этим единственным обязательным аргументом при обращении к сервлету у браузера является ID сессии. Управление жизненным циклом объектов класса «Сессия» осуществляется сервлетом (или несколькими), которые выполняют следующие задачи:

- • определяют экземпляр сессии, к которой адресовать пришедшие аргументы
- • в зависимости от аргументов, определяют какие методы конкретной сессии нужно вызвать и как сложить полученные от них HTML'и.

Заключение

Изложенная архитектура была реализована в серверной версии системы, на текущий момент работает опытная WEB - версия, состоящая и загрузчика, сервера и сервера приложений. В

ближайшем будущем система должна поддерживать все заявленные в спецификации компоненты, что позволит ей стать завершённой Настольной Издательской Системой для Электронных Изданий.

Литература

[1] eBook News and Community http://www.ebooknet.com/
[2] Open eBook Publication Structure http://www.openebook.org/specification.htm
[3] Elsevier Science Ltd., http://www.elsevier.com/
[4] Springer Science Online Ltd., http://www.springer.de/
[5] Orthogon Publishing System, http://www.xpanion.de/
[6] Интегрированная Система Информационных Ресурсов Российской Академии Наук http://isir.ras.ru
[8] «Кирилл и Мефодий» Co, http://www.megakm.ru/
[9] Britannica Encyclopedia , Inc. http://www.eb.com
[10] Microsoft Encarta Encyclopedia – http://www.encarta.msn.com
[11] Ibm World Book – www.worldbook.com
[12] Компания "Гарант" - http://www.garant.ru
[13] Система “КонсультантПлюс” - http://www.consultantplus.ru
[14] Система “Кодекс” - http://www2.kodeks.net
[15] Яндекс – Поисковая система http://www.yandex.ru/
[16]В.Суховеев (рецензия) *"Энциклопедический словарь" Брокгауза и Ефрона- "Биографии: Россия" на CD-ROM*, http://www.ci.ru/inform6_98/suchov.htm (Компьютер- Информ Inc., 6#98).
[17] Association for Computing Machinery http://www.acm.org
[18] Университетская информационная система «Россия» http://www.cir.ru
[19] The Institute of Electrical and Electronics Engineers, Inc. http://www.ieee.org
[20] HyperMethod ePublisher http://www.hypermethod.com/epub.html
[21] Kalinichenko L.A., Skvortsov N.A., Briukhov D.O., Kravchenko D.V., Chaban I.A. Designing Personalized Digital Libraries "Programming and Computer Software" Vol. 26, N3, 2000, pp. 123-133
[22]Михаил Агеев, Сергей Журавлев, Виктор Ламбурт Подготовка Web-версий традиционных изданий "Открытые Системы" №12, 2000г.
[23]Дж. У. Ольсен Бум электронных книг Журнал «Publish», #08/2000
[24]Джон Свенсон На страже цифровых рубежей Журнал «Publish», #07/2000
[25]С.В.Агошков, К.В. Вигурский, А.Е.Поляков, А.В.Котов, В.А.Серебряков, А.А.Штольберг Технология и программная поддержка создания Электронных Изданий конференция "Электронные библиотеки 2000", Протвино
[26]А.Г.Марчук Система поддержки работы с удаленными XML-документами конференция "Электронные библиотеки 2000", Протвино
[27]Дмитрий Барашев, Екатерина Горшкова, Борис Новиков Оптимизация представления XML документов в реляционной базе данных конференция "Электронные библиотеки 2000", Протвино
[28]Java 2 Enterprise Edition http://java.sun.com/j2ee

ELECTRONIC BOOK SUPPORTING SYSTEM ARCHITECTURE PLANNING

Dmitriev A.P.
Computing Centre of Russian Academy of Sciences
Moscow

This paper describe existed types of Electonic Books (EBooks) and qualify it by different parameters. Also it list the basic directions in industry and scientific researches in the EBook industry like particular questions of Digital Libraries knowledge area. The main conclusion of this paper concern "complex" system of EBook realization like the most effective way of Ebook supporting systems developing. The main attribute of the "complex" system is that it include supporting of all steps of EBook creating (prepare materials, loading, storing and representing) obey to single meta-paradigm.