Учреждение Российской академии наук
Институт вычислительных технологий
Сибирского отделения РАН

О методах сравнения информационных систем, их классификации и специальных сервисах

Жижимов О.Л.

Типы информационных ресурсов

Обилие программного обеспечения для построения электронных библиотек, большое количество инсталляций этого программного обеспечения и наполнение этих инсталляций разнообразным информационным содержанием принуждает рассматривать мир электронных библиотек как множество слабосвязанных сущностей, объединяемых на первый взгляд только неудачным общим названием «электронные библиотеки».

Каждый элемент этого множества имеет как некоторые общие характеристики, совокупность которых позволяет относить его к множеству электронных библиотек, так и уникальные характеристики, позволяющие различать элементы между собой.

Цель настоящей работы – определение набора характеристик, позволяющего идентифицировать любую электронную библиотеку с точностью, достаточной для описания основных ее свойств.

При этом термин «электронная библиотека» здесь будет употребляться вместе с термином «информационная система».

Типы информационных ресурсов

Сегодня термин «электронная библиотека» используется в литературе в различных смыслах. Этим термином могут обозначаться

- Набор специального программного обеспечения, предназначенного для создания информационных систем и хранилищ электронных документов.
- Набор реализованных сервисов на базе определенного программно-аппаратного комплекса – конкретная инсталляция и конфигурация специального программного обеспечения.
- Совокупность информационного контента конкретной информационной системы и сервисов доступа к этому контенту.

Категории параметров

Исходя из этого, набор параметров, обеспечивающий исчерпывающую характеристику ЭБ, можно рассматривать как схему метаданных для описания ЭБ как

- Пакета специального программного обеспечения
- Локальной инсталляции набора специализированных пакетов
- Информационного контента текущей инсталляции

```
<elibInfo>
   <softwareInfo> . . . . </softwareInfo>
   <installInfo> . . . . </installInfo>
   <contentInfo> . . . . </contentInfo>
</elibInfo>
```

Категории параметров

Для формирования детальной структуры **<softwareInfo>** полезны таблицы сравнительных характеристик

1. A Guide to Institutional Repository Software. 3rd Edition // Open Society Institute. – August 2004. – p.28.

2. Dion Hoe-Lian Goh, Alton Chua, Davina Anqi Khoo, Emily Boon-Hui Khoo, Eric Bok-Tong Mak andMaple Wen-Min Ng. A checklist for evaluating open source digital library software..// Online Information Review, Vol. 30 No. 4, 2006. - pp. 360-379.

3. Cyzyk Mark, Choudhury Sayeed. A Survey and Evaluation of Open-Source Electronic Publishing Systems // The Open Society Institute (OSI). - Technical Report. - 28 Apr 2008. - p.70 - [http://jhir.library.jhu.edu/handle/1774.2/32737]

Категории параметров

Для формирования детальной структуры элементов **<installInfo>** и **<contentInfo>** следует обратиться к унифицированным описаниям баз данных и сетевых сервисов, структура которых зафиксирована в стандарте ISO-23950 (ANSI/NISO Z39.50) так называемых категориях сервиса Explain.

Более простой подход может быть основан не на схеме данных Explain, а на схеме данных ZeeRex, (включающей только некоторые элементы из схемы Explain, упрощая тем самым представление описания информационной системы и всех ее компонент.

Категории Z39.50 Explain

targetInfo	Информация о конкретном сервере. Должна существовать одна запись этой категории.
databaseInfo	Информация о каждой базе данных, поддерживаемой конкретным сервером. Количество записей этой категории совпадает с количеством баз данных.
schemaInfo	Информация о каждой схеме данных, поддерживаемой конкретным сервером. Количество записей этой категории совпадает с количеством поддерживаемых схем.
tagSetInfo	Информация о каждом наборе меток, поддерживаемом конкретным сервером. Количество записей этой категории совпадает с количеством поддерживаемых наборов.
recordSyntaxInfo	Информация о каждом формате внешнего представления, поддерживаемом конкретным сервером. Количество записей этой категории совпадает с количеством форматов.
attributeSetInfo	Информация о каждом наборе поисковых атрибутов, поддерживаемом конкретным сервером. Количество записей этой категории совпадает с количеством поддерживаемых наборов.
termListInfo	Списки термов, поддерживаемые для базы данных
extendedServicesInfo	Расширенный сервис
attributeDetails	Информация о каждом поисковом атрибуте. Количество записей этой категории совпадает с количеством баз данных.
termListDetails	Список термов
elementSetDetails	Информация о каждом наборе элементов, поддерживаемом конкретным сервером. Запись этой категории существует для каждого формата внешнего представления и для каждой базы данных.
retrievalRecordDetails	Информация об элементах в каждом формате внешнего представления. Запись этой категории существует для каждого формата внешнего представления, для каждой базы данных и каждой схемы.
sortDetails	Спецификации сортировки для базы данных
processing	Исполняемые инструкции
variants	Определения наборов вариантов: классы, типы и значения для поддерживаемых сервером наборов.
units	Определения единиц, поддерживаемых сервером

Категория <installInfo>

Структура **<installInfo>** должна включать следующие информационные блоки

- Общее описание сервера
- Описание сервисов
- Описание интерфейсов доступа к сервисам
- Фиксация правообладателя
- Фиксация административной контактной информации

Видно, что в схеме описания любой информационной системы, в том числе и в описании электронной библиотеки, должны присутствовать элементы, содержащие информацию, аналогичную информации из категории `targetInfo` схемы данных Explain. При этом некоторые элементы `targetInfo` должны быть включены в структуру `<installInfo>`, описывающую конфигурацию конкретной информационной системы.

Категория <installInfo>

```
<installInfo>
    <name> . . . </name>
    <description> . . . </description>
    <accesInfo>
        <host> . . . </host>
        <port> . . . </port>
        <authentication> . . .
</authentication>
    </accesInfo>
    <administrator>
        <name> . . . </name>
        <mail> . . . </mail>
        <phone> . . . </phone>
    </administrator>
    . . . .
 </installInfo>
```

Категория <contentInfo>

Структура <contentInfo> должна включать

- Описание информационных ресурсов
- Описание интерфейсов доступа к ресурсам

При этом информационными ресурсами могут являться

- Базы данных
- Коллекции
- Отдельные документы

Категория <contentInfo>

Фактическое информационное наполнение ЭБ характеризуется параметрами, которые не относятся к ее техническим характеристикам. Ниже перечислены группы наиболее значимых параметров ЭБ, характеризующих ее информационное наполнение.

- Количество различных коллекций и баз данных
- Количество элементов или записей в каждой коллекции и базе данных
- Характеристики, в том числе и тематическая направленность каждой коллекции и базы данных, в том числе описание ограничений на доступ.
- Описание связей между коллекциями, базами данных и записями
- Описание словарей, справочников, рубрикаторов и тезаурусов, используемых при формировании информационного контента ЭБ
- Описание возможностей поиска информации
- Описание возможностей навигации
- Описание возможностей извлечения и просмотра информации

Категория <softwareInfo>

Как и любые информационные системы, электронные библиотеки могут быть классифицированы по характеристикам, которые можно отнести к следующим группам

1. Типы хранимых ресурсов
2. Типы обрабатываемых метаданных
3. Возможности поиска информации
4. Интерфейсы доступа к информации
5. Интерфейсы ввода информации
6. Системная архитектура
7. Аппаратная среде исполнения
8. Программная среде исполнения
9. Способ хранения и обработки данных
10. Ограничения доступа к ресурсам
11. Обеспечение интероперабельности
12. Мониторинг, сбор статистики и обработка статистики
13. Реализация, лицензирование и поддержка
14. Используемые стандартам и рекомендациям (профиль)

Хранимые ресурсы

Одна из основных характеристик ЭБ – перечень типов поддерживаемых электронных документов и файлов. При этом эти типы должны определять

1. Смысловой тип документа – монография, статья, презентация, изображение, видеозапись и пр.
2. Физический формат представления документа – документы PDF, MS Word, MS PowerPoint, MS Excel, TeX, текстовые документы, файлы в графических растровых или векторных форматах, файлы в ауди- и видеоформатах, и т.п.
3. Поддерживаемые типы текстового кодирования документов (языки и кодовые таблицы)
4. Поддерживаемые типы сжатия документов, работа с архивами (zip, gz, rar и пр.) «на лету».
5. Максимальный размер обрабатываемых документов с привязкой к формату представления.
6. Возможность распознавания внутренней структуры документов.
7. Возможность преобразования документов из одного физического формата в другой (pdf – txt, doc - txt, doc - rtf, jpg – txt, jpg – gif и т.п.) по требованию пользователя.
8. Максимальное количество документов, сопоставляемых с единицей учета и хранения в ЭБ.
9. Максимальное количество документов, обработка которого допустима в ЭБ.
10. Максимальный суммарный размер всех документов, т.е. максимально допустимый размер хранилища ЭБ.
11. Возможность определения множественных коллекций в рамках одной инсталляции системы

Зачем это нужно?

Описание **<softwarInfo>** как паспорт пакета программного обеспечения было бы очень полезно администраторам информационных систем при планировании развертывания той или иной системы.

Сравнивая между собой эти формальные описания можно было бы получать объективную сравнительную оценку соответствующих систем и, исходя из этого, принимать решение о развертывании.

Отсутствие формализованных единообразных описаний существенно затрудняет этот непростой процесс и заставляет администраторов самостоятельно строить соответствующие таблицы сравнения на основе программной документации.

Описания **<installInfo>** и **<contentInfo>** были бы полезны при сравнении инсталлированных и развернутых информационных систем.

Зачем это нужно?

Эволюция мировой информационной инфраструктуры имеет тенденцию к интеграции разрозненных информационных систем в нечто единое, но распределенное (DATA-GRID - ?).

Заставить совокупность отдельных информационных систем функционировать как нечто связанное можно лишь основе их полной интероперабельности.

Эта интероперабельность должна кроме всего прочего (стандарты, протоколы, запросы, схемы, форматы и т.п.) включать возможность взаимного информирования систем о своих функциональных возможностях и о своем информационном наполнении.

Адаптивность информационных систем

Без этого информирования невозможно обеспечить свойство **адаптивности** информационной системы при интеграции ее в какой-либо «DATA-GRID».

Информация, представленная в структурах **<installInfo>** необходима для обеспечения **функциональной адаптивности** информационных систем.

Информация, представленная в структурах **<contentInfo>** необходима для обеспечения **информационной адаптивности**.

Зачем это нужно?

Справедливости ради следует заметить сдедующее:

На обеспечение функциональной адаптивности информационных систем в части WEB-сервисов направлена технология на основе WSDL

На обеспечение функциональной в части описания программных интерфейсов направлены технологии, основанные на IDL.

На обеспечение информационной адаптивности информационных систем ориентированы технологии на основе XML, RDF, OWL. Они решают частные задачи обеспечения адаптивности для специальных (на основе XML) систем.

Задача обеспечения адаптивности информационных систем, несомненно, намного шире.

Специальные информационные системы

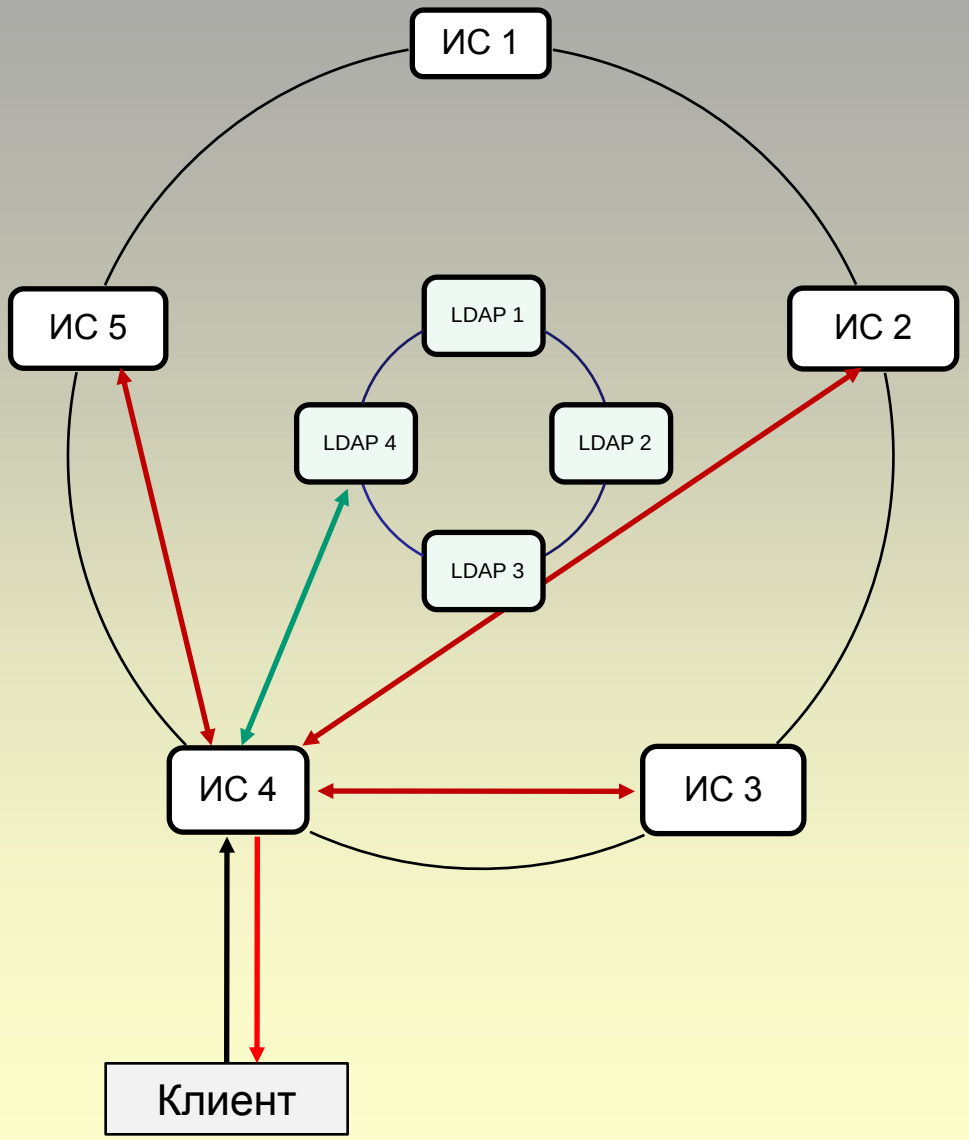
Что касается специальных информационных систем, то существуют упоминавшиеся выше решения для систем на основе Z39.50 (CIP) с использованием специальных стандартизованных системных сервисов (Explain).

Через сервис Explain системы на основе Z39.50 могут обмениваться информацией, составляющей в нашем случае содержание структур **<installInfo>** и **<contentInfo>**.

Более того, минимальная информация о функциональных возможностях любого сервера Z39.50 присутствует в APDU в виде битовых строк при инициализации любого сеанса Z39.50, что обеспечивает минимальный уровень адаптивности для субъектов сетевого взаимодействия в Z39.50.

Идеология систем на основе Z39.50 была частично сохранена при попытке реализовать функциональность Z39.50 в WEB-системах в виде технологий XML/SOAP/SRW (см. ZeeRex).

Один из сценариев работы распределенной информационной системы



1. Получение ИС запроса от клиента
2. Запрос ИС к LDAP о конфигурации и маршруте исполнения запроса
3. Получение списка серверов и их весов
4. Параллельное исполнение запроса в выбранных
5. Получение ИС ответа от других ИС
6. Формирование ответа и возврат его клиенту

Преимущества

Приведенная модель распределенной информационной системы позволяет

- Организовать гибкую распределенную информационную систему с простым добавлением новых узлов и удалением существующих
- Интегрировать разнородные информационные системы
- Организовать распределенное хранилище конфигурационной информации с автоматической репликацией данных
- Организовать единую систему идентификации и аутентификации пользователей
- Организовать прозрачную распределенную обработку поисковых запросов
- Организовать систему мониторинга доступности узлов и ресурсов
- Организовать сбор статистики работы всей системы и ее отдельных частей
- И другое …

О методах сравнения информационных систем, их классификации и специальных сервисах

Жижимов О.Л.

Благодарю за внимание!