# Электронная библиотека по научному наследию как фактографическая система

В.Б.Барахнин, А.М.Федотов, О.А.Федотова

*Институт вычислительных технологий СО РАН,*
*Государственная публичная научно-техническая библиотека СО РАН,*
*Новосибирский государственный университет*

# Особенности информационных систем по научному наследию

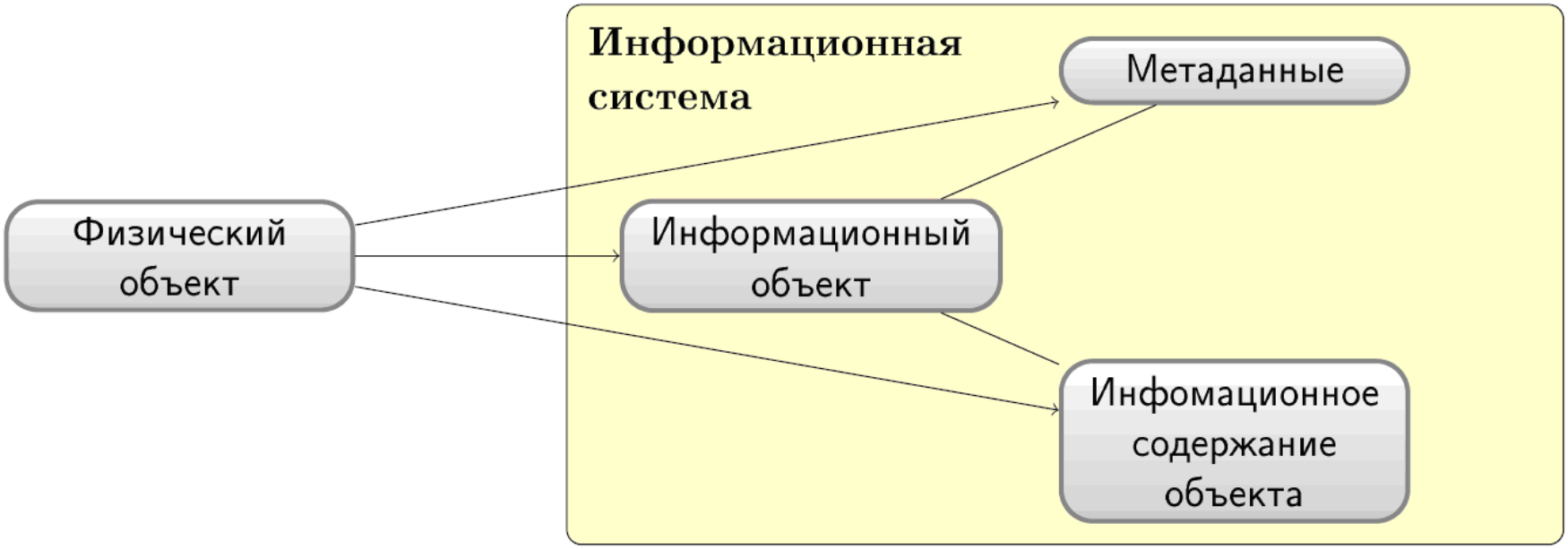
Научное наследие — это опубликованные результаты научных исследований и экспериментов, библиографические и фактографические базы данных, сведения об ученых, их научной деятельности, публикациях, проектах и т.п., а также большое количество неопубликованных документов, таких как отчеты, письма, воспоминания, записки, фотоматериалы и т. п. Эти ресурсы представляют большой интерес для научного сообщества и представителей общественности.

Для информационных систем (ИС) по научному наследию важной проблемой является идентификация ресурсов, определяющая конкретно для каждого факта, кто имеет к нему отношение (например, является его автором), где и когда он имел место, с какими другими фактами и объектами он связан. Для этого необходима поддержка различных уровней абстракции при создании метаописаний информационных объектов: от кратких описаний до очень подробных. Далее будем использовать следующее понимание факта: **“входящая в текст документа характеристика сущности, описываемой в онтологии информационной системы, представляемая как единичное значение данных”**. Факт может быть извлечен из информационного содержания объекта, либо определен экспертом. Факт может определять как свойства (атрибуты) объекта, так и его связь с другими объектами.

Ввиду того, что информация в ИС отображает некоторые сущности (предметы, процессы, явления, персоны, публикации, факты, ключевые термины и т. п.), следует рассматривать информационную систему как множество информационных объектов наборов данных, представляющих (описывающих) эти сущности в ИС.

# Структура информационной системы

Эффективным средством описания информационных объектов в ИС являются **метаданные** — данные, являющиеся неотъемлемой частью информационного объекта и




Для поддержки сложных функций поиска и классификации информации недостаточно хранить только полнотекстовые описания. Необходимы поддержка поиска по атрибутам, полнотекстового поиска, а также просмотр ресурсов по категориям и словарям-классификаторам. При этом выбор классификаторов определяется степенью специализации системы.

# Постановка проблемы

В большинстве существующих ИС документы являются слабо структурированными: хотя и снабженными метаданными, но содержащими неструктурированные элементы. Поэтому актуальной задачей является разработка теоретических основ и моделей создания ИС, способных в автоматизированном режиме извлекать метаданные и факты из электронных документов достаточно произвольной структуры.

В настоящий момент значительная часть информационных ресурсов по научному наследию хотя и переведена в цифровую форму, но недоступна широкому кругу научной общественности, а ресурсы, представленные в Интернет, разрознены, недостаточно систематизированы и структурированы. При создании их описаний недостаточное внимание уделяется вопросам *интероперабельности*: слабо применяются соглашения и рекомендации по стандартизации представления документов и средства интеграции разнородных информационных ресурсов. Под интероперабельностью ИС понимается степень ее способности взаимодействовать с другими ИС, в том числе и с человеком. Но если при взаимодействии с человеком (как с информационной системой) основная нагрузка на обеспечение взаимопонимания ложится на человека, который в состоянии обработать даже плохо организованную информацию, то для обеспечения эффективного взаимодействия между собственно информационными системами требуются специальные технологические методы и общие соглашения. Это приводит к требованию соответствия всех схем данных, интерфейсов и протоколов соответствующим международным стандартам и рекомендациям.

# Электронная библиотека как особый тип информационной системы

**Электронная библиотека (ЭБ)** — это структурированная каталогизированная коллекция разнородных электронных документов, снабженная средствами навигации и поиска (в отличие от печатных изданий, микрофильмов и других носителей). ЭБ способна не только обеспечить многосторонний поиск в каталоге, но и предоставить пользователю непосредственно найденный ресурс (публикацию, документ, фотографию, описание факта и др.), а также дополнительные сведения о нем, например, информацию об авторах, библиографию, организации и т. п.

**Основные цели, стоящие перед ЭБ:**

- обеспечение доступа к информации;
- сохранение научного и культурного наследия;
- повышение эффективности научных исследований и обучения.

В существующих разработках ЭБ, как правило, поиск и доступ к информации обеспечиваются только посредством визуальных графических интерфейсов. Это хорошо для пользователя-человека, но не годится для пользователя-системы. Для обеспечения функций поиска вне графических интерфейсов требуется поддержка специальных сетевых сервисов и языков запросов. В идеальном случае все ИС должны поддерживать единый поисковый профиль и единый язык запросов.

# Требования к профилю ЭБ

**Профиль ЭБ** — набор из одного или нескольких базовых нормативно-технических документов (стандартов и спецификаций), ориентированных на решение определенной задачи (реализацию заданной функции либо группы функций приложения или среды) с указанием, при необходимости, выбранных классов, подмножеств, опций базовых стандартов, которые являются необходимыми для выполнения конкретной функции. Наиболее важным является профиль метаданных информации, циркулирующей в системе.

**Профиль ЭБ должен:**

- включать в себя основные типы информации, требующейся для поддержки научной работы;
- быть открытым, т. е. обеспечивать доступ к соответствующей информации по этим описаниям;
- быть расширяемым, т. е. обеспечивать возможность детализации описаний;
- обеспечивать возможности интеграции информации;
- обеспечивать возможности уникальной идентификации информации;
- обеспечивать возможности размещения и поиска информации в распределенной среде;
- быть ориентированным на современные и перспективные технологии описания и использования информации;
- обеспечивать возможность интероперабельности с внешней средой.

# Функциональные требования к ЭБ по научному наследию

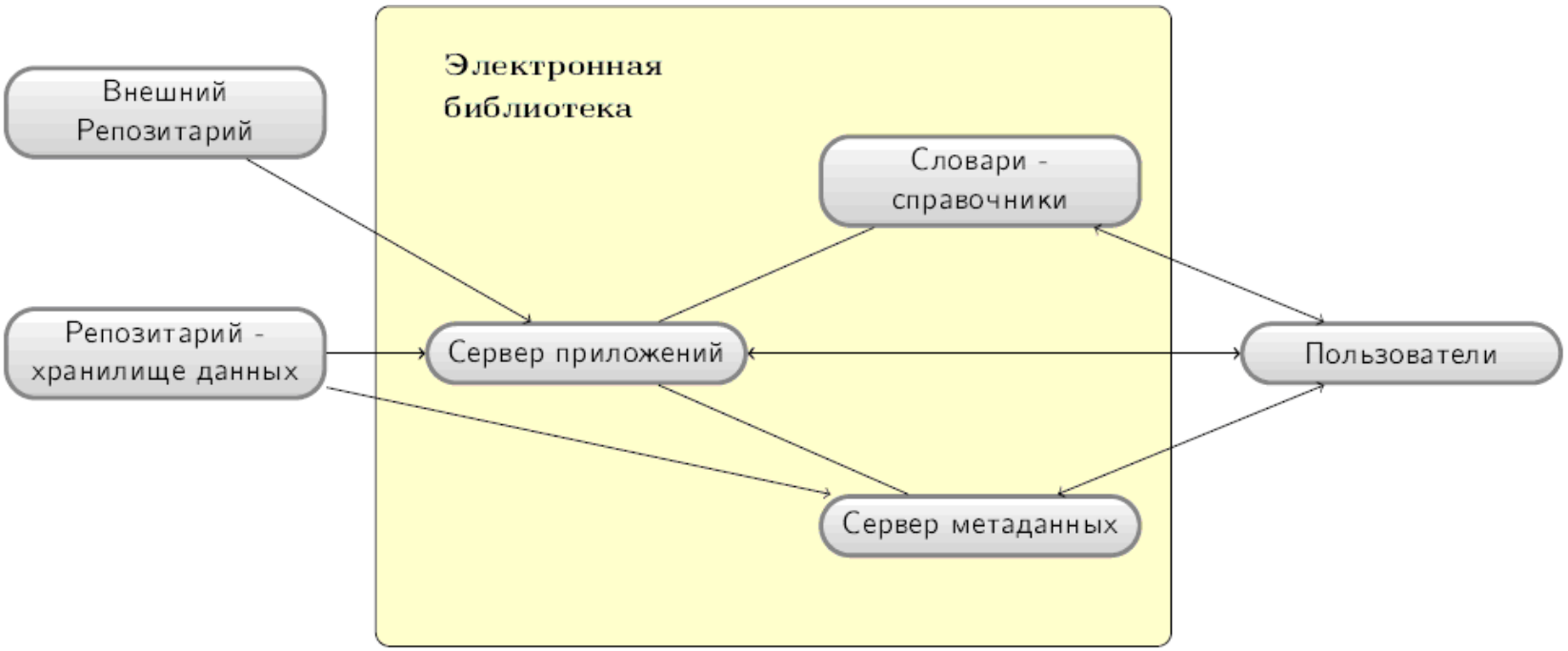
Для документов по научному наследию важной проблемой является идентификация информационных ресурсов, определяющая конкретно для каждого факта, кто является его автором, где и когда он получен, с какими другими фактами он связан. Для этого необходима поддержка различных уровней абстракции при описании информации от кратких описаний, до очень подробных описаний информационных объектов.

Исходя из целей ЭБ по научному наследию и анализа существующих систем, направленных на поддержку научных исследований, можно сформулировать следующие **функциональные требования к модели ЭБ по научному наследию:**

- надежное долговременное и защищенное от исчезновения хранение информации;
- актуальность, полнота, достоверность происхождения документов;
- историчность информации;
- географическая привязка информации;
- наличие большого числа словарей-классификаторов (справочников) для обеспечения идентификации и классификации ресурсов;
- поддержка неоднородных и слабоструктурированных информационных ресурсов;
- поддержка взаимосвязей информационных ресурсов;
- предоставление информации пользователю в виде, выбранном пользователем;
- наличие интеллектуальных служб обслуживания запросов пользователя;
- наличие программных интерфейсов для поддержки аналитической работы пользователя с помощью программных приложений;
- поддержка требований интероперабельности как на программном, так и на семантическом уровне;
- поддержка работы с внешними источниками.

# Архитектура электронной библиотеки

Наиболее важным выводом из вышесказанного является то, что информационная модель ЭБ должна быть многоуровневой и состоять как минимум из следующих компонент: хранилище данных — репозиторий, сервер метаданных, сервер приложений, словари-справочники.



**Метаданные необходимы для решения следующих задач:**

- предоставление сведений об объекте для получения представления о его содержании, структуре, способах использования и т. д.;
- сбор и систематизация информации об объектах описания;
- выбор из множества объектов определенного подмножества по формальным признакам и сопоставление объектов по формальным признакам;
- внутрисистемные технологические задачи, связанные с обеспечением подготовки объектов, размещением объектов в информационном фонде и т. п.;
- внешние технологические задачи, связанные, прежде всего, с обменом данными с внешними информационными системами.

Таким образом, метаданные целесообразно рассматривать как особого рода информационные ресурсы, выполняющие в использующих их ИС весьма разнообразные функции.

# Структура ЭБ по научному наследию

**Основу содержания ЭБ по научному наследию составляют информационные объекты, которые представляют следующие основные типы сущностей:**

- **субъекты:** персоны, организации и т. п.;
- **объекты — единицы хранения:** публикация, документ, факт, научный результат, мероприятие, фотография и др.;
- **отношения:** понятие, ключевой термин, событие, время, место.

В отличие от общепринятых документных (библиографических) ЭБ указание на субъекты дается ссылкой на экземпляр сущности субъект, что позволяет корректно решать задачу идентификации объектов. Схема отношений в ЭБ по научному наследию является персоноцентричной: все объекты и отношения, понятия, факты, мероприятия, публикации и др. жестко привязываются к персонам.

Используемый профиль определяет список элементов данных (полей), необходимых для создания записи соответствующего типа и раскрывает содержание элементов данных. Для эффективной работы сервера приложений необходимо использовать набор словарей-классификаторов, содержащих как классификационные признаки, так и наборы ключевых терминов (с отношениями порядка), по которым производится систематизация и классификация материала.

Для формирования метаданных применяются несколько стандартов, являющихся расширениями рекомендаций Dublin Core и Qualified Dublin Core (QDC). Для документов нами была расширена стандартная схема метаданных QDC полями, включающими основные требования государственного стандарта МЕКОФ.

# Словари в ЭБ по научному наследию

**Словари (ключевые признаки, ключевые термины)** — особый вид метаданных, которые отражают наиболее существенные свойства объекта, имеющие наибольшее значение с точки зрения ИС, и их специфика определятся терминологией конкретной предметной области, которой посвящена ЭБ. Необходимо рассматривать различные типы ключевых терминов, а именно:

- ключевые термины в стандартном понимании;
- ключевые термины, описывающие персону;
- ключевые термины, описывающие временные периоды;
- ключевые термины, описывающие географические понятия;
- тематические словари-классификаторы, тезаурусы, описания предметной области научной школы, классификаторы документов по стандарту МЕКОФ.

Метаданные существенным образом зависят от природы и структуры объектов реального мира, от способа представления их в виде информационных объектов и от специфики ИС. Учитывая это, необходимо классифицировать описываемые объекты. Законченная совокупность правил, достаточная для формирования метаданных в определенном классе ИС и (или) для решения определенного класса задач над информационными объектами представляет собой систему метаданных.

Функционирование ИС связано с разнообразными процессами по созданию метаданных, их модификации, проверке корректности, предоставлению метаданных пользователю и решению прикладных задач. Все эти процессы являются взаимосвязанными, их выполнение усложняется большим количеством объектов, на представление и работу с которыми направлена ИС. Реализация этих процессов и управление ими требуют специальных средств и методов, которые в совокупности с метаданными рассматриваются как отдельная подсистема — **система метаинформационного сопровождения**.

# Практическая реализация

Рассмотренная модель информационной системы, работающей с материалами научного наследия, реализуется на примере научной школы Алексея Андреевича Ляпунова основателя теоретического программирования и отечественной кибернетики.

Основной каталог информационных ресурсов сервера метаданных информационной системы строится в соответствии со схемой метаданных МЕКОФ. Для долговременного хранения документов использовался репозиторий DSpace . Стандартная схема метаданных DSpace была расширена полями, отвечающими основным требованиям МЕКОФ. Для поддержки процесса наполнения полнотекстовых баз созданные профили метаданных были зарегистрированы в системе DSpace и в соответствии с ними были настроены рабочие процессы, а также пользовательский интерфейс системы. Для того чтобы выполнять обмен метаданными между DSpace в соответствии с расширенным профилем, был создан сервис, выполняющий преобразование схем метаданных из внутренней схемы DSpace в схему сервера метаданных и в схему Dublin Core с использованием квалификаторов. Реализован также OAI-сервис, который в пакетном режиме периодически, в соответствии с расписанием, проводит синхронизацию метаданных репозитория и сервера метаданных. Для заполнения основного каталога метаданных в соответствии с созданными схемами метаданных используется контролируемые словари из справочного блока сопровождения.

Разработанная модель информационной системы может быть использована как типовая модель системы для работы с документами, связанных с научным наследием, поскольку решает основные задачи, предъявляемые к этим системам: обеспечение системы надежного долговременного хранения цифровых (электронных) документов с сохранением всех смысловых и функциональных характеристик исходных документов; обеспечение “прозрачного” поиска и доступа пользователей к документам, как для ознакомления, так и для анализа содержащихся в них фактов; организация сбора информации по удаленным ЦД, поддерживающих протокол OAI.

Спасибо за внимание!