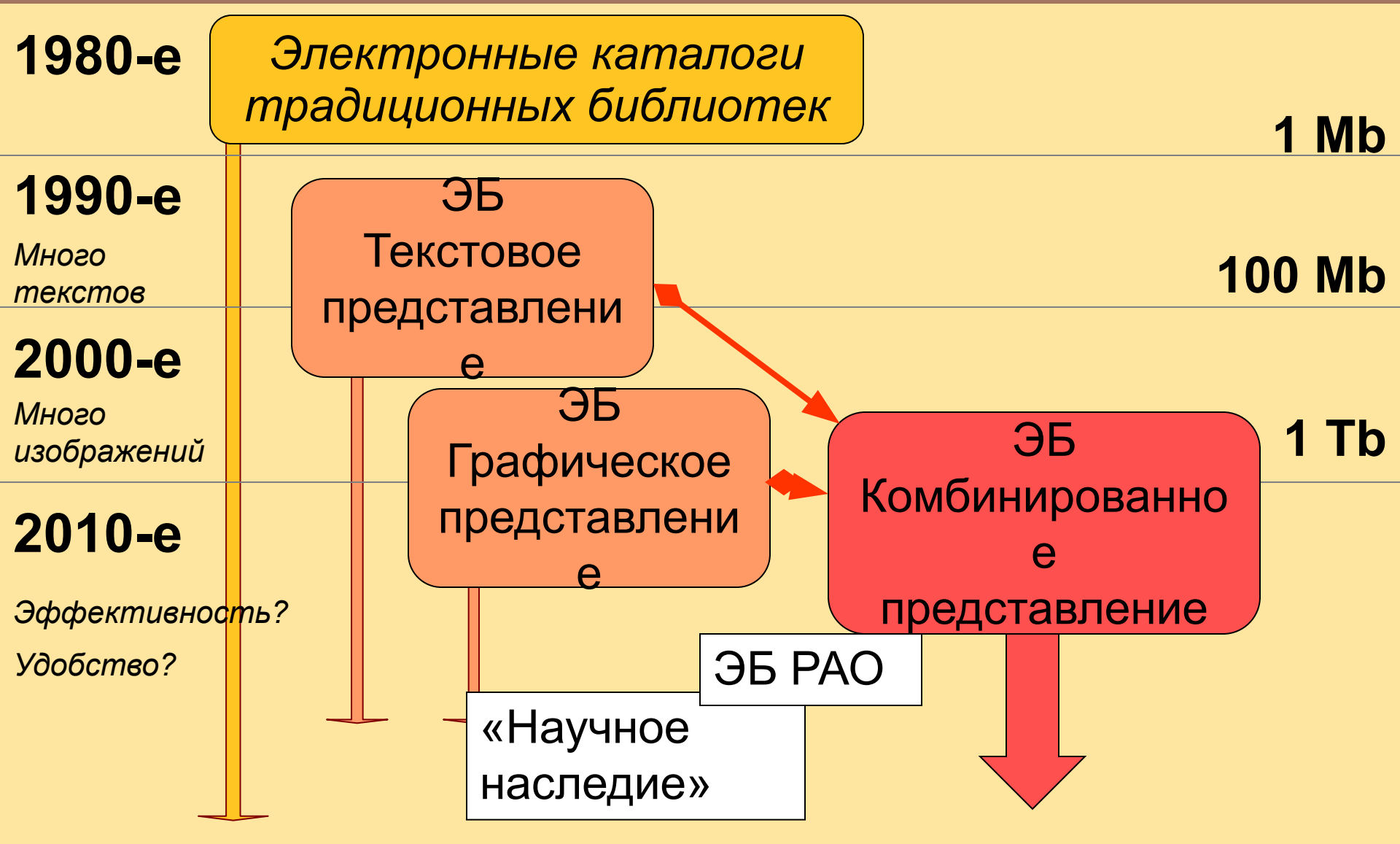
Комбинированное электронное представление печатных изданий

С.И. Трифонов

Лаборатория Разработки и Внедрения
Информационных Технологий
Научная педагогическая Библиотека
им. К.Д. Ушинского РАО

RCDL-2011

Электронные библиотеки: эволюция понятия



Технологии для электронной библиотеки

Подготовка изданий для электронной библиотеки требует разработки **технологии**, поддерживающей:

- Большие объёмы работы, большую долю ручного квалифицированного труда
- Единые требования к качеству подготовки электронного вида, по возможности гибкие
- Средства интеграции изданий в библиотеку - систему, с единой навигацией и поддержкой возможностей информационного поиска

Комбинированное представление — одно из ключевых решений, которые должна поддерживать любая *перспективная* технология

Подготовка электронного издания: операции

Форматы представления изданий

- Текстовое представление: **XHTML**

  единый документ на всё издание
- Графическое представление: **JPG**

  изображение каждой страницы
- Комбинированное представление:

  XHTML + JPG + <*графический индекс*>

Графический индекс имеет два вида:

- текстовый — понятный редактору, *постраничный*
- бинарный — оптимизированный под задачи сервера библиотеки, *единый*

Графический индекс

Вопросы, на которые предстоит ответить:

- Какие задачи решает графический индекс?
- Какую информацию он содержит?
- Зачем нужны два вида (текстовый и бинарный)?
- Как устроен формат индекса?
- Почему не PDF или DjVu?

Основная задача — показать результаты поиска

Задача: выделять найденные слова на графическом изображении - *подсветки*

арисовка, которою начат этот раздел, — один из очень не частых примеров выступления уровня А в роли ведущего уровня. В далеко преобладающем числе движений он уступает ведущее положение более молодым своим собратьям, но не стушевывается окончательно. Напротив, вряд ли найдутся такие движения, в которых не лежала бы в самом их основании работа этого «фона всех фонов». То, что она не бросается сразу в глаза, вполне вяжется

139

Запрос: «движение преобладает»

Содержание графического индекса

Для каждого слова текстового представления графический индекс должен хранить соответствие, где оно находится:

на какой странице и в каком прямоугольнике

Дополнения:

- Прямоугольников может быть несколько
- Координаты могут быть приблизительными
- Могут быть переносы между страницами

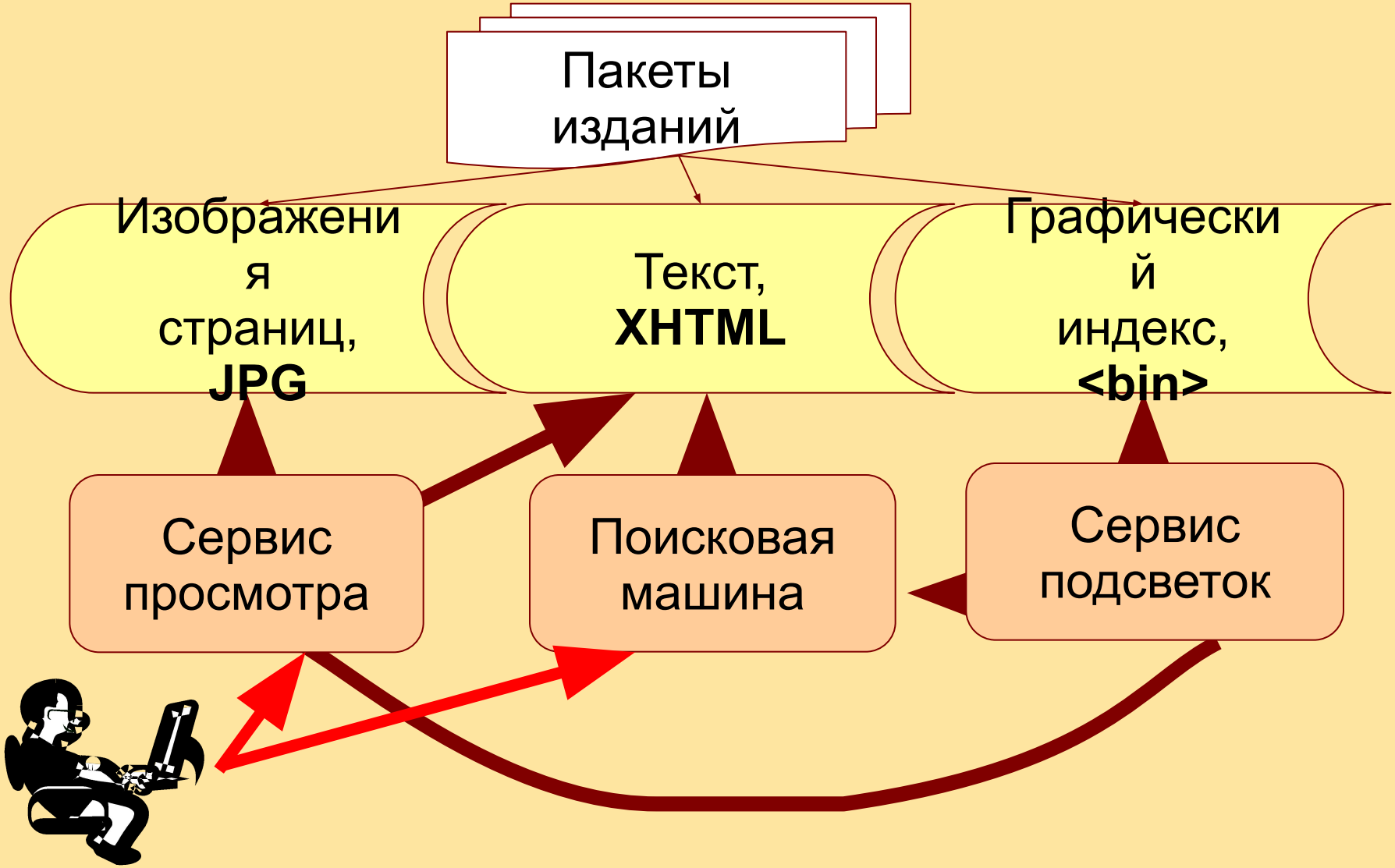
Пример текстового вида графического индекса

```
 W  177  386  217  587 |АКАДЕМИЯ
_W  177  600  217  953 |ПЕДАГОГИЧЕСКИХ
_W  177  965  217 1066 |НАУК
_W  177 1079  217 1180 |СССР

 W  477  378  579  440 |Л
 .  477  440  579  464 |.
_W  477  489  579  551 |С
 .  477  551  579  574 |.
_W  477  598  579 1186 |ВЫГОТСКИЙ

 W  643  541  739 1023 |СОБРАНИЕ

 W  754  500  851 1069 |СОЧИНЕНИЙ

 W  869  717  929  840 |ТОМ

 W  937  643  994  909 |ПЕРВЫЙ

 W 1974  654 2034  880 |МОСКВА
```

Флаги графического индекса

«_»	перед атомом стоит пробел
«**W**»	атом словный
«**.**»	атом несловный
«**-**»	перенос слова
«**!**»	пометка «проверить текст»
«**?**»	пометка «проверить графику»
«**/**»	разрыв абзаца

Архитектура сервера библиотеки



Почему не PDF или DjVu?

- Форматы PDF и DjVu — прекрасные решения для **отдельных книг и их автономного использования**.
- Но они не адекватны для создания электронных библиотек — потому, что они разработаны для других целей, см. выше
- Решение для комбинированного представления — это **технологическое** решение, специально разрабатываемое под электронные библиотеки.
- Различий в «идеологии» — нет.

Вопросы?