Лекция 21

Логическая и физическая схема организации пространства в документальных БД. Примеры моделей хранения и организации доступа

Отличия, обусловленные информационной природой элементов данных

Запись базы данных – ***документ***, который задается как набор в общем случае *необязательных* полей: «***форматных***» (числовые, символьные и другие величины) и ***текстовых*** (переменная длина, композиционная структура)

текстовое поле → параграф → предложение → слово

Поле **- атомарный адресуемый элемент данных** с точки зрения хранения

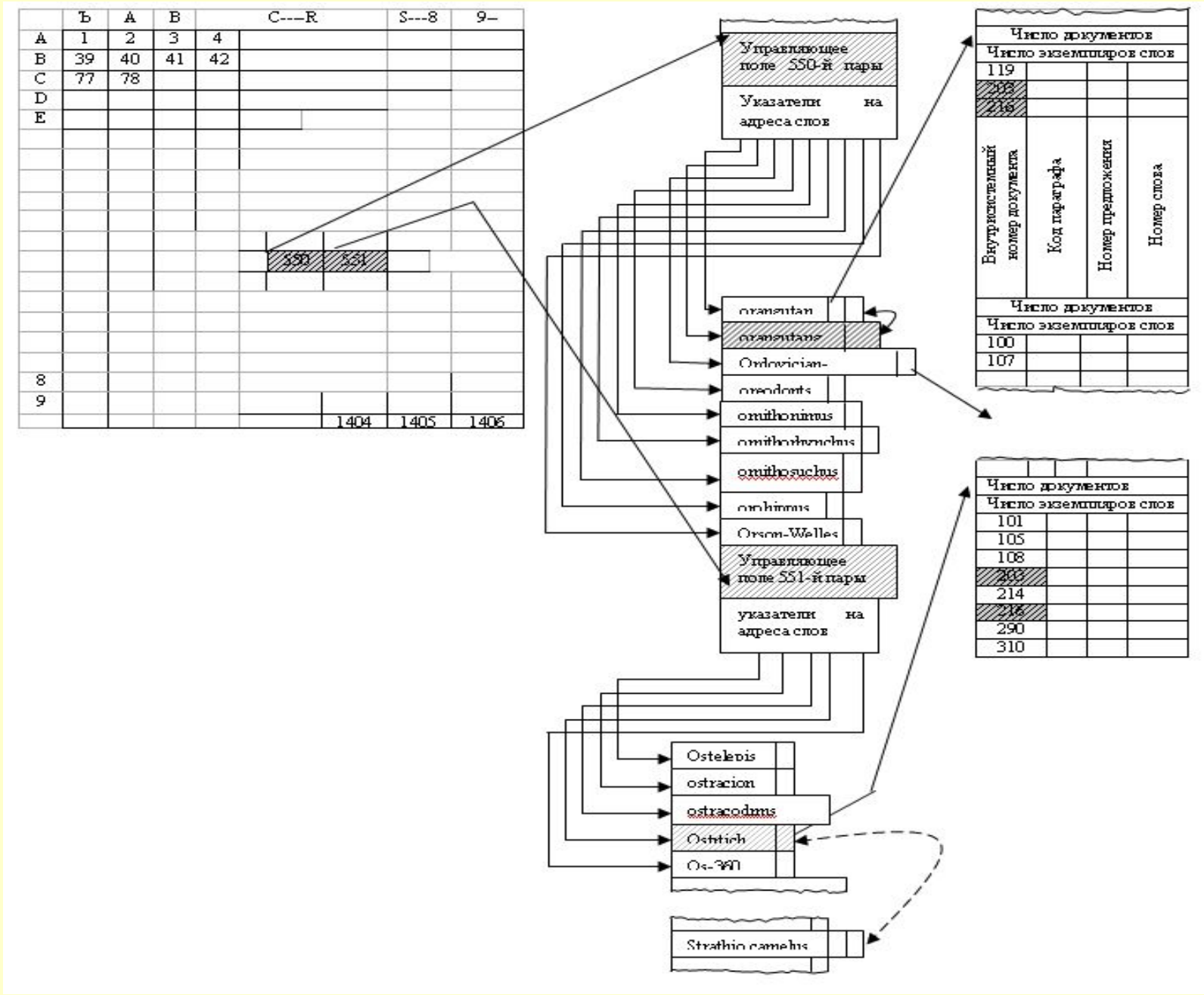
Слово **- атомарный семантически значимый элемент данных** с точки зрения поиска.

Семантическая природа текстовых полей: синонимия, полисемия, омонимия, контекстная обусловленность смысла отдельного слова, возможность выразить один смысл многими способами

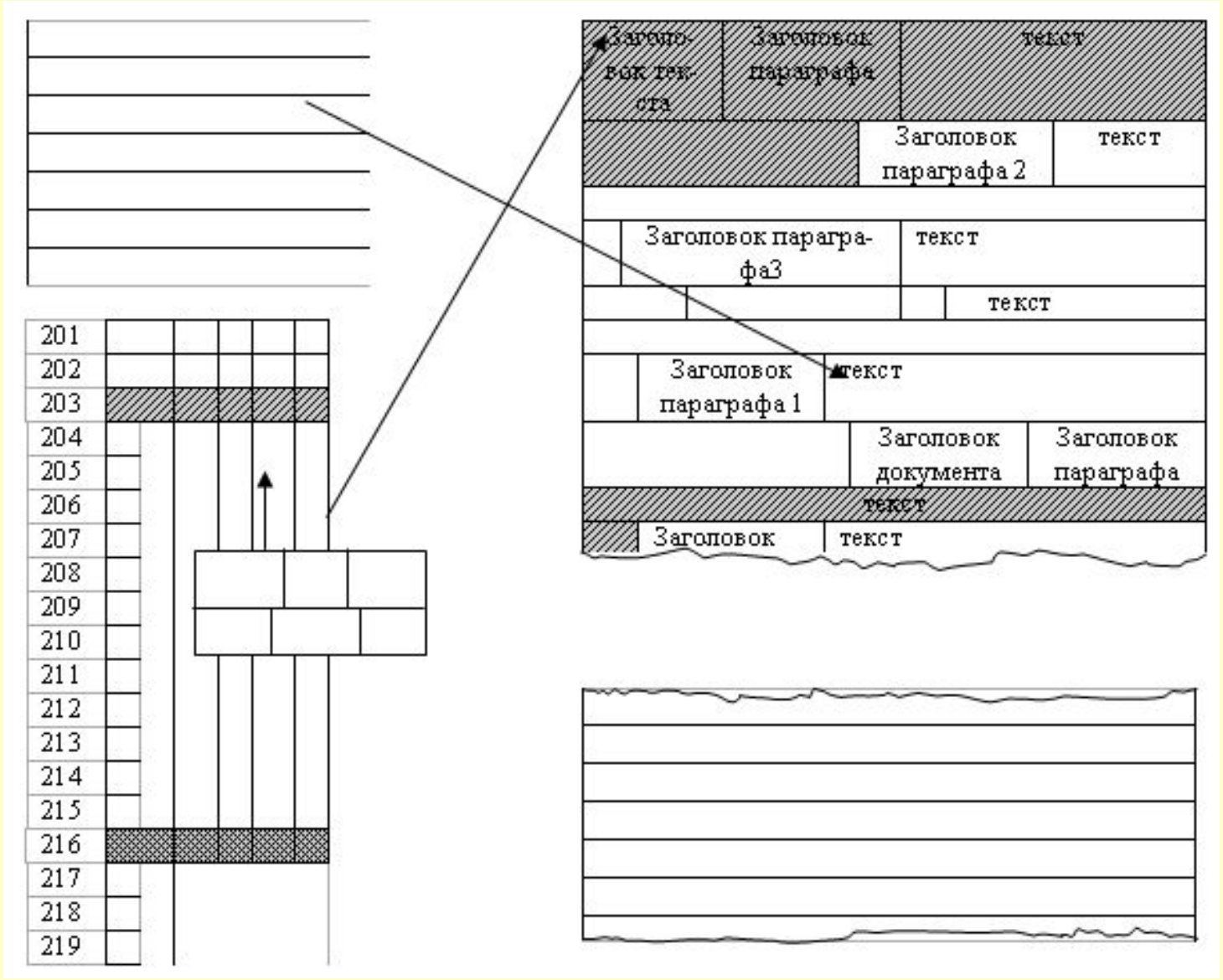
индексы <> словоформы поля

Организация данных в диалоговой системы поиска документов STAIRS (Storage and Information Retrieval System)

Организация индексных файлов документов АИПС STAIRS

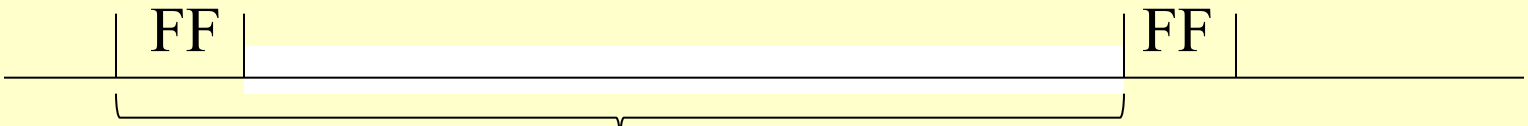


Организация поисковых файлов документов АИПС STAIRS

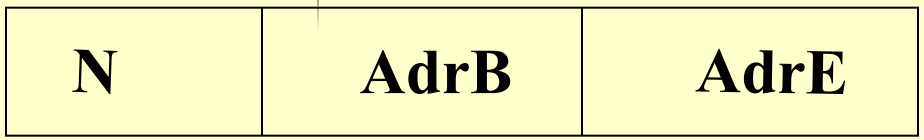


Организация доступа к документу в ИПС IRBIS

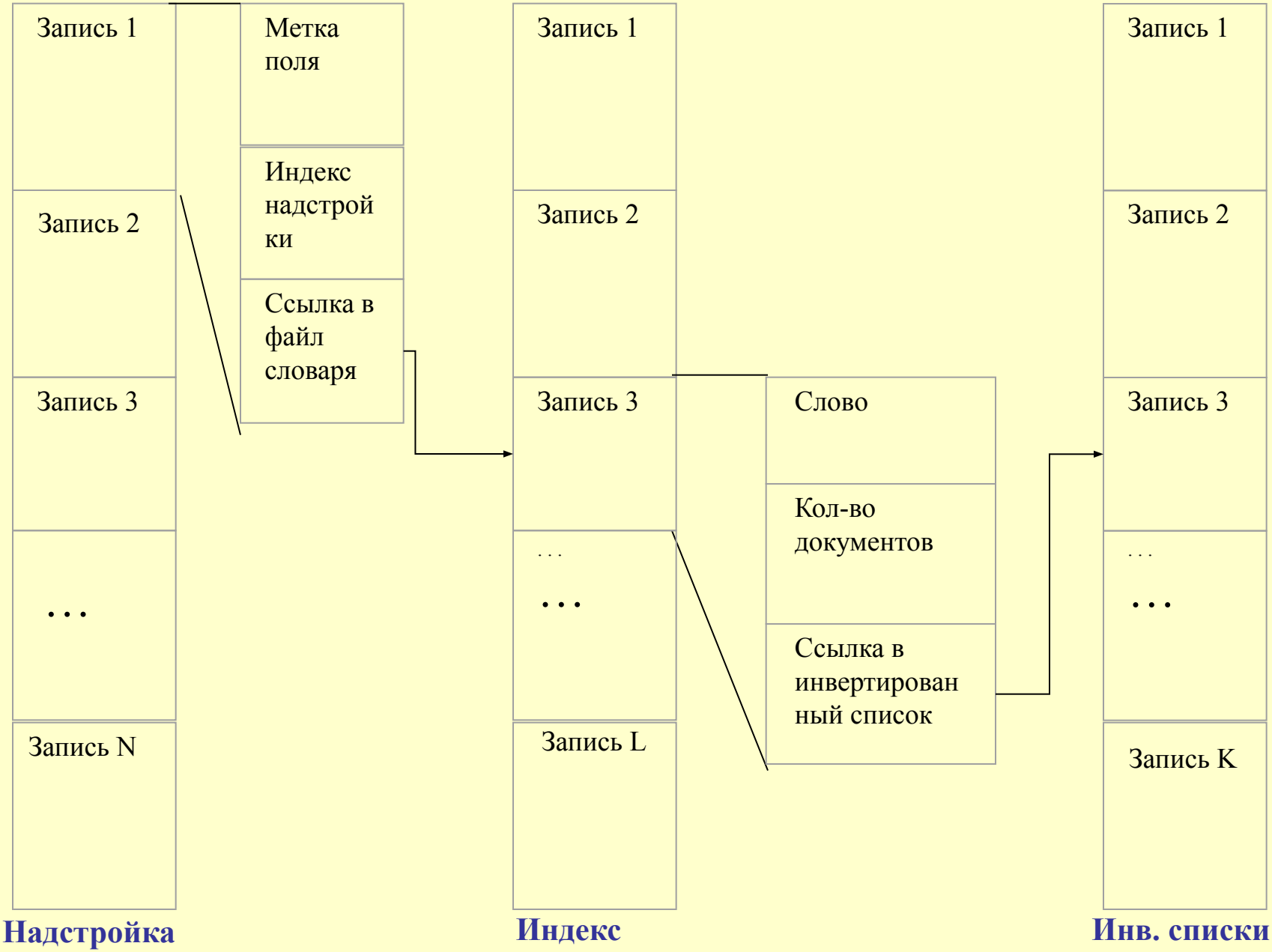
Массив документов



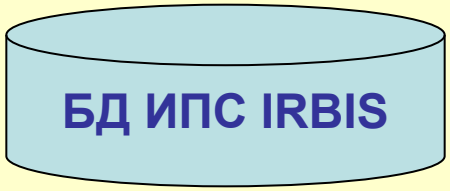
Индекс



Инвертированные индексы БД ИПС IRBIS



Физическая организация данных в ИПС IRBIS



Файлы БД

файлы данных и файлы инвертированных структур

Экстент

8 последовательных страниц

Страница

Инвертированные структуры БД

- Индексные страницы
- Страницы текстового представления словарных структур
- Страницы инвертированных списков

Заголовок страницы

- ✔ Тип страницы
- ✔ Идентификатор (номер) страницы
- ✔ Идентификатор (номер) следующей страницы
- ✔ Идентификатор (номер) предыдущей страницы
- ✔ Число вхождений, размещенных на странице
- ✔ Длина фиксированной части вхождения

Индексные страницы

Подзаголовок:

- ✔ число вхождений (указателей), размещенных на странице;
- ✔ номер первой страницы инвертированных списков для множества страниц текстового представления словарных структур, описываемых индексной страницей.

Указатели на отдельные страницы текстового представления словарных структур:

- ✔ метка сегмента (для представления общего словаря в виде объединения непересекающихся подмножеств);
- ✔ буква (символ), с которой начинается первое слово на странице;
- ✔ идентификатор (номер страницы).

Страницы текстового представления словарных структур

Подзаголовок:

- ✔ метка сегмента;
- ✔ номер первой страницы инвертированных списков;
- ✔ количество страниц инвертированных списков;
- ✔ размер свободного пространства;
- ✔ начало первого слова на странице (первые 4 буквы);
- ✔ начало последнего слова на странице (первые 4 буквы).

***Карта размещения словарных структур*:**

- ✔ длина слова (текстового выражения словарной структуры);
- ✔ количество документов (или длина инвертированного списка для словарной структуры);
- ✔ идентификатор страницы инвертированных списков, содержащей инвертированный список словарной структуры (по крайней мере, его начало);
- ✔ смещение начала инвертированного списка от начала списка страницы в целом.

Страницы инвертированных списков

Подзаголовок:

- ✔ метка сегмента (для представления общего словаря в виде объединения непересекающихся подмножеств);
- ✔ номер первой страницы текстового представления словарных структур (для текущей страницы инвертированных списков);
- ✔ количество страниц текстового представления словарных структур (которым соответствует текущая страница инвертированных списков);
- ✔ размер свободного пространства.