Открытые стандарты обработки документов. SGML и XML

Борис Тоботрас,

«Инфосистемы Джет»

Почему SGML?

Проблемы с обработкой документов:

- множество различных закрытых *несовместимых* форматов
- постоянная смена форматов и приложений
- трудности автоматической генерации и обработки документов
- непереносимость документов

Что такое SGML?

- международный стандарт разметки документов (ISO 8879:1986)
- метаязык для создания языков разметки - *приложений SGML* (например, HTML)
- документы хранятся в текстовом виде
- документы состоят из текста и элементов разметки
- структура документа строго определена

Что можно в SGML?

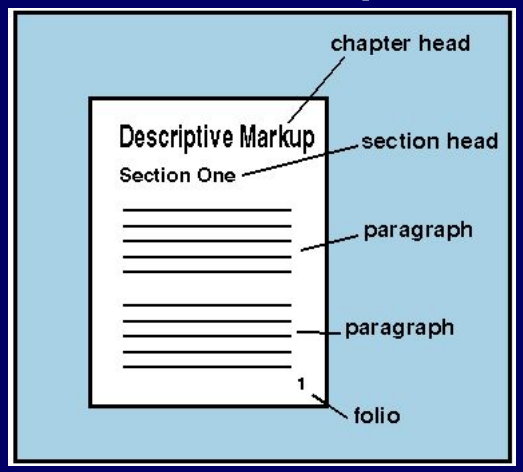
- один источник - много выходных форматов
- Web, связанные документы
- управление документами, версии, контекстный поиск
- управление данными

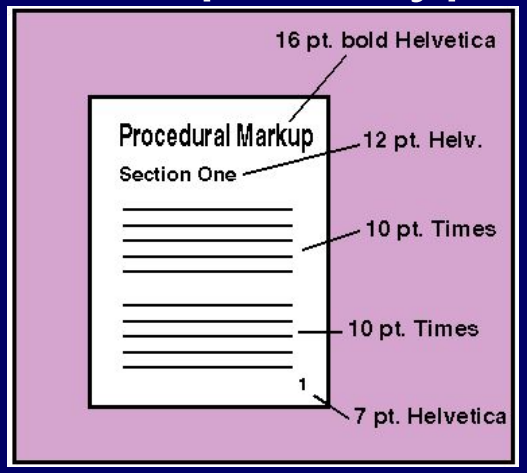
В чем суть SGML?
Отличия SGML

- разделение информации и представления
- типизированные документы
- выявление структуры информации
- управление данными
- связывание документов

Информация и представление

- описательная разметка вместо процедурной





- жесткая структура документа
- разные способы обработки документа
- стили

Типизированные документы

File Edit Motion Apps Options Buffers Tools Recent Files SGML Modify Move Markup View DTD

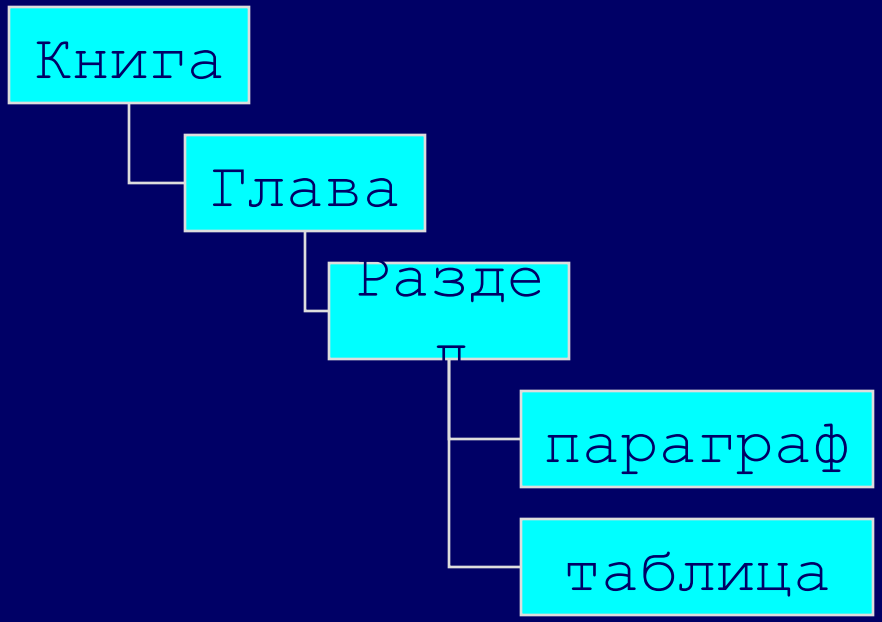
Open Dired Save Print Cut Copy Paste Undo Spell Replace Mail Info Compile Debug News

```
        <!doctype tz public "-//Jet//DTD TZ 2//RU">

-->


<!element tz - - (
        titlePage,
        body,
        back)>
<!attlist tz psgmlhelp CDATA "Техническое задание">

<!element titlePage - - (
        titleHead,
        customerApproved,
        performerApproved,
        (titlePart)+,
        startDate,
        (concordance)+
        )>
<!attlist titlePage psgmlhelp CDATA "Титульный лист">

<!element titleHead - - (quote|#PCDATA)+>
<!attlist titleHead psgmlhelp CDATA "Название фирмы">

<!element (customerApproved,performerApproved,concordance)
```

- понятие DTD
- анализаторы
- какие бывают DTD
  - универсальные
  - специализированные
- как сделать свой DTD?
  - элементы и их структура
  - атрибуты

Структура информации

- структурные элементы
- обработка текста в контексте структуры
- Пример:



Управление данными

- смысловые выделения
- обработка текста в контексте содержания
- Пример:

Связывание документов

- гарантия целостности
- двусторонние связи
- связи «один-ко-многим» и «многие-ко-многим»
- связи с произвольными точками документа
- Пример:
  - «ссылка на 3-ю главу 4-й части Руководства Администратора»

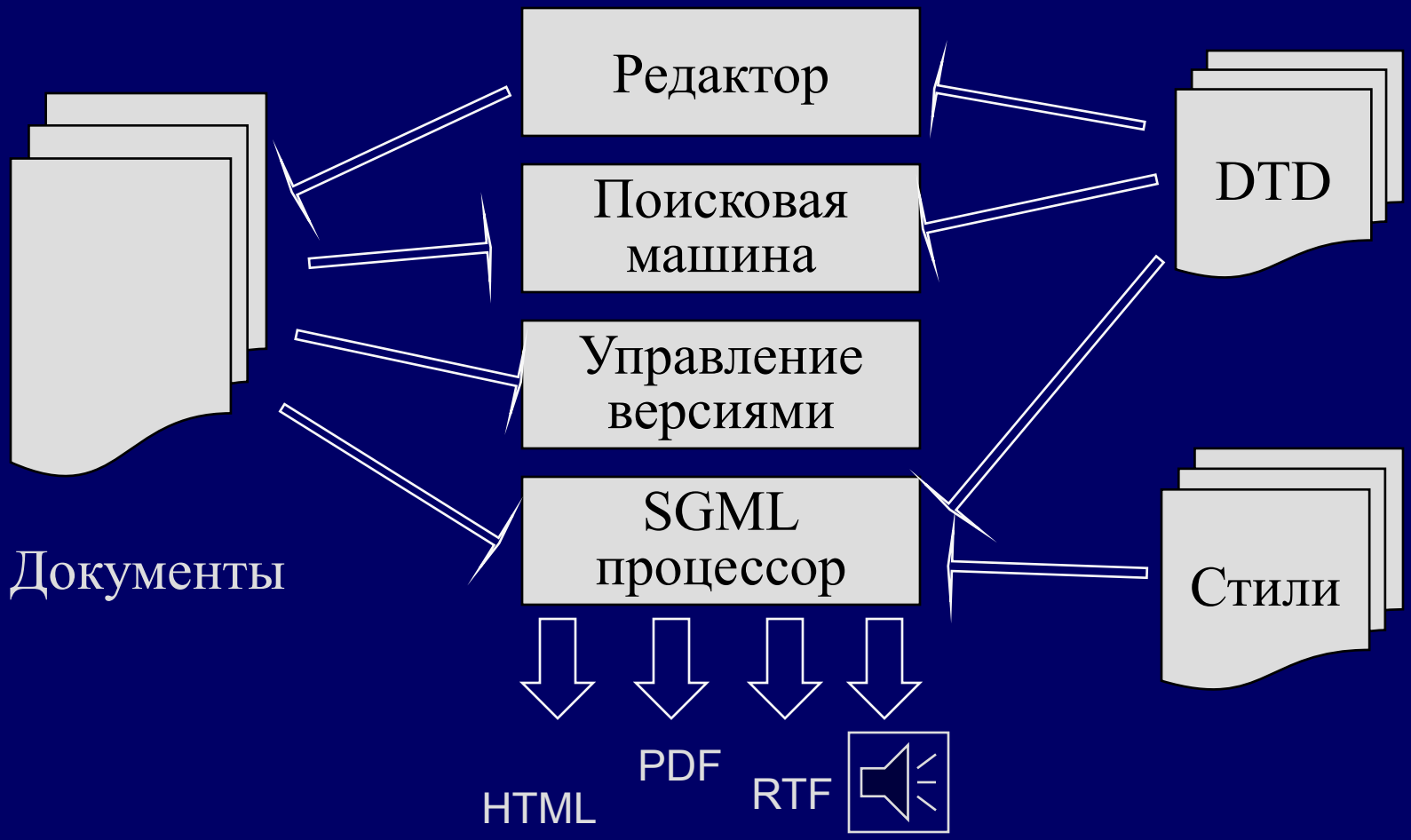
Преимущества SGML

- Продуктивность
- Единая стилистика
- Повторное использование
- Долговечность информации
- Разделяемость
- Мобильность
- Гибкость

SGML и другие

- HTML
  - уклон в сторону представления
  - размытость стандарта
  - нет возможностей расширения
- MS Word
  - закрытый
  - слабые средства автоматизации
  - нет смысловой и структурной разметки
- TeX
  - сложный, низкоуровневый
  - плохо экспортируется в Word

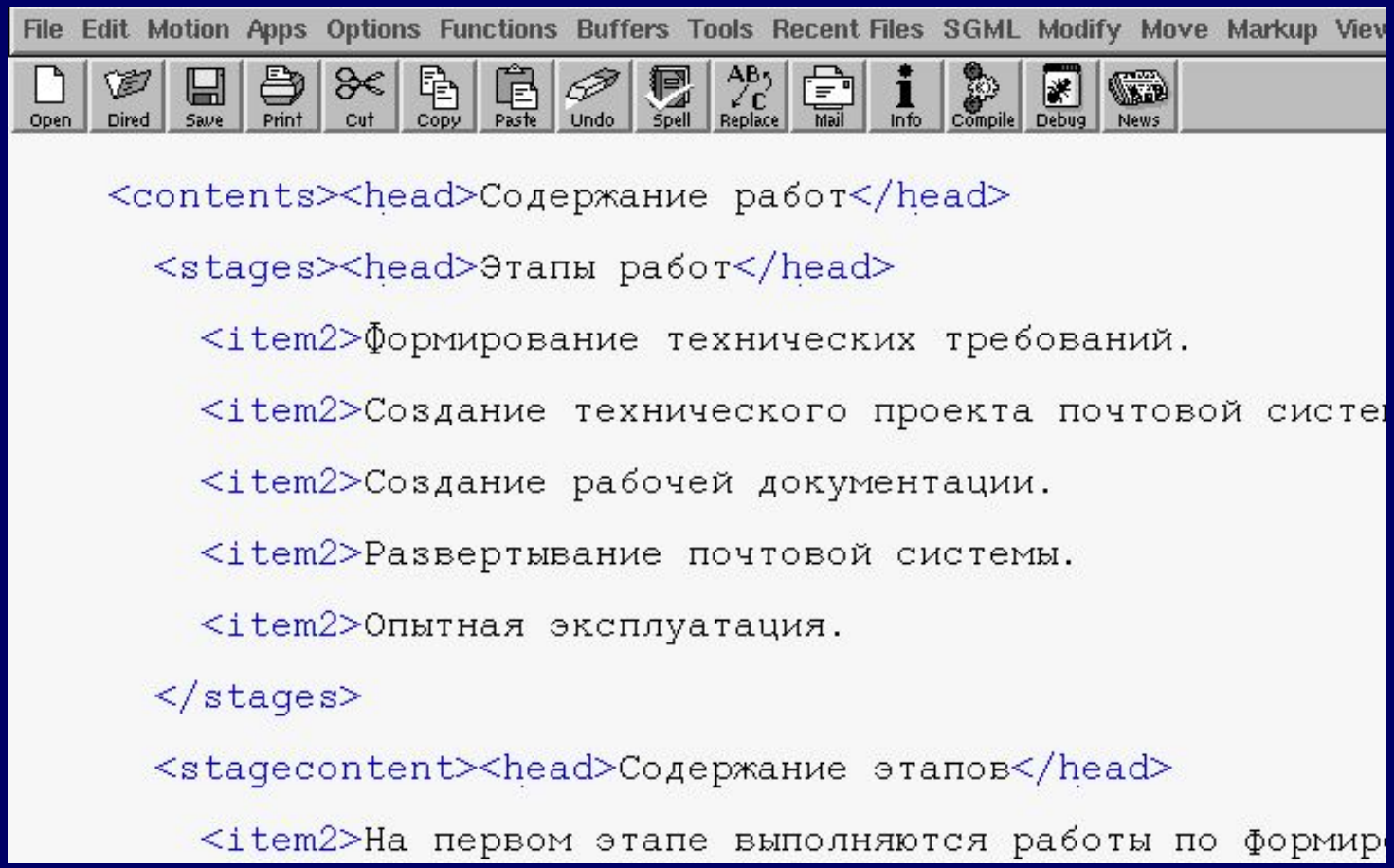
Как это делается



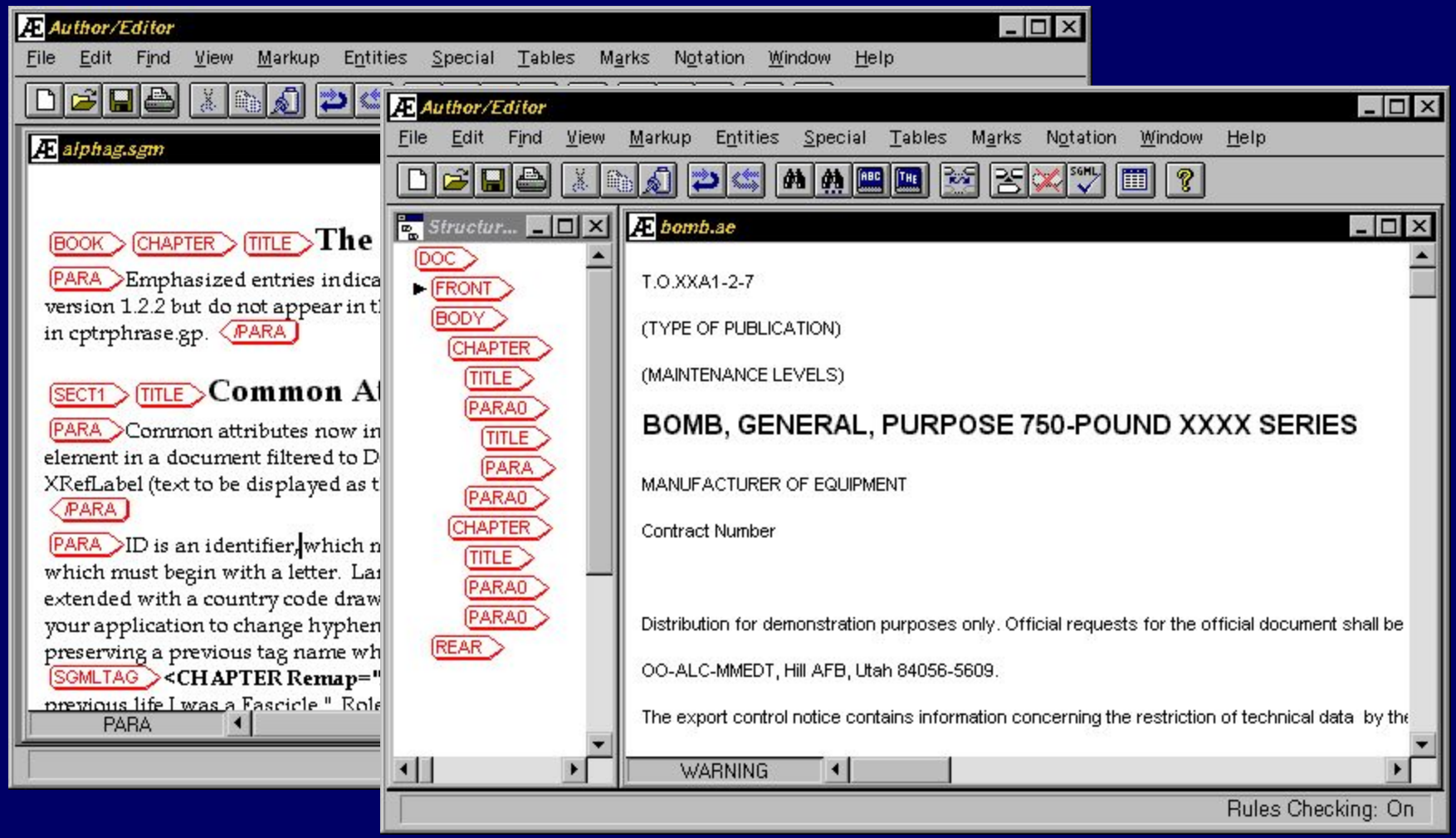
Как это делается

- SGML-редакторы
  - ArborText ADEPT*Editor, SoftQuad Author/Editor, Xemacs+psgml, Corel WordPerfect/SGML
- SGML-процессоры
  - SP, CoST, Jade, OmniMark, Balise,...
- Стили
  - DSSSL (Jade), XSL, CSS,...
- Выходные форматы
  - HTML, RTF, TeX, PostScript, PDF,...

Xemacs



Author/Editor



Контроль версий

- Текстовые файлы
- CVS/RCS
- diff
- Web-интерфейс

Групповая работа над документами

- Внешние объекты (документы, рисунки...)
- Параллельная работа
- Библиотека иллюстраций
- Пакетная обработка (сборка документа)

Поиск в документах

- Текстовые файлы
- Полнотекстовый поиск
- Поиск в контексте (структурном и смысловом)
- glimpse, CGI, Web

Генерация Web-сервера

- Общее дерево сервера в SGML
- Мастер-документ
- Взаимные ссылки
- Средства верификации
- Единый стиль

XML

- SGML, ориентированный на Web
- Упрощенный синтаксис
- Не обязательно наличие DTD
- Простые анализаторы

XML vs. HTML

- Автоматизация формирования страниц
- Возможность экспорта с WWW
- Точность поиска
- Неограниченное количество элементов

```
<FAQ>
        <Q>Что такое XML?</Q>
        <A>eXtensible Markup Language</A>
</FAQ>
```

- XSL и XLL - дополнение к XML

Инфосистемы Джет

Тел. 973-48-57, 973-48-58
info@jet.msk.su

Борис Тоботрас, tobotras@jet.msk.su