Технологии обработки документов

1. Текстовая информация. Модель документа

- Известно, что существуют различные типы текстовых файлов (плоские, размеченные, ASCII и пр.). Соответственно, для ввода, работки, представления информации в таких файлах требуются различные программные возможности. Для работы с текстами на компьютере используются программные средства, называемые текстовыми редакторами, или текстовыми процессорами.

Разновидности текстовых форматов

- Существует большое количество разнообразных текстовых редакторов, различающихся по своим возможностям, – от очень простых учебных до мощных, многофункциональных программных средств, называемых издательскими системами, которые используются для подготовки к печати книг, журналов и газет. Эти программы позволяют работать с различными типами и форматами текстовых файлов, по необходимости преобразуя их друг в друга.
- Например, в текстовом формате (плоский текст – .TXT) работают редактор Notepad, встроенные редакторы оболочек Norton Commander и Far Manager, в то время как Word (а также WordPad) позволяют работать с размеченными текстовыми файлами в коммуникативном (тип файла .RTF – rich text format, или «обогащенный формат текста»), внутреннем (.DOC), и текстовом (.TXT) форматах. Распространен редактор документов Adobe Acrobat, использующий коммуникативный формат .PDF (portable document format). Также широко применяются форматы разметки текстов HTML.
- Наиболее развитые редакторы позволяют обрабатывать не просто тексты, а документы (тексты, содержащие встроенные или внедренные объекты или файлы других типов – табличные, графические, мультимедиа и пр.).

Типы файлов для размещения документов

- текстовые файлы – обобщенное название для простых и размеченных текстов, ASCII-файлов и других наборов данных символьной информации, которые интерпретируются и обрабатываются текстовыми редакторами, процессорами, анализаторами (Lexicon, Word, TEC, анализаторы SGML, HTML);
- текст без разметки (планарный) – файл, содержащий только отображаемые (воспроизводимые на всех печатающих устройствах и терминалах) символы кода ASCII, а также простейшие управляющие символы: CR – возврат каретки; LF – перевод строки; TAB – символ табуляции, иногда LF – новая страница;
- текст с разметкой – планарный файл, содержащий бинарную и символьную разметку, управляющую отображением информации (программно и/или аппаратно);
- ASCII-файл – содержит только отображаемые коды левой части кодовой таблицы ASCII (латиница и служебные символы), обычно применяется для хранения документов с символьной разметкой (RTF, SGML, HTML).

Форматы полнотекстовых документов. Модель документа

- Понятие *модель документа* охватывает аспекты создания, преобразования, хранения, поиска, передачи и отображения документов. Принято рассматривать структуру документа в двух аспектах: логическом (содержание) и физическом (макет).
- ***Логическая структура*** определяет составные компоненты и их соотношения в понятиях, отвечающих взгляду на документы как смысловые структуры. Например, к основным смысловым компонентам относятся: авторские данные (имя автора, место работы), аннотация, оглавление, главы, разделы, параграфы, рисунки, сноски.

- На рисунке приведен пример документа «Пояснительная записка к дипломному проекту (работе)». Здесь выделены такие базовые понятия структуры документа, как обязательность/необязательность элемента, уникальность или повторяемость, вхождение нижестоящих элементов в вышестоящие по принципу И (оба типа данных должны или могут входить в элемент) либо ИЛИ (только какой-либо один из типов данных может или должен входить в элемент).

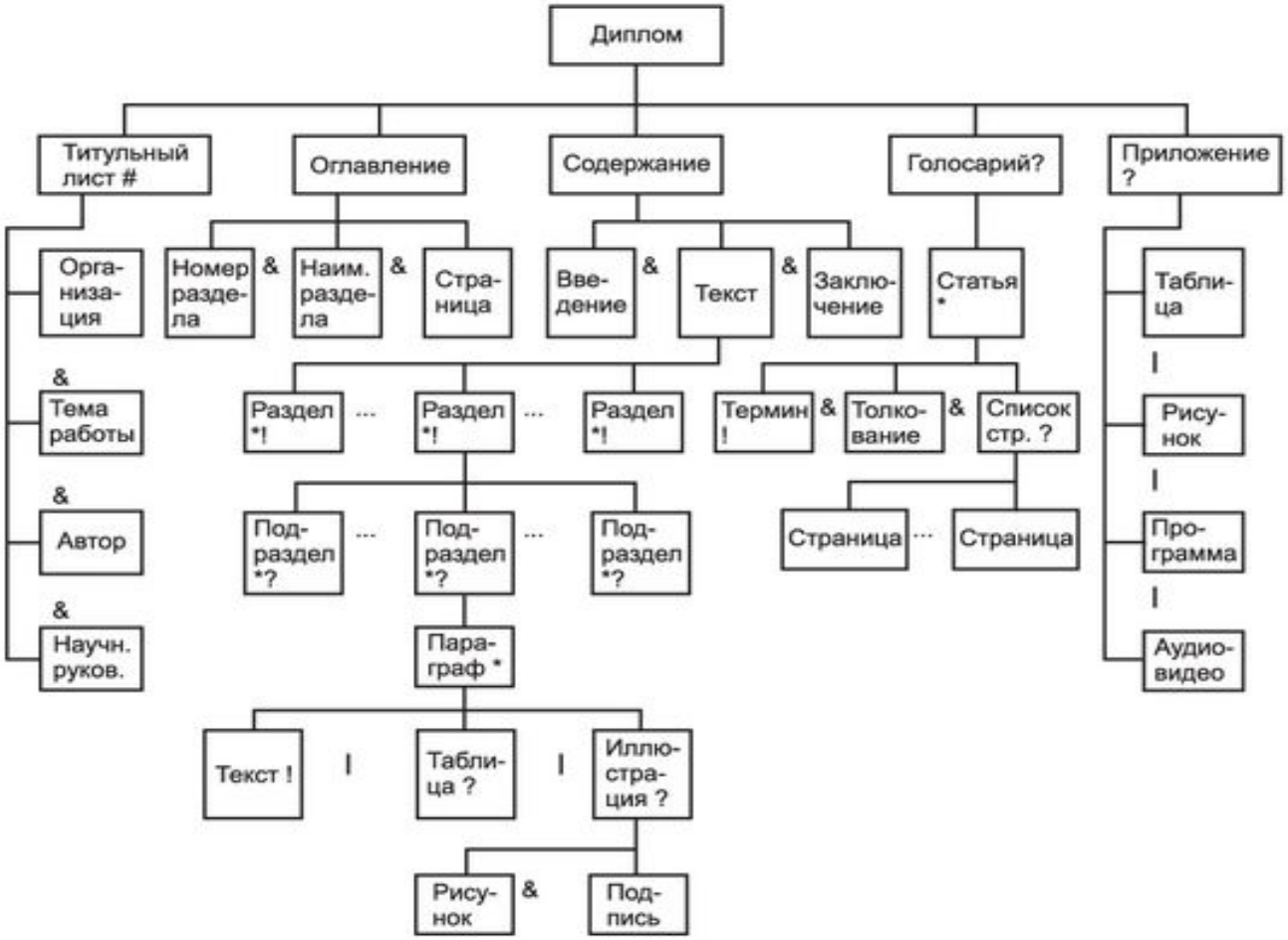

Рис. 2.5. Иерархическая структура документа «Пояснительная записка к дипломной работе»: # — уникальный элемент; * — повторяющийся элемент; ? — необязательный элемент; ! — обязательный элемент; & — вхождение типа «И»; | — вхождение типа «ИЛИ»

Макетная структура содержит описание документа в терминах физических единиц – страниц, полос, колонок, колонтитулов, рамок для рисунков, шрифтов, стилей и пр.

Подходы к моделированию документов опираются на два стандарта – ISO 8613 (ODA — Office Document Architecture — архитектура управленческой документации и ISO 8879 (SCM – Standard Generalized Markup Language — стандартный обобщенный язык разметки).

Документ в ODA представлен в виде профиля и собственно документа, организованных в форме древовидной структуры. Профиль содержит информацию о документе в целом и его прохождении; формальные признаки – дата составления, вид, регистрационный номер и т. д.

Собственно документ содержит текст и сведения о его структуре и стиле, а именно:

- структуру документа – заглавие, параграфы, оглавление и т. п. (логическая структура), а также абзацы, расположение текста, шрифты (физическая структура);
- архитектуру содержания – набор графических элементов, выделение определенных слов, строк и т. п.;
- коммуникативный формат – способы кодирования объектов, признаков и содержания документов.

2. Языки разметки документов

В системах обработки текстов в документ включается дополнительная информация, называемая разметкой и выполняющая следующие функции:

- выделение логических элементов данного документа;
- задание функций обработки выделенных элементов.

В обычных текстовых процессорах существуют встроенные команды включения/выключения шрифтов и др., аналогичные командам управления размещением информации на экране или при печати. Такой подход называется командной или процедурной разметкой.

Альтернативный способ разметки заключается в выделении части текста без указания способа обработки выделения. Затем другие команды назначают фрагментам способ обработки. Такая разметка называется *описательной* (дескриптивной). Она включает метки *(tags, маги)* начала и окончания элемента текста и указывает, как интерпретировать данный фрагмент.

Изменяя набор процедур, соответствующий описательной разметке, можно изменить внешнее представление одного и того же документа.

Основным достоинством описательной разметки является ее гибкость, поскольку фрагменты текста отмечены как «чем они являются» (а не «как они должны быть отображены).

Описательная разметка также облегчает задачу переформатирования документа при необходимости, поскольку описание формата не связано с содержанием.

В настоящее время существует множество языков разметки, например, DocBook, MathML, SVG, Open eBook, XBRL и др. В основном они предназначены для представления различных текстовых документов но специализированные языки могут использоваться во многих

Таблица 2.2. **Основные языки разметки**

Язык	Разработчик	Год выпуска	Редактирование	Просмотр	Основное назначение	На чем основан	Тип разметки	Структурная разметка	Управление представлением
Troff (typesetter runoff), groff (GNU runoff)	Joe Ossanna	1973	Текстовый редактор	Groffer или вывод в PostScript	Техническая документация	RUNOFF	Управляющие коды	Да	Да
TeX	Donald Knuth	1978	Текстовый редактор	DVI или конвертер в PDF	Научная литература		Управляющие коды	Да	Да
Maker Interchange Format (MIF)	Frame Technology acquired by Adobe Systems in 1995	1986	Текстовый редактор, FrameMaker	FrameMaker	Техническая документация		Теги	Да	Да
Rich Text Format (RTF)	Microsoft	1987	Текстовый редактор, Word processor	Word processor	Форматированные документы		Управляющие коды	Да	Да
Text Encoding Initiative (TEI)	Text Encoding Initiative Consortium	1990	Текстовый редактор /XML	Web-браузер (через конвертирование в XHTML), PDF или Word Processor (конвертирование в ODF)	Научная, техническая и пр. документация	XML	Теги	Да	Нет
DocBook	The Davenport Group	1992	Редактор XML	Вывод в формат HTML, PDF, CHM, javadoc, others	Техническая документация	SGML / XML	Теги	Да	Нет
HyperText Markup Language (HTML)	Tim Berners-Lee	1993	Текстовый редактор /HTML	Web-браузер	Гипертекстовые документы	SGML	Теги	Да	Да

Окончание табл. 2.2

Язык	Разработчик	Год выпуска	Редактирование	Просмотр	Основное назначение	На чем основан	Тип разметки	Структурная разметка	Управление представлением
Encoded Archival Description (EAD)	Berkeley Project	1998	Текстовый редактор	Web-браузер	Поиск информации	XML	Теги	Да	Нет
Math Markup Language (MathML)	W3C	1999	Текстовый редактор /XML конвертер в TeX	Web-браузер, Word processor	Математическая литература	XML	Теги	Да	Да
Office Open Extensible Markup Language (MS OOXML)	Microsoft	2000	Office suite	Office suite	Многоцелевой	XML/ZIP	Теги	Да	Да
Extensible HyperText Markup Language (XHTML)	W3C	2000	Текстовый редактор /XML /HTML	Web-браузер	Гипертекстовые документы	XML	Теги	Да	Нет
Open Mathematical Documents (OMDoc)	Michael Kohlhase	2000	Текстовый редактор /XML	Вывод в формате XHTML+MathML, TeX и др.	Математическая литература	XML	Теги	Да	Да
Wireless Markup Language (WML)	WAP Forum	2000	Текстовый редактор /XML	Microbrowser	Гипертекстовые документы	XML	Теги	Да	Да
Music Extensible Markup Language (MusicXML)	Recordare	2002	Scorewriter	Scorewriter	Запись музыки	XML	Теги	Да	Да
Darwin Information Typing Architecture (DITA)	OASIS	2005	Текстовый редактор /XML	Вывод в формат HTML, PDF, CHM и др.	Техническая документация	XML	Теги	Да	Нет
OpenDocument Format (ODF)	OASIS	2005	Office suite	Office suite	Многоцелевой	XML/ZIP	Теги	Да	Да

RUNOFF

RUNOFF была первой системой форматирования текстов, которая получила значительную известность. Она была разработана в 1964г. для операционной системы CTSS Джеромом X. Салтзером (Jerome H. Saltzer) с использования ассемблера MAD.

Продукт фактически состоял из пары программ:

- TYPSET, который был в основном редактором документов;
- RUNOFF – процессор вывода.

RUNOFF осуществлял поддержку разбиения на страницы и размещения заголовков, а также выравнивания текста.

TeX

TeX – наборная система, созданная Дональдом Нутом (Donald Knuth). Вместе с языком METAFONT для описания шрифта и Computer Modern typeface (Компьютерного Современного шрифта) он был спроектирован для двух основных целей:

1) представить каждому пользователю возможность создавать высококачественные книги в пределах разумных трудозатрат
2) чтобы такая система давала идентичные результаты на любых компьютерах как в настоящее время, так и в будущем. TeX – бесплатное программное обеспечение, популярное в академическом сообществе, особенно среди математиков, физиков информатиков, экономистов, и в технических сообществах.

TeX хорошо используется для создания и распечатки сложных математических формул и других программных средств форматирования.

- ***PostScript***

язык программирования, реализующий функцию описания страниц, использующийся в электронных изданиях и настольных издательских системах.

Концепция языка PostScript была создана в 1976 г. Джоном Вонок (John Warnock). Данный язык систем проектирования задумывался для обработки графической информации.

- PostScript скомбинировал лучшие особенности принтеров и плоттеров. Как и плоттеры, PostScript предложил высококачественную штриховую графику и единый язык управления, который мог использоваться на принтерах любых марок. Как матричные печатающие устройства, PostScript предложил простые способы генерировать страницы текста и растровой графики. Но, в отличие от обоих, PostScript мог располагать все эти данные на единой странице, что предлагало гораздо больше гибкости, чем любой принтер или плоттер.

- ***Portable Document Format (PDF)***

Переносимый формат документов – PDF – формат файла, созданный Adobe Systems в 1993 г. для использования в настольных издательских системах. Формат PDF позволяет представлять двумерные документы в форме, независимой от разрешающей способности устройств печати (или дисплеев). Каждый файл формата PDF содержит полное описание двумерного документа (с появлением Acrobat 3D – трехмерных документов), который включает текст, шрифты, изображения и двумерную векторную графику, которые образуют документ.

Формат файла формата PDF подвергся нескольким изменениям с выпуском новых версий Adobe Acrobat. Известно восемь версий формата PDF - 1.0 (1993 г.), 1.1 (1994 г.), 1.2 (1996 г.), 1.3 (1999 г.), 1.4 (2001 г.), 1.5 (2003 г.), 1.6 (2005 г.) и 1.7 (2006 г.) которые соответствуют выпускам Adobe Acrobat от 1.0 до 8.0

Формат PDF использует следующие технологии:

- подмножество языка программирования и описания страниц PostScript, чтобы генерировать размещение и графику;
- систему встраивания и замены шрифтов для обеспечения перемести мости документов;
- структурированную систему хранения, позволяющую связывать эти элементы в отдельный файл, с использованием сжатия данных при необходимости.

- ***Язык SGML***

разработан на базе программного продукта DCF GML фирмы IBM и представляет собой метод создания структурированных документов, а также языков для их разметки.

В языке SGML каждый документ имеет три части:

- декларации (объявления, определения) языка SGML, привязывающие к определенным значениям параметры обработки, а также имена синтаксиса;
- пролог, состоящий из деклараций о типе документа. Они определяют типы элементов, взаимосвязи между элементами и их атрибуты, а также условные обозначения, которые могут быть задействованы при разметке;
- данные, которые состоят из разметки документа и собственно информации.

- ***HTML – язык разметки гипертекста***

Hypertext Markup language формулируется в терминах языка SGML.

HTML ориентирован на решение нескольких важных задач, в которых участвуют его различные конструкции и элементы:

- описание структуры документа;
- адресация ресурсов;
- создание гипертекстовых ссылок и управление навигацией в БД локальных и WWW Internet ;
- реализация интерфейсов с пользователем
- .

- **Формат DocBook**

DocBook – язык разметки технических документов (1991г.). Он был первоначально предназначен для того, чтобы разрабатывать техническую документацию, связанную с компьютерной аппаратурой и программным обеспечением, однако может использоваться и для любых других видов документации.

Одно из основных преимуществ DocBook – он дает возможность пользователям создавать содержание документа в нейтральной форме, которая описывает только логическую структуру содержания, которое затем может быть отображено в разнообразных форматах, включая HTML, формат PDF, страницы руководства и помощи, не требуя от пользователей каких-либо изменений в исходном тексте.

DocBook первоначально начал существование как приложение SGML, однако затем было разработано эквивалентное приложение XML, которое заменило SGML в большинстве применений.

- ***ODF***

OpenDocument (ODF) (Открытый формат документов для офисных приложений), является форматом файла документа, используемым для того, чтобы описать электронные документы, например письма, сообщения, книги, электронные таблицы, диаграммы, презентации и файлы текстовых процес соров. Стандарт основан на формате XML, был разработан техническим комитетом OASIS (Organization for the Advancement of Structured Information Standards) и первоначально воплощен в офисном комплексе OpenOffice.org.

Основная цель таких открытых форматов: гарантировать долгосрочный доступ к данным без юридических или технических барьеров. OpenDocument является альтернативой закрытым форматам (doc, xls и ppt).

4. Текстовые редакторы

- Редакторы, предназначенные для подготовки текстов условно можно разделить на обычные (подготовка писем и других простых документов) и сложные (оформление документов с разными шрифтами, включающие графики, рисунки и др.).

Редактор Word

Основные функции. Текстовый редактор Word реализует следующие функции:

- создание, открытие, закрытие, сохранение текстовых документов*;*
- задание параметров страниц*;*
- набор текста (режим прописных букв, гарнитура, кегль и цвет шрифта, страница, работа с выделенным фрагментом "текста, межстрочный интервал, способы выравнивания, буфер обмена);
- форматирование абзаца (задание параметров абзаца, красная строка, межстрочный интервал) *;*
- задание шрифтов;
- установка рамки и заливки абзаца;
- создание нумерованных и маркированных списков, настройка нумерованных списков;
- ссылки, заголовки, оглавления;
- проверка правописания, расстановка переносов;
- создание, заполнение и форматирование статических таблиц; рамки, заливка. Изменение структуры таблицы (добавление и удаление строк и столбцов, объединение ячеек, изменение размеров ячеек). Преобразование текста в таблицу и наоборот;
- вставка и редактирование объектов — рисунков, клипов, MIDI-файлов, математических формул*;*
- деловая графика (построение диаграмм и графиков). Вставка рисунков, настройка положения, размера и способа обтекания рисунка (в тексте, перед текстом, за текстом и пр.) ;
- работа с автофигурами (линии, фигуры, стрелки и пр.), использование WordArt;
- печать текста.

Редактор документов OpenOffice.org Writer

В OpenOffice.org Writer, пользователь может создавать любые текстовые документы, составлять личные и официальные письма, брошюры, факсы и профессиональные учебные пособия. Документы, которые используются часто, можно сохранять как шаблоны. Имеется проверка орфографии и тезаурус, а при необходимости может быть задействована Автозамена и расстановка переносов во время ввода текста с клавиатуры. В OpenOffice.org нет ограничений на длину текстового документа.

OpenOffice.org реализует следующие функции:

- Создание и структурирование документов;
- Подготовка публикации;
- Вычисления;
- Создание чертежей;
- Вставка изображений;
- Изменяемый интерфейс приложения.

5. Работа с электронными таблицами

- Электронная таблица — интерактивная система обработки информации, упорядоченной в виде таблицы с поименованными строками и столбцами. Прототипом современных электронных таблиц послужила разработанная в 1979 г. специалистами США программа Visual Calc. Ныне наиболее часто используются электронные таблицы Quatro Pro, MS Excel и Lotus 1-2-3

Основные характеристики программного продукта Excel

- Excel представляет собой мощный арсенал средств ввода, обработки и вывода в удобных для пользователя формах фактографической информации. Эти средства позволяют обрабатывать фактографическую информацию, используя большое число типовых функциональных зависимостей: финансовых, математических, статистиченских, логических и т. д., строить объемные и плоские диаграммы, обрабатывать информацию по пользовательским программам, анализировать ошибки, возникающие при обработке информации, выводить на экран или печать результаты обработки информации в наиболее удобной для пользователя форме.
- Структура таблицы включает нумерационный и тематический заголовки, головку (шапку), боковик (первая графа таблицы, содержащая заголовки строк) и прографку (собственно данные таблицы). На пересечении столбца и строки устанавливается графическая смысловая связь между понятием, объединяющим материал в строку, и понятием, объединяющим материал в стол бец, что позволяет выявить ее без мысленного перевода в словесную форму и существенно облегчить усвоение и анализ организованных в таблицу данных

Структура таблиц и основные операции:

- в нижней части электронной таблицы расположен алфавитный указатель (регистр), обеспечивающий доступ к рабочим листам. Пользователь может задавать названия листам в папке (вместо алфавитного указателя), что делает наглядным содержимое регистра, облегчает поиск и переход от документа к документу;
- в режиме оформления и модификации экрана можно фиксировать заголовки строк, столбцов, оформлять рабочие листы и т. д.;
- для оформления рабочих листов в табличном процессоре предусмотрены возможности: выравнивания данных внутри клетки, выбора цвета фона клетки и шрифта, изменения высоты строк и ширины колонок, черчения рамок различного вида, определения формата данных внутри клетки (например: числовой, текстовый, финансовый, дата и т. д.), а также обеспечения автоматического форматирования, когда в систему уже встроены различные варианты оформления таблиц, и пользователь может выбрать наиболее подходящий формат;
- для вывода таблиц на печать предусмотрены функции, обеспечивающие выбор размера страницы, разбивку на страницы, установку размера полей страниц, оформление колонтитулов, а также предварительный просмотр получившейся страницы;
- связывание данных – абсолютная и относительная адресации являются характерной чертой всех табличных процессоров. Они дают возможность работать одновременно с несколькими таблицами, которые могут быть тем или иным образом связаны друг с другом;

- вычисления – для удобства вычисления в табличных процессорах имеются встроенные функции: математические, статистические, финансовые, даты и времени, логические и др. Менеджер функций позволяет выбрать нужную и, проставив значения, получить результат;
- деловая графика – возможность построения различного типа двухмерных, трехмерных и смешанных диаграмм (более 20 различных типов и подтипов), которые пользователь может строить самостоятельно. Многообразны и доступны возможности оформления диаграмм, например, вставка и оформление легенд, меток данных; оформление осей – возможность вставки линий сеток и другие;
- выполнение табличными процессорами функций баз данных – обеспечивается заполнением таблиц аналогично заполнению БД, т. е. через экранную форму защитой данных, сортировкой по ключу или по нескольким ключам, обработкой запросов к БД, созданием сводных таблиц. Кроме этого, осуществляется обработка внешних БД, позволяющая работать с файлами, созданными например, в формате dBase, Paradox или других форматах;
- программирование – в табличном процессоре существует возможность использования встроенного языка программирования макрокоманд. Разделяют макрокоманды и макрофункции. При использовании макрокоманд упрощается работа с табличным процессором и расширяется список его собственных команд. С помощью макрофункций определяют собственные формулы и функции, расширив, таким образом, набор функций, предоставляемый системой.