Модель текстового документа

Особенности

- Информационные технологии отождествляют с технологиями работы с документом.
- Так ли это?

Типы текстовых файлов

- Плоские (.txt)
- Размеченные (коммуникативный формат .rtf «обогащенный формат текста», внутренний .doc,
- ASCII
- Др.

Представление информации разное.

- Текстовый файл представляет из себя последовательность символов
- Символы обычно сгруппированы в строки.
- Строки разделяются разделителями строк.

Типичные файлы данных

- текстовые файлы — обобщенное название для простых и размеченных текстов, ASCII-файлов и других наборов данных символьной информации
- 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 — файл, содержащий только отображаемые коды левой части кодовой таблицы ASCII (латиница и служебные символы) и обычно применяющийся для хранения документов с символьной разметкой (RTF, SGML, HTML).

Бинарная разметка

Вид разметки	Управляющие символы принтера EPSON	Редактор Лексикон
Полужирный шрифт	ESC G...ESC H	chr(255)2... chr(255)0
Курсив (италик)	ESC 4...ESC 5	chr(255)1... chr(255)0
Подчеркивание	ESC-1...ESC-0	_chr(255)... chr(0)
Индекс верхний	ESC S 0 ...ESC T	chr(255)5... chr(255)0
Индекс нижний	ESC S 1...ESC T	chr(255)4... chr(255)0
Выбор вида шрифта	NLQ – ESC x 1 DRAFT – ESC x 0 Отмена chr(18)	
Перевод страницы	chr(12)	.chr(12)
Выравнивание		
Параграф (абзац)	Табуляция (TAB, chr(9))	TAB

Символьная разметка

Формат RTF (Rich Text Format), в том числе WinWord	Формат электронной почты (стандарт MIME)	HTML
{\b...}	<bold>...</bold>	...
{\i...}	<italic>...<italic>	...
{ul\...}	<underline>... </underline>	<u>...</u>
{\super...}	<superscript>... </superscript>	^{...}
{\sub...}	<subscript>... </subscript>	_{...}
\f11 – Courier, \f4 –Times, \f5 – Arial		
\page	пр	
\qc – по центру \ql – влево \qr – вправо \qj – по краям	<Center> <FlushLeft> <FlushRight>	<aligh =center> =left> =right> =justify>
\par	Paragraph	<p>

ASCII

- Аббревиатура от ***A****merican* ***S****tandard* ***C****ode for* ***I****nformation* ***I****nterchange* - Стандартный американский код обмена информацией. ASCII - это код для представления символов английского алфавита в виде чисел, каждой букве сопоставлено число от 0 до 127. В большинстве компьютеров код ASCII используется для представления текста, что позволяет передавать данные от одного компьютера на другой.

ASCII Code Chart

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

Макет документа

- Логическая (содержание)
  - Логическая структура – составные элементы

— уникальный элемент;* — повторяющийся элемент; ? — необязательный элемент; ! — обязательный элемент; & — вхождение типа «И»; | — вхождение типа «ИЛИ»

- Физическая (макет) –
  - описание документа в физических единицах: страница, полоса, колонка, колонтитул и др.

Подходы к моделированию документов опираются на два стандарта — ISO 8613 (ODA — Office Document Architecture — архитектура управленческой документации) и ISO 8879 (SGML — 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.

В системах обработки текстов в документ включается дополнительная информация, называемая р а з м е т к о й и выполняющая следующие функции:

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

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

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

- **Модель документа Microsoft Word** http://www.monographies.ru/ru/book/section?id=3330
- **Автоматическая классификация текстовых документов с использованием нейросетевых алгоритмов и семантического анализа** **http://www.inteltec.ru/publish/articles/textan/RCDL2003.shtml**
-
-

- Максимов Н. В.

- Современные информационные технологии: Учебное пособие / Н.В. Максимов, Т.Л. Партыка, И.И. Попов. - М.: Форум, 2008. - 512 с.: ил.; 60x90 1/16. (переплет) ISBN 978-5-91134-239-5
- про модель текстового документа
  стр 45