## Тема 1: Гипертекстовые способы хранения и представления информации. Информационные ресурсы Интернета

**Основные понятия:**

- Гипертекст, Гиперссылка, Гиперсвязь и Гипертекстовый документ;
- WWW и HTTP;
- HTML, XML.

### *1. Гипертекстовые информационные технологии*

***Гипертекст*** (нелинейный текст) – это организация текстовой информации, при которой текст представляет множество фрагментов с явно указанными связями между этими фрагментами.

Термин “гипертекст” ввёл Т. Нельсон, определивший его как:

- соединение текста на естественном языке с создаваемой компьютером возможностью интерактивного формирования внутри него новых ветвей;
- динамичную организацию нелинейного текста, который уже не может быть напечатан обычным образом на обычной странице.

Гипертекст можно рассматривать как своеобразную базу данных, организуемую в виде открытой, свободно наращиваемой и изменяемой сети, узлы которой (линейные тексты) соединяются пользователем. От обычной базы данных гипертекст отличается, прежде всего, тем, что в нём отсутствуют заранее заданные ограничения на характер связей и структуру.

В 1987 году фирма Apple выпустила первую для ПК гипертекстовую систему – пакет HyperCard для Macintosh.

Эффективно применять гипертекстовые технологии стало возможным с 1992 года, когда появился WWW и возникли веб-технологии.

**WWW** (сокращение от англ. “World Wide Web” – “мировая паутина”, “всемирная паутина” или “всемирная сеть”) – это глобальный механизм обмена информацией; информационная система и популярная служба Интернета. Это самое распространённое приложение Интернета.

Основой WWW являются протокол передачи гипертекстовых данных (HTTP) и язык гипертекста (HTML), т.е. гипертекстовые технологии.

**HTTP** – это гипертекстовый транспортный протокол для связи веб-серверов и веб-клиентов. Он предназначен для построения распределённых информационных сетей коллективного пользования, поддерживающих различные типы данных (текст, изображение, аудио- и видеоинформация) и загрузки веб-страниц (файлов).

Элементы гипертекста (текстовые фрагменты) называются узлами. Узлы, между которыми возможен переход, называют смежным, а возможность перехода – “связью”. Совокупность смежных узлов образует “окрестность” данного узла. В общем случае в качестве узла могут выступать: слово; словосочетание; предложение; абзац; параграф; документ; собрание документов, относящихся к одной теме; отдельные сообщения и т.п. Создание гипертекста, прежде всего, состоит в формировании системы переходов от узла к узлу (системы ссылок). Движение в гипертекстовой сети, совершаемое в процессе чтения гипертекста, называется “навигацией”.

Гипертекстовая технология реализуется в конкретной гипертекстовой системе, состоящей из гипертекста (базы данных) и гипертекстовой оболочки. Гипертекст содержит не только информацию, но также аппарат её эффективного поиска и просмотра. Путешествие по WWW (всемирной паутине) начинается с ввода электронного адреса в строку местоположения (Locator) и нажатия затем кнопки “Enter”.

Просмотр (браузинг) – это операция, характерная только для гипертекста. Он означает поиск информации посредством просмотра гипертекстовой сети, при этом возможно запоминание пути следования для того, чтобы при последующем аналогичном запросе поиск проходил по зафиксированному пути следования. Ныне чаще всего используются следующие программы-браузеры: Internet Explorer, Netscape Navigator, Opera и др.

Гипертекстовые технологии широко используются в:

- настольных издательских системах для создания документов большого объёма со свойствами гипертекста (т. е. с системой ссылок);
- системах управления документами (СУД), например, для сведения в один итоговый документ информации, содержащейся в разнородных документах;
- системах подготовки электронных документов, позволяющих составлять гипертекстовые документы с возможностью осуществления навигации и др.

Одним из перспективных направлений развития гипертекстовых систем является технология гипермедиа – соединение технологий гипертекста и мультимедиа (интеграция текста, графики, звука, видео).

Использование гипертекста позволяет создать информационную инфраструктуру распределённой организации и упростить диалоговый интерфейс пользователя, что важно при разработке информационных приложений. Гипертекстовые технологии способствовали созданию в информационных сетях информационных хранилищ, которым могут получать доступ самые различные категории пользователей. В результате в Интернете сформированы огромные и свободно доступные широким массам пользователей (открытые) самые разнообразные информационные ресурсы.

Основная идея гипертекстовых технологий состоит в том, что поиск документальной информации происходит с учётом множества взаимосвязей, имеющихся между документами, а значит, более эффективно, чем при традиционных методах поиска. При этом доступ к информации осуществляется не последовательным просмотром текста, как в обычных информационно-поисковых системах, а путём движения от одного фрагмента к другому.

В общем случае взаимодействие пользователя с гипертекстовой системой заключается в том, что он видит на экране компьютера некоторый текст и (или) графический объект и, в зависимости от возникающих у него ассоциаций, может выполнять ряд действий (главным образом переходов к другим, предлагаемым на данной странице, объектам).

***2. Языки гипертекстовой разметки документов***

Гипертекстовая технология изначально использовала специальный язык, получивший название HTML (Hyper Text Markup Language). Благодаря ему,

можно не только формировать гипертекстовые документы, но осуществлять связь текста и изображения с документами, расположенными на другом веб-сервере и др.

***Гиперссылка*** или ***Гиперсвязь*** (Hyperlink) – фрагмент HTML-документа, указывающий на другой файл, который может быть расположен в Интернете или содержать полный путь к этому файлу. Гиперссылка для пользователя – это графическое изображение, текст на сайте или в письме электронной почты, устанавливающие связь и позволяющие переходить к другим объектам Интернета.

**Гипертекстовый документ** - это файл, содержащий различные виды информации и имеющий в своей структуре ссылки (гиперссылки) на другие файлы или сам являющийся документом, на который есть ссылка в другом файле, расположенном на некотором сервере в любой точке планеты.

Гипертекстовый документ, размещённый на сервере с использованием WWW, называют ***Web-страницей*** (веб-страницей). Это минимальный фрагмент гипертекста, который можно за один раз загрузить и прочитать.

Используя предлагаемые связи можно читать материал в любом порядке. Таким образом, текст становится “открытым”. В него можно вставлять любые объекты, указывая для них связи с имеющимися объектами. При этом структура текста не разрушается. Поэтому гипертекст стал новой информационной технологией представления неструктурированного легко наращиваемого знания.

Структура гипертекстовых документов формируется с помощью последовательности *тегов* – элементов языка HTML, включающих тексты в формате ASCII. Они позволяют управлять шрифтом, цветом фона и текста, определять ссылки, вставлять объекты (графику, звук и видео). При этом сами тэги на странице в режиме её просмотра через браузер не видны.

**Тег** или **тэг** (англ. “Tag” – метка) является признаком объекта, который управляет соответствующим кодом.

Любая страница начинается с тега ***<HTML>*** и заканчивается тегом ***</HTML>***.

Легко заметить, что теги представляют собой некоторые операторы (команды, дескрипторы), заключенные в скобки типа “< >”, причём ими начинается и заканчивается целая страница или некоторый фрагмент. В конце команды ставится слеж “/” с именем, используемым в её начале. Таким образом, любая команда включает некоторое действие, а аналогичная команда со знаком “/” – выключает его. Это правило касается парных команд (дескрипторов). Существуют операторы, не требующие их выключения (закрытия).

В качестве примера приведём запись следующих заголовков “Министерство образования и науки Российской Федерации” и “Московская финнансово-промышленная академия”, последовательно размещённых на одной странице. Запись в файле может выглядеть следующим образом:

<HTML>

<HEAD>

<TITLE> Министерство образования и науки Российской Федерации </TITLE>

</HEAD>
<BODY>
Московская финнансово-промышленная академия
</BODY>
</HTML>

При этом оператор <TITLE> используется как заголовок страницы. Он не виден в программе просмотра гипертекстовых страниц, но отображается в верхней строке браузера. Обычно эта команда используется как идентификатор страницы и служит поисковым образом документа.

Оператор <HEAD> ограничивает область заголовка, включающую другие операторы, отражающие название программы, в которой создавалась страница, код страницы, ключевые слова (мета тэги) и др.

Оператор <BODY> включает область, где размещаются все основные команды, входящие в состав данной страницы. Все эти операторы парные и они обязательно включают команды их закрытия.

Ссылка обозначается парными тегами ***<a>*** и ***</a>***.

Некоторые теги могут быть разовыми, например, для обозначения перехода на другую строку в одном абзаце используется непарный тег ***<br>***.

Точно также внутри таблицы, образуемой парными тегами <***Table***> и <***/Table***>, могут присутствовать непарные теги <***tr***> и <***td***>, отражающие строку и ячейку в строке.

Для указания признака объекта и (или) его состояния к тегам добавляются атрибуты. Так, для ссылок в качестве атрибута используется адрес фрагмента текста в документе или собственно документа. Таким образом, команда на переход к другому ресурсу будет выглядеть следующим образом: ***<a href=”адрес”>имя ссылки</a>***.

Чтобы при установке мыши на пиктограмме появлялся поясняющий текст, следует в тэг (<***a href***=...) включить параметр (alt="текст пояснения"), например,

*<a href="help-1.htm" alt=“Помощь и полезные советы”>*.

В таком тексте не должно быть кавычек, символов “<” (меньше), “>” (больше), "/" и т.п. Их заменяют записями типа " < >.

Кроме того, используется ***CGI*** (Common Gateway Interface – общий шлюзовой интерфейс) – программа для связи HTML-браузеров с другими прикладными программами и (или) текстами, находящимися на стороне сервера.

Для создания гипертекстовых документов и веб-страниц применяют как обычное ПО, так и специальные программы. В простейшем случае для создания гипертекстовых документов можно использовать любые текстовые редакторы (Блокнот, Notepad, WordPad и др.), а также текстовые процессоры типа Word. Так, в MS Word97 легко и просто получить компактный гипертекстовый документ в формате “html”, выбрав для сохранения текстового документа именно этот тип формата. При этом не требуется знание HTML. В других последующих версиях того же MS Word также можно создавать подобные файлы, но при этом файл “обрастает” большим количеством порой совершенно ненужных тегов. В результате подобный файл, конечно же, можно использовать как веб-страницу и размещать на сайтах или порталах. Однако его размер становится в несколько раз больше оптимального. А это не очень хорошо, так как, во-первых, в несколько

раз увеличится объём хранимых на сервере страниц сайта или портала. Во-вторых, что более важно, значительно возрастёт время загрузки таких станиц на компьютеры пользователей. Последнее обстоятельство вызывает у пользователей нежелание обращаться к таким сайтам и порталам.

Заметим, что увеличение времени загрузки страниц на компьютеры пользователей зависит и от других причин. Например, важно не перегружать страницы, особенно графическими и мультимедийными данными. Кроме того, замедление загрузки файлов возникают и по техническим причинам, включающим низкое быстродействие компьютера и модема пользователя, а также самой сети, подключение к которой предоставляет интернет-провайдер. Данного недостатка порой можно избежать путём выбора более надёжного провайдера, предоставляющего более быстрый канал связи.

Обычно же для создания гипертекстовых документов целесообразно использовать специальные программы, в которые по мере необходимости загружать материалы для веб-страниц, сделанные в других программах, например, текст из Word, табличные данные из Excel, а графические объекты из Photoshop и PowerPoint. Такие программы позволяют ускорить создание страниц за счёт автоматизации ряда простых рутинных процедур, например, создания шаблона страницы и таблиц, изменения стилей и размеров шрифтов, цвета, включения наиболее часто используемых тегов и т.д., а также возможности не переходя в другую программу периодически просматривать полученные результаты. В качестве специальных программ можно использовать, например, FrontPage, Publisher, Hot Dog, HomeSite, Reamweaver, Noteped, SiteCreate и др.

На смену HTML пришёл новый язык разметки документов – **XML** (eXtensible Markup Language). Прогнозируется, что использование этого языка позволит значительно упростить процессы взаимодействия между информационными системами, обеспечить эффективное управление предприятиями, использовать системы электронной коммерции и др.

В отличие от HTML, он ориентирован не на форматирование, а на данные. XML позволяет создавать оригинальные теги, поддерживать произвольную структуру данных и обеспечивать проверку корректности документа.

XML не связан с типом хранимых данных. Независимость от СУБД организуется путём представления запросов к ним на языке XML Query.

***3. Информационные ресурсы Интернета***

Под *ресурсами Интернета* **(**англ. “Internet Resources”) понимается совокупность информационных запасов всей совокупности сетей, предоставляемых пользователю Интернета. Отметим, что к концу 2004 года в России насчитывалось около 15 млн., а в Москве – около двух млн. пользователей Интернета.

Цель размещения информационных ресурсов Интернета заключаться в качественном и оперативном предоставлении их пользователям, создании им возможности работы с разнообразной информацией, размещённой в различных регионах планеты. Для достижения этой цели, кроме создания эффективно действующих правовых механизмов, необходимо решить следующие задачи: классификации и систематизации обширной информации по отраслям знаний; определения способа нахождения необходимого пользователю ресурса. К

сожалению, полной классификационной картины информационных ресурсов Интернета пока не существует, но исследования в этом направлении ведутся.

Предлагаются различные их деления. В Интернете множество видов информационных ресурсов, отличающихся: тематикой (по археологии, музыке и т. п.), направленностью, содержанием, доступностью или охватом предлагаемой пользователям информации (например, ресурсы правительства).

*Тематические информационные ресурсы* обычно отражают коллекции художественных произведений, стихов, фантастики, фотографий и др. Они могут включать много или несколько различных тем, а также быть посвящёнными одной проблеме и при этом затрагивать различные типы, виды, содержание и др. Кроме общепринятых, к ним следует отнести БД рефератов, курсовых и дипломных работ.

*По виду информационные ресурсы Интернета* можно разделить на: документы, аналогичные традиционным (документы, периодические издания, книги, брошюры и т.п.); графические изображения (картины, фотографии, диаграммы и т.п.); интерактивные (банеры и т.п.) и мультимедийные ресурсы (звук, видео, анимация).

*По принадлежности и содержанию* можно выделить: государственные (Правительства РФ и Субъектов Федерации), территориальные (региональные), политические, военные, общественные, научные, научно-технические, правовые, коммерческие, образовательные, художественные, СМИ (новости, ТВ), развлекательные, рекламные и иные информационные ресурсы.

Существуют информационные ресурсы, представляющие организации, учреждения, фирмы, в том числе библиотеки и информационные службы.

*По назначению* информационные ресурсы можно разделить на: возрастные, гетерогенные, национальные, религиозные, территориальные, коммерческие, производственные, технологические, учебные, туристические и спортивные, развлекательные, справочные и др.

Ряд информационных ресурсов трудно отнести к какому-либо виду, например, появившийся в 1994 году частный сайт “Библиотека Мошкова” (http://lib.ru), представляет значительные собрания условно систематизированных электронных документов, в большинстве художественного характера.

Информационные ресурсы в Интернете размещаются на серверах и представляют сайты, порталы (государственные, политические, общественные, коммерческие, религиозные, учебные, частные и иные) и БД.

Для поиска в Интернете широко используют поисковые системы (поисковые машины). Это специальный веб-сайт, на котором пользователь по заданному запросу может получить соответствующие ссылки на сайты и документы. Поисковая система состоит из:

1) поискового робота;
2) индекса системы;
3) программы, которая: обрабатывает запрос пользователя, находит в индексе документы, отвечающие критериям запроса, и выводит список найденных документов в порядке убывания релевантности.

В мире насчитывается 45 тыс. поисковых машин для открытого (“поверхностного”) Интернета. Среди множества поисковых отечественных и зарубежных машин, предоставляющих информационные ресурсы в Интернете, отметим: AltaVista, Google и Рамблер, Яндекс и другие.

В общем случае информация, представленная на сайтах и порталах в Интернете – потребительские информационные ресурсы. По мнению специалистов в сети на порядок больше документов, которые скрыты от широкого круга пользователей, что на 99% видимое содержание (контент) сайтов не представляет никакой ценности. Однако цели и назначение предлагаемых информационных ресурсов, а также интересы пользователей многогранны и разделяются по различным критериям.

Среди информационных ресурсов, ориентированных на большую аудиторию пользователей, совершенно очевидно выделяются средства массовой информации (СМИ).

СМИ достаточно быстро осознав преимущества и достоинства Интернета, активно ведут в нём различные виды деятельности. Важным аспектом представления пользователям публикаций являются создаваемые в Интернете электронные газеты и журналы. Они могут отражать материалы существующих печатных оригинальных изданий или быть полностью электронными, т.е. не имеющими печатных аналогов.

В середине 1995 года около 300 коммерческих газет работали или планировали работу с электронными видами услуг (из них 199 находились в США и 55 в Европе). Это электронные доски объявлений, электронные информационные службы (в том числе новостей), электронные газеты и сайты. Например, сайт *Global Network Navogator* (http://www.gnn.com) является одной из первых публикаций в WWW и содержит множество интересных соединений, путеводителей, тематических статей. *New York Times* (http://nytimesfax.com) – сетевой вариант газеты “New York Times”, а GAZETA.RU (http://www.gazeta.ru) – электронная газета Российских политических новостей и т.д.

Одним из представителей электронных изданий, не имеющих печатных аналогов, является научный электронный журнал “Электронные библиотеки” (www.iis.ru/el-lib/).

Также многообразна развлекательная информация в Интернете. Кроме текстовых и графических данных, в Интернете можно слушать музыку, смотреть фильмы и мультфильмы, играть в игры и др.