

Въведение в XML. Общ преглед на XML технологиите и на приложението на XML



Интернет и Уеб

Web 1.0/2.0/3.0

Маркър езици

От SGML през HTML до XML

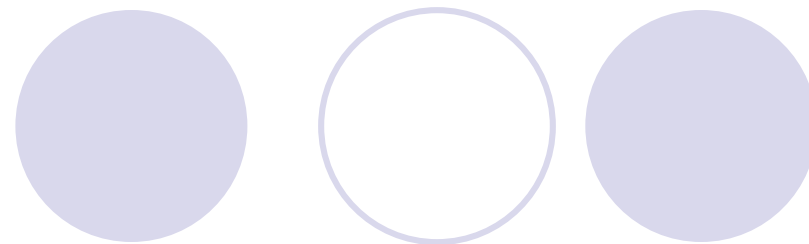
Цели на XML

XML езици и технологии

Интернет и Уеб

- Роля на World Wide Web (Световна паяжина), съкр. WWW, Уеб или Световната мрежа
- Нов начин на общуване между хората, общностите и дори начините за комуникация между компютрите
- Услуги за е-бизнес, електронна търговия, е-икономика, електронно правителство, електронна демокрация, електронно обучение, ...
- Информационен обмен и взаимодействие между различни актьори – лица, организации, Уеб приложения, интелигентни агенти, Интернет на нещата, Голем(е)ми данни...

Интернет и Уеб



- ❖ World Wide Web често се счита за синоним на Интернет, но това не е така
- ❖ Интернет - електронна съобщителна мрежа и структура, на която се базира WWW
- ❖ Уеб - част на Интернет, достъпна посредством графичен потребителски интерфейс и съдържаща приложения, услуги и документи, които често са свързани чрез хипервръзки

Развитие на Уеб

- Тим Бърнърс-Лий, Церн, 1991 г. - разработва основите на Уеб
- Проектира прости средства за пренос на взаимосвързани документи със структурирана информация до всякакви компютри, свързани в Интернет и работещи с различни операционни системи :
 - ✓ език за маркиране на хипертекст (**Hypertext Markup Language, или съкр. HTML**), и
 - ✓ протокол за трансфер на хипертекст (**Hyper Text Transfer Protocol, или съкр. HTTP**)
- По-късно - спецификация на URI (**Uniform Resource Identifier**) - нотация за уникално идентифициране на обекти в целия Интернет.

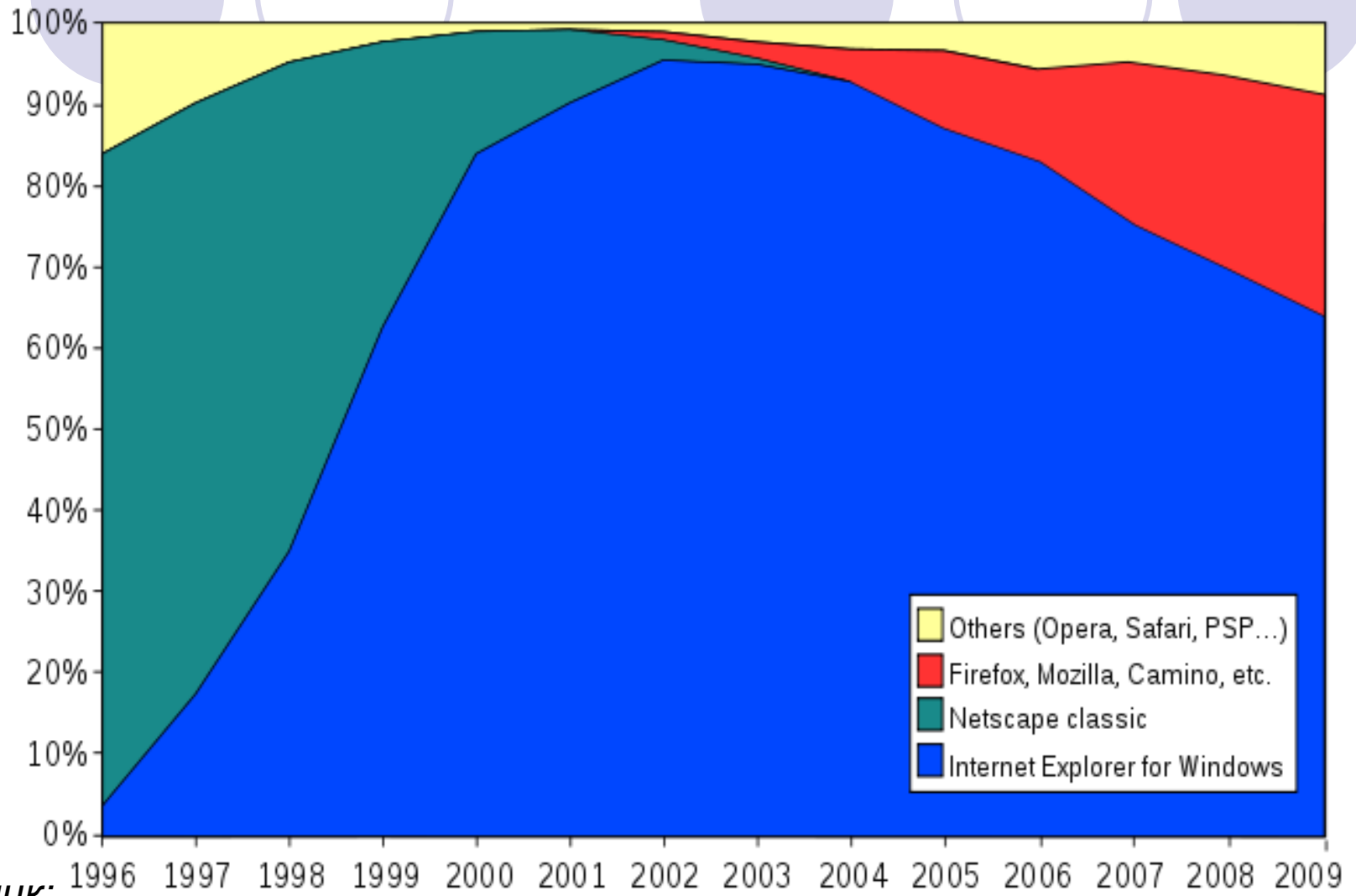
Исторически възход

- 1992 год. - повече от един милион компютри вече са свързани към Интернет
- 1993 год. - свободна експлоатация на първия браузер NCSA Mosaic
- 1994 г. - Tim Berners-Lee основава **World Wide Web Consortium (W3C)** в Massachusetts Institute of Technology, Laboratory for Computer Science (MIT/LCS) с поддръжка от **Defense Advanced Research Projects Agency (DARPA)**, чиято ARPANET е основата на съвременния Internet
- 1994 год. - Netscape Navigator - "well on its way to becoming the world's standard interface", според Gary Wolfe в статия в сп. *Wired*

Войни на браузърите

- 1995 год. - Microsoft лицензира Mosaic като Internet Explorer 1.0, пуснат на пазара като част от Microsoft Windows 95 (август 1995)
- Ноември 1995 г. - Microsoft Internet Explorer 2.0 е пуснат за свободна употреба
- 1996 г. - Netscape 3.0 с първа в света поддръжка на стилове - Cascading Style Sheets (CSS)
- 1997 г. - Internet Explorer 4 е интегриран в Microsoft Windows
- Нови видове браузъри с поддръжка на разширени (макар често несъвместими) версии на HTML заливат пазара – виж http://upload.wikimedia.org/wikipedia/commons/7/74/Timeline_of_web_browsers.svg
- Февруари 2011 г. - Internet Explorer 45%, Firefox 30%, Chrome 17%, Safari 5%, Opera 2%, други браузери - около 1%
(източник: StatCounter global market)

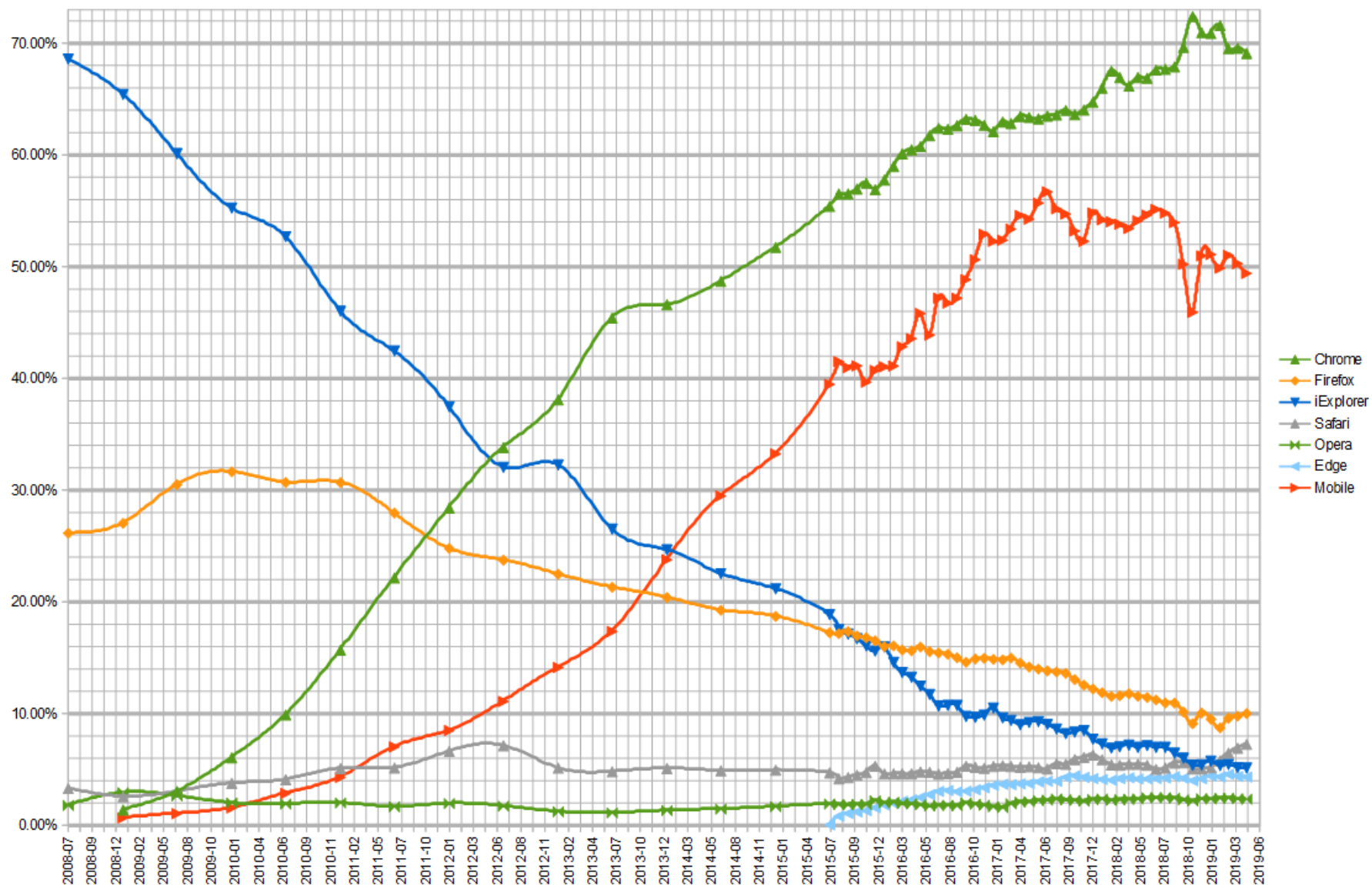
Browser Wars



Источник:
Wikipedia

XML

Usage share of browsers (source StatCounter)



Последствие от войните на браузърите

- ❑ Надпревара в предлагането на нови функционалности и по-бавно коригиране на проблемите
- ❑ Добавяне на уникални за браузера нови маркери (тагове) на съдържание вместо придържане към стандартите – напр. `<marquee>` в Microsoft's Internet Explorer и `<blink>` в Netscape
- ❑ Изоставане в развитието на стандарта HTML, поддържан от W3C
- ❑ Проблеми със сигурността – зареждане на вируси при зареждане на документи с активно съдържание (документи-приложения)

Справка за нестандартен HTML

- ❖ Непрепоръчителни (deprecated) и патентовани (proprietary):
 - Елементи
 - Атрибути
- ❖ Поддръжка от различни layout engines за съответните браузъри -
[http://en.wikipedia.org/wiki/Comparison_of_layout_engines_\(Non-standard_HTML\)](http://en.wikipedia.org/wiki/Comparison_of_layout_engines_(Non-standard_HTML))

Етапи в развитието на Уеб

- Web 1.0
- Web 2.0
- Web 3.0
- ...

Уеб 1.0 (традиционен Уеб) 1/3

- **Web 1.0** - термин за етап от еволюцията на World Wide Web; обхваща периода от 1993 до 2001г.
- **Бизнес модел:**
 - top-down подход за изграждане и използване на WWW - статични страници с хипертекст (Hyperlinks е стандарт на WWW от 1993г.)
 - страници на малцина автори на съдържание (webmasters), зареждани от голям брой потребители с глобален достъп;
 - фокус върху презентацията, а не върху създаването на съдържание;
 - печалби от броя посещения (most visited webpages)

Уеб 1.0 (традиционен Уеб) 2/3

- **Технически характеристики:**

- статичен хипертекст – без динамика в браузера
- липса на редактиране на страниците от външни потребители
- използване на Framesets - рамката (frame) е начин за представяне на няколко Уеб страници и/или медия елементи в един прозорец (или таб) на брауъра:
 - характерен за HTML 3 и 4;
 - липса на поддръжка от много брауъри, лоша индексация от търсачките, трудни връзки и bookmark към рамкираните страници, лошо скролиране при ниска резолюция;
 - изключени от HTML 5

```
<frameset cols="65%, 35%">
  <frame src="URL OF FRAME PAGE 1">
  <frame src="URL OF FRAME PAGE 2">
  <noframes> Sorry but your browser do not support frames ☹
</noframes>
</frameset>
```

Уеб 1.0 (традиционен Уеб) 3/3

- **Още технически характеристики:**

- използване на таблици (<table>) за подравняване на съдържанието на страницата
- отделяне на съдържание с прозрачни 1x1 pixel изображения в GIF format
- патентовани нестандартни HTML елементи като <blink> и <marquee>
- онлайн книги за гости
- изпращане на HTML форми като ел. поща от статичен хипертекст
- сървърни технологии като PHP, Ruby, Perl, Python, JSP, and ASP.NET

Уеб 2.0 (социален Уеб) 1/3

- **Web 2.0:**

- ✓ термин за втория етап от еволюцията на World Wide Web;
- ✓ от началото на века до наши дни;
- ✓ въведен от Tim O'Reilly на Web 2.0 conference през 2004

- **Бизнес модел:**

- ✓ top-down + bottom-up подход за изграждане и използване на WWW - динамични хипертекст страници с авторско съдържание и на самите потребители;
- ✓ добавена стойност от споделянето на информация и сътрудничеството между организации и хора;
- ✓ фокус върху създаването на съдържание и персонализирана презентация

Уеб 2.0 (социален Уеб) 2/3

- **Технически характеристики:**

- ✓ динамичен хипертекст
- ✓ съдържание от външни потребители
- ✓ модел "Network as platform" - потребителски интерфейси за достъп до разл. услуги като напр. публични сайтове с галерии на потребителя (Flickr, Picasa Web Albums, ...), частни и споделени хранилища за данни (DropBox), споделени документи (Goodle Docs), представяне на геогр. обекти върху карта (Google Mail API), ...
- ✓ оперативен обмен на данни (interoperability)
- ✓ Rich Internet Application (RIA)

Уеб 2.0 (социален Уеб) 3/3

- **Нови технологии:**

- ✓ клиентски (client-side/web browser):
 - XML или JSON (JavaScript Object Notation)
 - asynchronous JavaScript (Ajax)
 - Adobe Flash
 - Adobe Flex
 - JavaScript/Ajax frameworks като jQuery
 - HTML5 - изисква по-малко изчислителни ресурси отколкото Adobe's Flash; по-малко ел. мощност (батерия при мобилни устройства); замразяване на публичните мобилни Adobe's Flash приставки (plugins)
- ✓ сървърни

Социални феномени в Уеб 2.0

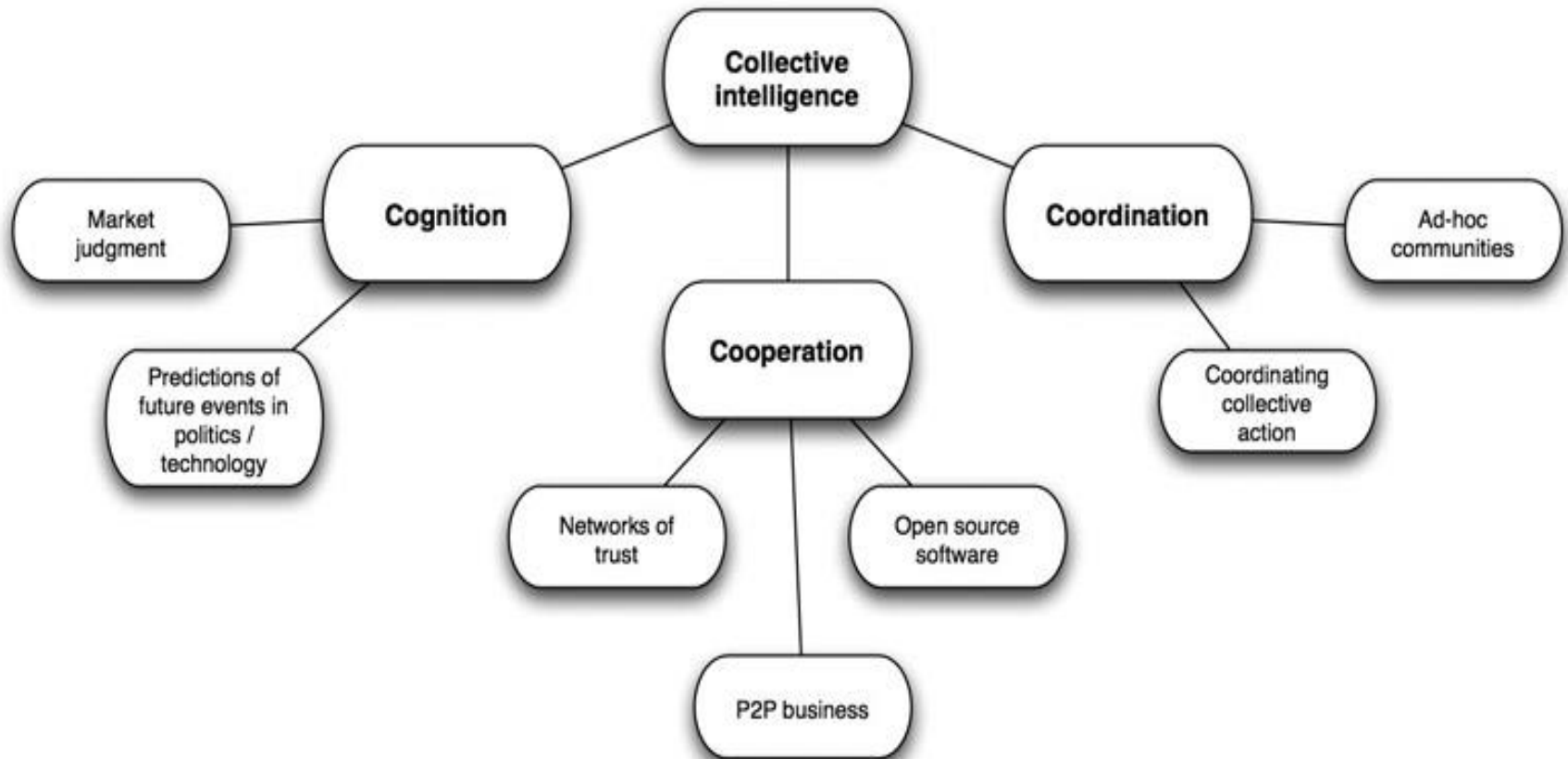
- **Podcasting**
- **Bloggging**
- **Tagging**
- **Folksonomy**
- **Social bookmarking**
- **Social networking**

Опишете ги...

Социални феномени в Уеб 2.0

- **Podcasting** - от broadcast и (i)Pod - сваляне на онлайн видео или аудио съдържание от настолни или мобилни компютри
- **Blogging** - web log - поддържане на личен журнал, публикуван в Уеб на дискретни порции (т.нар. posts), показвани в ред, обратен на хронологичния
- **Tagging** - добавяне на метаданни (описания с ключови думи и термини) към съдържание или части от него, с цел да се ползват при търсене и разглеждане (browsing)
- **Folksonomy** (social tagging) - много потребители добавят метаданни като кл. думи към споделени ресурси - Golder, Scott; Huberman, Bernardo A. (2006). "Usage Patterns of Collaborative Tagging Systems". Journal of Information Science 32 (2): 198–208.
- **Social bookmarking** - социални отметки - организиране, поддръжка и търсене на отметки към онлайн ресурси
- **Social networking** - сътрудничество и съревнование в Уеб

Колективна интелигентност



- *Източник:* Olga Generozova, по книгите 'The wisdom of crowds' и 'Smart mobs'

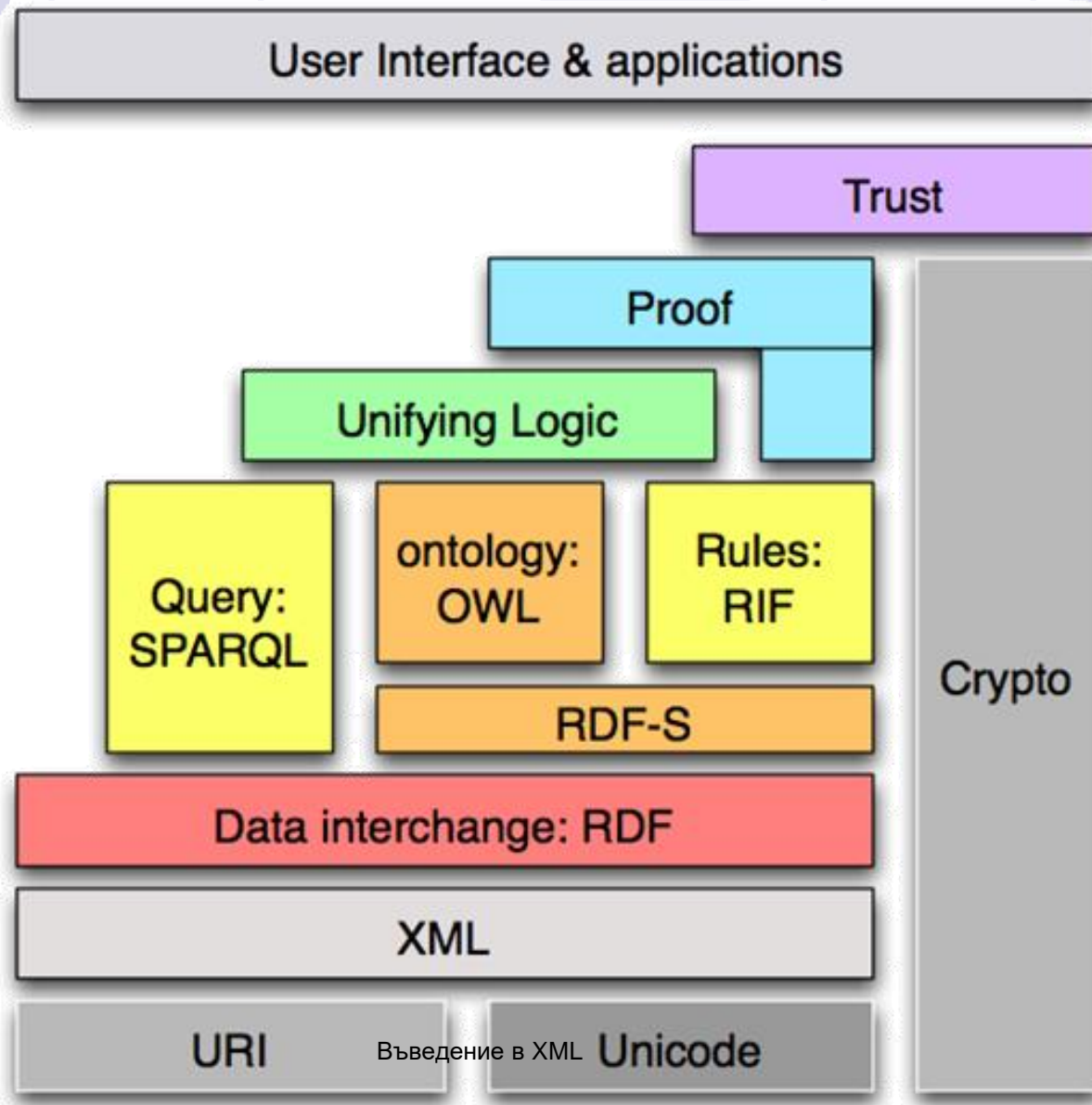
Уеб 3.0 (семантичен Уеб) 1/2

- **Еволюция и преход от сегашното състояние на Световната мрежа към семантичен Уеб**
- ✓ Семантични услуги – базирани на онтологии за представяне на знанието за дадена предметна област:
 - Семантично аотиране на съдържание
 - Семантично търсене
 - Семантично разглеждане
 - Семантично препоръчване
 - ...

Уеб 3.0 (семантичен Уеб) 2/2

- **Семантични езици – базирани на XML**
 - ✓ Resource Description Framework (RDF) - за описание на модела на метаданните относно Уеб ресурси
 - ✓ RDF Schema
 - ✓ SPARQL Protocol and RDF Query Language - език за заявки към RDF графи
 - ✓ Web Ontology Language (OWL) - фамилия от езици за представяне на знания чрез онтологии
 - ✓

Стек на семантичния Уеб (W3C, 2006)

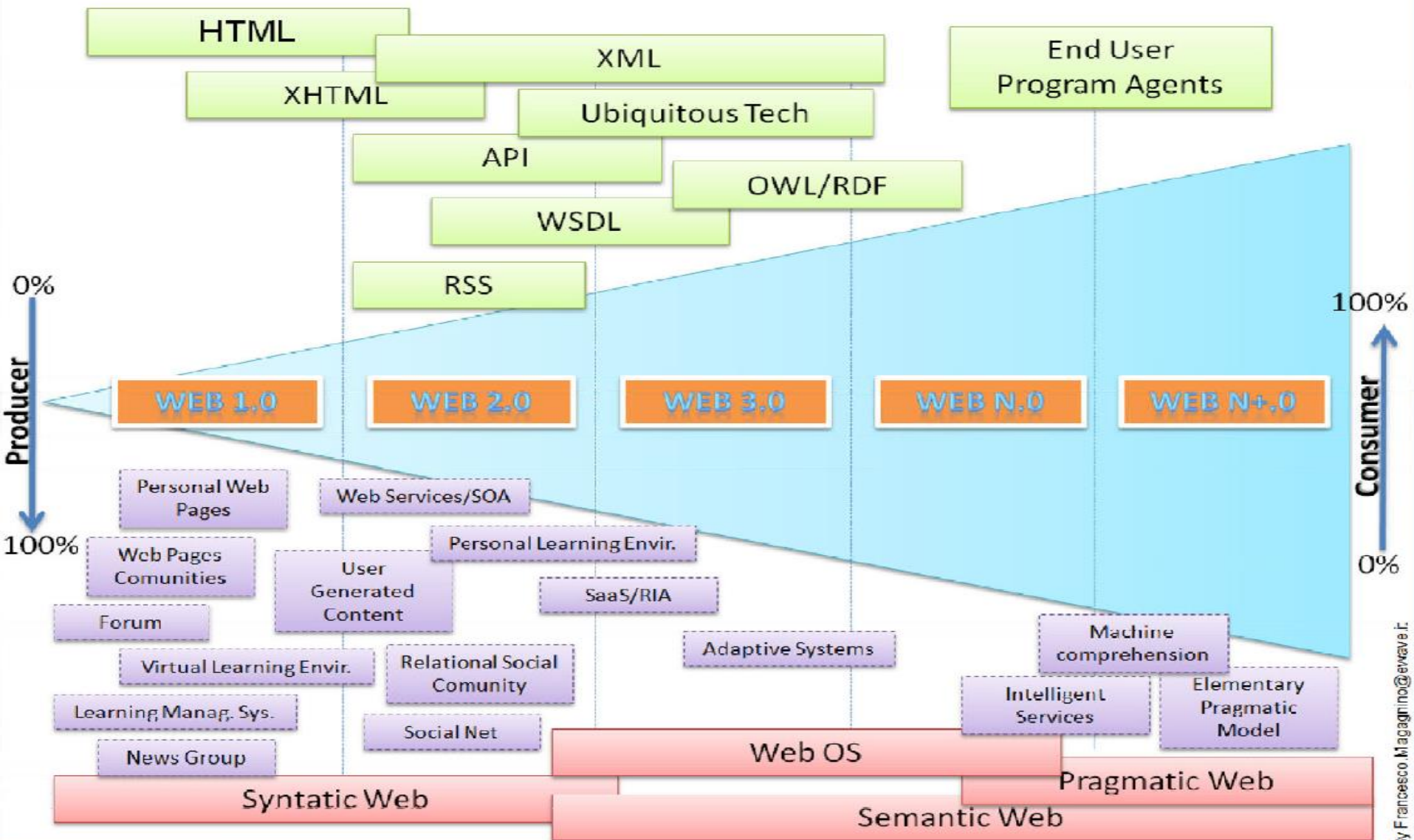


Адаптирано от http://en.wikipedia.org/wiki/Semantic_Web_Stack

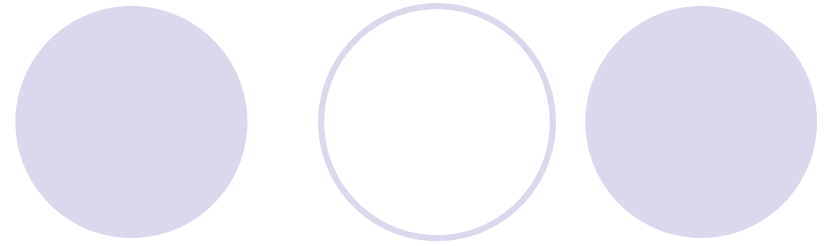
XML

Въведение в XML Unicode

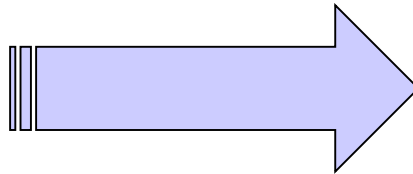
Към Уеб N.0??



Период на преход



● HTML

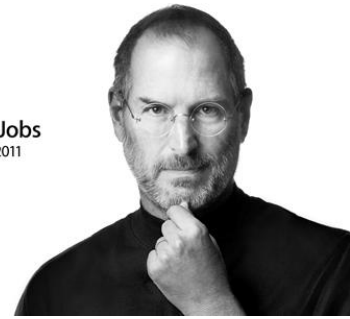


● XML

Струва ли си?

Steve Jobs, 1955-2011

Steve Jobs
1955-2011



- “That’s been one of my mantras — **focus and simplicity. Simple can be harder than complex:** You have to work hard to get your thinking clean to make it simple. But it’s worth it in the end because once you get there, you can move mountains.”

[BusinessWeek, May 25, 1998]

See also:

http://www.ted.com/talks/steve_jobs_how_to_live_before_you_die.html



ML = Markup Language = език за маркиране

- ❑ Маркъп езиците (езици за маркиране) възникват на един сравнително късен етап.
- ❑ Понятието markup се дефинира от популярния речник Уебстър като подробни инструкции, които обикновено се записват в ръкописа на дадено типографско издание относно типа и стила на шрифта, начина на странициране и други печатни оформления (1915-1920 год.)
- ❑ Интуитивен пример за такъв маркъп може да бъдат ръчно записани коментари от коректора на изданието относно правописни грешки или граматически и стилови корекции, включващи изтриване и добавяне на текст, прехвърляне на съдържание на друго място, представяне на думите в курсив, и др.

Езици за маркиране

- Езиците за маркиране (*markup languages*) служат за описание на данни относно структурирането, форматирането, значението и начина за представянето на тези данни.
- Самото маркиране представлява добавяне на допълнителна стойност към основното съдържание – от гледна точка на неговото оформление като структура, значение и представяне.
- По този начин общият обем на текста се увеличава с допълнителен, маркиращ текст, включен в документа и разграничен от основното съдържание по определен начин.

Черти на езиците за маркиране

- **СТИЛИСТИЧНИ** - отнасящи се до външния вид, или представянето на текстовото съдържание - напр. таговете в HTML: **<I> <U>**
- **структурни** - определящи макета на структурата (layout) - напр. таговете в HTML: **<H2> <P>
**
- **семантични** - служещи за определяне на смисъла на съдържанието - напр. таговете в HTML:
<TITLE> <META NAME=keywords CONTENT = " " >
- **функционални** - задаващи определено действие представяне на съдържанието в дадена медия - напр. таговете в HTML:
<BLINK> Щракнете тук</ A>

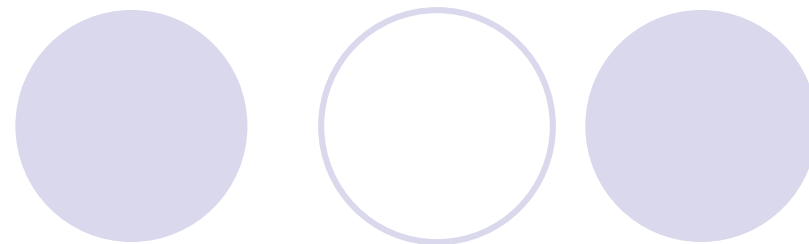
Компоненти на език за маркиране

- Пример 1: Hello, world!
 - = **start tag**
 - = **end tag**
 - Hello, world! = **content (or data)**
 - < > = **delimiter characters (identify tags)**
- Hello = **element**
 - element = **tags** + **content**
- Пример 2:
 - name/value pairs identify **attributes** and give **values**

Какво представлява XML?

- XML 1.0 е препоръка на W3C (World Wide Web Consortium: <http://www.w3.org>) от 10.02.1998
- XML наследява Standard Generalized Markup Language (SGML) - през 1986 год. SGML е утвърден като стандарт ISO 8879 и се налага в световен мащаб като език за структурно, стилистично и семантично маркиране на текстови документи.
- Специално за стилистичното и семантичното маркиране към него е разработен стиловия (stylesheet) език DSSSL (Document Style Semantics and Specification Language).
- SGML е много тежък и труден за употреба

XML е метаязык



- XML, както и SGML, е *метаязык* за маркиране.
- Това е разширяем език, който позволява на потребителя да създава таговете (респ. инструкциите за маркиране), които са му необходими за дадена предметна област, и да ги използва заедно.

HTML



- ❖ HTML е приложение на SGML и поради това е подобен на него, но е значително по-опростен и ограничен.
- ❖ За разлика от SGML, HTML не е разширяем.
- ❖ HTML 4.01 спецификацията определя 91 елемента за използване в HTML документи, като нови елементи не могат да бъдат създавани от потребителя.
- ❖ Крайните тагове могат да бъдат пропуснати за 15 от дефинираните от спецификацията елементи, което показва друга отличителна черта на HTML в сравнение с SGML - липсата на строги синтактични правила.
- ❖ Езикът е прост, лесен за научаване и използване за гъвкаво представяне на съдържание в Уеб браузер.

Недостатъци на HTML при използването му в Уеб 1/2

- HTML не е разширяем
- HTML не позволява представянето на отношения (релации) между елементи и в такъв смисъл има "плосък" синтаксис. HTML представя само двуизмерни таблици, а многомерни таблици трябва да се изобразяват по отделните измерения. Също така, няма как да бъдат представени релациите
- **name = име + фамилия**
- **client = име + адрес (улица + номер + код и др.)**
- Не е възможно да се представи ефективно структурата на документа, макар че таговете за стил като <H1>, <H2> и др. могат да се тълкуват като структурни тагове, което е обаче крайно недостатъчно.

Недостатъци на HTML при използването му в Уеб 2/2

- HTML притежава много малко възможности за описание на значението на съдържанието. Използването на елементи като **<TITLE>** и **<META NAME="keywords" CONTENT="metadata ..." >** се оказва крайно недостатъчно.
- Повечето тагове на HTML имат чисто процедурен характер – те описват какво трябва да направи браузъра при представяне на съдържанието, напр. **
** за пренасяне на нов ред, **<hr>** за извеждане на хоризонтална линия, **<i>** за показване на текста в курсив.
- Това определя HTML като процедурен, а не като дескриптивен език за маркиране.

XML



- През 1996 г. Джон Босак, Тим Брей, Джеймс Кларк и няколко други Уеб специалисти започват работа по създаването на олекотена версия на SGML
- W3C утвърждава разработката на този език, който не притежава както недостатъците на както на SGML, така и на HTML, през февруари 1998 г. под името:

XML (Extensible Markup Language) 1.0

- През последните години развитието и използването на XML се очерта като най-перспективното направление за развитие на Уеб. XML е основният двигател на процесите на прерастването на съвременния Уеб в Семантична мрежа.

Предимства на XML

- XML е разширяем мета език за дескриптивно маркиране
- Валидиране на типа на документа - XML и всички XML-базирани езици са формални и имат формални граматики
- Оперативен обмен на данни.

A screenshot of an Internet Explorer browser window. The title bar reads "C:\Documents and Settings\Boyan\My Document...". The address bar shows the file path "Lbook\examples\example1.xml". The main content area displays an XML document with the following structure:

```
- <client>
+ <name>
- <address>
  <street>Oborishte St.</street>
  <streetNo>57</streetNo>
  <city>Sofia</city>
  <postCode>Oborishte St.</postCode>
</address>
</client>
```

XML спрямо HTML



Database publishing

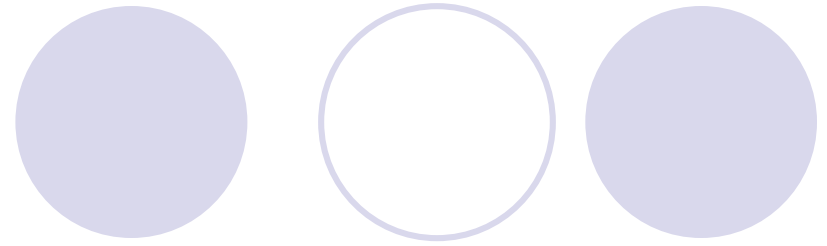
Integration of data from different sources

Extended usability of metadata

Multi-directional and –dimensional links

Reusability of information

XML йерархии



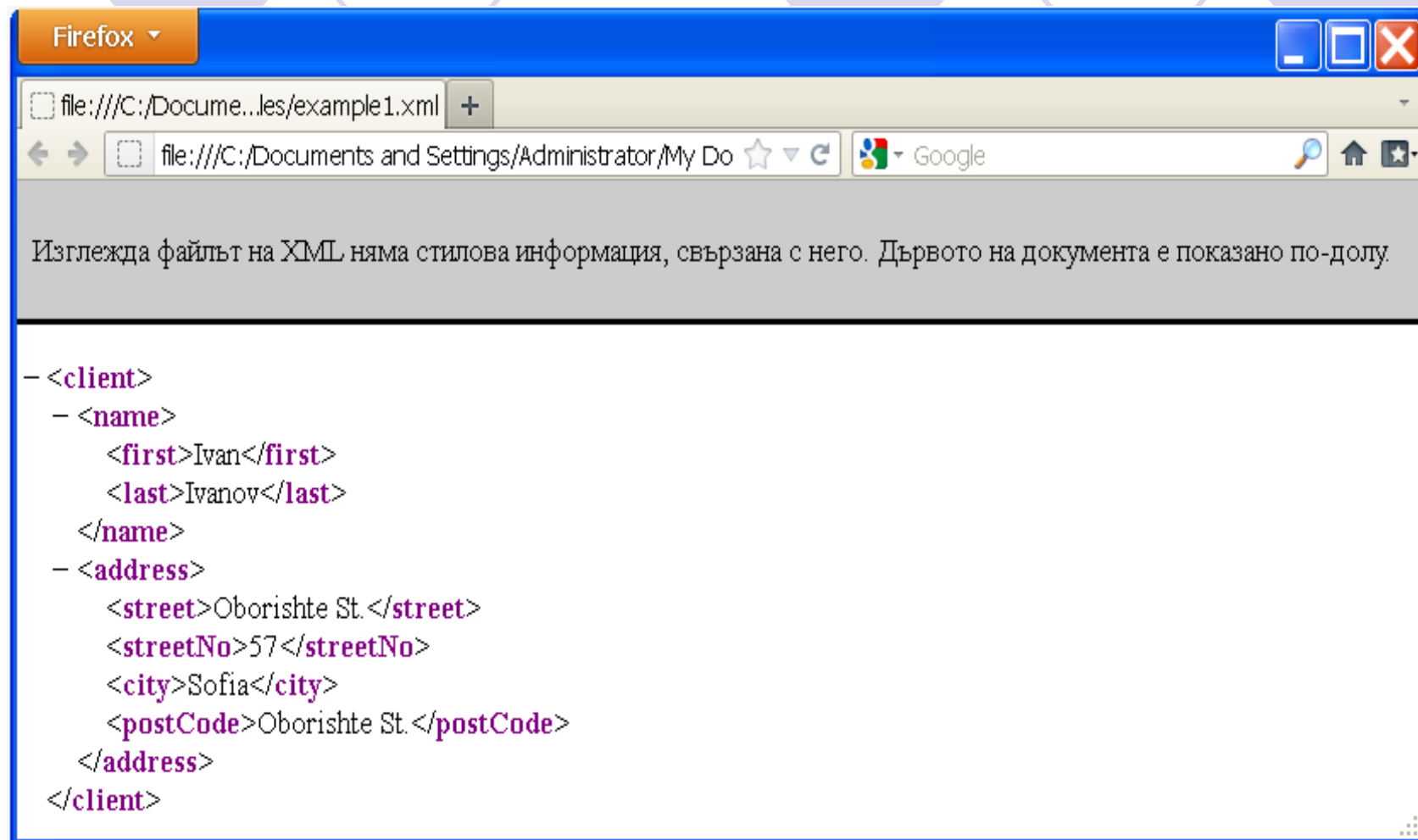
- textbooks (root element)
 - book (repeating element)
 - title
 - ISBN
 - authors
 - author_name (repeating element)
 - description
 - price

Създаване на XML документи

- Подобно на HTML, но без ограничения за елементите и атрибутите им:

```
<?xml version="1.0"?><!--required-->  
<books><!--root element-->  
  <textbooks>  
    <btitle>Beginning ASP 3.0</btitle>  
    <!--other items-->  
  </textbooks>  
</books>
```

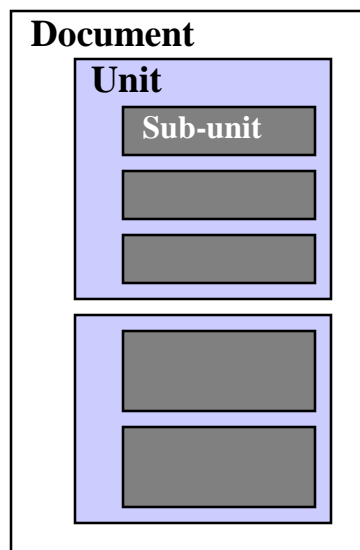

XML в браузър (без стилове)



Структура на XML документ

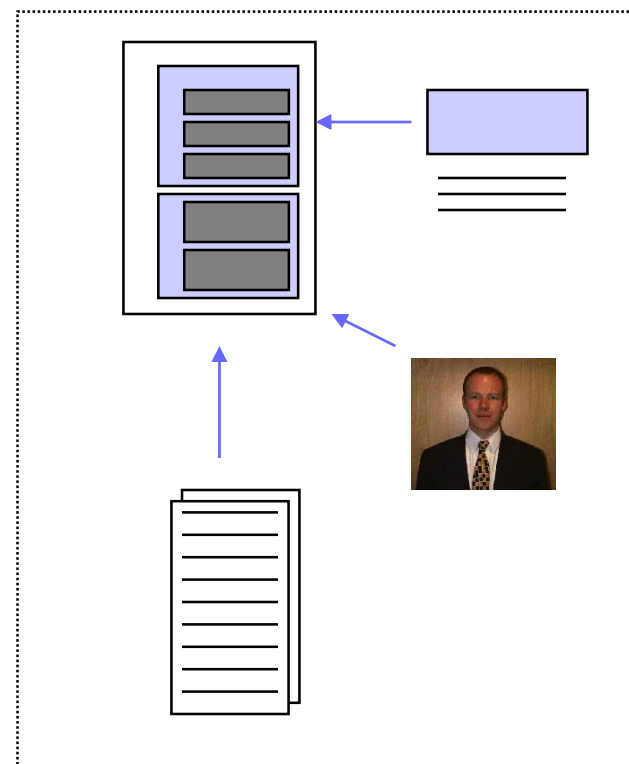
- Логическа структура

- Елементи



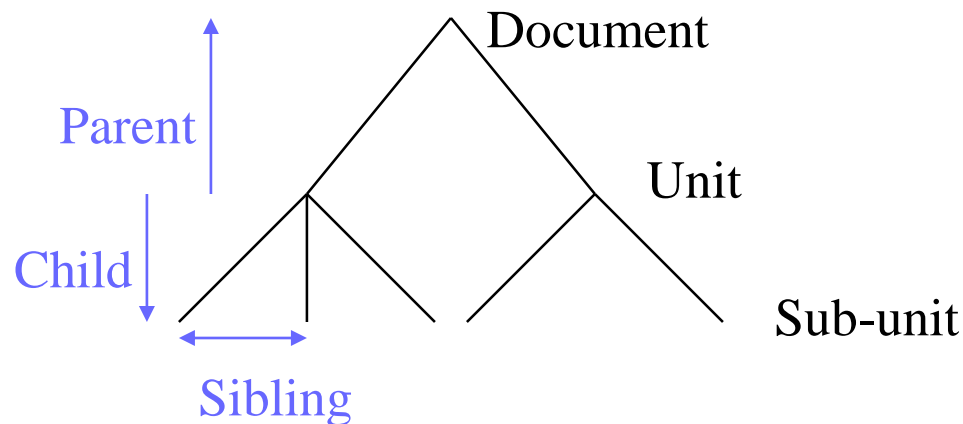
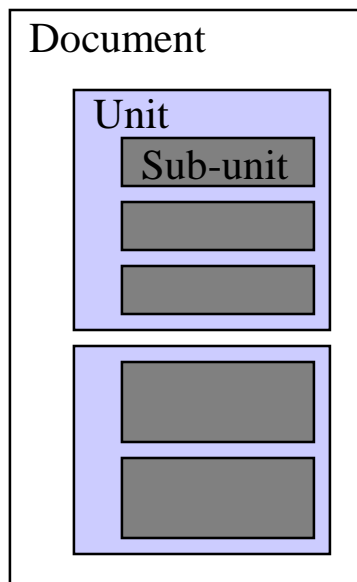
- Физическа структура

- Единици (Entities)



XML йерархия

Планарно и дървовидно представяне



N.B. All elements must be nested

Средства за работа с XML

- Минимум

- Text editor and XML parser (e.g. msxml-Parser by Microsoft)
- XML-Browser (XML, XSLT) (e.g. Internet Explorer)

- Повече

- XML-Editor (e.g. XML Spy by Altova)
(<http://www.xmlspy.com>)
- XML-Browser with XSLT (IE 6)
- XSL:Fo Formatter (Antenna by Antennahouse)
(<http://www.antennahouse.com>)

XML технологии

- DTD
- XML Schema
- XLink and XPointer
- XSL and XSLT/XSL-FO
- XQL
- XPath
- Namespaces
- SAX
- DOM
- StAX
- JAXB
- Други ... 😊