Surfing inside the Web

From SGML to HTML … and back

Hans C. Arents

s.a. OFFIS n.v.

“Office Future International Services”
Atlas Park, Weiveldlaan 41 B. 32, B-1930 Zaventem, Belgium
Tel: +32 (0)2 725 40 25 - Fax: +32 (0)2 725 40 12
Email: info@offis.be - Web: www.offis.be
From SGML to HTML ... and back

- Introduction
- A nice try: HTML
- The real thing: SGML
- HTML or SGML?
- Going beyond SGML
- How and where to use SGML?
- But what's really next?
- Conclusions
Introduction: HTML vs SGML

- The World-Wide Web: the world’s largest and most successful SGML application
  - now more than 35 million documents
  - still doubling in size every 6 months
  → growing awareness of the possibilities of SGML

- The Web language: HTML
  - Hypertext Markup Language
  - focus on linking and presentation
  - HTML pragmatics: guided by looks
  vs
  SGML purists: guided by contents
What is HTML?

**Hypertext Markup Language**

- an Internet RFC (Request For Comments)
- an *Internet “standard”*
  for hypermedia document access and display
- a *DTD* written in SGML
  for creating hypermedia documents
  and accessing them on the World-Wide Web
- goals:
  - specify the content and presentation of hypermedia documents
  - specify *simple* hyperlinking and *basic* interactive behaviour
  - define document addressing and locating mechanisms
HTML documents

n What?
– Web documents in their “standard” format
– using open Internet standards

n Why?
– support full Web functionality
  l hyperlinks, multimedia, interactivity, …
– simple and intuitive graphical user interface
– free or inexpensive clients / servers for document delivery

n Why not?
– HTML is a continuously moving target
  l NS Navigator extensions, MS Internet Explorer extensions, ...
– HTML is a presentation format, not a real data storage format
HTML capabilities

- content
  - text
  - media
    - images, sound, 3D, …
  - scripts
    - JavaScript, Visual Basic Script
  - objects
    - Java applets, ActiveX controls
**HTML capabilities**

- **content**
  - text
  - media
    - images, sound, 3D, ...
  - scripts
    - JavaScript, Visual Basic Script
  - objects
    - Java applets, ActiveX controls

- **presentation**
HTML capabilities

- contents
  - text
  - media
    - images, sound, 3D, ...
  - scripts
    - JavaScript, Visual Basic Script
  - objects
    - Java applets, ActiveX controls

- presentation

- hyperlinking
HTML capabilities

**contents** = HTML \((\text{Hypertext Markup Language})\)
- recently approved version 3.2
- improved image and table support
- in the future: embedding / controlling objects

**presentation** = CSS \((\text{Cascading Style Sheets})\)
- new standard for Web style sheets
- specify fonts, set margins, change colours, ...
- in the future: control page layout (columns, margin text, …)

**hyperlinking** = URLs \((\text{Universal Resource Locators})\)
- remains a simple addressing mechanism
- still no support for serious hyperlink management
- in the future: hopefully results from the work on URIs
**HTML browsers**

- **Netscape Navigator 4.0**
  - availability of more than 50 plug-ins
  - support for Java and JavaScript
  - support for JSS and layered HTML
  - 75% market share (but falling)

- **Microsoft Internet Explorer 3.0**
  - support for ActiveX and Visual Basic Script
  - support for Java and JScript
  - support for CSS and layout control
  - 20% market share (but rising rapidly)

   "the future de facto standard?"
So what's the problem?

Limitations of HTML

- no validation of document structure
- navigational links are difficult to generate
- document dependencies are hard to maintain
- tools are hardwired to a particular version of HTML
- document contents cannot be restructured or reused

HTML is *bad* (broken as designed)

- object granularity is fixed and file-based
- "link rot" is endemic and ever increasing
  - fragile addressing scheme
  - addressing based on identification of instances, not on the abstract naming of objects
So what's the problem?

n “HTML is like a baby SGML, but it is a baby born without a brain”
   – content cannot be marked up for its meaning
   – tags cannot be extended for specific uses
   – hyperlink behaviour cannot be modified

n Result: the Web is a world-wide electronic library
   – where books have no ISBNs,
   – where there are no bibliographic records,
   – where there is no agreed set of subject descriptors,
   – and where the only librarian available has committed suicide
What is SGML?

Standard Generalized Markup Language
ISO 8879:1986

- an international standard
  for electronic document interchange

- a meta-language
  for formally specifying different markup languages
  - HTML is just a (very simple) example of such a language

- goals:
  - system and application independent data storage format
  - support document exchange and data longevity
  - capture the meaning of document contents
The SGML view of a document

A document is a combination of:

- **content**
  - the actual data inside a document
  - which information objects are present

- **structure**
  - the logical organization of a document
  - how information objects relate to one another

- **presentation**
  - the look and feel of a document
  - how information objects are visually presented

- **structure ≠ presentation**!
Structure ≠ presentation!

layout markup styles

- font: Helvetica
  - size: 18 pt
  - justification: left

- font: Palatino
  - size: 10 pt

- font: Helvetica
  - size: 12 pt
  - weight: bold

- font: Palatino
  - size: 8 pt

DOCUMENT

logical markup objects

- document title
- level 1 heading
- paragraph
- numbered list
- list element
- address
Strengths and shortcomings of SGML

**n strengths**

- standard immune to software vendor politics
- markup under *your* control, for *your* documents
- simplifies administration of document repositories
- create a document once, publish in many formats
- industrial-strength tools are readily available

**n shortcomings**

- creation of documents (conversion/authoring)
- DTD creation and validation
- dealing with presentation issues
- handling of graphics and multimedia
- solutions have to be made to measure
What is HyTime?

**Hypermedia/Time**-based Structuring Language
ISO/IEC 10744:1992

- an *international standard*
  for hypermedia document interchange
- a *meta-DTD* written in SGML
  for specifying locations for addressing
  in document logical structure and time or space

**goals:**
- system and application independent link specification
- support inter/intra document and multimedia linking
- capture the "relatedness" of document contents
**HyTime concepts**

- **locators**
  - specifying locations for addressing
    - addressing by name
    - addressing by relative location
    - addressing by position in a coordinate space
    - specifying a sequence of locations (“location ladders”)

- **architectural forms**
  - adding information to elements
    - adding attributes to elements
    - adding relationships between elements
    - defining “object-oriented” inheritance between elements
HyTime links

n a hyperlink in HyTime:
- can link to other documents
- can link to other hyperlinks
- can be late-binding
  (rendered at presentation time)

n hyperlinked documents can be:
- read-only
- HyTime or non-HyTime
- unstructured (non-SGML) or SGML

n a hyperlink is an association of document locations
- can be used to define Topic Maps
HyTime links

LINKING

independent link

role #1 role #2 role #n

anchor #1 anchor #2 anchor #n

aggregate address address address

LOCATING
What are Topic Maps?

- express a set of *relationships* between *topics* (portions of information with a common semantics)

- used for:
  - cross-document indexes and glossaries
  - virtual tables of contents
  - knowledge bases
  - thesauri

- advantages:
  - can be created above existing documents, without altering the documents themselves
  - can add meaning to structured or non-structured documents
What is DSSSL?

Document Style Semantics and Specification Language

n an international standard
for electronic document interchange

n a meta-DTD written in SGML
for describing document presentation
and transformations of document structure

n goals
  – system and application independent document presentation
  – system and application independent representation of
document structure (tree of elements and attributes)
**DSSSL concepts**

n *document presentation*
   - document has associated style sheet
   - tag has associated style
     1. fonts
     1. colours
     1. positioning
   - tag has associated presentation instructions

n *structure transformation*
   - from one set of element attributes into another
   - from one document style sheet into another
   - from one document structure into another
From: <B>GDT</B>
To: <B>MDL</B>
Subject: <I>Results</I>
Keyword: <A HREF="http://www.lib.be/catalog/biology/rDNA/">rDNA</A>

The first test results in our rDNA manipulation …
How to use SGML?

Using HTML as an output format:
- **SGML** = the “back-end” *content markup* language
- **HTML** = the “front-end” *presentation markup* language
- batch conversion or on-the-fly generation

What you need:
- document design tools
  - designing page layouts, linking and navigation strategies
  - SoftQuad *HoTMetaL Pro*, InContext *Spider*, ...
- document downtranslation tools
  - converting SGML to HTML 3.2, NS-N HTML, MS-IE HTML
  - Exoterica *OmniMark*, Sema Group *Mark-It*, ...
Using HTML as an output format

Web server ➔ HTML ➔ Web browser

flat HTML ➔ flat HTML ➔ flat HTML

s.a. OFFIS n.v.
Using HTML as an output format

- Document database
  - Rich SGML
  - Downtranslate to HTML
  - Web browser
    - Print to paper
    - Printer
    - Convert to proprietary format
      - CD-ROM
Using HTML as an output format

- extending the functionalities of the Web server
  - on-the-fly creation of HTML
  - context-sensitive search engine
  - automatic creation of reliable links
  - on-the-fly creation of tables of contents
  - automatic chunking of large documents
  - centralized management of SGML data

Electronic Book Technologies *DynaWeb*

- *DynaWeb* server
  - integrates with Microsoft and Netscape Web servers
  - delivers SGML functionality in an HTML browser
DynaWeb\textsuperscript{(tm)} Server Benefits

\textit{Search for DynaWeb produced 52 hits.}

\begin{enumerate}
\item \textbf{Terminology} \ 3
\item \textbf{DynaWeb Makes Web Publishing Easy} \ 24
\begin{enumerate}
\item DynaWeb and DynaTag Support Existing Authoring Software \ 4
\item DynaWeb Supports All SGML Document Types \ 2
\item DynaWeb Optimizes Large Document Access Automatically \ 4
\item DynaWeb Makes it Easy to Track HTML as it Evolves \ 2
\item Table of Contents Views Are Generated Automatically \ 1
\item Non-redundant Fulltext Indices Are Generated Automatically \ 1
\item Publications With Common Content Can Be Supported by the Same Source \ 2
\item DynaWeb Hyperlinks are Easier to Maintain \ 3
\item One Publishing Process Serves all Mediums and all Platforms \ 1
\item DynaWeb is a Commercially Supported Software Package \ 1
\end{enumerate}
\item \textbf{DynaWeb Makes Web Access More Efficient for End Users} \ 13
\begin{enumerate}
\item DynaWeb TOC Views Expand and Collapse on Demand \ 2
\item DynaWeb Supports Fast Fulltext Searching Across Collections of Large Publications \ 2
\item DynaWeb Supports Context-Sensitive Fulltext Queries \ 5
\item End Users Don't Have to Master a New Fulltext Query Syntax \ 1
\item Hit Counts are Displayed in the TOC View \ 1
\end{enumerate}
\end{enumerate}
How to use SGML?

**n** Using SGML as a Web data format:
- SGML documents separate from HTML documents
- SGML file is just another downloadable data format
- find a balance between HTML and SGML functionality

**n** What you need:
- Web browser extensions
  - Navigator plug-ins or Internet Explorer ActiveX controls
  - SoftQuad HoTMetaL Intranet Publisher
- SGML-aware browsers
  - helper applications: external viewers
  - SoftQuad Panorama Publisher
Using SGML as a Web data format

- extending the functionalities of the Web browser
  - one-to-many links
  - context-sensitive search
  - different document views
  - dynamic tables of contents
  - user-defined HTML extensions
  - annotations and pop-up windows

- SoftQuad *HoTMeta*l *Intranet Publisher*
  - HiP Viewer
    - add-on for Microsoft Internet Explorer and Netscape Navigator
  - HiP Content Creator
  - HiP Publisher & Site Manager
Conflict of Interest and Confidentiality

Conflict of Interest

Although rare, there may be occasions at work when employees may be drawn into potential conflict of interest situations. A need to be alert is therefore paramount. As a general definition, an employee's interests conflicts with those of Exemplar where he/she profits, or places him/herself in a position to profit, directly or indirectly, through a misuse of the company's position. Therefore it is unacceptable to appropriate Company property, sell or trade on company information or accept rebates, fees or commissions from suppliers. Conflicts of interest may be subtle and sometimes it is just a matter of degree between an acceptable and unacceptable activity. As a rule however, no employee who is in a position to make or influence a decision regarding a business transaction between Exemplar and a third party should accept anything of substantial value from that party. Further clarification relating to to what is, and what is not acceptable is available from Human Resources.

The key points relating to Conflict of Interest are:
In terms of references for prospective employees, Human Resources staff are responsible for providing this information. First, we confirm that the candidate has the necessary experience and qualifications. Then, we verify the candidate's previous employment and contact their former employers. We refer to our employment references system, which contains detailed records of each employee's performance, which we maintain for the period of their employment. The references will be current as of the candidate's last employment and their date of reference. If an employer is not available, we will contact any alternative contact person provided by the candidate and report any status changes to you. A summary of the candidate's employment particulars, their references, and any other relevant material will be provided to you. Any questions or concerns you have can be addressed to Sandy Black.
Where to use SGML?

- **High value documents (heavily linked)**
  - e.g. scientific encyclopedia, reference works, ...
  - collections of documents converted to SGML + HyTime
  - SGML is used in its full richness, using SGML-aware browsers

- **High value documents (lightly linked)**
  - e.g. course material, scientific journals,...
  - well-structured documents converted to rich SGML
  - SGML is downtranslated to HTML, for use in Web browsers

- **Low value documents**
  - e.g. progress reports, lab notes, ...
  - simple documents converted to flat HTML
  - HTML managed using a Web site management tool
Where to use SGML?

- **High value document (heavily linked)**
  - WfW doc to SGML
  - SGML database
- **High value document (lightly linked)**
  - WfW doc to SGML
  - Document database
  - SGML gateway
- **Low value document**
  - WfW doc to HTML
  - Document server
  - Web site manager
  - Search engine
  - Document database
  - File system
  - Web server
  - Web browser

**From SGML to HTML ... and back**

© 1997 s.a. OFFIS n.v.
But what's really next?

If SGML is so great, why hasn't it taken over the world already?
- it has taken over the world
- the world is not yet ready for SGML
- we are waiting for XML to take over the world

XML (eXtensible Markup Language)
- a leaner, meaner subset of SGML for use on the Internet
- features of the SGML elephant which have been cast to the wolves:
  - need for a DTD
  - tag minimization
  - white space rules
  - ...

s.a. OFFIS n.v. From SGML to HTML … and back
© 1997
The XML timeline

What needs to be done:

- Phase I: a specification for XML
  - draft ready at SGML '96 Conf. (Boston, November 1996)
- Phase II: a specification of XML hyperlink mechanisms
  - draft ready by the 6th WWW Conf. (Santa Clara, April 1997)
- Phase III: a specification of XML stylesheet mechanisms
  - draft ready by SGML '97 Conf. (Washington, December 1997)

What has been done:

- the draft XML specification will go final by the end of March
- a prototype XML parser is already available on the Internet
- major SGML vendors are rumored to be working on tools
- a lot of excitement in the SGML world, but not (yet) outside it
Conclusions

- SGML is strong where HTML is weak
  - capturing meaning of information
  - handling complex, dynamic information
  - targeted towards the information provider

- HTML is strong where SGML is weak
  - low start-up, rapid return
  - ubiquitous, cheap and simple tools
  - targeted towards the information user

- XML may be the final answer
  (but does everybody understand the questions?)
  - meaning before content before presentation