Document engineering for the Intranet

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

Document engineering for the Intranet

- Intranet documents
  - native / HTML / dead / live
- Creating Intranet documents
  - authoring / conversion
- Managing Intranet documents
  - corrective / preventive
- Searching Intranet documents
  - search engines / web spiders
- HTML or SGML?
- The future of Intranet documents

Intranet documents

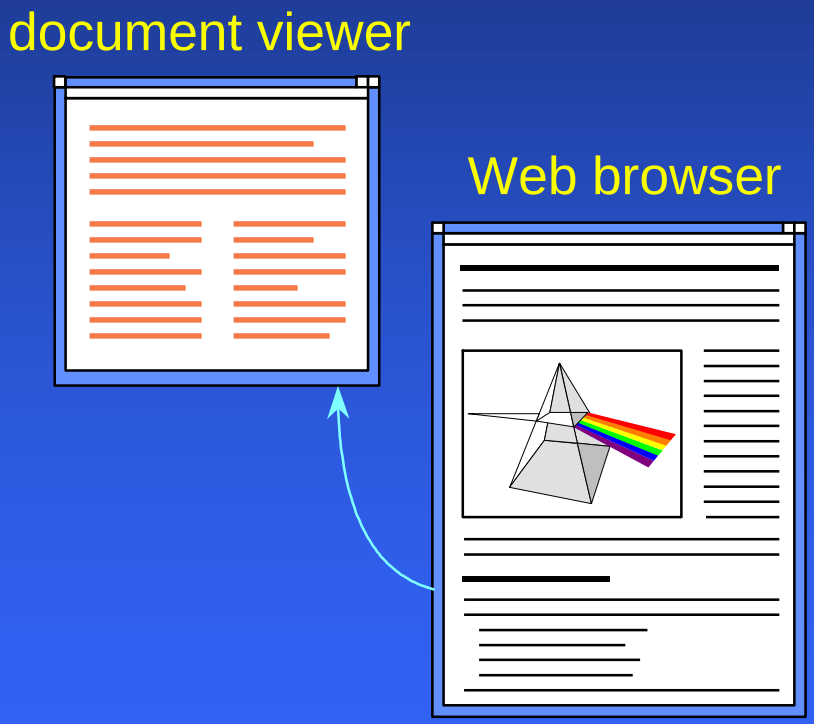
- Intranet = a tool for document delivery
  - deliver documents “on demand” / “just in time”
  - guarantee accurateness / timeliness of information
- Types of Intranet documents:
  - *native* = documents in original format
  - *HTML* = documents with full Web functionality
  - *dead* = document contents is created only once
  - *live* = document contents changes very frequently
- Documents are the foundation for groupware
  - integration with e-mail and newsgroups, workflow support, ...

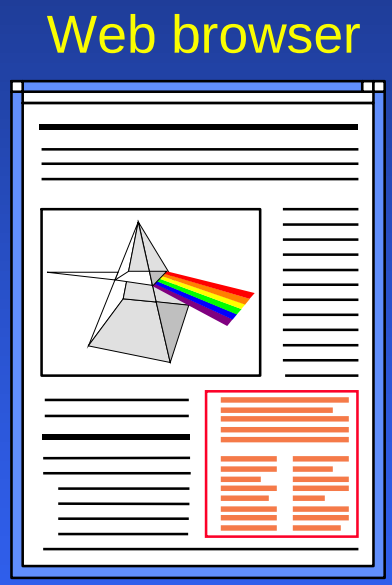
Native format documents

- What?
  - documents in their proprietary format
  - using closed vendor formats
- Why?
  - familiar document production tools
  - no training or support costs
  - guaranteed delivery
- Why not?
  - limited configurability and no extensibility
  - do not exploit full Web functionality
  - held hostage by the vendor

Viewing native format documents

- helper applications
  - viewers *outside* the browser
  - off-the-shelf applications
- inline plug-ins
  - viewers *inside* the browser
  - custom-built extensions





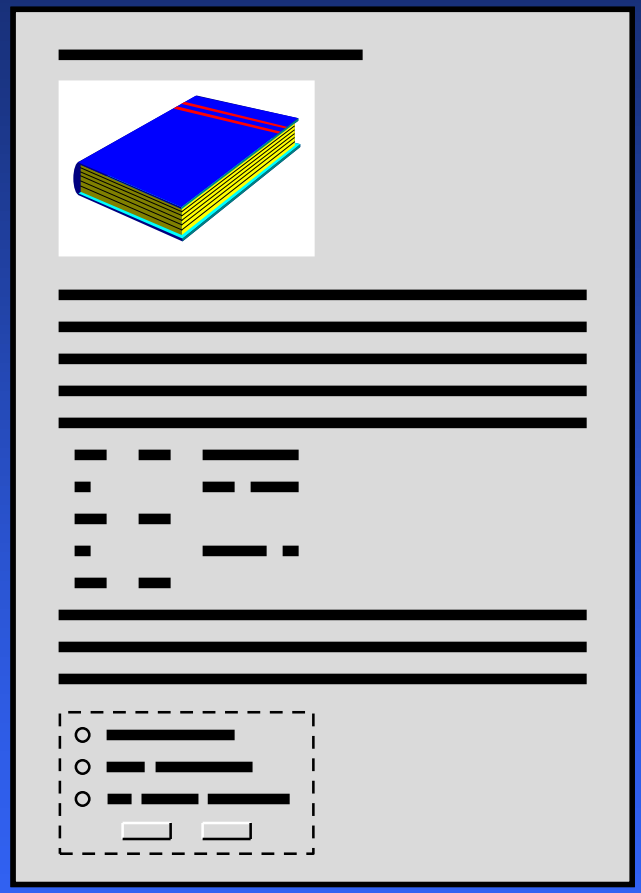
Document viewers and plug-ins

- document viewers
  - Adobe *Acrobat Reader*, MS *Viewers* for Word, Powerpoint, ...
  - **C** complete viewing, annotation, printing functionality
  - **D** not optimized for on-line delivery
- document plug-ins
  - Adobe *Amber*, Tumbleweed *Envoy, ...*
  - **C** optimized for on-line delivery
  - **D** not stand-alone, but dependent upon Web browser
- considerations:
  - document delivery is free
  - document creation can be expensive
  - time-consuming installation & maintenance

HTML format documents

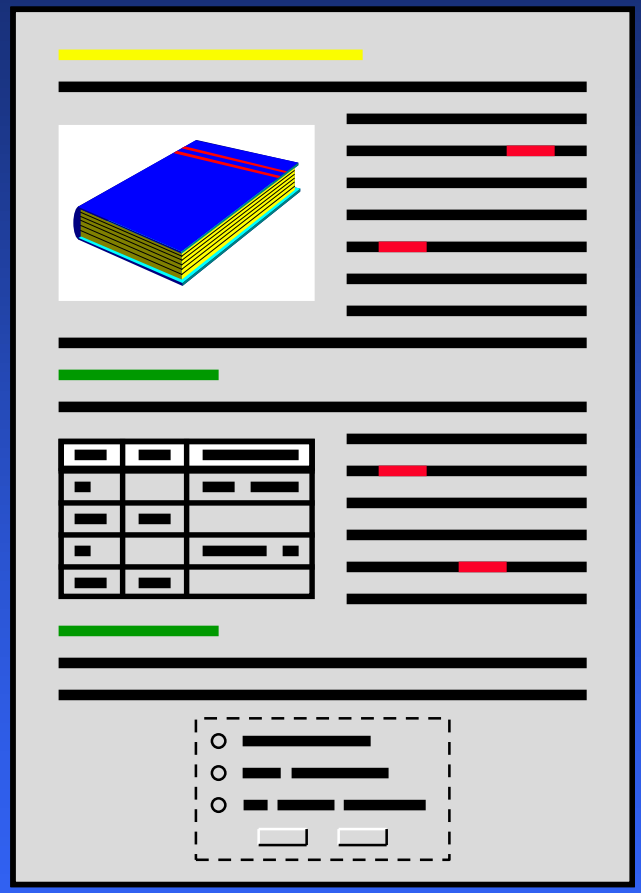
- What?
  - Web documents in their “standard” format
  - using open Internet standards
- Why?
  - support full Web functionality
    - hyperlinks, multimedia, interactivity, …
  - simple and intuitive graphical user interface
  - free or inexpensive clients / servers for document delivery
- Why not?
  - HTML is a moving target
    - NS Navigator extensions, MS Internet Explorer extensions, ...
  - HTML is a presentation format, not a real data storage format

Viewing HTML format documents



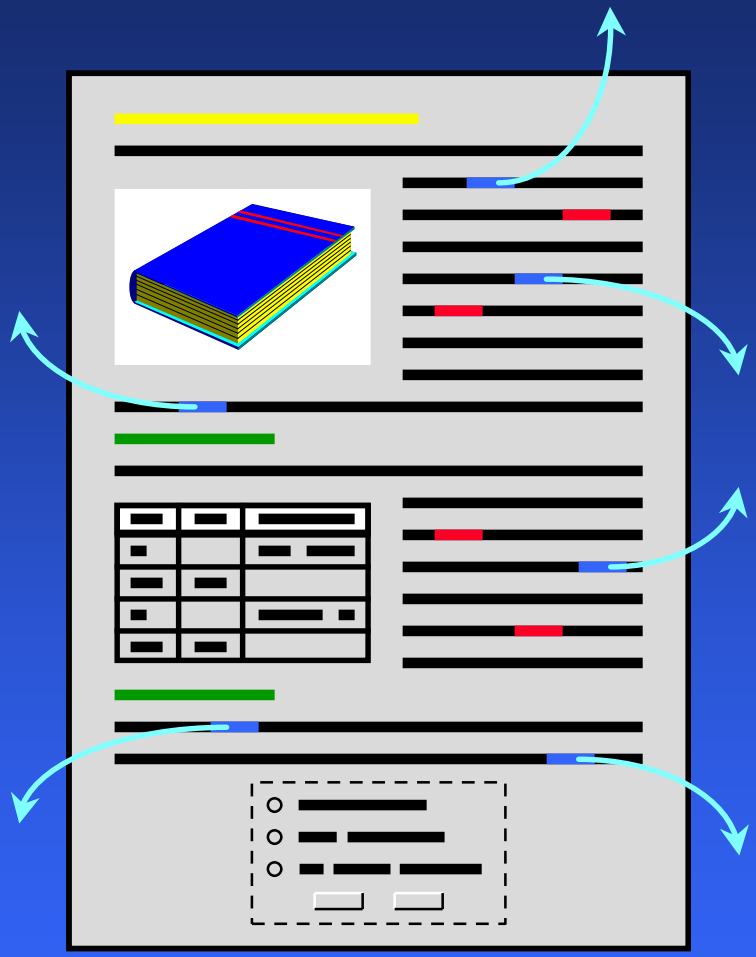
- contents
  - text
  - media
    - images, sound, 3D, …
  - scripts
    - JavaScript, Visual Basic Script
  - objects
    - Java applets, ActiveX controls

Viewing HTML format documents



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

Viewing HTML format documents



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

Viewing HTML format documents

- contents = HTML (**H**yper**t**ext **M**arkup **L**anguage)
  - recently approved version 3.2
  - improved image and table support
  - in the future: embedding / controlling objects
- presentation = CSS (**C**ascading **S**tyle **S**heets)
  - new standard for Web style sheets
  - specify fonts, set margins, change colours, ...
  - in the future: control page layout (columns, margin text, …)
- hyperlinking = URLs (**U**niversal **R**esource **L**ocators)
  - remains a simple addressing mechanism
  - still no support for serious hyperlink management
  - in the future: hopefully results from the work on URIs

Web browsers

- Netscape Navigator 3.0
  - C availability of more than 50 plug-ins
  - C support for Java and JavaScript
  - C more than 80% market share
  - D lost their HTML focus
  - è the present de facto standard
- Microsoft Internet Explorer 3.0
  - C support for ActiveX and Visual Basic Script
  - C future support for Java and JavaScript
  - C support for CSS and layout control
  - D only 10% market share
  - è the future de facto standard?

Dead or live Intranet documents

- “dead” Intranet documents
  - Intranet is a means of accessing a document repository
  - documents are created once and then stored forever
  - focus is on *consulting* documents

  examples: newsletters, tutorials, procedure manuals, ...
- “live” Intranet documents
  - Intranet is a means of keeping track of business processes
  - documents are continuously created, modified, deleted
  - focus is on *sharing* documents

  examples: project reports, design specs, product sheets, ...
- in practice: *a mix between “dead” and “live”*

Creating Intranet documents

- authoring
  - creating new documents from scratch
  - static documents: created manually
    - writing contents, specifying presentation, defining hyperlinks
  - dynamic documents: created on-the-fly
    - designing template documents, standardized navigational aids
- conversion
  - getting existing documents on the Web
  - uptranslation: from poorer to richer format
    - legacy documents, OCR documents, ...
  - downtranslation: from richer to poorer format
    - DTP documents, database publishing documents, ...

Authoring of Intranet documents

- Do-it-yourself HTML editors
  - Sausage *HotDog Pro*, Nesbitt Software *WebEdit*
  - è low-level tag editing, authors have to know HTML well
- Wysiwyg HTML editors
  - Adobe *PageMill*, SoftQuad *HoTMetaL Pro*
  - è Web desktop publishing, hide HTML from the authors
- Web site HTML editors
  - Adobe *SiteMill*
  - MS *FrontPage*
    - WebWizards = automate page creation
    - WebBots = automate script installation
  - è Wysiwyg, basic link maintenance and Web site management

Conversion to Intranet documents

- add-ons to conventional document applications
  - – MS *Internet Assistants* for Word, Excel, Powerpoint, ...
  - **C** hardwired conversion of simple business documents
  - **D** “quick and dirty” conversion
- off-the-shelf conversion applications
  - – InfoAccess *HTML Transit*, Stattech *Epublish Internet*
  - **C** interactive conversion of wordprocessor documents
  - **D** “best try” conversion
- custom-built conversion tools
  - – Exoterica *OmniMark*, AIS *Balise*, Sema *Mark-It*, ...
  - **C** batch conversion of large amounts of legacy documents
  - **D** “do it yourself” conversion

Managing Intranet documents

- managing contents
  - keep contents up-to-date
  - maintain complete version history
  - author supervision / reader notification
- managing presentation
  - maintain a consistent look for the whole site
  - associate a specific look to specific documents
  - create site maps, what's new, what's changed, ...
- managing hyperlinking
  - avoid undefined or "dangling" links
  - allow migration of documents and Web sites
  - support abstract naming and addressing conventions

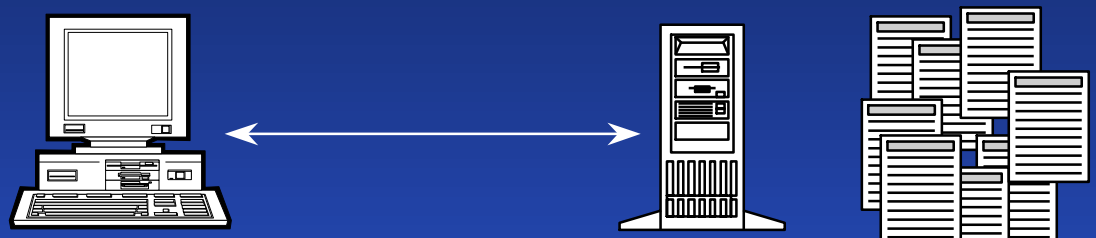
Corrective Web site management

- What?
  - manage contents and verify hyperlinks “post-mortem”
  - on top of existing Web server file system
  - edit HTML files, view thumbnail images
  - hyperlink verification engine
- Used where?
  - fast-growing, high-turnaround Web sites
  - documents created manually
- Products
  - InContext *WebAnalyzer*

Preventive Web site management

- What?
  - manage contents and verify hyperlinks “before birth”
  - on top of (object-)relational document database
  - drag-and-drop editing, version control, re-use
  - hyperlink resolution engine
- Used where?
  - mission-critical document applications
  - documents created on-the-fly
- Products
  - Electronic Book Technologies *DynaBase*
  - Inforium *LivePage WebMaster*, Arachnid Software *WebPower*

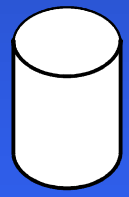
Search engines



Web browser

Web server
with search engine

native and HTML
documents

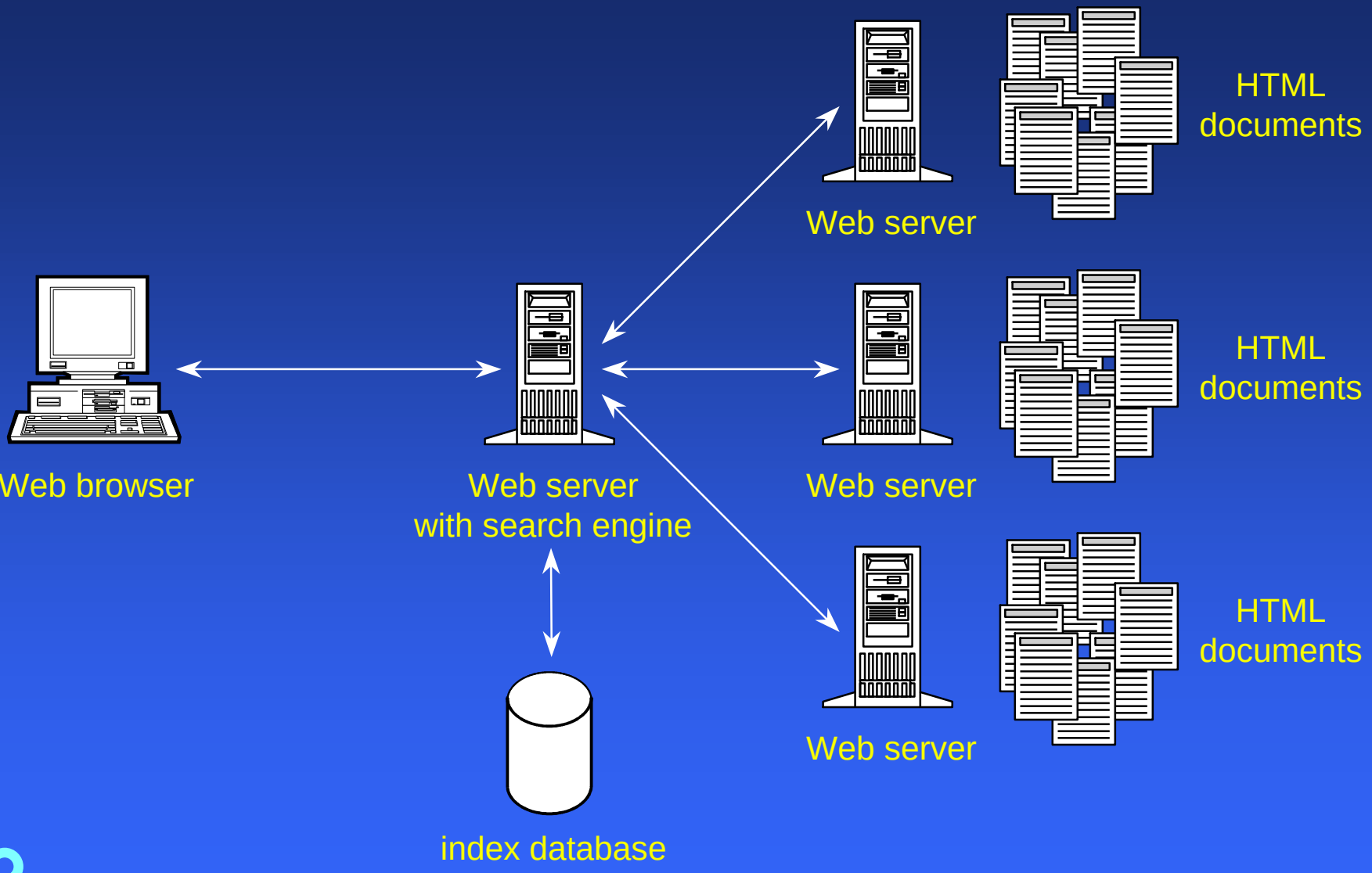


index database

Search engines

- What?
  - build a collection of Intranet documents for a specific use
  - index every Intranet document in that collection
- Used where?
  - centralized Web infrastructure
  - single publishing site
  - “top-down” Intranets
- Products
  - Verity *Topic*
    - topicACCESS, topicSEARCH, topicAGENTS
  - Open Text *Livelink Search*, Fullcrum *SearchServer*, MS *Tripoli*

Web spiders



Web spiders

- What?
  - traverse the Intranet by following hyperlinks
  - index every Intranet document found
- Used where?
  - distributed Web infrastructure
  - multiple publishing sites
  - “bottom-up” Intranets
- Products
  - Digital *AltaVista Search*
    - Enterprise, Workgroup, Personal edition
  - Open Text *Livelink Spider*

HTML or SGML?

- What is SGML?
  - ISO international standard for electronic document interchange
  - a meta-language for specifying different markup languages
  - HTML is just a (simple) example of such a language
- 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
- *“HTML is like a baby SGML,*
  but it is a baby born without a brain”

HTML or SGML?

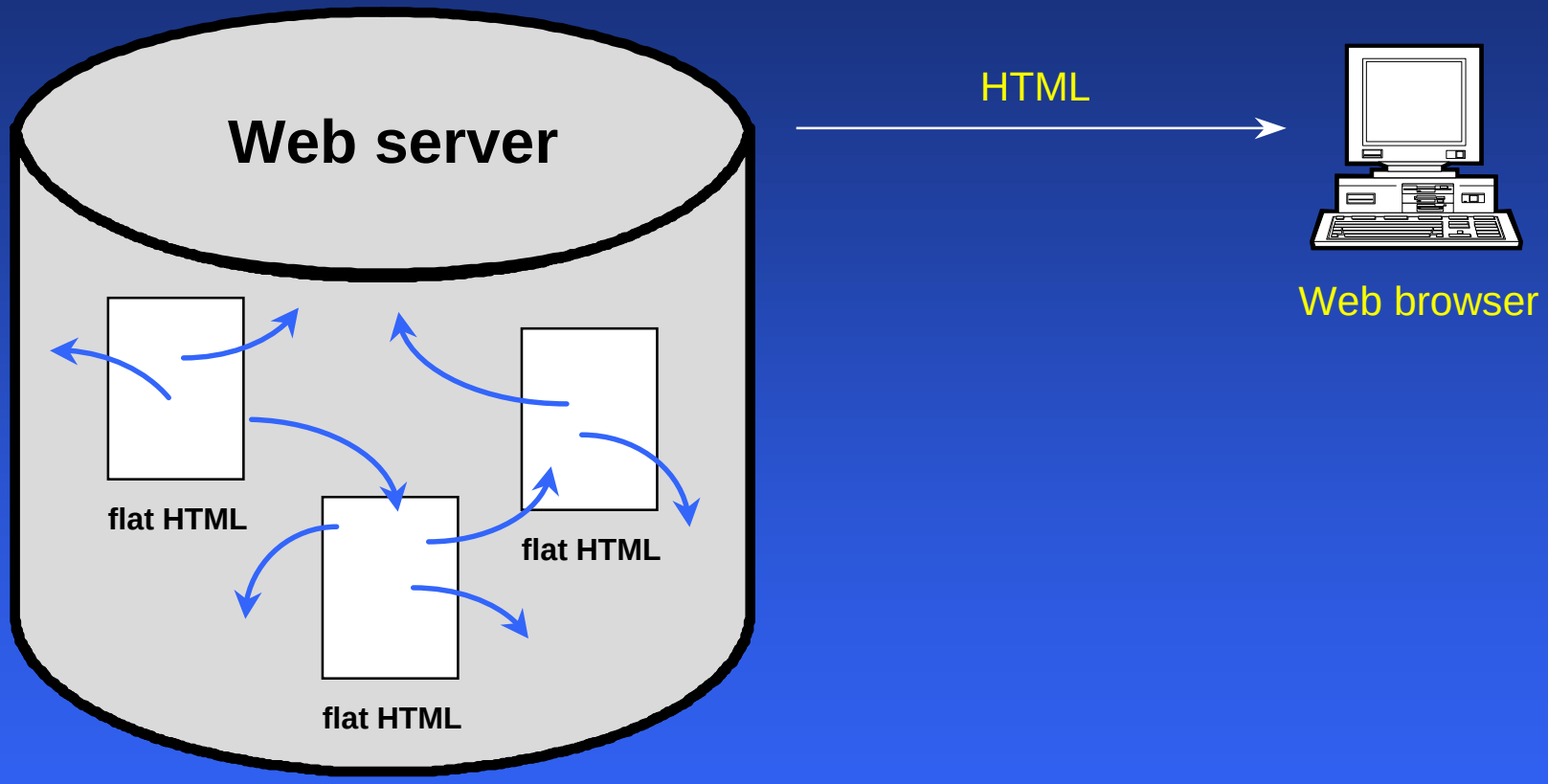
```
<HTML>
<HEAD> … </HEAD>
<BODY>
<P>From: <B>GDT</B></P>
<P>To: <B>MDL</B></P>
<P>Subject: <I>Strategy</I></P>
<P>Keyword:
   <A HREF="http://www.nv.be/
      catalog/products/x2000/">
      X2000
   </A>
<HR>
<P>I believe our strategy in
   selling the X2000 … </P>
</BODY>
</HTML>
```

```
<MEMO>
<HEAD> …
<NAMELOC ID=id123>
<NMLIST>catalogproducts</NMLIST>
</NAMELOC>
</HEAD>
<BODY>
<FROM>GDT</FROM>
<TO>MDL</TO>
<SUBJECT>Strategy</SUBJECT>
<KEYWORD LINKEND=id123>
 X2000
</KEYWORD>
<P>I believe our strategy in
   selling the X2000 … </P>
</BODY>
</MEMO>
```

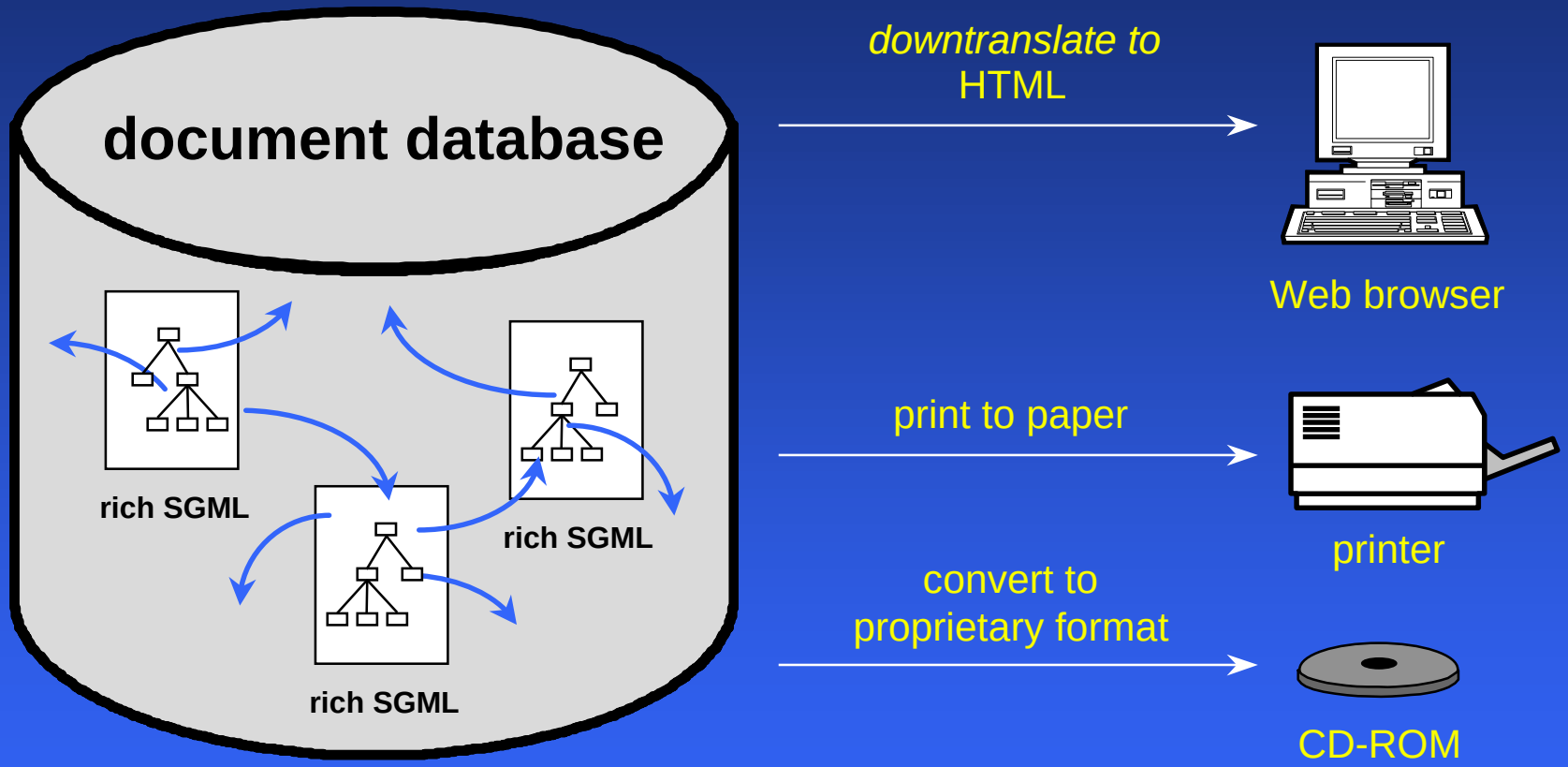
HTML or SGML?

- Advantages of SGML
  - standard immune to present and future Internet politics
  - markup under *your* control, suited to *your* documents
  - simplifies administration of document repositories
  - documents can be reused for different purposes
  - different versions of a document can be built
  - industrial-strength tools are readily available
- The best of both worlds:
  - SGML = the “back-end” *content markup* language
  - HTML = the “front-end” *presentation markup* language
  - batch conversion or on-the-fly generation

HTML or SGML?



HTML or SGML?



Guidelines for Intranet documents

- analyse your documents
  - which documents are really business-critical?
  - what is your optimal document mix?
  - authoring or conversion?
- analyse your information flows
  - top-down or bi-directional?
  - documents for consultation or for sharing?
  - formally approved or “publish-as-you-please” documents?
- analyse your production processes
  - writing / editing / validating / publishing
  - different persons with different responsibilities?
  - separate Web sites with distinct document functionalities?

The future of Intranet documents

- "information at your fingertips"
  - the computer on your desk is important not for the data it can process, but for the information it can give access to
  - Intranet document engineering possibilities are exciting, but off-the-shelf industrial-strength tools are only just appearing
- the browser = the operating system
  - the Netscape model: browser ≥ operating system
  - the Microsoft model: operating system ≥ browser
  - Microsoft owns the desktop (Windows 95) and is now conquering the network (Windows NT)
  - è will Microsoft own the Intranet too?
- *Intranet documents are the windows to your business*