



XML: a brief introduction

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



The need to go beyond HTML

Arguments pro HTML

- | simple
- | widespread
- | interoperable
- | user-controlled
- | fault-tolerant

Arguments contra HTML

- | simplistic
- | unmanageable
- | browser dependent
- | application opaque
- | almost anything goes

n Conclusion: HTML has reached its built-in limits

- HTML is a *presentation* format, not a *document/data* format
- the browser wars are over, so we can start doing something useful
 - building document/data management/transaction solutions
 - for the networked enterprise (intranet / extranet / internet)



XML motivation

n Beyond HTML, instead of SGML:

- problem: extending HTML
- suggested solution: SGML
 - ü extensible by definition
 - ü has all the necessary mechanisms to address HTML's shortcomings
 - û big and (sometimes very) complicated
 - û hated by Web developers and designers, misunderstood by Web users
- real solution: XML (Extensible Markup Language)
 - throw the hard parts of SGML away → is SGML - -
 - optimize SGML for Web creation and delivery
 - an extended, richer version of HTML → is *not* HTML + +
 - a leaner, meaner subset of SGML for use on the Internet

n Initiative of WG8 of the World Wide Web Consortium (W3C)

- vision expanded to include push, metadata, transactions, ...



XML history

n Milestones:

- Jul '96 W3C XML Working Group is formed
- Nov '96 First draft of XML standard published
- Mar '97 Microsoft announces CDF push format
- Apr '97 Netscape accepts XML as a new data format
- Oct '97 Microsoft ships IE 4.0 with 2 built-in XML parsers
- Dec '97 Draft XML 1.0 standard submitted to W3C
- Feb '98 Final XML 1.0 standard approved by W3C**
- Jun '98 First XML-aware beta versions of NS and IE 5.0
- Q3 '98 Availability of commercial XML tools and technologies
- Q4 '98 XML will become a native data format of MS Office '98



XML design goals

- Ⓔ XML shall be straightforwardly usable over the Internet
 - XML shall support a wide variety of applications
- Ž XML shall be fully compatible with SGML
 - It shall be easy to write programs which process XML
 - The number of optional features in XML is to be kept to the absolute minimum, ideally zero
 - The XML design should be prepared quickly
 - The design of XML shall be formal and concise
 - “ XML documents shall be easy to create
 - “ Terseness is of minimal importance



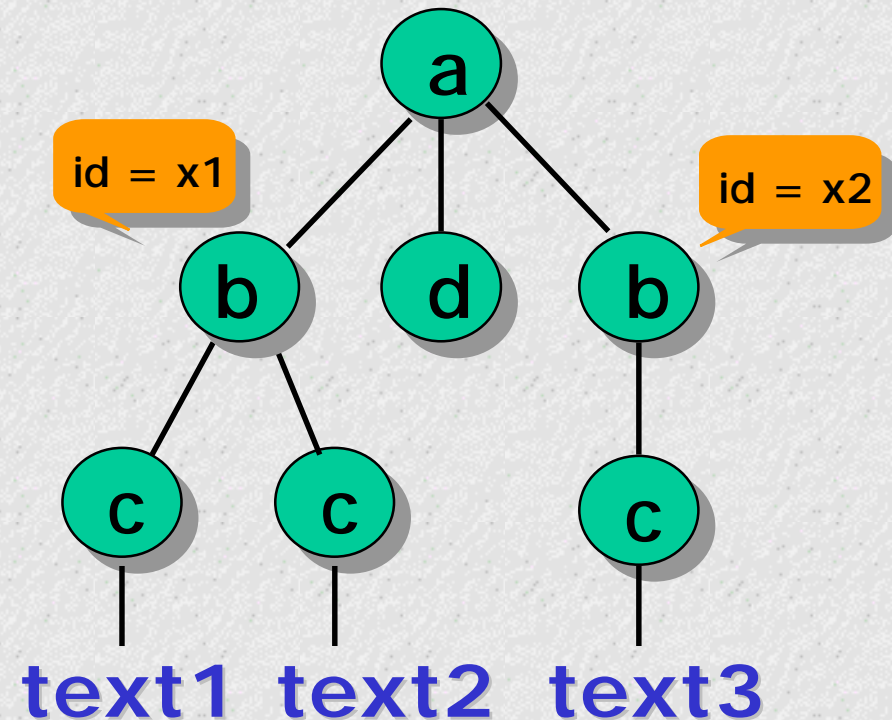
Key concept: XML is SGML

- n XML looks a lot like HTML, but really behaves like SGML
 - simple markup language, but adheres to strong conventions
 - focus on *structure* and *representation* instead of *presentation*
 - obscure and never-used SGML features have been left behind
- n **Valid** documents
 - respect a DTD (Document Type Definition)
 - can be parsed without errors by an XML parser
 - the document structure can be passed to an application
- n **Well-formed** documents
 - start and end tags must match
 - elements must nest hierarchically
 - there must be (only) one root element



Key concept: XML is SGML

```
<?XML VERSION="1.0"?>  
  
<a>  
  <b id="x1">  
    <c>text1</c>  
    <c>text2</c>  
  </b>  
  <d att="xyz"/>  
  <b id="x2">  
    <c>text3</c>  
  </b>  
</a>
```



data markup
“transport format”

data structure
“element grove”



Key concept: XML markup is about meaning

```
<order>
  <sold-to>
    <person>
      <lastname>Layman</lastname><firstname>Andrew</firstname>
    </person>
  </sold-to>
  <sold-on>19970317</sold-on>
  <item>
    <price>5.95</price>
    <book>
      <title>Number, the Language of Science</title>
      <author>Dantzig, Tobias</author>
      <isbn>0-452-01030-6</isbn>
    </book>
  </item><item>
    <price>12.95</price>
    <record>
      <title><composer>Tchaikovsky</composer>'s First Piano Concerto</title>
      <style>classical music</style>
      <artist>Janos</artist>
    </record>
  </item><item>
    <price>1.50</price>
    <coffee >
      <size>small</size>
      <style>cafe macchiato</style>
    </coffee>
  </item>
</order>
```




Key concept: XML is a document/data format

n XML is a **document** format

- markup to capture the meaning of content
 - intelligent searching, filtering, ...
- markup to verify the correctness of structure
- è open document computing applications

n XML is a **data** format

- markup to capture the meaning of information
 - intelligent processing, extracting, ...
- markup to verify the completeness of definition
- è open data processing applications

n XML will end the distinction between

- networked document distribution
- networked data transactions



Key concept: XML is a set of solutions

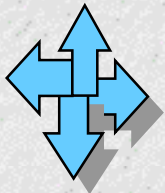
Related standards



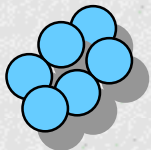
XML : Extensible Markup Language



XSL : Extensible Stylesheet Language



XLL : Extensible Linking Language



DOM : Document Object Model



XSL: Extensible Stylesheet Language

n XML stylesheet mechanism

- based on DSSSL
- compatible with CSS 1.0 and 2.0
- advanced capabilities:
 - reordering contents
 - automatically generated text
 - for both printing and online display

n is useful for:

- handling presentation separate from content
- single source, multiple output media
- reformatting / re-using content

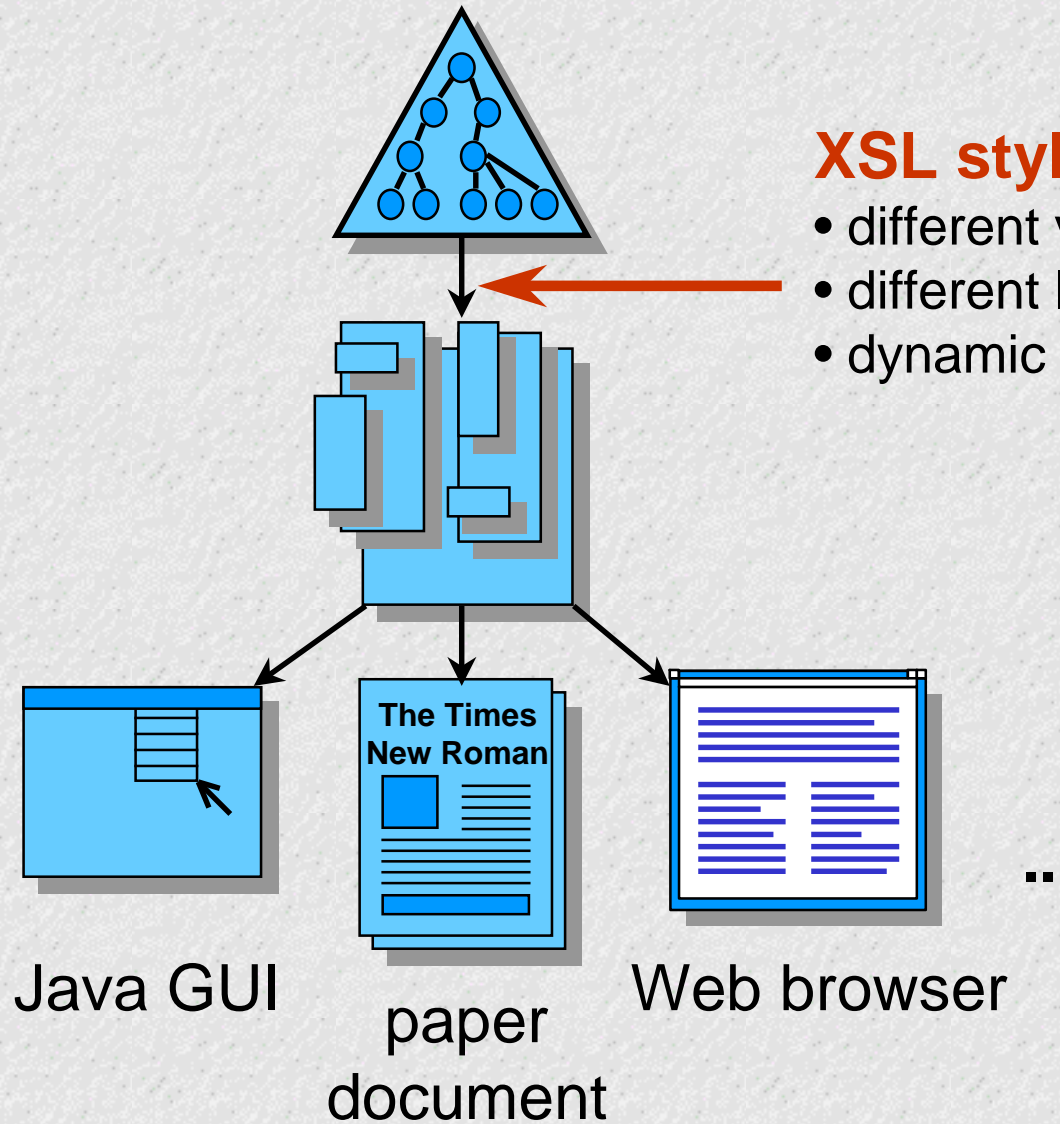


XSL: Extensible Stylesheet Language

element
grove

flow
object
tree

output
media

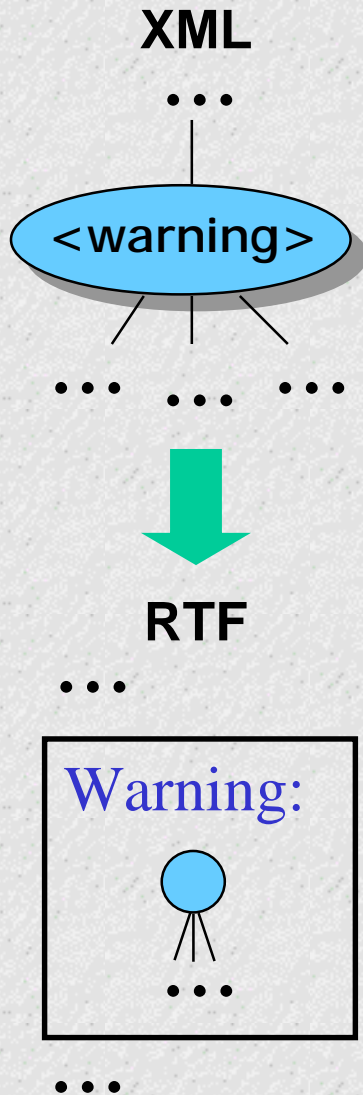


XSL stylesheet

- different views
- different layouts
- dynamic contents



XSL: Extensible Stylesheet Language



```

<rule>
  <!-- pattern in element grove -->
  <target-element
    type="warning"/>
  <!-- object in flow objects tree -->
  <box>
    <paragraph
      font-size="24pt"
      font-family="serif">
      Warning:
      <children/>
    </paragraph>
  </box>
</rule>

```

rule

pattern

action



XLL: Extensible Linking Language

n XML linking mechanism: XLink and XPointer

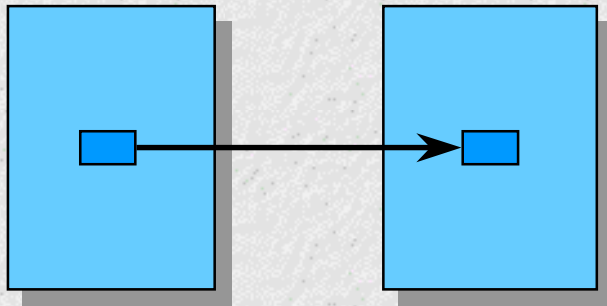
- subset of HyTime and TEI
- compatible with existing URL linking
- additional functionality:
 - bi-directional links
 - conditional links
 - indirect links

n is useful for:

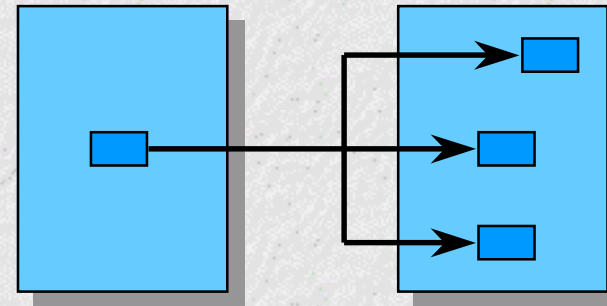
- rich hypertext functionality
- computer-based training
- Web site management



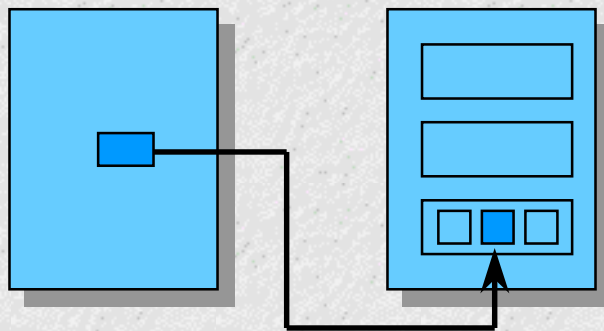
XLL: Extensible Linking Language



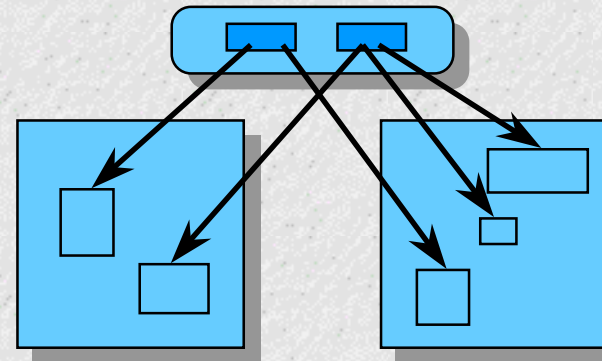
simple link



extended link



symbolic link



link group



DOM: Document Object Model

n electronic document API

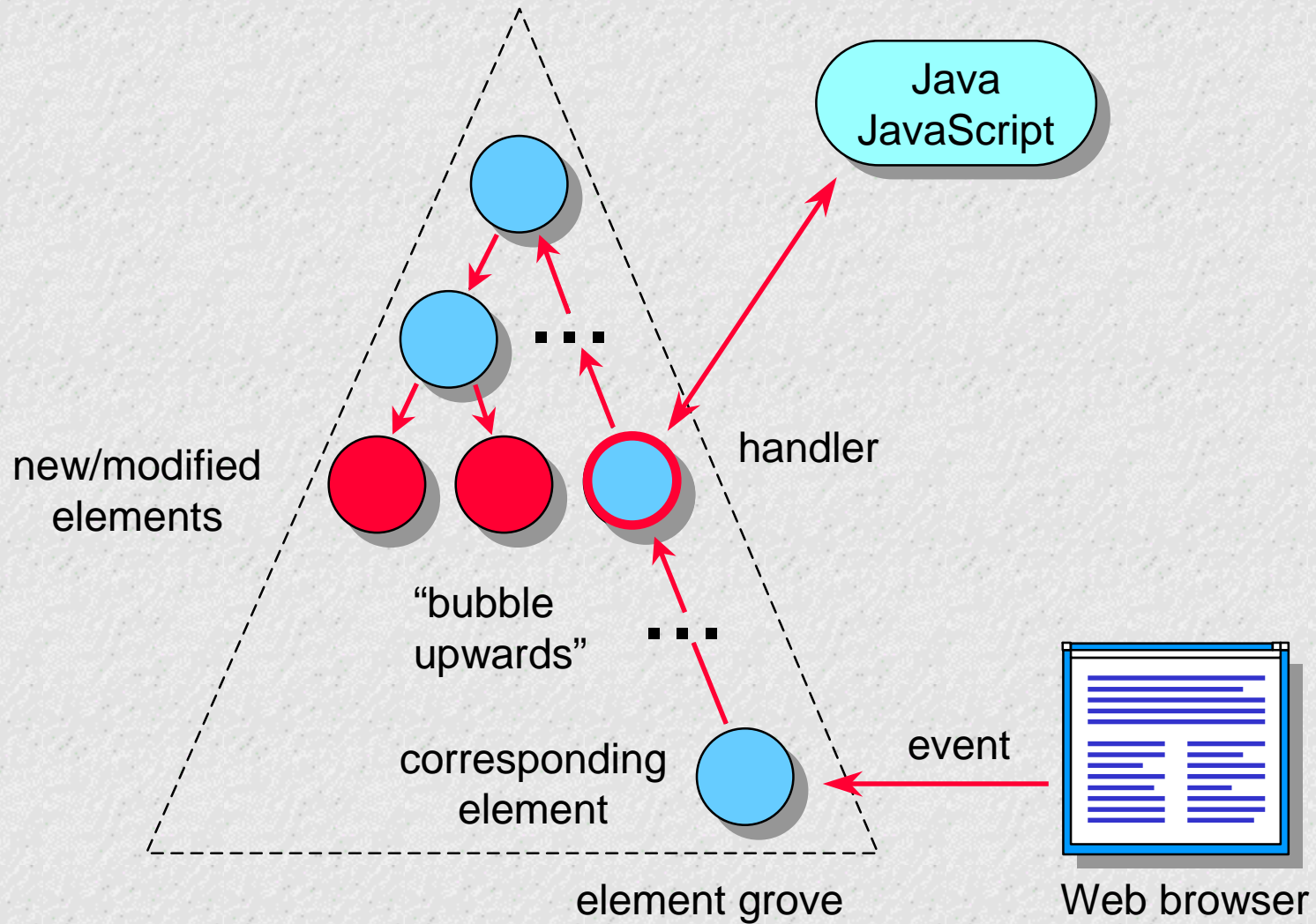
- standard API for electronic documents and GUI event handling
- access to document contents as an attributed tree of elements
- advanced capabilities:
 - expanding/collapsing text
 - dynamic tables of contents
 - client-side active documents

n is useful for:

- interactive parts catalogs
- electronic self-service manuals
- online process/procedures documentation



DOM: Document Object Model





XML: beyond HTML, instead of SGML

Hans C. Arents

s.a. OFFIS n.v.

“Office Future International Services”

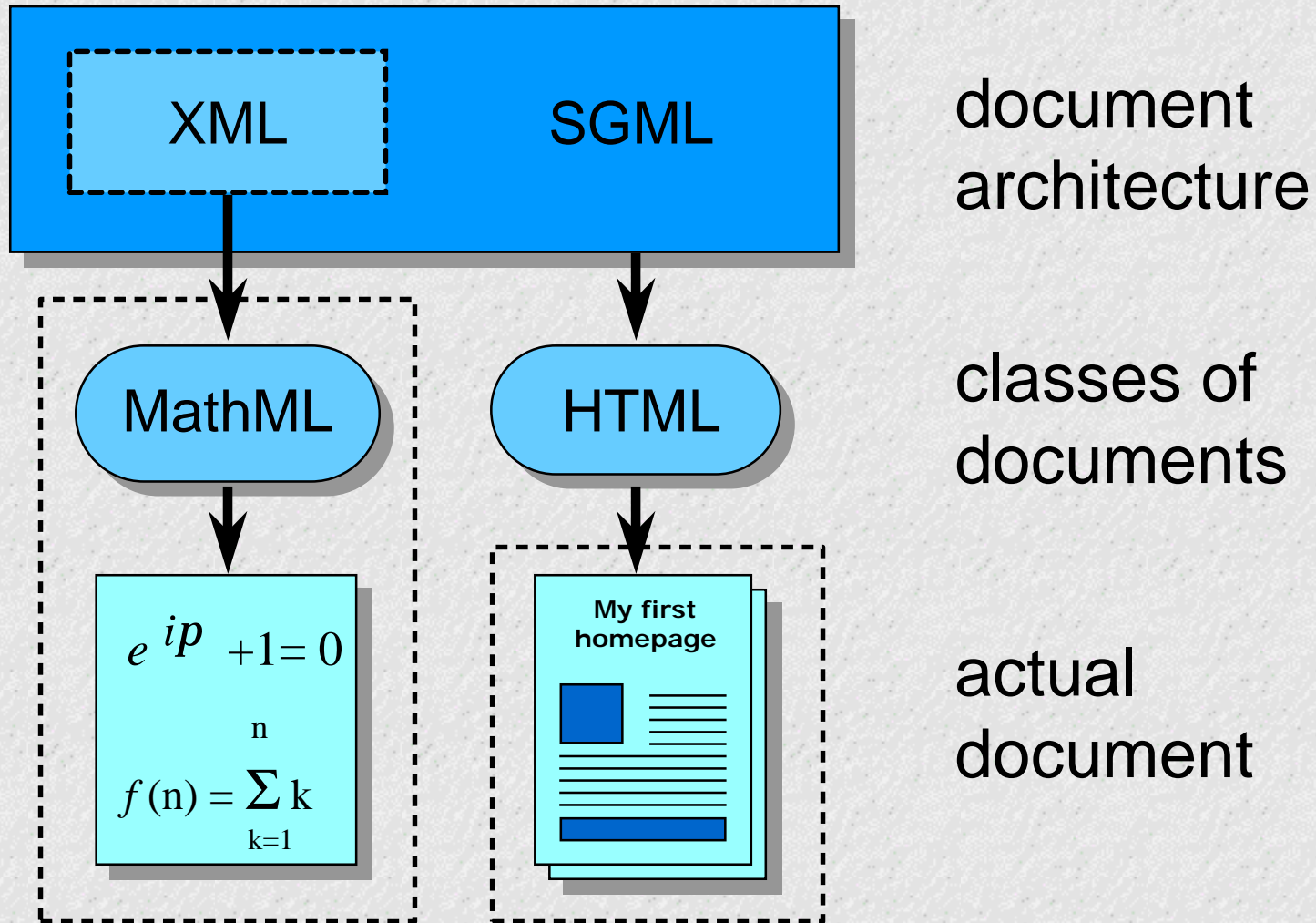
Atlas Park, Weiveldlaan 41 B. 32, B-1930 Zaventem, Belgium

Tel: +32 (0)2 725 40 25 - Fax: +32 (0)2 725 40 12

Email: info@offis.be - Web: www.offis.be

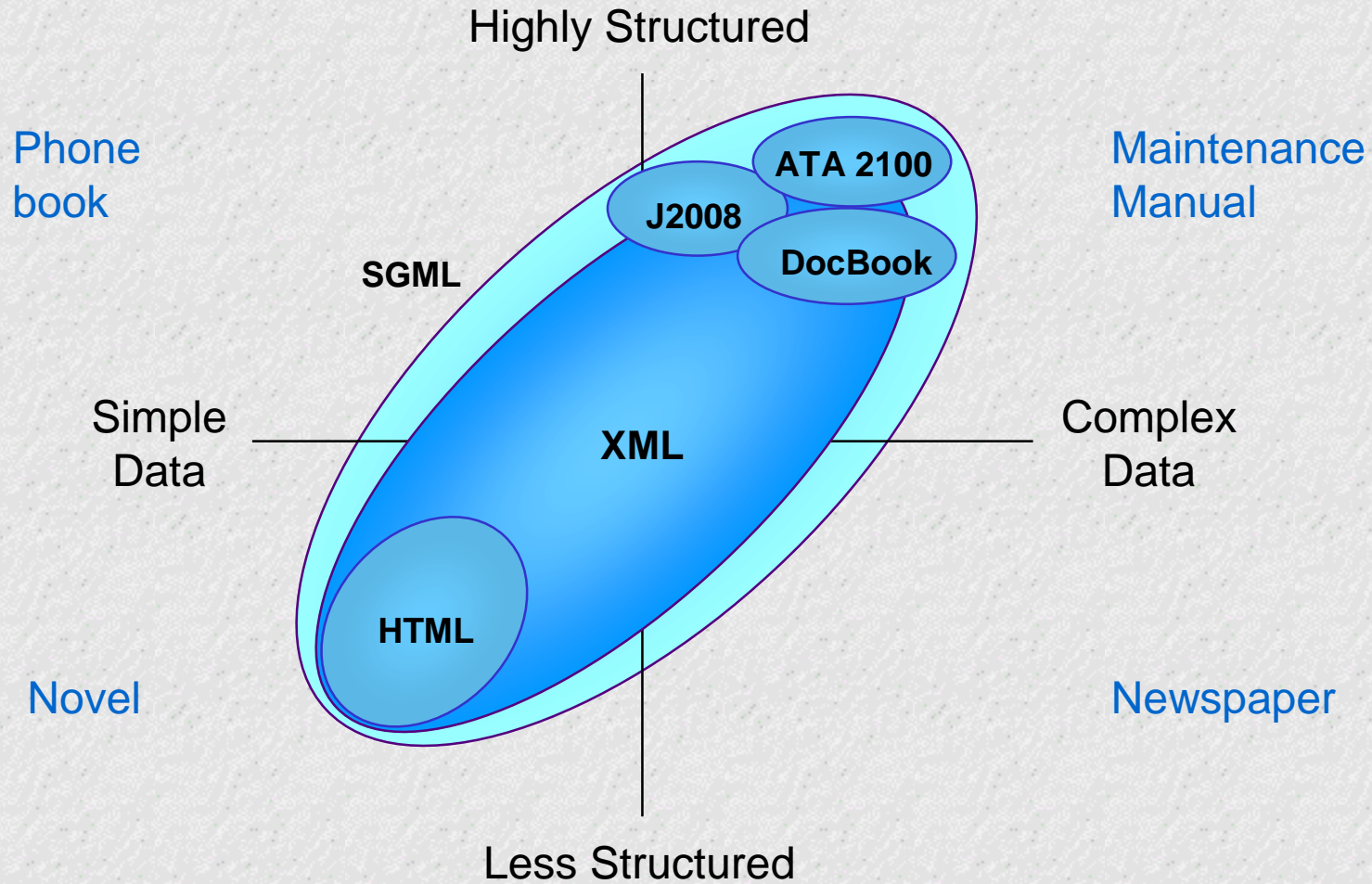


XML compared to HTML / SGML





XML compared to HTML / SGML





XML compared to HTML / SGML

n XML matches SGML

- captures structure & meaning
- future-proof standard
- repurpose and reuse
- strength of validation
- ease of automation

n XML improves on SGML

- reduces optional and advanced features
- machine-processable without a DTD
- mainstream browser and tool support
- standardized stylesheets (XSL)
- standardized linking (XLL)



XML compared to HTML / SGML

- n 80% of the SGML functionality ...



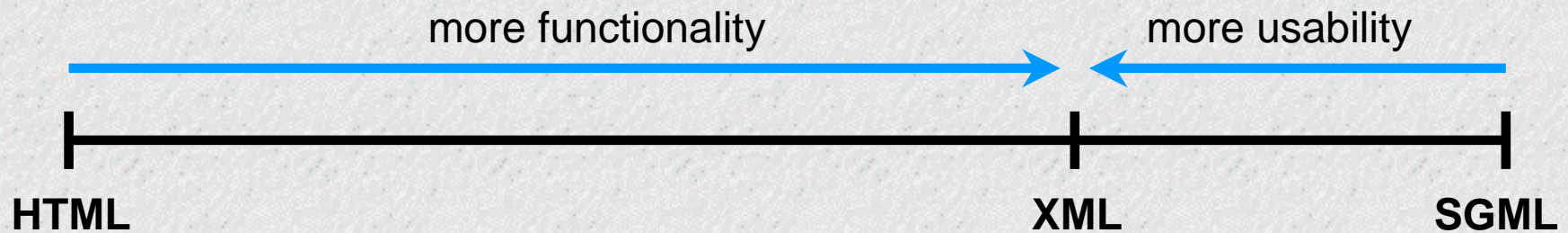
- n ... for 20% of the cost in time and effort





XML compared to HTML / SGML

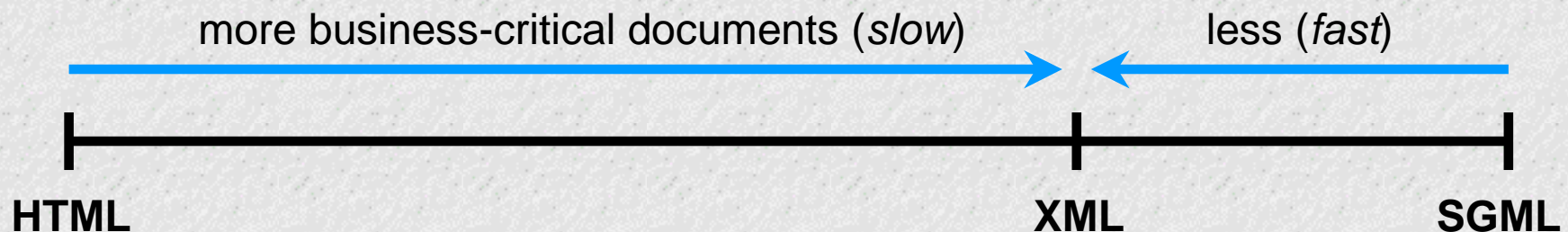
n Application capabilities?



n Application price?



n Application uses?





Where & when to use HTML / XML / SGML

Web Application:

Create in:

Deliver in:

Home page/personal website

HTML

HTML

Huge number of pages/large sites

XML/SGML

XML/HTML

Rich pages/highly interactive sites

XML/SGML

XML/DHTML

Interactive/automated documents

XML/SGML

XML

Formal processes (workflow, ISO 9000)

XML/SGML

XML

Data for reuse and interchange

XML/SGML

XML

Complex document repositories

SGML

XML

Non-document, data exchange

XML

XML



XML tools and technologies

n XML parsers

- Microsoft XML parsers (C, Java) + XML DSO and XML OM in IE4
- IBM XML for Java parser, DataChannel XML parser (Java), ...

n XML converters/databases/middleware

- Inso *DynaTag*, *DynaBase* and *DynaWeb*
 - convert wordprocessing documents to XML, manage and distribute
- AIS *Balise*
 - programming environment for building XML-based information systems
- OmniMark *Konstruktor*
 - development suite for XML content management & delivery applications

n XSL tools

- Microsoft *XSL processor* : converts XML to HTML on-the fly
- ArborText *XSL Styler* : XML stylesheet editor (Windows 95/NT)



XML as a document format

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



XML as a document format: what's the problem?

n The problem with document representation

- proprietary document formats
 - vendor-specific
 - application-specific
- proprietary document management solutions
 - store-and-forget document stores
 - high-end document managers
 - custom-built solutions

n Typical operations on documents

- the use of structured documents
- document workflow
 - sharing
 - tracing
 - archival
 - retrieval



XML as a document format: when to use?

n preserving documents **over time**

- future-proofing documents against tools and context
- è machine-readable format: for parsers & generators
- è self-descriptive format: for extraction and validation

n preserving documents **over space**

- distributed systems with decentralized documents
- è common documents definitions
- è link to networked resources

n preserving documents **across companies**

- companies defined by their way of speaking and way of working
- è shared terminology, defining common terms
- è document-driven workflow



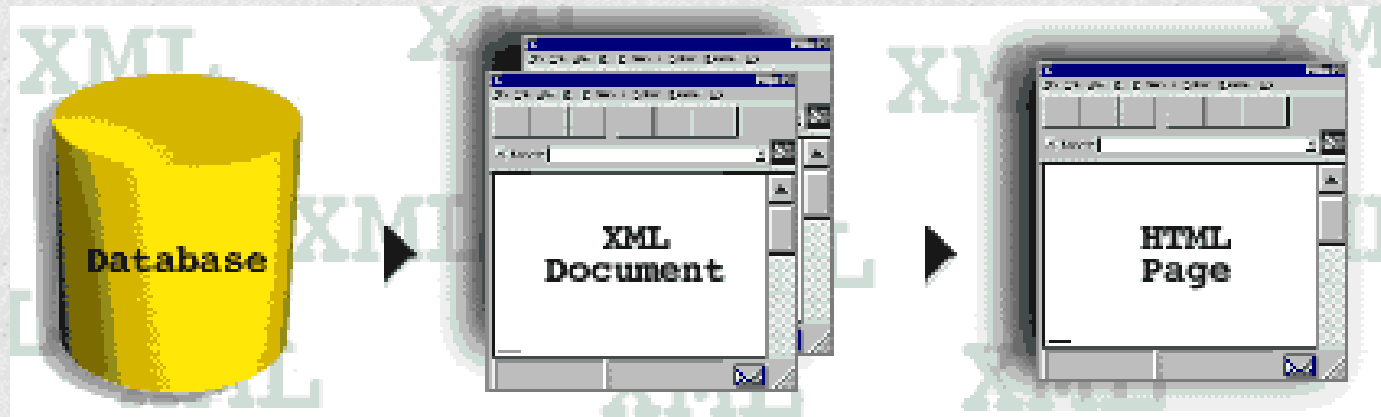
XML as a document format: what's available?

- n Focus on intelligent document handling on the Web
- n Existing applications:
 - CDF: Channel Definition Format
 - download directives for push channels (Microsoft Internet Explorer)
 - RDF: Resource Description Framework
 - metadata about Web pages/program objects (Netscape Navigator)
- n Future applications:
 - ICE: Information and Content Exchange
 - data format to facilitate the process of automatically exchanging, updating, supplying, and controlling information assets
 - PICS-NG: Platform for Internet Content Selection - Next Generation
 - content rating of Web site pages
 - SDML: Signed Document Markup Language
 - creating, processing, and displaying electronic "signed writings"



Intranet XML use: database pull

n XML for consultation of electronic documents



- database of modular document components
- XML-based virtual document
- HTML-based actual document

n Used for:

- self-service manuals
- process/procedures documentation



Example self-service maintenance manual

n server-side

- parts information (XML fragments)
- configuration information (XML hyperlinks)

n browser-side

- fragment selection (JavaScript/VBscript)
- fragment assembly (XML document)
- fragment display (XSL stylesheet)
- document display (HTML document)

The top screenshot shows a browser window with the title 'Balise Xml Plugin - Microsoft Internet Explorer'. The address bar contains 'fragment: xml/391610.xml'. The page title is 'HYDRAULIC CIRCUIT BREAKER'. Below the title is a table with columns for quantity, part number, and description. The table contains the following data:

1	95 588 686	01 CIRCUIT BREAKER
	RP 5274 16	01
		EXCEPT POWER STEERING OPR. before 6893.
		BYA POWER STEERING PRESS PUMP 6+2 + NON-SINKING OPR. before 6894.
1	5274 16	01 CIRCUIT BREAKER
		INJECTION XU7JP BYA POWER STEERING PRESS PUMP 6+2 + NON-SINKING OPR. after 6895.
		INJECTION XU10D2 BYA POWER STEERING PRESS PUMP 6+2 + NON-SINKING OPR. after 6892.
		DIESEL XUD9 BYA POWER STEERING PRESS PUMP 6+2 + NON-SINKING AIR CONDITIONING OPR. after 6895.
		DIESEL XUD9 BYA POWER STEERING PRESS PUMP 6+2 +

The bottom screenshot shows the same browser window with the address bar containing 'fragment: xml/391610.xml'. The page title is 'HYDRAULIC CIRCUIT BREAKER'. Below the title is a table with columns for quantity, part number, and description. The table contains the following data:

1	5274 18	01 CIRCUIT BREAKER
2	95 666 763	01 SPHERE + GASKET
3	5 412 801	01 DRAIN SCREW
4	22 709 009	01 BALL
5	96 157 738	01 HYD OIL O-RING
6	95 630 952	01 SUSP CYL O-RING
	RP 5272 15	01

Arrows point from the text 'partly specified parts configuration' to the top screenshot and 'completely specified parts configuration' to the bottom screenshot.

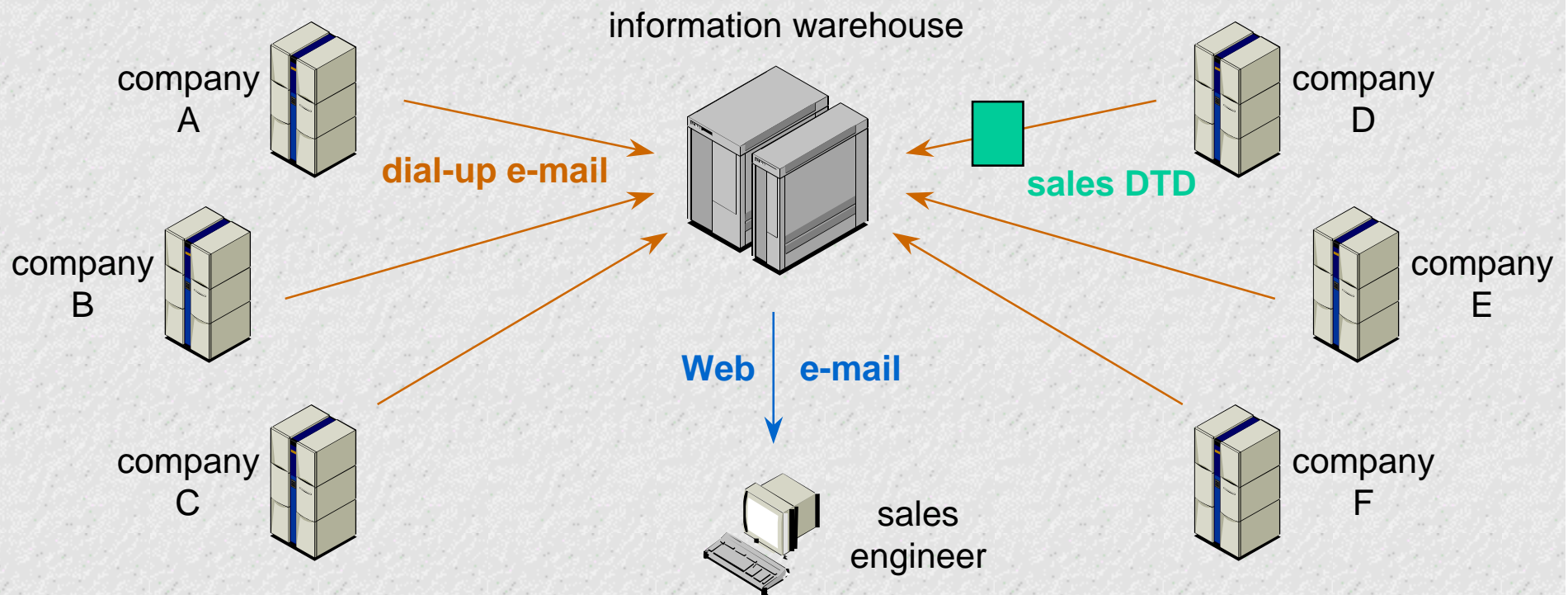
partly specified parts configuration

completely specified parts configuration



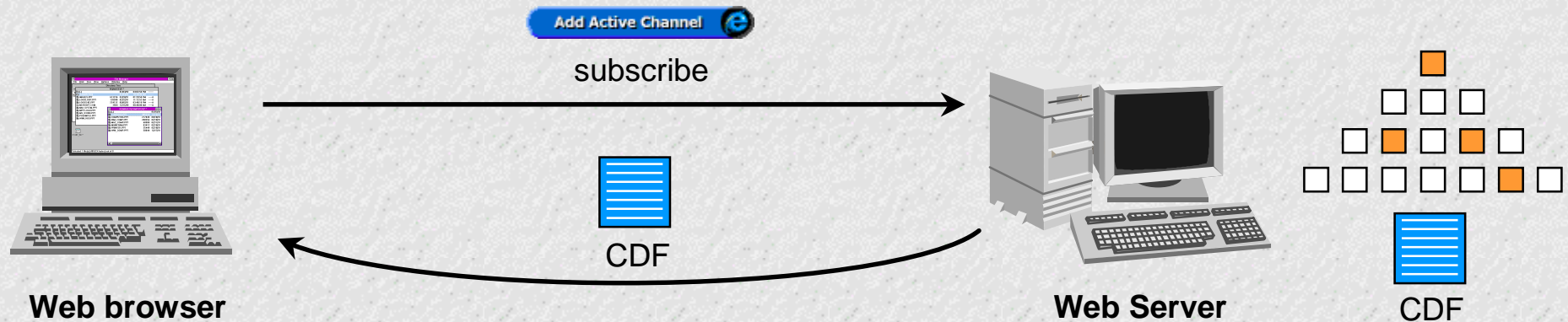
Extranet XML use: information aggregation

- n group of companies collaboratively exchanging information
 - not just data, but rich documents
 - agreeing on specific but common document format





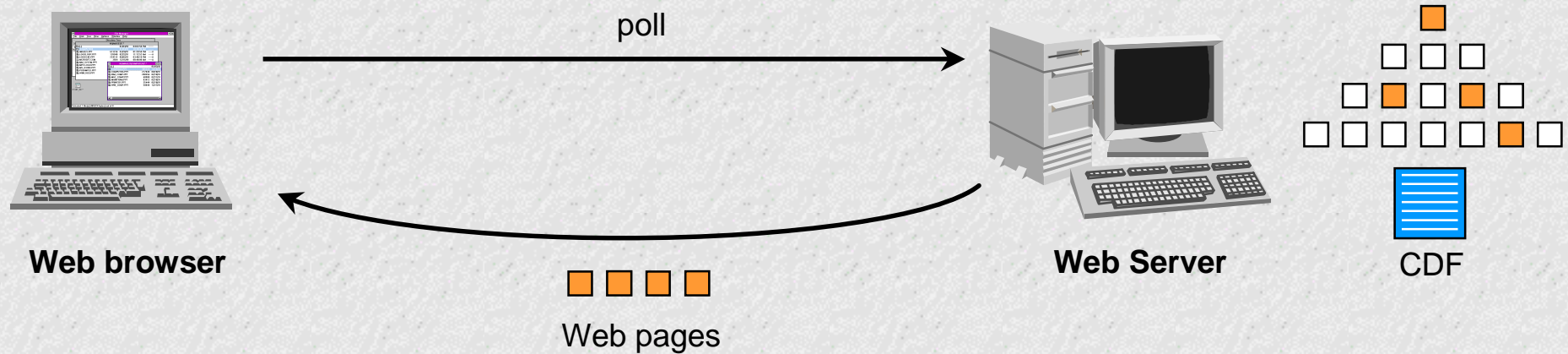
Internet XML use: information push



- Ⓔ client subscribes to Web site
- server sends CDF file with “smart pull” directives



Internet XML use: information push



- client polls Web site for push of changed content
- server sends changed pages according to CDF directives



Example CDF file

```
<?XML VERSION="1.0"?>
```

```
<CHANNEL HREF="http://msie.cmpnet.com/news/home"  
        SELF="http://msie.cmpnet.com/CMPnet.cdf">
```

```
<SCHEDULE STARTDATE="1998-02-17T00:00+0000" TIMEZONE="-0700">  
  <INTERVALTIME HOUR="6" />  
  <EARLIESTTIME HOUR="2" />  
  <LATESTTIME HOUR="6" />  
</SCHEDULE>
```

```
<LOGTARGET HREF="http://msie.cmpnet.com/logging"  
          METHOD="POST" SCOPE="OFFLINE">  
</LOGTARGET>
```

```
<LOGO HREF="http://img.cmpnet.com/msie/CMP19432.gif" STYLE="IMAGE-WIDE" />  
<LOGO HREF="http://img.cmpnet.com/msie/CMP8032b.gif" STYLE="IMAGE" />
```

```
<TITLE>CMPnet Active Channel</TITLE>
```

```
<ABSTRACT>Welcome to the CMPnet Active Channel. The best place on the Web to  
find technology and internet information.</ABSTRACT>
```

```
...
```



Example CDF file parsed by MS XML parser

The screenshot shows the XMLViewer application window. The title bar reads "XMLViewer". The address bar shows the file path: "file:/C:/msxml/viewer/CMPnet.cdf". There are "Parse" and "View" buttons to the right of the address bar. The main content area displays a tree view of the XML document structure under the root element "CHANNEL".

- CHANNEL
 - SCHEDULE
 - INTERVALTIME
 - EARLIESTTIME
 - LATESTTIME
 - LOGTARGET
 - LOGO
 - LOGO
 - TITLE
 - CMPnet Active Channel
 - ABSTRACT
 - Welcome to the CMPnet Active Channel. The best place on the Web to find technology and internet information.
 - ITEM
 - LOGO
 - TITLE
 - News
 - LOG

At the bottom of the window, there is a status bar that says "Done." and a warning message: "Warning: Applet Window".



XML as a data format

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



XML as a data format: what's the problem?

n The problem with data representation

- proprietary data formats
 - client/server
 - transaction-based processing
- proprietary middleware solutions
 - data converters
 - gateway applications
 - migration tools

n Typical operations on data

- the definition of data models
- transactional processing
 - data transmission
 - verification
 - archival
- migration



XML as a data format: when to use?

n Description of syntactical data schema

- DTD's
- appropriate document formats
 - 'well-formed'
 - hierarchical

n Description of conceptual data schema

- concepts
 - classes
 - objects
 - properties
- ... and their relationships
 - RDBMS
 - OODBMS



XML as a data format: when to use?

n Description of metadata

- data about data
 - MCF: **M**eta **C**ontent **F**ormat
 - XMI: **X**ML **M**etadata **I**nterchange Format
- enhanced retrieval
- autogeneration of hub documents
 - keyword-based
 - thesauri

n Description of data locators

- where to find what kind of data on the Web
 - RDF: **R**esource **D**escription **F**ramework

n Message-based transactions

- application interactions between servers over standard Web protocols
 - WIDL: **W**eb **I**nterface **D**efinition **L**anguage



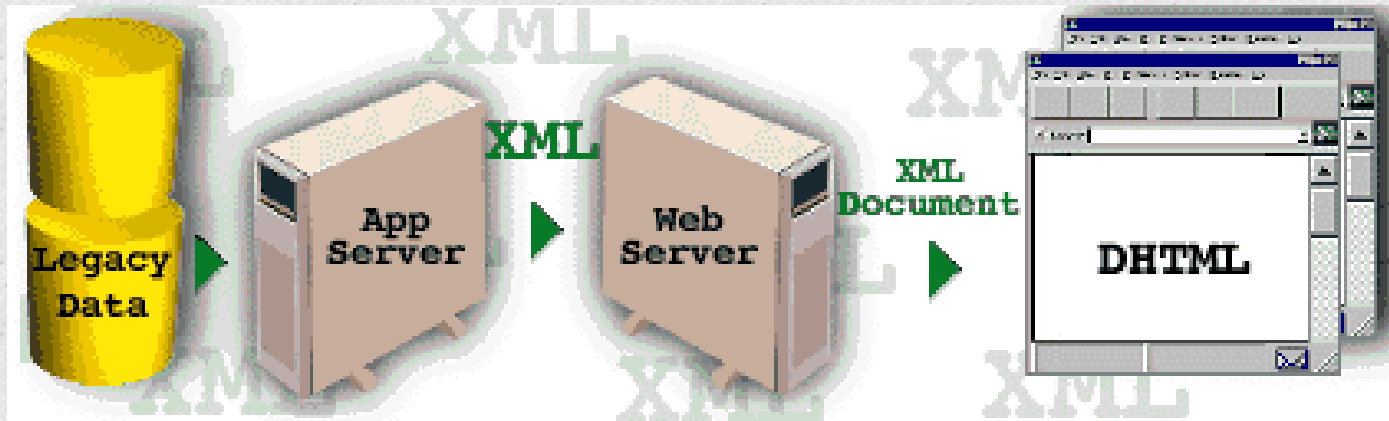
XML for data exchange: what's available?

- n Focus on open data exchange between different applications
- n Existing applications:
 - OSD: Open Software Description Format
 - a vocabulary used for describing software packages and their dependencies for automated software distribution (Marimba, Microsoft)
 - OFX: Open Financial Exchange
 - for exchanging financial data and instructions between customers and their financial institutions (Microsoft Money, Quicken)
- n Future applications:
 - XML/EDI: EDI using XML
 - for replacing expensive VAN-based EDI by cheap Internet-based EDI
 - OTP: Internet Open Trading Protocol
 - transaction protocol for encapsulating the whole e-buying process



Intranet XML use: legacy database access

- n XML for dynamic interaction with legacy data



- legacy database
- XML-based transaction server
- DHTML-based browser user interface or Java client

- n Used for:

- web-enabling legacy databases
- data-enriching of Web sites



Intranet XML use: legacy database access

```
<?XML VERSION="1.0" ENCODING="UTF-8" STANDALONE="yes"?>
<REGEDOC>
<TABLE NAME="DOSSIER" NUMRECORDS="59" NUMFIELDS="2">
<RECORDSPEC>
<FIELDSPEC TYPE="Character" OCCURS="single">NUMBER</FIELDSPEC>
<FIELDSPEC TYPE="Character" OCCURS="single">DUTCH_TITLE</FIELDSPEC>
</RECORDSPEC>
<RECORD>
<FIELD><VALUE>89A01050.050</VALUE></FIELD>
<FIELD><VALUE>Programmawet-1989</VALUE></FIELD>
</RECORD>
<RECORD>
<FIELD><VALUE>89A02620.001</VALUE></FIELD>
<FIELD><VALUE>W-gedwongen medeëigendom</VALUE></FIELD>
</RECORD>
<RECORD>
<FIELD><VALUE>89A03310.002</VALUE>
<FIELD><VALUE>KB-dienstplichtige
</RECORD>
<RECORD>
<FIELD><VALUE>89A11030.030</VALUE>
<FIELD><VALUE>W-Staats hervorming
</RECORD>
<RECORD>
<FIELD><VALUE>89A20030.041</VALUE>
<FIELD><VALUE>Ristorno's 1988-G
</RECORD>
...
</TABLE>
</REGEDOC>
```

REGEDOC - Consultation

Nombre de dossiers trouvés : 59

Dossier	Titre
89A01050.050	Programmawet-1989
89A02620.001	W-gedwongen medeëigendom
89A03310.002	KB-dienstplichtigen-vrijlating-broederdienst
89A11030.030	W-Staats hervorming-Duitstalige Gemeenschap
89A20030.041	Ristorno's 1988-Gemeenschappen en Gewesten
89A20400.008	W+KB-financiële transacties en markten
89A20400.013	W-Nationale Delcrederedienst
89A20410.010	KB-BTW-RTBF-Koninklijke Munt van België
89A20410.011	KB-belastingsvrijstelling-"Fondation eco.soc.B.W."
89A20410.013	KB-accijnzen.benzine-bier -BTW vruchtesappen

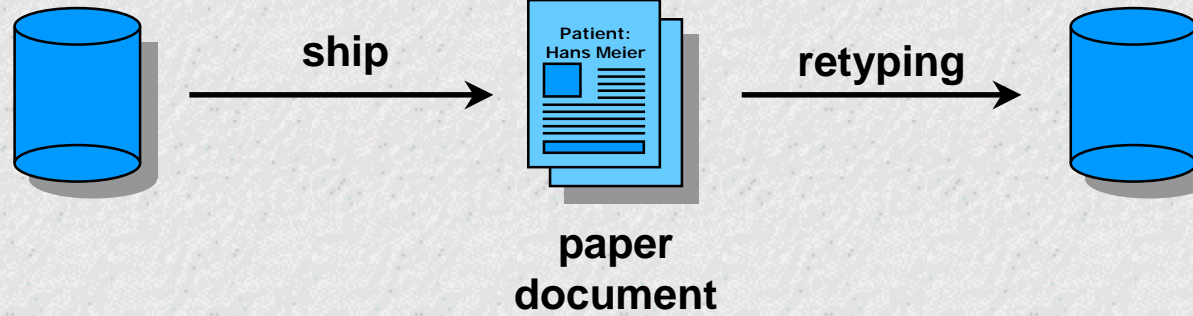
Relations
Documents
Historique
Notes Privées
Infos
Titres longs

Recherche par Mots-Clés Recherche Structurée Rapports Fermer

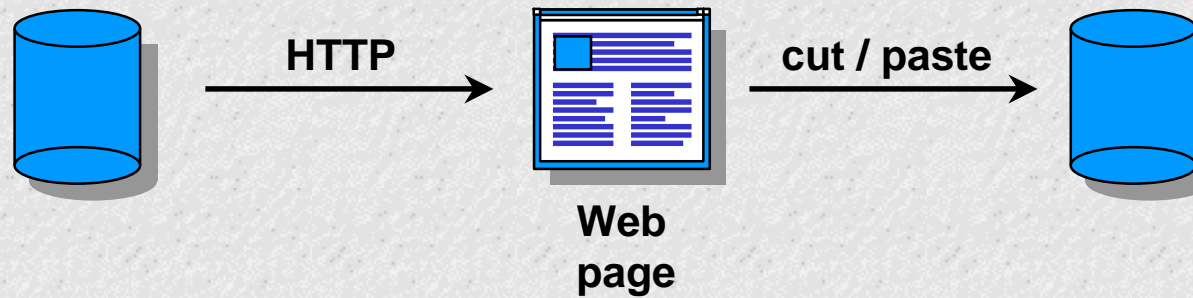


Extranet XML use: healthcare records

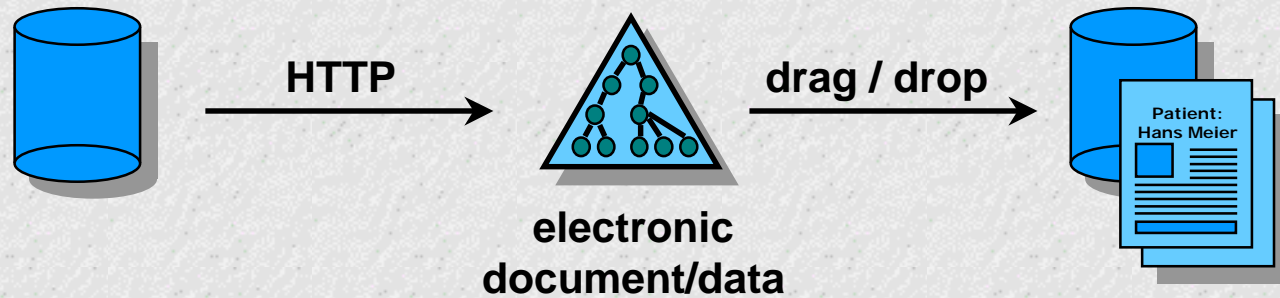
yesterday



today

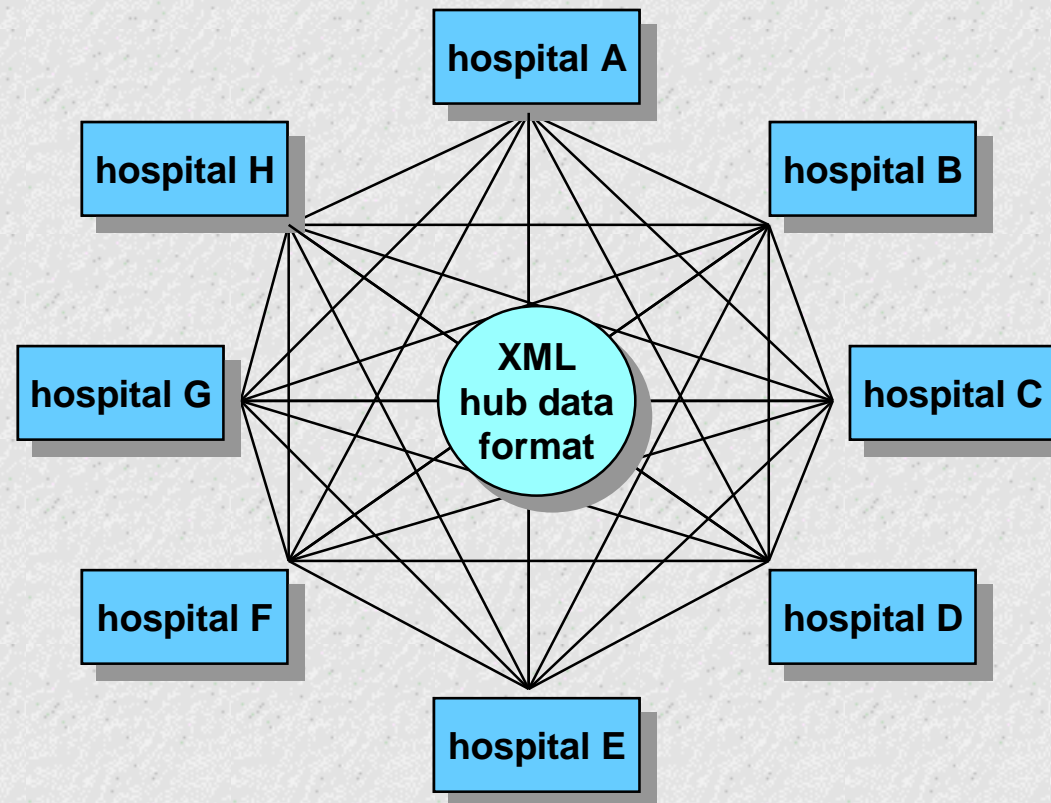


tomorrow





Extranet XML use: healthcare records



- n *Interchange98* XML Transaction Server (Sequoia, Microsoft)
 - manage, index, and transport patient data to / from client applications ranging from radiology equipment to hospital billing systems
 - uses the **M**aster **P**atient **I**ndex (MPI), based on XML, under development by the U.S. Commerce Department



Internet XML use: enhanced retrieval

n Aeneid's *Internet Research Assistant*

- rapid collection and navigation of Internet content based on XML meta-information about Web sites

The screenshot shows the 'Narrowcast Search' application window. The title bar reads 'Narrowcast Search' and 'High-Tech Catalog'. The interface includes a toolbar with icons for 'Expand', 'Expand all', 'Collapse', 'Collapse all', 'Up Level', 'Find', and 'Help'. Below the toolbar are tabs for 'Catalog', 'Search Sets', and 'My Search Sets'. The main area is divided into two panes. The left pane shows a tree view of categories: 'Government resources', 'Hardware', 'Legal', 'News and Publications', 'Business & Financial Sources', 'Financial Coverage', 'Economics', 'Investing', 'General Business News', and 'International'. The right pane is a table with columns 'Name' and 'Description', listing various sources like 'Alert-IPD', 'Barron's', 'Briefing.com', 'Business Week', 'Business Wire', 'CBS MarketWatch', 'CIO', 'CNN Interactive', and 'East Company'. Below the panes is a 'Selected Search Domains' table with columns 'Domain' and 'Catalog Origin', listing URLs and their corresponding sources. At the bottom right are buttons for 'Add New...', 'Remove', 'Save Set...', 'Previous', 'Next', 'Finish', and 'Cancel'.

Name	Description
Alert-IPD	Online and email notification of v
Barron's	Reporting on business, economi
Briefing.com	News, commentary and analysis
Business Week	In-depth coverage of broad-bas
Business Wire	Sections on technology, health,
CBS MarketWatch	General business, finance, ecor
CIO	CIO Online is directed at informa
CNN Interactive	Global news topics covering US
East Company	East Company covers emerging

Domain	Catalog Origin
http://www.asiaecon.com/APER/	Asia Pacific Economic Review
http://www.frontiernet.net/~cahners/	Cahners Economics
http://www.cipe.org/ert.html	Economic Reform Today
http://www.feer.com/	Far Eastern Economic Review
http://www.hbsp.harvard.edu/noframes/groups/	Harvard Business Review
http://mof.com/moharom.htm	Media General Daily Market Barometer



Conclusions and questions

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



Steps towards XML in your enterprise

n Step 1: Learn XML

- convert/create and edit XML
- view and manage XML

n Step 2: Think XML

- create open document formats
- create open data formats
- create open applications
 - company-specific or industry-general

n Step 3: Use XML

- intranet document management
- extranet data exchange
- internet e-commerce
- ...



The XML revolution

- n XML is *the*
 - open document format
 - open data format
- n we are only at the very start of the **networked information revolution**
- n with XML, networked documents and data will become the foundation of the **networked enterprise**

**XML is the ASCII
of the 21st century**



n **Networked document engineering**

- document conversion & processing
- Internet & intranet information systems
- XML/SGML-based document management

n **XML courses:**

- management track: where / when / how to best use XML (1/2 day)
- technical track: hands-on XML/XSL training (2 or 3 days)

n **XML support site: www.xmlcenter.com**

- please send your XML questions to info@xmlcenter.com!