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

- 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

- 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 **à** is SGML - -
    - optimize SGML for Web creation and delivery
    - an extended, richer version of HTML **à** is *not* HTML + +
    - a leaner, meaner subset of SGML for use on the Internet
- Initiative of WG8 of the World Wide Web Consortium (W3C)
  - vision expanded to include push, metadata, transactions, ...

XML history

- 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

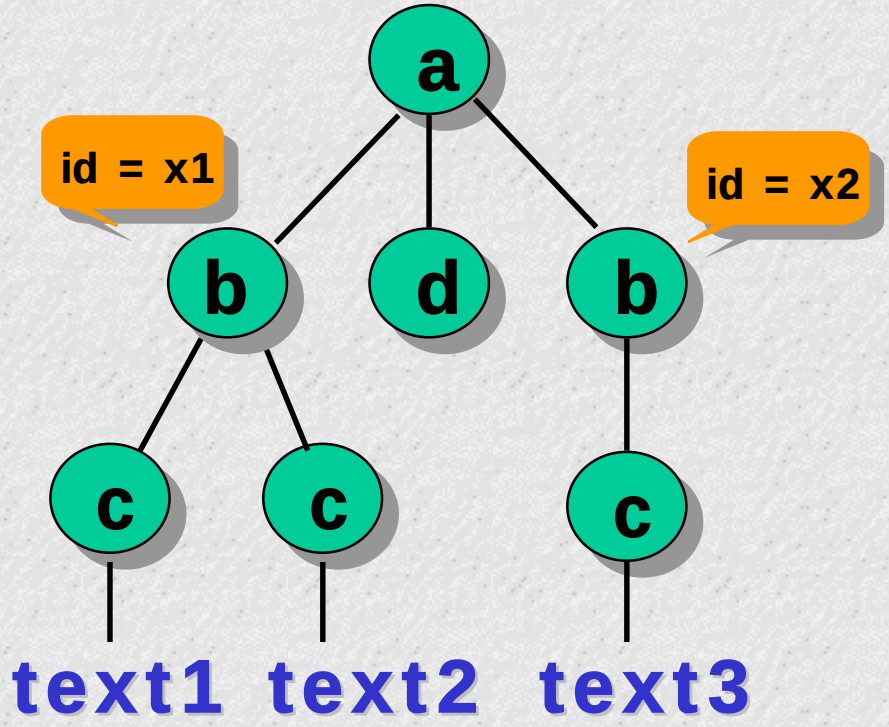
- XML shall be straightforwardly usable over the Internet
- XML shall support a wide variety of applications
- XML shall be fully compatible with SGML
- It shall be easy to write programs which process XML
- The number of optional features in XML is to be kept to the absolute minimum, ideally zero
- 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

- 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
- 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
- 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

- XML is a document format
  - markup to capture the meaning of content
    - intelligent searching, filtering, …
  - markup to verify the correctness of structure

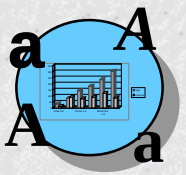
  è open document computing applications
- XML is a data format
  - markup to capture the meaning of information
    - intelligent processing, extracting, …
  - markup to verify the completeness of definition

  è open data processing applications
- XML will end the distinction between
  - networked document distribution
  - networked data transactions

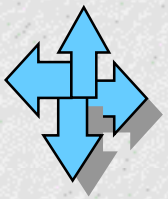
Key concept: XML is a set of solutions

Related standards

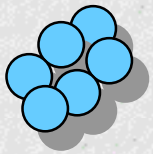
XML : Extensible Markup Language



XSL : Extensible Stylesheet Language



XLL : Extensible Linking Language



DOM: Document Object Model

XSL: Extensible Stylesheet Language

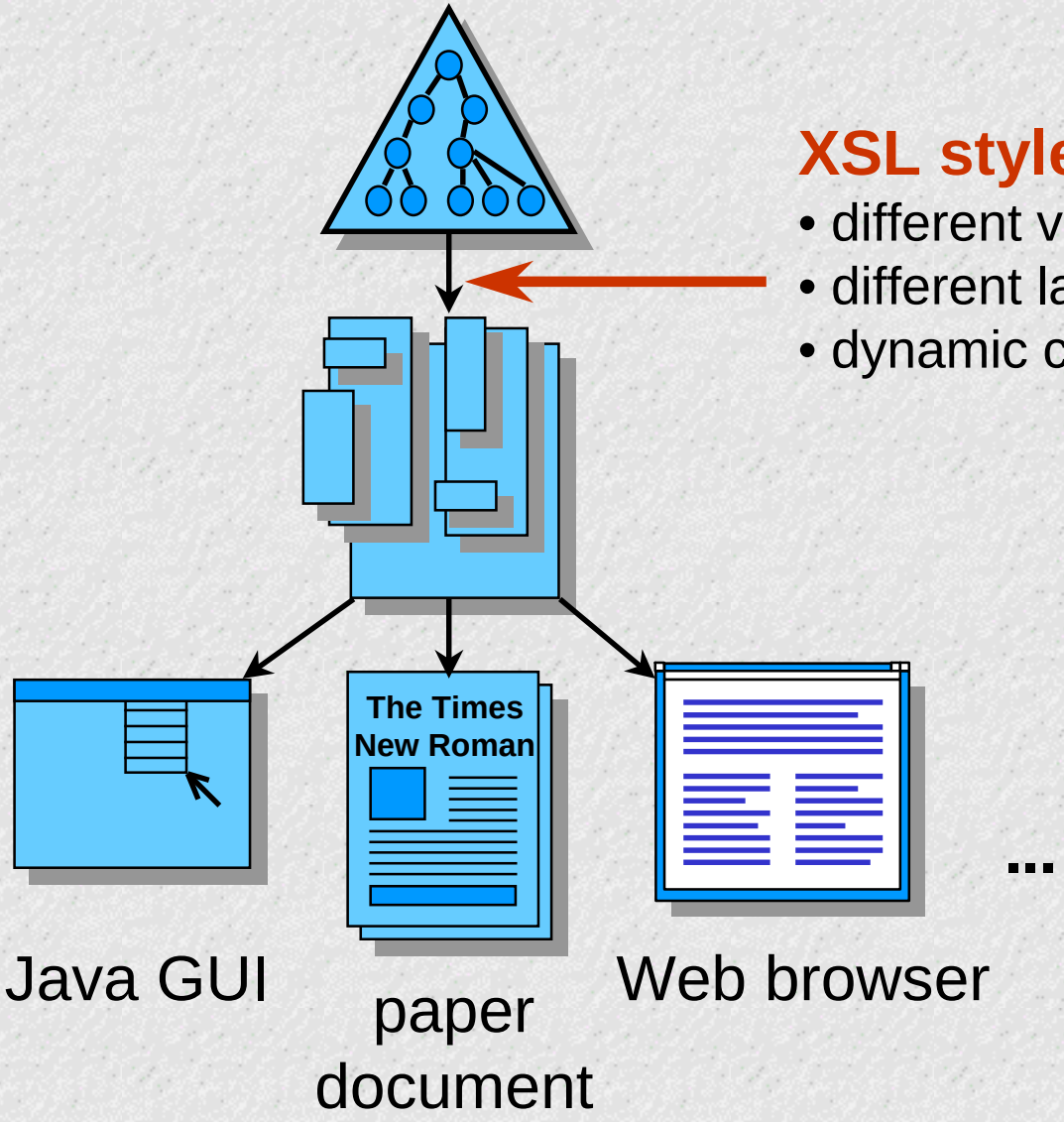
- XML stylesheet mechanism
  - based on DSSSL
  - compatible with CSS 1.0 and 2.0
  - advanced capabilities:
    - reordering contents
    - automatically generated text
    - for both printing and online display
- is useful for:
  - handling presentation separate from content
  - single source, multiple output media
  - reformatting / re-using content

XSL: Extensible Stylesheet Language

element grove

flow object tree

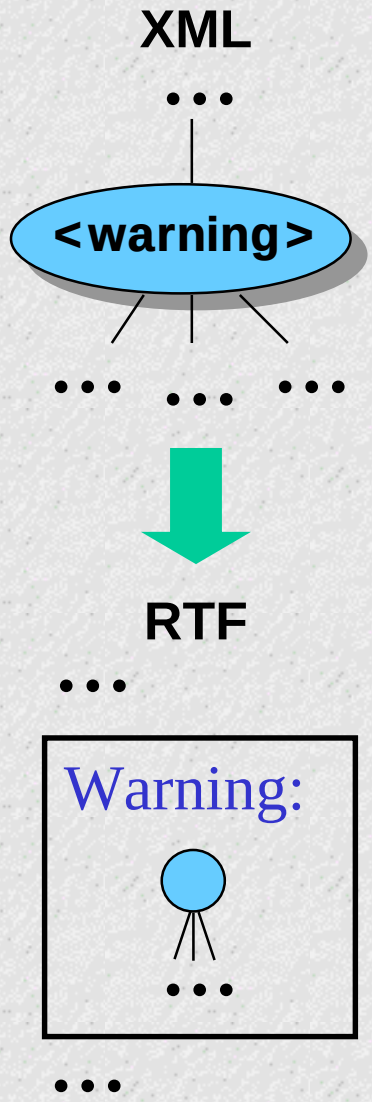
output media



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

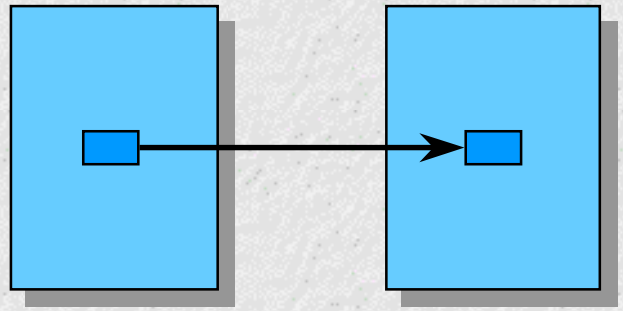
pattern

action

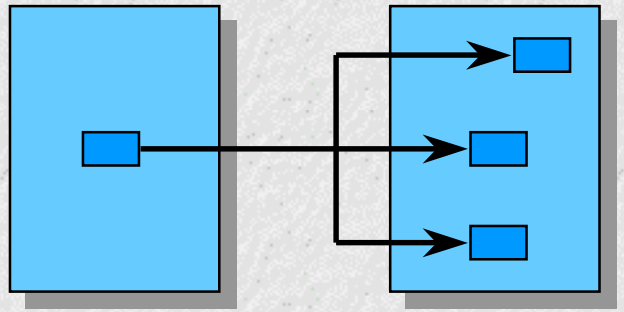
XLL: Extensible Linking Language

- 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
- is useful for:
  - rich hypertext functionality
  - computer-based training
  - Web site management

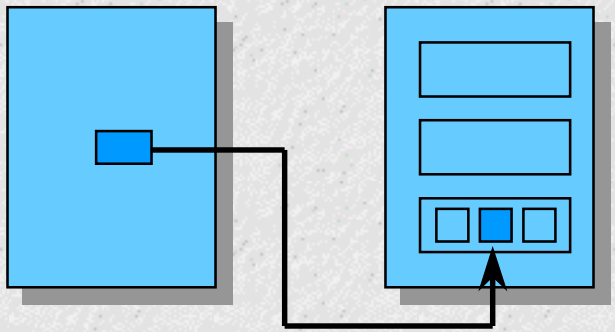
XLL: Extensible Linking Language



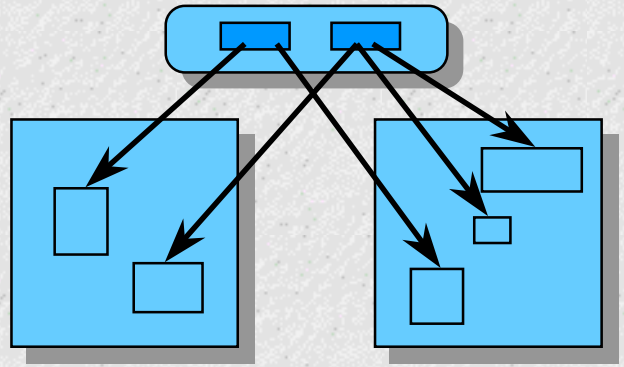
simple link



extended link



symbolic link

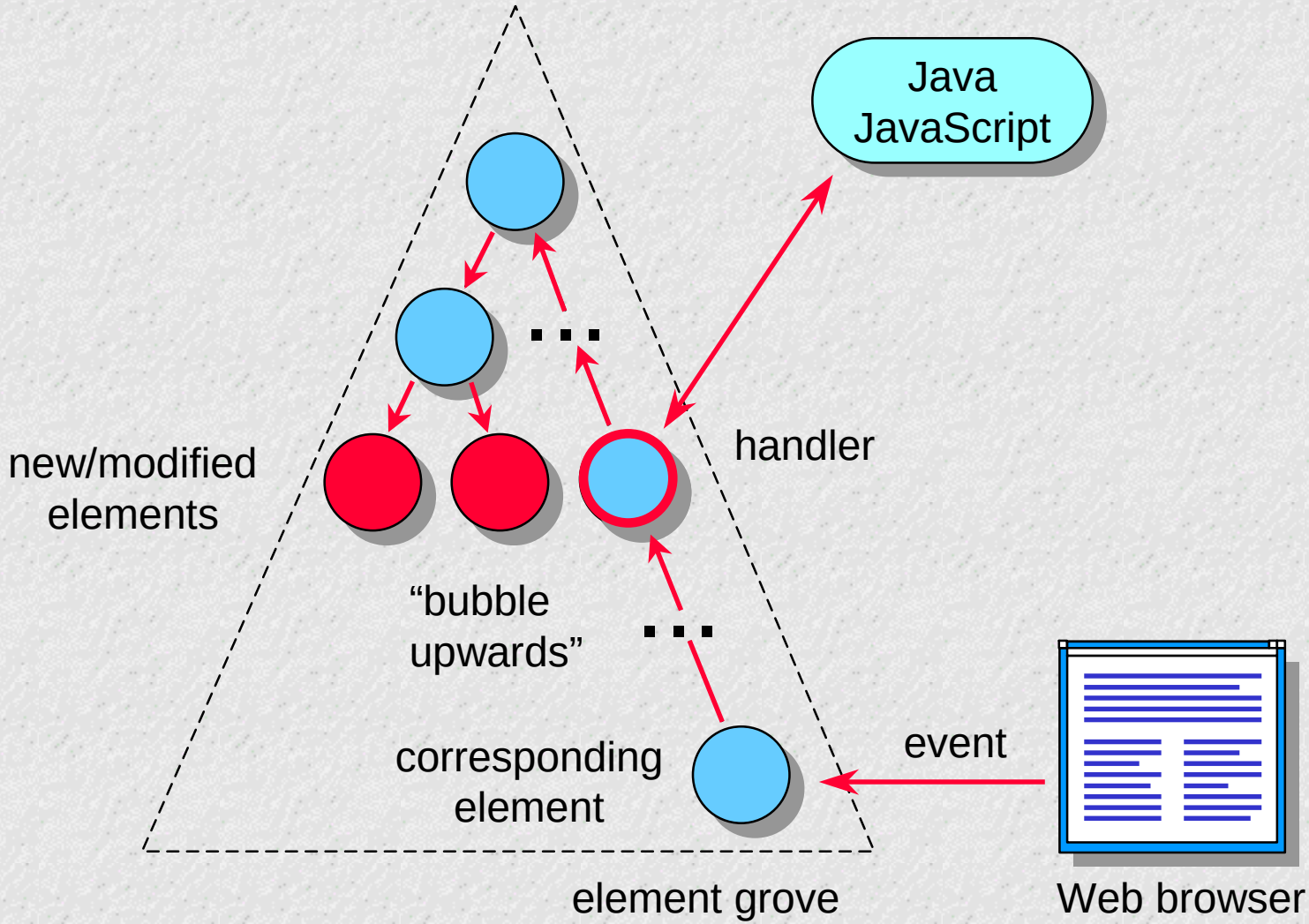


link group

DOM: Document Object Model

- 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
- 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.

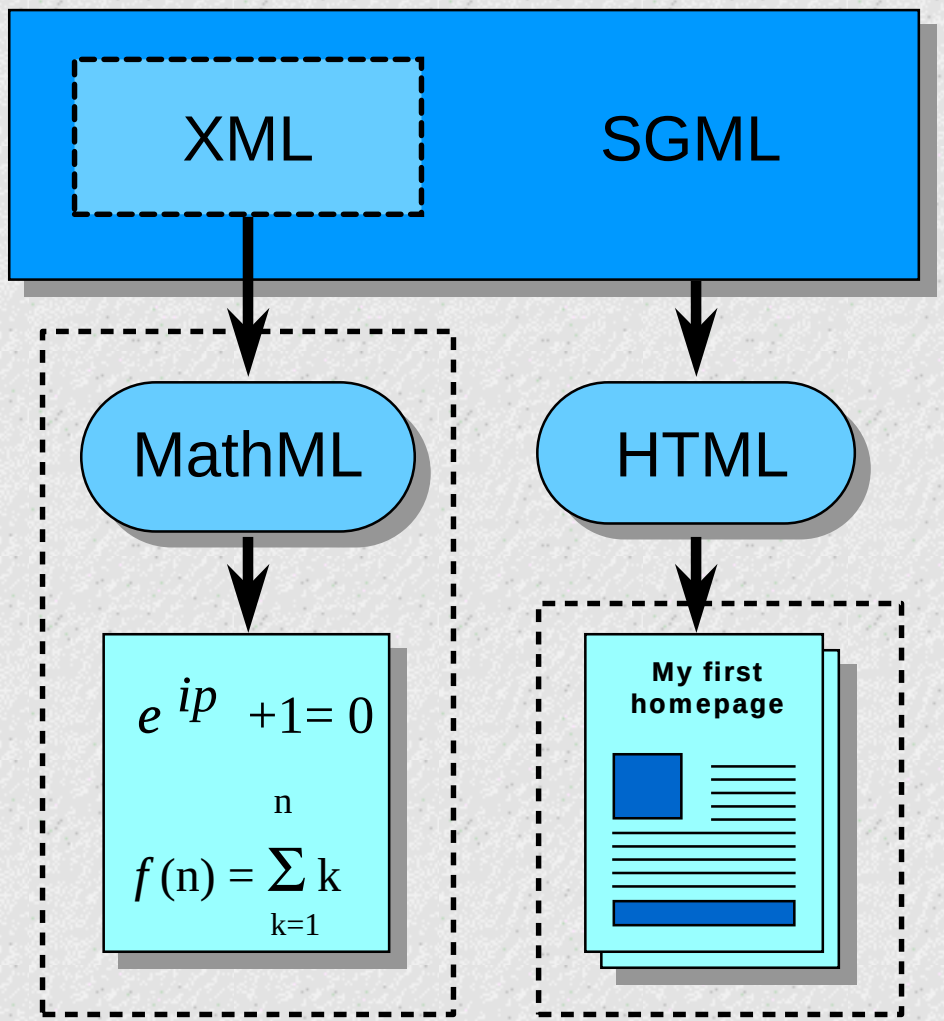
"Office Future International Services"

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

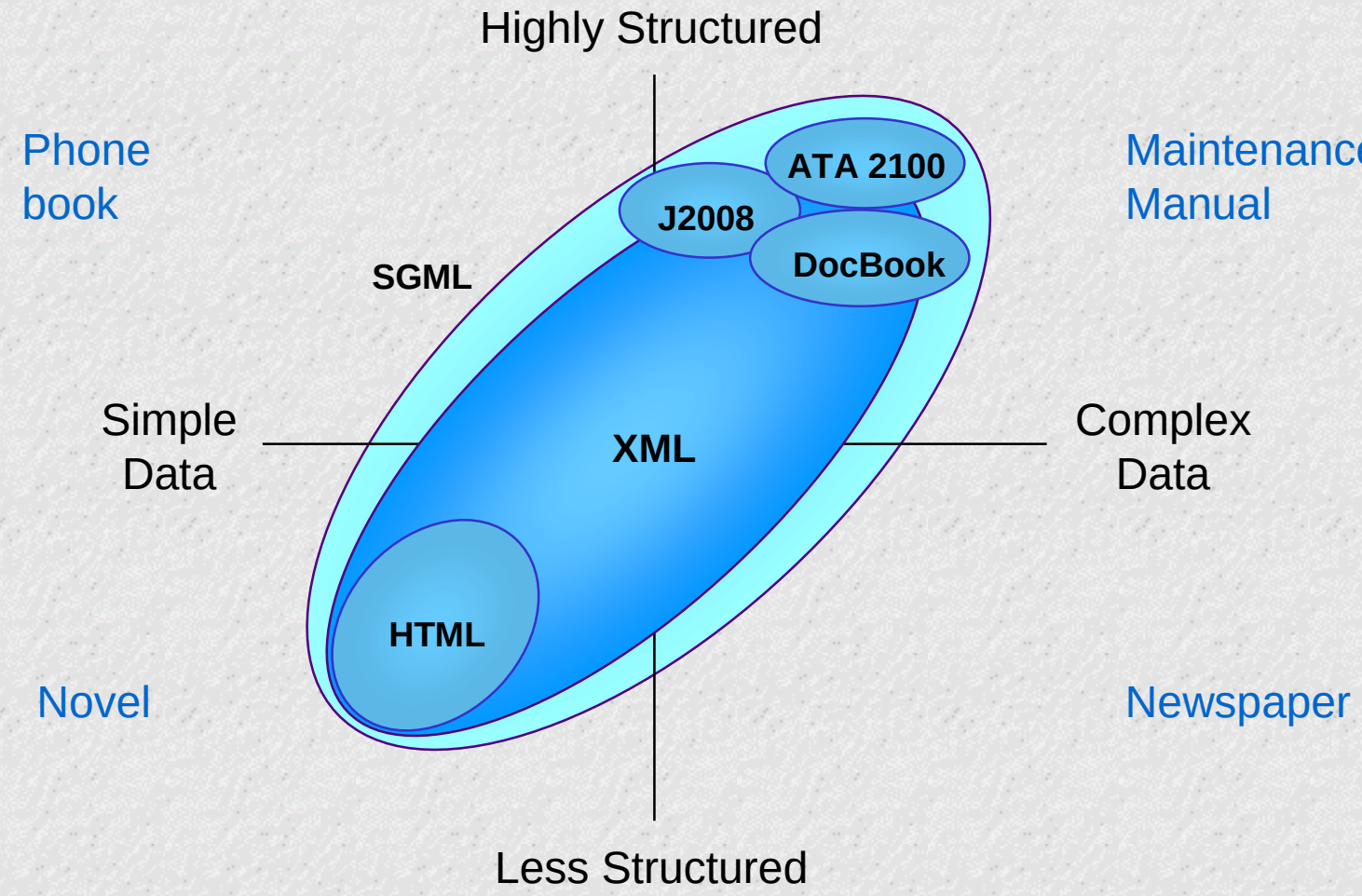


document architecture

classes of documents

actual document

XML compared to HTML / SGML

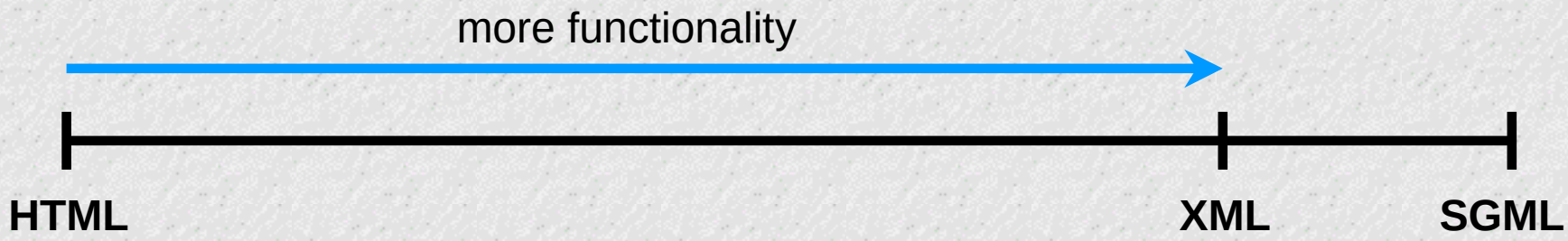


XML compared to HTML / SGML

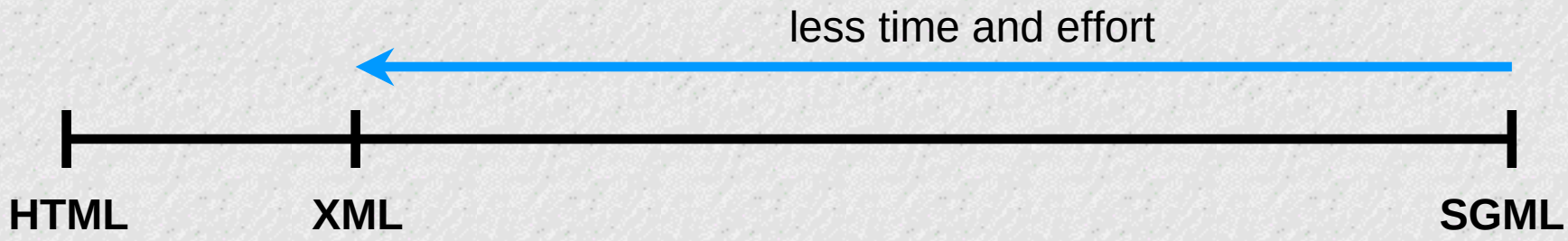
- XML matches SGML
  - captures structure & meaning
  - future-proof standard
  - repurpose and reuse
  - strength of validation
  - ease of automation
- XML improves on SGML
  - reduces optional and advanced features
  - machine-processable without a DTD
  - mainstream browser and tool support
  - standardized stylesheets (XSL)
  - standardized linking (XLL)

XML compared to HTML / SGML

- 80% of the SGML functionality ...

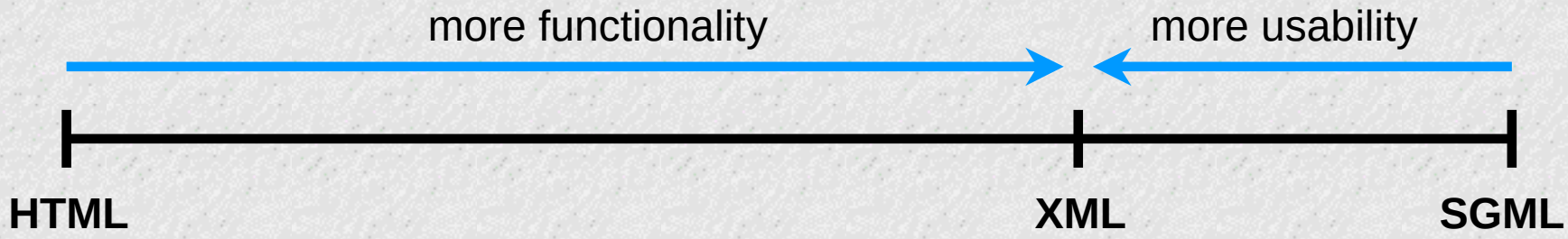


- … for 20% of the cost in time and effort

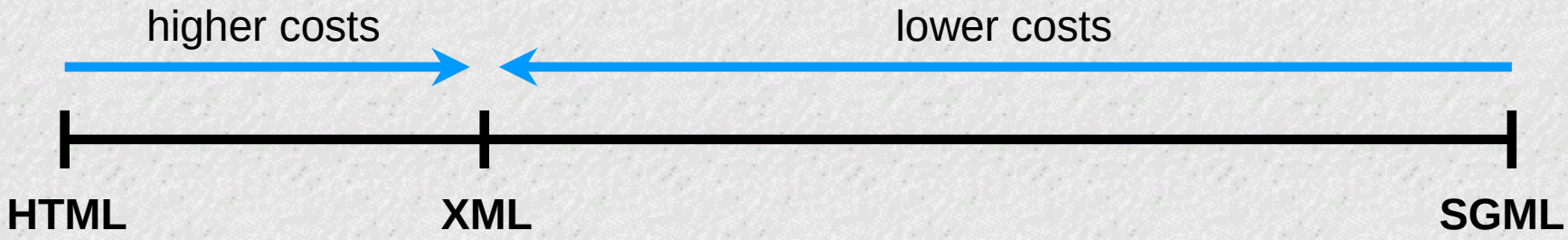


XML compared to HTML / SGML

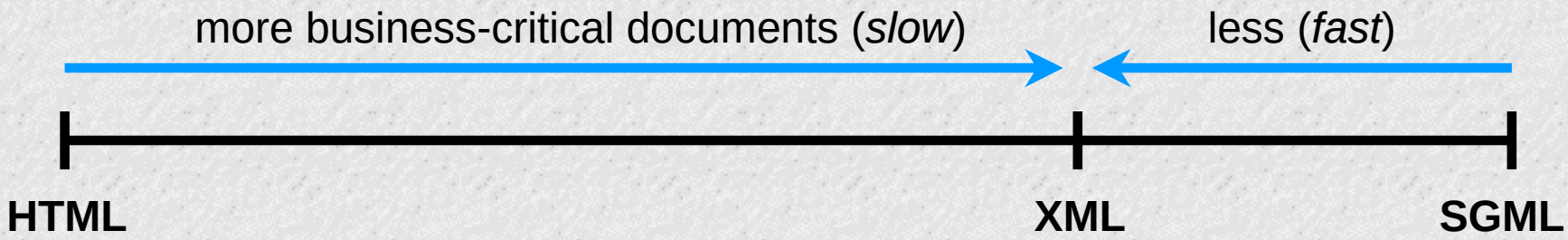
- Application capabilities?



- Application price?



- 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

- XML parsers
  - Microsoft XML parsers (C, Java) + XML DSO and XML OM in IE4
  - IBM XML for Java parser, DataChannel XML parser (Java), ...
- 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 *Konstructor*
    - development suite for XML content management & delivery applications
- 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?

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

XML as a document format: when to use?

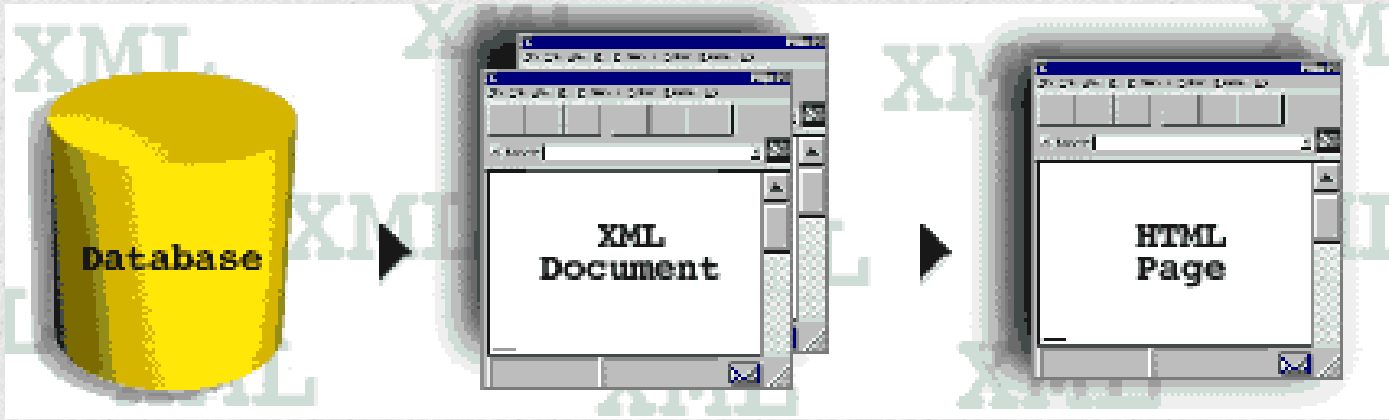
- preserving documents over time
  - future-proofing documents against tools and context
    - è machine-readable format: for parsers & generators
    - è self-descriptive format: for extraction and validation
- preserving documents over space
  - distributed systems with decentralized documents
    - è common documents definitions
    - è link to networked resources
- 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?

- Focus on intelligent document handling on the Web
- 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)
- 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

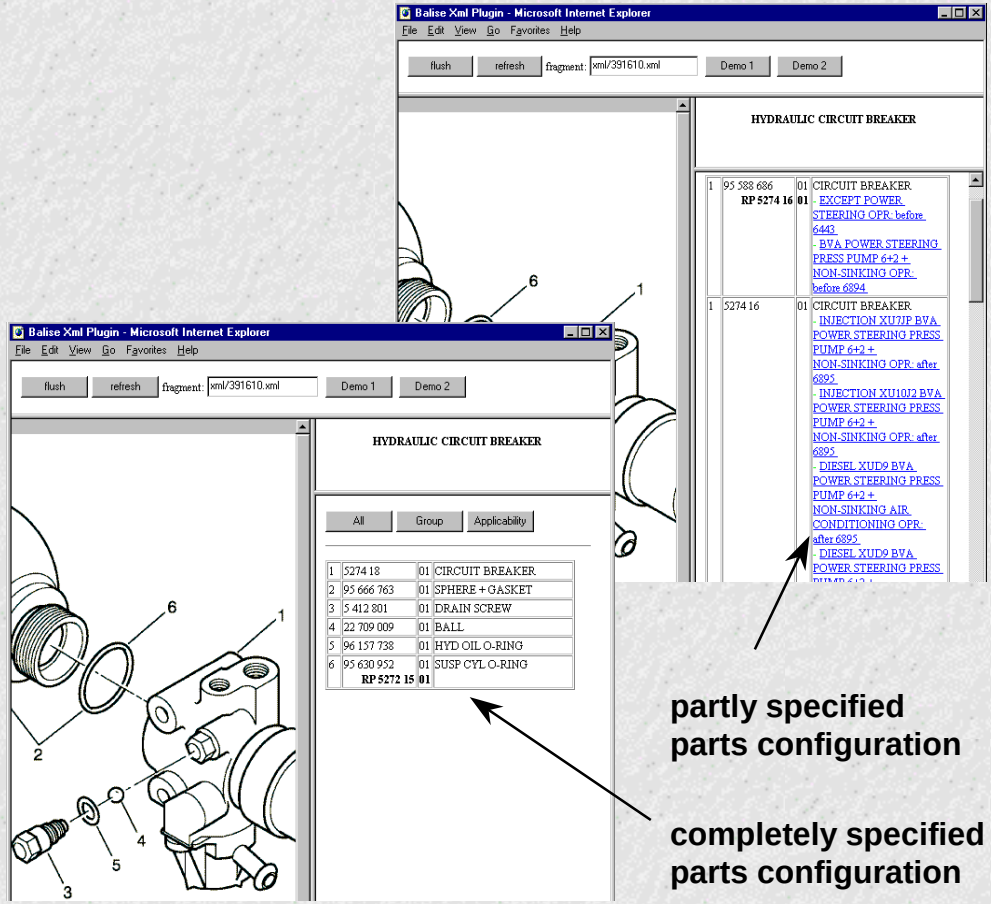
- XML for consultation of electronic documents



  - database of modular document components
  - XML-based virtual document
  - HTML-based actual document
- Used for:
  - self-service manuals
  - process/procedures documentation

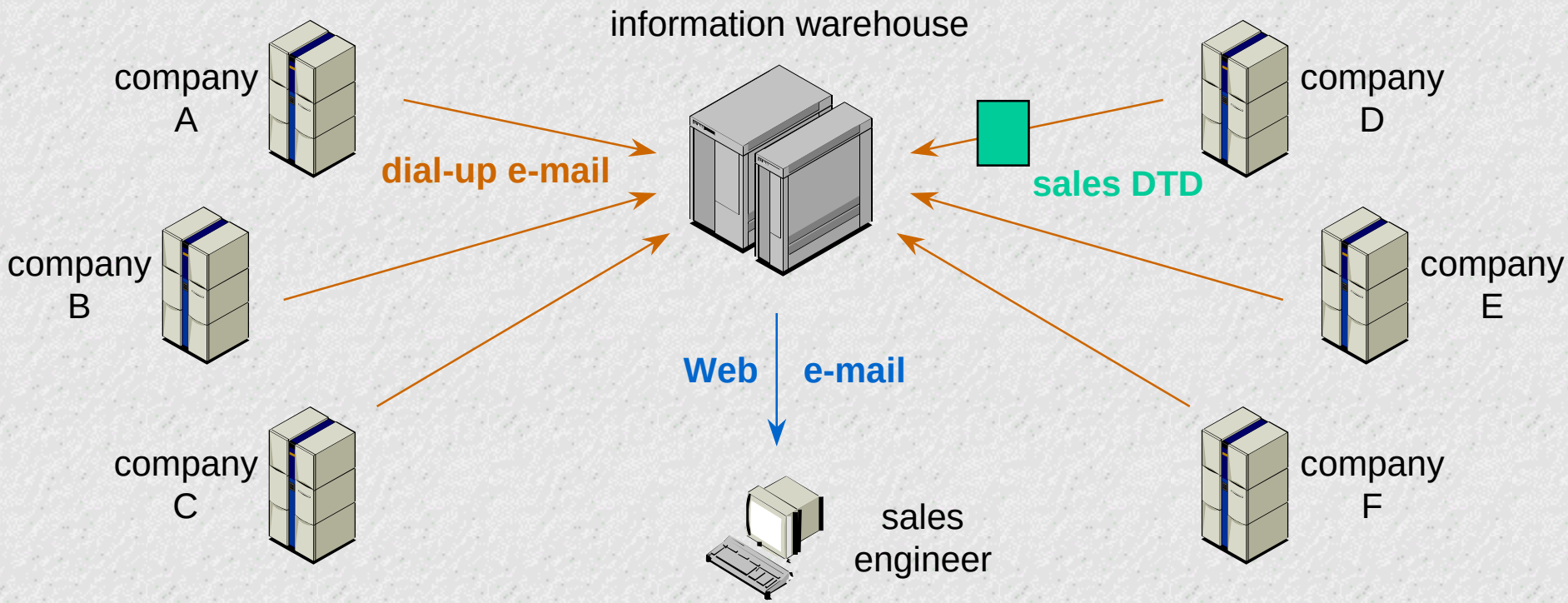
Example self-service maintenance manual

- server-side
  - parts information (XML fragments)
  - configuration information (XML hyperlinks)
- browser-side
  - fragment selection (JavaScript/VBscript)
  - fragment assembly (XML document)
  - fragment display (XSL stylesheet)
  - document display (HTML document)

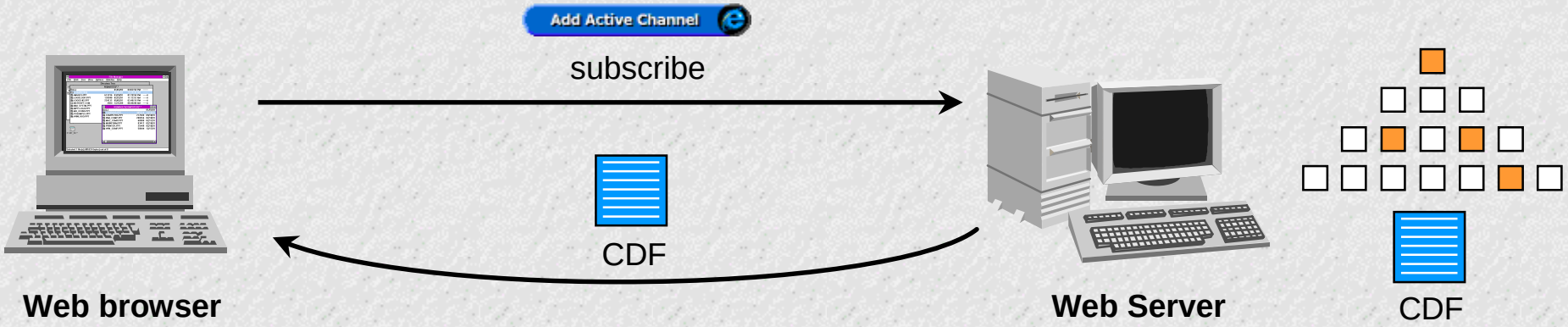


Extranet XML use: information aggregation

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

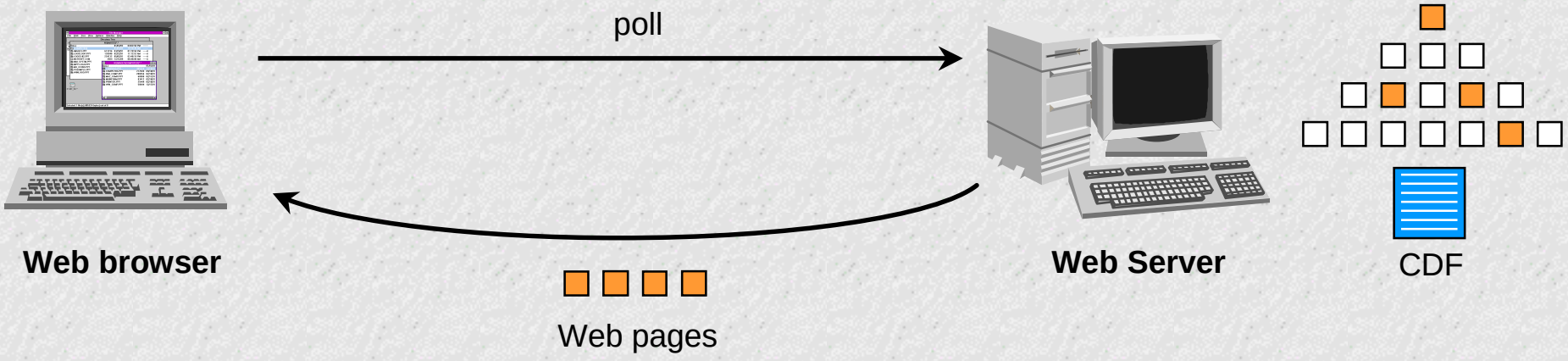


Internet XML use: information push



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

Internet XML use: information push

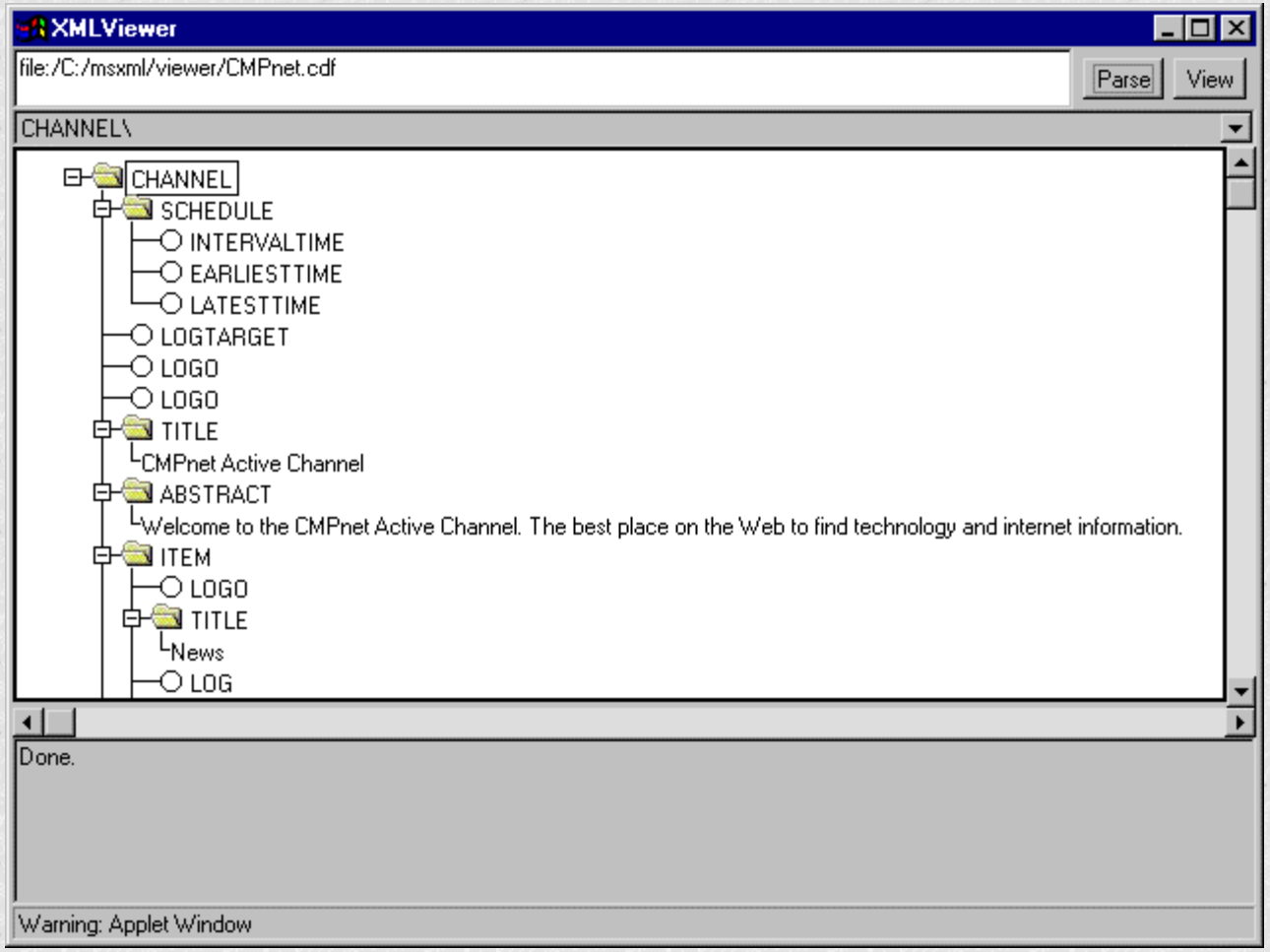


- client polls Web site for push of changed content
- server sends changed pages according to CDF directives

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



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?

- The problem with data representation
  - proprietary data formats
    - client/server
    - transaction-based processing
  - proprietary middleware solutions
    - data converters
    - gateway applications
    - migration tools
- Typical operations on data
  - the definition of data models
  - transactional processing
    - data transmission
    - verification
    - archival
  - migration

XML as a data format: when to use?

- Description of syntactical data schema
  - DTD's
  - appropriate document formats
    - 'well-formed'
    - hierarchical
- Description of conceptual data schema
  - concepts
    - classes
    - objects
    - properties
  - … and their relationships
    - RDBMS
    - OODBMS

XML as a data format: when to use?

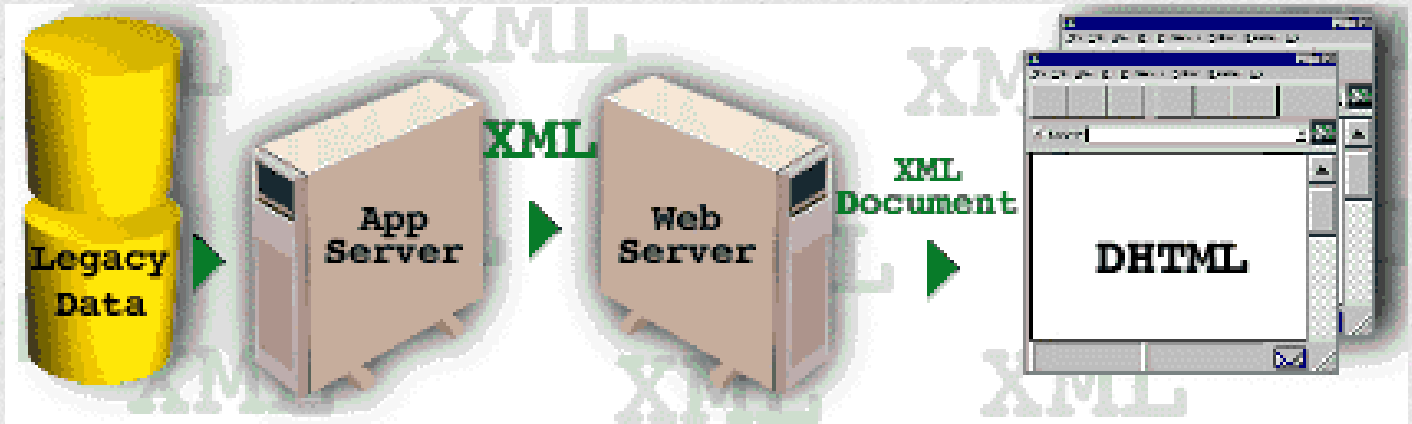
- Description of metadata
  - data about data
    - MCF: Meta Content Format
    - XMI: XML Metadata Interchange Format
  - enhanced retrieval
  - autogeneration of hub documents
    - keyword-based
    - thesauri
- Description of data locators
  - where to find what kind of data on the Web
    - RDF: Resource Description Framework
- Message-based transactions
  - application interactions between servers over standard Web protocols
    - WIDL: Web Interface Definition Language

XML for data exchange: what's available?

- Focus on open data exchange between different applications
- 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)
- 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

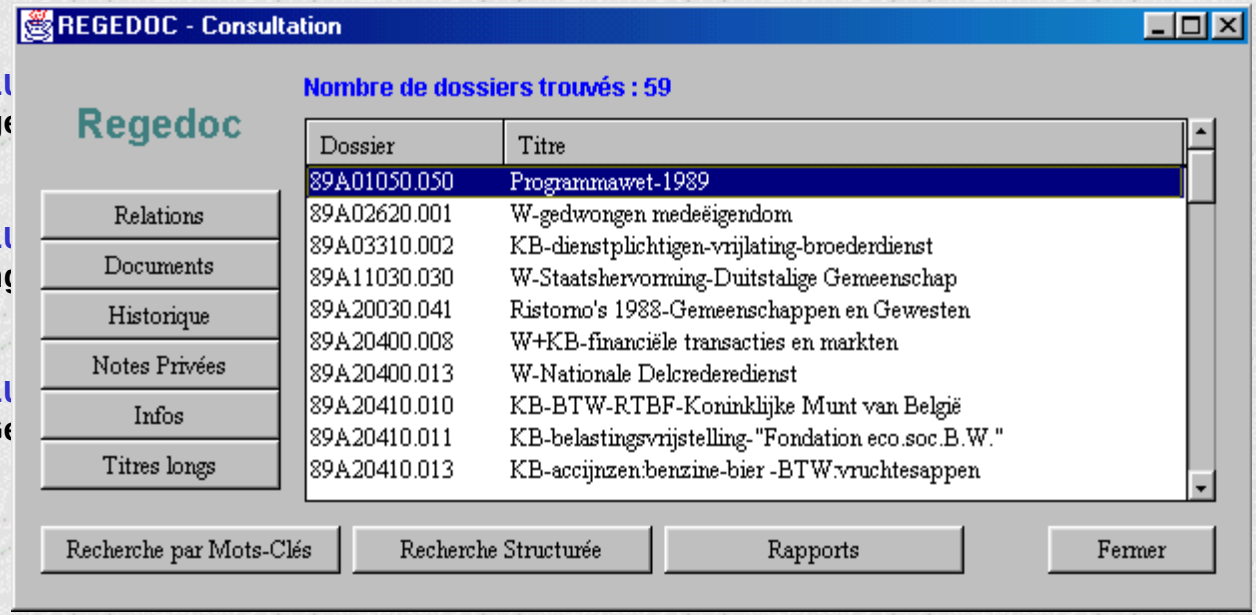
- XML for dynamic interaction with legacy data



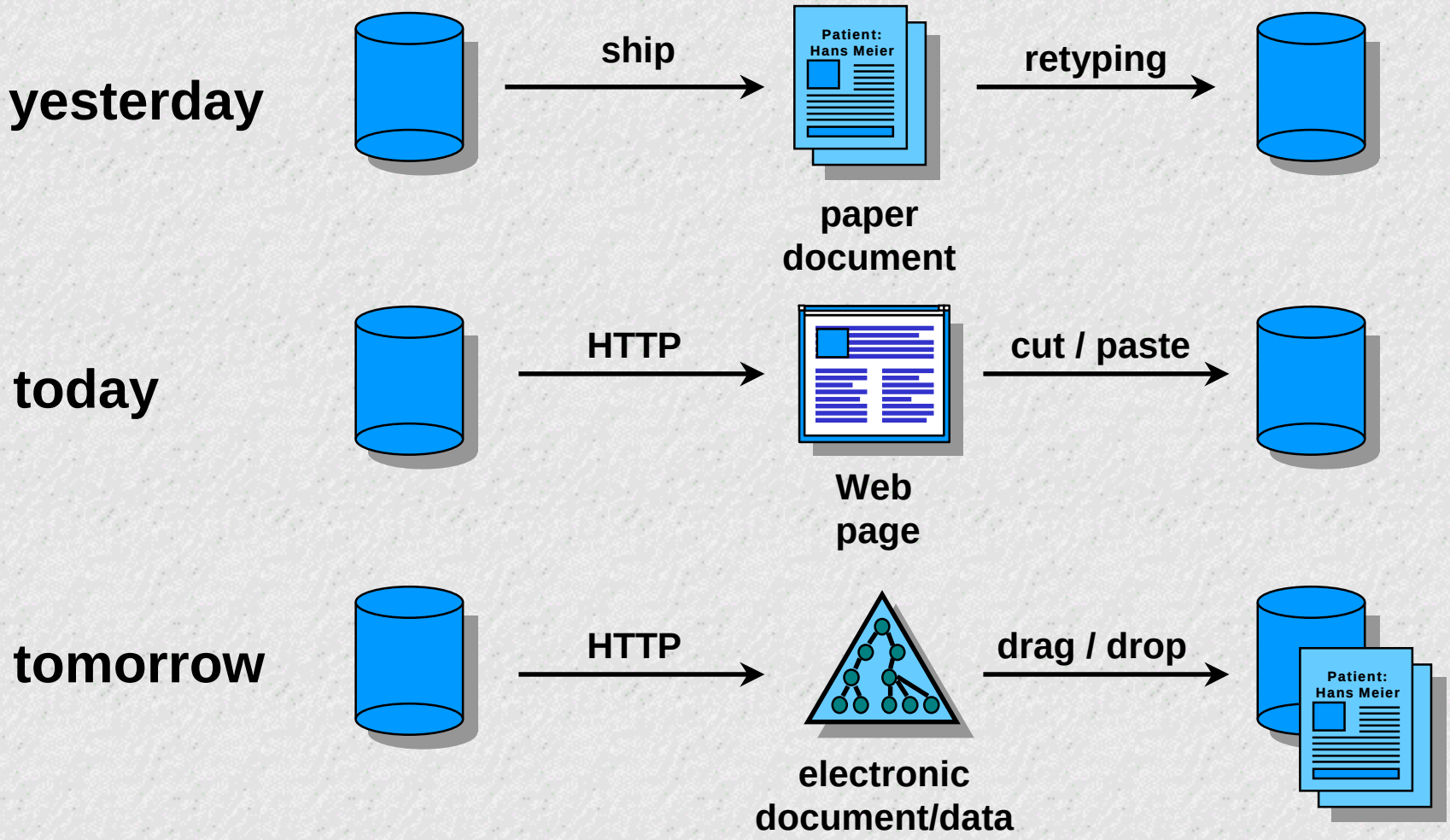
  - legacy database
  - XML-based transaction server
  - DHTML-based browser user interface or Java client
- Used for:
  - web-enabling legacy databases
  - data-enriching of Web sites

Intranet XML use: legacy database access

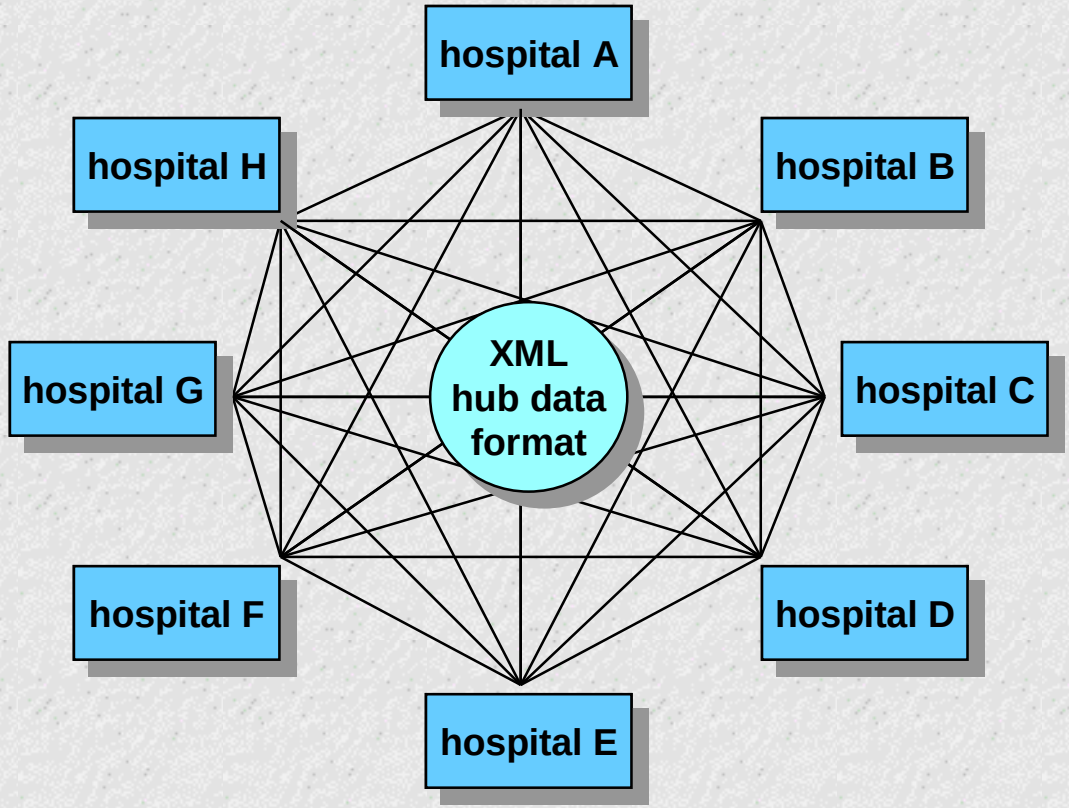
```
<?XML VERSION="1.0" ENCODING="UTF-8" STANDALONE="yes"?>
<REGEDOC>
<TABLE NAME="DOSSIER" NUMRECORDS="59" NUMFIELDS="2">
<RECORDSPEC>
<FIELDSPEC TYPE="Character" OCCURS="single">NUMBER</FIELDSPEC>
<FIELDSPEC TYPE="Character" OCCURS="single">DUTCH_TITLE</FIELDSPEC>
</RECORDSPEC>
<RECORD>
<FIELD><VALUE>89A01050.050</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</VALU
<FIELD><VALUE>KB-dienstplichtige
</RECORD>
<RECORD>
<FIELD><VALUE>89A11030.030</VALU
<FIELD><VALUE>W-Staatshervorming
</RECORD>
<RECORD>
<FIELD><VALUE>89A20030.041</VALU
<FIELD><VALUE>Ristorno's 1988-Ge
</RECORD>
…
</TABLE>
</REGEDOC>
```



Extranet XML use: healthcare records



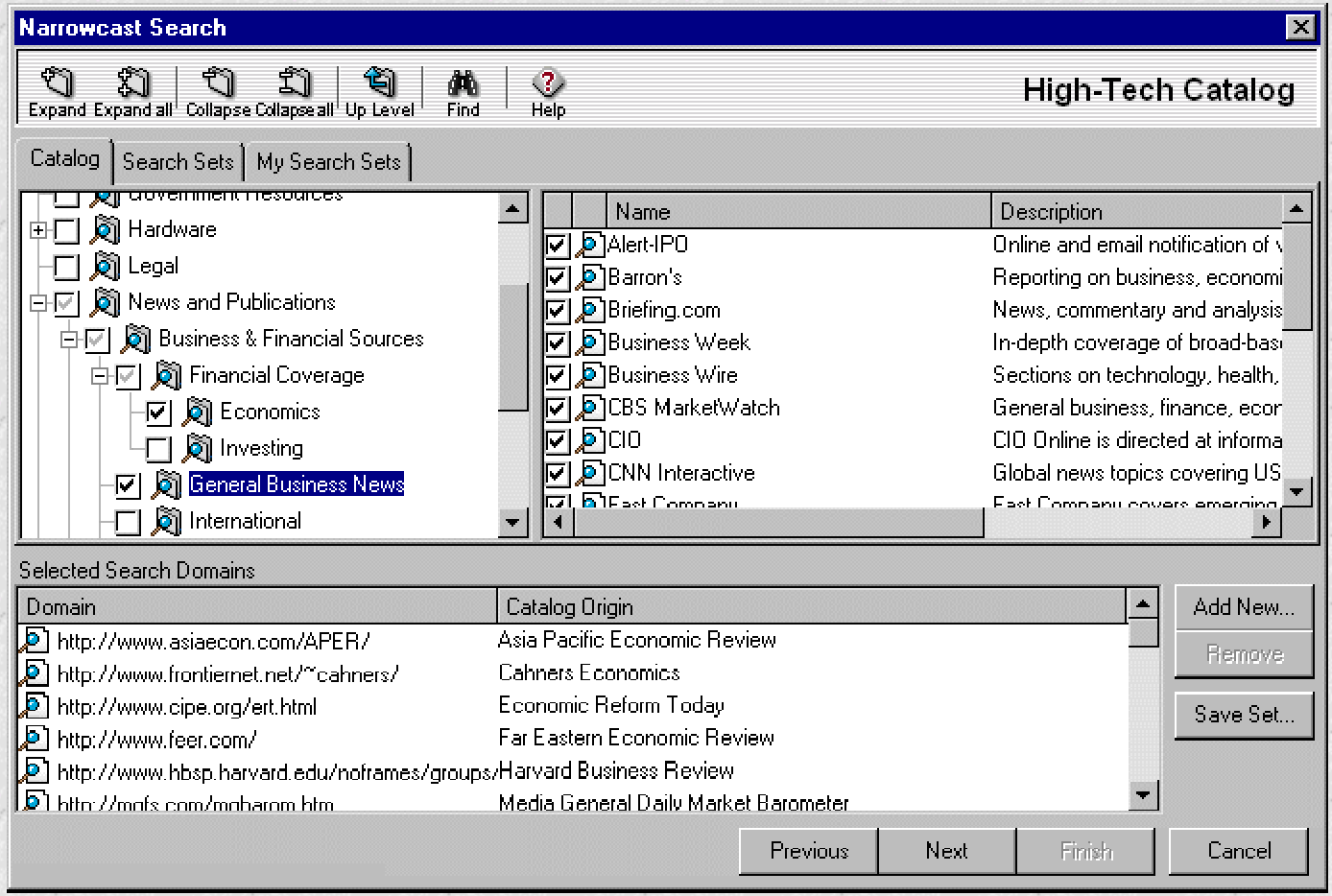
Extranet XML use: healthcare records



- *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 Master Patient Index (MPI), based on XML, under development by the U.S. Commerce Department

Internet XML use: enhanced retrieval

- Aeneid's *Internet Research Assistant*
  - rapid collection and navigation of Internet content based on XML meta-information about Web sites



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

- Step 1: Learn XML
  - convert/create and edit XML
  - view and manage XML
- Step 2: Think XML
  - create open document formats
  - create open data formats
  - create open applications
    - company-specific or industry-general
- Step 3: Use XML
  - intranet document management
  - extranet data exchange
  - internet e-commerce
  - …

The XML revolution

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

XML is the ASCII of the 21st century

OFFIS services

- **Networked document engineering**
  - document conversion & processing
  - Internet & intranet information systems
  - XML/SGML-based document management
- **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)
- **XML support site: www.xmlcenter.com**
  - please send your XML questions to info@xmlcenter.com!