XML: the document format of the future?

Hans C. Arents

senior IT market analyst

I.T. Works

"Guiding the IT Professional"

Innovation Center, Technologiepark 3, B-9052 Gent (Belgium), Tel: +32 (0)9 241 56 21 - Fax: +32 (0)9 241 56 56

E-mail: hca@itworks.be - Site: http://www.itworks.be/ - Home: http://www.arents.be/

XML: the document format of the future?

- Why XML?
- What is XML?
- Examples of XML use
  - eBook
  - WAP / WML
- What's the best document format?
  - HTML vs. PDF vs. XML
- What are the document design goals?
  - from the reader's perspective
  - from the publisher's perspective
- Comparing XML with HTML and PDF
- Conclusions

Why XML?

- Beyond HTML, instead of SGML:
  - problem: extending HTML (**H**yper**t**ext **M**arkup **L**anguage)
    - too simplistic
    - unmanageable
    - useless for applications
  - suggested solution: SGML (**S**tandardized **G**eneral **M**arkup **L**anguage)
    - ü extensible by definition
    - ü has all the necessary mechanisms
    - û big, complex and (sometimes very) hard to use
    - û hated by Web developers, misunderstood by Web users
  - real solution: **XML** (E**x**tensible **M**arkup **L**anguage)
    - throw the hard parts of SGML away **à** is SGML – –
    - an extended, richer version of HTML **à** is *not* HTML + +

Why XML?

- Milestones:

Jul	‘96	W3C XML Working Group is formed
Nov	‘96	First draft of XML standard published
Oct	‘97	Microsoft ships IE 4.0 with 2 built-in XML parsers
Feb	‘98	Final XML 1.0 standard officially approved by W3C
Q3	‘98	Availability of commercial XML tools and technologies
Q4	‘98	Announcements of XML support: IBM, Sun, Oracle, SAP, ...
Jan	‘99	Microsoft announces XML native data format in Office 2000
Apr	‘99	Microsoft ships IE 5.0 with (not quite) complete XML support
Q4	‘99	Netscape will ship Navigator 5.0 with complete XML support
Q1	Y2K	Industry-wide and cross-platform adoption of XML - for document applications - for data applications

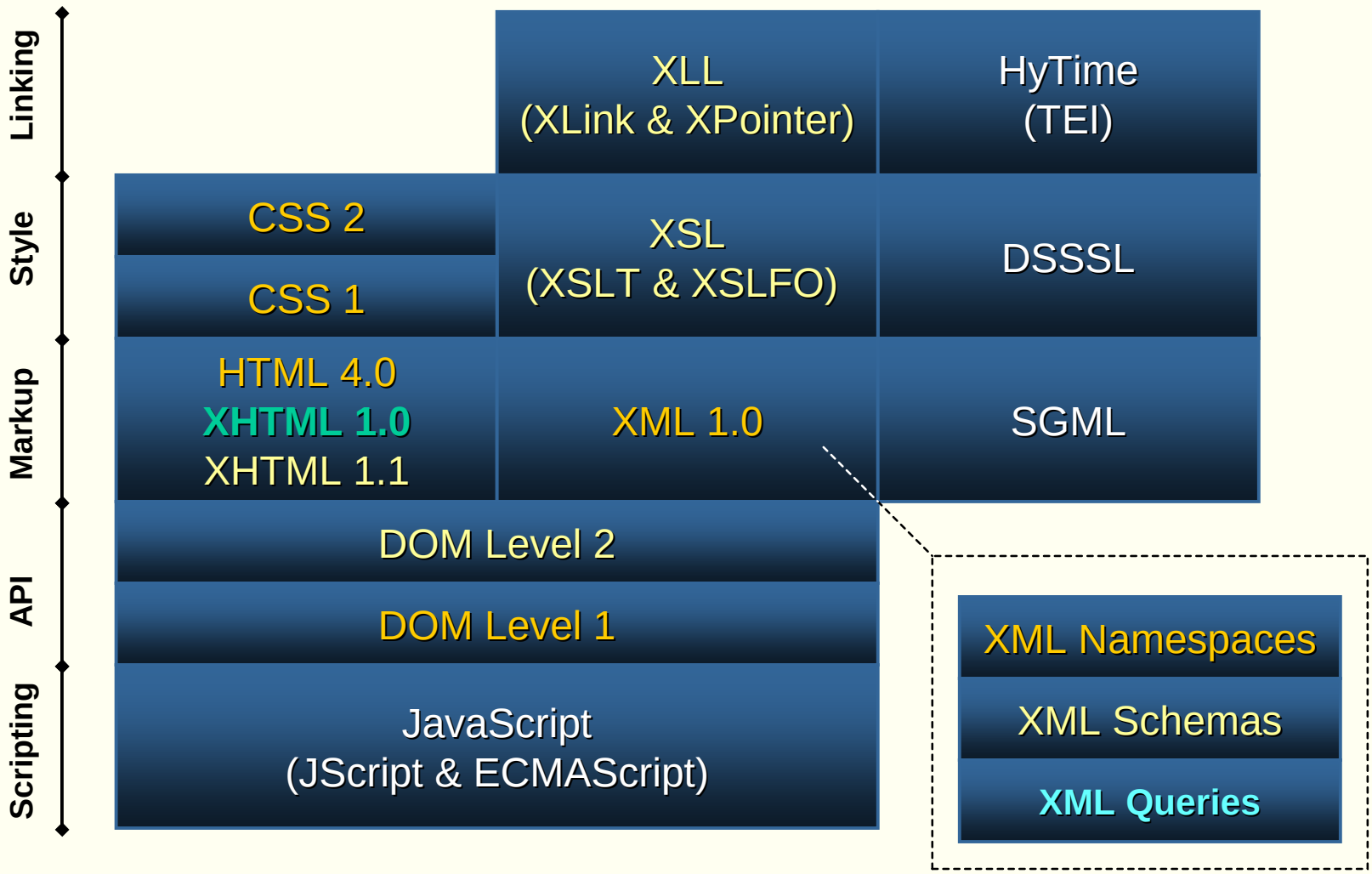
XML is about structure and meaning

```
<?xml version="1.0" standalone="yes"?>
<order>
  <customer>
    <person><lastname>Layman</lastname><firstname>Andrew</firstname></person>
  </customer>
  <sold-on>19970317</sold-on>
  <item>
    <price>5.95</price>
    <book>
      <title>Mathematics, the Language of Science</title>
      <author>Dantzig, Tobias</author><isbn>0-452-01030-6</isbn>
    </book>
    <review>Once more, Dantzig explores the fascinating world of mathematics,
    this time by examining how mathematics is used as a tool for ...</review>
  </item>
  <item>
    <price>12.95</price>
    <record>
      <title><composer>Tchaikovsky</composer>'s First Piano Concerto</title>
      <style>classical music</style><artist>Janos</artist>
    </record>
    <review>As so many recordings of this classical masterpiece, this version
    of Tchaikovsky's First Piano Concerto suffers from severe ...</review>
  </item>
</order>
```

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
  - **à** open document computing applications
- XML is a **data format**
  - markup to capture the meaning of information
    - intelligent processing, extracting, …
  - markup to enable the exchange of information
  - **à** open data processing applications
- XML will **end the distinction** between
  - networked document distribution
  - networked data transactions

XML is a whole family of standards



W3C: Note > Working Draft > Proposed Recommendation > Recommendation / Not W3C

Examples of XML use

- In active use:
  - XHTML: Extensible Hypertext Markup Language
    - reformulating HTML to ensure its survival in an XML world
  - WAP-WML: Wireless Access Protocol - Wireless Markup Language
    - for use in specifying content and user interface for narrowband devices, including personal digital assistants, cellular phones and pagers
- Under development:
  - Open eBook Publication Structure
    - for creating reading material for portable e-books
  - XFA: XML Forms Architecture

    XFDL: Extensible Forms Description Language
    - for the capture, presentation, transfer, signing, and processing of e-forms
  - MathML: Mathematical Markup Language
    - for describing mathematical formulas (both presentation and content)

Open eBook Publication Structure

- **eBook**
- What?
  - single, universal format for electronic books
  - based on the use of a smart combination of HTML and XML
- Why?
  - accelerate the wide availability of cheap, secured electronic reading material
  - support a healthy e-publishing industry

WAP / WML

- **W**ireless **M**arkup **L**anguage
- What?
  - specifying document content and user interface for narrowband wireless devices
- Why?
  - wireless networks and devices have specific needs and requirements not addressed by existing Internet technologies
  - WAP uses plain Web HTTP 1.1 servers
  - WML + WMLScript support
    - user interface metaphor: cards with hyperlinks
    - state management, input validation, extensions, ...

  for pushed content, real-time news bulletins, ...

What's the best document format?

- HTML = **H**yper**t**ext **M**arkup **L**anguage
  - ü reasons to love: cheap, usable, light, whole world is already using it
  - û reasons to hate: weak presentation, weak markup, not extensible
- PDF = **P**ortable **D**ocument **F**ormat
  - ü reasons to love: cheap, portable, preserves layout in preview/print
  - û reasons to hate: heavy, not flexible, poor search/navigation
- XML = E**x**tensible **M**arkup **L**anguage
  - ü reasons to love: portable, strong markup, extensible by definition
  - û reasons to hate: expertise is hard to find, design is difficult to do

Document goals of the reader

- Ease of use
  - intuitive interface, zero time to learn
  - “our’s better” is not better, standard is better
  - “in your face” interface is bad, invisible interface is good
- Quality of experience
  - fast and lightweight
  - rich in interactivity and functionality
  - support a range of users: novice **à** expert
- Accessibility of information
  - good retrieval: by content / by subject / by linkage
  - appropriate size: static / whole **ß à** dynamic / parts
  - appropriate format: visually impaired / bandwidth impaired

Document goals of the publisher

- Low cost
  - cheap to produce
  - easy to manage
  - light to distribute
- Protect investment
  - different application, same data
  - protect the document from printing / cutting / pasting
  - data conversion is almost as expensive as data creation
- Manageability & flexibility
  - scaleable from a couple of pages to a million pages
  - automated validation of hyperlinks, versions, security
  - easy reuse and repurpose: different sizes, different formats

Reader goal: ease of use

HTML

- hypertext is a powerful way of working
- the browser already sits on their PC
- users already surf every day

PDF

- harder to read and move around in
- documents are very similar to paper
- respects look and feel of original document

XML

- no commercial XML viewer
- partial support for XML in browsers
- (promise of) richer hyperlinking / interactivity

Reader goal: quality of experience

HTML

- excellent performance, with rich interactivity
- document can be linked, stored, reused
- hard to print and exchange

PDF

- good performance, but big download
- document can only be viewed
- easy to print and exchange

XML

- in theory, anything will be possible:
  multipoint / virtual hypertext, multimedia, …
- in practice, the viewers do not yet exist

Reader goal: accessibility of information

HTML

- not as much structure as XML, just enough
- metadata can be added, but nobody does
- proper design guarantees accessibility

PDF

- indexing depends on originating application
- no structure-based search possible
- one size / format fits all

XML

- structure-based search possible
- has it all: modular, metadata, hyperlinks
- can deliver in a variety of shapes and forms

Publisher goal: low cost

HTML

- so simple everybody thinks he can do it
- high cost of creation and conversion
- managing hyperlinks is a nightmare

PDF

- low cost of creation and conversion
- integrates into existing document production
- allows both high-fidelity viewing and printing

XML

- very high cost of creation and conversion
- document storage and display formats have to be designed and managed

Publisher goal: protect investment

HTML

- terrible long-term storage format
- preserves neither structure nor presentation
- no mechanisms for hyperlinks quality control

PDF

- 100% if created from legacy data
- 0% if created from rich structured data
- only content and presentation are captured

XML

- open non-proprietary standard
- also structure and meaning are captured
- adheres to the “SGML live-long promise”

Publisher goal: manageability & flexibility

HTML

- only when generated dynamically
- impossible to automate quality control
- to achieve flexibility you have to go dynamic

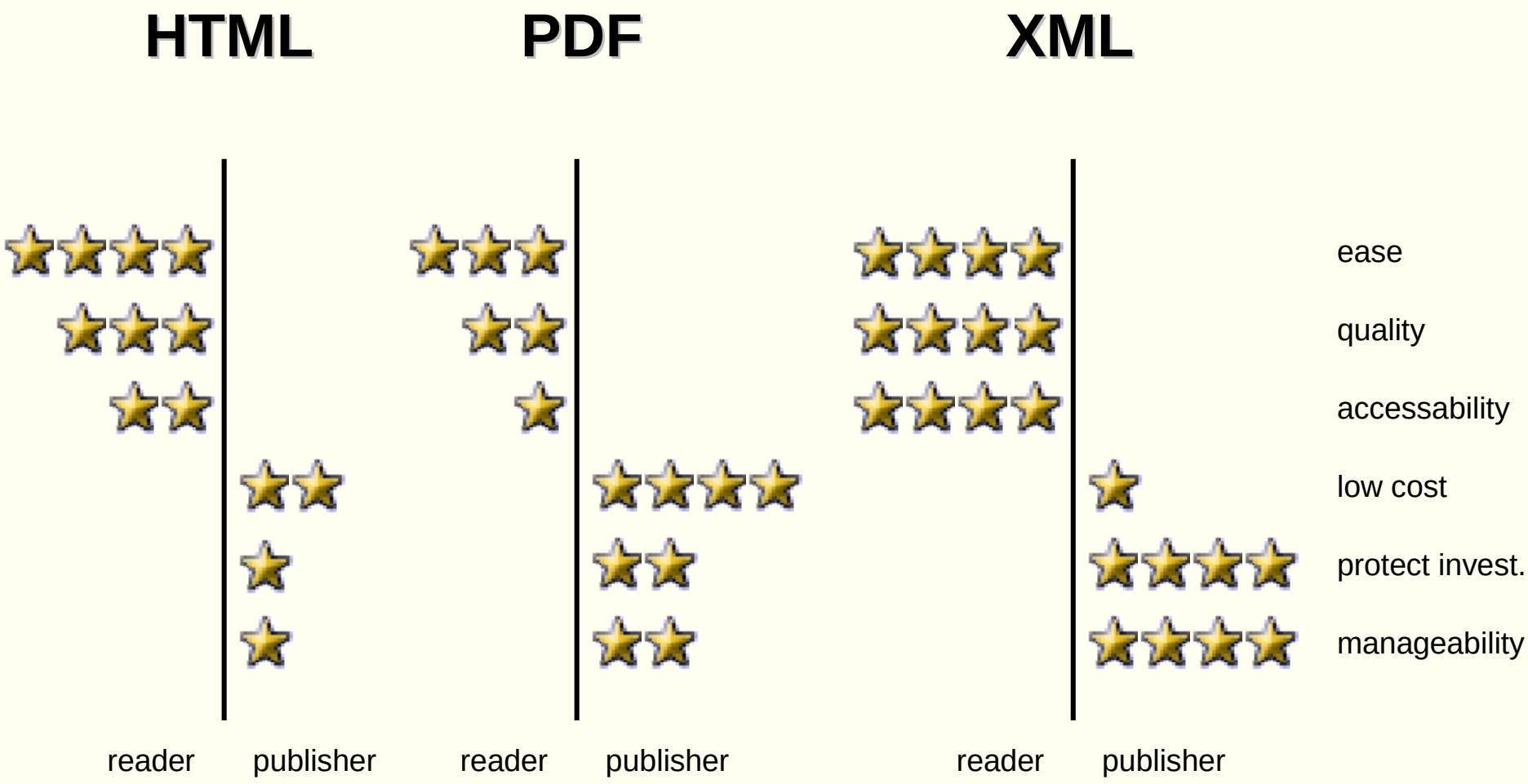
PDF

- good for static legacy documents
- not good for fast-changing new documents
- still managing documents, not information

XML

- can be as rich (metadata, links) as needed
- create for multiple platforms, multiple media, multiple purposes, all from one single source

What's the best document format?



Conclusions

- Not HTML *or* PDF *or* XML but HTML *and* PDF *and* XML
  - HTML and PDF are display formats
  - XML is a storage format
- For “dead” documents, use PDF
- For “living” documents, use XML
  - and generate HTML
  - (and generate PDF too …)
  - and generate whatever you want
- It’s still early days for XML use as a document format

  … but sooner rather than later we’ll all be using it!