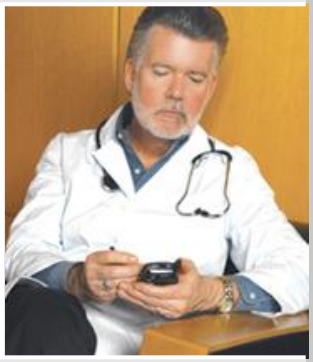
From Word to XML to Mobile Devices

David A. Lee
Senior member of the technical staff





From Word to XML to Mobile Devices

- Introduction
- Background
- Microsoft Word Authoring
- Getting XML Out of Word
- XML / XQUERY Pipeline
- Conclusions & Lessons Learned

Introduction

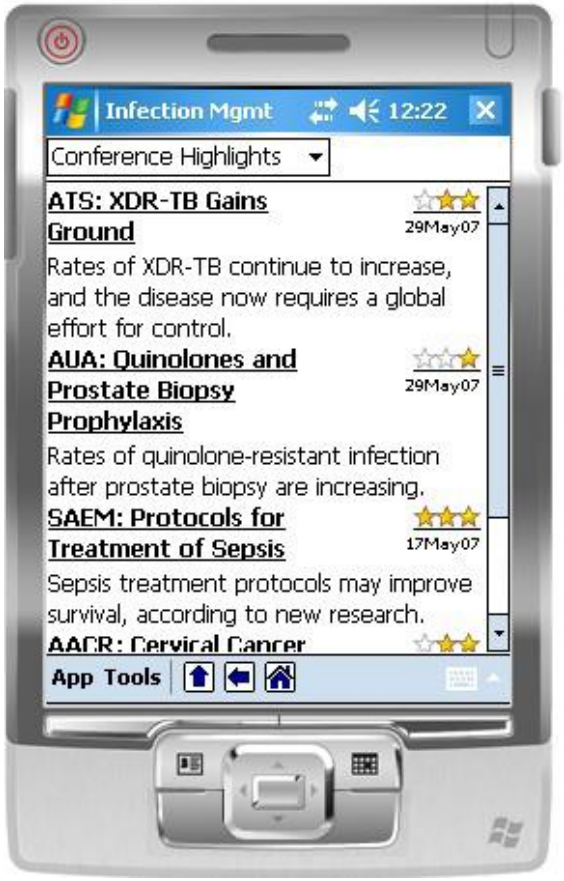
- Who is David Lee ?
  - 20+ years in software development
  - Sun, IBM, Centura, WebGain, Premenos, Epiphany ...
  - Currently Epocrates
- What is Epocrates ?
  - Epocrates is an industry leader in providing clinical references on handheld devices.
  - 500,000 active subscribers
  - Subscription based clinical publishing

Introduction
Common Terminology

- **PDA** - "Personal Digital Assistant".
- **Palm** – A PDA device running the Palm OS® operating system
- **PPC** - A PDA device running Microsoft's Pocket PC or Windows Mobile operating system.
- **Syncing** - The process of synchronizing a server's database with a PDA
- **PDB** - "Palm Database". A very simple variable length record format with a single 16 bit key index.
- **KOL** – “Key Opinion Leader”, a person who is considered an expert in their field.

Core Application
“Mobile Resource Center” (MRC)

A constantly updated
mobile reference
to clinical publications



Background
“Extreme Problem”

‘Impedance mismatch'

vs.

```
<?xml version="1.0" encoding="US-ASCII"?>
<ARTICLE
  mrc_id="eo00"
  target_epoc_publish_date="2007-06-19"
  clinical_significance_rating="2"
  article_id="1022"
  content_category="Conference Highlights"
  date_of_article="2007-05-29"
  version="1"
 last_update_date="2007-06-11">
<FULL_TITLE>ATS: XDR-TB Gains Ground Around the
World</FULL_TITLE>
<SHORT_TITLE>ATS: XDR-TB Gains Ground</SHORT_TITLE>
<EXPERT_COMMENT><P>It should be emphasized that the
total number of
XDR TB cases in the US
from 1993-06 was 49 or about 3 per year.
Of these, 25 (52%) were foreign born.
The big problem with XDR TB is in countries where TB
is endemic -
especially in those where HIV rates are also high.
- John Bartlett - </P>
</EXPERT_COMMENT>
```

Clinical Authors

XML

‘Impedance Mismatch’ Clinical Authors

- Insist on using Microsoft Word
  - Even when forced to use another tool
- Are not trained in markup
- Unstructured data
- Copy & Paste from divergent environments
- Difficult to teach new tools and techniques
- Working from remote locations
  - Difficult to oversee and help with mistakes

‘Impedance Mismatch’ XML Content

- Highly structured
- Editing tools not well established in community
- Intolerant of errors
- Unfamiliar to authors
- “Too Complicated” for authors

Background
Device Application

- Rich text and markup on device
- Severe resource constraints
  - Low memory
  - Low storage
  - Slow CPU
- Device Requirements
  - Efficient Binary XML
  - Fast to parse
  - Low storage and resource use

Background
Human Workflow

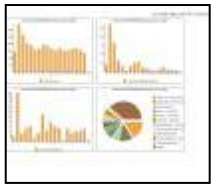
Epocrates Editor selects articles, writes summaries

KOL selects articles to publish from this set and writes expert commentary

Production Assistant(s) converts word doc to XTEXT

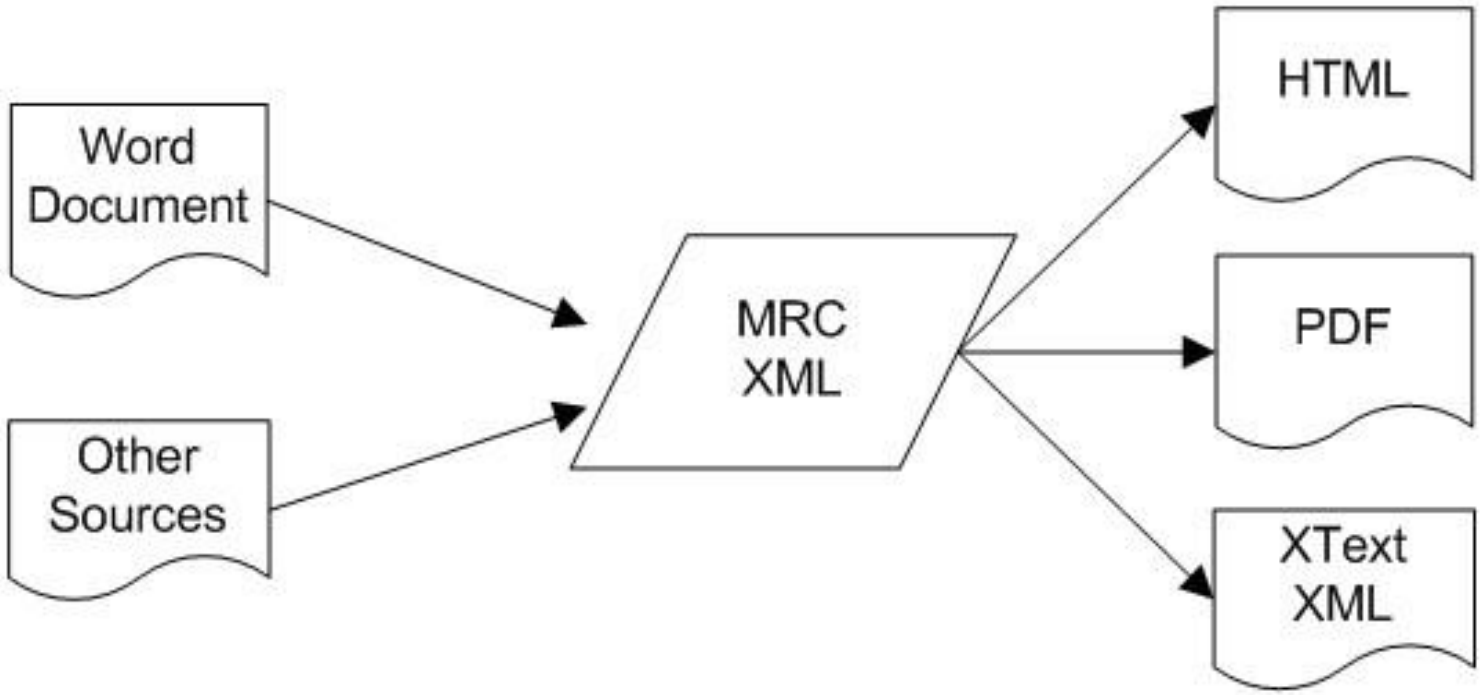
Editor approves edition. Med Info advisor performs sanity check

Publish content on a weekly basis

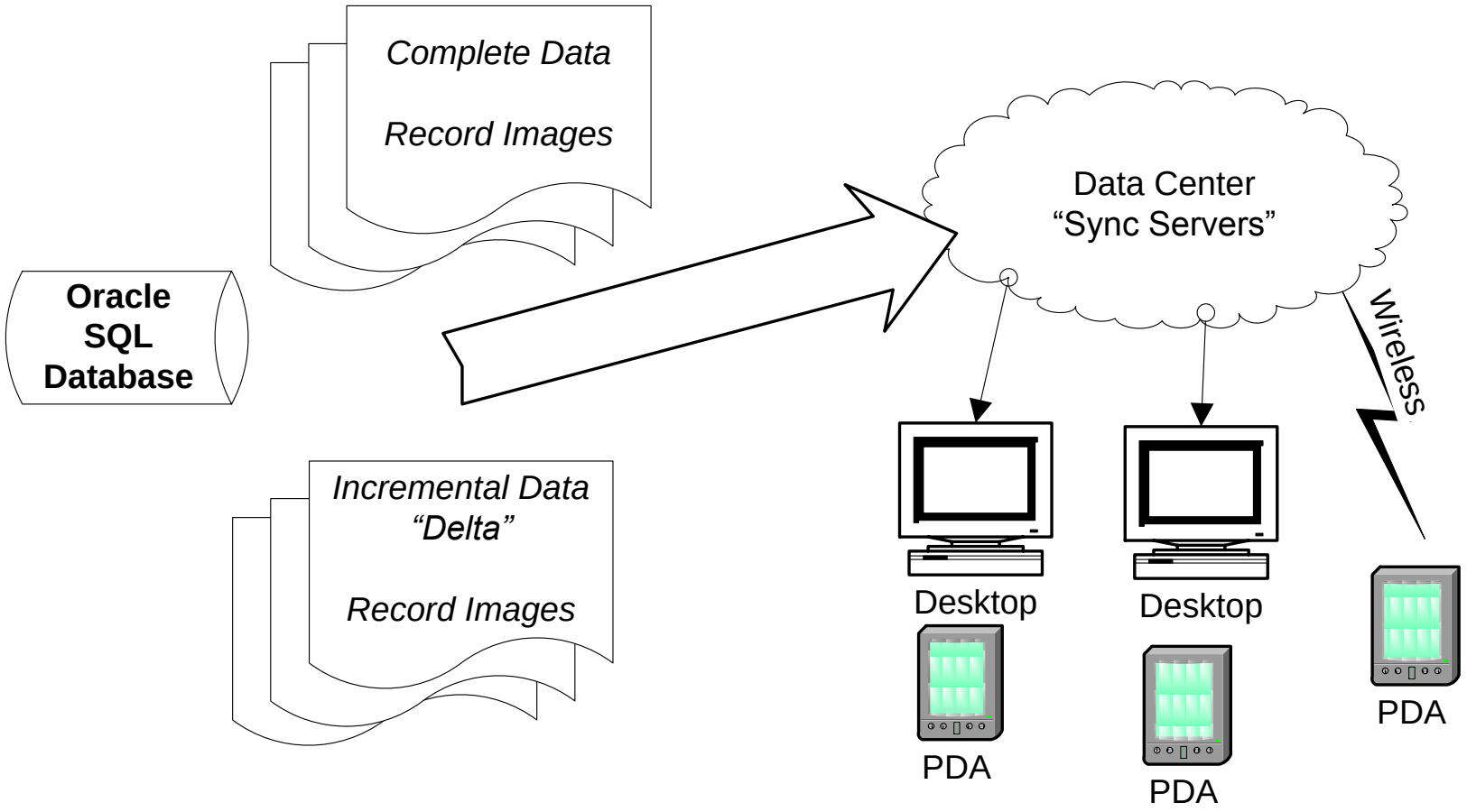


Account Manager distributes monthly internal reports, quarterly external reports, and plans promotions

Background
XML Conversion Workflow



Background
Deployment Workflow



Possible Strategies

- Word Authoring Strategies
  - Design patterns for authoring in Word
- XML Extraction
  - Getting the XML out of Word

Microsoft Word Authoring

- Word Styles
- Tagged Sections
- Form Fields
- Special Symbols and formatting
- Word Macros
- Tables

Getting XML out of Word Conversion Strategies

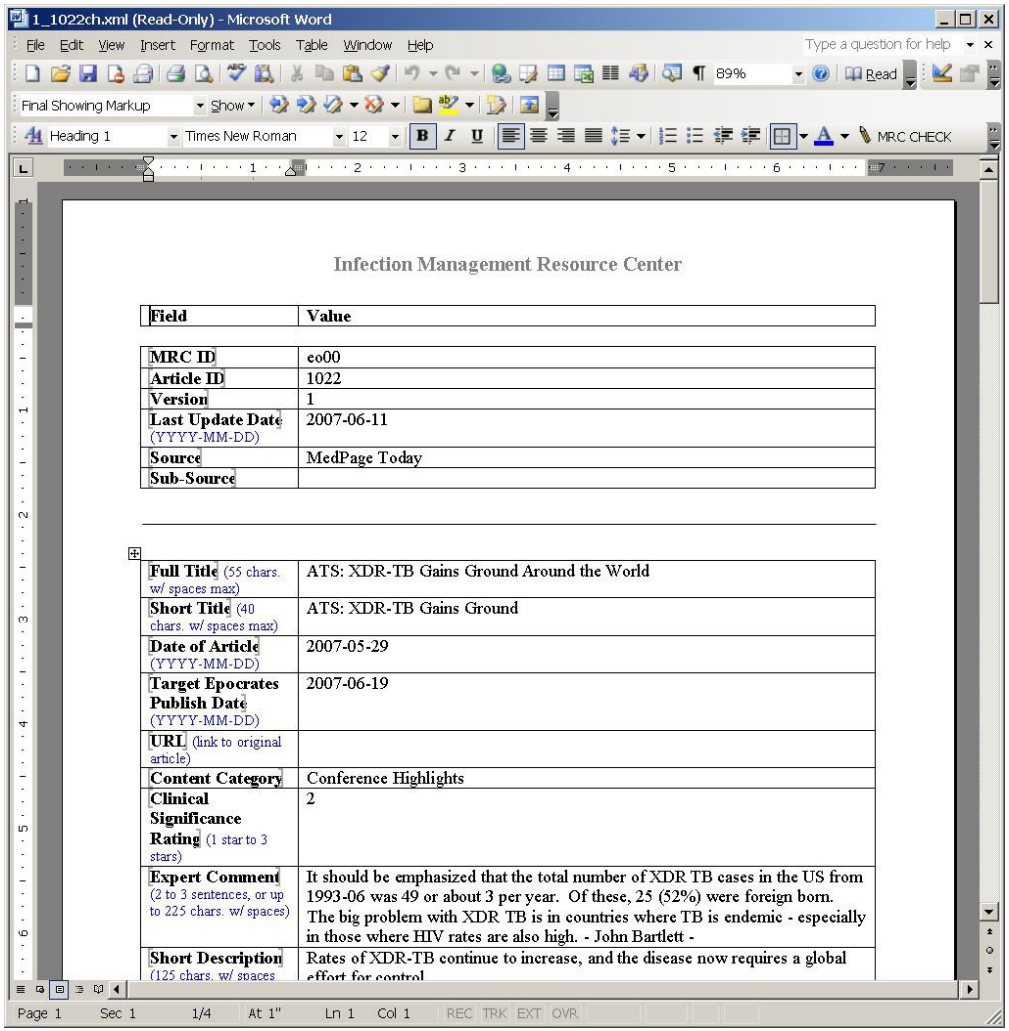
- RTF
- HTML
- Native Word 2003 XML
- Word Macros (VBSCRIPT)

Getting XML out of Word
Selected Strategies

- Trial & error
  - Mostly Error ....
  - Good ideas didn't always work well

Getting XML out of Word Selected Input Strategy

- Word Tables
  - fielded data
- Word Macros
  - First Level validity checks
  - Auto correct
  - (optional) Generate XML



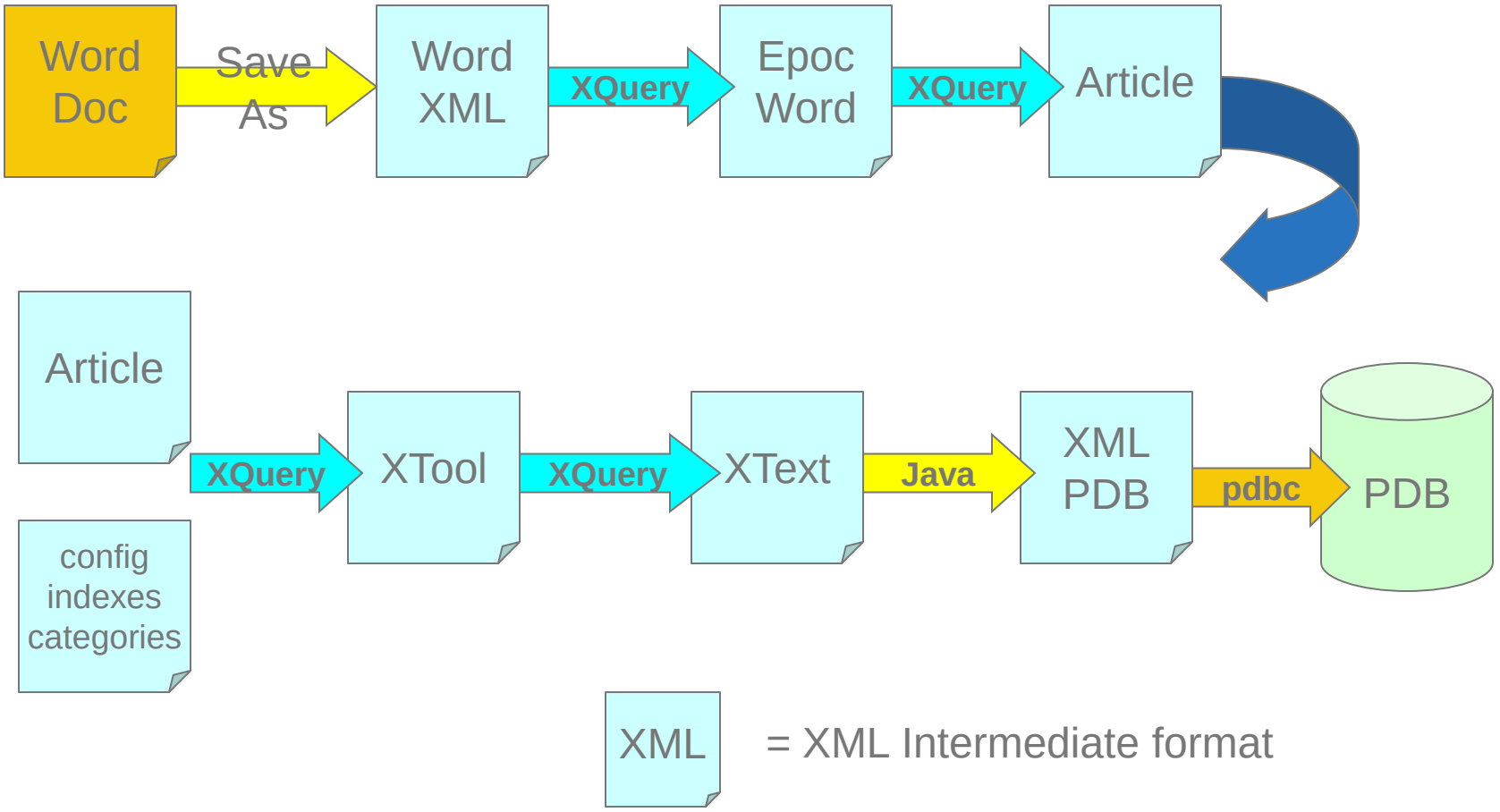
Getting XML out of Word
Selected Extraction Strategies

- Project 1 - Word Basic (Macros)
  - Uses Word object model to extract data
  - Save directly as “Epocrates Word” XML schema
- Project 2 - “Save As XML”
  - Saves as Word XML
  - XQuery converts to “Epocrates Word” XML schema

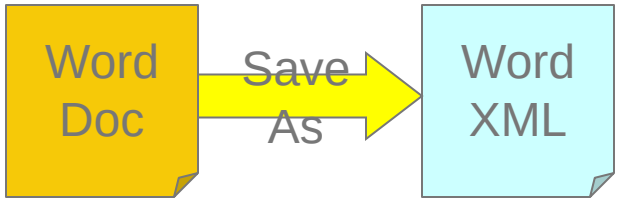
Getting XML out of Word "Epocrates Word" schema

- Intermediate XML format
- Very simple, represents basic structure
  - Tables
  - Paragraphs
  - Line breaks
  - Formatting
  - Bookmark
  - "XML Like" embedded markup Tags
  - Plain text

XML Pipeline Overview



XML Pipeline
Word “Save As”



Word “Save As”

Input

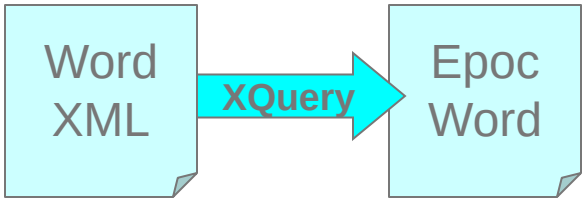
- Microsoft Word 2003 or greater Document

Output

- Microsoft Word XML

http://schemas.microsoft.com/office/word/2003/wordml

XML Pipeline
XQuery



XQuery

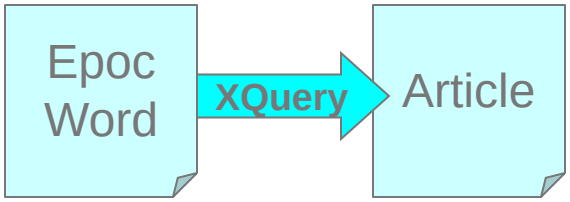
Input

- Microsoft Word XML

Output

- “Epocrates Word” XML

XML Pipeline
XQuery



XQuery

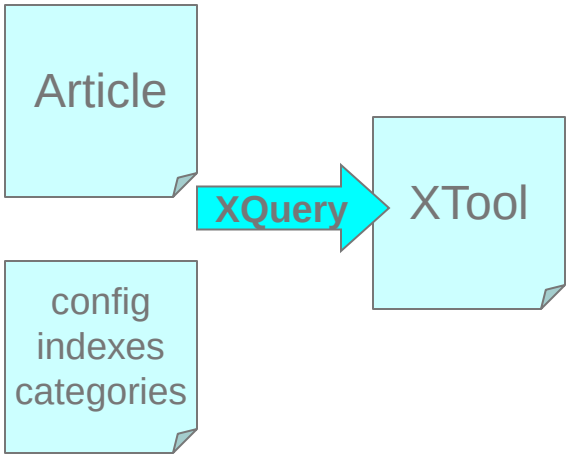
Input

- “Epocrates Word” XML

Output

- Article XML

XML Pipeline
XQuery



XQuery

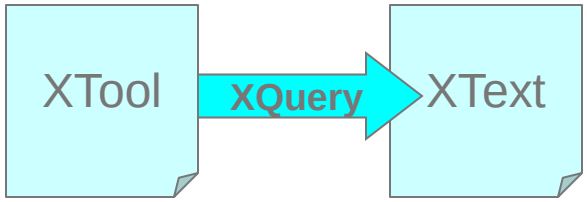
Input

- Article XML
- Config files
- indexes, categories
- other XML

Output

- “XTool” XML

XML Pipeline
XQuery



XQuery

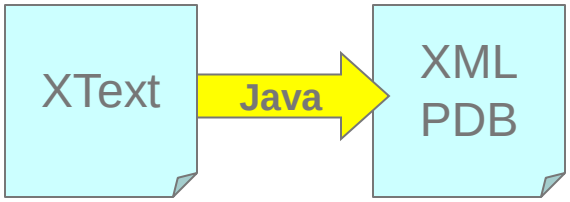
Input

- “XTool” XML

Output

- “XText” (similar to HTML)

XML Pipeline XQuery



Java

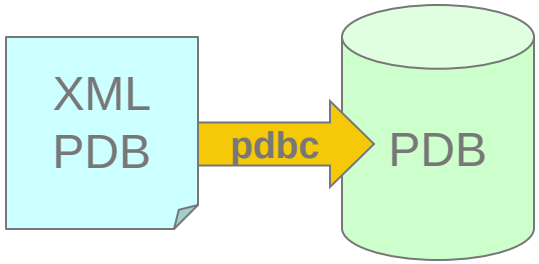
Input

- "XText" XML

Output

- Compiled binary "XText" data in "PDB" record format
  - Encoded as XML

XML Pipeline
Java, pdbc



java , pdbc

Input

- “XML PDB”

device record format

Output

- Device ready “PDB” files

Summary
Lessons Learned

- Design evolved as much by failures as successes

- Failures
  - incorrect assumptions
    - Human component
    - How authors really work
  - Word macros overused
  - Many word features highly error prone
  - “Cant teach old dog new tricks”

Summary
Lessons Learned

- Successes
  - Word Tables for fielded data
  - Word Macros – limited use
    - Extracting XML - OK but tedious
    - Early validation
    - Auto correction
  - Word 2003 “Save As” XML
  - XML Pipeline architecture
    - Multiple intermediate formats
  - XQuery for XML transformation
    - “Proof of concept” – worked well

For further information

- Many more details in paper
  - Full schemas for “Epocrates Word” schema
  - Fragments of XML from various stages
  - Full XQuery source for Word XML to “Epocrates Word” transformation
  - References

Questions ?

Contact Info

David A. Lee

Epocrates, Inc

dlee@epocrates.com