TeX and XML

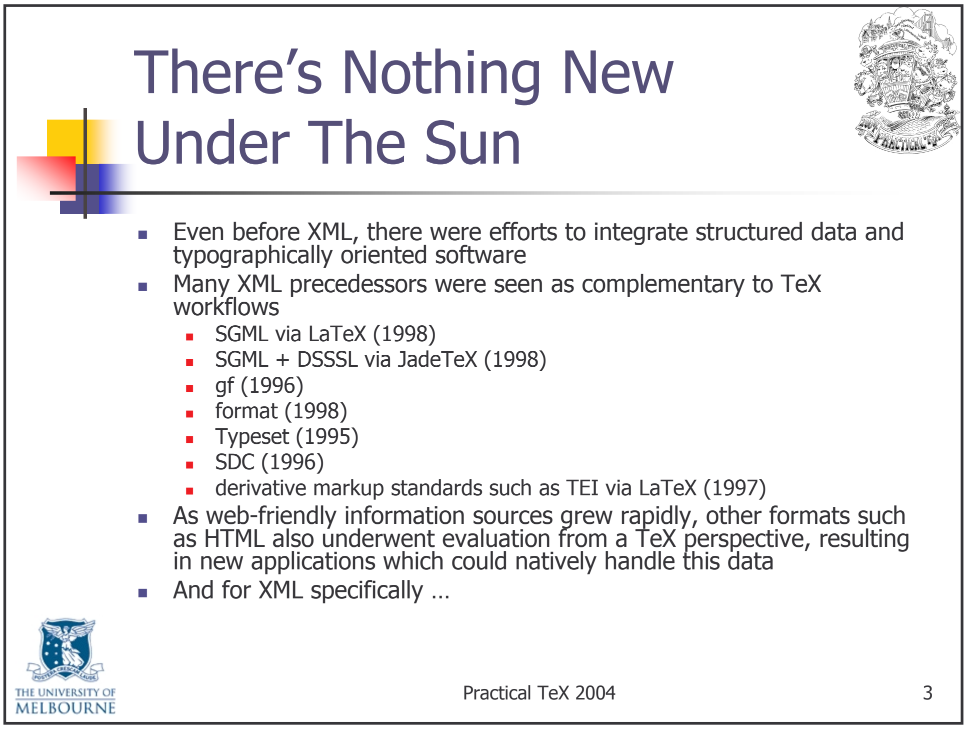
Baden Hughes
Department of Computer Science and Software Engineering
University of Melbourne
badenh@cs.mu.oz.au

THE UNIVERSITY OF
MELBOURNE

Agenda

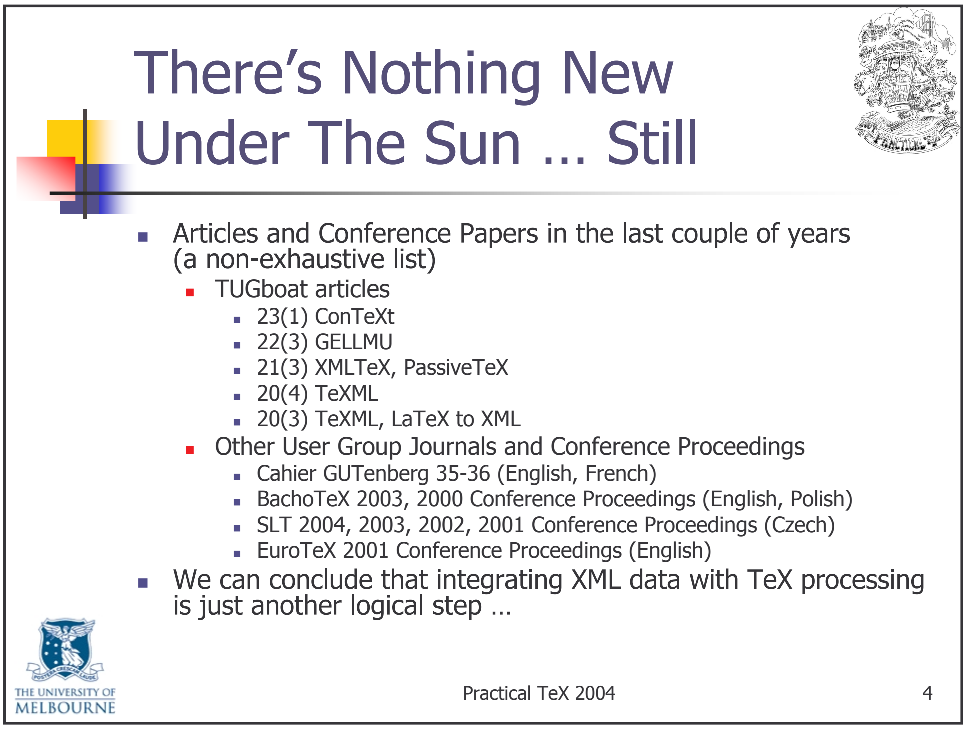
- There’s Nothing New Under the Sun
- Typology of TeX and XML Relations
- The State of the Art and Challenges
- Recent Developments
- Conclusion

THE UNIVERSITY OF
MELBOURNE



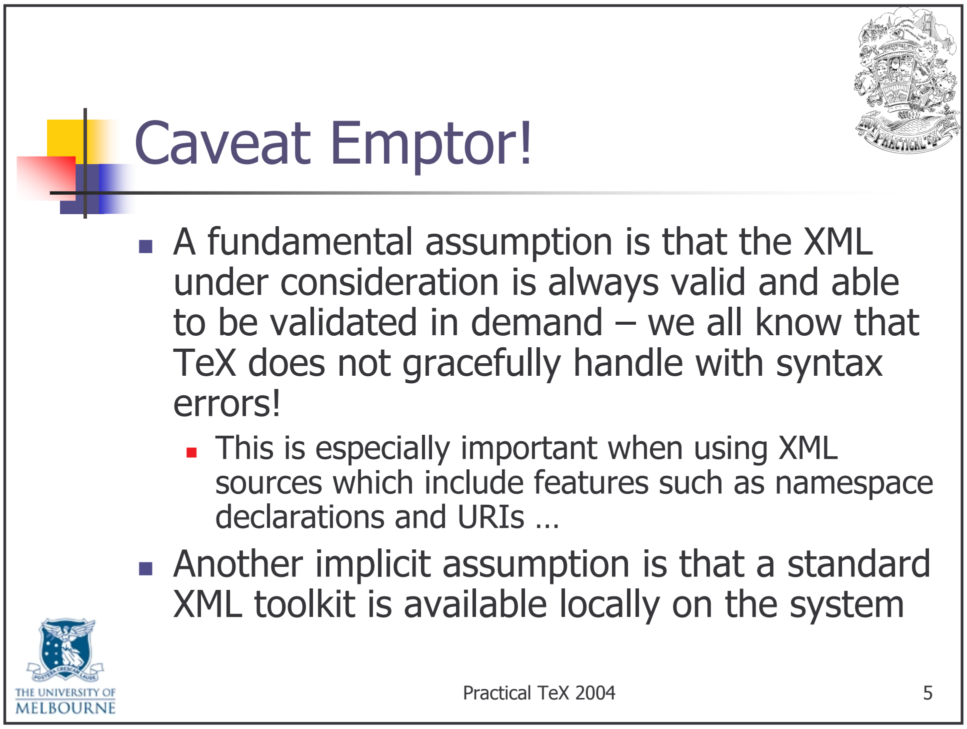
There’s Nothing New Under The Sun

- Even before XML, there were efforts to integrate structured data and typographically oriented software
- Many XML precedessors were seen as complementary to TeX workflows
  - SGML via LaTeX (1998)
  - SGML + DSSSL via JadeTeX (1998)
  - gf (1996)
  - format (1998)
  - Typeset (1995)
  - SDC (1996)
  - derivative markup standards such as TEI via LaTeX (1997)
- As web-friendly information sources grew rapidly, other formats such as HTML also underwent evaluation from a TeX perspective, resulting in new applications which could natively handle this data
- And for XML specifically …



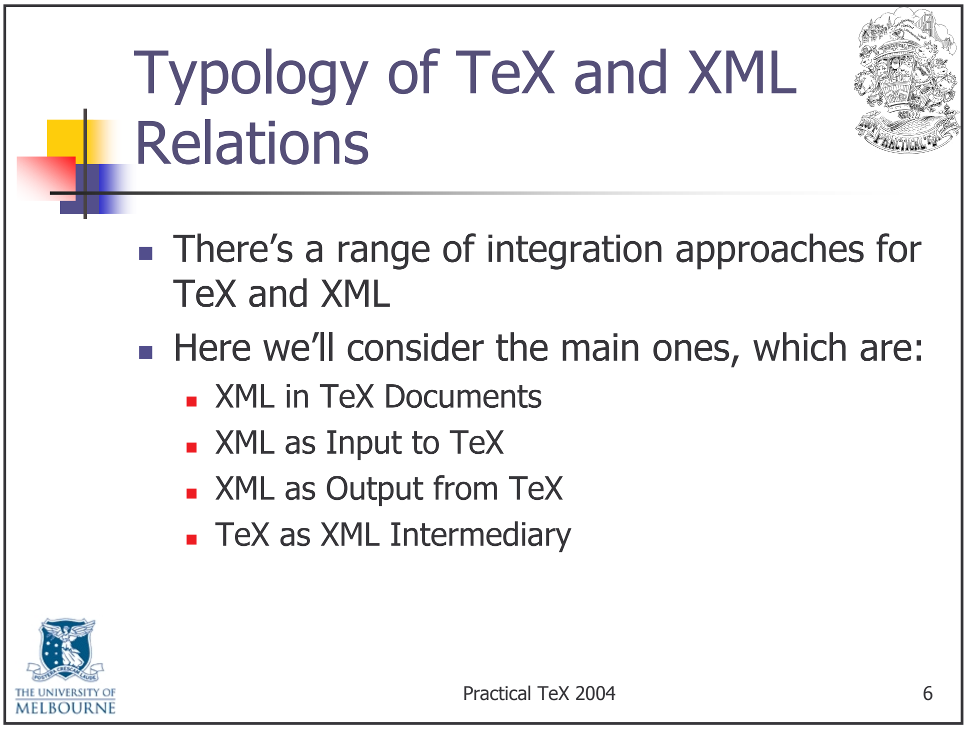
There’s Nothing New Under The Sun … Still

- Articles and Conference Papers in the last couple of years (a non-exhaustive list)
  - TUGboat articles
    - 23(1) ConTeXt
    - 22(3) GELLMU
    - 21(3) XMLTeX, PassiveTeX
    - 20(4) TeXML
    - 20(3) TeXML, LaTeX to XML
  - Other User Group Journals and Conference Proceedings
    - Cahier GUTenberg 35-36 (English, French)
    - BachoTeX 2003, 2000 Conference Proceedings (English, Polish)
    - SLT 2004, 2003, 2002, 2001 Conference Proceedings (Czech)
    - EuroTeX 2001 Conference Proceedings (English)
- We can conclude that integrating XML data with TeX processing is just another logical step …



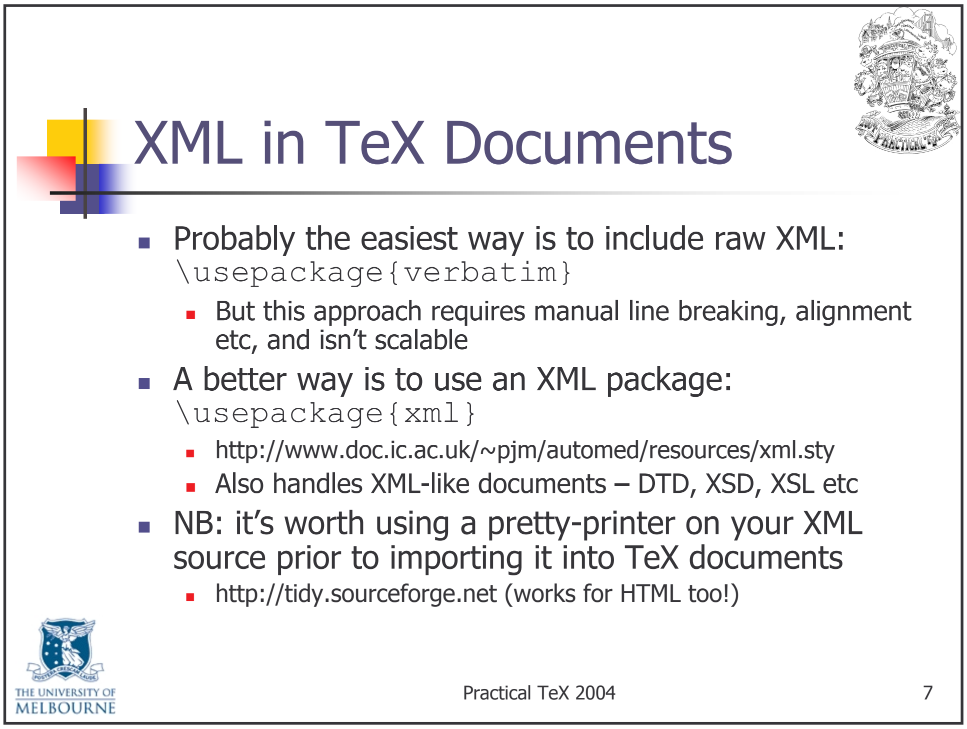
Caveat Emptor!

- A fundamental assumption is that the XML under consideration is always valid and able to be validated in demand – we all know that TeX does not gracefully handle with syntax errors!
  - This is especially important when using XML sources which include features such as namespace declarations and URIs …
- Another implicit assumption is that a standard XML toolkit is available locally on the system



Typology of TeX and XML Relations

- There’s a range of integration approaches for TeX and XML
- Here we’ll consider the main ones, which are:
  - XML in TeX Documents
  - XML as Input to TeX
  - XML as Output from TeX
  - TeX as XML Intermediary

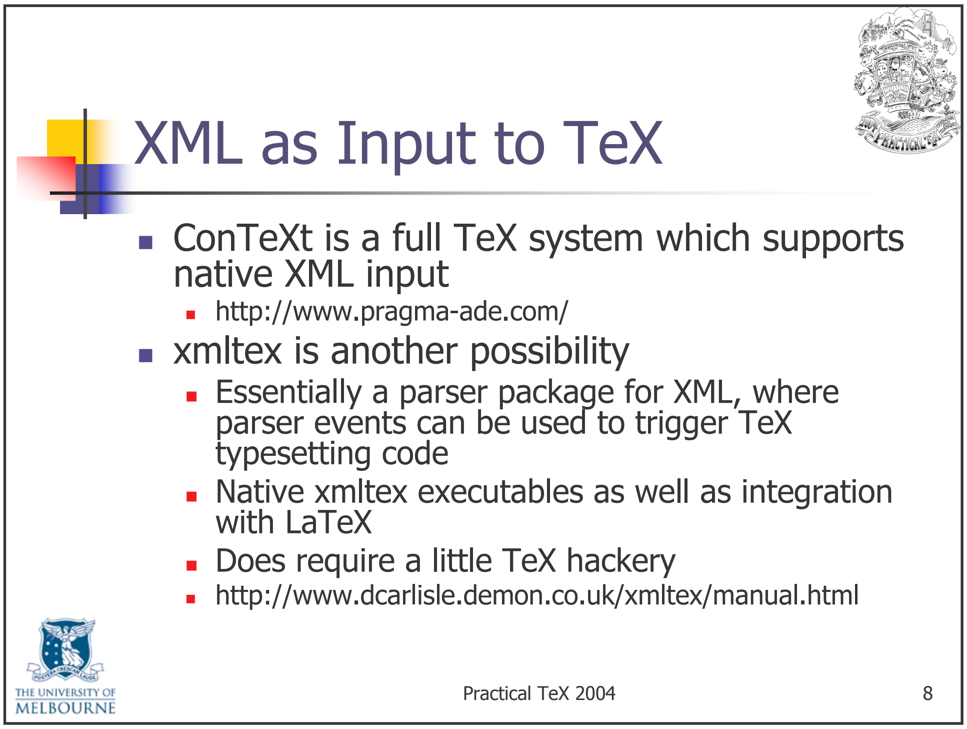


XML in TeX Documents

- Probably the easiest way is to include raw XML: `\usepackage{verbatim}`
  - But this approach requires manual line breaking, alignment etc, and isn't scalable
- A better way is to use an XML package: `\usepackage{xml}`
  - http://www.doc.ic.ac.uk/~pjm/automed/resources/xml.sty
  - Also handles XML-like documents – DTD, XSD, XSL etc
- NB: it's worth using a pretty-printer on your XML source prior to importing it into TeX documents
  - http://tidy.sourceforge.net (works for HTML too!)

THE UNIVERSITY OF MELBOURNE

Practical TeX 2004 7

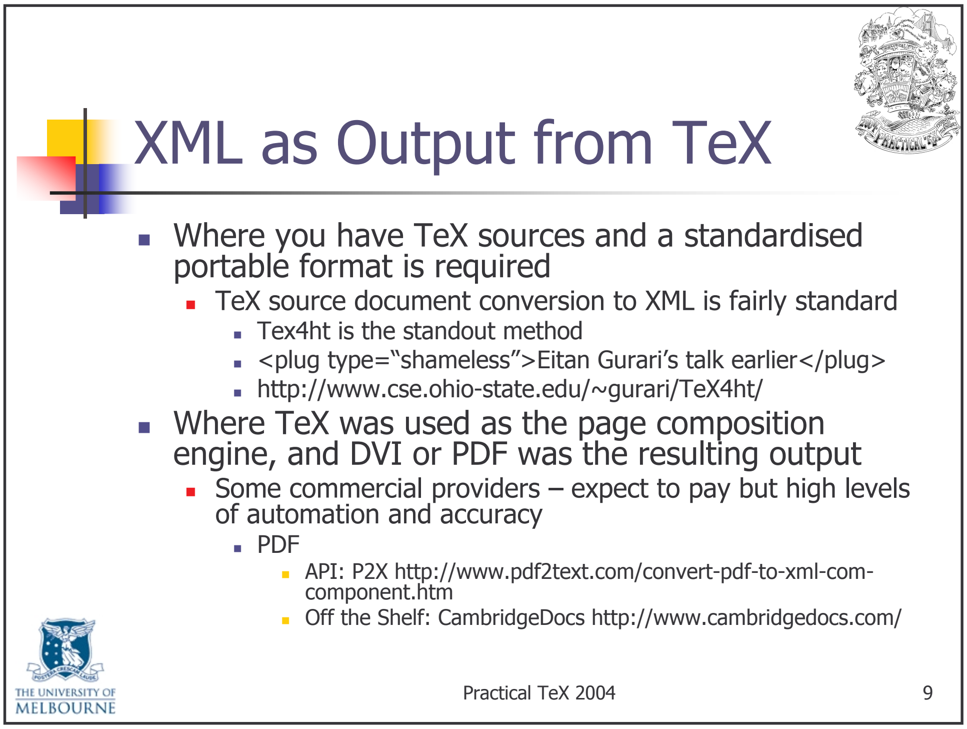


XML as Input to TeX

- ConTeXt is a full TeX system which supports native XML input
  - http://www.pragma-ade.com/
- xmltex is another possibility
  - Essentially a parser package for XML, where parser events can be used to trigger TeX typesetting code
  - Native xmltex executables as well as integration with LaTeX
  - Does require a little TeX hackery
  - http://www.dcarlisle.demon.co.uk/xmltex/manual.html

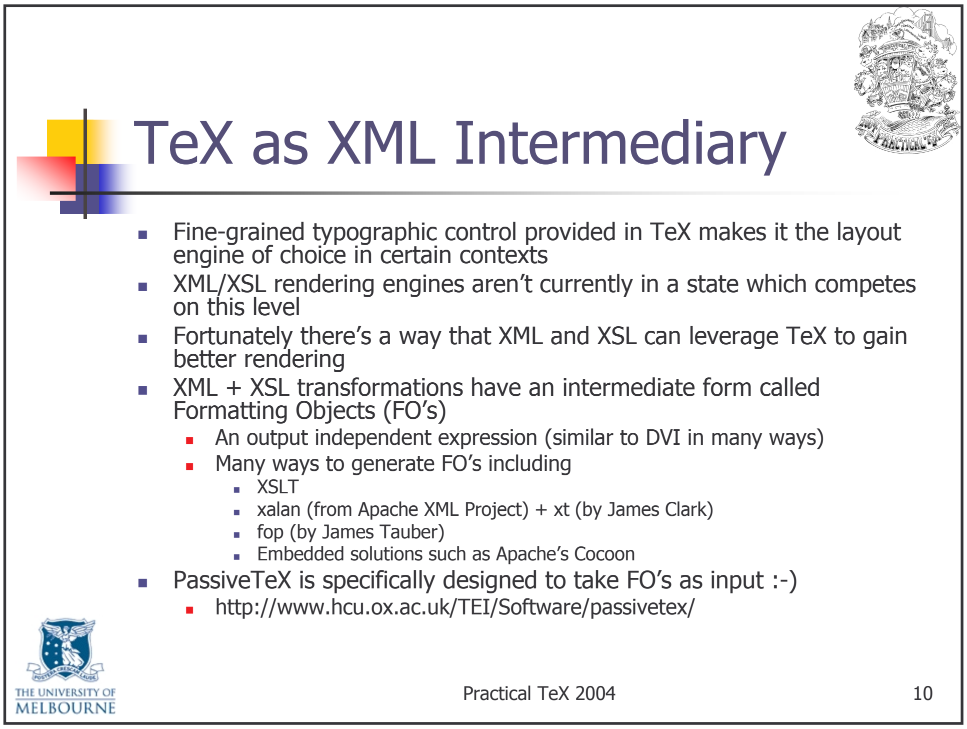
THE UNIVERSITY OF MELBOURNE

Practical TeX 2004 8



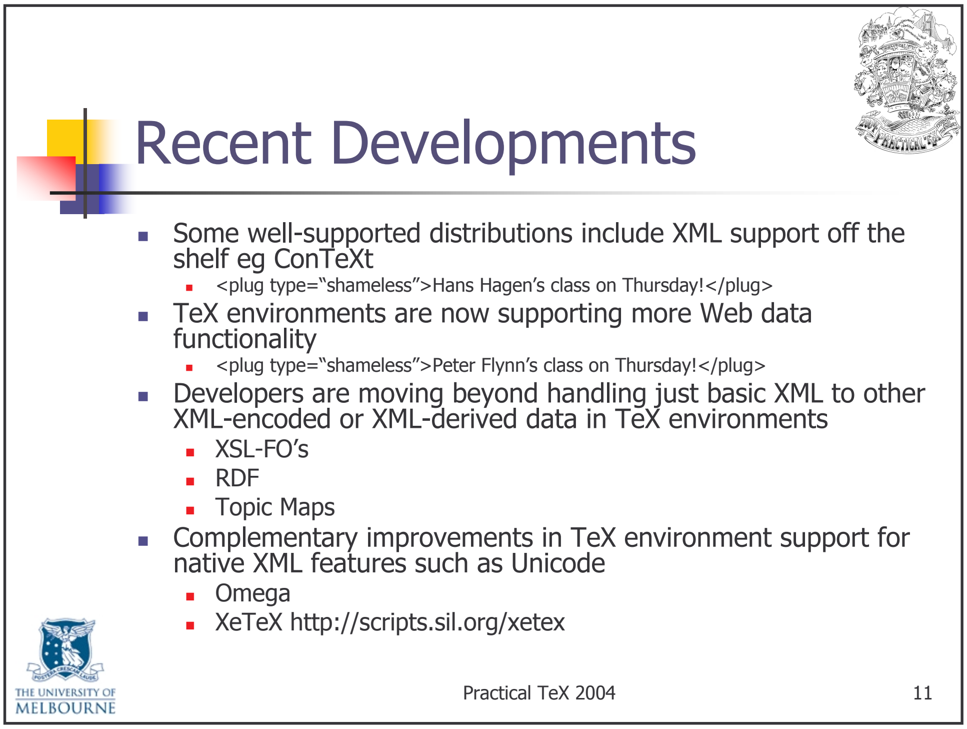
XML as Output from TeX

- Where you have TeX sources and a standardised portable format is required
  - TeX source document conversion to XML is fairly standard
    - Tex4ht is the standout method
    - <plug type="shameless">Eitan Gurari's talk earlier</plug>
    - http://www.cse.ohio-state.edu/~gurari/TeX4ht/
- Where TeX was used as the page composition engine, and DVI or PDF was the resulting output
  - Some commercial providers – expect to pay but high levels of automation and accuracy
    - PDF
      - API: P2X http://www.pdf2text.com/convert-pdf-to-xml-com-component.htm
      - Off the Shelf: CambridgeDocs http://www.cambridgedocs.com/



TeX as XML Intermediary

- Fine-grained typographic control provided in TeX makes it the layout engine of choice in certain contexts
- XML/XSL rendering engines aren't currently in a state which competes on this level
- Fortunately there's a way that XML and XSL can leverage TeX to gain better rendering
- XML + XSL transformations have an intermediate form called Formatting Objects (FO's)
  - An output independent expression (similar to DVI in many ways)
  - Many ways to generate FO's including
    - XSLT
    - xalan (from Apache XML Project) + xt (by James Clark)
    - fop (by James Tauber)
    - Embedded solutions such as Apache's Cocoon
- PassiveTeX is specifically designed to take FO's as input :-)
  - http://www.hcu.ox.ac.uk/TEI/Software/passivetex/

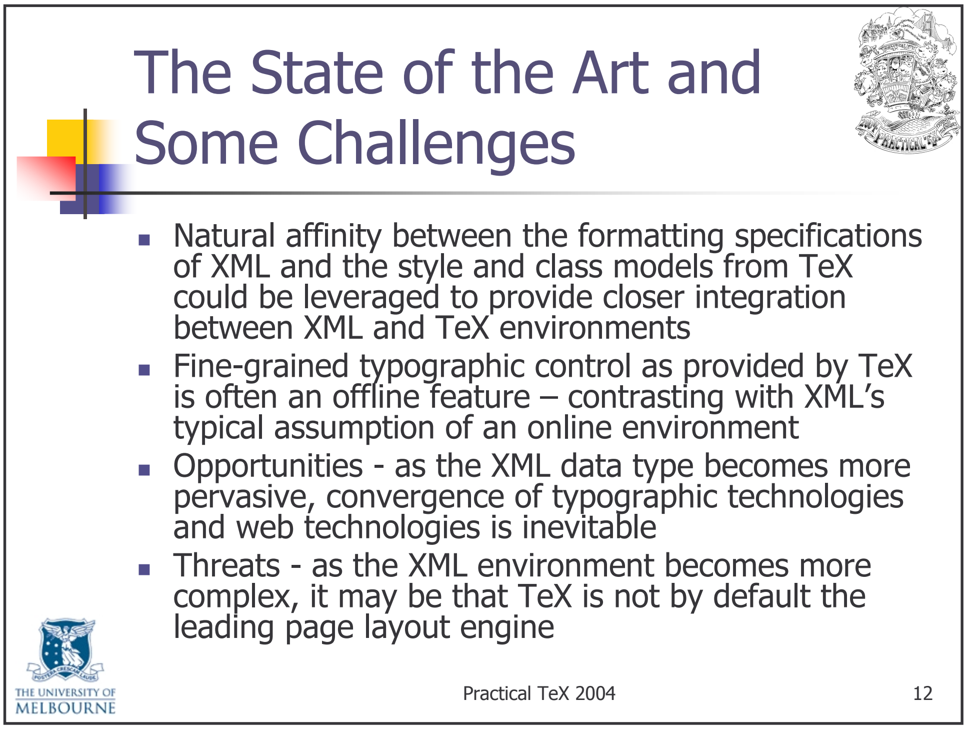


Recent Developments

- Some well-supported distributions include XML support off the shelf eg ConTeXt
  - <plug type="shameless">Hans Hagen's class on Thursday!</plug>
- TeX environments are now supporting more Web data functionality
  - <plug type="shameless">Peter Flynn's class on Thursday!</plug>
- Developers are moving beyond handling just basic XML to other XML-encoded or XML-derived data in TeX environments
  - XSL-FO's
  - RDF
  - Topic Maps
- Complementary improvements in TeX environment support for native XML features such as Unicode
  - Omega
  - XeTeX http://scripts.sil.org/xetex

THE UNIVERSITY OF MELBOURNE

Practical TeX 2004 11

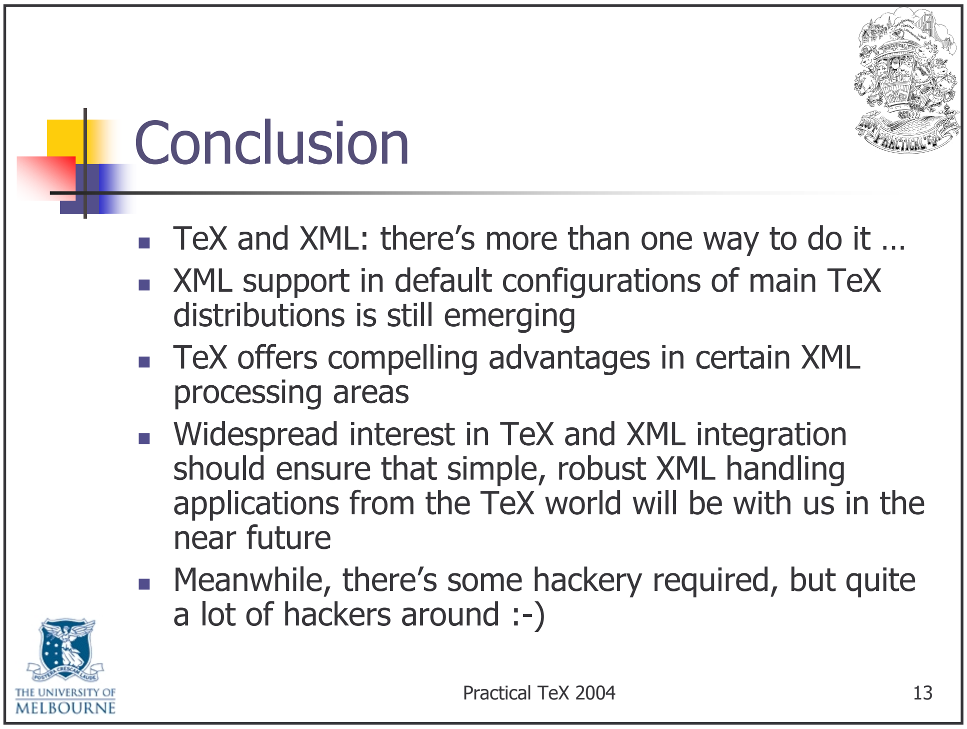


The State of the Art and Some Challenges

- Natural affinity between the formatting specifications of XML and the style and class models from TeX could be leveraged to provide closer integration between XML and TeX environments
- Fine-grained typographic control as provided by TeX is often an offline feature – contrasting with XML's typical assumption of an online environment
- Opportunities - as the XML data type becomes more pervasive, convergence of typographic technologies and web technologies is inevitable
- Threats - as the XML environment becomes more complex, it may be that TeX is not by default the leading page layout engine

THE UNIVERSITY OF MELBOURNE

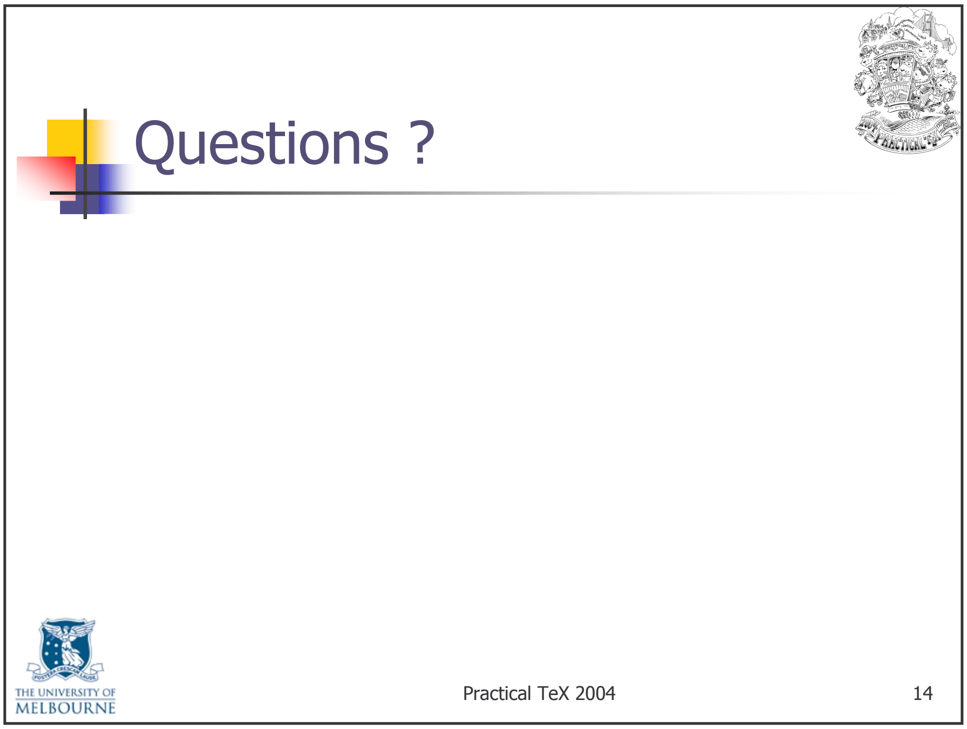
Practical TeX 2004 12



Conclusion

- TeX and XML: there's more than one way to do it …
- XML support in default configurations of main TeX distributions is still emerging
- TeX offers compelling advantages in certain XML processing areas
- Widespread interest in TeX and XML integration should ensure that simple, robust XML handling applications from the TeX world will be with us in the near future
- Meanwhile, there's some hackery required, but quite a lot of hackers around :-)

THE UNIVERSITY OF MELBOURNE



Questions ?

THE UNIVERSITY OF MELBOURNE