Technical challenges, opportunities, goals, strategies

FWHDML 2012, Washington, USA

Alan P. Sexton

School of Computer Science
University of Birmingham, UK

June 2012

Digital Mathematical Libraries

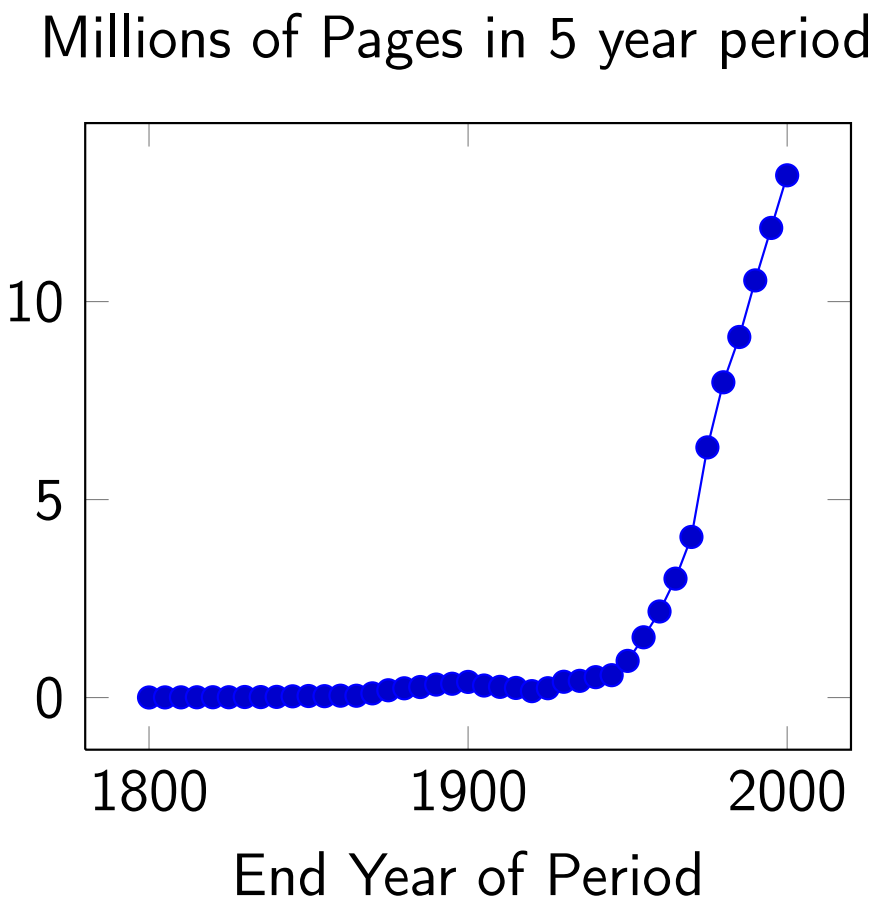
- Project Gutenburg, 1971
  - Digital, but not mathematical, library
- ArXiv, 1991, preprints and technical reports rather than peer-reviewed archival publication
- JSTOR, 1993, now contains significant number of Math items
- Project Euclid, 1999: Dedicated DML
- NUMDAM 1999: Dedicated DML

DML Sizes

- Ulf Rehman's DML web site refers to about 5 Million pages
- Project Euclid reports approx. 1.5 Million pages
- NUMDAM reports approx 760,000 pages
- EuDML has about 2.6 Million pages
- Many small DMLs with much less that 100,000 pages
- But much duplication (e.g. EuDML aggregates items available elsewhere)
- **Very generous estimate: Maybe 8 Million pages in DMLs after 12 years**

Size of the Mathematical Literature

- Keith Dennis estimated the size of the published mathematical literature as approx 50 Million pages in June 2000
- In 2002, He updated missing data and made the data available



Growth of Mathematical Literature

- $\approx 3.7M$ pages by 1900
- $\approx 7.7M$ pages by 1950
- $\approx 77M$ pages by 2000
- $\approx .9M$ pages 1945–1950
- Since 1965, approx linear increase (avg 1.4M) per 5 year period
- $\approx 13.2M$ pages 1995–2000
- $\Rightarrow 108M$ pages to 2010?

- Recently, Patrick Ion (MR), Michael Jost (ZBL) and I have started to dig deeper.
- Probably around 9 Million pages published in 1995–2000.

More published in that 5 year period than we have been able to get into DMLs in 12 years.

Technical Challenges, Opportunities, Goals, Strategies

- DMLs as PDF collections is not good enough
  - Searchability (formula search, diagram search, semantic similarity browsing...),
  - Accessibility (visually impaired, dyslexics,...),
  - Interactivity (cut and paste formulae, plot functions, explore appendix material and formal proofs such as Flyspeck, 4-colour,...)
  - This in turn needs knowledge rich representations:

    We are not done when we have obtained the PDF.
- Fresh items
  - Need to make math items "DML-ready" at production time and automatically ingested at publication (Concept of Copyright Library?)

    Without this we drown
- Vintage items
  - Born Digital (e.g. LaTeX sources)
  - Retro Born Digital (e.g. PDF from LaTeX)
  - Retro Digitised (e.g. Scanned images)

The Problem with MathML

- MathML (and variants) is a suitable representation that gives us a good basis for providing much of the enhanced functionality we want.
- But, MathML is premised on an assumed workflow directionality: If you have Math, you can use MathML to represent it: Mathematics $\longrightarrow$ Presentation
  - Fine for writing math papers
  - Not so good when trying to capture maths from papers
  - Automatic Presentation to Content MathML translation is unworkable
  - Even OCR to Presentation MathML is fraught as one has to commit to interpretation choices before there is sufficient information to guide those choices
    - $a(b+c)$
    - $\int f \ d\mu$
  - Possible solution: Partially Instantiated MathML?, Constraint MathML?