#ArchivePress:
A Really Simple Solution to Archiving Blog Content

Maureen Pennock
Web Archive Preservation Project Manager

iPRES (San Francisco, Sept 2009)
Blogs: New Medium: Old Genre

Samuel Pepys Diary: Manuscript Volumes
Blogs: New Medium: Old Genre

Samuel Pepys Diary:

Sunday 23 September 1666
(Lord's day). Up, and after being trimmed, all the morning at the office with my people about me till about one o'clock, and then home, and my people with me, and Mr. Wayth and I eat a bit of victuals in my old closet, now my little dining-room, which makes a pretty room, and my house being so clean makes me mightily pleased, but only I do lack Mercer or somebody in the house to sing with. Soon as eat a bit Mr. Wayth and I by water to White Hall, and there at Sir G. Carteret's lodgings Sir W. Coventry met, and we did debate the whole business of our accounts to the Parliament; where it appears to us that the charge of the war from September 1st, 1664, to this Michaelmas, will have been but 3,200,000l., and we have paid in that time somewhat about 2,200,000l.; so that we owe above 900,000l.: but our method of accounting, though it cannot, I believe, be far wide from the mark, yet will not abide a strict examination if the Parliament should be troublesome. Here happened a pretty question...
But the web is already being archived… isn’t it?
Blogs: New Medium: New Possibilities

- Can continue to collect ‘per object’
  OR

- Can develop new corpora of aggregated content for different purposes, eg:
  - Institutional recordkeeping
  - Cultural heritage themed collections
  - Academic research
ArchivePress: aims & objectives

- Research ‘significant properties’ of blogs
- Develop and release open source plug-in
- Create demonstrator instances from participating pilot institutions:
  - UK Digital Curation Centre
  - UKOLN
  - University of Lincoln
  - British Library
Part I: User survey & SigProps
ArchivePress survey

- Academic Attitudes to Blogging & Blog Archiving
  - Short survey, 10 questions
  - Produced data on existing practices & value, including:
    - If blogs are used for academic research
    - How blog content is typically digested
    - Most valuable functions served by blogs
    - Most valuable elements/parts of blogs
    - Measures currently used to archive content
    - Which elements/parts of blogs that blog archiving applications should capture
AP Survey: What elements should blog archiving app capture?

Posts!
AP Survey:
What elements should blog archiving app capture?

- Posts!
- Comments
AP Survey: What elements should blog archiving app capture?

- Posts!
- Comments
- Blog name & URLs
AP Survey: What elements should blog archiving app capture?

- Posts!
- Comments
- Blog name & URLs
- Tag & category names
AP Survey:
What elements should blog archiving app capture?

- Posts!
- Comments
- Blog name & URLs
- Tag & category names
- Embedded objects
Significant properties (first round)

- **Content**
  - Posts, Comments, Embedded objects,

- **Context**
  - Blog title & URL; Primary authors; Author profiles; Content dates, Tags, Categories

- **Structure**
  - Post IDs, Comment IDs, Component relationships

- **Rendering**
  - Text formatting

- **Behaviour**
  - Hyperlinks
Part II: Technical Details
“Use the Feeds...”
AP1 Demonstrator: The DCC collection
Next steps

- AP2: the UKOLN collection (underway)
  - Harvesting comments
  - Resolving rendering & configuration issues from AP1
- Clarifying differences of Atom & RSS feeds
- ID core set of metadata req’ments
- Enabling administrators to add new categories
- Involving participants in validation exercises
- Producing installation & config guides
Conclusions (so far)

- Our premise and approach is valid!
- Greater variations in feed content than expected
- Configuration is tricky: must make it easier
- Development of new tools to respond to changes in the web are essential
Your input is welcome –

Website: http://archivepress.ulcc.ac.uk/

Twitter: @archivepress

Thank You!