RLG Programs

Big issues and barriers in web archiving

Background to issues

- Dealing with 21st century collections and users
- Trying to apply 20th (and 19th!) century training and techniques, metal model for how paper based collections are used.

- Jump in and do it – save what's on the web now, sort out issues later.
- Results in a lot of unknowns.

Business cases for web archiving

- Subject-based (area studies, etc.)
- Mandated collection/retention
  - State and agency publications
  - Government websites / publications
  - University archives
- Institutional repository/intellectual output of an institution
- Event based (9/11, Katrina, Virginia Tech shootings)
- Preserving virtual environments (Second Life)

Barriers to shared collection development

- Creating a community of practice
- Collection development policies and documentation for shared business models
  - North Carolina presentation – a lot of interest

- Not knowing who else is collecting in your area in a formal way
- Possible solution – global registry of harvested URLs?
  - Who is harvesting, frequency, depth, harvest list
  - Could be generated by Heritrix and other harvesters

Dealing with robots.txt, javascript issues

- How to open doors?
- Approach information officers in agencies rather than web masters (move up the food chain!)
- Work with webmasters to let them know about usability issues (most possible for those in agencies, on campus)
- Larger awareness of web archiving could help here

Descriptive metadata

- Best practices? First practices!
- Importance of descriptive metadata when text can be searched?
- Ability to incorporate metadata into other search systems (metasearch tools, OPAC) so that web harvested materials are not a separate silo?
- Important that metadata plays well with others, especially when unified with other like resources in something like WorldCat.
- Knowing the audience

Frequency of harvest?

- How to make sure important documents don't get lost in the shuffle, while staying in budget and on target?

End user issues

- Not enough known about current and potential users of web archived material
- What do they want? (goes to collection development issues)
  - Mining searches to find out what current users are doing?
- How will they find it? (goes to descriptive metadata issues)
- Synthesis of existing research on user studies