COLLECTION DEVELOPMENT FOR THEMATIC WEB ARCHIVES

Pamela Graham
Columbia University

RESAW Conference
London, June 14-16, 2017
THE NATURE OF THE PROBLEM

“. . .we have not transcended the foundational preoccupation with the efficacy of our technical methods for Web capture, preservation and replay”

(Taylor 2017, p. 4)
More on the problem

- Lack of communities of practice, shared knowledge and best practices

- Low profile of Collection Development

- Challenges of Selector engagement
  - “start-up” “first box”

- Underspecified in programmatic development
Sampling the literature
“Selection morphs from a human selection of discrete and stable units . . . to a more versatile and dynamic selection of paths to be followed with certain depth and time.” (Masanès, 2006, p. 74)
Take-aways

- Phases of selection
  - Exploration, Discovery, Sense-Making

- Interconnectedness with subject and technical work

- Participation in the full life cycle

- Research, Assessment, Analysis
Thematic Web Archiving Selection Life Cycle

- Establish Theme
- Access & Tool Development
- Test Crawl / Collect / QA
- Metadata

- Research Information Space
- Explore / Sample
- Assess / Analyze / Support Research
- Categorize / Describe

The diagram illustrates the life cycle of thematic web archiving, including stages for establishing a theme, researching information space, exploring and sampling, assessing and analyzing support research, categorizing and describing, and metadata development.
• Establish Theme

- Nature of the web, as a primary source, platform and/or repository of knowledge in a subject area or topic
- Connection to existing collection programs and strengths?
- Applicability of existing collection development skills/knowledge
- Researcher needs and practices

Output: High-level statement documenting rationale and need for creating a collection.
How do scholars scope and define the “web” in their studies of this field or topic?

“Exogenous sources” to guide the selection of sites

Exclusions, what NOT to collect

Scholarly practices and metadata schema and choices?

Degrees of “at-risk-ness”?

Scoping and frequency of capture?

Archivability of sample websites?

Legal and ethical issues in the collecting process; authority and trust of the web archive

Outputs: Seed lists, exogenous resources for selection and benchmarking; decisions about what NOT to collect; scoping, frequency and permissions policies.
Are there categories or typologies of websites emerging from the exploratory work?

How do these categories inform metadata decisions, access and ongoing collecting?

Visualize, represent the archive in new ways; inform ongoing collecting and documentation

Support researcher use; identify collecting and access needs

Monitor research generated through the use of web archives to inform collecting and access

Outputs: Ongoing benchmarking and comprehensiveness metrics; Additional seeds; clearer documentation; use cases and input on access and discovery.
Thematic Web Archiving Selection Life Cycle

- Establish Theme
- Access & Tools
- Research Information Space
- Explore / Sample
- Assess / Analyze / Support Research
- Categorize / Describe
- Test Crawl / Collect / QA
- Metadata

Circular process:
- Research Information Space
- Explore / Sample
- Assess / Analyze / Support Research
- Categorize / Describe
- Return to Research Information Space