Capturing Websites - Insight From Experience, Looking Toward Tomorrow

Mark Conrad
Archives Specialist, National Archives and Records Administration

Education Code: TR1-1227

53rd Annual Conference & Expo

Las Vegas Convention Center

Learning Objectives

Upon completion of this session, participants will be able to:

- Identify policy decisions to address when managing their websites
- Evaluate benefits and drawbacks of various methods and tools used in capturing websites
- Describe the challenges of preserving the components of a website over the long term

W,W,W,W,W, and H

- Who, What, When, Where, Why, and How
- Not in that order!

Why are You Capturing Websites?

- Answers to this question will help form the other questions (and their answers!)
  - Define objectives
  - Define customers (and their expectations)
  - Define scope
  - Determine available resources
  - Determine what's good enough

Why are You Capturing Websites?

- Website as recordkeeping system?
  - Sufficient functionality?
    - DoD 5015.2
    - MoReq2

Why are You Capturing Websites?

- Website as record(s)?
  - Best copy
  - Retention
    - Long term
    - Medium term
    - Short term
  - Good idea???

Why are You Capturing Websites?

- For the information they contain
  - Internal websites
  - External websites
  - Best source of the information
    - Timeliness
    - Completeness

Why are You Capturing Websites?

- Websites as ESI
  - See your General Counsel
- Other reasons
  - Objectives
  - Scope

What Will You Capture?

- Boundaries
  - Where does the website “end”
  - External content
  - Dynamic content

What Will You Capture?

- Identifying all of the components
  - Files, viewers, plug-ins, scripts, services, databases, data streams, etc, etc, etc
- Identifying all of the functionality
  - Dynamic content, streaming audio/video, presentation, sorting, etc

When Will You Capture It?

- Webpages change
  - Some are dynamic
  - Some are query-driven
  - Even “static” pages change over time

When Will You Capture It?

- How much of the change do you need to capture
  - Timeliness
  - Version control
  - Completeness

How Will You Capture It?

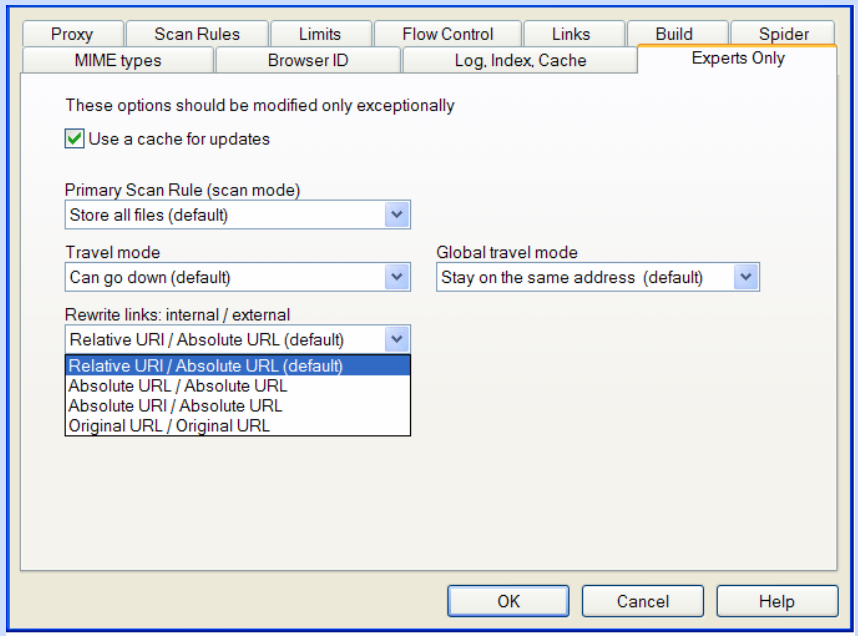
- Selecting tools to effect the capture
  - Can they capture all of the identified components
  - Is any of the functionality of the website lost
  - Can you tell what changes the tools are making

How Will You Capture It?

- Some potential obstacles on the path
  - Firewalls
  - Robots.txt
  - Query limits
  - ECM

How Will You Capture It?

- Selecting the settings for the tools
  - A web crawl…is not a web crawl



Where Will You Keep It?

- “Native habitat” vs “the zoo”
  - Duplicate the native habitat
    - Full copy of the web infrastructure
    - That changes over time, too
    - Configuration control

Where Will You Keep It?

- The “Zoo”
  - Store the files
    - Off-line
    - In a file system
    - In a repository

Your Websites in the Zoo

- Moving from the native habitat to the zoo
  - How will components interact in the zoo
  - Will they find everything they need
  - What else was introduced
    - Tool-specific files
    - Stubs/pointers (especially for streaming data)

Your Websites in the Zoo

- Do you rewrite the links
  - Make a “self-contained” website
  - Maintain the “original” links
  - Rewrite the links to work in a repository system

Your Websites in the Zoo

- How do you validate the capture
  - Thousands of files
    - Too little
    - Too much
  - Lots of functionality to check
  - Multiple browsers
- How do you do it at scale

Keeping Your Websites Alive

- Thousands of files
- Dozens of software-dependent formats
- Preservation requirements will often vary from one website to the next
- How long do I have to keep these?!

Keeping Your Websites Alive

- Longevity by design
- Choose wisely
  - File formats
  - Bells and Whistles

Who Will Capture Your Websites?

- The “Creator(s)”
  - Carefully negotiate the answers to previous questions
  - Test transfers are good

Who Will Capture Your Websites?

- A “Hired Gun”
  - Carefully negotiate the answers to previous questions
  - Service level agreements
  - Develop your assessment criteria

Who Will Capture Your Websites?

- You
  - Carefully negotiate the answers to previous questions
  - Customer expectations
  - Resource allocator signoff

Before Your Safari

- Understand WHY you are going
- Understand what you are up against
- Let others know where you are going
- Make sure you are properly equipped
- Make adequate provisions for your captured prey
- Make sure someone has your back

Capturing Websites - Insight From Experience, Looking Toward Tomorrow

Please Complete Your Session Evaluation

Mark Conrad
National Archives and Records Administration
http://www.archives.gov/era/research/

Education Code: TR1-1227