

# Accessioning 2.0: Documenting Institutional Outreach in the 21st Century

The Smithsonian Institution has more than 100 public websites and nearly 300 social media sites with different stories to share

## What's out there



Smithsonian Institution museums had 30 million visits in FY 2009 while the Smithsonian's numerous websites, blogs, virtual museums, and other social media sites had more than 115 million unique visits.

Most of SI's websites have been appraised as permanent institutional records. The Smithsonian Institution Archives has been archiving the Institution's websites through a variety of methods since the late 1990s. This included copying files from the content management system and using the HTTrack crawler.

By 2009, a more automated approach was needed to address the growing number of websites and social media sites.



Smithsonian Tropical Research Institute homepage in 1996

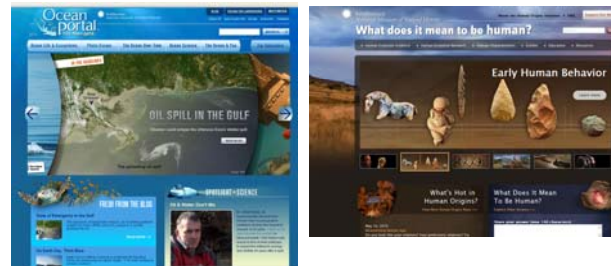


Smithsonian Tropical Research Institute homepage in 2010

## Appraisal and transfer

Appraisal is based on content. Most traditional websites are deemed permanently valuable because of the information they convey. Appraisal of social media sites must be done at the account-level because every unit uses these tools in different ways.

- Capture full baseline of SI websites
- Capture full one-time baseline of most SI social media sites as documentation that they existed
- Work with webmasters to determine frequency of website crawls
- Appraise current social media accounts each time the office's website is crawled



The Smithsonian National Museum of Natural History launched the Ocean Portal and Human Origins websites in 2010.

## Doing it on our own

Challenges of archiving online Smithsonian

- Technology – Mandated-Windows shop at Smithsonian Institution made Heritrix and Wayback install and usage difficult. Heritrix supported in Linux.

*Solution involved using older documentation found on Web regarding Heritrix install on Windows.*

- Rich media – Cannot capture everything as Flash remains problematic for web crawlers.

- Frequency of captures – Constantly changing content  
*Solution involves SI Archives determining when to crawl such as a new website or major redesign*

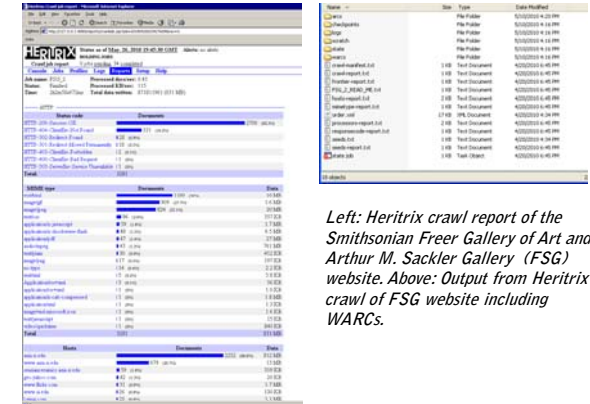
- Keeping up – SI units are constantly creating new sites  
*Solution involved creating site registry in SharePoint that is maintained by SI Archives and can be updated by webmasters*

- Legal issues – Content on third-party sites  
*Solution involves working with General Counsel to establish agreements and using guidance from General Services Administration and NARA*

- SI Social Media sites from YouTube to Facebook to Twitter

## What worked and what didn't

New Windows server could not run Heritrix 1.14.3 (crawler). Had to go back to Windows 2000 machine to conduct crawls. Unsuccessful in installing open-source Wayback to view the crawled content (WARCs – Web ARChive format) in Windows environment. Continuing to work to get Wayback operational in Linux



Left: Heritrix crawl report of the Smithsonian Freer Gallery of Art and Arthur M. Sackler Gallery (FSG) website. Above: Output from Heritrix crawl of FSG website including WARCs.

## A little help from the community



National Science Resources Center website page viewed in open-source Wayback at Library of Congress. Site was crawled by the Smithsonian Institution Archives with Heritrix.

*Note: The three menu items not appearing are due to Wayback display issue. The What's New not displaying because Heritrix couldn't find due to link construction.*

## Next steps

- Continue work to get Wayback functional
- Continue crawls of sites using Heritrix