From the foundational web to founding a web archives

Creating a formalized web archiving program at the MIT Libraries

IIPC Web Archiving Conference | 2019-06-06
Joe Carrano | Digital Archivist | Department of Distinctive Collections
MIT Libraries
@joecar25
@joe@digipres.club

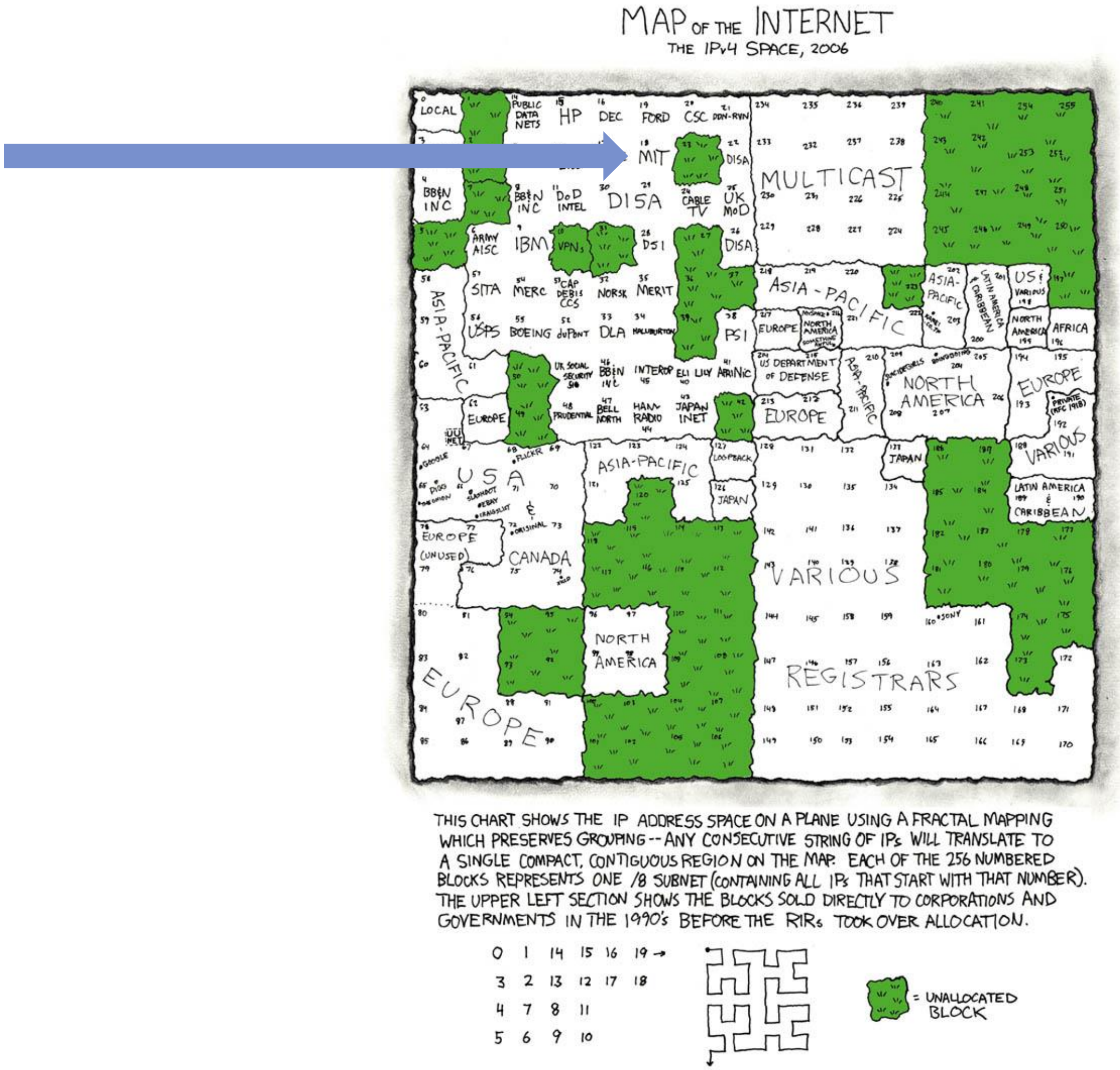
The Massachusetts Institute of Technology has a long history with the Internet and the Web.

Web archiving of active sites at MIT only began in 2015-2016 and only started increasing into a full program in 2018.

Ad Hoc → Formalized

1. Prioritize
2. Know what there is to crawl
3. Appraise

1. We need to: prioritize



MAP OF THE INTERNET

THE IPv4 SPACE, 2006

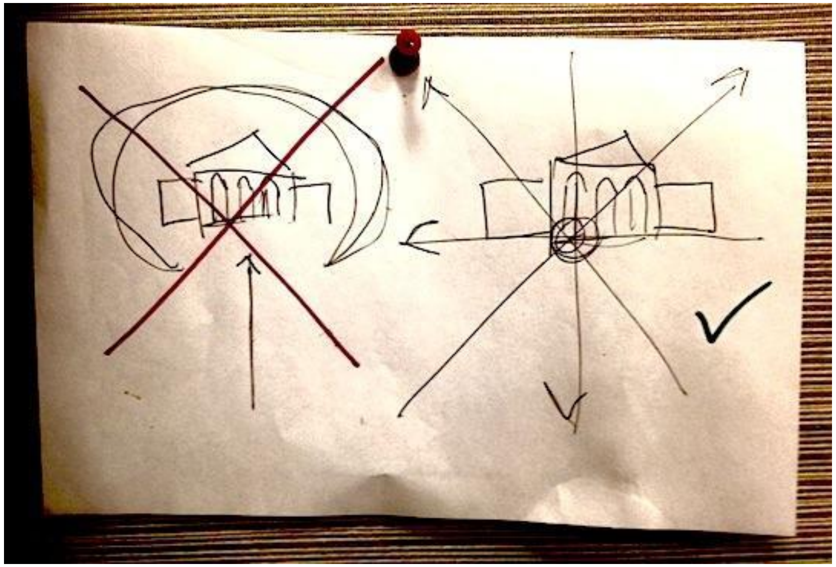
THIS CHART SHOWS THE IP ADDRESS SPACE ON A PLANE USING A FRACTAL MAPPING WHICH PRESERVES GROUPING -- ANY CONSECUTIVE STRING OF IPs WILL TRANSLATE TO A SINGLE COMPACT, CONTIGUOUS REGION ON THE MAP. EACH OF THE 256 NUMBERED BLOCKS REPRESENTS ONE /8 SUBNET (CONTAINING ALL IPs THAT START WITH THAT NUMBER). THE UPPER LEFT SECTION SHOWS THE BLOCKS SOLD DIRECTLY TO CORPORATIONS AND GOVERNMENTS IN THE 1990's BEFORE THE RIRs TOOK OVER ALLOCATION.

At MIT, the community uses its large IP range and ease of hosting to the fullest

With a large amount of web content created at MIT, where to begin?

Determining initial scope of our collecting

- Focus on Departments, Labs, and Centers (DLCs) of the Institute due to a clear policy to collect.
- Target unique content to the Institute.
- Build on pre-existing archival collecting strengths.



2. We need to: know what there is to crawl

Developing a seed list

- We did not want to create a list manually
- Sometimes all you have to do is ask a web developer

Me: *“Can you run a query on your back-end database and give me the list of all the websites listed?”*

Them: *“Sure, here you go!”*

Me: *“Thanks!”* *starts sweating thinking about where to start with these hundreds of urls*

3. We need to: appraise

INTERNET ARCHIVE

Brewster Kahle and IA love MIT.edu

Brewster Surton Kahle

MIT Class of ‘82

Are other groups crawling these sites?

- The simplest method was to create a script to query the Wayback CDX server API for each seed in the list we gathered from our outreach efforts
- We started looking at pages with less than 50 crawls of their home page
- If warranted we would investigate further with memento, waybackprov, and browsing IA crawls to see how well covered they are

MIT Libraries Values

“The MIT Libraries contribute to a better world by pursuing Social Justice and an Ethic of Care”

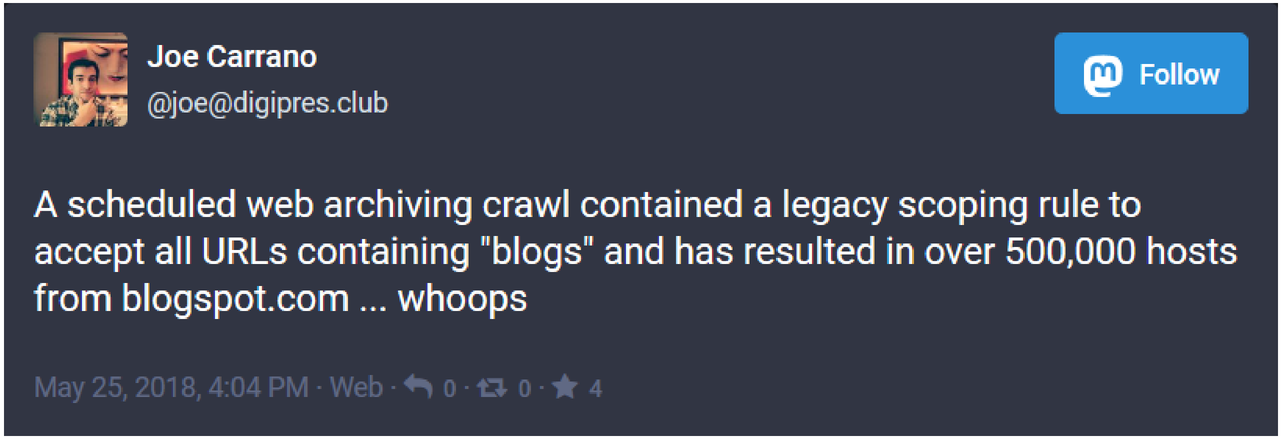
- “We strive to promote many voices, and to reflect diversity of both knowledge itself and ways of knowing in our collections, and in our approach to information management and organization.”
- “We aspire to leverage the work, values, and resources of libraries and archives as forces for social justice in our communities.”

More here: https://libraries.mit.edu/about/organization/

How and which sites should we collect based on MIT Libraries values?

- Consider areas that have not been well covered in our archival collections in the past: women, people of color, non-faculty staff members, and students.
- Somewhat limited by our initial scope, how to collect sites that document these themes when Institute websites are largely administrative?
- Starting a culture of informing and informed consent.

Making appraisal decisions after crawls


Joe Carrano
@joe@digipres.club

A scheduled web archiving crawl contained a legacy scoping rule to accept all URLs containing "blogs" and has resulted in over 500,000 hosts from blogspot.com ... whoops

May 25, 2018, 4:04 PM · Web · 0 · 0 · 4

- Often appraisal choices or how to scope a crawl do not arise until after an initial crawl.
- Evaluate what needs to be included in a crawl and what level of effort to put into QA.
- Video was one problem area that led to other archival collecting efforts.

What about access and description?

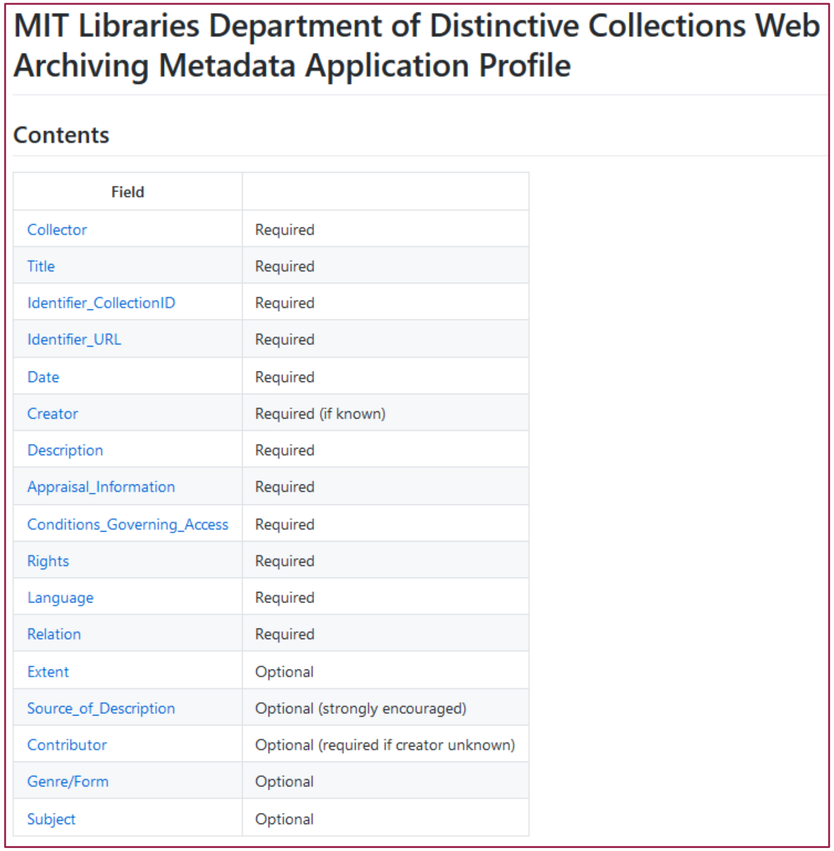
Describing Web Archives

A number of descriptive frameworks have been released in the past year in the United States that informed my thinking

- OCLC Descriptive Metadata for Web Archiving report
- University of Virginia Web Archiving Application Profile
- Growing community of practice - Describing Web Archives Users Group
- Jennifer Douglas, “Towards More Honest Description”

MIT Web Archiving Metadata Application Profile

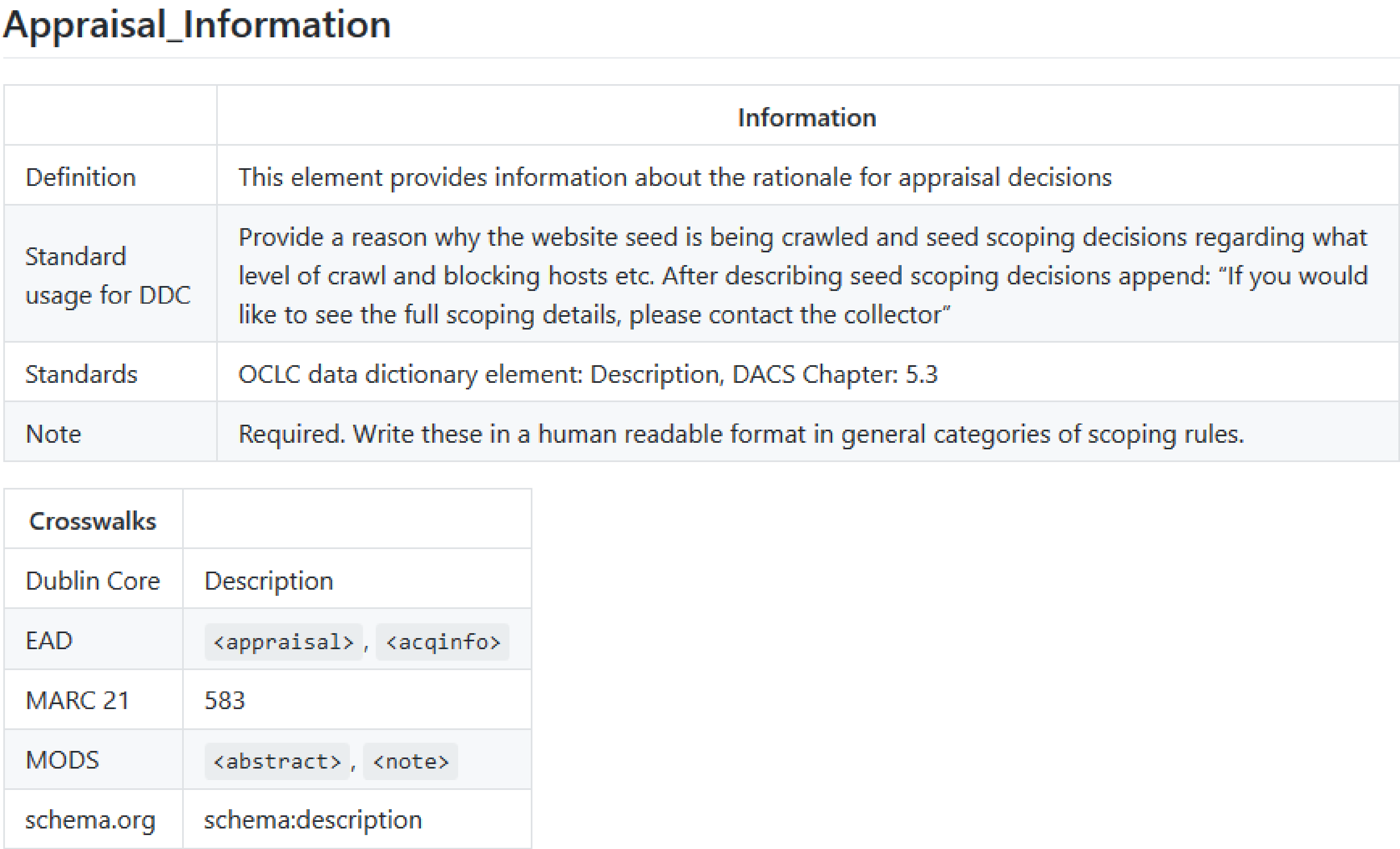
- Created a metadata profile for local use.
- Largely pulling together information from OCLC and UVA.
- Available for reuse on Github
- Added additional fields for archival description.



MIT Libraries Department of Distinctive Collections Web Archiving Metadata Application Profile

Contents

Field	
Collector	Required
Title	Required
Identifier_CollectionID	Required
Identifier_URL	Required
Date	Required
Creator	Required (if known)
Description	Required
Appraisal_Information	Required
Conditions_Governing_Access	Required
Rights	Required
Language	Required
Relation	Required
Extent	Optional
Source_of_Description	Optional (strongly encouraged)
Contributor	Optional (required if creator unknown)
Genre/Form	Optional
Subject	Optional

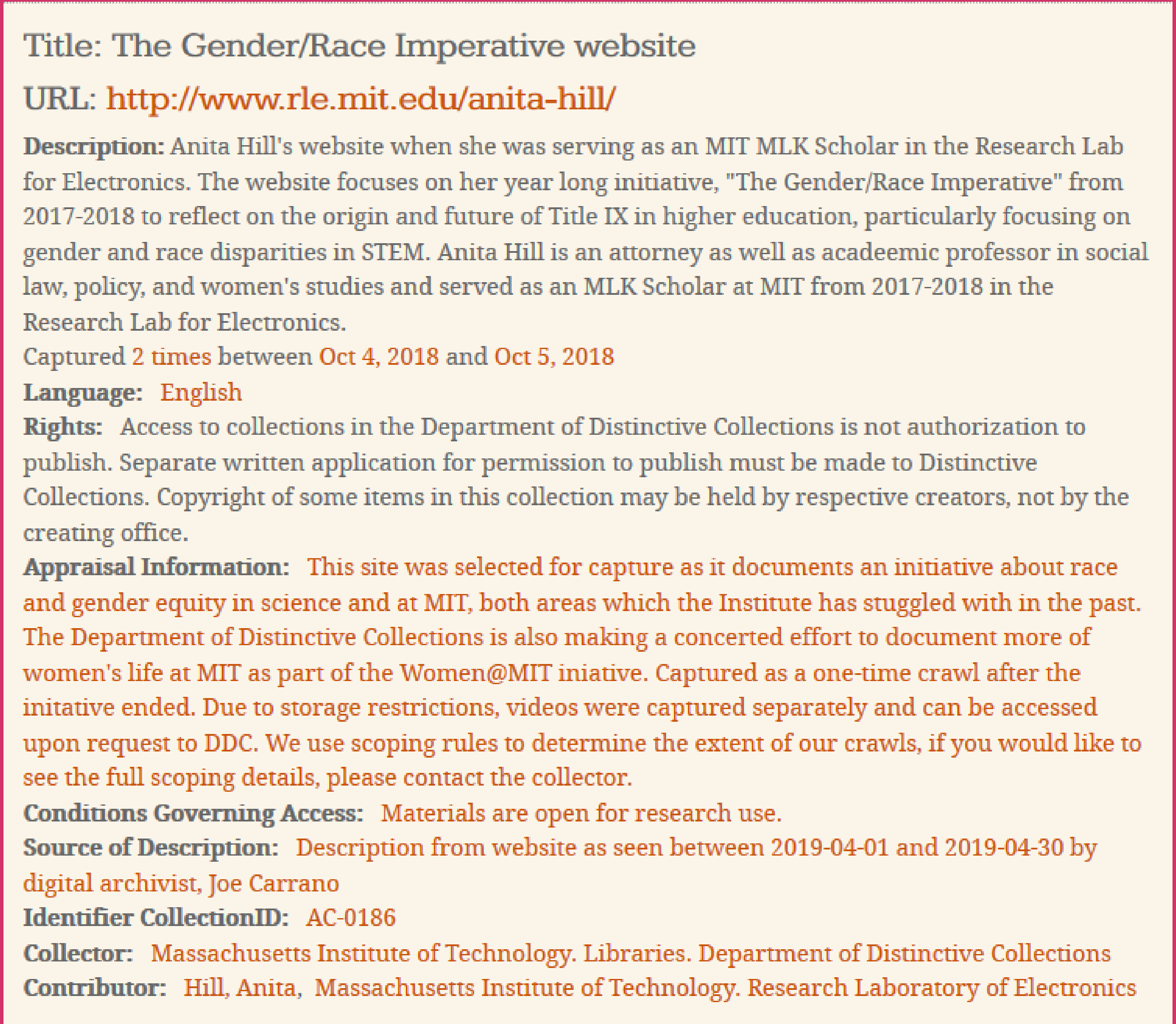


Appraisal_Information

	Information
Definition	This element provides information about the rationale for appraisal decisions
Standard usage for DDC	Provide a reason why the website seed is being crawled and seed scoping decisions regarding what level of crawl and blocking hosts etc. After describing seed scoping decisions append: “If you would like to see the full scoping details, please contact the collector”
Standards	OCLC data dictionary element: Description, DACS Chapter: 5.3
Note	Required. Write these in a human readable format in general categories of scoping rules.

Crosswalks	
Dublin Core	Description
EAD	`<appraisal>`, `<acqinfo>`
MARC 21	583
MODS	`<abstract>`, `<note>`
schema.org	schema:description

Archive-It example

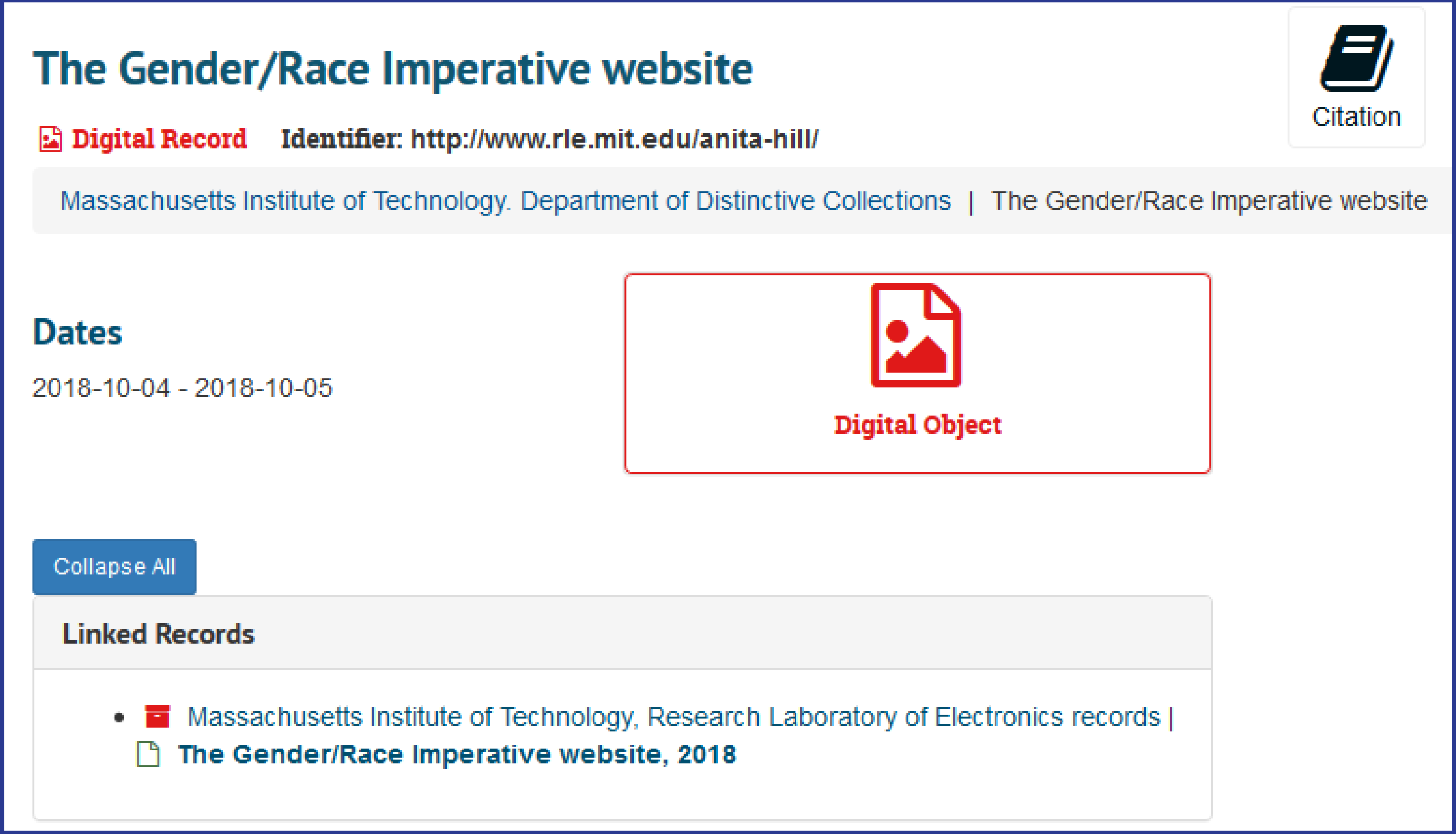


Title: The Gender/Race Imperative website

URL: http://www.rle.mit.edu/anita-hill/

Description: Anita Hill's website when she was serving as an MIT MLK Scholar in the Research Lab for Electronics. The website focuses on her year long initiative, "The Gender/Race Imperative" from 2017-2018 to reflect on the origin and future of Title IX in higher education, particularly focusing on gender and race disparities in STEM. Anita Hill is an attorney as well as acadeemic professor in social law, policy, and women's studies and served as an MLK Scholar at MIT from 2017-2018 in the Research Lab for Electronics.
Captured 2 times between Oct 4, 2018 and Oct 5, 2018
Language: English
Rights: Access to collections in the Department of Distinctive Collections is not authorization to publish. Separate written application for permission to publish must be made to Distinctive Collections. Copyright of some items in this collection may be held by respective creators, not by the creating office.
Appraisal Information: This site was selected for capture as it documents an initiative about race and gender equity in science and at MIT, both areas which the Institute has stuggled with in the past. The Department of Distinctive Collections is also making a concerted effort to document more of women's life at MIT as part of the Women@MIT iniative. Captured as a one-time crawl after the initative ended. Due to storage restrictions, videos were captured separately and can be accessed upon request to DDC. We use scoping rules to determine the extent of our crawls, if you would like to see the full scoping details, please contact the collector.
Conditions Governing Access: Materials are open for research use.
Source of Description: Description from website as seen between 2019-04-01 and 2019-04-30 by digital archivist, Joe Carrano
Identifier CollectionID: AC-0186
Collector: Massachusetts Institute of Technology. Libraries. Department of Distinctive Collections
Contributor: Hill, Anita, Massachusetts Institute of Technology. Research Laboratory of Electronics

ArchivesSpace example



The Gender/Race Imperative website

Digital Record **Identifier: http://www.rle.mit.edu/anita-hill/**

Massachusetts Institute of Technology. Department of Distinctive Collections | The Gender/Race Imperative website

Dates

2018-10-04 - 2018-10-05

Linked Records

- Massachusetts Institute of Technology, Research Laboratory of Electronics records | **The Gender/Race Imperative website, 2018**

Future Growth

Continuing to scale up

- Distribute the work
  - Curators take the lead of selection
  - Metadata experts take the lead on most description
  - Involve more archivists in QA
- Distribute metadata in more systems
- Distribute the collections as data to researchers

Photo credit

“Map of the Internet”, Randall Munroe https://xkcd.com/195/ Used under Creative Commons Attribution-NonCommercial 2.5 License

NYPL Labs doodle, NYPL Labs website (RIP)
https://www.nypl.org/collections/labs

Photo of Brewster Kahle, *Technique* Yearbook 1982, Department of Distinctive Collections, Massachusetts Institute of Technology, Cambridge, MA

Hvala

Thank you!