# Taming Cerberus:

## Balancing Legacy Digitization and Emerging Needs

Carnegie Mellon University | March 2022

Julia Corrin, Ann Marie Mesco

1994

## HELIOS

**1 collection**

First of its kind to provide deep access to archival materials.  Used NLP for search.

1999

## DIVA

**16 collections**

The next generation of HELIOS, built and managed entirely by the CMU libraries.

2011

## ArchivalWare

**26 collections**

A vended system, designed to let digitized content be added quickly without technical support.

2021

## Islandora

The next generation of access to digitized content -- incorporating linked open data, exhibits, and improved user access.

# Collections by the Numbers

**26** years

**2.2** million pages

**37+** terabytes

# Digitization is more than just scanning…

- Manage legacy collections to ensure ongoing accessibility
- Adapt to evolving standards and technologies for new collections
- Scanning as a skill set

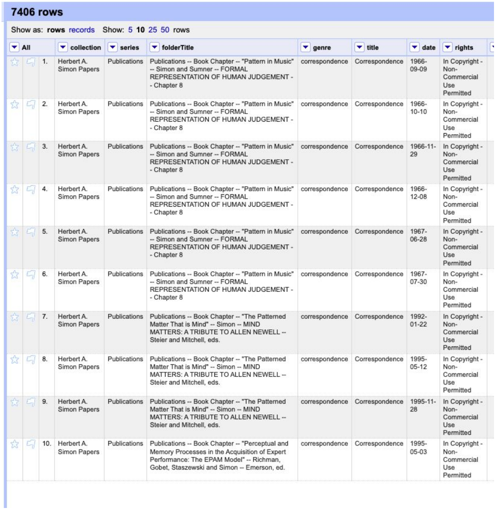
*Hans Sebald Beham, Public domain, via Wikimedia Commons*

# Access is more than just a platform…

- User expectations & knowledge
  - Different access needs for different audiences
- Accessibility
  - WCAG, ADA
  - Annotated PDFs
- Metadata
  - Search and retrieval is limited to what a computer can understand

# Whither Metadata?

- Metadata is under-discussed when it comes to digitization
- Metadata is frequently the most resource intensive aspect of digitization
- Metadata can be the source of most long-term technical debt issues for digital projects



**7406 rows**

Show as: **rows** records Show: 5 **10** 25 50 rows

| All | collection | series | folderTitle | genre | title | date | rights |
|---|---|---|---|---|---|---|---|
| 1. | Herbert A. Simon Papers | Publications | Publications -- Book Chapter -- "Pattern in Music" -- Simon and Sumner -- FORMAL REPRESENTATION OF HUMAN JUDGEMENT -- Chapter 8 | correspondence | Correspondence | 1966-09-09 | In Copyright - Non-Commercial Use Permitted |
| 2. | Herbert A. Simon Papers | Publications | Publications -- Book Chapter -- "Pattern in Music" -- Simon and Sumner -- FORMAL REPRESENTATION OF HUMAN JUDGEMENT -- Chapter 8 | correspondence | Correspondence | 1966-10-10 | In Copyright - Non-Commercial Use Permitted |
| 3. | Herbert A. Simon Papers | Publications | Publications -- Book Chapter -- "Pattern in Music" -- Simon and Sumner -- FORMAL REPRESENTATION OF HUMAN JUDGEMENT -- Chapter 8 | correspondence | Correspondence | 1966-11-29 | In Copyright - Non-Commercial Use Permitted |
| 4. | Herbert A. Simon Papers | Publications | Publications -- Book Chapter -- "Pattern in Music" -- Simon and Sumner -- FORMAL REPRESENTATION OF HUMAN JUDGEMENT -- Chapter 8 | correspondence | Correspondence | 1966-12-08 | In Copyright - Non-Commercial Use Permitted |
| 5. | Herbert A. Simon Papers | Publications | Publications -- Book Chapter -- "Pattern in Music" -- Simon and Sumner -- FORMAL REPRESENTATION OF HUMAN JUDGEMENT -- Chapter 8 | correspondence | Correspondence | 1967-06-28 | In Copyright - Non-Commercial Use Permitted |
| 6. | Herbert A. Simon Papers | Publications | Publications -- Book Chapter -- "Pattern in Music" -- Simon and Sumner -- FORMAL REPRESENTATION OF HUMAN JUDGEMENT -- Chapter 8 | correspondence | Correspondence | 1967-07-30 | In Copyright - Non-Commercial Use Permitted |
| 7. | Herbert A. Simon Papers | Publications | Publications -- Book Chapter -- "The Patterned Matter That is Mind" -- Simon -- MIND MATTERS: A TRIBUTE TO ALLEN NEWELL -- Steier and Mitchell, eds. | correspondence | Correspondence | 1992-01-22 | In Copyright - Non-Commercial Use Permitted |
| 8. | Herbert A. Simon Papers | Publications | Publications -- Book Chapter -- "The Patterned Matter That is Mind" -- Simon -- MIND MATTERS: A TRIBUTE TO ALLEN NEWELL -- Steier and Mitchell, eds. | correspondence | Correspondence | 1995-05-12 | In Copyright - Non-Commercial Use Permitted |
| 9. | Herbert A. Simon Papers | Publications | Publications -- Book Chapter -- "The Patterned Matter That is Mind" -- Simon -- MIND MATTERS: A TRIBUTE TO ALLEN NEWELL -- Steier and Mitchell, eds. | correspondence | Correspondence | 1995-11-28 | In Copyright - Non-Commercial Use Permitted |
| 10. | Herbert A. Simon Papers | Publications | Publications -- Book Chapter -- "Perceptual and Memory Processes in the Acquisition of Expert Performance: The EPAM Model" -- Richman, Gobet, Staszewski and Simon -- Emerson, ed. | correspondence | Correspondence | 1995-05-03 | In Copyright - Non-Commercial Use Permitted |

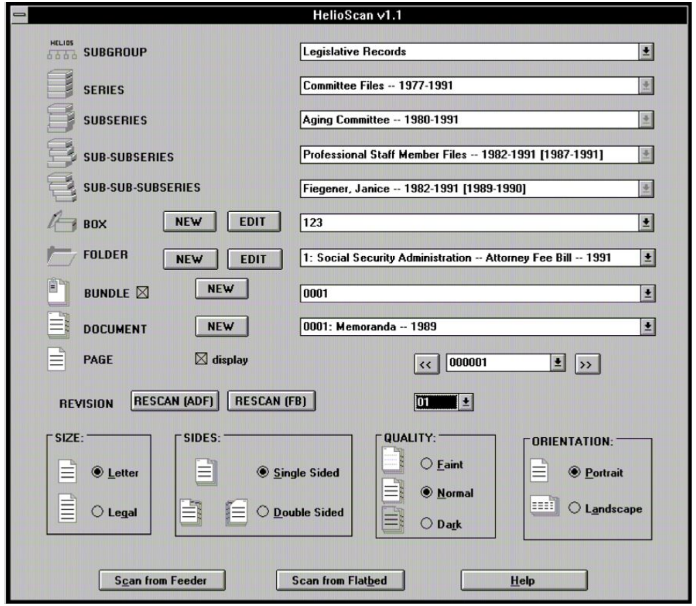
# Case Study 1: Sen. H. John Heinz III Papers

**Goal**: Replicate the experience of using the physical collection in a digital space.

- Received by the archives in 1991, digitized in 1994
- Over $1.5 million in project funding
- Over 300,000 documents / 390 cubic feet

# 1994 Challenges - The Wild West

- Lack of best practice
  - No established scanning guidelines for this type of scanning
    - FADGI - 2007
  - Limited understanding of metadata
    - DublinCore - 1995; EAD 1.0 - 1998
- Technology Limitations
  - Scanners were expensive and limited in what they could do
    - 400 DPI, black and white TIFFs
  - No widely adopted content management systems



HelioScan v1.1

SUBGROUP: Legislative Records
SERIES: Committee Files -- 1977-1991
SUBSERIES: Aging Committee -- 1980-1991
SUB-SUBSERIES: Professional Staff Member Files -- 1982-1991 [1987-1991]
SUB-SUB-SUBSERIES: Fiegener, Janice -- 1982-1991 [1989-1990]
BOX: 123
FOLDER: 1: Social Security Administration -- Attorney Fee Bill -- 1991
BUNDLE ☒: 0001
DOCUMENT: 0001: Memoranda -- 1989
PAGE: ☒ display 000001
REVISION: 01

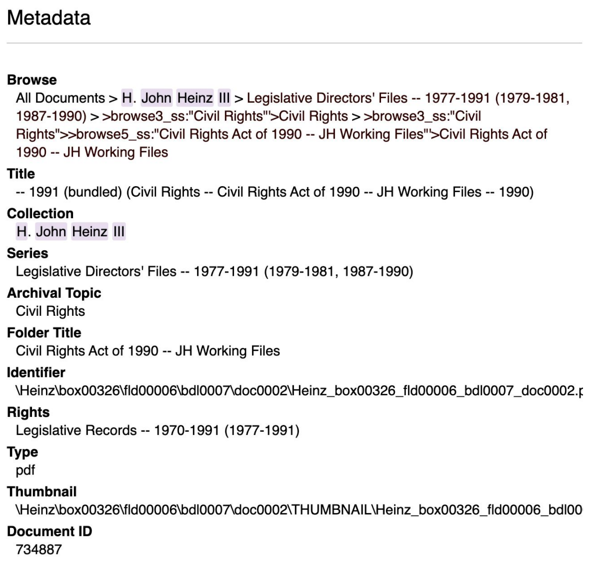
SIZE: ● Letter ○ Legal
SIDES: ● Single Sided ○ Double Sided
QUALITY: ○ Faint ● Normal ○ Dark
ORIENTATION: ● Portrait ○ Landscape

# 2022 Challenges

- TIFFs can no longer be read by modern TIFF libraries
  - Required to use modern viewers like Mirador
- Periodic re-OCR to leverage technology improvements
  - Re-OCRing 700,000 pages is resource intensive
- Long term file management
  - Master vs. derivative
  - Corruption overtime

# More 2022 Challenges

- Metadata remediation
  - Increased “noise” over time
- Structural assumptions
  - The way users interact with the collection has evolved
  - 1994 assumptions about users are baked into the collection
- Persistency
  - DOIs, PURLs need to be maintained overtime
  - Limits changes that can be made



Metadata

**Browse**
All Documents > H. John Heinz III > Legislative Directors' Files -- 1977-1991 (1979-1981, 1987-1990) > >browse3_ss:"Civil Rights"'>Civil Rights > >browse3_ss:"Civil Rights">>browse5_ss:"Civil Rights Act of 1990 -- JH Working Files"'>Civil Rights Act of 1990 -- JH Working Files

**Title**
-- 1991 (bundled) (Civil Rights -- Civil Rights Act of 1990 -- JH Working Files -- 1990)

**Collection**
H. John Heinz III

**Series**
Legislative Directors' Files -- 1977-1991 (1979-1981, 1987-1990)

**Archival Topic**
Civil Rights

**Folder Title**
Civil Rights Act of 1990 -- JH Working Files

**Identifier**
\Heinz\box00326\fld00006\bdl0007\doc0002\Heinz_box00326_fld00006_bdl0007_doc0002.p

**Rights**
Legislative Records -- 1970-1991 (1977-1991)

**Type**
pdf

**Thumbnail**
\Heinz\box00326\fld00006\bdl0007\doc0002\THUMBNAIL\Heinz_box00326_fld00006_bdl00

**Document ID**
734887

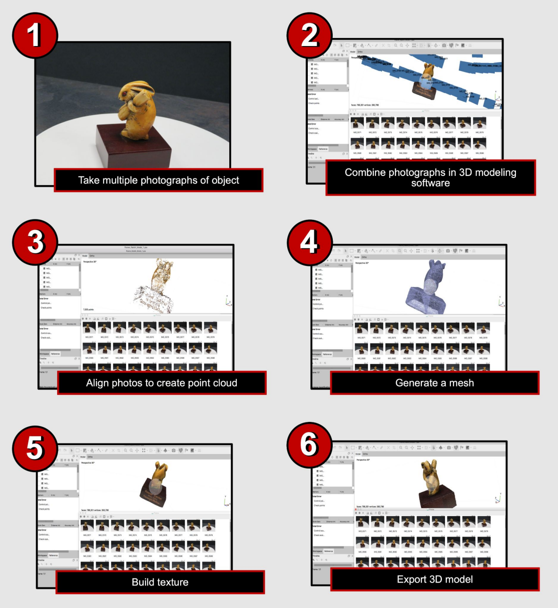
# Case Study 2: New Formats

**Goal:** Capture “non-traditional” theses, dissertations, and research outputs

- Equitable preservation of scholarly output, regardless of discipline
- Expand scanning capabilities without expensive purchases
- Types of objects: Architectural models, design projects

# Method 1: Photogrammetry

- Create a 3D model from a series of 2D photographs
- Lower cost entry point than dedicated 3D scanners
- Preservation advantages
- Leverages technology and skills already available in many digitization labs



1 Take multiple photographs of object

2 Combine photographs in 3D modeling software

3 Align photos to create point cloud

4 Generate a mesh

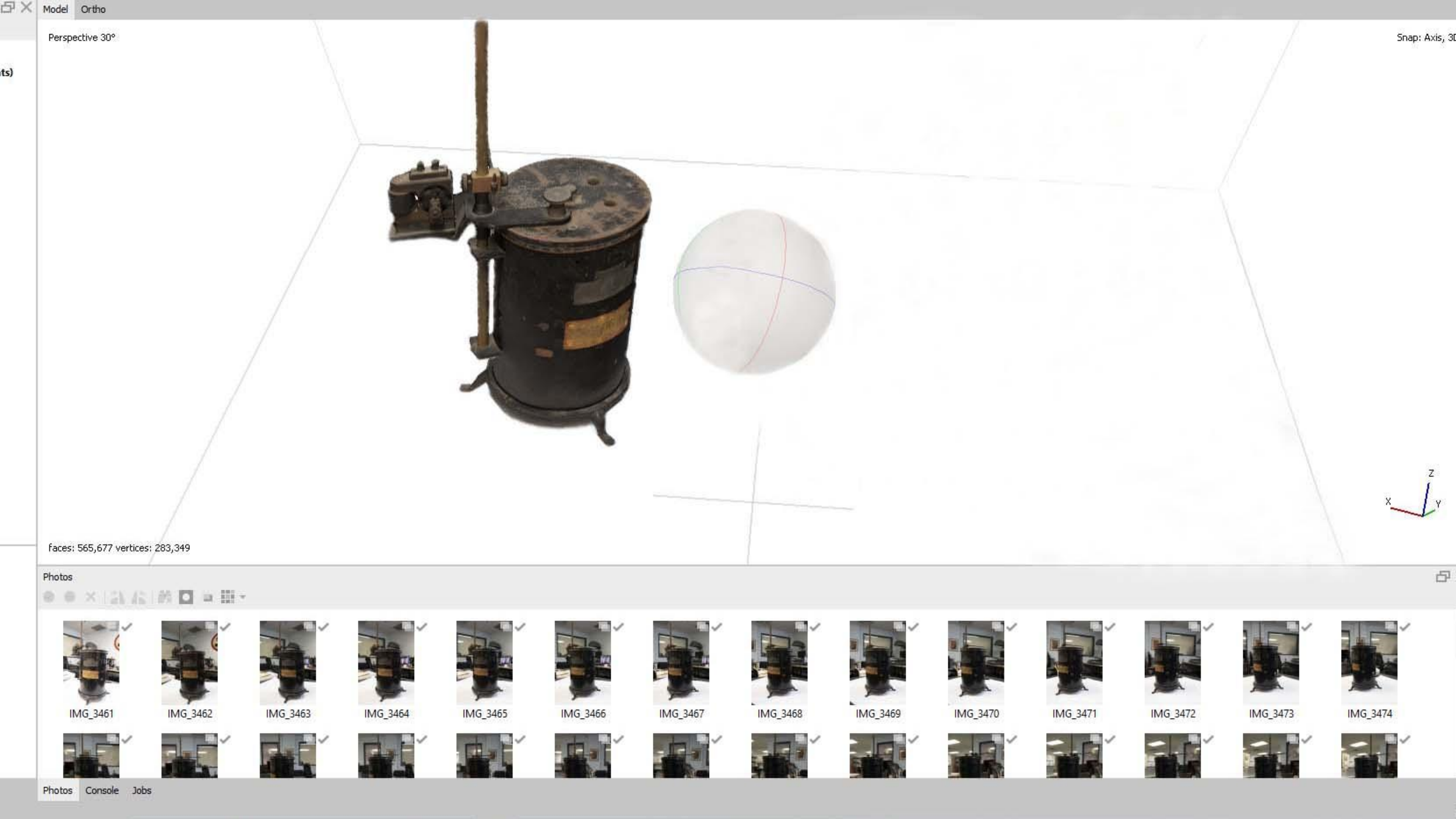
5 Build texture

6 Export 3D model

*Credit: Ann Marie Mesco, Emma Slayton, Jon McIntire, Joe Mesco*

# Photogrammetry Starter Kit

- Decent digital camera
    - Cannon DSLR EOS 80D
- Computer
    - Digital Storm Velox
    - Gaming computer, VR Computer
- Software
    - Agisoft Metashape
    - Meshmixer
- Turntable



Perspective 30°

Snap: Axis, 30

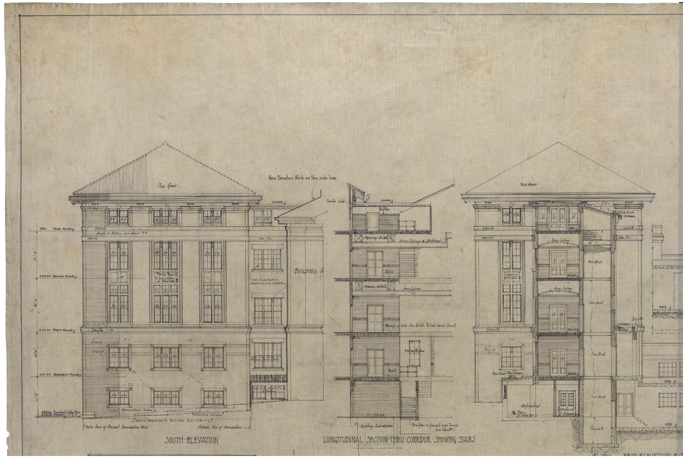
faces: 565,677 vertices: 283,349

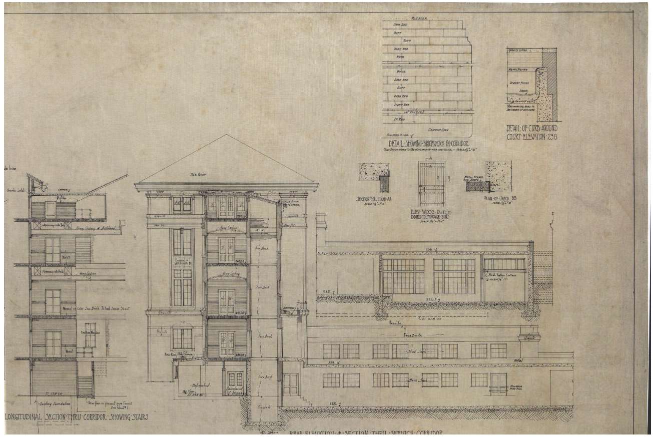
# Method 2: Image Stitching

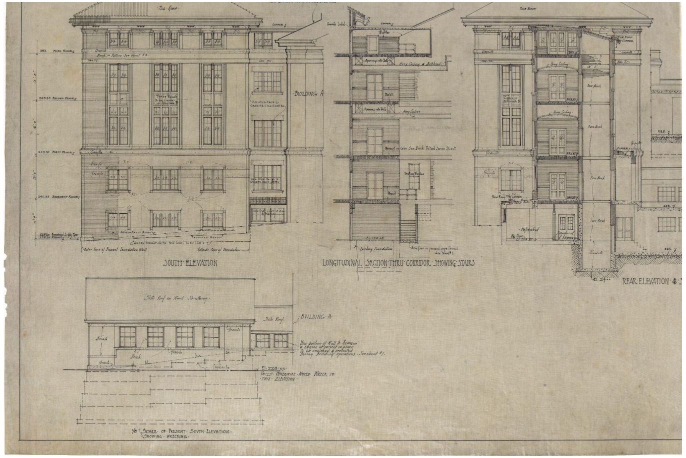
- Capture large format works that will not fit on a traditional scanner
- Digitize fragile large format works without damaging the original
- Utilize skills of experienced scanner operators

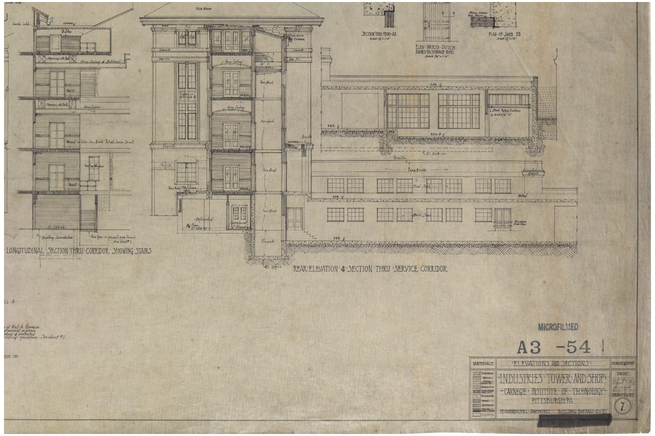
# Large Format Starter Kit

- Existing overhead scanner
  - CopiBook scanner; 23x16 in scan area
  - Large documents require 4-6 captures
- Adobe Photoshop
- People!

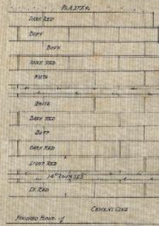

Tile Roof
New Elevation Work on this side line
BUILDING A
SOUTH·ELEVATION
LONGITUDINAL·SECTION·THRU·CORRIDOR·SHOWING·STAIRS
Tile Roof
REAR·ELEVATION


PLASTER
DARK RED
BUFF
BUFF
DARK RED
WHITE
WHITE
DARK RED
BUFF
DARK RED
LIGHT RED
14 COURSES
LT. RED
CEMENT COVE
FINISHED FLOOR
DETAIL·SHOWING·BRICKWORK·IN·CORRIDOR
OLD BRICK WORK TO BE MATCHED IN SIZE AND COLOR – SCALE 1½"=1'-0"
GRANITE COPING
METAL NOSING
CEMENT FINISH
DETAIL·OF·CURB·AROUND
COURT·ELEVATION·238
SECTION·THRU·HEAD·AA
SCALE 1½"=1'-0"
ELEV. WOOD·DOOR
PLAN·OF·JAMB BB
SCALE 1½"=1'-0"
METAL STRAPS AND BOLTS
LONGITUDINAL·SECTION·THRU·CORRIDOR·SHOWING·STAIRS
REAR·ELEVATION·&·SECTION·THRU·SERVICE·CORRIDOR


Tile Roof
BUILDING A
SOUTH·ELEVATION
LONGITUDINAL·SECTION·THRU·CORRIDOR·SHOWING·STAIRS
REAR·ELEVATION·&·S
Outer Face of Present Foundation Wall
Outside Face of Foundation
Slate Roof on Wood Sheathing
Slate Roof
BUILDING·A
Brick
Granite
Concrete
This portion of Wall to Remain & Engine at present in place to be enclosed & protected During Building operations. See sheet #
El. 228·00
UNLESS·OTHERWISE·NOTED·WRECK·TO·THIS·ELEVATION·
⅛" SCALE· OF PRESENT· SOUTH·ELEVATION· SHOWING·WRECKING·

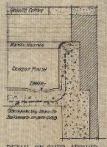

SECTION·THRU·HEAD·AA
ELEV. WOOD·DOOR
PLAN·OF·JAMB BB
LONGITUDINAL·SECTION·THRU·CORRIDOR·SHOWING·STAIRS
REAR·ELEVATION·&·SECTION·THRU·SERVICE·CORRIDOR
MICROFILMED
A3 -54
MATERIALS
·ELEVATIONS AND SECTIONS·
SCALE ⅛"=1'-0"
CONCRETE
BRICK
STONE GRANITE
ARCH. T.C.
FIREPROOF
PLASTER
WOOD
MARBLE
METAL
·INDUSTRIES·TOWER·AND·SHOP·
–CARNEGIE·INSTITUTE·OF·TECHNOLOGY–
PITTSBURGH PA
H. HORNBOSTEL·ARCHITECT BUILDING·BUREAU·C·I·T·
DATE
April 7 1915
Revised
May 3, 1915
DRAWING NO
7.

[redacted]

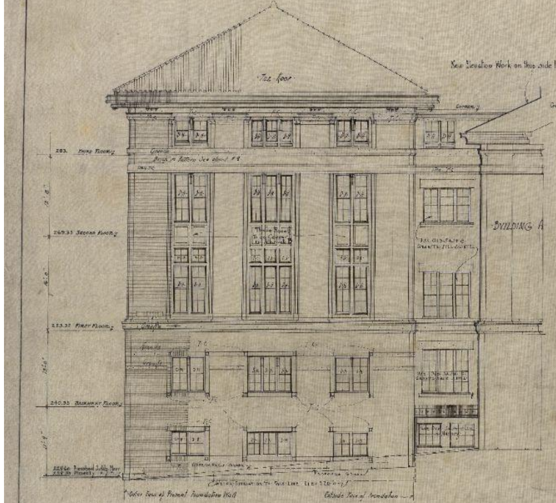


DETAIL: SHOWING BRICKWORK IN CORRIDOR

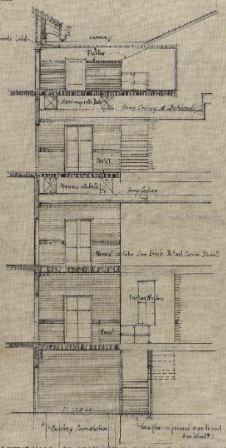
Old brick work to be made good of such are cutting. Scale 1 1/2" = 1'-0"



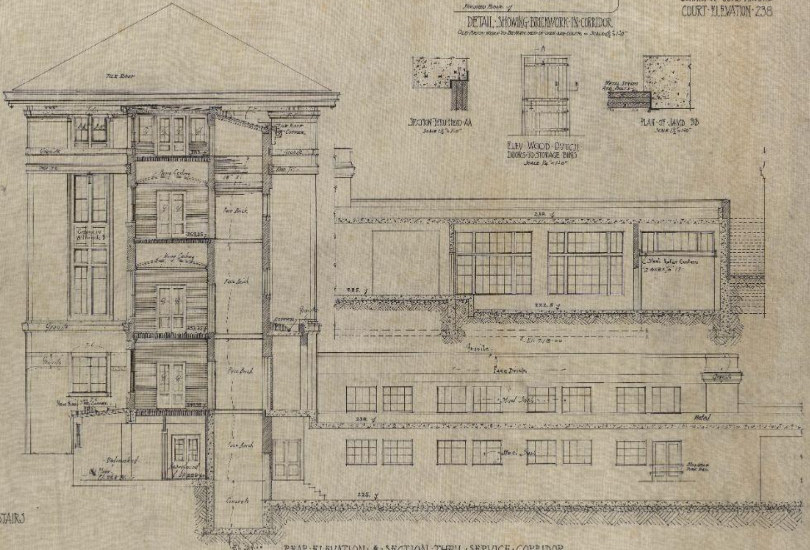
DETAIL: OF CURB AROUND COURT ELEVATION 238


Tile Roof
New Elevation Work on this side line
Coppery
Granite Label
283. Third Floor
Granite
Brick in Pattern See sheet #8
269.35 Second Floor
Three Panels to be Copper
Building A
253.35 First Floor
Granite
240.35 Basement Floor
228.60 Finished Subway Floor
Proposed Grade
Outer Face of Present Foundation Wall
Outside Face of Foundation

SOUTH ELEVATION



LONGITUDINAL SECTION THRU CORRIDOR SHOWING STAIRS



SECTION THRU HEAD AA

Scale 1 1/2" = 1'-0"

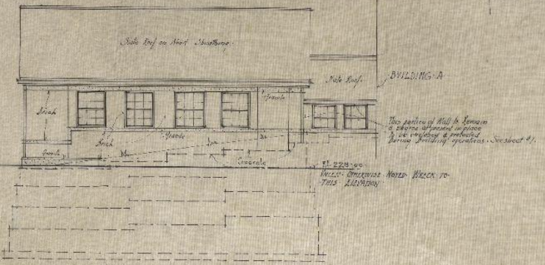
ELEV. WOOD DUTCH DOORS TO STORAGE BINS

Scale 3/4" = 1'-0"

PLAN OF JAMB BB

Scale 1 1/2" = 1'-0"

REAR ELEVATION & SECTION THRU SERVICE CORRIDOR


Slate Roof on Wood Sheathing
Slate Roof
BUILDING A
Brick
Granite
Concrete
This portion of Wall to Remain & Frame at present in place to be enclosed & protected During Building operations. See sheet #1.
El. 228.00
Unless Otherwise Noted Wreck to This Elevation

1/8" Scale of Present South Elevation Showing Wrecking

| MATERIALS | ELEVATIONS AND SECTIONS | SCALE 1/8" = 1'-0" |
|---|---|---|
|  | INDUSTRIES TOWER AND SHOP<br>CARNEGIE INSTITUTE OF TECHNOLOGY<br>PITTSBURGH PA<br>H HORNBOSTEL ARCHITECT BUILDING BUREAU C.I.T. | DATE<br>April 1915<br>DRAWING NO.<br>7 |

[redacted]

# Conclusions

- Don’t start a digitization project without a metadata plan in place
- Always consider long term needs - every new digitization project reduces your bandwidth overtime
- Invest in people, not just technology

# Acknowledgements

Ann Marie Mesco, Digitization Projects Manager

Joe Mesco, Scanner Operator

Jon McIntire, Scanner Operator

Lina Crowe, Metadata Specialist