Mass Digitization on Demand

Automation and Terrible Metadata

UNIVERSITY AT ALBANY
State University of New York

Gregory Wiedeman
University Archivist
GWiedeman@albany.edu
@GregWiedeman

We Digitize for Remote Requests

- 1-3 requests for scanning per week
- Performed by student assistants

Simple Fact:

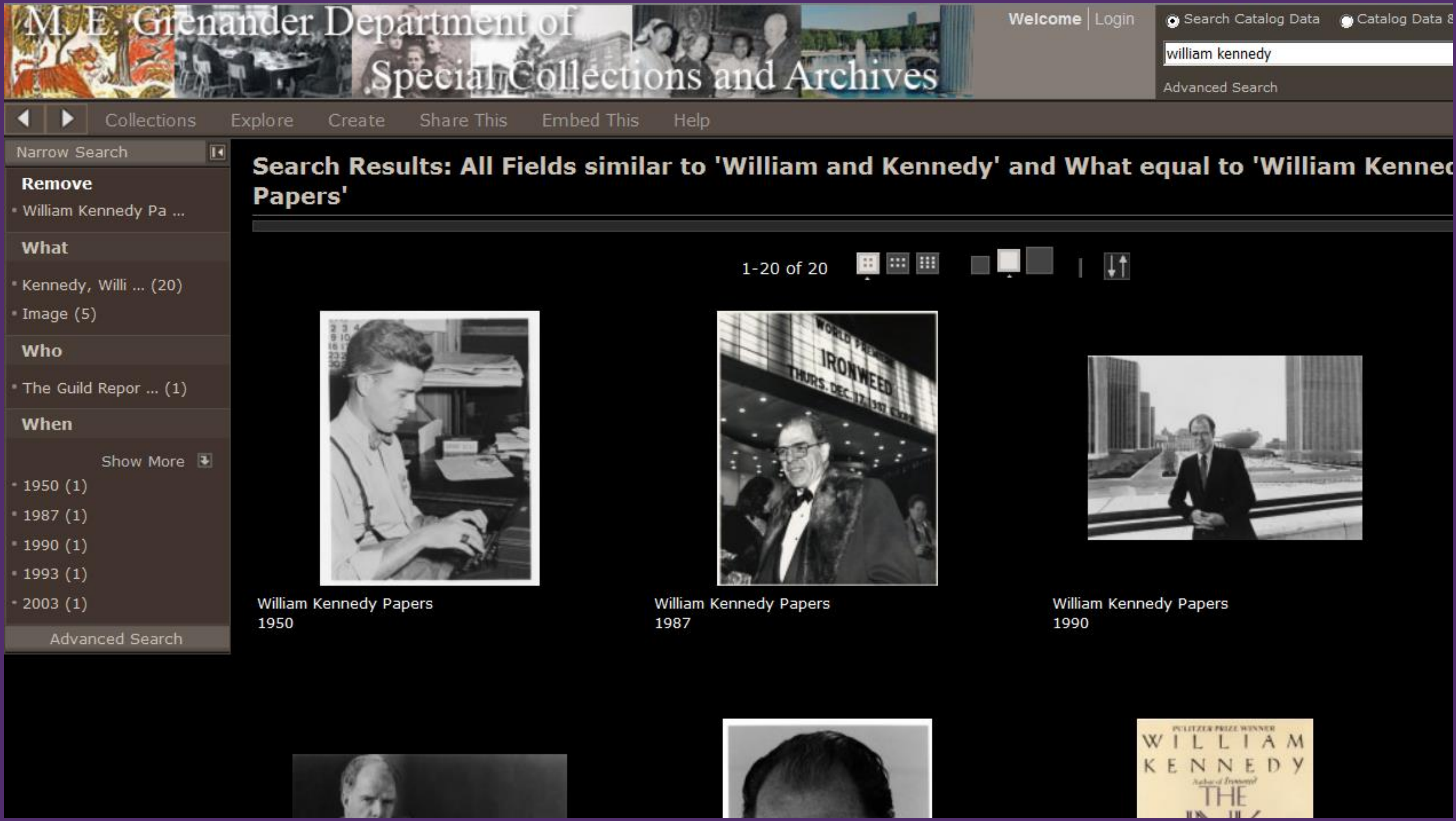
> The most costly part of any traditional digitization project is <u>metadata creation</u>

- We don’t have resources to add metadata sustainably

UNIVERSITY AT ALBANY
State University of New York

Gregory Wiedeman
University Archivist
GWiedeman@albany.edu
@GregWiedeman

We Need Descriptive Metadata for Discovery



Search box: william kennedy

Search Results: All Fields similar to 'William and Kennedy' and What equal to 'William Kenned Papers'

1-20 of 20

Remove
- William Kennedy Pa ...

What
- Kennedy, Willi ... (20)
- Image (5)

Who
- The Guild Repor ... (1)

When
- 1950 (1)
- 1987 (1)
- 1990 (1)
- 1993 (1)
- 2003 (1)

William Kennedy Papers
1950

William Kennedy Papers
1987

William Kennedy Papers
1990

UNIVERSITY AT ALBANY
State University of New York

Gregory Wiedeman
University Archivist
GWiedeman@albany.edu
@GregWiedeman

Archives Principles are Designed for ~~Terrible~~ Minimal Metadata

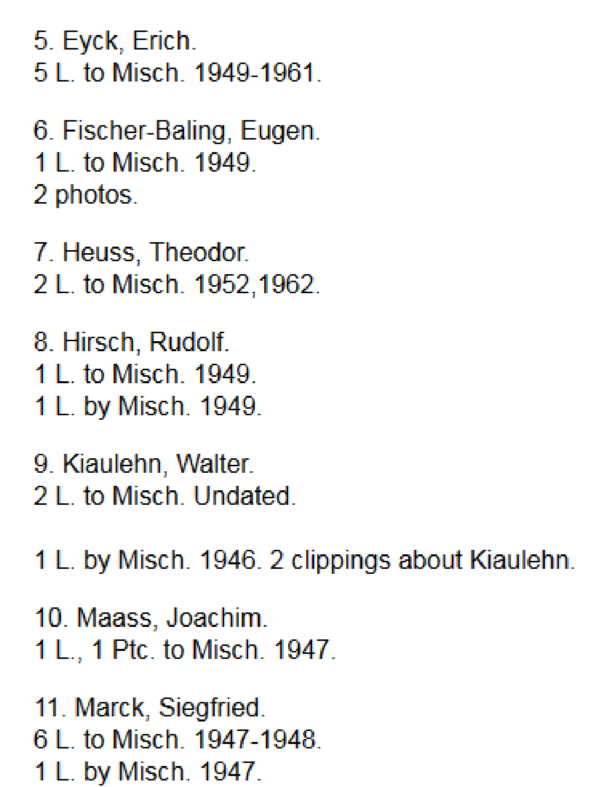
- Hierarchy
  - describe things once
  - describe by grouping, top-down
- Original Order
  - context aids discovery

UNIVERSITY AT ALBANY
State University of New York

Gregory Wiedeman
University Archivist
GWiedeman@albany.edu
🐦@GregWiedeman

Archival Collections Already Have Metadata!

(But it's terrible)



5. Eyck, Erich.
5 L. to Misch. 1949-1961.

6. Fischer-Baling, Eugen.
1 L. to Misch. 1949.
2 photos.

7. Heuss, Theodor.
2 L. to Misch. 1952,1962.

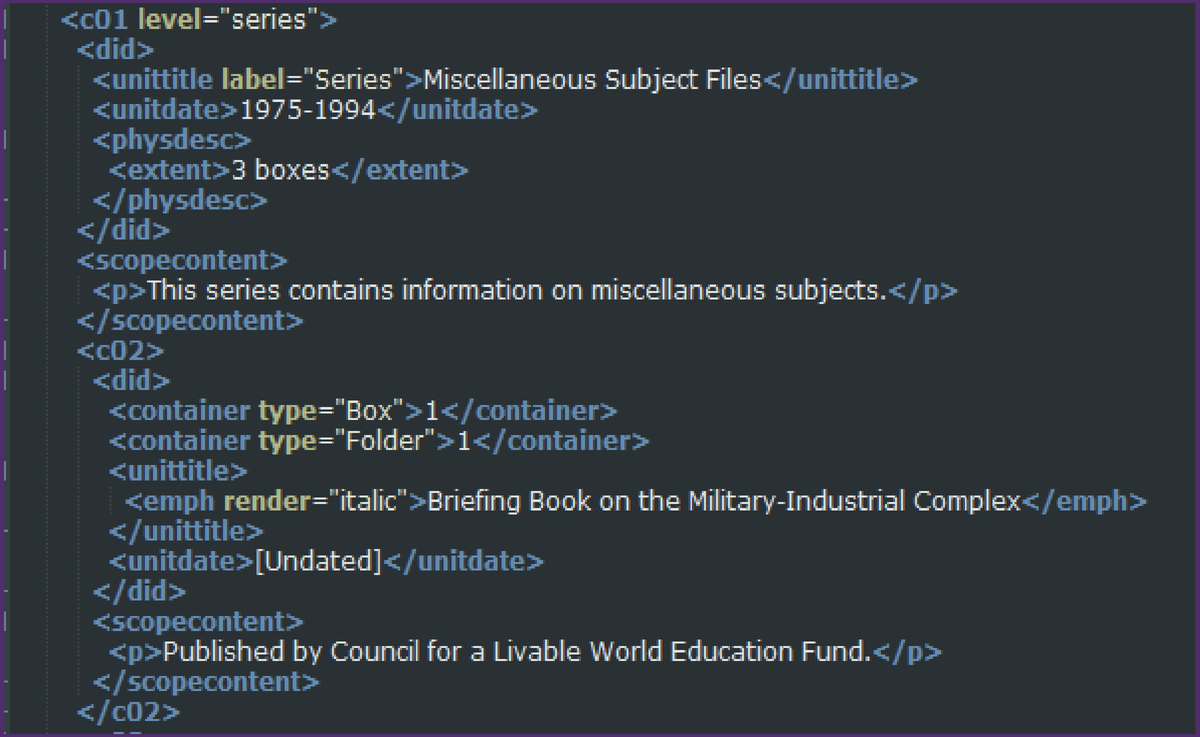
8. Hirsch, Rudolf.
1 L. to Misch. 1949.
1 L. by Misch. 1949.

9. Kiaulehn, Walter.
2 L. to Misch. Undated.

1 L. by Misch. 1946. 2 clippings about Kiaulehn.

10. Maass, Joachim.
1 L., 1 Ptc. to Misch. 1947.

11. Marck, Siegfried.
6 L. to Misch. 1947-1948.
1 L. by Misch. 1947.



```
<c01 level="series">
  <did>
    <unittitle label="Series">Miscellaneous Subject Files</unittitle>
    <unitdate>1975-1994</unitdate>
    <physdesc>
      <extent>3 boxes</extent>
    </physdesc>
  </did>
  <scopecontent>
    <p>This series contains information on miscellaneous subjects.</p>
  </scopecontent>
  <c02>
    <did>
      <container type="Box">1</container>
      <container type="Folder">1</container>
      <unittitle>
        <emph render="italic">Briefing Book on the Military-Industrial Complex</emph>
      </unittitle>
      <unitdate>[Undated]</unitdate>
    </did>
    <scopecontent>
      <p>Published by Council for a Livable World Education Fund.</p>
    </scopecontent>
  </c02>
```

UNIVERSITY AT ALBANY
State University of New York

Gregory Wiedeman
University Archivist
GWiedeman@albany.edu
@GregWiedeman

Archival Metadata

- Uncontrolled at lower levels
- Messy history of finding aids
- Legacy data (yuck)
  - doesn’t meet current standards
- Technical Barriers
  - may not be machine-readable
  - may not be easily discoverable at low levels

Gregory Wiedeman
University Archivist
GWiedeman@albany.edu
@GregWiedeman

Getting Archival Metadata in Shape for Automation

- STRICT Format Controls
- Hierarchical relationships must be machine-readable
- Each archival object at every level must have unique identifier
  - Hierarchical and automated
    - `nam_ua150-3.1_155.3`

UNIVERSITY AT ALBANY
State University of New York

Gregory Wiedeman
University Archivist
GWiedeman@albany.edu
@GregWiedeman

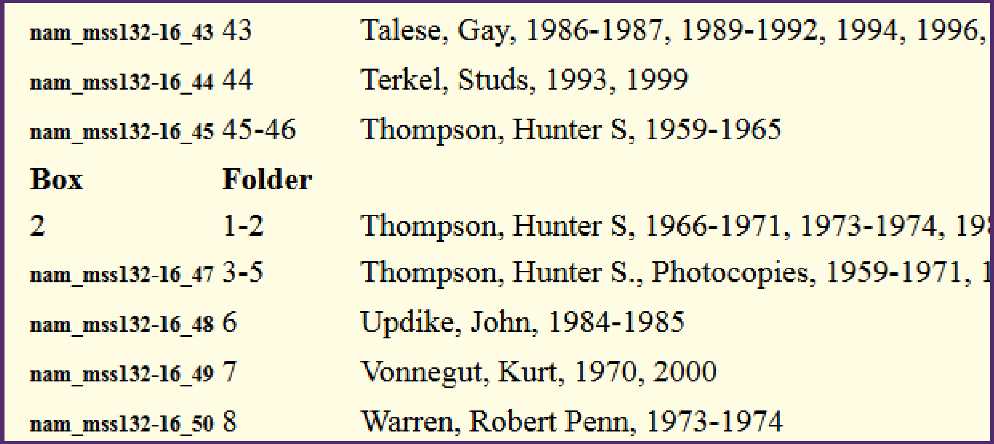
EADValidator

- Python script packaged as .EXE
- Produces HTML report
- Line by line rule-based validation
  - 300+ Detailed Rules:
    - 183 at collection-level
    - 34 at series-level
    - 47 at file-level
    - 25 at item-level
    - 12 for each @normal date
- Not all data is standardized
- Documented set of elements that can be automated

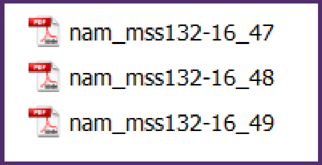
UNIVERSITY AT ALBANY
State University of New York

Gregory Wiedeman
University Archivist
GWiedeman@albany.edu
@GregWiedeman

AutoUpload.py



nam_mss132-16_43	43	Talese, Gay, 1986-1987, 1989-1992, 1994, 1996,
nam_mss132-16_44	44	Terkel, Studs, 1993, 1999
nam_mss132-16_45	45-46	Thompson, Hunter S, 1959-1965
Box	**Folder**	
2	1-2	Thompson, Hunter S, 1966-1971, 1973-1974, 19
nam_mss132-16_47	3-5	Thompson, Hunter S., Photocopies, 1959-1971, 1
nam_mss132-16_48	6	Updike, John, 1984-1985
nam_mss132-16_49	7	Vonnegut, Kurt, 1970, 2000
nam_mss132-16_50	8	Warren, Robert Penn, 1973-1974



nam_mss132-16_47
nam_mss132-16_48
nam_mss132-16_49

- ID is entered as filename
- Script runs hourly to check for new files
- Finds matching object record in EAD XML

UNIVERSITY AT ALBANY
State University of New York

Gregory Wiedeman
University Archivist
GWiedeman@albany.edu
@GregWiedeman

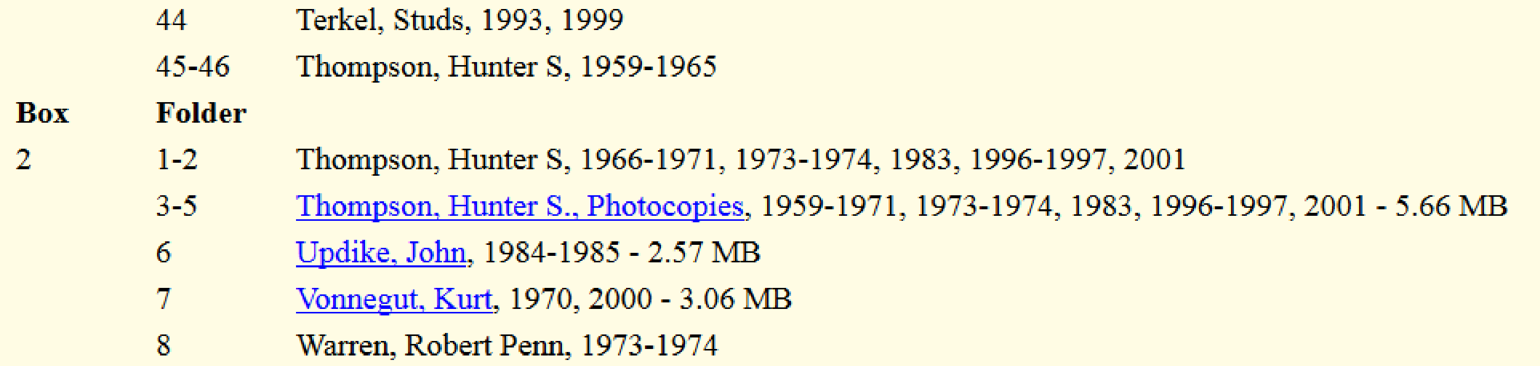
AutoUpload.py

- Manages digital object
  - Uses Bag-it to make preservation copy
  - For preservation TIFFs uses ImageMagik to make PDF access files
  - Moves access copy web server

Gregory Wiedeman
University Archivist
GWiedeman@albany.edu
@GregWiedeman

AutoUpload.py

- Edits metadata record
  - Updates running XML log of all actions
  - Stores copy of original EAD XML
  - Enters digital object record in EAD
  - Transforms to EAD to live HTML



	44	Terkel, Studs, 1993, 1999
	45-46	Thompson, Hunter S, 1959-1965
Box	**Folder**	
2	1-2	Thompson, Hunter S, 1966-1971, 1973-1974, 1983, 1996-1997, 2001
	3-5	Thompson, Hunter S., Photocopies, 1959-1971, 1973-1974, 1983, 1996-1997, 2001 - 5.66 MB
	6	Updike, John, 1984-1985 - 2.57 MB
	7	Vonnegut, Kurt, 1970, 2000 - 3.06 MB
	8	Warren, Robert Penn, 1973-1974

UNIVERSITY AT ALBANY
State University of New York

Gregory Wiedeman
University Archivist
GWiedeman@albany.edu
@GregWiedeman

Mass Digitization on Demand

- Selection based on actual use
- Benefits of making our body of materials more accessible as a whole
- Making our collections more valuable but giving them a wider reach

UNIVERSITY AT ALBANY
State University of New York

Gregory Wiedeman
University Archivist
GWiedeman@albany.edu
@GregWiedeman