COLUMBIA UNIVERSITY LIBRARIES

CUL ArchivesSpace Migration Project

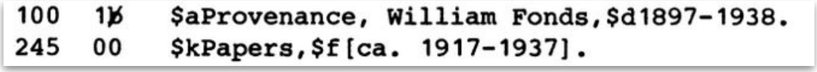
David Hodges and Kevin Schlottmann

Project phases:

1. Migrate collection-level data (~8,000 collection and accession records), and integrate with Voyager ILS (2017–18)
2. Remediate and migrate ~1,400 EAD finding aids (2019)
3. Implement public-facing interface for finding aids, and integrate with other discovery services (2019–20)

archivists’ toolkit

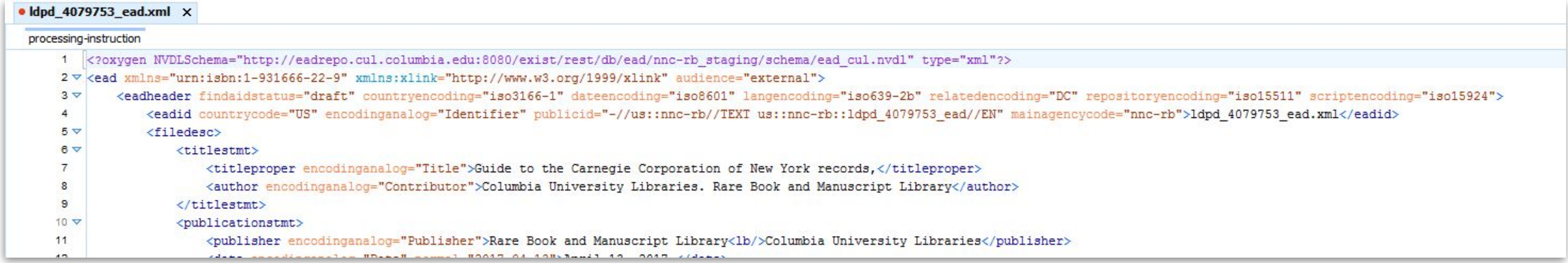
for archivists by archivists



```
100    1b̸    $aProvenance, William Fonds,$d1897-1938.
245    00    $kPapers,$f[ca. 1917-1937].
```

Data and workflow complexities:

- Several archival repositories (RBML/Univ. Archives, Avery, Burke, Starr) with divergent tools/workflows
- ~4K MARC records
- ~4K accessions in Archivists’ Toolkit
- ~1,400 legacy EAD finding aids
- Bespoke finding aids in PDF or HTML
- Linked spreadsheets and other binary formats



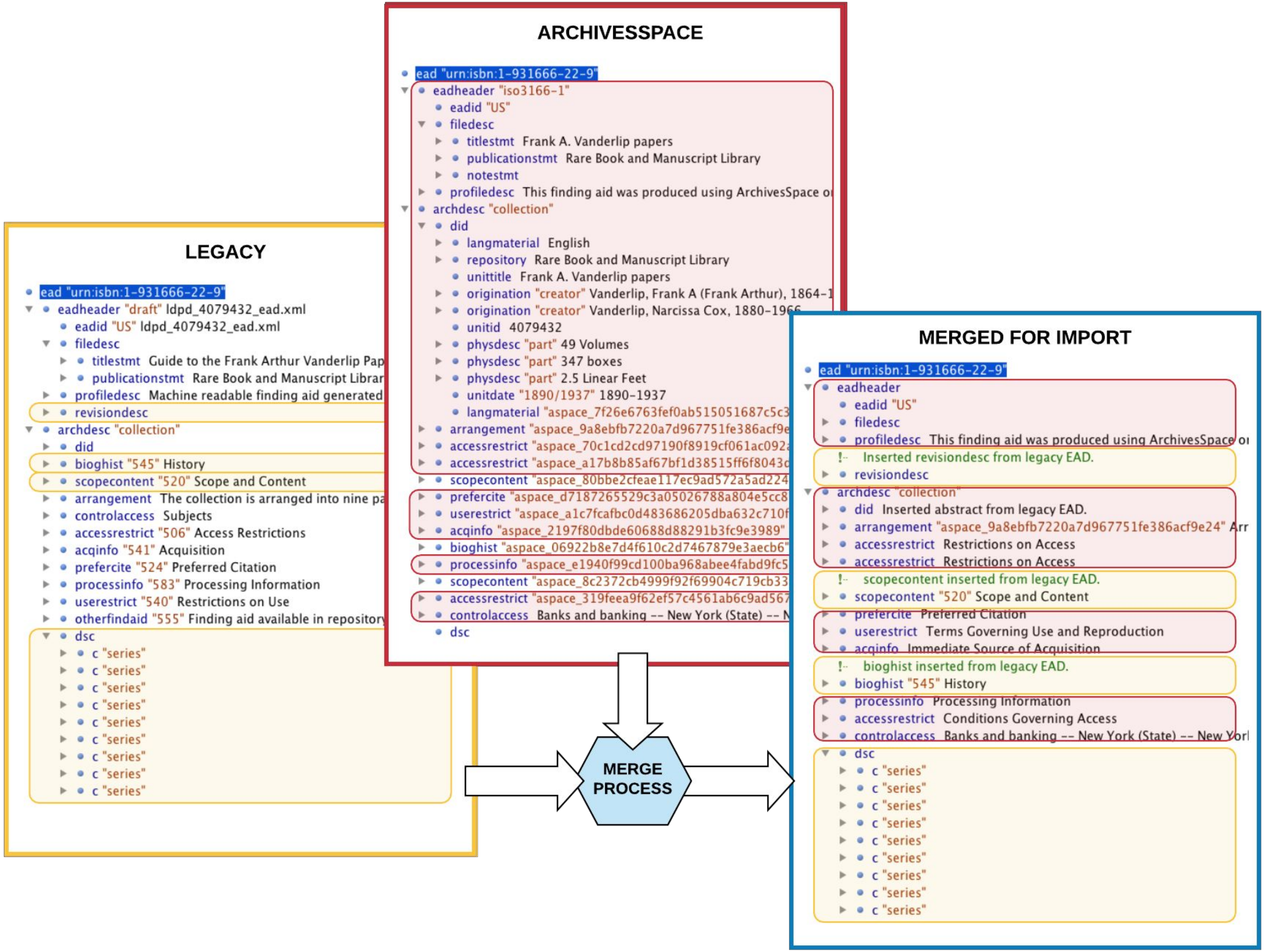
```
<?oxygen NVDLSchema="http://eadrepo.cul.columbia.edu:8080/exist/rest/db/ead/nnc-rb_staging/schema/ead_cul.nvdl" type="xml"?>
<ead xmlns="urn:isbn:1-931666-22-9" xmlns:xlink="http://www.w3.org/1999/xlink" audience="external">
    <eadheader findaidstatus="draft" countryencoding="iso3166-1" dateencoding="iso8601" langencoding="iso639-2b" relatedencoding="DC" repositoryencoding="iso15511" scriptencoding="iso15924">
        <eadid countrycode="US" encodinganalog="Identifier" publicid="-//us::nnc-rb//TEXT us::nnc-rb::ldpd_4079753_ead//EN" mainagencycode="nnc-rb">ldpd_4079753_ead.xml</eadid>
        <filedesc>
            <titlestmt>
                <titleproper encodinganalog="Title">Guide to the Carnegie Corporation of New York records,</titleproper>
                <author encodinganalog="Contributor">Columbia University Libraries. Rare Book and Manuscript Library</author>
            </titlestmt>
            <publicationstmt>
                <publisher encodinganalog="Publisher">Rare Book and Manuscript Library<lb/>Columbia University Libraries</publisher>
                <date encodinganalog="Date" normal="2017-04-13">April 13, 2017 </date>
```

Motivations:

- Single-source of truth for description
- Improved creation and editing of archival description
- Unify practices and workflows across CU libraries
- Modernize infrastructure
- Improve discoverability and UX

Phase 2 EAD Migration

- Combine container structure (<dsc>) from finding aids with collection-level data already in AS and MARC records.
- Some collection-level description was better in legacy finding aids, but varied from record to record.
- How to determine the source of truth for conflicting data? Review process could be extremely time consuming.



LEGACY

```
ead "urn:isbn:1-931666-22-9"
eadheader "draft" ldpd_4079432_ead.xml
  eadid "US" ldpd_4079432_ead.xml
  filedesc
    titlestmt  Guide to the Frank Arthur Vanderlip Pap
    publicationstmt  Rare Book and Manuscript Librar
  profiledesc  Machine readable finding aid generated
  revisiondesc
archdesc "collection"
  did
  bioghist "545" History
  scopecontent "520" Scope and Content
  arrangement  The collection is arranged into nine pa
  controlaccess  Subjects
  accessrestrict "506" Access Restrictions
  acqinfo "541" Acquisition
  prefercite "524" Preferred Citation
  processinfo "583" Processing Information
  userestrict "540" Restrictions on Use
  otherfindaid "555" Finding aid available in repository
  dsc
    c "series"
    c "series"
    c "series"
    c "series"
    c "series"
    c "series"
    c "series"
    c "series"
    c "series"
```

ARCHIVESSPACE

```
ead "urn:isbn:1-931666-22-9"
eadheader "iso3166-1"
  eadid "US"
  filedesc
    titlestmt  Frank A. Vanderlip papers
    publicationstmt  Rare Book and Manuscript Library
    notestmt
  profiledesc  This finding aid was produced using ArchivesSpace oı
archdesc "collection"
  did
    langmaterial  English
    repository  Rare Book and Manuscript Library
    unittitle  Frank A. Vanderlip papers
    origination "creator" Vanderlip, Frank A (Frank Arthur), 1864-1
    origination "creator" Vanderlip, Narcissa Cox, 1880-1966
    unitid  4079432
    physdesc "part" 49 Volumes
    physdesc "part" 347 boxes
    physdesc "part" 2.5 Linear Feet
    unitdate "1890/1937" 1890-1937
    langmaterial "aspace_7f26e6763fef0ab515051687c5c3
  arrangement "aspace_9a8ebfb7220a7d967751fe386acf9e
  accessrestrict "aspace_70c1cd2cd97190f8919cf061ac092a
  accessrestrict "aspace_a17b8b85af67bf1d38515ff6f8043d
  scopecontent "aspace_80bbe2cfeae117ec9ad572a5ad224
  prefercite "aspace_d7187265529c3a05026788a804e5cc8
  userestrict "aspace_a1c7fcafbc0d483686205dba632c710f
  acqinfo "aspace_2197f80dbde60688d88291b3fc9e3989"
  bioghist "aspace_06922b8e7d4f610c2d7467879e3aecb6"
  processinfo "aspace_e1940f99cd100ba968abee4fabd9fc5
  scopecontent "aspace_8c2372cb4999f92f69904c719cb33
  accessrestrict "aspace_319feea9f62ef57c4561ab6c9ad567
  controlaccess  Banks and banking -- New York (State) -- N
  dsc
```

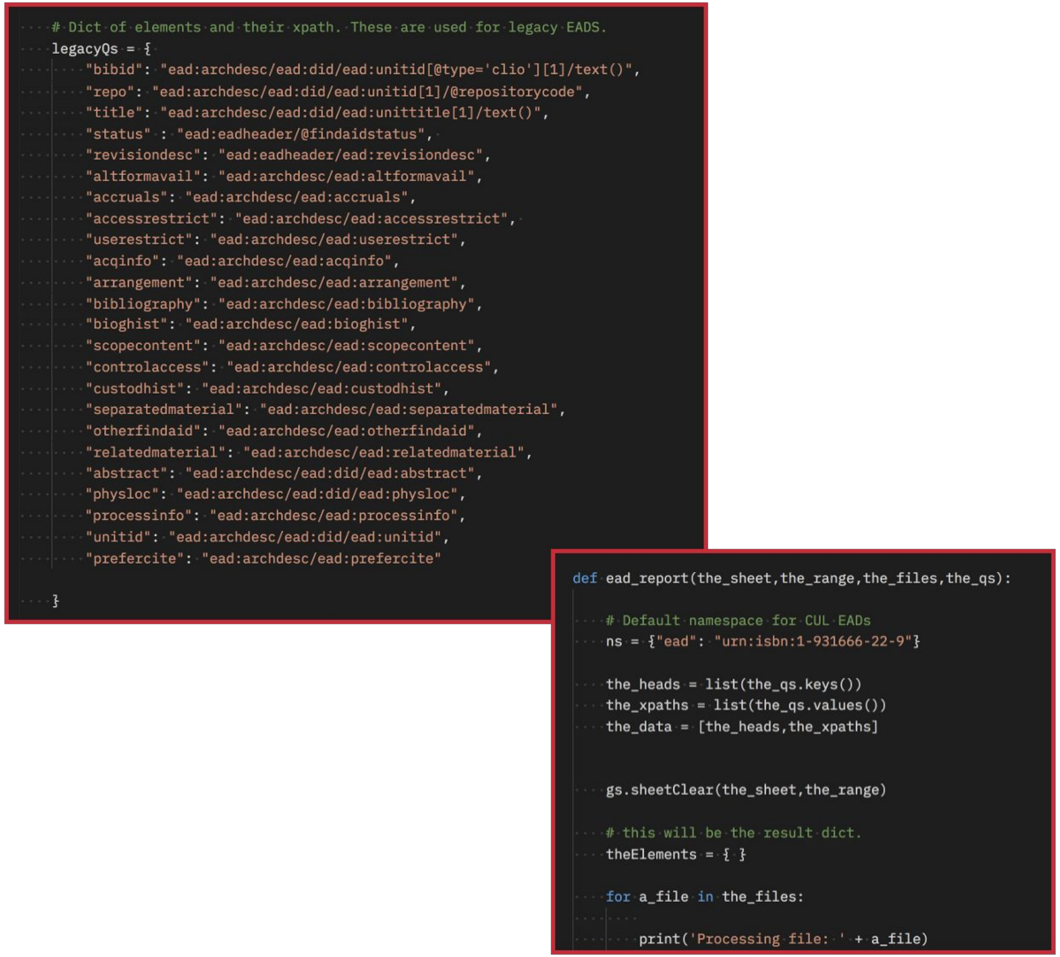
MERGE PROCESS

MERGED FOR IMPORT

```
ead "urn:isbn:1-931666-22-9"
eadheader
  eadid "US"
  filedesc
  profiledesc  This finding aid was produced using ArchivesSpace oı
  !--  Inserted revisiondesc from legacy EAD.
  revisiondesc
archdesc "collection"
  did  Inserted abstract from legacy EAD.
  arrangement "aspace_9a8ebfb7220a7d967751fe386acf9e24" Arr
  accessrestrict  Restrictions on Access
  accessrestrict  Restrictions on Access
  !--  scopecontent inserted from legacy EAD.
  scopecontent "520" Scope and Content
  prefercite  Preferred Citation
  userestrict  Terms Governing Use and Reproduction
  acqinfo  Immediate Source of Acquisition
  !--  bioghist inserted from legacy EAD.
  bioghist "545" History
  processinfo  Processing Information
  accessrestrict  Conditions Governing Access
  controlaccess  Banks and banking -- New York (State) -- New Yorl
  dsc
    c "series"
    c "series"
    c "series"
    c "series"
    c "series"
    c "series"
    c "series"
    c "series"
    c "series"
```

EAD_Process (Python + Google Sheets)

- Python + Google Sheets + XSLT: extract data, flag elements to merge, process all files.
- Two data sources: legacy finding aid and EAD exported from AS processed in pairs.
- Defined list of XPATH queries for each.
- Data is extracted from both data sets and sent to a Google Sheet, which collates into side-by-side view for each of targeted elements.

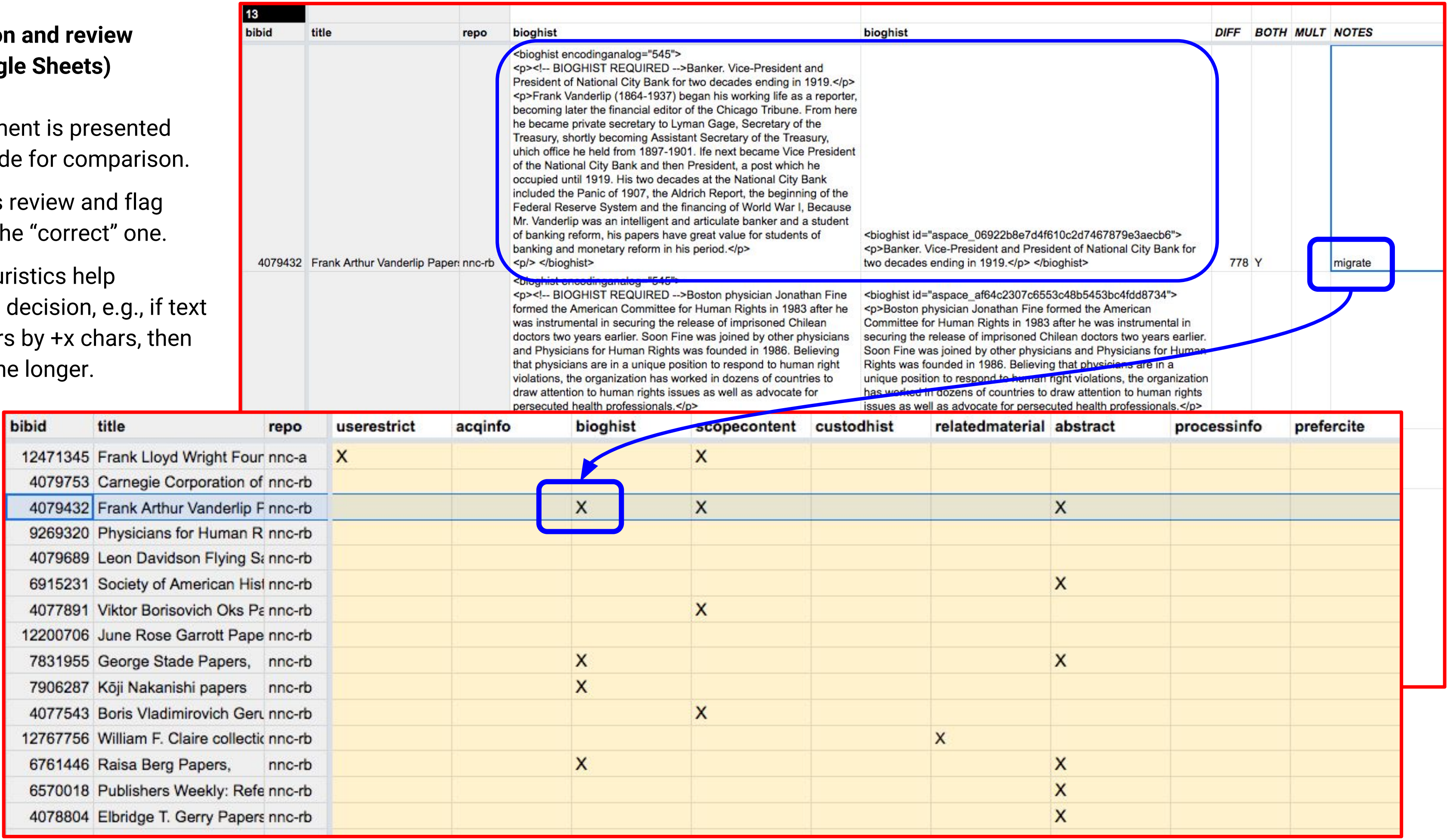


```
    # Dict of elements and their xpath. These are used for legacy EADS.
    legacyQs = {
        "bibid": "ead:archdesc/ead:did/ead:unitid[@type='clio'][1]/text()",
        "repo": "ead:archdesc/ead:did/ead:unitid[1]/@repositorycode",
        "title": "ead:archdesc/ead:did/ead:unittitle[1]/text()",
        "status" : "ead:eadheader/@findaidstatus",
        "revisiondesc": "ead:eadheader/ead:revisiondesc",
        "altformavail": "ead:archdesc/ead:altformavail",
        "accruals": "ead:archdesc/ead:accruals",
        "accessrestrict": "ead:archdesc/ead:accessrestrict",
        "userestrict": "ead:archdesc/ead:userestrict",
        "acqinfo": "ead:archdesc/ead:acqinfo",
        "arrangement": "ead:archdesc/ead:arrangement",
        "bibliography": "ead:archdesc/ead:bibliography",
        "bioghist": "ead:archdesc/ead:bioghist",
        "scopecontent": "ead:archdesc/ead:scopecontent",
        "controlaccess": "ead:archdesc/ead:controlaccess",
        "custodhist": "ead:archdesc/ead:custodhist",
        "separatedmaterial": "ead:archdesc/ead:separatedmaterial",
        "otherfindaid": "ead:archdesc/ead:otherfindaid",
        "relatedmaterial": "ead:archdesc/ead:relatedmaterial",
        "abstract": "ead:archdesc/ead:did/ead:abstract",
        "physloc": "ead:archdesc/ead:did/ead:physloc",
        "processinfo": "ead:archdesc/ead:processinfo",
        "unitid": "ead:archdesc/ead:did/ead:unitid",
        "prefercite": "ead:archdesc/ead:prefercite"

    }
```

```
def ead_report(the_sheet,the_range,the_files,the_qs):

    # Default namespace for CUL EADs
    ns = {"ead": "urn:isbn:1-931666-22-9"}

    the_heads = list(the_qs.keys())
    the_xpaths = list(the_qs.values())
    the_data = [the_heads,the_xpaths]


    gs.sheetClear(the_sheet,the_range)

    # this will be the result dict.
    theElements = { }

    for a_file in the_files:

        print('Processing file: ' + a_file)
```

EAD comparison and review workflow (Google Sheets)

- Each element is presented side by side for comparison.
- Archivists review and flag which is the “correct” one.
- Some heuristics help automate decision, e.g., if text size differs by +x chars, then migrate the longer.

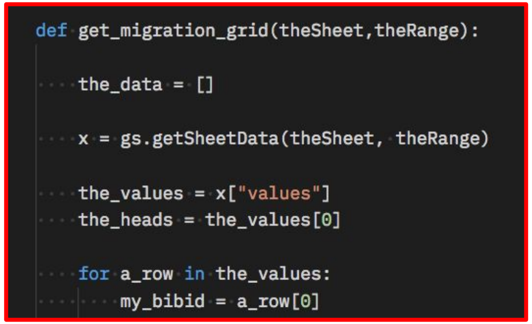


13								
bibid	title	repo	bioghist	bioghist	DIFF	BOTH	MULT	NOTES
4079432	Frank Arthur Vanderlip Paper:	nnc-rb	<bioghist encodinganalog="545"> <p><!-- BIOGHIST REQUIRED -->Banker. Vice-President and President of National City Bank for two decades ending in 1919.</p> <p>Frank Vanderlip (1864-1937) began his working life as a reporter, becoming later the financial editor of the Chicago Tribune. From here he became private secretary to Lyman Gage, Secretary of the Treasury, shortly becoming Assistant Secretary of the Treasury, uhich office he held from 1897-1901. Ife next became Vice President of the National City Bank and then President, a post which he occupied until 1919. His two decades at the National City Bank included the Panic of 1907, the Aldrich Report, the beginning of the Federal Reserve System and the financing of World War I, Because Mr. Vanderlip was an intelligent and articulate banker and a student of banking reform, his papers have great value for students of banking and monetary reform in his period.</p> <p/> </bioghist>	<bioghist id="aspace_06922b8e7d4f610c2d7467879e3aecb6"> <p>Banker. Vice-President and President of National City Bank for two decades ending in 1919.</p> </bioghist>	778	Y		migrate
			<bioghist encodinganalog="545"> <p><!-- BIOGHIST REQUIRED -->Boston physician Jonathan Fine formed the American Committee for Human Rights in 1983 after he was instrumental in securing the release of imprisoned Chilean doctors two years earlier. Soon Fine was joined by other physicians and Physicians for Human Rights was founded in 1986. Believing that physicians are in a unique position to respond to human right violations, the organization has worked in dozens of countries to draw attention to human rights issues as well as advocate for persecuted health professionals.</p>	<bioghist id="aspace_af64c2307c6553c48b5453bc4fdd8734"> <p>Boston physician Jonathan Fine formed the American Committee for Human Rights in 1983 after he was instrumental in securing the release of imprisoned Chilean doctors two years earlier. Soon Fine was joined by other physicians and Physicians for Human Rights was founded in 1986. Believing that physicians are in a unique position to respond to human right violations, the organization has worked in dozens of countries to draw attention to human rights issues as well as advocate for persecuted health professionals.</p>				

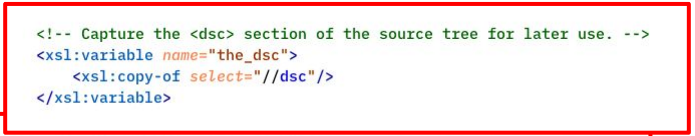
bibid	title	repo	userestrict	acqinfo	bioghist	scopecontent	custodhist	relatedmaterial	abstract	processinfo	prefercite
12471345	Frank Lloyd Wright Four	nnc-a	X			X					
4079753	Carnegie Corporation of	nnc-rb									
4079432	Frank Arthur Vanderlip F	nnc-rb			X	X			X		
9269320	Physicians for Human R	nnc-rb									
4079689	Leon Davidson Flying Sa	nnc-rb									
6915231	Society of American Hist	nnc-rb							X		
4077891	Viktor Borisovich Oks Pa	nnc-rb				X					
12200706	June Rose Garrott Pape	nnc-rb									
7831955	George Stade Papers,	nnc-rb			X				X		
7906287	Kōji Nakanishi papers	nnc-rb			X						
4077543	Boris Vladimirovich Geru	nnc-rb				X					
12767756	William F. Claire collectic	nnc-rb						X			
6761446	Raisa Berg Papers,	nnc-rb			X				X		
6570018	Publishers Weekly: Refe	nnc-rb							X		
4078804	Elbridge T. Gerry Papers	nnc-rb							X		

EAD_Merge (Python + Google Sheets + XSLT)

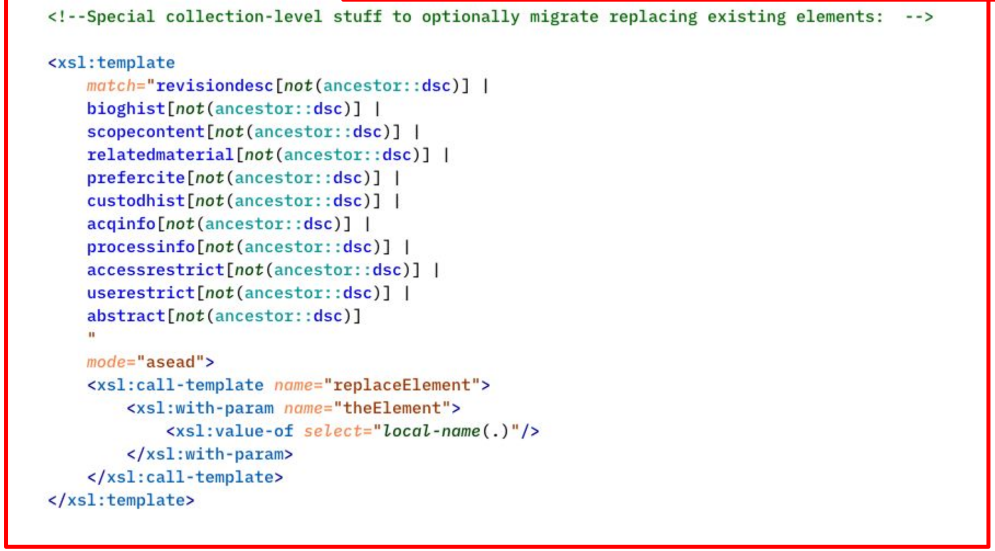
- After archivists have completed review, a Python script reads Google Sheet migration grid rows as lists ([['4079432', 'bioghist', 'scopecontent', 'abstract'],...]).
- Legacy EAD file is sent to XSLT with parameters of which additional elements to migrate (other than <dsc>).
- XSLT merges two EAD trees, incorporating <dsc> plus selected additional elements per the parameters.
- Another XSLT stylesheet in pipeline performs other global cleanup functions.
- Output is validated and QC'd before importing into AS.



```
def get_migration_grid(theSheet,theRange):

    the_data = []

    x = gs.getSheetData(theSheet, theRange)

    the_values = x["values"]
    the_heads = the_values[0]

    for a_row in the_values:
        my_bibid = a_row[0]
```



```
<!-- Capture the <dsc> section of the source tree for later use. -->
<xsl:variable name="the_dsc">
    <xsl:copy-of select="//dsc"/>
</xsl:variable>
```



```
<!--Special collection-level stuff to optionally migrate replacing existing elements:  -->

<xsl:template
    match="revisiondesc[not(ancestor::dsc)] |
    bioghist[not(ancestor::dsc)] |
    scopecontent[not(ancestor::dsc)] |
    relatedmaterial[not(ancestor::dsc)] |
    prefercite[not(ancestor::dsc)] |
    custodhist[not(ancestor::dsc)] |
    acqinfo[not(ancestor::dsc)] |
    processinfo[not(ancestor::dsc)] |
    accessrestrict[not(ancestor::dsc)] |
    userestrict[not(ancestor::dsc)] |
    abstract[not(ancestor::dsc)]
    "
    mode="asead">
    <xsl:call-template name="replaceElement">
        <xsl:with-param name="theElement">
            <xsl:value-of select="local-name(.)"/>
        </xsl:with-param>
    </xsl:call-template>
</xsl:template>
```

Thank you!

⇔

Code repo

https://github.com/cul/rbml-archivesspace

Contact info

- @archivistkevin // kws2126@columbia.edu
- dwh2128@columbia.edu