# Metadata Quality Analysis

### Tools & Scripts to Check Your Data

Christina Harlow, UT Knoxville ShareFest 2015, Nashville Tennessee

# Slides & Examples

http://bit.ly/MetadataQA

# Introduction

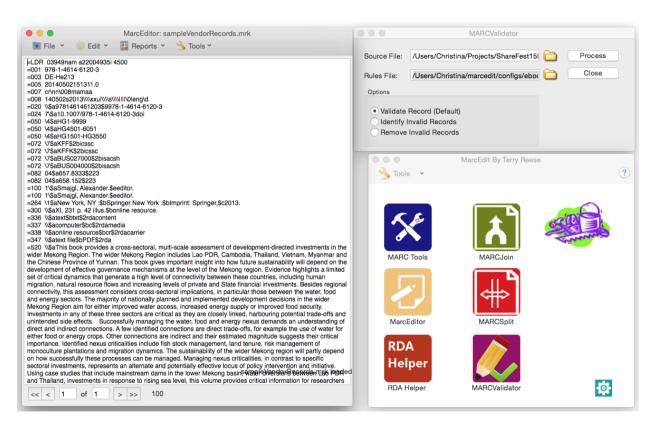
- Need for metadata QC review tools
- Use cases => Examples
- GUIs to Scripting
- Working with:
  - ∘ MARC, XML, Json
  - ∘ DC, MODS, DPLA MAP v.4
  - ∘ RDA, AACR2, DACS

# Method 1: MarcEdit

MarcEdit = Freely available (but not open source) tool for working with MARC records.

http://marcedit.reeset.net/

### MarcEdit Screenshot



### MarcEdit

- Easy to learn
- Work with .mrc, .mrk, or MARC/XML
- Standard conversions
- Validates MARC, RDA Helper
- Generates reports
- Performance issues for bigger (>50k) sets
- Not always/easily customizable

# MarcEdit Reports at UTK

Review Vendor Supplied, Legacy Records:

- Field Usage, Item Type Reports
- Generating CSV files
- MARC Validation Report

# MarcEdit Reports at UTK

### Field Usage Report

examples >

MARCEdit >

FieldandTypeReports.md

Field	Subfield	In Records	Total
650		46416	156927
	\$v		19946
	\$d		16
	\$0		1338
	\$c		2
	\$a		156926
	\$2		1415
	\$1		2
	\$z		41253
	<b>\$</b> y		5781
	\$x		34059
651		11221	17323
	\$v		2159
	\$0		330
	\$a		17323
	\$2		337
	\$6		9
	\$z		487
	<b>\$</b> y		4773
	\$x		12665
653		310	509
	\$a		529
655		33586	80463

### MarcEdit Reports at UTK

### Item Type Report

```
Material Type Report Results
Visual Materials: 50876
     10 meter & 30 meter coverage of 48 conterminous states: 1
     2007, 2008, 2009: 1
     Area size: 3837 sq. miles, coverage of Nashville, Memphis, Knoxville & Chattanooga, color 1 foot
resolution, Tennessee State-Plane NAD83 MrSID format: 1
     Bildtonträger: 1
     Chart: 5
     Direct Electronic: 12
     DVD: 6
     Electronic: 10643
     Electronic resource: 752
     Elektronische Ressource: 1
     Filmstrip: 9
     Flash card: 3
     Graphic: 3
     Kit: 22
```

examples > MARCEdit > FieldandTypeReports.md

# MarcEdit Reports at UTK

# Generated CSV files

examples >

MARCEdit >

GenerateCSVreports.md

001	007\$00	245\$a	245\$h	336	337	338	500	538
991262280102311	v	Equine laryngostomy and ventriculectomy, and emergency tracheostomy.	[videorecording]					U-matic.
994938970102311	v	Day for night	[videorecording]				"English language version."; Original screenplay by François Truffaut.	
994403500102311	v	Alexander spots	[videorecording] /					U-matic.
994478180102311	v	Elvis in the fifties	[videorecording]					
993287420102311	g	[Kefauver miscellany.	[videorecording]				Reproduction from motion picture.;A compilation of various motion picture shorts.	DVD
992088700102311	v	The shirt off your back	[videorecording].				U Standard.; Orginally broadcast as television program by WGBH.;Title on container: Inside Europe; the shirt off our backs.	

### MarcEdit Reports at UTK

#### MARC Validation Report

```
Record #: 1
001 (if defined): 978-1-4614-6120-3
245 (if defined): The Water-Food-Energy Nexus in the Mekong Region$h[electronic resource]:
$bAssessing Development Strategies Considering Cross-Sectoral and Transboundary Impacts /$cedited
by Alexander Smajgl, John Ward.
Errors:
One 245 field must be present in a MARC21 record.
245: Has been marked as a non-repeating field.

Record #: 16
001 (if defined): 978-3-642-41704-7
245 (if defined): Microfinance 3.0$h[electronic resource]: $bReconciling Sustainability with Social
Outreach and Responsible Delivery /$cedited by Doris K¶hn.
Errors:
506-ind1: Invalid data (0) Indicator should be undefined.
```

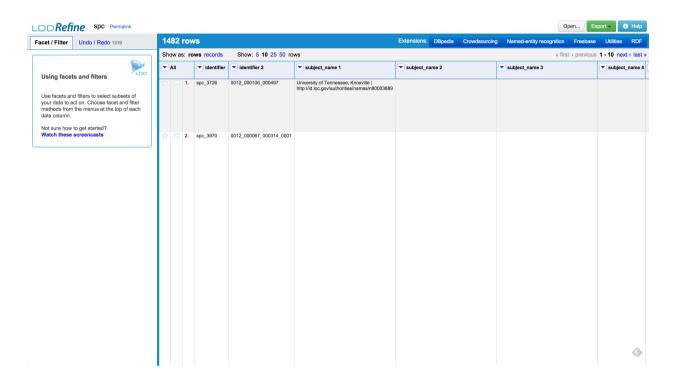
examples > MARCEdit > MARCvalidation.md

# Method 2: OpenRefine

OpenRefine = Freely available, open source tool for working with all kinds of data.

http://openrefine.org/

# OpenRefine Screenshot



- OpenRefine . Helpful Graphical User Interface for understanding data
  - Works with many types of data (better with flatter)
  - Faceting, Clustering, GREL, Reconciliation
  - Exports a number of formats/encodings
  - Performance issues for bigger (>150k) sets

# OpenRefine Reports at UTK, DLTN

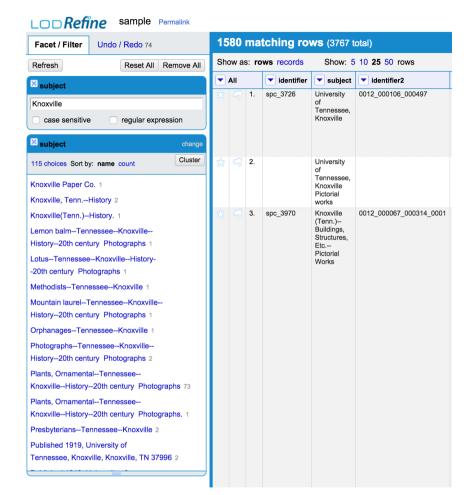
Review non-MARC Metadata for Mapping, Normalization, Enhancement:

- Facet & Cluster Fields
- Target, Export Geographic Terms
- Review Fields Used

# OpenRefine Reports at UTK, DLTN

Facet & Cluster Fields

examples >
OpenRefine >
SortFacetFilters.md



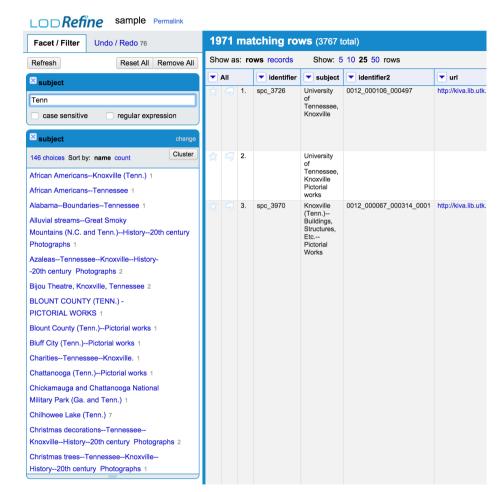
# OpenRefine Reports at UTK, DLTN

### Target Geographic Terms

examples >

OpenRefine >

SortFacetFilters.md



### OpenRefine Reports at UTK, DLTN

#### Export Geographic Terms

identifier	subject	identifier2	url	item type	internet media type	date_created
spc_3726	University of Tennessee, Knoxville	0012_000106_000497	http://kiva.lib.utk.edu/spc/ items/show/3726	still image	image/jpeg	03/08/15
	University of Tennessee, Knoxville Pictorial works					
spc_3970	Knoxville (Tenn.)Buildings, Structures, EtcPictorial Works	0012_000067_000314_0001	http://kiva.lib.utk.edu/spc/ items/show/3970	still image	image/jpeg	1916 - 1922
	Knoxville (Tenn.)Industries					
spc_3856	TennesseePictorial works.	0012_000067_000200_0001	http://kiva.lib.utk.edu/spc/ items/show/3856	still image	image/jpeg	1955
	Blount County (Tenn.)Pictorial works					
spc_3644	University of Tennessee, Knoxville	0012_000106_000415	http://kiva.lib.utk.edu/spc/ items/show/3644	still image	image/jpeg	194-
	University of Tennessee, Knoxville Pictorial works					
	Knoxville (Tenn.)Buildings, Structures, EtcPictorial Works					
	Knox County (Tenn.)Industries					

# Method 3: Python

Python = a popular, high-level programming language used for a variety of purposes.

PyMARC is a Python library helpful for working with MARC data.

python wiki: https://wiki.python.org/moin/ pymarc: https://github.com/edsu/pymarc

# Python Scripting

- Highly customizable for special projects' data
- Works with all types of data through use of libraries
- Runs through larger datasets more quickly
- Requires programming knowledge
- Requires often building scripts locally

# Standing on Shoulders of Giants...

Python OAI parsing scripts from Mark Phillips at the University of North Texas

http://github.com/vphill/metadata\_breakers https://github.com/vphill/pyoaiharvester

- UNT-based OAI scripts
- DPLA Json objects review

#### OAI-DC script excerpt

```
def main():
    stats_aggregate = {
        "record count": 0.
        "field info": {}
    element stats aggregate = {}
    parser = ArgumentParser(usage='%(prog)s [options] data_filename.xml')
    parser.add_argument("-e", "--element", dest="element",
                        help="element to print to screen")
    parser.add_argument("-i", "--id", action="store_true", dest="id",
                        default=False, help="prepend meta_id to line")
    parser.add argument("-s", "--stats", action="store true", dest="stats",
                        default=False, help="only print stats for repository")
    parser.add_argument("-p", "--present", action="store_true", dest="present",
                        default=False, help="print if there is value of defined element in record")
    parser.add_argument("-d", "--dump", action="store_true", dest="dump",
                        default=False, help="Dump all record data to a tab delimited format")
    parser.add_argument("datafile", help="put the datafile you want analyzed here")
    args = parser.parse args()
```

qithub.com/cmh2166/metadataOA

#### OAI-DC script report

```
{http://purl.org/dc/elements/1.1/}contributor:
                                                                                        6528/8101
                                                                                                     80%
            {http://purl.org/dc/elements/1.1/}coverage:
                                                                                        4730/8101 |
                                                                                                     58%
             {http://purl.org/dc/elements/1.1/}creator:
                                                                                       2162/8101
                                                                                                     26%
                {http://purl.org/dc/elements/1.1/}date:
                                                                                       6517/8101
                                                                                                     80%
         {http://purl.org/dc/elements/1.1/}description:
                                                                                        6272/8101 |
                                                                                                     77%
                                                                                                     95%
              {http://purl.org/dc/elements/1.1/}format:
                                                                                        7721/8101
          {http://purl.org/dc/elements/1.1/}identifier:
                                                                                        8095/8101
                                                                                                     99%
{http://purl.org/dc/elements/1.1/}identifier.thumbnail:
                                                                                                     99%
                                                                                        8086/8101 |
            {http://purl.org/dc/elements/1.1/}language:
                                                                                        6543/8101
                                                                                                     80%
           {http://purl.org/dc/elements/1.1/}publisher:
                                                                                        2628/8101
                                                                                                     32%
            {http://purl.org/dc/elements/1.1/}relation:
                                                                                        7666/8101 I
                                                                                                     94%
              {http://purl.org/dc/elements/1.1/}rights:
                                                                                        8055/8101
                                                                                                     99%
              {http://purl.org/dc/elements/1.1/}source:
                                                                                       2107/8101
                                                                                                     26%
             {http://purl.org/dc/elements/1.1/}subject:
                                                                                       7522/8101 I
                                                                                                     92%
               {http://purl.org/dc/elements/1.1/}title:
                                                                                        8095/8101
                                                                                                     99%
                {http://purl.org/dc/elements/1.1/}type:
                                                                                        8004/8101
                                                                                                     98%
                        {http://www.modstest.org/}test:
                                                                                           1/8101 |
```

```
dc_completeness 82.896762
collection_completeness 77.714943
    wwww_completeness 76.746698
    dpla_completeness 99.761346
average completeness 84.279937
```

github.com/cmh2166/metadataQA

### OAI-MODS script report

mods:abstract:		152698/155164	l 98%
mods:accessCondition:	1	155134/155164	90%   99%
	 	116045/155164	55%   74%
	 	139213/155164	74%   89%
<b>3</b> · · · · · · · · · · · · · · · · · · ·	 	150219/155164	05%   96%
	 	127452/155164	90%   82%
mods:location/mods:holdingExternal/mods:holding/mods:physicalAddress/mods:text:		2090/155164	02%   1%
mods:location/mods:holdingSimple/mods:copyInformation/mods:shelfLocator:	 	569/155164	13   0%
mods: totalion/mods:notalingsimple/mods:copyIntormation/mods:shellCotator:		279/155164	0%
mods:location/mods:physicalLocation: mods:location/mods:physicalLocation:	  -===================================	143166/155164	05   92%
mods:tocation/mods:physicatiocation: mods:location/mods:publisher:		342/155164	92%   0%
mods:location/mods:publisher: mods:location/mods:shelfLocator:		3581/155164	৩%   2%
	  -===================================		23   99%
	=====================================	155155/155164 23037/155164	99%   14%
	===  -==	23031/155164	14%   14%
mods:note:		8054/155164	5%
mods:originInfo/mods:dateCreated:	======================================	146675/155164	94%
mods:originInfo/mods:place:	!	3307/155164	2%
mods:originInfo/mods:publisher:	========	120399/155164	77%
mods:part/mods:detail/mods:number:	!	219/155164	0%
mods:physicalDescription/mods:digitalOrigin:	!	3094/155164	1%
	===	22235/155164	14%
mods:physicalDescription/mods:form:		144448/155164	93%
mods:physicalDescription/mods:internetMediaType:		140894/155164	90%
mods:physicalDescription/mods:note:	!	3243/155164	2%
moust cool alive, moust tanguages to a to tanguage to tan		155164/155164	100%
mods:recordInfo/mods:recordChangeDate:		155164/155164	100%
mods:recordInfo/mods:recordContentSource:		155164/155164	100%
mods:recordInfo/mods:recordCreationDate:	!	2090/155164	1%
mods:recordInfo/mods:recordIdentifier:	!	2154/155164	1%
mods:recordInfo/mods:recordOrigin:		155164/155164	100%
	====	29360/155164	18%
mods:relatedItem/mods:identifier:	!	1267/155164	0%
mods:relatedItem/mods:location/mods:url:		155154/155164	99%
mods:relatedItem/mods:part/mods:detail/mods:number:		1/155164	0%
<pre>mods:relatedItem/mods:titleInfo/mods:title:</pre>		155154/155164	99%
mods:subject/mods:cartographics/mods:coordinates:		127769/155164	82%
mods:subject/mods:geographic:	========	137056/155164	88%
mods:subject/mods:name:	!	712/155164	0%
mods:subject/mods:name/mods:namePart:	!	3118/155164	2%
mods:subject/mods:temporal:		129068/155164	83%
mods:subject/mods:titleInfo/mods:title:		114968/155164	74%
	====	34577/155164	22%
mods:subject/mods:topical:	!	29/155164	0%
mods:titleInfo/mods:nonSort:	!	88/155164	0%
mods:titleInfo/mods:title:		155163/155164	99%
mods:typeOfResource:		152159/155164	98%

#### OAI script field report

```
brighid - NT MmetadataOA Mpython oaidc analysis.py test/output.xml -e format | sort | unig -c
 82 1 digital image
  2 1 digital image: 1 artifact
  2 1 digital image; 1 beehive
  1 1 digital image; 1 black and white photograph
  4 1 digital image; 1 building
  2 1 digital image; 1 cartoon
  1 1 digital image; 1 cartoon with caption; 15 x 15 inches
  1 1 digital image; 1 cartoon; 5.25 x 10 inches
  2 1 digital image; 1 cartoon; 6.5 x 10 inches
  1 1 digital image; 1 cartoon; 6.75 x 10.75 inches
  1 1 digital image; 1 cartoon; 6.75 x 8.5 inches
  1 1 digital image; 1 cartoon; 7 x 10.5 inches
  1 1 digital image; 1 cartoon; 7 x 11 inches
  1 1 digital image; 1 cartoon; 7 x 11.5 inches
  1 1 digital image: 1 cartoon: 7 x 12 inches
  1 1 digital image; 1 cartoon; 7.25 x 10 inches
  1 1 digital image: 1 cartoon: 8.25 x 12 inches
  1 1 digital image; 1 cartoon; 8.5 x 10.75 inches
  1 1 digital image; 1 cartoon; 8.5 x 11 inches
  1 1 digital image; 1 cartoon; 8.75 x 10.25 inches
  2 1 digital image; 1 certificate
  1 1 digital image; 1 cloth piece
  1 1 digital image; 1 color photograph
  1 1 digital image: 1 desk
  1 1 digital image; 1 diagram
```

### OAIMODS script XPath report

```
brighid 🗠 🗑 T 🖫 metadataQA 🖫 python oaimods_analysis.py test/DLTNphase1.mods.xml -x 'mods:originInfo/mods:dateCreated[@encoding="edtf"]' | uniq
1877
uuuu
1877/1878
1876
1843
uuuu
1855
1915
1930
1869
1910
1938
1888
uuuu
1876
1923
1911
1913
1874
1892
1910
1881
1882
uuuu
1886
1837/1838
1840-01
1897
1879
```

github.com/cmh2166/metadataQA

#### DPLA Json objects script excerpt

```
def get elements(self):
    out = []
    try:
        record = objectpath.Tree(self.elem)
        response = record.execute(self.args.element)
        if response != None:
            if isinstance(response, list):
                for item in response:
                    if isinstance(item, list):
                        for item2 in item:
                            if isinstance(item2, list):
                                for item3 in item2:
                                    out.append((', '.join(item3)).encode("utf-8").strip())
                            elif isinstance(item2, dict):
                                for key3, value3 in item2.iteritems():
                                    out.append((', '.join(value3)).encode("utf-8").strip())
                            else:
                                out.append(item2.encode("utf-8").strip())
                    elif isinstance(item, dict):
                        for key, value in item.iteritems():
                            if isinstance(value, list):
                                for item2 in value:
                                    if isinstance(item2, list):
                                        for item3 in item2:
                                            out.append((', '.join(item3)).encode("utf-8").strip())
                                    elif isinstance(item2, dict):
                                         for key3, value3 in item2.iteritems():
                                            out.append((', '.join(valu3)).encode("utf-8").strip())
                                    else:
                                        out.append(item2.encode("utf-8").strip())
```

/ 10400/ 1 1 1 04

37

#### DPLA Json objects report excerpt

```
brighid ▼~ ▼T YmetadataQA Ypython dpla_analysis.py test/dplaafter2020.json -e '$.dataProvider' | sort | uniq -c
  1 Anaheim Public Library
  5 Archives Center - NMAH
  1 Arizona State Library, Archives and Public Records
  7 Banning Library District
  2 Bead Museum
  1 Boston Public Library
 14 Brigham Young University - Harold B. Lee Library
  1 Bryn Mawr College
  1 Cochise County (AZ) Clerk of the Superior Court
 27 Colby College
2778 Cooper Hewitt, Smithsonian Design Museum
708 Cornell University
 12 Dallas Museum of Art
 20 Deer Valley Rock Art Center
  1 Dorot Jewish Division. The New York Public Library
  4 East Carolina University
 261 Freer Gallery of Art and Arthur M. Sackler Gallery
  2 Georgia State University. Libraries. Special Collections
  1 HathiTrust
 12 Historic Huguenot Street
  1 Hobart and William Smith Colleges. Warren Hunting Smith Library
 12 Idaho Commission for Libraries
  2 Indiana Department of Natural Resources, Division of Historic Preservation and Archaeology
  3 Indiana Memory
 30 Indianapolis Museum of Art
 341 Irma and Paul Milstein Division of United States History. Local History and Genealogy. The New York Public Library
  1 Johnson C. Smith University
 34 Linda Hall Library
  1 Los Angeles Public Library
```

github.com/cmh2166/metadataQA

# Method 4: Catmandu

Catmandu = data toolkit (command line client and Perl modules) developed as part of the LibreCat project

http://librecat.org/Catmandu/

# Catmandu Perl Scripts

- Highly customizable
- Works with all types of data
- Can be part of ETL process with many different data stores, engines, platforms, etc.
- Works well with larger datasets (>100k records)
- Requires some programming knowledge
- Installing Perl, Catmandu & dependencies can be tricky  $_{31}$  /  $_{37}$

# Catmandu Reports at UTK

- Saved fix routines
- MARC value reports for big sets (when MARCEdit has trouble)

# Catmandu Reports at UTK

MARC URI Fix routine excerpt

```
marc_map("001", recordIdentifier)
marc_map("020a",isbn.$append)
join_field(isbn," | ")
split_field(isbn, " | ")
marc map("245ab",title, join:" ")
# Subject Identifiers?
if marc_match('6**0', "\w+")
  set field("subjectRecon", "exists")
else
  set_field("subjectRecon", "does not exist")
end
# Name Identifiers?
if marc_match('1**0','\w+')
        set_field("nameRecon", "exists")
else
        set_field("nameRecon", "does not exist")
end
# Primary Name
marc map('100abcdegg024','primaryName.$append', join:" ")
sort_field('primaryName',-unig=>1)
```

# Catmandu Reports at UTK

MARC URI
- Fix
routine
YAML
report

```
id: '994790850102311'
addlCorpName:
- Walker & Associates (Memphis, Tenn)
addlName:
- Butcher, Jacob Franklin, 1936-
nameRecon: does not exist
recordIdentifier: '994790850102311'
subjectName:
- Butcher, Jacob Franklin, | 1936-
subjectRecon: does not exist
subjectTopical:
- Advertising,

    Political—Tennessee

title: Jake Butcher for governor
_id: '991856830102311'
addlCorpName:
- American Animal Hospital Association.
- Pitman-Moore, Inc.
- Tuskegee Institute. School of Veterinary Medicine. Dept. of Radio
addlName:
- Carey, JoAnne.
nameRecon: does not exist
primaryName:
- Hall, Ellis.
recordIdentifier: '991856830102311'
subjectRecon: does not exist
subjectTopical:
Veterinary
```

# Catmandu Reports at UTK

#### MARC General CSV Report, Fix routine Excerpt

```
add_field('type', 'mixed materials')
end
if marc_match("LDR/6","r")
do | add_field('type', 'three-dimensional artifact or naturally occurring object')
end
if marc_match("LDR/6","t")
add_field('type', 'manuscript language material')
end

marc_map('020az',isbn.$append, join:' | ')
join_field('isbn', ' | ')

marc_map("035a",sysNumbers.$append, join:' | ')
join_field('sysNumbers', ' | ')

marc_map('100', 'author')

marc_map('110', 'corpAuthor')

marc_map('245ab','title', join:" ")
```

## Catmandu Reports at UTK, DLTN

### Fix routine for MARC Report

_id	publisher	statementOfResp	sysNumbers	title	type
994790850102311	Walker & Associates,	by Walker & Associates.	(TU)000479085UTK01   504822   (OCoLC)09379175	Jake Butcher for governor	projected medium
991856830102311	Tuskegee Institute, School of Veterinary Medicine, Dept. of Radiology	Ellis Hall, JoAnne Carey.	(TU)000185683UTK01   202056   (OCoLC)07266005	The use of air as a contrast medium in small animal radiology	projected medium
991262280102311	College of Veterinary Medicine, University of Tennessee,	by Dennis Geiser.	(TU)000126228UTK01   139387   (OCoLC)06969820	Equine laryngostomy and ventriculectomy, and emergency tracheostomy.	projected medium
994938970102311	Warner Home Video,	released by Warner Brothers.	(TU)000493897UTK01   520081   (OCoLC)09452899	Day for night	projected medium
994403500102311	Bailey, Deardourff & Associates?,	Bailey, Deardourff & Associates.	(TU)000440350UTK01   464891   (OCoLC)09208626	Alexander spots	projected medium
994478180102311	King of Video,		(TU)000447818UTK01   472632   (OCoLC)08812014	Elvis in the fifties	projected medium
993287420102311	s.n.,		(TU)000328742UTK01   350430   (OCoLC)08192209	[Kefauver miscellany.	projected medium

examples > Catmandu > MARCgeneralReport.csv

# Links & Thank you

Christina Harlow

cmharlow@gmail.com

@cm\_harlow

http://bit.ly/MetadataQA