Online Northwest

Online Northwest 2019

Mar 29th, 11:15 AM - 12:00 PM

HTML to MARC: Webscraping Using Googlesheets

Brianne N. Hagen
Humboldt State University, bnh215@humboldt.edu

Hagen, Brianne N., "HTML to MARC: Webscraping Using Googlesheets" (2019). *Online Northwest*. 12.
https://pdxscholar.library.pdx.edu/onlinenorthwest/2019/schedule/12

HTML to MARC: Web Scrapi
using Google Sheets

Brianne N. Hagen
Discovery & Metadata Services Librarian
Humboldt State University Library

Session Goals

Who this session might be useful for

What web scraping is

When to use it, why it’s useful (and when to use bigger/better tools)

How to do it

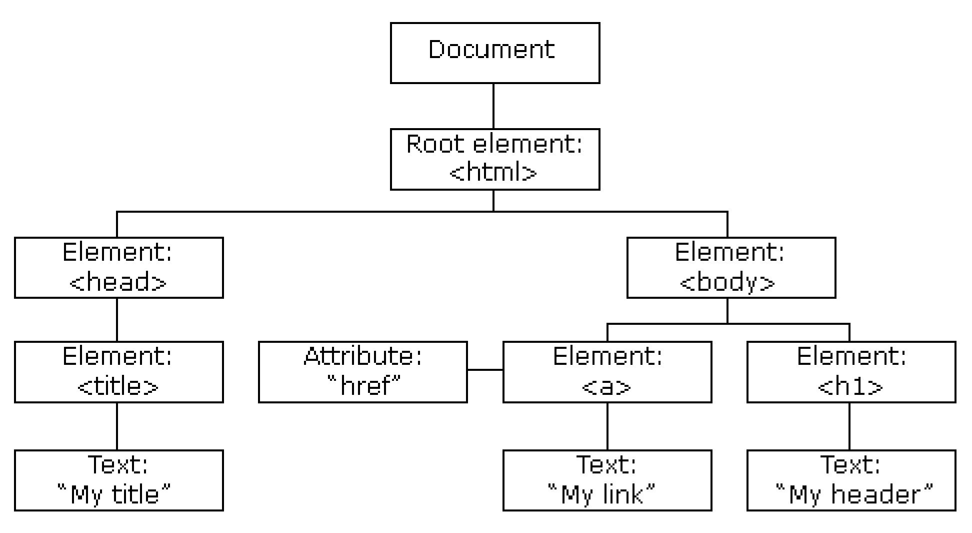
Where to find resources

Who should use web scraping?

- Catalogers
- Electronic resource managers
- Discovery Librarians
- Anyone who wants to make their lives easier

“Web scraping allows you to convert non-tabular or poorly structured data into a usable, structured format, such as a .csv file or spreadsheet”

When to Web Scrape with Google Sheets



Document

Root element:
<html>

Element:
<head>

Element:
<body>

Element:
<title>

Attribute:
“href”

Element:
<a>

Element:
<h1>

Text:
“My title”

Text:
“My link”

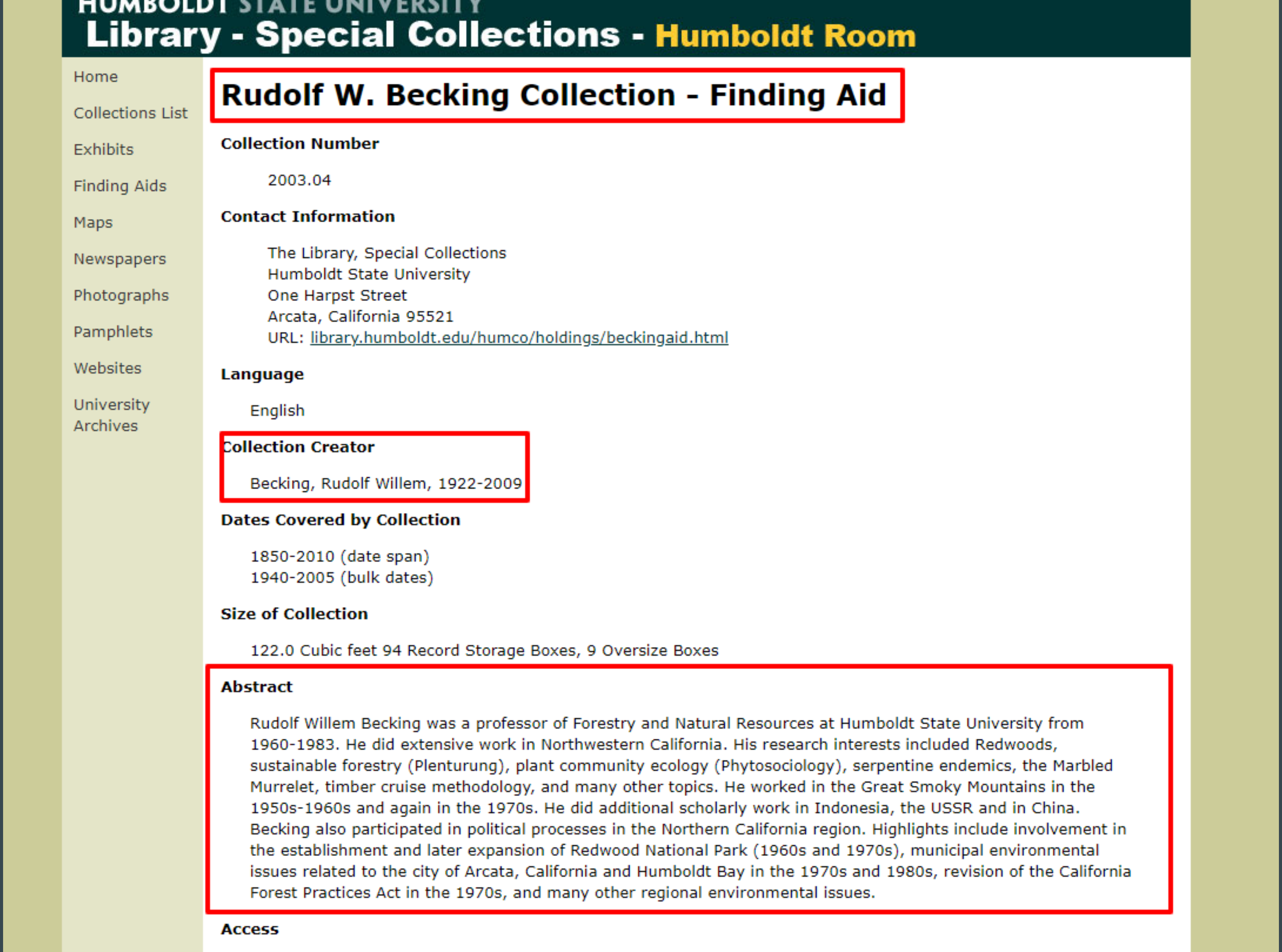
Text:
“My header”

Syntax

=IMPORTXML(“URL”, “query”)

Example:

```
IMPORTXML("https://en.wikipedia.org/wiki/Moon_landing", "//a/@href")
```



Rudolf W. Becking Collection - Finding Aid

Collection Number

2003.04

Contact Information

The Library, Special Collections
Humboldt State University
One Harpst Street
Arcata, California 95521
URL: library.humboldt.edu/humco/holdings/beckingaid.html

Language

English

Collection Creator

Becking, Rudolf Willem, 1922-2009

Dates Covered by Collection

1850-2010 (date span)
1940-2005 (bulk dates)

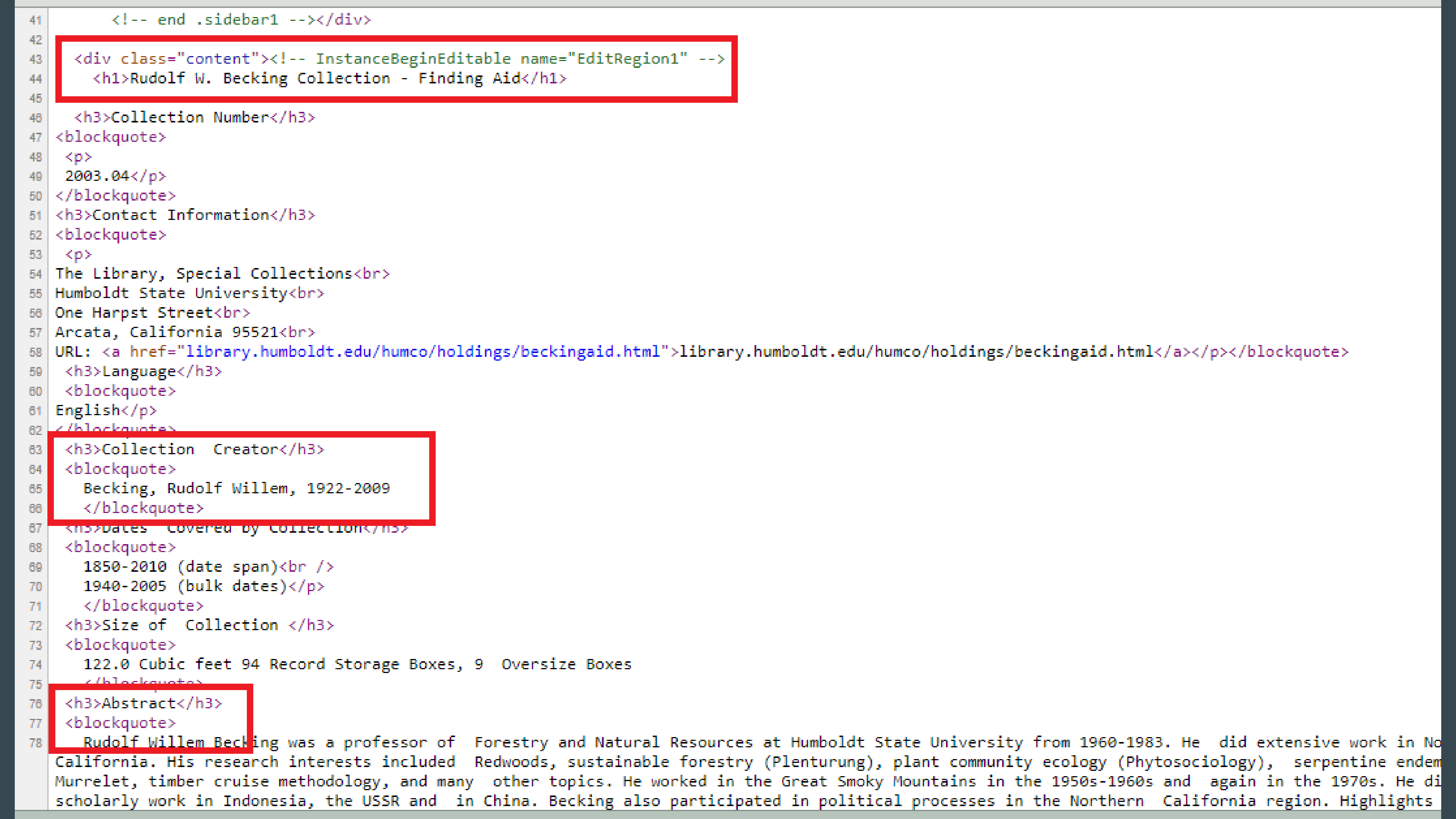
Size of Collection

122.0 Cubic feet 94 Record Storage Boxes, 9 Oversize Boxes

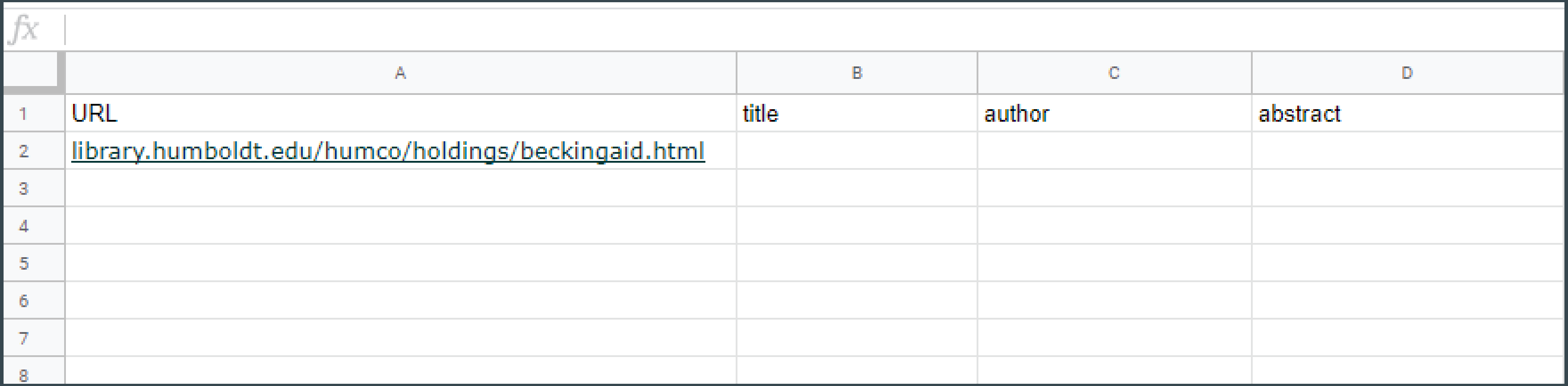
Abstract

Rudolf Willem Becking was a professor of Forestry and Natural Resources at Humboldt State University from 1960-1983. He did extensive work in Northwestern California. His research interests included Redwoods, sustainable forestry (Plenturung), plant community ecology (Phytosociology), serpentine endemics, the Marbled Murrelet, timber cruise methodology, and many other topics. He worked in the Great Smoky Mountains in the 1950s-1960s and again in the 1970s. He did additional scholarly work in Indonesia, the USSR and in China. Becking also participated in political processes in the Northern California region. Highlights include involvement in the establishment and later expansion of Redwood National Park (1960s and 1970s), municipal environmental issues related to the city of Arcata, California and Humboldt Bay in the 1970s and 1980s, revision of the California Forest Practices Act in the 1970s, and many other regional environmental issues.

Access



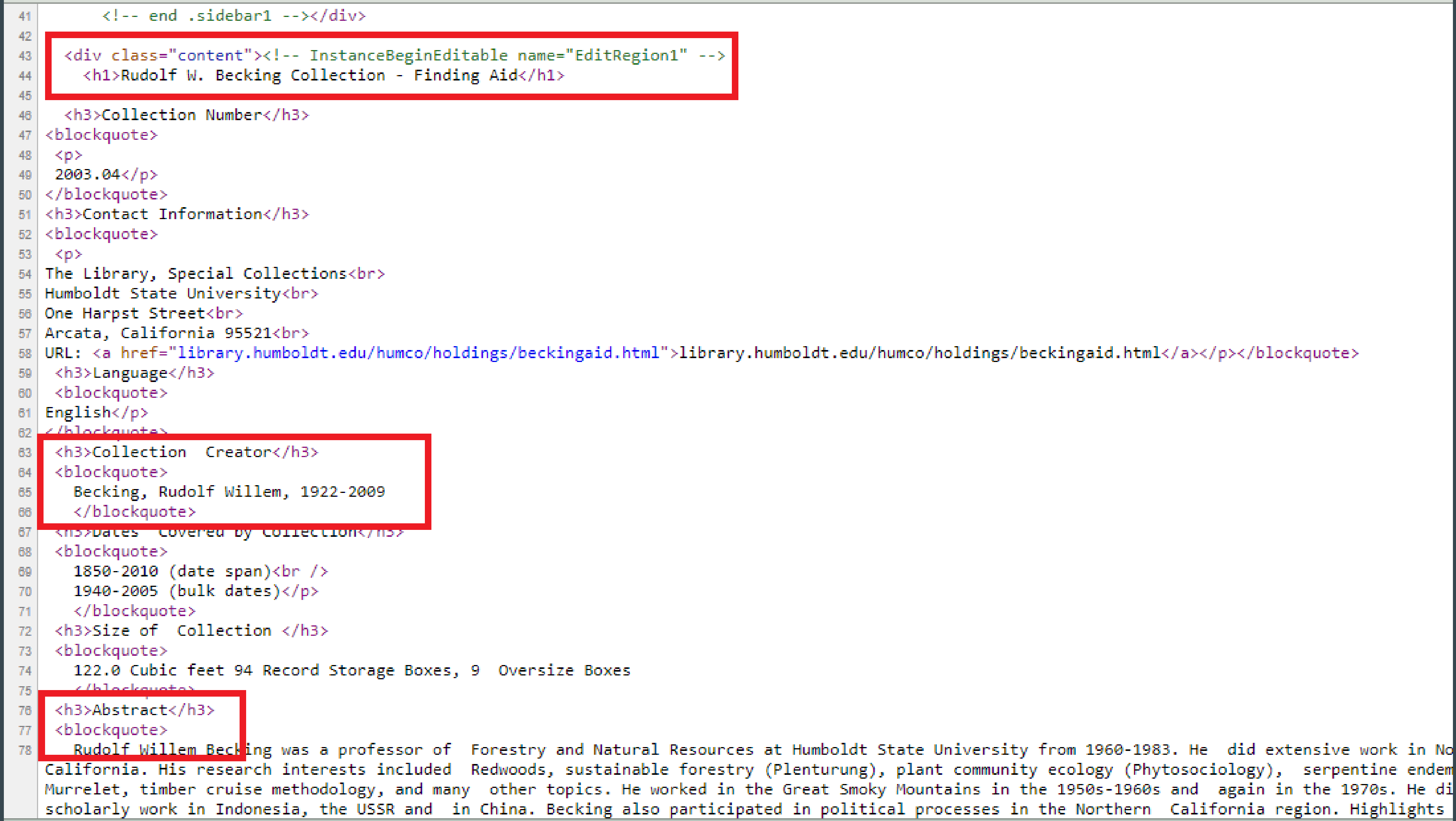
```
      <!-- end .sidebar1 --></div>

  <div class="content"><!-- InstanceBeginEditable name="EditRegion1" -->
    <h1>Rudolf W. Becking Collection - Finding Aid</h1>

  <h3>Collection Number</h3>
<blockquote>
 <p>
 2003.04</p>
</blockquote>
<h3>Contact Information</h3>
<blockquote>
 <p>
The Library, Special Collections<br>
Humboldt State University<br>
One Harpst Street<br>
Arcata, California 95521<br>
URL: <a href="library.humboldt.edu/humco/holdings/beckingaid.html">library.humboldt.edu/humco/holdings/beckingaid.html</a></p></blockquote>
 <h3>Language</h3>
 <blockquote>
English</p>
</blockquote>
 <h3>Collection  Creator</h3>
 <blockquote>
   Becking, Rudolf Willem, 1922-2009
   </blockquote>
 <h3>Dates  Covered by Collection</h3>
 <blockquote>
   1850-2010 (date span)<br />
   1940-2005 (bulk dates)</p>
   </blockquote>
 <h3>Size of  Collection </h3>
 <blockquote>
   122.0 Cubic feet 94 Record Storage Boxes, 9  Oversize Boxes
   </blockquote>
 <h3>Abstract</h3>
 <blockquote>
   Rudolf Willem Becking was a professor of  Forestry and Natural Resources at Humboldt State University from 1960-1983. He  did extensive work in No
California. His research interests included  Redwoods, sustainable forestry (Plenturung), plant community ecology (Phytosociology),  serpentine endem
Murrelet, timber cruise methodology, and many  other topics. He worked in the Great Smoky Mountains in the 1950s-1960s and  again in the 1970s. He di
scholarly work in Indonesia, the USSR and  in China. Becking also participated in political processes in the Northern  California region. Highlights
```



	A	B	C	D
1	URL	title	author	abstract
2	library.humboldt.edu/humco/holdings/beckingaid.html			
3				
4				
5				
6				
7				
8				

Selecting nodes

Expression	Description
/	Selects from the root node
//	Selects nodes in the document from the current node that match the selection no matter where they are
.	Selects the current node
..	Selects the parent of the current node
@	Selects attributes



```
      <!-- end .sidebar1 --></div>

  <div class="content"><!-- InstanceBeginEditable name="EditRegion1" -->
    <h1>Rudolf W. Becking Collection - Finding Aid</h1>

  <h3>Collection Number</h3>
<blockquote>
 <p>
 2003.04</p>
</blockquote>
<h3>Contact Information</h3>
<blockquote>
 <p>
The Library, Special Collections<br>
Humboldt State University<br>
One Harpst Street<br>
Arcata, California 95521<br>
URL: <a href="library.humboldt.edu/humco/holdings/beckingaid.html">library.humboldt.edu/humco/holdings/beckingaid.html</a></p></blockquote>
 <h3>Language</h3>
 <blockquote>
English</p>
</blockquote>
 <h3>Collection  Creator</h3>
 <blockquote>
   Becking, Rudolf Willem, 1922-2009
   </blockquote>
 <h3>Dates  Covered by Collection</h3>
 <blockquote>
   1850-2010 (date span)<br />
   1940-2005 (bulk dates)</p>
   </blockquote>
 <h3>Size of  Collection </h3>
 <blockquote>
   122.0 Cubic feet 94 Record Storage Boxes, 9  Oversize Boxes
   </blockquote>
 <h3>Abstract</h3>
 <blockquote>
   Rudolf Willem Becking was a professor of  Forestry and Natural Resources at Humboldt State University from 1960-1983. He  did extensive work in No
California. His research interests included  Redwoods, sustainable forestry (Plenturung), plant community ecology (Phytosociology),  serpentine endem
Murrelet, timber cruise methodology, and many  other topics. He worked in the Great Smoky Mountains in the 1950s-1960s and  again in the 1970s. He di
scholarly work in Indonesia, the USSR and  in China. Becking also participated in political processes in the Northern  California region. Highlights
```

Title -

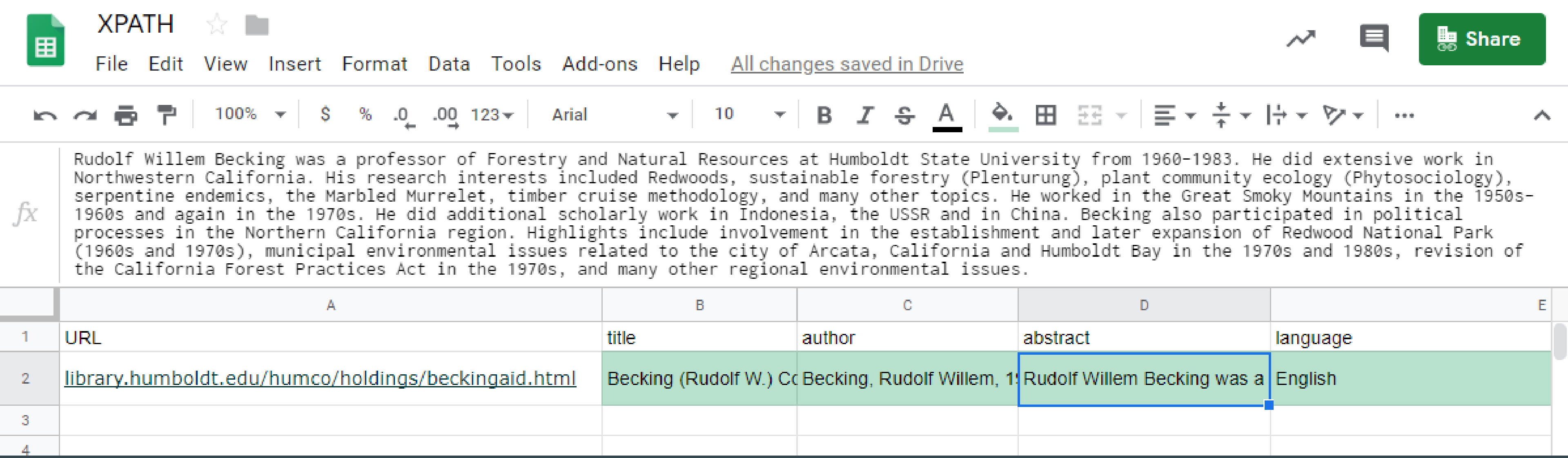
```
=IMPORTXML("http://library.humboldt.edu/humco/holdings/beckingaid.htm","//div[@class='content']//h1")
```

Collection Creator/Author

```
=IMPORTXML("http://library.humboldt.edu/humco/holdings/beckingaid.htm","//div[@class='content']//h3[contains(text(),'Collection Creator')]/following-sibling::blockquote[1]")
```

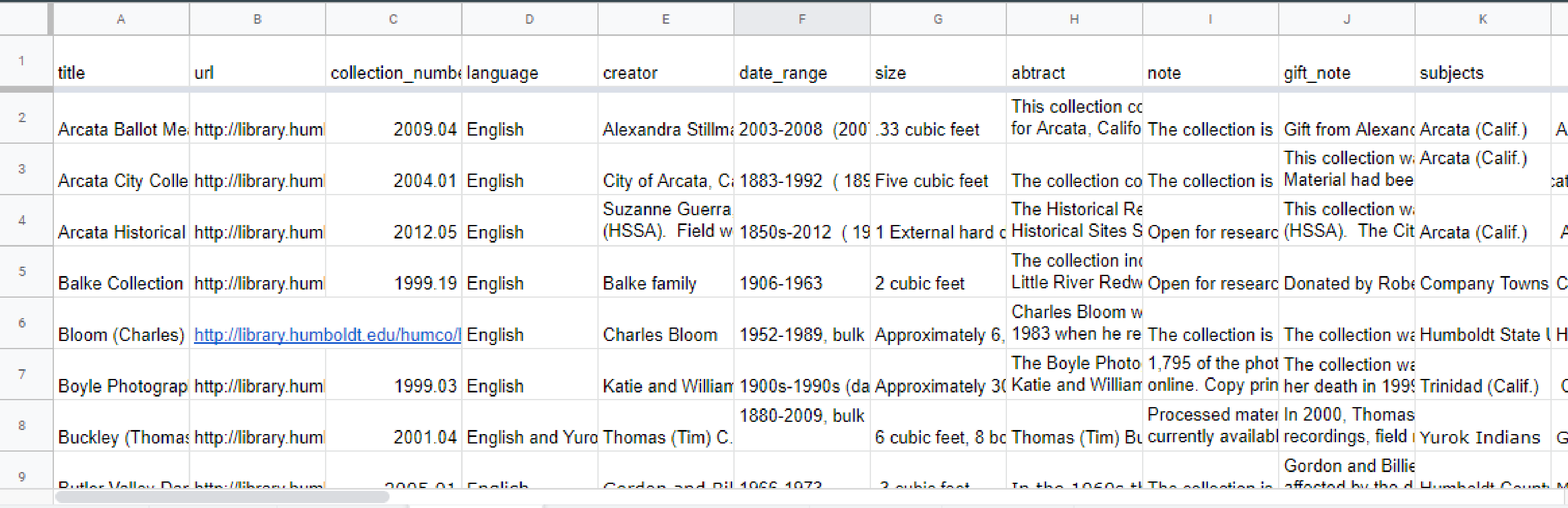
Abstract -

```
=IMPORTXML("http://library.humboldt.edu/humco/holdings/beckingaid.htm","//div[@class='content']//h3/a[@name='abstract']/following::blockquote[1]")
```



Rudolf Willem Becking was a professor of Forestry and Natural Resources at Humboldt State University from 1960-1983. He did extensive work in Northwestern California. His research interests included Redwoods, sustainable forestry (Plenturung), plant community ecology (Phytosociology), serpentine endemics, the Marbled Murrelet, timber cruise methodology, and many other topics. He worked in the Great Smoky Mountains in the 1950s-1960s and again in the 1970s. He did additional scholarly work in Indonesia, the USSR and in China. Becking also participated in political processes in the Northern California region. Highlights include involvement in the establishment and later expansion of Redwood National Park (1960s and 1970s), municipal environmental issues related to the city of Arcata, California and Humboldt Bay in the 1970s and 1980s, revision of the California Forest Practices Act in the 1970s, and many other regional environmental issues.

	A	B	C	D	E
1	URL	title	author	abstract	language
2	library.humboldt.edu/humco/holdings/beckingaid.html	Becking (Rudolf W.) Co	Becking, Rudolf Willem, 1	Rudolf Willem Becking was a	English
3					
4					



	A	B	C	D	E	F	G	H	I	J	K	
1	title	url	collection_numbe	language	creator	date_range	size	abtract	note	gift_note	subjects	
2	Arcata Ballot Me	http://library.huml	2009.04	English	Alexandra Stillma	2003-2008 (200	.33 cubic feet	This collection cc for Arcata, Califo	The collection is	Gift from Alexanc	Arcata (Calif.)	A
3	Arcata City Colle	http://library.huml	2004.01	English	City of Arcata, C	1883-1992 (189	Five cubic feet	The collection co	The collection is	This collection w Material had bee	Arcata (Calif.)	cat
4	Arcata Historical	http://library.huml	2012.05	English	Suzanne Guerra, (HSSA). Field w	1850s-2012 (19	1 External hard c	The Historical Re Historical Sites S	Open for researc	This collection w (HSSA). The Cit	Arcata (Calif.)	A
5	Balke Collection	http://library.huml	1999.19	English	Balke family	1906-1963	2 cubic feet	The collection inc Little River Redw	Open for researc	Donated by Robe	Company Towns	C
6	Bloom (Charles)	http://library.humboldt.edu/humco/l		English	Charles Bloom	1952-1989, bulk	Approximately 6,	Charles Bloom w 1983 when he re	The collection is	The collection wa	Humboldt State	H
7	Boyle Photograp	http://library.huml	1999.03	English	Katie and Willian	1900s-1990s (da	Approximately 3(	The Boyle Photo Katie and Willian	1,795 of the phot online. Copy prin	The collection wa her death in 199	Trinidad (Calif.)	C
8	Buckley (Thomas	http://library.huml	2001.04	English and Yuro	Thomas (Tim) C.	1880-2009, bulk	6 cubic feet, 8 bc	Thomas (Tim) Bu	Processed mater currently availabl	In 2000, Thomas recordings, field	Yurok Indians	G
9	Butler Valley Dar	http://library.huml	2005.01	English	Gordon and Bil	1966-1973	3 cubic feet	In the 1960s th	The collection is	Gordon and Billie affected by the d	Humboldt Count	M

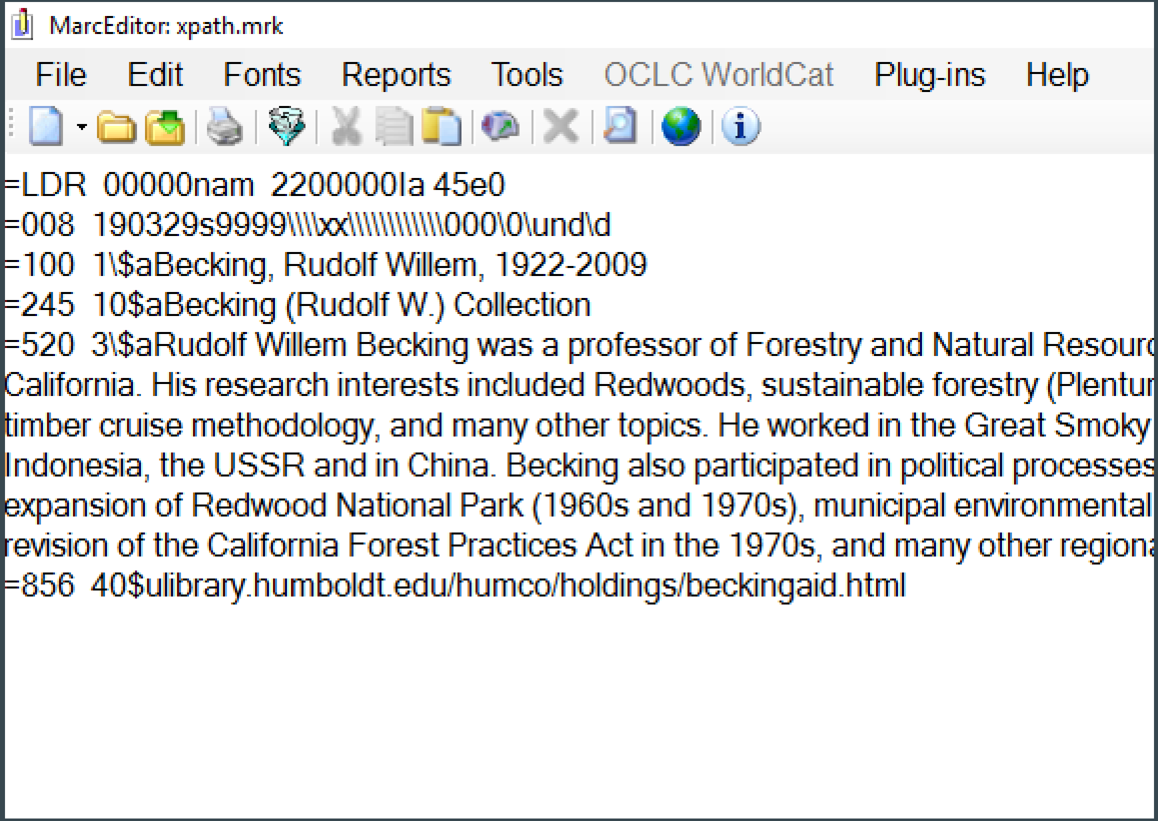
MARCEdit

Download as tsv or csv

Import into MARCEdit Delimited Text Translator

Map fields

Download .mrk file



```
=LDR  00000nam  2200000la 45e0
=008  190329s9999\\\\xx\\\\\\\\\\\\000\0\und\d
=100  1\$aBecking, Rudolf Willem, 1922-2009
=245  10$aBecking (Rudolf W.) Collection
=520  3\$aRudolf Willem Becking was a professor of Forestry and Natural Resourc
California. His research interests included Redwoods, sustainable forestry (Plentur
timber cruise methodology, and many other topics. He worked in the Great Smoky
Indonesia, the USSR and in China. Becking also participated in political processes
expansion of Redwood National Park (1960s and 1970s), municipal environmental
revision of the California Forest Practices Act in the 1970s, and many other regiona
=856  40$ulibrary.humboldt.edu/humco/holdings/beckingaid.html
```

Perfect is the enemy of the good

Problem: Meatdata

More than one page

Set up can be time consuming

Denial of Service Attacks (DoS)

Copyright - is it fair use?

Resources

- Introduction to Web Scraping by Library Carpentry

- Neumann, Mandy, Jan Steinberg, and Philipp Schaer. “Web-Scraping for Non-Programmers: Introducing OXPath for Digital Library Metadata Harvesting *The Code4Lib Journal*, no. 38 (October 18, 2017). https://journal.code4lib.org/articles/13007.

- XPATH Syntax by W3C