

In Search of Useful Collection Metadata

**Using OpenRefine to Create
Accurate, Complete, and Clean
Title Level Collection Information**



- Why you should care
- What other libraries are doing
- What I am doing
- Demonstration



So what is

OPEN
Refine 

The wilds of messy data can be...





Transformed

Clustered

Explored via Facets

Extended

Conquered!





Big Data Friendly

Data Entry Friendly

Totally Coding Free

	A	B	C	D
1	Bib Sysno	245#	506#	520#
2	150809	SSaCurrent index to legal periodicalsSSH[electronic resource].	SSaAccess restricted to Duke University network, or via library proxy server with valid NetID.SS5NcD	SSaSubscription-based current awareness se
3	824737	SSaFulltext sources onlineSSH[serial] /SScBiblioData.		
4	1209094	SSaPAIS InternationalSSH[electronic resource].	SSaAvailable through Cambridge Scientific Abstracts' Internet Database Service (IDS). Access restrict	SSaThis database chronicles issues in the pul
5	1212326	SSaNTIS bibliographic databaseSSH[electronic resource].		
6	1212466	SSaEconLitSSH[electronic resource].		SSa"EconLit, compiled by the American Econ
7	1213178	SSaEthnic NewsWatch and ENW : a historySSH[electronic resource].		SSaContains the full text of more than 450,0
8	1540393	SSaINIS databaseSSH[electronic resource].	SSaFree, open and unrestricted access (formerly restricted to subscribers).	SSaCovers all aspects of the peaceful uses o
9	1995726	SSaDYABOLASSSH[electronic resource].		SSaIndexes books, articles and essays on ant
10	2024673	SSaEncyclopaedia Britannica onlineSSH[electronic resource].		SSaConsists of a fully searchable and brows
11	2066976	SSaMathSciNetSSH[electronic resource] /SScAmerican Mathematical Society.		SSaContains more than 1.7 million citations
12	2232368	SSaProject MuseSSH[electronic resource] :SSbscholarly journals online.	SSaRestricted to institutions with an electronic subscription and requires a site/user ID and passwor	SSaA database of e-journals published by vai
13	2232521	SSaJSTORSSH[electronic resource]		SSaProvides page images of back issues of th
14	2273271	SSaLiterature online :SSH[electronic resource]SSbLION.		SSaA searchable collection of over a quarter
15	2316701	SSaLatin America Data BaseSSH[electronic resource] :SSbLADB : a news and educational service on	SSaAccess restricted to subscribers.	
16	2372299	SSaUlrich's international periodicals directorySSH[electronic resource]		
17	2372985	SSaBiography and genealogy master indexSSH[electronic resource].	SSaAccess limited to 2 simultaneous users.SS5NcD	SSaAccess restricted to subscribing institutions.
18	2372987	SSaAssociations unlimitedSSH[electronic resource] /SScGale Research.	SSaSubscription to GaleNet required for access.	SSaIndexes current, readily-available biograp
19	2375833	SSaCPL. QSSH[electronic resource] :SSbCanadian periodical index.	SSaAccess limited to 2 simultaneous users.SS5NcD	SSaInformation about associations and prof
20	2380605	SSaLe Monde diplomatiqueSSH[electronic resource].		SSaProvides comprehensive Canadian and in
21	2380790	SSaBHA and RILASSSH[electronic resource].	SSaFreely available database.	
22	2395550	SSaPeriodicals index onlineSSH[electronic resource].	SSaAccess is restricted to licensed subscribers.	SSaContains more than 487,000 records that
23	2397327	SSaColumbia International Affairs OnlineSSH[electronic resource] :SSbCIAO.	SSaAccess restricted to subscribers.	SSaIndexes thousands of selected periodical
24	2397364	SSaMarciveWeb DOCSSSSH[electronic resource].	SSaAccess restricted to subscribers.	SSa"Columbia International Affairs Online (C
25	2404372	SSaLinguistics & language behavior abstractsSSH[electronic resource] :SSbLLBA	SSaAccess restricted to subscribers.	SSaOnline catalog of U.S. government public
26	2428971	SSaProQuestSSH[electronic resource].	SSaAccess restricted to subscribers.	SSaSubjects covered include the nature, use,
27	2429000	SSaIndex of Christian artSSH[electronic resource]/SScPrinceton University.	SSaAccess restricted to subscribers.	SSaIncludes summaries of articles from over
28	2431435	SSaASFASSH[electronic resource] :SSbaquatic sciences & fisheries abstracts.	SSaAccess restricted to subscribers.	SSaA thematic and iconographic index of ear
29	2431437	SSaEnvironmental science and pollution managementSSH[electronic resource].	SSaAccess restricted to subscribers.	SSaCovers aquaculture, aquatic biology, aqu
30	2437461	SSaAccessUNSSH[electronic resource].	SSaAccess restricted to subscribers.	SSaProvides citations and abstracts to the lit

Research Databases

Keywords Find by title & subject

Advanced Search Trial Databases

General Databases

[Academic Search Complete](#) [JSTOR](#) [LexisNexis Academic](#) [Web of Science](#) [Proquest](#) [WorldCat](#)

Browse Databases by Subject

[Reference & Primary Sources](#)

[Arts & Humanities](#)

[Business](#)

[Data & Statistics](#)

[Engineering, Computer Science & Mathematics](#)

[Government & Law](#)

[Health & Medical Science](#)

[History](#)

[International Studies](#)

[Sciences](#)

[Social Sciences](#)

[Biography](#)

[Book Reviews](#)

[Dictionaries and Encyclopedias](#)

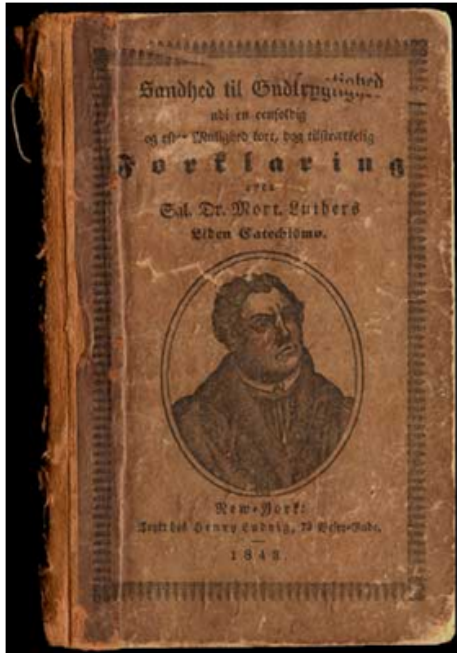
[Directories](#)

[Newspapers](#)

[Primary Sources](#)

Images
from
Jackie
Samples

Norwegian-American Book Cover Art: A Visual Guide



Sandhed til gudfrygtighed: en ... forklaring over sal. Dr. Morten Luthers Liden katekismus. Befordret til trykken af Elling Eielson [sic]

Author: Pontoppidan, Erich, 1698-1764

Publisher: New York :Trykt hos H. Ludvig, 1842 (cover date: 1843)

[Bridge catalog record](#)

Spine height: 16 cm. Notes: Elling Eielson--founder of the first Norwegian Lutheran synod in the New World--came to America some fifteen years after the first immigrants from Norway had arrived. One year earlier he had his first book, Luther's Small Catechism, published in English, thinking that all the Norwegian children were speaking only English by then. He was wrong. The Norwegian language would be the official language spoken in Norwegian Lutheran churches in America well into the twentieth century.

Project name **Sample data.xml**

Create Project **Start Over** Configure Parsing Options

Open Project

Import Project

```
<<:ID>Sample Vendor</:ID>
<<:Contact>
  <c:Contact>Sample Customer Support</c:Contact>
  <c:E-mail>support@sample.com</c:E-mail>
</:Contact>
<c:WebSiteUrl>http://www.website.com</c:WebSiteUrl>
</:Vendor>
<<:Customer>
  <c:Name>Sample College</c:Name>
  <c:ID>555555</c:ID>
  <c:Consortium>
    <c:Code>12</c:Code>
    <c:WellKnownName>OhioLink</c:WellKnownName>
  </c:Consortium>
  <c:ReportItems>
    <c:ItemIdentifier>
      <c:Type>Proprietary</c:Type>
      <c:Value>Sample Citation Index</c:Value>
    </c:ItemIdentifier>
    <c:ItemPlatform>Sample Platform</c:ItemPlatform>
    <c:ItemPublisher>Sample Publisher</c:ItemPublisher>
```

Images
by
Brooklyn
Ludlow

Title

American Journal of Medical Quality (Modify title through variants below)

Status [Current](#)

Status Reason [Empty](#)

Edit Status [Approved](#)

Latest Publisher Sage (Organization 8723)

Imprint [Empty](#) [Follow Link](#)

Published From [1986/02/01](#)

Published To [2016/03/07](#)

Title History

On this date	This/These titles	Changed to This/These titles
--------------	-------------------	------------------------------

Available actions

-- Select an action

Support

Title Details

[Alternate Names](#) 0

[Add to Title History](#)

[Identifiers](#) 2

[Publishers](#) 2

[Availability](#) 2

[Custom Fields](#) 0

[Review Tasks](#) 1



My Own OpenRefine Journey





Openrefine.org

You will find on this page a list of OpenRefine distributions and extensions available for download. Are we missing something? Want to fix a typo? You can submit changes (pull request) [from here](#).

[Home](#)

[Download](#)

[Documentation](#)

[Community](#)

[Post archive](#)

[OpenRefine News:](#)
[December 2015](#)

Official Distribution


Read the [installation instructions](#)

OpenRefine 2.6

This is the first beta release of OpenRefine 2.6 on Aug 27, 2013. A change log is provided on [the release page](#).

- **Windows kit**, Download, unzip, and double-click on *google-refine.exe*. If you're having issues with the above, try double-clicking on *refine.bat* instead.
- **Mac kit**, Download, open, drag icon into the Applications folder and double click on it.
- **Linux kit**, Download, extract, then type `./refine` to start.

2. Create a Project

 *A power tool for working with messy data.*

- Create Project
- Open Project
- Import Project
- GOKb

Create a project by importing data. What kinds of data files can I import?

TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data Google Refine extensions.

Get data from

This Computer

Web Addresses (URLs)

Clipboard

Google Data

Locate one or more files on your computer to u

Browse...

No files selected.

Next »

2. Create a Project



Google refine *A power tool for working with messy data.*

Create Project

Open Project

Import Project

GOKb

Create a project by importing data. What kinds of data files can I import?

TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data
Google Refine extensions.

Get data from

This Computer

Web Addresses (URLs)

Clipboard

Google Data

Paste data from clipboard here:

2. Create a Project



A power tool for working with messy data.

Create Project

Open Project

Import Project

GOKb

Create a project by importing data. What kinds of data files can I import?

TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data document Google Refine extensions.

Get data from

Locate one or more files on your computer to upload:

This Computer

Browse...

Wiley_ARDI_Collection.xlsx

Web Addresses (URLs)

Next »





Parse data as

Excel (.xlsx) files

XML files

Open Document Format spreadsheets (.ods)

RDF/XML files

JSON files

Line-based text files

Worksheets to Import

All Lists Merged Full 6900 rows

- Ignore first 0 line(s) at beginning of file
- Parse next 1 line(s) as column headers
- Discard initial 0 row(s) of data
- Load at most 0 row(s) of data

Update Preview

- Store blank rows
- Store blank cells as nulls
- Store file source (file names, URLs) in each row

Live Demo

- **OpenRefine Main Documentation:**
 - Main Refine Page: <http://openrefine.com>
 - Official OpenRefine FAQ: <https://github.com/OpenRefine/OpenRefine/wiki/FAQ>
 - Screencasts introducing OpenRefine: <https://github.com/OpenRefine/OpenRefine/wiki/Screencasts>
 - OpenRefine Wiki: <https://github.com/OpenRefine/OpenRefine/wiki/>
- **General Tutorials:**
 - Free Your Metadata: <http://freeyourmetadata.com>
 - Getting Started with OpenRefine: <https://wikis.utexas.edu/pages/viewpage.action?pageId=46631837>)
 - A Librarian's Guide to OpenRefine: <http://acrl.ala.org/techconnect/?p=4253>
 - Using OpenRefine to Clean Multiple Documents in the Same Way: <http://schoolofdata.org/2013/07/26/using-openrefine-to-clean-multiple-documents-in-the-same-way/>
 - Chitchat with New Datasets: Facets in OpenRefine:
– <https://blog.ouseful.info/2012/11/06/chit-chat-with-new-datasets-facets-in-open-was-google-refine>

Get out there and explore!

