The Magic of OpenRefine

PRESENTER: Andrea Knabe-Schönemann

INTRO:

- increasing number of open data sets, more data in machine readable formats – a treasure trove for Free Knowledge
- there are less people with technical skills (technical experts) to handle the data than people with the knowledge to understand the data (domain experts)
- if we can close this gap we can get more and more diverse data for our projects

Data Pipeline

- 1. finding data sources/acquiring data
- 2. processing data
 - discovering
 - extracting
 - cleansing, transforming, integrating
 - modeling
- 3. using the data

Processing data takes up 60% - 80% of the time.

Spreadsheets

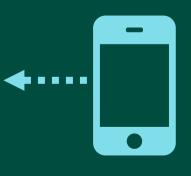
- mostly used by domain experts to handle data
- set relational data sets to "flat"
- weak for large amounts of data
- less features for data processing
- less features for data discovery

OpenRefine

- works reliably for up to 5 million rows of data
- gives coding power to non-coders
- exports to common data formats
- Offers extensions, libraries, and interfaces for many use cases and external services

OpenRefine can do more than just upload data to Wikidata. It can help you get more and more diverse data for your projects – empower yourself!





Take a picture to find out even more

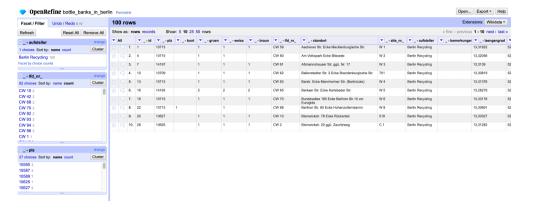
Examples

acquiring data

- 1. stored locally on your computer (many common data formats)
- 2. load data via API or URL
- 3. copy & paste from any source (Wikipedia lists e.g.)

exploring data

download via URL, clustering



preparing data use powerful features

acet / Filter Undo / Redo 070		100 row	/s									Extensions:	Wikidata
efresh Reset All	Remove All	Show as: I	ows records	Show: 5	10 25 50 ro	ws					« firs	t oprevious 1-50) next⇒ la
- aufsteller	change	💌 bunt	💌 gruen	💌 weiss	💌 braun	💌 Ifd_nr_	💌 standort	💌 alte_nr_	💌 aufsteller	💌 bemerkunger		💌 breitengrad	💌 st
hoices Sort by: name count	Cluster		1	1	1	CW 59	Aachener Str. Ecke Mecklenburgische Str.	W 1	Berlin Recycling		13,31822	^E Facet	klenbr 89, 10
fin Recycling 100			1	1		CW 60	Am Volkspark Ecke Blissestr.	W 2	Berlin Recycling		13,32056	E Text filter	Volksp 15 Ber
at by choice counts			1	1	1	CW 61	Aßmannshauser Str. ggü. Nr. 17	W 3	Berlin Recycling		13,3139	Edit cells	hanna
-	change		1	1	1	CW 62	Ballenstedter Str. 2 Ecke Brandenburgische Str.	701	Berin Recycling	plit into several colum	ns	Edit column	+ 17, 14 ensted
- Ifd_nr_ thoices Sort by: name count	Cluster		1	1	1	CW 63	Barstr. Ecke Mannheimer Str. (Barbrücke)	W 4	Berlin Recycling A	dd column based on ti	nis column	Transpose	▶ 09 Be str. 43
10 2	Ordenie		2	2	2	CW 65	Berkaer Str. Ecke Karlsbader Str.	W 5	Berin Recycline A	dd column by fetching	URLs	Sort	lin
/ 42 2						CW 70	Bundesallee 180 Ecke Berliner Str.10 vor Eurogida	W6		dd columns from reco	nciled values	View	 93 Be idesa
/ 68 2			1	1	1					ename this column		Reconcile	15 Be
(70 2 (82 2		1		1		CW 68	Berliner Str. 80 Ecke Hohenzollerndamm	W 8		emove this column		6710010	10713 Be
/ 93 2			1	1	1	CW 10	Bismarckstr. 78 Ecke Rückertstr.	519	Berlin Recyclin	love column to beginn	ina	2,51112	Bismarci 10627 Bi
/ 94 2			1	1	1	CW 2	Bismarckstr. 20 ggü. Zauritzweg	C 1	Berlin Recycline	lave column to end	- 9	2,51216	Bismarck 10525 Be
V 96 2 V 1 1			1	1	1	CW 3	Brahestr. 6 - 7 / Sportplatz	C 2	Berlin Recycling N	love column left		2,52701	Brahestr. Berlin
_			1	1	1	CW 68	Brandenburgische Str. 1 - 2 Ecke Sigmaringer Str.	W 9	Berlin Recyclin N	tove column right		2,48725	Brandent Str. 1, 10
piz	change		1	1		CW 86	Konstanzer Str. 18	W 10	Berlin Recycling		13,3099	52,49395	Konstanz
thoices Sort by: name count	Cluster		1	1	1	CW 69	Bundesallee 188 Ecke Tharandter Str.4	W 11	Berlin Recycling		13.33119	52,48992	10707 Be Tharandt
85 3			1	1	1	CW 70	Bundesallee 178 Ecke Berliner Str. ön der	W 12	Berlin Recycling		13.33107	52,4869	10717 Be Bundesa
87 4						01110	Wendeschiefe)		benn nesyong			08/1000	10717 B

combining data

join datasets or enrich your source with new data

Facet / Filter Undo / Redo or 6 rows

	s	how	as: 1	rows records	Show: 5 10	25	50 rows				« first «	previous 1-6 next> last
Using facets and filters	·	AI		💌 code	💌 link			💌 thumb		💌 title	thumb_large	thumb_featured
•	12		1.	noj	Facet	٠	es/collection/fbj/	https://www.loc.gov/picture	as/static/data/tbjthumb.png	Johnston (Frances Benjamin) Collection	https://www.loc.gov/pictures/static/data/fbj/thumb_large.png	https://www.loc.gov/pictures.
Use facets and filters to select subsets of your data to act on. Choose facet and	2		2	highsm	Text filter		es/collection/highern/	https://www.loc.gov/picture	es/static/data/highem/thumb.png	Highsmith (Carol M.) Archive	https://www.loc.gov/pictures/static/data/highsm/thumb_large.png	https://www.loc.gov/pictures/
filter methods from the menus at the top	12		3.	. 10	Edit cells	•	es/collection/lij/	https://www.loc.gov/picture	as/static/data/lij/thumb.png	Liljenquist Family Collection of Civil	https://www.loc.gov/pictures/static/data/lii/thumb_large.png	https://www.loc.gov/pictures/
of each data column.					Edit column	•	Split into several or	lumns		War Photographs		
Not sure how to get started?	12		4.	. hh	Transpose Sort	•	Add column based on this column	on this column	/static/data/hh/thumb.png	Historic American Buildings Survey/Historic American Engineering	https://www.loc.gov/pictures/static/data/hhthumb_large.png	https://www.loc.gov/pictures
Watch these screencasts							Add column by feto	itching URLs				
					View	•	Add columns from	reconciled values		Record/Historic American		
	23 6				Reconcile		Rename this column	n		Landscapes Survey		
			6.		untre unwww.oc.gov	open	Remove this colum	n	/static/data/wpapos/thumb.png	Posters: WPA Posters	https://www.loc.gov/pictures/static/data/wpapos/thumb_large.png	https://www.loc.gov/pictures
	2		6.	fsa	https://www.loc.go	//picts	Move column to be	ginning	/static/data/fsa/thumb.prg	Farm Security Administration/Office of War Information	https://www.loc.gov/pictures/static/data/fsa/thumb_large.png	https://www.loc.gov/pictures
							Move column to en	d		Black-and-White Negatives		
										Neganies		
							Marine and more stable		1			

exporting data

export data with pre-defined scripts, with a custom tabular exporter for manual fine-tuning, or formatted for a direct upload to QuickStatements

OpenRefine list of all collections	Permalink					Open Expo	rt≖ H
acet / Filter Undo / Redo 0/0	6 rows					Export project Project data package	dat
Using facets and filters	Show as: rows records	Show: 5 10 25 50 rows	💌 thumb	💌 title	thumb_large	Tab-separated value Comma-separated value	> la
•	😭 🗐 1. fbj	https://www.loc.gow/pictures/collection/fbj/	https://www.loc.gov/pictures/static/data/tbj/thumb.png	Johnston (Frances Benjamin) Collection	https://www.loc.gov/pictures/static/da	HTML table	ctu
Use facets and filters to select subsets of your data to act on. Choose facet and	😭 ק 2. highsm	https://www.loc.gow/pictures/collection/highsm/	https://www.loc.gov/pictures/static/data/highsm/thumb.png	Highsmith (Carol M.) Archive	https://www.loc.gov/pictures/static/da	Excel (.xis)	ict.
filter methods from the menus at the top of each data column.	☆ 🤤 3. 태	https://www.loc.gov/pictures/collection/N/	https://www.loc.gov/pictures/static/data/Npthumb.png	Liljenquist Family Collection of Civil War Photographs	https://www.loc.gov/pictures/static/da	Excel 2007+ (.xisx) ODF spreadsheet	icti
Not sure how to get started? Watch these screencasts	☆ 🖓 4. hh	https://www.ioc.gov/pictures/collection/htv	https://www.loc.gov/pictures/static/data/hhthumb.png	Historic American Bulkings Survey/Historic American Engineering Record/Historic American Landscapes Survey	https://www.loc.gov/pictures/static/da	Custom tabular exporter SQL Exporter Tempfating Upload edits to Wkidata	icti
	😭 🗐 5. wpapos	https://www.loc.gov/pictures/collection/wpapos/	https://www.loc.gov/pictures/static/data/wpapos/thumb.png	Posters: WPA Posters	https://www.loc.gov/pictures/static/da	Export to QuickStatements	ict.
	😭 🖓 6. fsa	https://www.loc.gov/pictures/collection/fsa/	https://www.loc.gov/pictures/static/data/fsa/thumb.png	Farm Security Administration/Office of War Information Black-and-White	https://www.loc.gov/pictures/static/da	Export schema	eti

bonus

you can write code or use regex (but you don't have to...)

et / Filter Undo / R	tedo o / o	6 rows					Extensions: Wikida
		Show as: rows records	Show: 5 10 25 50 rows				t «previous 1-6 next» I
Using facets and fil		All code	💌 link	💌 thumb	💌 title	thumb_large	thumb_featured
-		会 🖓 1. 📾	https://www.loc.gov/pictures/collection/fbj/	https://www.loc.gov/pictures/statio/data/fbj/thumb.png	Johnston (Frances Benjamin) Collection	https://www.loc.gov/pictures/atatic/data/lbj/thumb_large.png	https://www.loc.gov/pictu
Jse facets and filters to of your data to act on.	select subsets	A C 2 Nghsm	https://www.loc.gov/pictures/collection/highsm	https://www.loc.gov/pictures/static/data/highsm/thumb.		https://www.loc.gov/pictures/statio/data/highsm/thumb_large.pr	ng_https://www.loc.gow/pictu
	Custom text tran	nsform on column _ ·	title				https://www.loc.gov/pictr
lot sure how to get s Vatch these screen	Expression value.contains("A:	rchive")			L	anguage Sentral Refine Expression Language (GREL) Python / Jython Clojure	https://www.loc.gov/pidu
	Preview His	tory Starred Help					https://www.loc.gov/pict
	row value			valu	e.contains("Archive")		https://www.loc.gov/pict
	1. Johnston (F	rances Benjamin) Collectio	n	false			
	2. Highsmith (Carol M.) Archive		true			
		Carol M.) Archive amily Collection of Civil Wa	r Photographs	true			
	3. Liljenquist F 4. Historic Am	amily Collection of Civil Wa erican Buildings Survey/His	ar Photographs storic American Engineering Record/Hist	false oric American Landscapes Survey false			
	3. Liljenquist F 4. Historic Am 5. Posters: Wi	amily Collection of Civil Wa erican Buildings Survey/His PA Posters	storic American Engineering Record/Hist	falsi oric American Landscapes Survey falsi falsi			
	3. Liljenquist F 4. Historic Am 5. Posters: Wi	amily Collection of Civil Wa erican Buildings Survey/His PA Posters		falsi oric American Landscapes Survey falsi falsi			
	3. Liljenquist F 4. Historic Am 5. Posters: Wi	amily Collection of Civil Wa erican Buildings Survey/His PA Posters	storic American Engineering Record/Hist	falsi oric American Landscapes Survey falsi falsi			