Digital humanities

2. Cvičení: OpenRefine

Jindřich Marek

OpenRefine

- dříve Google Refine
- nástroj na čištění a obohacování dat
- mnoho možností importu i exportu

Instalace a spuštění

- https://openrefine.org/download.html
  - pokud používáte Windows, nejlepší volbou je „ZIP file, with embedded Java install“
- po stažení archiv ve formátu zip rozbalte a spusťte *openrefine.exe* nebo *refine.bat*
  - okno s příkazovým řádkem nezavírejte
  - instrukce pro Mac a Linux tamtéž
- v prohlížeči navštivte http://127.0.0.1:3333
  - stránka běží lokálně na vašem počítači
  - měli byste vidět stránku na následujícím snímku

OpenRefine *A power tool for working with messy data.*

- Create project
- Open project
- Import project
- Language settings

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 documents are all supported. Support for other formats can be added with OpenRefine extensions.

Get data from

- **This Computer**
- Web Addresses (URLs)
- Clipboard
- Database
- Google Data

Locate one or more files on your computer to upload:

Procházet... Soubory nevybrány.

Next »

Version 3.6-SNAPSHOT [864ac30]

- Preferences
- Help
- About

První seznamení s programem

- podívejte se na manuál: https://openrefine.org/docs
  - zejména oddíly Exporing data, Transforming data a Reconciling

Načtení dat

- načtěte soubor Rozmberska_knihovna_rozpis.xlsx
  - soubor je uložený v Moodlu, stáhněte si jej do svého počítače
  - soubor ve formátu Excel obsahuje jeden list
  - po načtení souboru byste měli vidět stránku na následujícím snímku
- soubor obsahuje záznamy k rukopisům zapsaným v katalogu Rožmberské knihovny z roku 1608
  - vedle záznamu v katalogu jsou zde i další údaje včetně případné identifikace exemplářů

OpenRefine Rozmberska knihovna rozpis xlsx Permalink

Open... Export Help

Facet / Filter | Undo / Redo 0 / 0

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
Watch these screencasts

990 rows

Extensions: Wikidata

Show as: **rows** records Show: 5 **10** 25 50 100 500 1000 rows « first ‹ previous 1 of 99 pages next › last »

All		Č.	Katalog	Obor/číslo	Fol.	Autor	Název	Datace	D1	D2	Místo vzniku	Jazyk	Ps.l., vazba	Obor
1.	1	Březan	Teologie	10r		Biblí staročeské psané na pergameně (jenž někdy náležela pánům z Kunstátu) dílové dva I. II.	1400-1450	1400	1500	Morava?	ces	perg; ilum, neúpl.; obsahuje i Žďárskou kroniku. Vazba prvního i druhého svazku shodná (převazba rožmberské provenience z doby kolem roku 1608): vazba kožená, dochovány středové puklice, 6 nárožnic, na hřbetu papírové štítky (název díla, starší signatury MZK a Královské knihovny ve Stockholmu).	bible

Vytvoření textových faset

- inspirujte se v manuálu odkázaném výše
- vytvořte textovou fasetu (facet) pro sloupec Místo vzniku
- vytvořte textovou fasetu (facet) pro sloupec Jazyk

OpenRefine Rozmberska knihovna rozpis xlsx Permalink

Open... Export Help

Facet / Filter | Undo / Redo 0 / 0

Refresh | Reset all | Remove all

Místo vzniku change

38 choices Sort by: **name** count Cluster

- [Třeboň: rožmberský dvůr, Václav Březan písař] 1
- Čechy 34
- Čechy, Praetorius Havel, David - písař 1
- Čechy? 2
- Český Krumlov 1
- Český Krumlov? rožmberský dvůr? 1
- Ferrara 1
- Francie 1
- Itálie 1

Jazyk change

21 choices Sort by: **name** count Cluster

- ces 95
- ces lat 1
- ces, deu, lat 1
- ces, lat 3
- cze 1
- deu 83
- deu lat, ces 1
- deu, lat 3

990 rows

Extensions: Wikidata

Show as: **rows** records Show: 5 **10** 25 50 100 500 1000 rows « first ‹ previous 1 of 99 pages next › last »

All		Č.	Katalog	Obor/číslo	Fol.	Autor	Název	Datace	D1	D2	Místo vzniku	Jazyk	Ps.I., vazba	Obor
1.	1	Březan	Teologie	10r		Biblí staročeské psané na pergameně (jenž někdy náležela pánům z Kunstátu) dílové dva I. II.	1400-1450	1400	1500	Morava?	ces	perg; ilum, neúpl.; obsahuje i Žďárskou kroniku. Vazba prvního i druhého svazku shodná (převazba rožmberské provenience z doby kolem roku 1608): vazba kožená, dochovány středové puklice, 6 nárožnic, na hřbetu papírové štítky (název díla, starší signatury MZK a Královské knihovny ve Stockholmu).	bible

Seskupování a sjednocování hodnot

- pomocí funkce seskupování (cluster) na facetech odhalte různé zápisy stejných hodnot
- tyto hodnoty sjednoťte

Cluster & edit column "Misto vzniku"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more…

Method: key collision | Keying Function: Fingerprint | **5** clusters found

Cluster size	Row Count	Values in cluster	Merge?	New cell value
3	5	• Třeboň: rožmberský dvůr (2 rows) • Třeboň: rožmberský dvůr? (2 rows) • Třeboň - rožmberský dvůr	☐	Třeboň: rožmberský dvůr
2	3	• Praha? (2 rows) • Praha	☐	Praha?
2	7	• Německo (5 rows) • Německo? (2 rows)	☐	Německo
2	36	• Čechy (34 rows) • Čechy? (2 rows)	☐	Čechy
2	2	• Třeboň: rožmberský dvůr, písař Václav Březan • [Třeboň: rožmberský dvůr, Václav Březan písař]	☐	Třeboň: rožmberský dvůr, pís

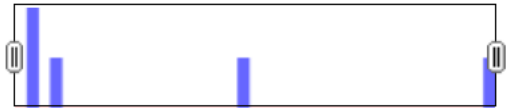
 # Choices in cluster

2 — 3

 # Rows in cluster

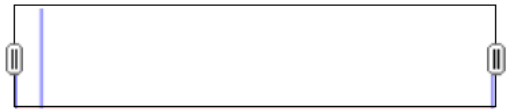
2 — 36

Average length of choices



5 — 45

Length variance of choices



0.47100000000000003 — 1

Select all | Unselect all | Export clusters | **Merge selected & re-cluster** | Merge selected & Close | Close

Různé metody seskupování

- v případě sloupce Místo vzniku vyzkoušejte i jiné metody seskupování (cluster) – viz následující snímek

- jaké jsou výhody (resp. nevýhody) této metody?

Cluster & edit column "Misto vzniku"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more…

Method: key collision | Keying Function: Daitch-Mokotoff | **5** clusters found

Cluster size	Row Count	Values in cluster	Merge?	New cell value
9	12	• Třeboň: rožmberský dvůr (2 rows) • Třeboň: rožmberský dvůr, písař Theobald Höck (2 rows) • Třeboň: rožmberský dvůr? (2 rows) • Třeboň - rožmberský dvůr • Třeboň: rožmberský dvůr (na základě účtů) • Třeboň: rožmberský dvůr, písař Václav Březan • Třeboň: rožmberský dvůr: písař Pavel Lhenický • Třeboň: rožmberský dvůr: písař Václav Burda • [Třeboň: rožmberský dvůr, Václav Březan písař]	☐	Třeboň: rožmberský dvůr
2	36	• Čechy (34 rows) • Čechy? (2 rows)	☐	Čechy
2	2	• Český Krumlov • Český Krumlov? rožmberský dvůr?	☐	Český Krumlov
2	3	• Praha? (2 rows) • Praha	☐	Praha?
2	7	• Německo (5 rows)	☐	Německo

 # Choices in cluster

2 — 9

 # Rows in cluster

2 — 36

Average length of choices

5 — 38

Length variance of choices

0.5 — 9.6

Select all | Unselect all | Export clusters | **Merge selected & re-cluster** | Merge selected & Close | Close

Fasety číselné osy

- převeďte hodnoty ve sloupcích D1 a D2 (limity datace) na datum
  - klikněte na zobáček v záhlaví sloupce: Edit cells > Common transforms > To date
- vytvořte fasety číselné osy (Timeline facet) pro sloupce D1 a D2

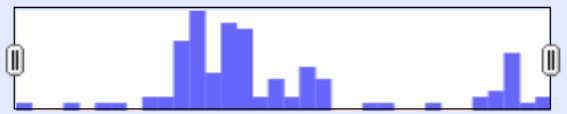
OpenRefine Rozmberska knihovna rozpis xlsx Permalink

Open... Export Help

Facet / Filter | Undo / Redo 17 / 17

Refresh | Reset all | Remove all

D1 change reset



1273-01-01 00:57:44 — **1608-01-01** 00:57:44

Time 109 | Non-Time 208 | Blank 673 | Error 0

D2 change reset

1273-01-01 00:57:44 — **1608-01-01** 00:57:44

Time 109 | Non-Time 207 | Blank 674 | Error 0

990 rows

Extensions: Wikidata

Show as: **rows** records Show: 5 **10** 25 50 100 500 1000 rows « first ‹ previous 1 of 99 pages next › last »

All	Č.	Katalog	Obor/číslo	Fol.	Autor	Název	Datace	D1	D2	Místo vzniku	Jazyk
1.	1	Březan	Teologie	10r		Biblí staročeské psané na pergameně (jenž někdy náležela pánům z Kunstátu) dílové dva I. II.	1400-1450	1400	1500	Morava?	ces

Více hodnot v jedné buňce

- v případě, že v jedné buňce je vloženo více hodnot (např. sloupce Jazyk a Obor v naší tabulce), je vhodné je rozdělit
  - klikněte na zobáček v záhlaví sloupce: Edit cells > Split multi-valued cells…
  - jako oddělovač (Separator) zvolte čárku následovanou mezerou
- po rozdělení bude možné pomocí faset a seskupování odhalit další duplicitní hodnoty
- před exportem je třeba hodnoty v buňkách analogickým způsobem znovu sjednotit, jinak se exportuje jen první hodnota

OpenRefine Rozmberska knihovna rozpis xlsx xlsx Permalink

Open... Export Help

Facet / Filter | Undo / Redo 33 / 33

Refresh | Reset all | Remove all

Jazyk change

8 choices Sort by: **name** count Cluster

- ces 111
- deu 91
- eng 1
- gre 1
- heb 1
- ita 3
- lat 803
- pol 1
- (blank) 431

Facet by choice counts

Obor change

265 choices Sort by: **name** count Cluster

- alegorie společenská 1
- Alexandr Veliký 1
- alchymie 30
- angelologie 1
- antika 2
- apokalyptika 7
- aristotelismus 7
- astrologie 8

1443 rows

Extensions: Wikidata

Show as: **rows** records Show: 5 **10** 25 50 100 500 1000 rows « first ‹ previous 1 of 145 pages next › last »

All		Č.	Katalog	Obor/číslo	Fol.	Autor	Název	Datace	D1	D2	Místo vzniku	Jazyk	Ps.I., vazba	Obor
☆	👎	1.	1	Březan	Teologie	10r		Biblí staročeské psané na pergameně (jenž někdy náležela pánům z Kunstátu) dílové dva I. II.	1400-1450	1400	1500	Morava?	ces	perg; ilum, neúpl.; obsahuje i Žďárskou kroniku. Vazba prvního i druhého svazku shodná (převazba rožmberské provenience z doby kolem roku 1608): vazba kožená, dochovány středové puklice, 6 nárožnic, na hřbetu papírové štítky (název díla, starší signatury MZK a Královské knihovny ve Stockholmu).	bible

Export dat

- zkuste pomocí faset (facet) a seskupování (cluster) vyčistit i data v dalších sloupcích

- soubor s vyčištěnými daty uložte ve formátu MS Excel (XSLX)
  - menu s možnostmi exportu je na následujícím snímku

OpenRefine Rozmberska knihovna rozpis xlsx Permalink

Open... Export Help

Facet / Filter | Undo / Redo 0 / 0

Refresh | Reset all | Remove all

Místo vzniku change

38 choices Sort by: **name** count Cluster

- [Třeboň: rožmberský dvůr, Václav Březan písař] 1
- Čechy 34
- Čechy, Praetorius Havel, David - písař 1
- Čechy? 2
- Český Krumlov 1
- Český Krumlov? rožmberský dvůr? 1
- Ferrara 1
- Francie 1
- Itálie 1

Jazyk change

21 choices Sort by: **name** count Cluster

- ces 95
- ces lat 1
- ces, deu, lat 1
- ces, lat 3
- cze 1
- deu 83
- deu lat, ces 1
- deu, lat 3

990 rows

Show as: **rows** records Show: 5 **10** 25 50 100 500 1000 rows « first ‹ pre last »

All	Č.	Katalog	Obor/číslo	Fol.	Autor	Název	Datace	D1	D2	Mo	Obor
1.	1	Březan	Teologie	10r		Biblí staročeské psané na pergameně (jenž někdy náležela pánům z Kunstátu) dílové dva I. II.	1400-1450	1400	1500		

Královské knihovny ve Stockholmu).

Export menu:

- OpenRefine project archive to file
- Tab-separated value
- Comma-separated value
- HTML table
- Excel (.xls)
- Excel 2007+ (.xlsx)
- ODF spreadsheet
- Custom tabular exporter…
- SQL Exporter…
- Templating…
- OpenRefine project archive to Google Drive…
- Google Sheets…
- Wikibase edits…
- QuickStatements file
- Wikibase schema