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

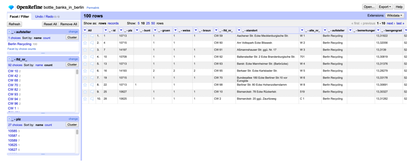
Examples

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

combining data

join datasets or enrich your source with new data

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

bonus

you can write code or use regex (but you don’t have to…)