Introduction to OpenRefine

Objectives

Describe OpenRefine’s uses and applications.

Differentiate data cleaning from data organization.

Experiment with OpenRefine’s user interface.

Locate helpful resources to learn more about OpenRefine.

History of OpenRefine

MetaWeb Technologies / Freebase

- Community curated database of well known people, places, things
- 2009 Freebase Gridworks
- Open source

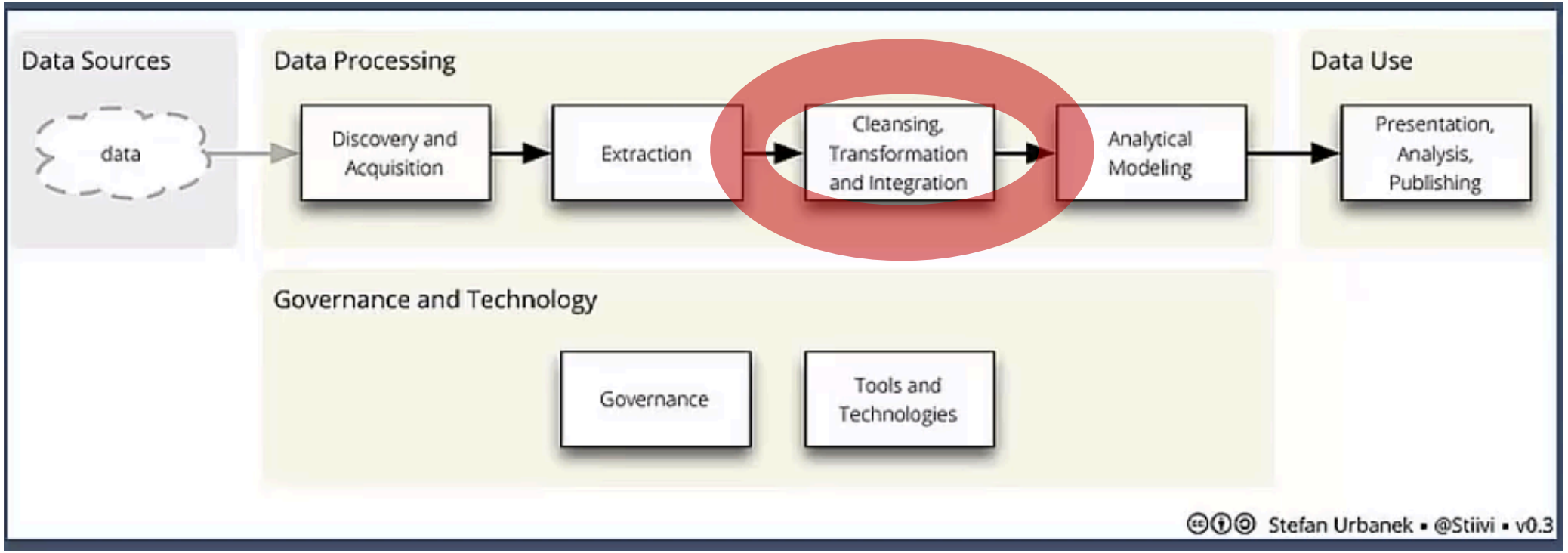
2010: Google acquires MetaWeb

- Integrates Freebase into Google Knowledge Graph
- Gridworks → Google Refine

2012: Google ends support of Google Refine

- Open source, moved to Github as OpenRefine

Data quality & integration



Data quality & integration

First Name	Last Name	Address	Birthday	City	Country	
John	Smith	1 King Street E.	1975-05-15	Toronto	CA	
Jhon	Smith	1 King Street East	15/05/75	Toronto	Canada	
Roger	Baker	33 Bloor East	Sept 12 1980	tornoto	Canada	
Rachel	Figo	25 Market St.	N/A	San	Francisco	USA
Mary and Joel	Green	305 3rd N St.		SAN JOSE	US	

- Duplicate records
- Typ;os
- Cells with multiple fields
- Data in wrong field
- Missing / partial data
- Encoding error

- Format
- Conversion from flat to relational data sets
- Schema alignment
- Transposition
- Join
- Data source enrichment

Data quality & integration

First Name	Last Name	Address	Birthday	City	Country	
John	Smith	1 King Street E.	1975-05-15	Toronto	CA	
Jhon	Smith	1 King Street East	15/05/75	Toronto	Canada	
Roger	Baker	33 Bloor East	Sept 12 1980	tornoto	Canada	
Rachel	Figo	25 Market St.	N/A	San	Francisco	USA
Mary and Joel	Green	305 3rd N St.		SAN JOSE	US	

- Duplicate records
- Typ;os
- Cells with multiple fields
- Data in wrong field
- Missing / partial data
- Encoding error

(That's supposed to be funny.)

- Format
- Conversion from flat to relational data sets
- Schema alignment
- Transposition
- Join
- Data source enrichment

Addressing the skill gap

Programming Software

OpenRefine

Spreadsheet

Database administrators

Data scientists

Programmers

Archivists

Librarians

OpenRefine

Free, open source tool for data cleaning (source on GitHub)

Large, growing community with diverse levels of expertise

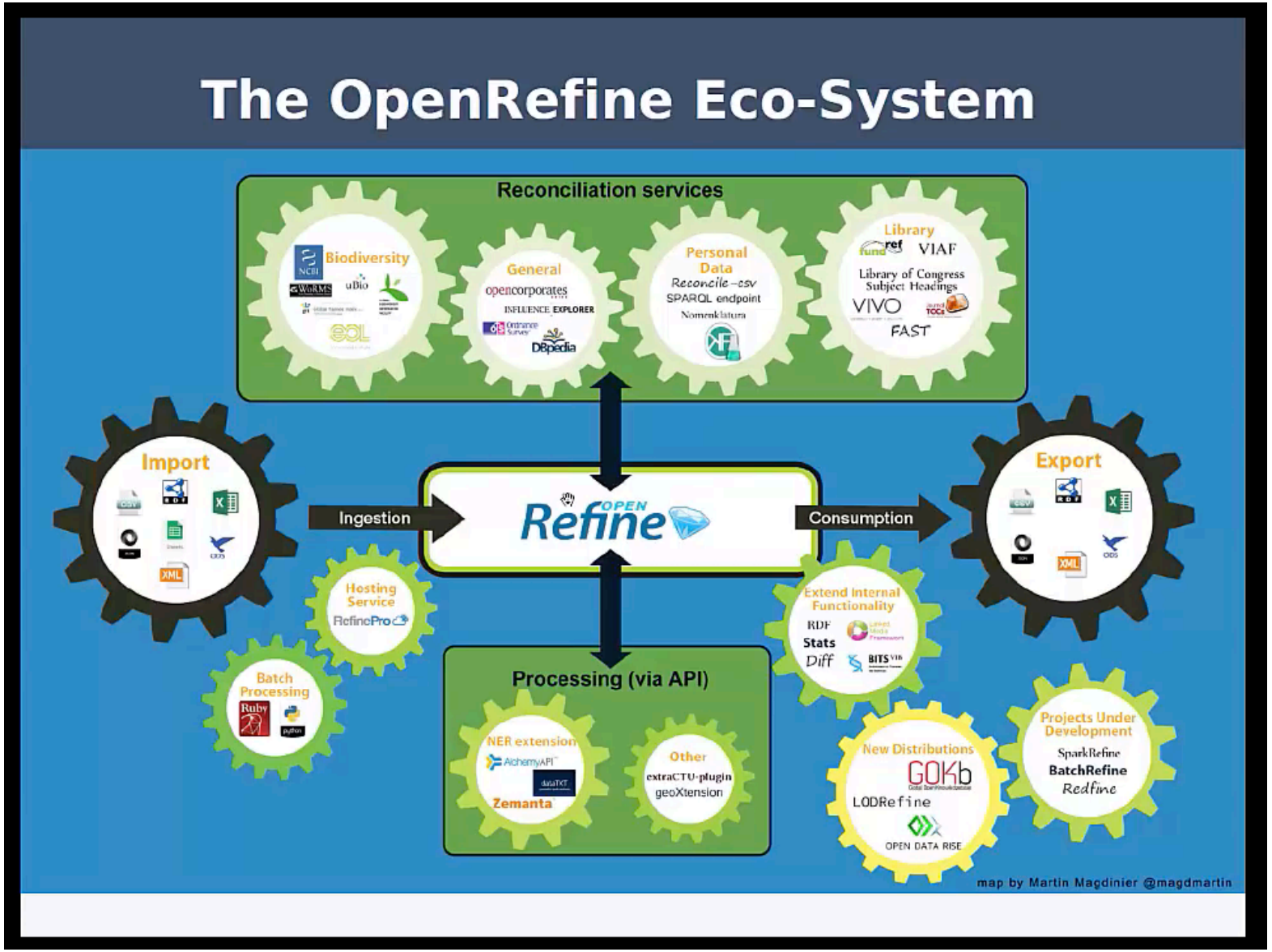
Works with large-ish datasets (100,000 rows).

Adjust memory allocation for larger datasets

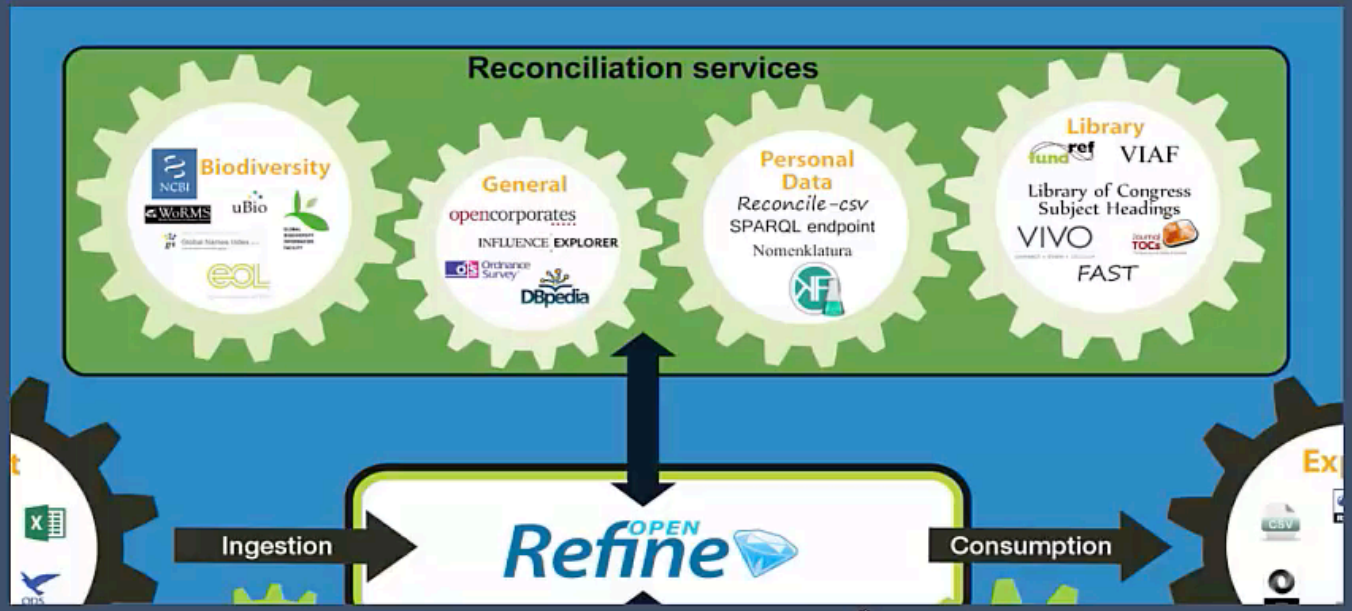
Auto-tracks every change

Unlimited undo

Work log


The OpenRefine Eco-System
Reconciliation services
Biodiversity
NCBI
WoRMS
uBio
eol
General
opencorporates
INFLUENCE EXPLORER
Ordnance Survey
DBpedia
Personal Data
Reconcile-csv
SPARQL endpoint
Nomenklatura
Library
fundref
VIAF
Library of Congress Subject Headings
VIVO
FAST
Import
CSV
RDF
XML
ODS
Ingestion
OpenRefine
Consumption
Export
Hosting Service
RefinePro
Batch Processing
Ruby
python
Processing (via API)
NER extension
AlchemyAPI
dataTXT
Zemanta
Other
extraCTU-plugin
geoXtension
Extend Internal Functionality
RDF
Stats
Diff
BITS VIB
New Distributions
GOKb
LODRefine
OPEN DATA RISE
Projects Under Development
SparkRefine
BatchRefine
Redfine
map by Martin Magdinier @magdmartin

The OpenRefine Eco-System

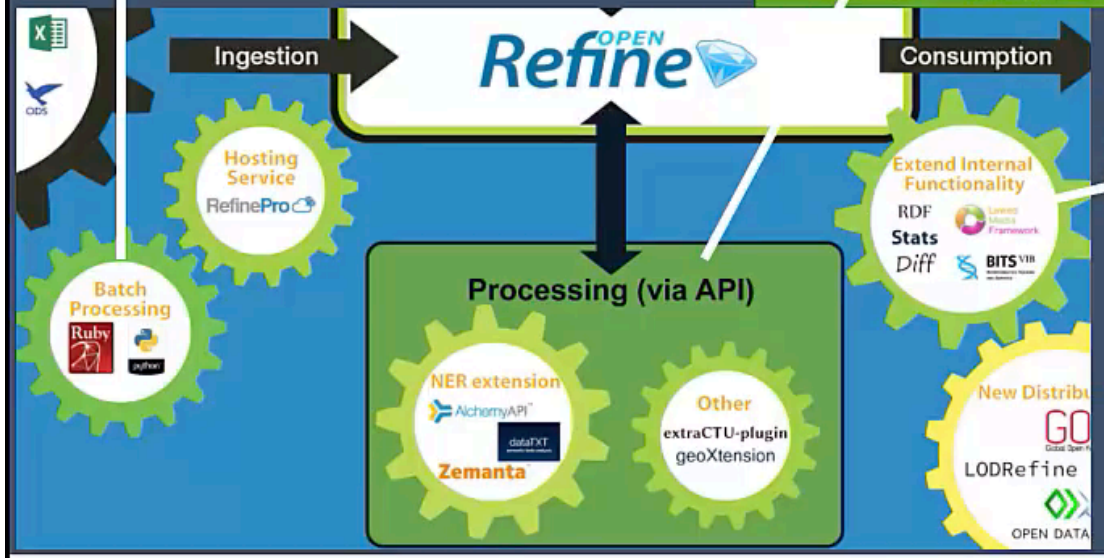


Reconciliation services sit outside of Refine and enable the user to align and enrich data against

The OpenRefine Eco-System

Batch Processing
libraries enable
lightweight ETL process

API processing plugins
enable seamless data
processing
with API-based services



Extensions add
functionality to
Refine's core.

Download

https://ndownloader.figshare.com/files/11502815

Objectives

- Create a new OpenRefine project from a CSV file.
- Understand potential problems with file headers.
- Use facets to summarize data from a column.
- Use clustering to detect possible typing errors.
- Demonstrate that different clustering algorithms give different results.
- Employ drop-downs to remove white spaces from cells.
- Manipulate data using previous steps with undo/redo.

Getting started

Follow the Setup instructions to install OpenRefine.

Launch the program

Open the OpenRefine app on your device	Point your your browser to **http://127.0.0.1:3333/** or **http://localhost:3333**

Note: this is a Java program that runs on your machine (not in the cloud). It runs inside your browser, but no web connection is needed.

Create a new OpenRefine project

Filtering & sorting

How can we select only a subset of our data?

Sort table by column

Sort by multiple columns

- we’ll browse our computer to the sample data file for this lesson. In this case, we will be using data obtained from interviews of farmers in two countries in eastern sub-Saharan Africa (Mozambique and Tanzania). Instructions on downloading the data are available here.

Trim leading & trailing whitespace

Create a new text facet for the column **respondent_wall_type**

Choose Edit cells > Common transforms > Trim leading and trailing whitespace

Key points

OpenRefine can import a variety of file types.

OpenRefine can be used to explore data using filters.

Clustering in OpenRefine can help to identify different values that might mean the same thing.

OpenRefine can transform the values of a column.

Parse data as

- **CSV / TSV / separator-based files**
- Line-based text files
- Fixed-width field text files
- PC-Axis text files
- JSON files
- RDF/N3 files
- XML files
- Open Document Format spreadsheets (.ods)
- RDF/XML files

Character encoding

Update Preview

Columns are separated by

- (•) commas (CSV)
- ( ) tabs (TSV)
- ( ) custom ,

Escape special characters with \

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

- [ ] Parse cell text into numbers, dates, ...
- [x] Quotation marks are used to enclose cells containing column separators
- [x] Store blank rows
- [x] Store blank cells as nulls
- [ ] Store file source (file names, URLs) in each row

Custom text transform on column F14_items_owned

Expression Language General Refine Expression Language (GREL)

```
value
```

No syntax error.

Preview History Starred Help

row	value	value
1.	['bicycle' ; 'television' ; 'solar_panel' ; 'table']	['bicycle' ; 'television' ; 'solar_panel' ; 'table']
2.	['cow_cart' ; 'bicycle' ; 'radio' ; 'cow_plough' ; 'solar_panel' ; 'solar_torch' ; 'table' ; 'mobile_phone']	['cow_cart' ; 'bicycle' ; 'radio' ; 'cow_plough' ; 'solar_panel' ; 'solar_torch' ; 'table' ; 'mobile_phone']
3.	['solar_torch']	['solar_torch']
4.	['bicycle' ; 'radio' ; 'cow_plough' ; 'solar_panel' ; 'mobile_phone']	['bicycle' ; 'radio' ; 'cow_plough' ; 'solar_panel' ; 'mobile_phone']
5.	['motorcyle' ; 'radio' ; 'cow_plough' ; 'mobile_phone']	['motorcyle' ; 'radio' ; 'cow_plough' ; 'mobile_phone']

On error
- (•) keep original
- ( ) set to blank
- ( ) store error

[ ] Re-transform up to 10 times until no change

OK Cancel