Analysis of the OpenRefine data preparation tool

Analisi del tool per la preparazione dei dati OpenRefine

Prof.ssa Sonia Bergamaschi
Dott. Luca Zecchini

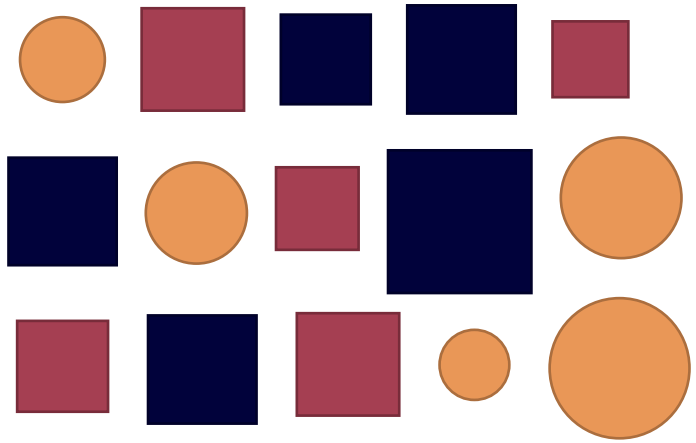
Enrico De Luca

Anno Accademico 2019/2020

DATA PREPARATION

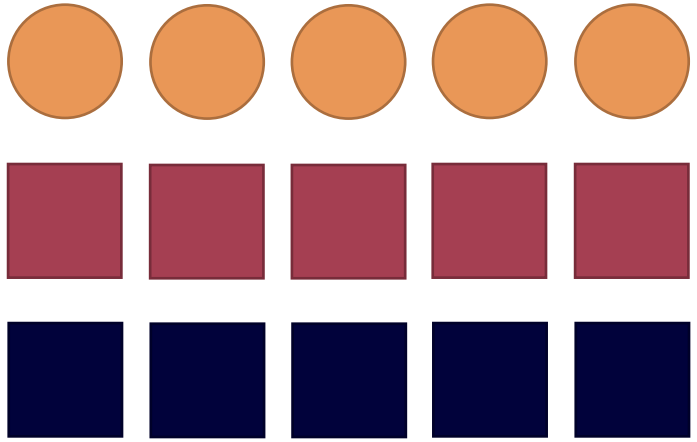
DEFINITIONS & MOTIVATIONS

RAW DATA



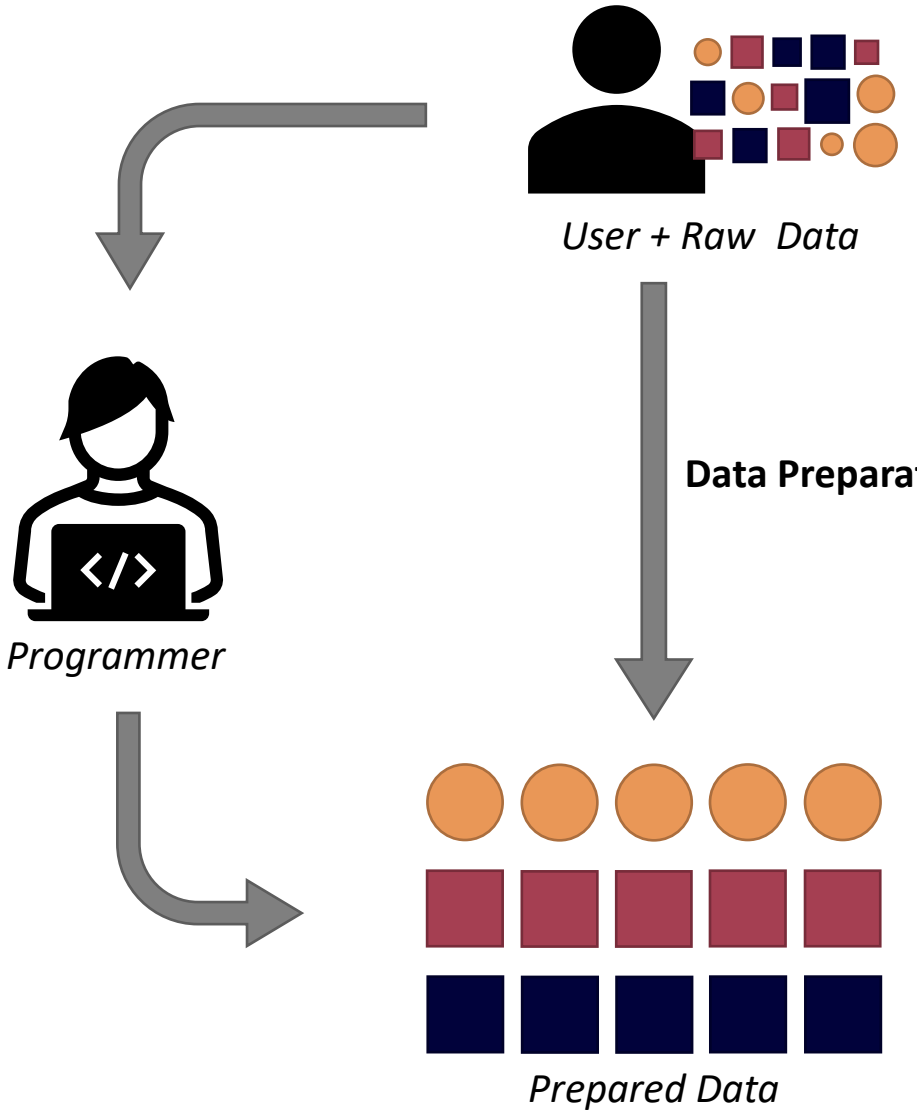
- Produced by different sources
- Unorganized and messy
- Contains errors and inconsistencies

PREPARED DATA



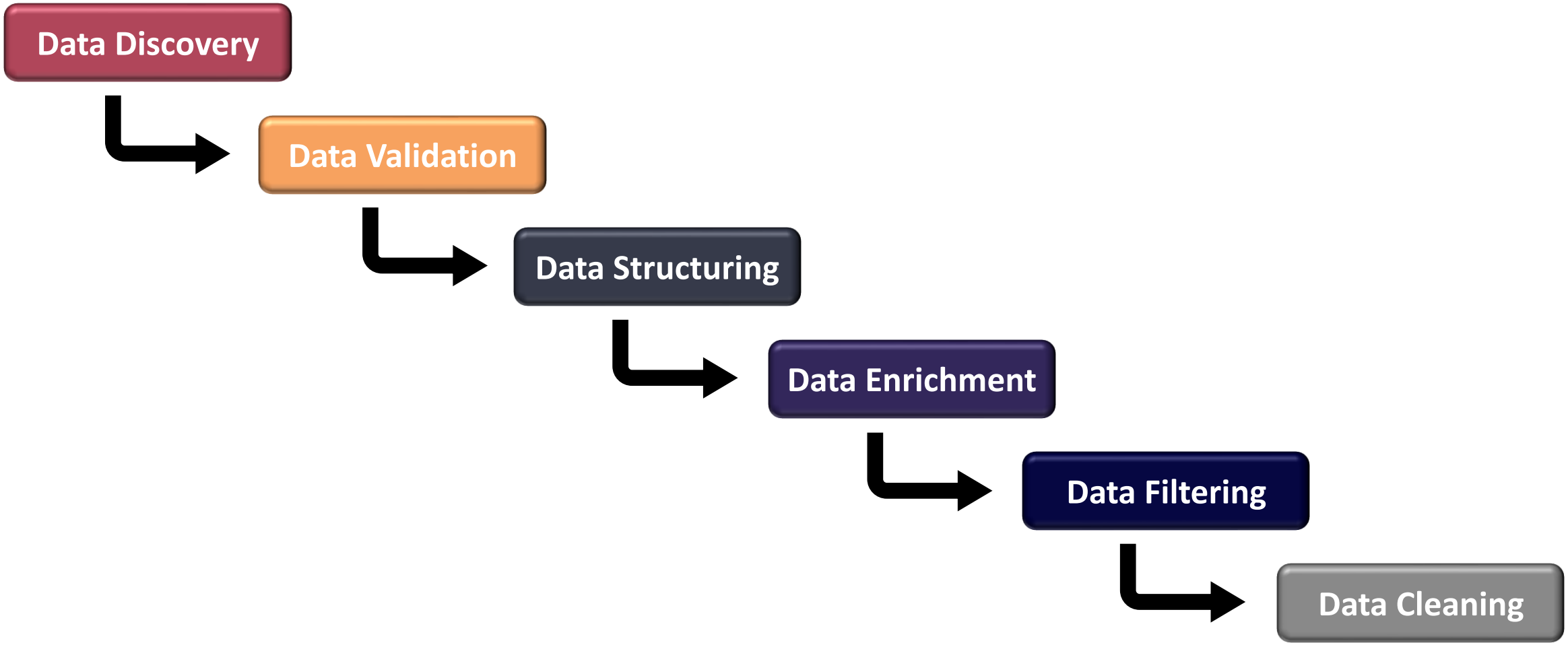
- Organized and structured
- Ready to be used (analysis or applications)

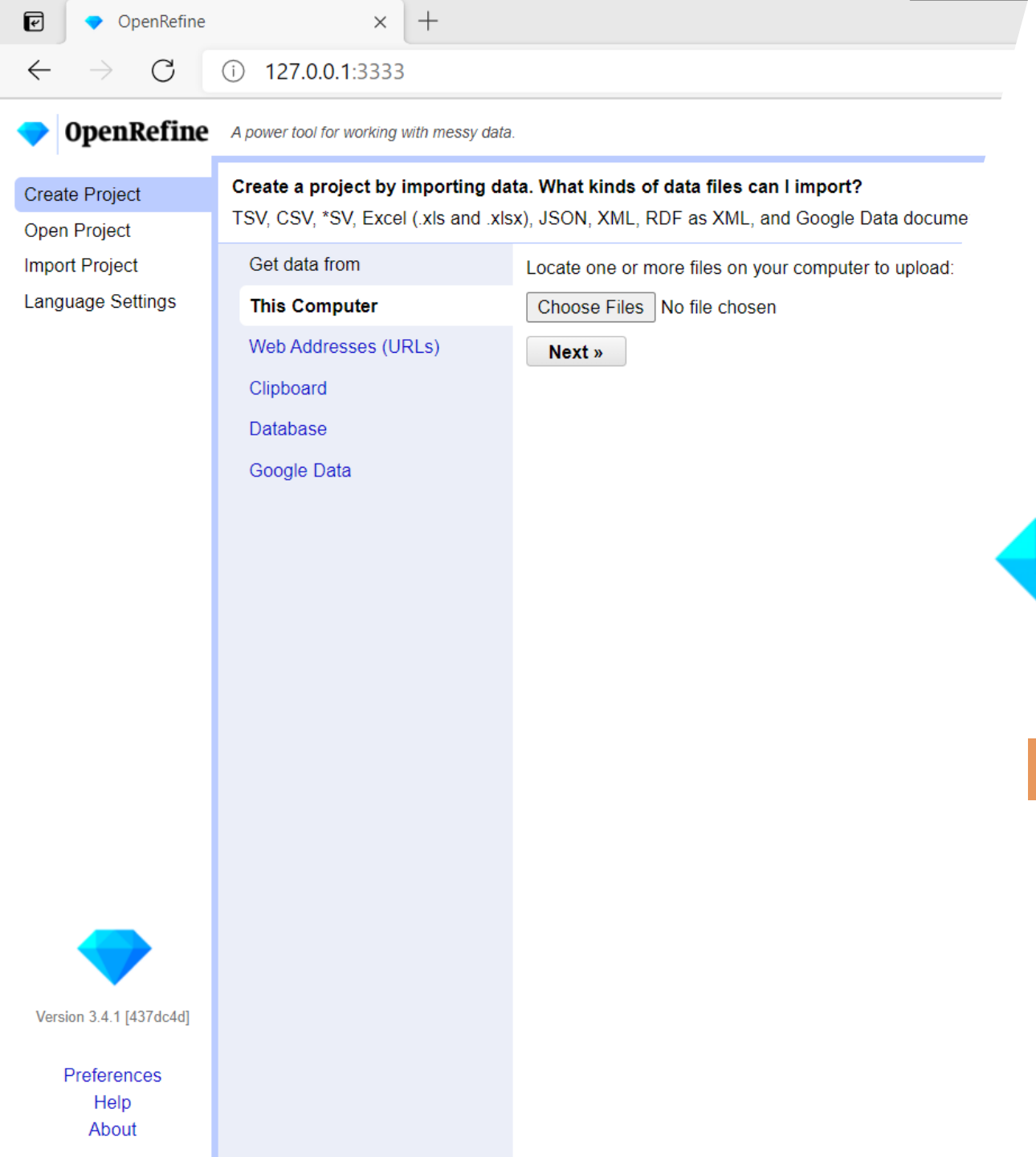
EVOLUTION OF THE APPROACH

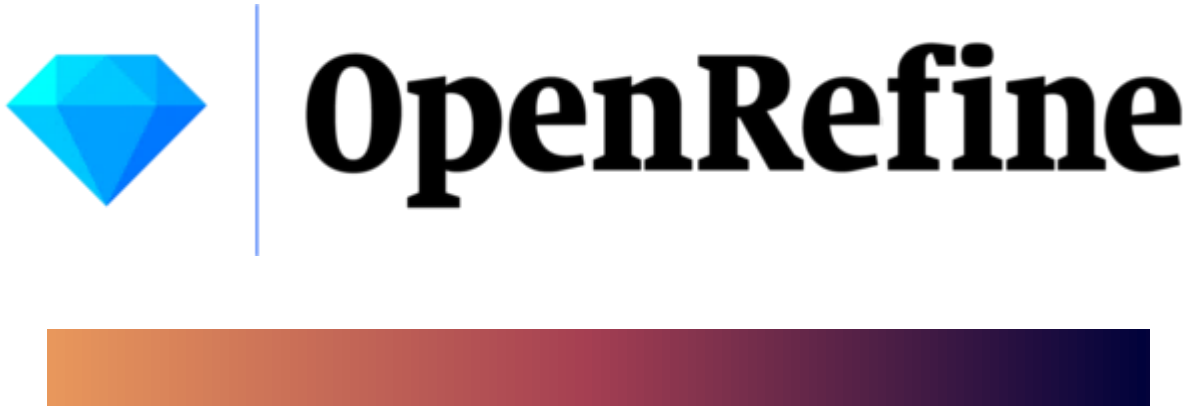


- Altair Monarch Data Preparation
- Paxata Self Service Data Preparation
- SAP Agile Data Preparation
- SAS Data Preparation
- Tableau Prep
- Talend Data Preparation
- Trifacta Wrangler
- **OpenRefine**

DATA PREPARATION TASKS




OpenRefine
127.0.0.1:3333
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 docume
Get data from
This Computer
Web Addresses (URLs)
Clipboard
Database
Google Data
Locate one or more files on your computer to upload:
Choose Files
No file chosen
Next »
Version 3.4.1 [437dc4d]
Preferences
Help
About


OpenRefine

HISTORY

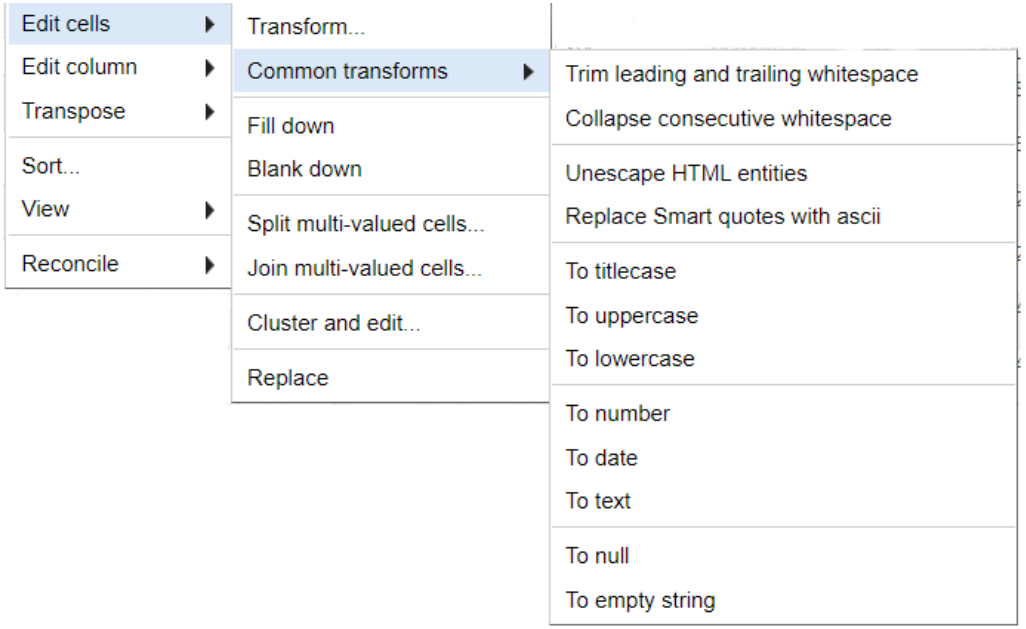
2010	2010	2013
PRODUCT LAUNCH	GOOGLE ACQUISITION	GOOGLE SUPPORT STOPPED
Freebase Gridworks (Metaweb Technologies)	**Google Refine** (Google)	**OpenRefine** (Open-Source Community)

OVERVIEW

- Desktop application for Linux, Mac, Windows
- Local web server
- supported formats: CSV, TSV, text file, spreadsheet, database, etc.
- Extensions
- Privacy is protected

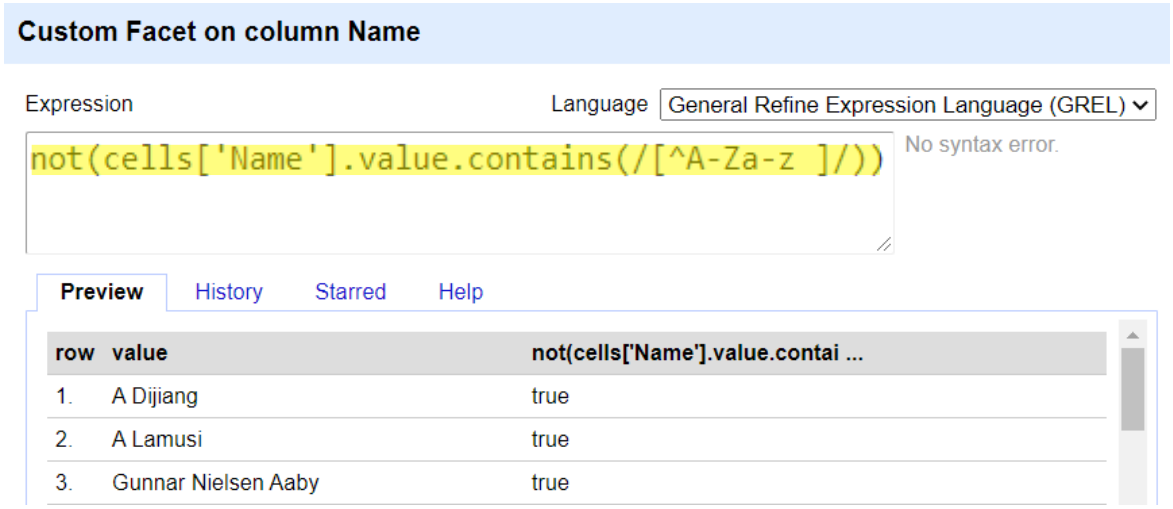
EXPRESSIONS

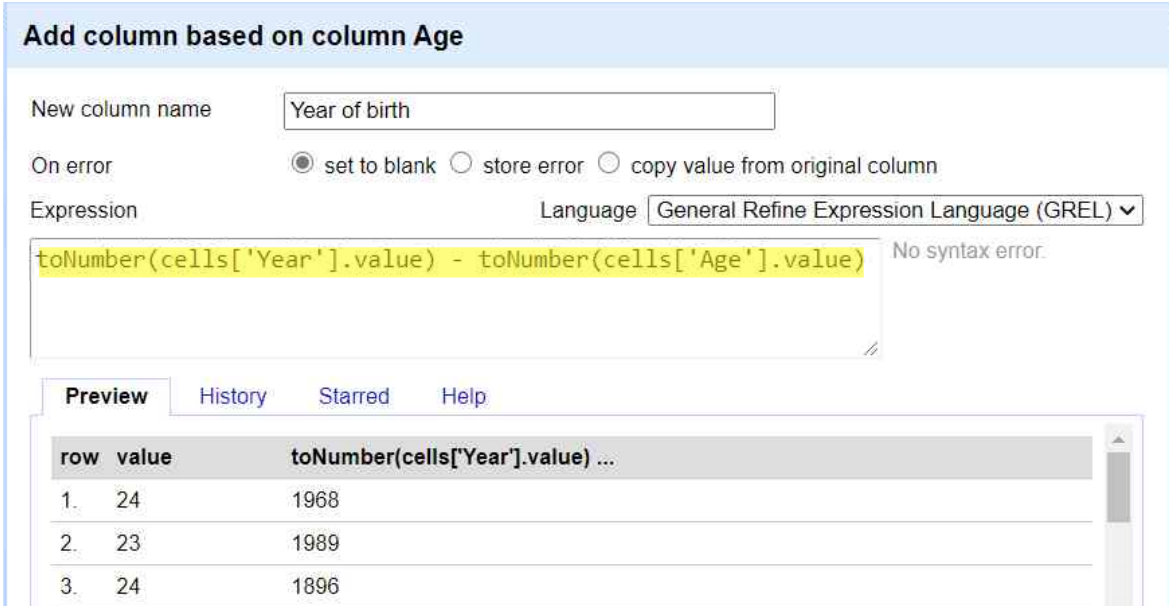
- DESCRIBE THE TRANSFORMATION
- DEFAULT LANGUAGE: GREL (GENERAL REFINE EXPRESSION LANGUAGE)



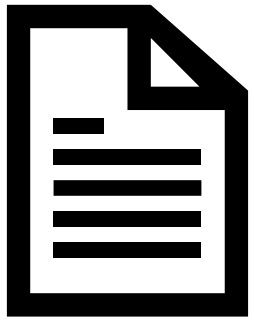
EXPRESSIONS

- DESCRIBE THE TRANSFORMATION
- DEFAULT LANGUAGE: GREL (GENERAL REFINE EXPRESSION LANGUAGE)



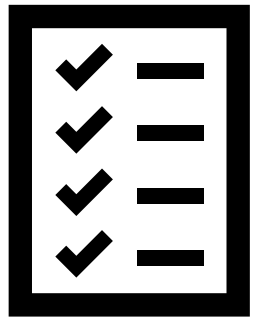


EXPERIMENTAL EVALUATION



Survey Paper
“Data Preparation: A Survey of Commercial Tools”
by *Mazhar Hameed* and *Felix Naumann*
(Hasso Plattern Institute,
University of Potsdam)

Search for documentation



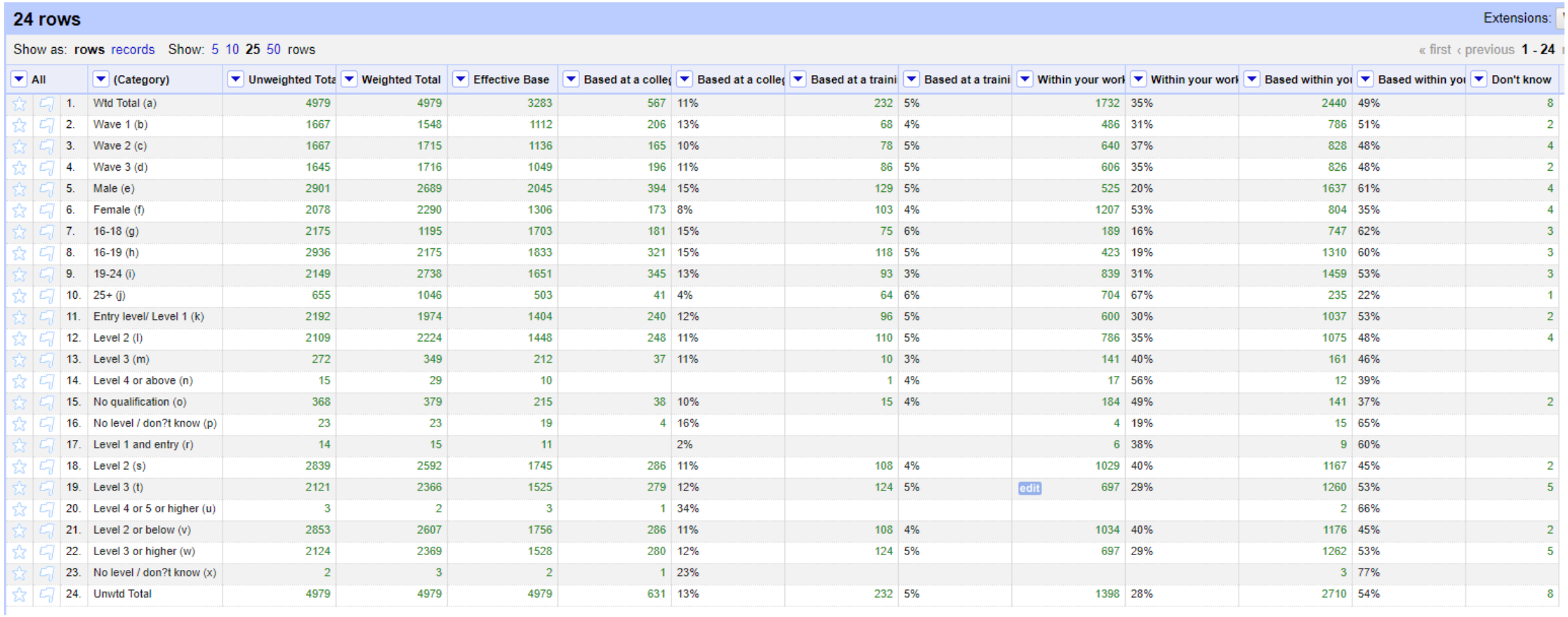
Test OpenRefine Functionalities

BEFORE

	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10	Column 11	Column 12	Column 13	Column 14	Column 15	Column 16	Column 17	Column 18	Column 19	Column 20	Column 21	Column 22	Column 23	Column 24	Column 25
1.													National Learner Satisfaction Survey 2009												
2.													(FINAL - ALL LEARNERS)												
3.																									Table 1
4.	Q.1. Please think about any time away from your day-to-day job that you spend in training. Is your training ...?																								
5.	Base : All apprentices																								
6.																									
7.			Wave			Gender		Age				Highest Prior Level						Current Level							
8.		Wtd Total (a)	Wave 1 (b)	Wave 2 (c)	Wave 3 (d)	Male (e)	Female (f)	16-18 (g)	16-19 (h)	19-24 (i)	25+ (j)	Entry level/ Level 1 (k)	Level 2 (l)	Level 3 (m)	Level 4 or above (n)	No qualification (o)	No level / don?t know (p)	Level 1 and entry (r)	Level 2 (s)	Level 3 (t)	Level 4 or 5 or higher (u)	Level 2 or below (v)	Level 3 or higher (w)	No level / don?t know (x)	Unwtd Total
9.	Unweighted Total	4979	1667	1667	1645	2901	2078	2175	2936	2149	655	2192	2109	272	15	368	23	14	2839	2121	3	2853	2124	2	4979
10.	Weighted Total	4979	1548	1715	1716	2689	2290	1195	2175	2738	1046	1974	2224	349	29**	379	23**	15**	2592	2366	2**	2607	2369	3**	4979
11.	Effective Base	3283	1112	1136	1049	2045	1306	1703	1833	1651	503	1404	1448	212	10	215	19	11	1745	1525	3	1756	1528	2	4979
12.	Based at a college only	567	206	165	196	394	173	181	321	345	41	240	248	37	-	38	4	*	286	279	1	286	280	1	631
13.		11%cf	13%ac	10%	11%	15%af	8%	15%aij	15%aij	13%aj	4%	12%	11%	11%	-	10%	16%	2%	11%	12%	34%	11%	12%	23%	13%
14.		j																							
15.	Based at a training provider only	232	68	78	86	129	103	75	118	93	64	96	110	10	1	15	-	-	108	124	-	108	124	-	232
16.		5%i	4%	5%	5%	5%	4%	6%ahi	5%ai	3%	6%i	5%	5%	3%	4%	4%	-	-	4%	5%	-	4%	5%	-	5%
17.	Within your workplace only	1732	486	640	606	525	1207	189	423	839	704	600	786	141	17	184	4	6	1029	697	-	1034	697	-	1398
18.		35%be	31%	37%ab	35%	20%	53%ae	16%	19%g	31%gh	67%aghi	30%	35%k	40%k	56%	49%ak	19%	38%	40%at	29%	-	40%atw	29%	-	28%
19.		ghikt														l			w						
20.		w																							
21.	Based within your workplace and at a college or training provider	2440	786	828	826	1637	804	747	1310	1459	235	1037	1075	161	12	141	15	9	1167	1260	2	1176	1262	3	2710
22.		49%fj	51%	48%	48%	61%af	35%	62%ahi	60%aij	53%aj	22%	53%al	48%o	46%	39%	37%	65%	60%	45%	53%asv	66%	45%	53%as	77%	54%
23.		osv						j				o											v		
24.	Don't know	8	2	4	2	4	4	3	3	3	1	2	4	-	-	2	-	-	2	5	-	2	5	-	8
25.		*	*	*	*	*	*	*	h *	*	*	*	*	-	-	1%	-	-	*	*	-	*	*	-	*
26.																									
27.	Fieldwork dates : 17 February 2009 - 31 July 2009																								
28.	Respondent Type : Learners																								
29.	Source : Ipsos MORI (J34262)																								
30.	*=Less than 0.5 %																								
31.	Proportions/Means: Columns Tested (5% risk level) - a/b/c/d - a/e/f - a/g/h/i/j - a/k/l/m/n/o/p - a/r/s/t/u/v/w/x																								
32.	* small base; ** very small base (under 30) ineligible for sig testing																								
33.																									

UK government web archive – Research on how much time employees spend on training besides their 9-5 job (Q1 Table 1)

AFTER



24 rows

Extensions:

Show as: **rows** records Show: 5 10 **25** 50 rows « first ‹ previous **1 - 24**

	All	(Category)	Unweighted Tota	Weighted Total	Effective Base	Based at a colleg	Based at a colleg	Based at a traini	Based at a traini	Within your work	Within your work	Based within you	Based within you	Don't know
1.		Wtd Total (a)	4979	4979	3283	567	11%	232	5%	1732	35%	2440	49%	8
2.		Wave 1 (b)	1667	1548	1112	206	13%	68	4%	486	31%	786	51%	2
3.		Wave 2 (c)	1667	1715	1136	165	10%	78	5%	640	37%	828	48%	4
4.		Wave 3 (d)	1645	1716	1049	196	11%	86	5%	606	35%	826	48%	2
5.		Male (e)	2901	2689	2045	394	15%	129	5%	525	20%	1637	61%	4
6.		Female (f)	2078	2290	1306	173	8%	103	4%	1207	53%	804	35%	4
7.		16-18 (g)	2175	1195	1703	181	15%	75	6%	189	16%	747	62%	3
8.		16-19 (h)	2936	2175	1833	321	15%	118	5%	423	19%	1310	60%	3
9.		19-24 (i)	2149	2738	1651	345	13%	93	3%	839	31%	1459	53%	3
10.		25+ (j)	655	1046	503	41	4%	64	6%	704	67%	235	22%	1
11.		Entry level/ Level 1 (k)	2192	1974	1404	240	12%	96	5%	600	30%	1037	53%	2
12.		Level 2 (l)	2109	2224	1448	248	11%	110	5%	786	35%	1075	48%	4
13.		Level 3 (m)	272	349	212	37	11%	10	3%	141	40%	161	46%	
14.		Level 4 or above (n)	15	29	10			1	4%	17	56%	12	39%	
15.		No qualification (o)	368	379	215	38	10%	15	4%	184	49%	141	37%	2
16.		No level / don?t know (p)	23	23	19	4	16%			4	19%	15	65%	
17.		Level 1 and entry (r)	14	15	11		2%			6	38%	9	60%	
18.		Level 2 (s)	2839	2592	1745	286	11%	108	4%	1029	40%	1167	45%	2
19.		Level 3 (t)	2121	2366	1525	279	12%	124	5%	edit 697	29%	1260	53%	5
20.		Level 4 or 5 or higher (u)	3	2	3	1	34%					2	66%	
21.		Level 2 or below (v)	2853	2607	1756	286	11%	108	4%	1034	40%	1176	45%	2
22.		Level 3 or higher (w)	2124	2369	1528	280	12%	124	5%	697	29%	1262	53%	5
23.		No level / don?t know (x)	2	3	2	1	23%					3	77%	
24.		Unwtd Total	4979	4979	4979	631	13%	232	5%	1398	28%	2710	54%	8

UK government web archive – Research on how much time employees spend on training besides their 9-5 job (Q1 Table 1)

BEFORE

Games	Year	Season
1992 Summer	1992	Summer
2012 Summer	2012	Summer
1920 Summer	1920	Summer
1900 Summer	1900	Summer
1988 Winter	1988	Winter
	1988	Winter
1992 Winter	1992	Winter
	1992	Winter
1994 Winter	1994	Winter
	1994	Winter
1992 Winter	1992	Winter

AFTER

Games	country	edition number	Year	Season
1992 Summer Olympics Choose new match	Spain Choose new match	25	1992	Summer
2012 Summer Olympics Choose new match	United Kingdom Choose new match	30	2012	Summer
1920 Summer Olympics Choose new match	Belgium Choose new match	7	1920	Summer
1900 Summer Olympics Choose new match	France Choose new match	2	1900	Summer
1988 Winter Olympics Choose new match	Canada Choose new match	15	1988	Winter
			1988	Winter
1992 Winter Olympics Choose new match	France Choose new match	16	1992	Winter
			1992	Winter
1994 Winter Olympics Choose new match	Norway Choose new match	17	1994	Winter
			1994	Winter
1992 Winter Olympics	France Choose new	16	1992	Winter

Data Enrichment Example

Kaggle – 120 years of Olympic history: athletes and results

RESULTS

OpenRefine Functionalities

Categories	ID	Preparators	OpenRefine
Data Discovery	I.a	Locate missing values (nulls)	✓
	I.b	Locate outliers	✓
	I.c	Search by pattern	✓
	I.d	Sort data	✓
Data Validation	II.a	Compare values	✓
	II.b	Check data range	✓
	II.c	Check permitted characters	✓
	II.d	Check column uniqueness	✓
	II.e	Find type-mismatched data	X
	II.f	Find data-mismatched datatypes	X
Data Structuring	III.a	Change column data type	X
	III.b	Delete column	✓
	III.c	Detect & change encoding	✓
	III.d	Pivot / unpivot	✓
	III.e	Rename column	✓
	III.f	Split column	✓
	III.g	Transform by example	X
Data Enrichment	IV.a	Assign semantic data type	X
	IV.b	Calculate column using expressions	✓
	IV.c	Discover & merge external data	✓
	IV.d	Duplicate column	✓
	IV.e	Generate primary key column	✓
	IV.f	Join & union	✓
	IV.g	Merge columns	✓
	IV.h	Normalize numeric values	X
Data Filtering	V.a	Delete/keep filtered rows	✓
	V.b	Delete empty and invalid rows	✓
	V.c	Extract value parts	✓
	V.d	Filter with regular expressions	✓
Data Cleaning	VI.a	Change date & time format	✓
	VI.b	Change letter case	✓
	VI.c	Change number format	X
	VI.d	Deduplicate data	✓
	VI.e	Delete by pattern	✓
	VI.f	Edit & replace cell data	✓
	VI.g	Fill empty cells	✓
	VI.h	Remove extra whitespace	✓
	VI.i	Remove diacritics	✓
	VI.j	Standardize strings by pattern	✓
	VI.k	Standardize values in clusters	✓

OpenRefine vs Other Tools


THANK YOU