DAPAAS

Data- and Platform-as-a-Service

Data Publishing with DaPaaS

~ Data-as-a-Service for Open Data ~

@ ALLDATA
April 23, 2015

http://dapaas.eu/

Dumitru Roman, SINTEF, Norway

What can open data do for you?

(Source: The ODI, https://vimeo.com/110800848)

Outline statements

- Open Data

  … is changing the nature of business and reflects a cultural shift to an open society

- Linked Data

  … is great technology for Open Data but has been ignored by the mainstream

- Data-as-a-Service (DaaS)

  … is emerging as a cost-effective solution for publishing and consuming Linked Open Data

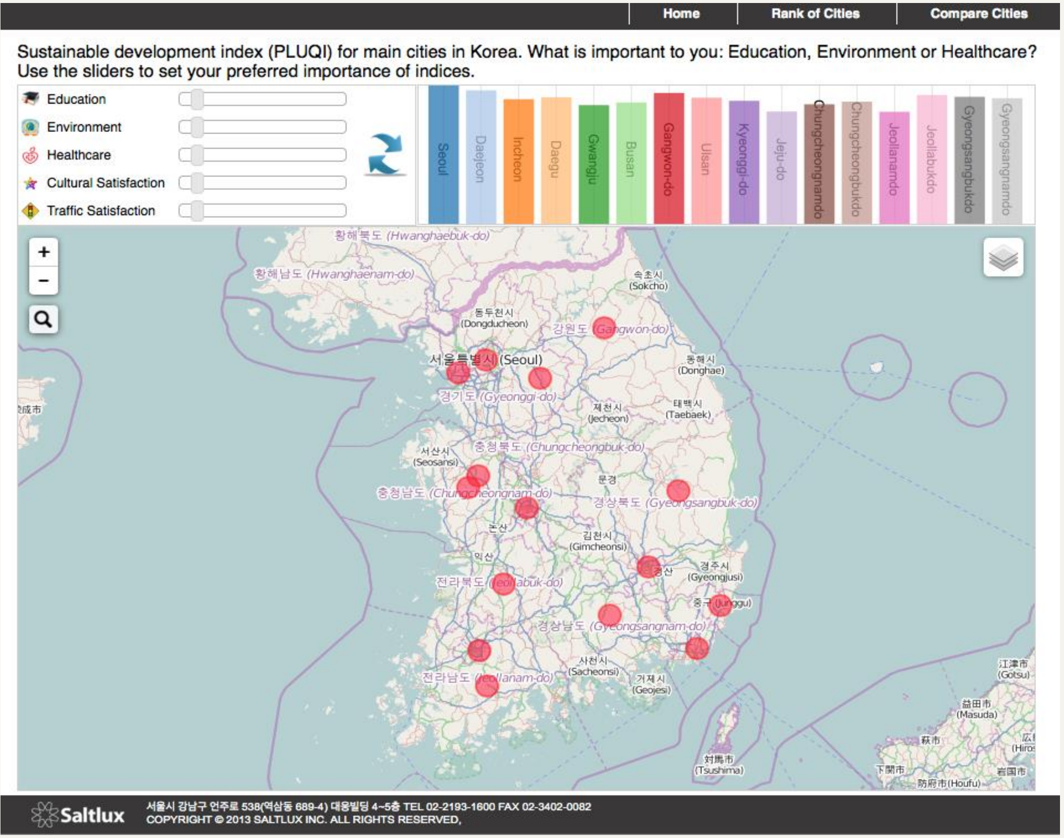
  … DaPaaS: an emerging solution for DaaS

Case study: PLUQI

PLUQI: Personalized and Localized Urban Quality Index
isA
Application (mobile/Web) showing a customizable index that represents and visualize the level of well-being and sustainability for given cities based on individual preferences.

The index model includes **various domains**:

Daily life satisfaction: weather, transportation, community etc.;
Healthcare level: number of doctors, hospitals, suicide statistics, etc.;
Safety and security: number of police stations, fire stations, crimes per capita, etc.;
Financial satisfaction: prices, incomes, housing, savings, debt, insurance, pension, etc.;
Level of opportunity: jobs, unemployment education, re-education, economic dynamics, etc.;
Environmental needs and efficiency: green space, air quality, etc.;



Sustainable development index (PLUQI) for main cities in Korea. What is important to you: Education, Environment or Healthcare? Use the sliders to set your preferred importance of indices.

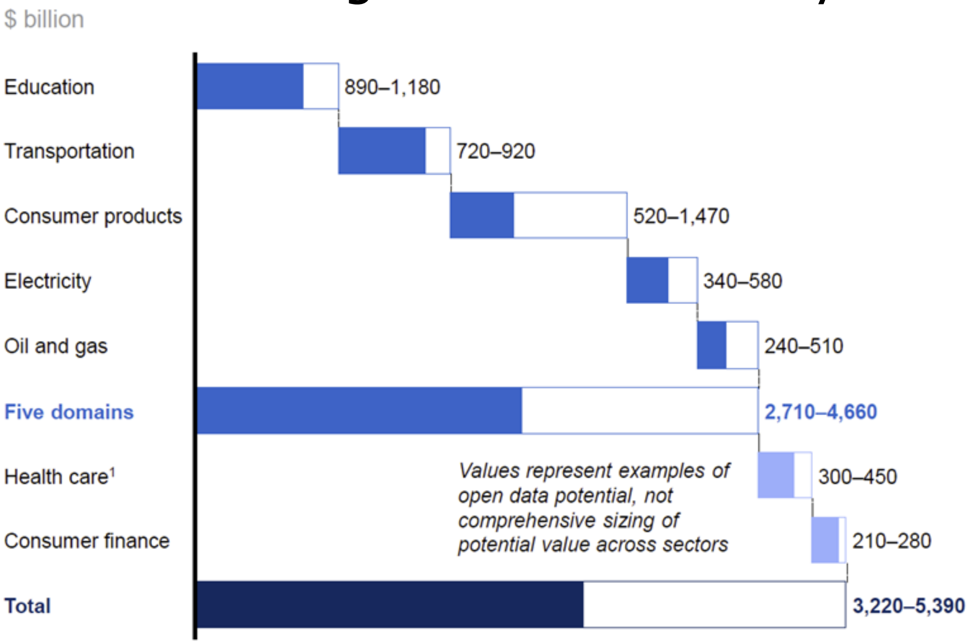
서울시 강남구 언주로 538(역삼동 689-4) 대웅빌딩 4~5층 TEL 02-2193-1600 FAX 02-3402-0082
COPYRIGHT © 2013 SALTLUX INC. ALL RIGHTS RESERVED,

PLUQI – potential usage

- Place recommendation for travel agencies or travelers
- Policy analysis and optimization for (local) government
- Understanding the citizen’s voice and demands regarding environmental conservation
- Commercial impact analysis for retailer and franchises
- Location recommendation and understanding local issues for real estate
- Risk analysis and management for insurance and financial companies
- Local marketing and sales force optimization for marketers

Open Data

- *Businesses* can develop new ideas, services and applications; improve decision making, cost savings
- Can increase *government* transparency and accountability, quality of public services
- *Citizens* get better and timely access to public services



Values represent examples of open data potential, not comprehensive sizing of potential value across sectors

Source: McKinsey

http://www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information

Gartner:

By 2016, the use of "open data" will continue to increase — but slowly, and predominantly limited to Type A enterprises.

By 2017, over 60% of government open data programs that do not effectively use open data internally, will be scaled back or discontinued.

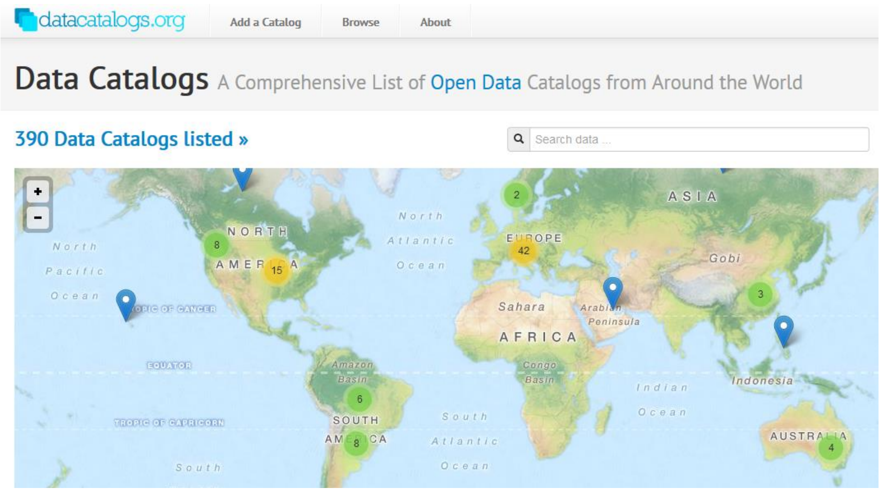
By 2020, enterprises and governments will fail to protect 75% of sensitive data and will declassify and grant broad/public access to it.

Source: Garner

http://training.gsn.gov.tw/uploads/news/6.Gartner+ExP+Briefing_Open+Data_JUN+2014_v2.pdf

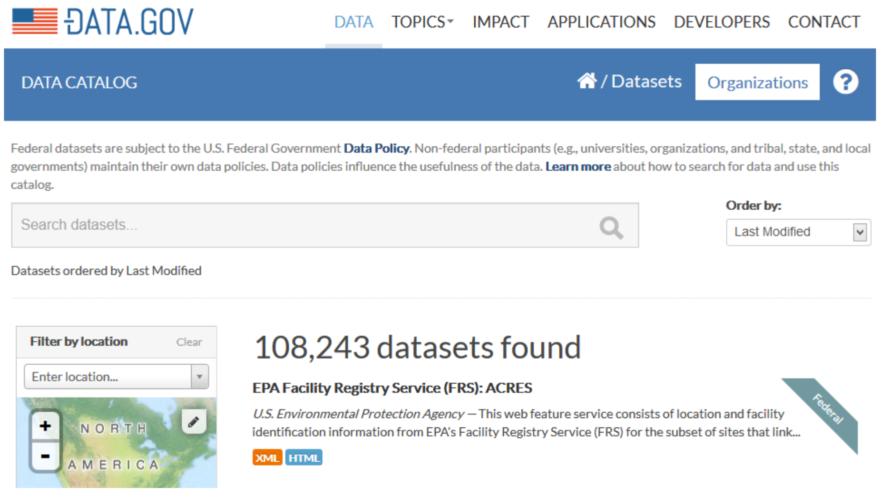
Lots of open datasets on the Web…

- A large number of datasets have been published as open data in the recent years



Data Catalogs A Comprehensive List of Open Data Catalogs from Around the World

390 Data Catalogs listed »



Federal datasets are subject to the U.S. Federal Government Data Policy. Non-federal participants (e.g., universities, organizations, and tribal, state, and local governments) maintain their own data policies. Data policies influence the usefulness of the data. Learn more about how to search for data and use this catalog.

Datasets ordered by Last Modified

108,243 datasets found

EPA Facility Registry Service (FRS): ACRES

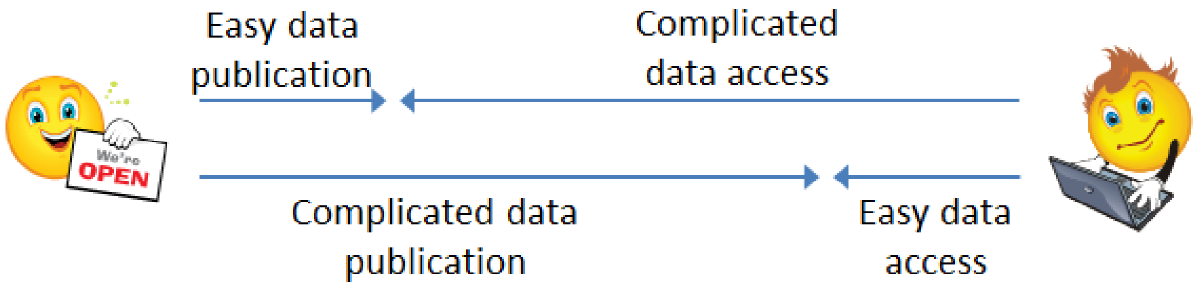
U.S. Environmental Protection Agency — This web feature service consists of location and facility identification information from EPA's Facility Registry Service (FRS) for the subset of sites that link...

- Many kinds of data: cultural, science, finance, statistics, transport, environment, …
- Popular formats: tabular (e.g. CSV, XLS), HTML, XML, JSON, …

…but few actually used

- Few applications utilizing open and distributed datasets at present
- **Challenges for data consumers**
  - Data quality issues
  - Difficult or unreliable data access
  - Licensing issues
- **Challenges for data publishers**
  - Lack of expertise & resources: not easily to publish & maintain high quality data
  - Unclear monetization & sustainability

Open Data Portal	Datasets	Applications
data.gov	~ 110 000	~ 350
publicdata.eu	~ 50 000	~ 80
data.gov.uk	~ 20 000	~ 350
data.norge.no	~ 300	~ 40



Easy data publication

Complicated data access

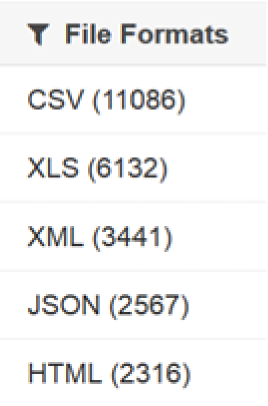
Complicated data publication

Easy data access

Open Data is mostly tabular data

Tabular datasets

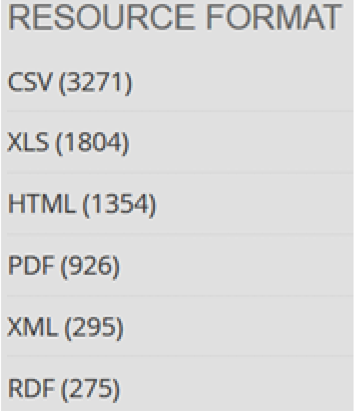
publicdata.eu



File Formats

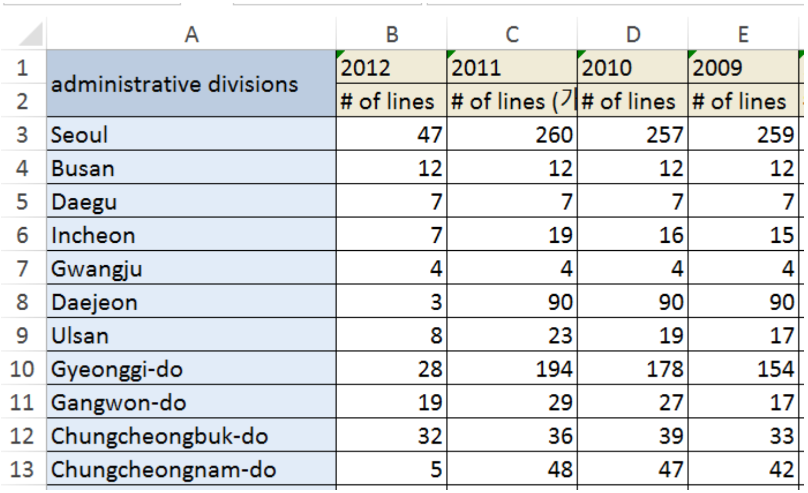
- CSV (11086)
- XLS (6132)
- XML (3441)
- JSON (2567)
- HTML (2316)

data.gov.uk



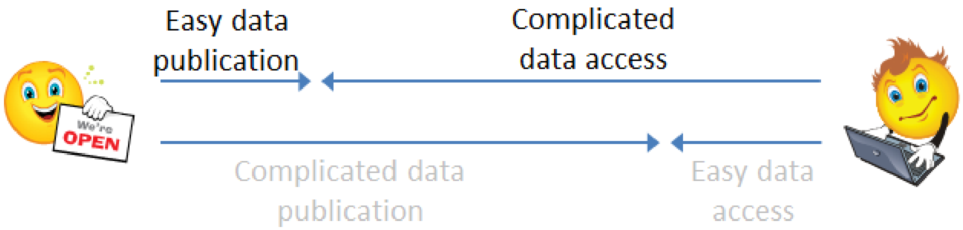
RESOURCE FORMAT

- CSV (3271)
- XLS (1804)
- HTML (1354)
- PDF (926)
- XML (295)
- RDF (275)



	A	B	C	D	E
1	administrative divisions	2012	2011	2010	2009
2		# of lines	# of lines (7	# of lines	# of lines
3	Seoul	47	260	257	259
4	Busan	12	12	12	12
5	Daegu	7	7	7	7
6	Incheon	7	19	16	15
7	Gwangju	4	4	4	4
8	Daejeon	3	90	90	90
9	Ulsan	8	23	19	17
10	Gyeonggi-do	28	194	178	154
11	Gangwon-do	19	29	27	17
12	Chungcheongbuk-do	32	36	39	33
13	Chungcheongnam-do	5	48	47	42

- Records organized in silos of collections
- Very few links within and/or across collections
- Difficult to understand the nature of the data
- Difficult to integrate / query



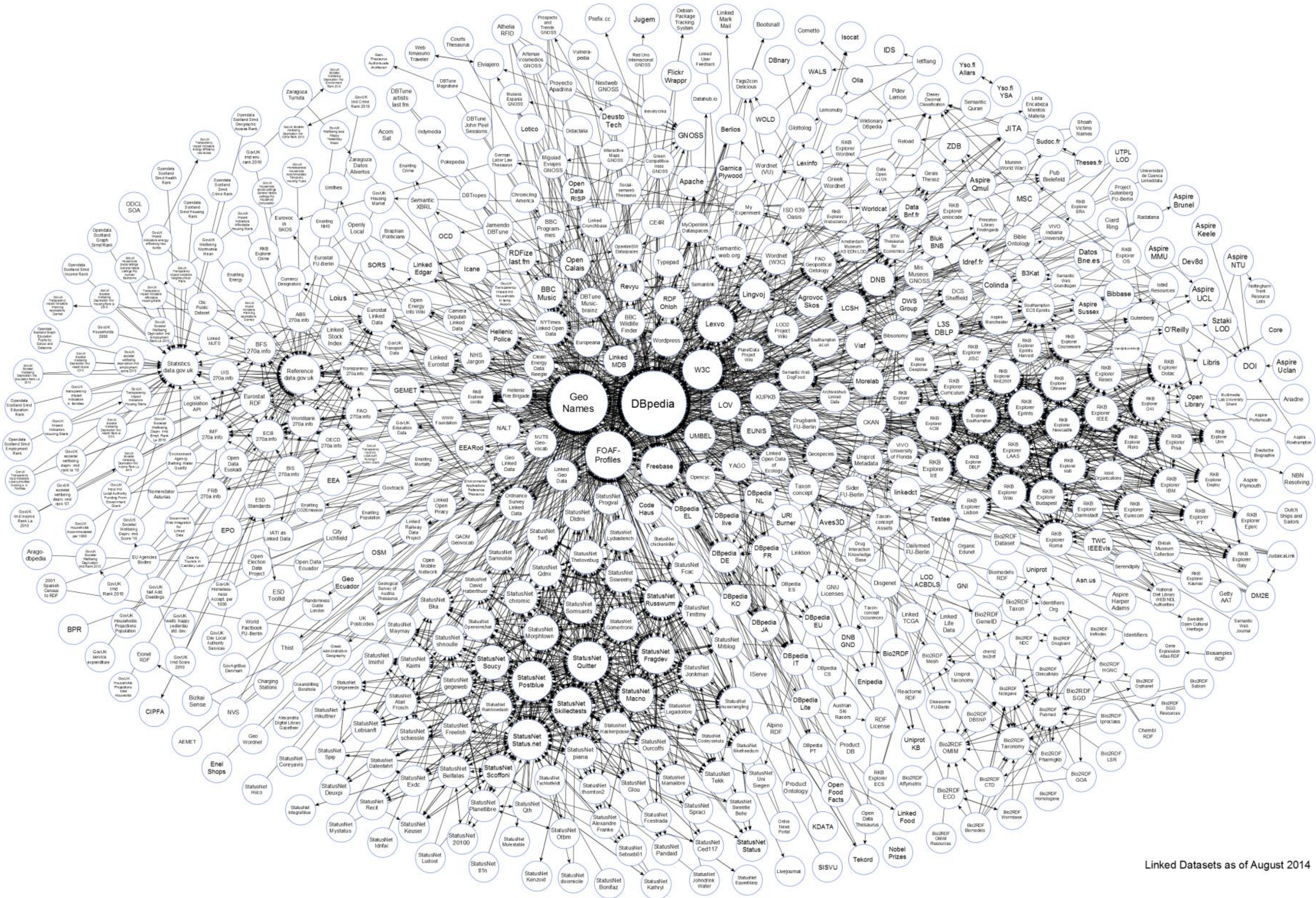
Linked Data

- Method for **publishing data on the Web**
- **Self-describing** data and relations
- **Interlinking**
- Accessed using **semantic queries**

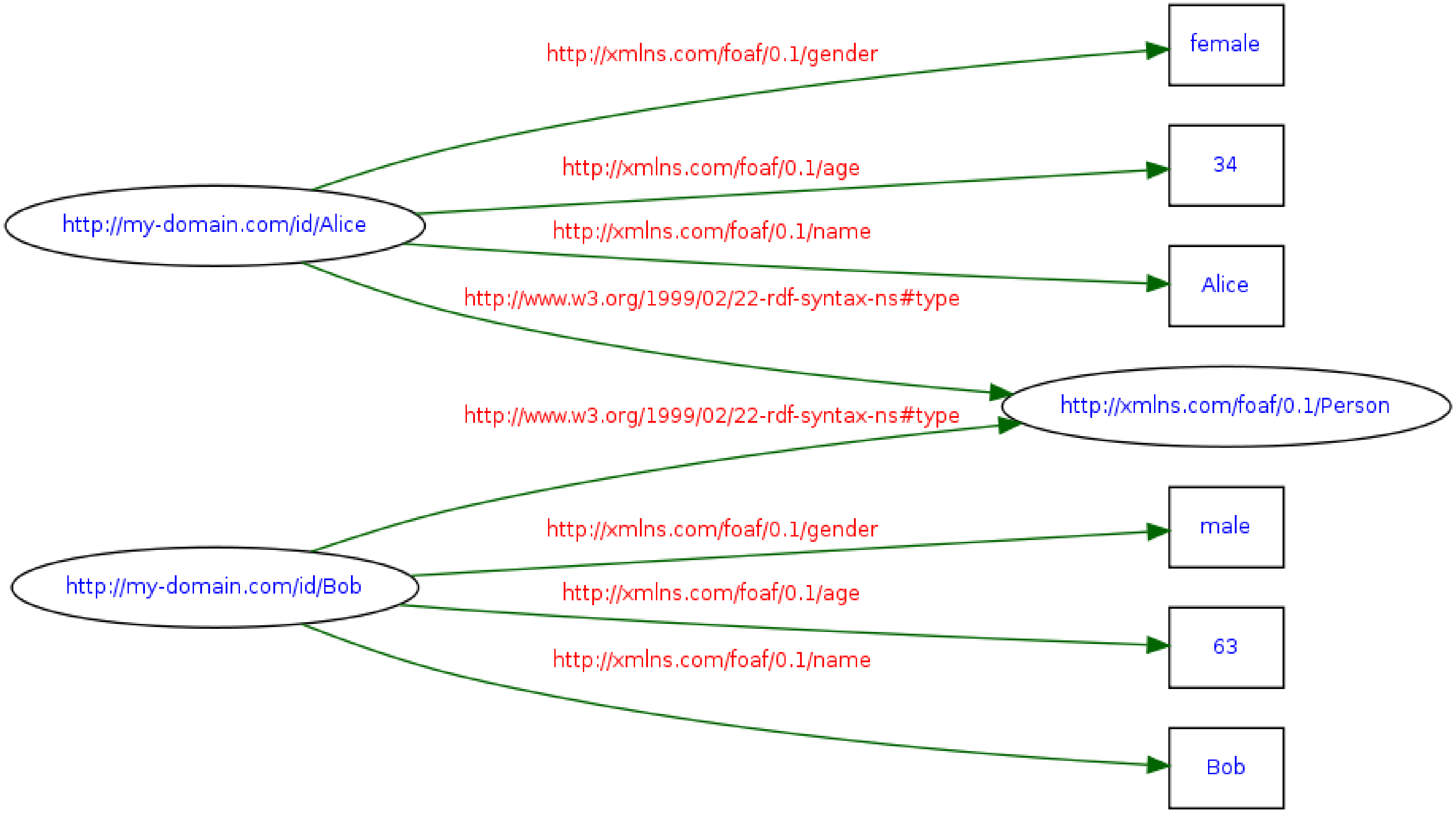
http://www.w3.org/standards/semanticweb/data

- A set of standards developed by W3C
  - Data format: RDF
  - Knowledge representation: RDFS/OWL
  - Query language: SPARQL
  - Linking medium: HTTP

Linked Open Data Cloud



Example



http://my-domain.com/id/Alice

http://xmlns.com/foaf/0.1/gender

female

http://xmlns.com/foaf/0.1/age

34

http://xmlns.com/foaf/0.1/name

Alice

http://www.w3.org/1999/02/22-rdf-syntax-ns#type

http://xmlns.com/foaf/0.1/Person

http://www.w3.org/1999/02/22-rdf-syntax-ns#type

http://my-domain.com/id/Bob

http://xmlns.com/foaf/0.1/gender

male

http://xmlns.com/foaf/0.1/age

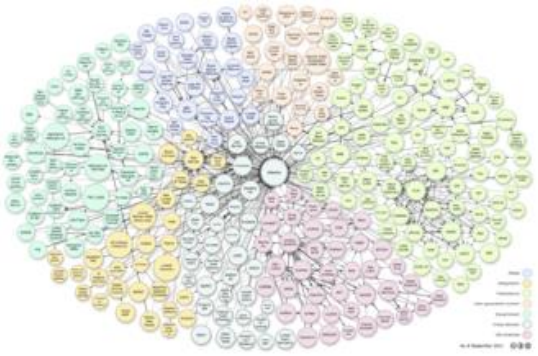
63

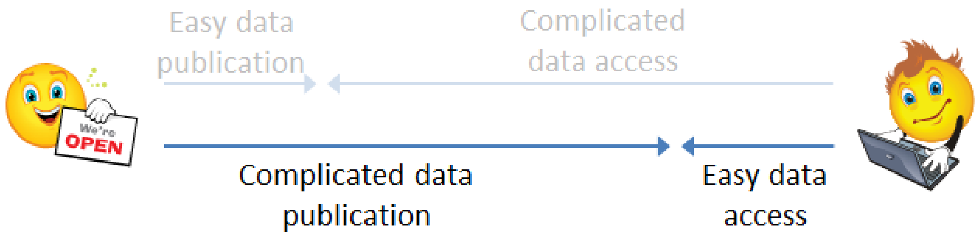
http://xmlns.com/foaf/0.1/name

Bob

Linked Data is great for Open Data

- Linked Data as a great means to represent and integrate disparate and heterogeneous open data sources
- How Linked Data can improve Open Data:
  - Easier integration, free data from silos
  - Seamless interlinking of data
  - Understand the data
  - New ways to query and interact with data
- Challenges with using Linked Data
  - Lack of tooling & expertise to publish high quality Linked Data
  - Lack of resources to host LOD endpoints / unreliable data access





Easy data publication

Complicated data access

Complicated data publication

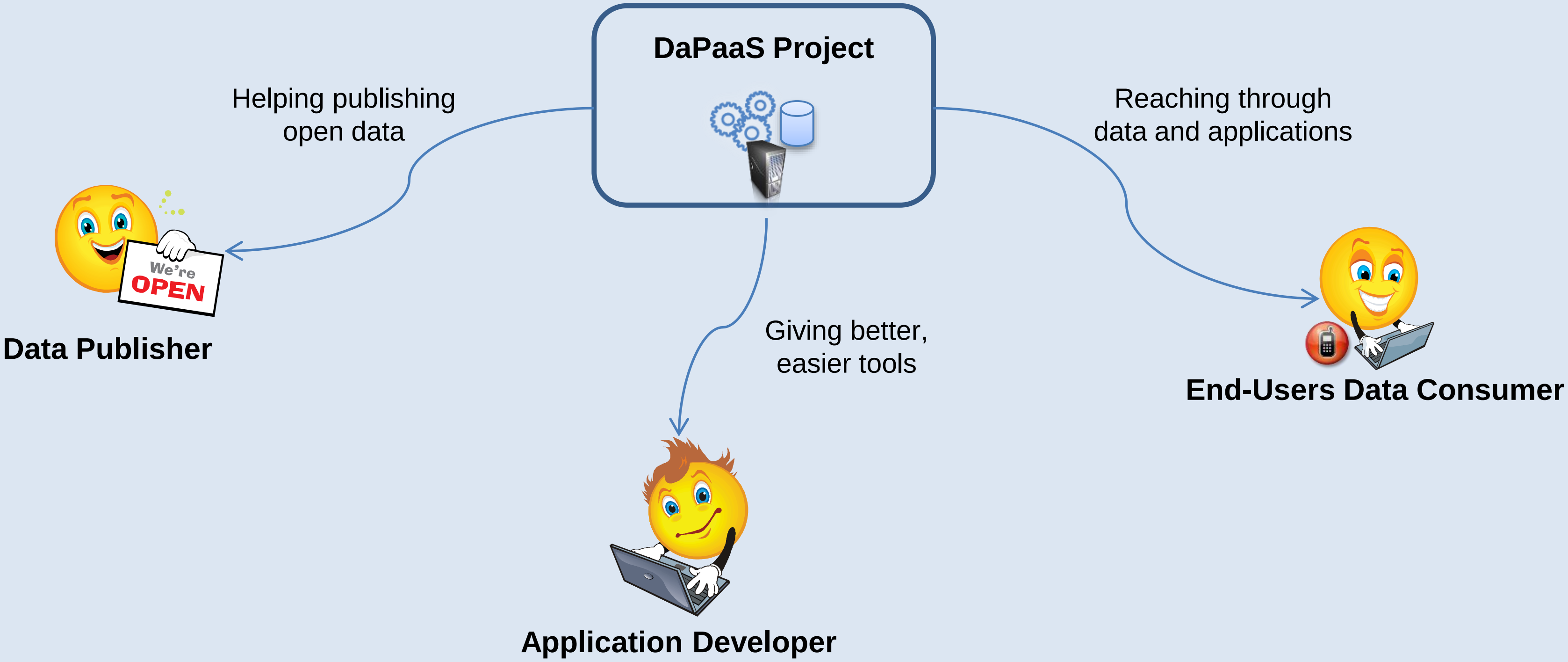
Easy data access

Linked Data has been ignored by the mainstream

- Difficult to make it accessible to people
  - Publishers
  - Developers
  - Data workers

- DaPaaS: packaging Linked Data to make it more approachable to the open data community

DaPaaS – one package 3 audiences



DaPaaS Project

Helping publishing open data

Data Publisher

Giving better, easier tools

Application Developer

Reaching through data and applications

End-Users Data Consumer

DaPaaS means to making Open (Linked) Data easier to use

- A **platform/hosting**: to make it easy for publishers to put data on the web, and developers to publish their applications
- A **portal**: to help advertising data and applications availability - and enticing new users
- **Tool-supported data transformation methodology**: to make it easy for people with Excel knowledge to publish large amounts of high quality data
- **API's with high-quality documentation**: for processing large amounts of data reliably in order to create interactives, visualisations and transformations

Make Linked Data more accessible to everyone!

DaPaaS – Data Value Chain

- **End-user Data Consumer**
  - Browse/Search Datasets&Apps Catalogue
  - App execution

- **App Developer**
  - Browse/Search Datasets Catalogue
  - App deployment and metadata creation

- **Data Publisher**
  - Dataset and Metadata creation
  - Data import and transformation
  - Data exploration
  - Data-driven portal configuration
  - Data export
  - Browse/Search Datasets Catalogue

Data
Value
Chain

Publishing and consuming data

- Data creates value when it is used:
  - help **users** find, understand and use data
  - help **data owners** publish it in the best way for re-use
  - support **intermediaries** to add value for end users by creating applications
  - reduce effort, increase quality during the publishing and consumption lifecycle

- **Rich structure of data allows development of rich applications**

Requirements for data publishing software

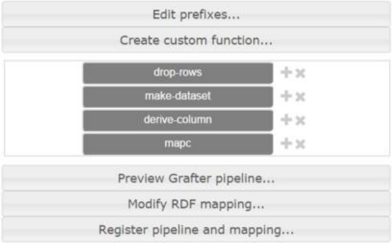
- Well-suited to producing RDF as the target output
- Already have a Graphical User Interface (GUI), or be suitable for one to be added
- Ability to use via an API, so that it can be automated and incorporated into other software tools
- Ability to serialise, export, version control and exchange transformation definitions
- Ability to accept a range of input types
  - CSV files, spreadsheets, relational database, geographical data formats, web form, copy of external RDF, extraction of data from an API
- Perform well with large datasets, both via API and via the GUI

DaPaaS Enablers

Grafter

Grafterizer

(Graphical Tool & DSL)

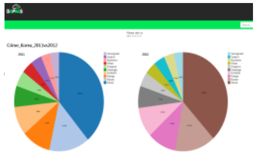

Edit prefixes...
Create custom function...
drop-rows
make-dataset
derive-column
mapc
Preview Grafter pipeline...
Modify RDF mapping...
Register pipeline and mapping...

DaPaaS platform

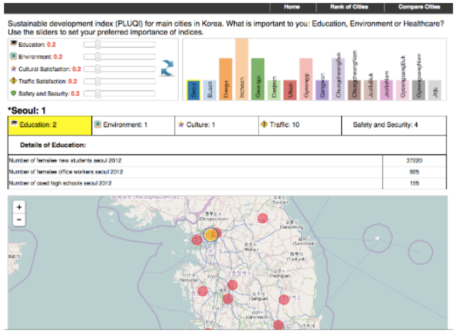
DAPAAS
Data- and Platform-as-a-Service

RDF database-as-a-service

RDF DDP



PLUQI



Sustainable development index (PLUQI) for main cities in Korea. What is important to you: Education, Environment or Healthcare? Use the sliders to set your preferred importance of indices.

Education: 0.2
Environment: 0.2
Cultural Satisfaction: 0.2
Traffic Satisfaction: 0.2
Safety and Security: 0.2

*Seoul: 1

Education: 2	Environment: 1	Culture: 1	Traffic: 10	Safety and Security: 4

Details of Education:

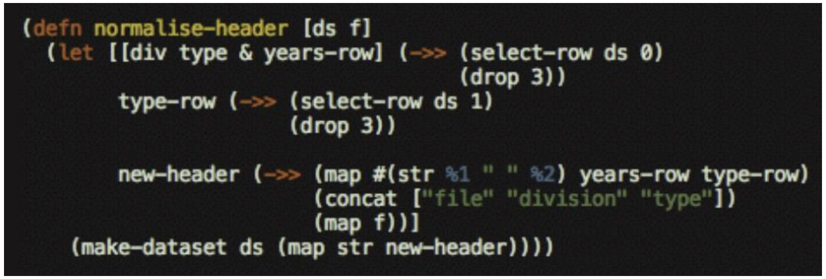
Number of females new students seoul 2012	37220
Number of females office workers seoul 2012	585
Number of coed high schools seoul 2012	155

Open Data Visualization-as-a-service

(Rainbow)

Grafter

- Grafter is a Clojure library, a DSL and a suite of tools for data transformation and processing
  - Clojure is a functional programming language similar to Lisp

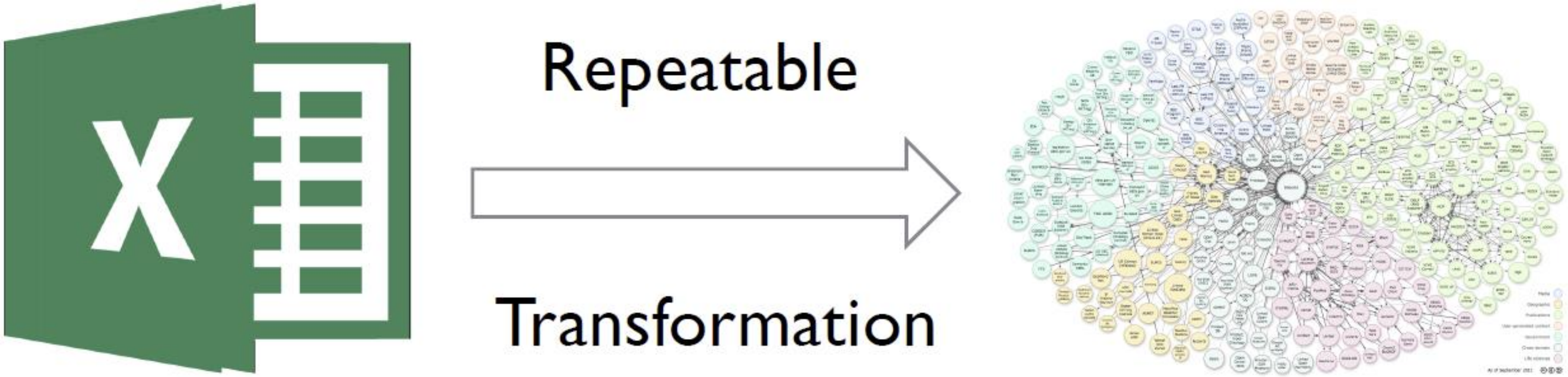


```
(defn normalise-header [ds f]
  (let [[div type & years-row] (->> (select-row ds 0)
                                    (drop 3))
        type-row (->> (select-row ds 1)
                      (drop 3))

        new-header (->> (map #(str %1 " " %2) years-row type-row)
                        (concat ["file" "division" "type"])
                        (map f))]
    (make-dataset ds (map str new-header))))
```

- Primarily used for handling data conversions from:
  - tabular data formats to tabular data formats
  - tabular data formats to RDF Linked Data format
- Open Source
  - Eclipse Public License (EPL)
  - http://github.com/dapaas/grafterizer

Tabular data (spreadsheet) to RDF Linked Data (graph)

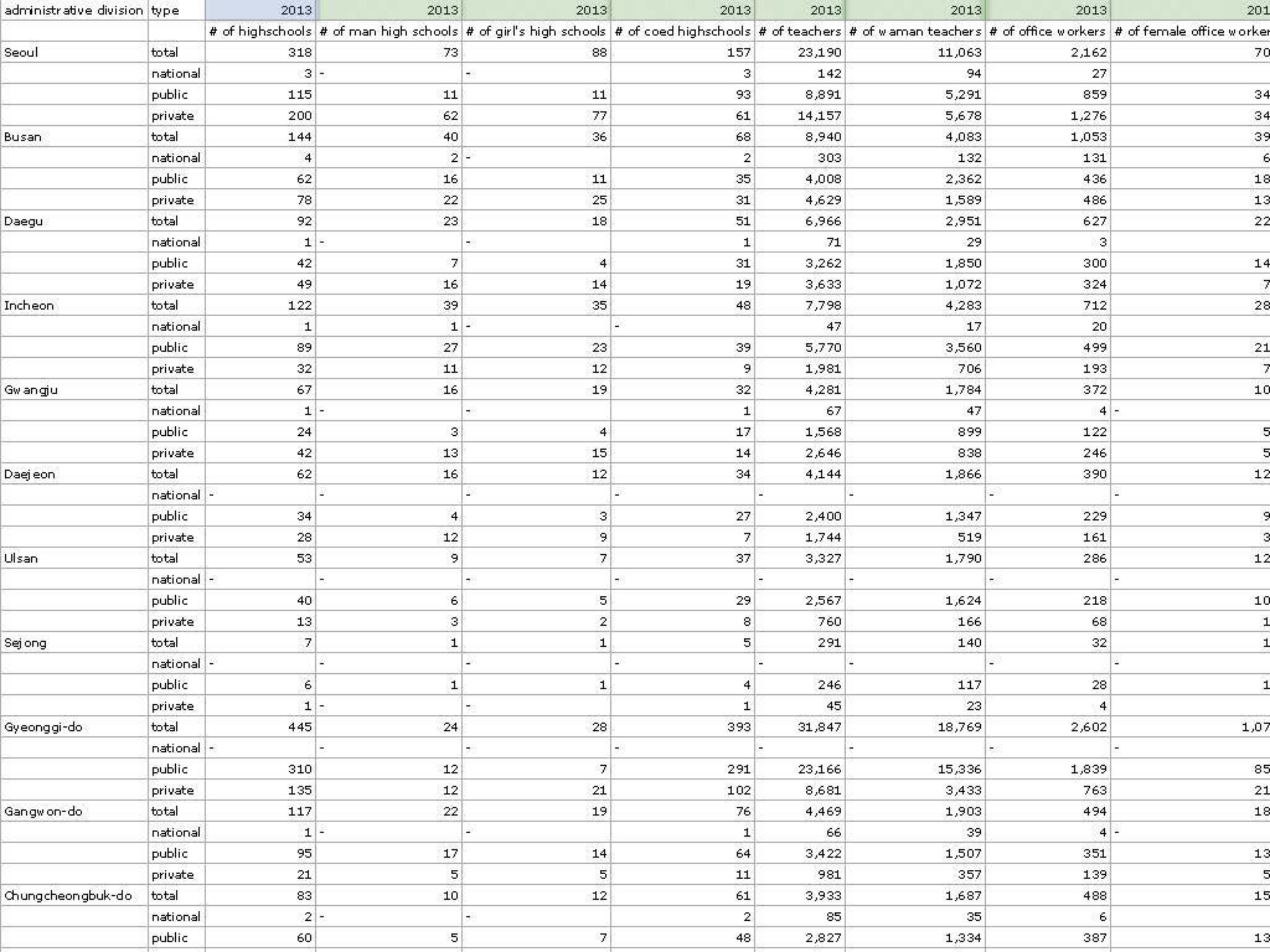


Repeatable

Transformation

1. **Specify a pipeline**, of tabular transformations for data cleaning and transformation
2. **Create the graph fragments**, resulting in the generation of an RDF graph

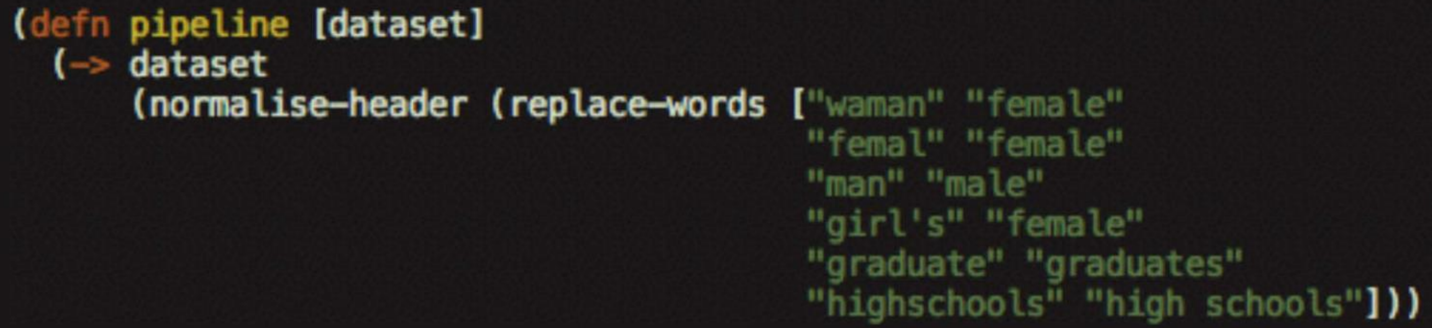
administrative division	type	2013							
		# of highschools	# of man high schools	# of girl's high schools	# of coed highschools	# of teachers	# of waman teachers	# of office workers	# of female office worker
Seoul	total	318	73	88	157	23,190	11,063	2,162	70
	national	3	-	-	3	142	94	27	
	public	115	11	11	93	8,891	5,291	859	34
	private	200	62	77	61	14,157	5,678	1,276	34
Busan	total	144	40	36	68	8,940	4,083	1,053	39
	national	4	2	-	2	303	132	131	6
	public	62	16	11	35	4,008	2,362	436	18
	private	78	22	25	31	4,629	1,589	486	13
Daegu	total	92	23	18	51	6,966	2,951	627	22
	national	1	-	-	1	71	29	3	
	public	42	7	4	31	3,262	1,850	300	14
	private	49	16	14	19	3,633	1,072	324	7
Incheon	total	122	39	35	48	7,798	4,283	712	28
	national	1	1	-	-	47	17	20	
	public	89	27	23	39	5,770	3,560	499	21
	private	32	11	12	9	1,981	706	193	7
Gwangju	total	67	16	19	32	4,281	1,784	372	10
	national	1	-	-	1	67	47	4	-
	public	24	3	4	17	1,568	899	122	5
	private	42	13	15	14	2,646	838	246	5
Daejeon	total	62	16	12	34	4,144	1,866	390	12
	national	-	-	-	-	-	-	-	-
	public	34	4	3	27	2,400	1,347	229	9
	private	28	12	9	7	1,744	519	161	3
Ulsan	total	53	9	7	37	3,327	1,790	286	12
	national	-	-	-	-	-	-	-	-
	public	40	6	5	29	2,567	1,624	218	10
	private	13	3	2	8	760	166	68	1
Sejong	total	7	1	1	5	291	140	32	1
	national	-	-	-	-	-	-	-	-
	public	6	1	1	4	246	117	28	1
	private	1	-	-	1	45	23	4	
Gyeonggi-do	total	445	24	28	393	31,847	18,769	2,602	1,07
	national	-	-	-	-	-	-	-	-
	public	310	12	7	291	23,166	15,336	1,839	85
	private	135	12	21	102	8,681	3,433	763	21
Gangwon-do	total	117	22	19	76	4,469	1,903	494	18
	national	1	-	-	1	66	39	4	-
	public	95	17	14	64	3,422	1,507	351	13
	private	21	5	5	11	981	357	139	5
Chungcheongbuk-do	total	83	10	12	61	3,933	1,687	488	15
	national	2	-	-	2	85	35	6	
	public	60	5	7	48	2,827	1,334	387	13



administrative division	type	2013	2013	2013	2013	2013	2013	2013	201
		# of highschools	# of man high schools	# of girl's high schools	# of coed highschools	# of teachers	# of waman teachers	# of office workers	# of female office worker
Seoul	total	318	73	88	157	23,190	11,063	2,162	70
	national	3	-	-	3	142	94	27	
	public	115	11	11	93	8,891	5,291	859	34
	private	200	62	77	61	14,157	5,678	1,276	34
Busan	total	144	40	36	68	8,940	4,083	1,053	39
	national	4	2	-	2	303	132	131	6
	public	62	16	11	35	4,008	2,362	436	18
	private	78	22	25	31	4,629	1,589	486	13
Daegu	total	92	23	18	51	6,966	2,951	627	22
	national	1	-	-	1	71	29	3	
	public	42	7	4	31	3,262	1,850	300	14
	private	49	16	14	19	3,633	1,072	324	7
Incheon	total	122	39	35	48	7,798	4,283	712	28
	national	1	1	-	-	47	17	20	
	public	89	27	23	39	5,770	3,560	499	21
	private	32	11	12	9	1,981	706	193	7
Gw angju	total	67	16	19	32	4,281	1,784	372	10
	national	1	-	-	1	67	47	4	-
	public	24	3	4	17	1,568	899	122	5
	private	42	13	15	14	2,646	838	246	5
Daej eon	total	62	16	12	34	4,144	1,866	390	12
	national	-	-	-	-	-	-	-	-
	public	34	4	3	27	2,400	1,347	229	9
	private	28	12	9	7	1,744	519	161	3
Ulsan	total	53	9	7	37	3,327	1,790	286	12
	national	-	-	-	-	-	-	-	-
	public	40	6	5	29	2,567	1,624	218	10
	private	13	3	2	8	760	166	68	1
Sej ong	total	7	1	1	5	291	140	32	1
	national	-	-	-	-	-	-	-	-
	public	6	1	1	4	246	117	28	1
	private	1	-	-	1	45	23	4	
Gyeonggi-do	total	445	24	28	393	31,847	18,769	2,602	1,07
	national	-	-	-	-	-	-	-	-
	public	310	12	7	291	23,166	15,336	1,839	85
	private	135	12	21	102	8,681	3,433	763	21
Gangw on-do	total	117	22	19	76	4,469	1,903	494	18
	national	1	-	-	1	66	39	4	-
	public	95	17	14	64	3,422	1,507	351	13
	private	21	5	5	11	981	357	139	5
Chungcheongbuk-do	total	83	10	12	61	3,933	1,687	488	15
	national	2	-	-	2	85	35	6	
	public	60	5	7	48	2,827	1,334	387	13

administrative division	type	2013	2013	2013	2013	2013	2013	2013	201
		# of highschools	# of man high schools	# of girl's high schools	# of coed highschools	# of teachers	# of waman teachers	# of office workers	# of female office worker
Seoul	total	318	73	88	157	23,190	11,063	2,162	70
	national	3	-	-	3	142	94	27	
	public	115	11	11	93	8,891	5,291	859	34
	private	200	62	77	61	14,157	5,678	1,276	34
Busan	total	144	40	36	68	8,940	4,083	1,053	39
	national	4	2	-	2	303	132	131	6
	public	62	16	11	35	4,008	2,362	436	18
	private	78	22	25	31	4,629	1,589	486	13
Daegu	total	92	23	18	51	6,966	2,951	627	22
	national	1	-	-	1	71	29	3	
	public	42	7	4	31	3,262	1,850	300	14
	private	49	16	14	19	3,633	1,072	324	7
Incheon	total	122	39	35	48	7,798	4,283	712	28
	national	1	1	-	-	47	17	20	
	public	89	27	23	39	5,770	3,560	499	21
	private	32	11	12	9	1,981	706	193	7
Gwangju	total	67	16	19	32	4,281	1,784	372	10
	national	1	-	-	1	67	47	4	-
	public	24	3	4	17	1,568	899	122	5
	private	42	13	15	14	2,646	838	246	5
Daejeon	total	62	16	12	34	4,144	1,866	390	12
	national	-	-	-	-	-	-	-	-
	public	34	4	3	27	2,400	1,347	229	9
	private	28	12	9	7	1,744	519	161	3
Ulsan	total	53	9	7	37	3,327	1,790	286	12
	national	-	-	-	-	-	-	-	-
	public	40	6	5	29	2,567	1,624	218	10
	private	13	3	2	8	760	166	68	1
Sejong	total	7	1	1	5	291	140	32	1
	national	-	-	-	-	-	-	-	-
	public	6	1	1	4	246	117	28	1
	private	1	-	-	1	45	23	4	
Gyeonggi-do	total	445	24	28	393	31,847	18,769	2,602	1,07
	national	-	-	-	-	-	-	-	-
	public	310	12	7	291	23,166	15,336	1,839	85
	private	135	12	21	102	8,681	3,433	763	21
Gangwon-do	total	117	22	19	76	4,469	1,903	494	18
	national	1	-	-	1	66	39	4	-
	public	95	17	14	64	3,422	1,507	351	13
	private	21	5	5	11	981	357	139	5
Chungcheongbuk-do	total	83	10	12	61	3,933	1,687	488	15
	national	2	-	-	2	85	35	6	
	public	60	5	7	48	2,827	1,334	387	13

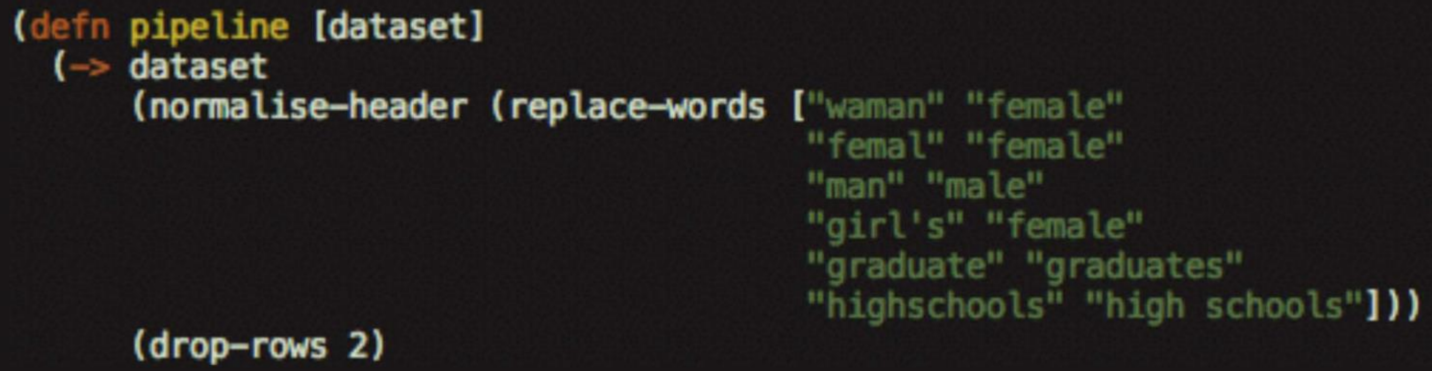
division	type	2013 # of high schools	2013 # of male high schools	2013 # of female high schools	2013 # of coed high schools	2013 # of teachers	2013 # of female
administrative division	type	2,013	2,013	2,013	2,013	2,013	
시도별	구분별	# of highschools	# of man high schools	# of girl's high schools	# of coed highschools	# of teachers	# of waman
Seoul	total	318	73	88	157	23,190	
	national	3	-	-	3	142	
	public	115	11	11	93	8,891	
	private	200	62	77	61	14,157	
Busan	total	144	40	36	68	8,940	
	national	4	2	-	2	303	
	public	62	16	11	35	4,008	
	private	78	22	25	31	4,629	
Daegu	total	92	23	18	51	6,966	
	national	1	-	-	1	71	
	public	42	7	4	31	3,262	
	private	49	16	14	19	3,633	
Incheon	total	122	39	35	48	7,798	
	national	1	1	-	-	47	
	public	89	27	23	39	5,770	
	private	32	11	12	9	1,981	
Gwangju	total	67	16	19	32	4,281	
	national	1	-	-	1	67	
	public	24	3	4	17	1,568	
	private	42	13	15	14	2,646	
Daejeon	total	62	16	12	34	4,144	
	national	-	-	-	-	-	
	public	34	4	3	27	2,400	
	private	28	12	9	7	1,744	
Ulsan	total	53	9	7	37	3,327	
	national	-	-	-	-	-	
	public	40	6	5	29	2,567	
	private	13	3	2	8	760	
Sejong	total	7	1	1	5	291	



```clojure
(defn pipeline [dataset]
  (-> dataset
      (normalise-header (replace-words ["waman" "female"
                                        "femal" "female"
                                        "man" "male"
                                        "girl's" "female"
                                        "graduate" "graduates"
                                        "highschools" "high schools"]))
```

division	type	2013 # of high schools	2013 # of male high schools	2013 # of female high schools	2013 # of coed high schools	2013 # of teachers	2013 # of female t
administrative division	type	2,013	2,013	2,013	2,013	2,013	
시도별	구분별	# of highschools	# of man high schools	# of girl's high schools	# of coed highschools	# of teachers	# of waman
Seoul	total	318	73	88	157	23,190	
	national	3	-	-	3	142	
	public	115	11	11	93	8,891	
	private	200	62	77	61	14,157	
Busan	total	144	40	36	68	8,940	
	national	4	2	-	2	303	
	public	62	16	11	35	4,008	
	private	78	22	25	31	4,629	
Daegu	total	92	23	18	51	6,966	
	national	1	-	-	1	71	
	public	42	7	4	31	3,262	
	private	49	16	14	19	3,633	
Incheon	total	122	39	35	48	7,798	
	national	1	1	-	-	47	
	public	89	27	23	39	5,770	
	private	32	11	12	9	1,981	
Gwangju	total	67	16	19	32	4,281	
	national	1	-	-	1	67	
	public	24	3	4	17	1,568	
	private	42	13	15	14	2,646	
Daejeon	total	62	16	12	34	4,144	
	national	-	-	-	-	-	
	public	34	4	3	27	2,400	
	private	28	12	9	7	1,744	
Ulsan	total	53	9	7	37	3,327	
	national	-	-	-	-	-	
	public	40	6	5	29	2,567	
	private	13	3	2	8	760	
Sejong	total	7	1	1	5	291	

division	type	2013 # of high schools	2013 # of male high schools	2013 # of female high schools	2013 # of coed high schools	2013 # of teachers	2013 # of female teachers	20
Seoul	total	318	73	88	157	23,190	11,063	
	national	3	-	-	3	142	94	
	public	115	11	11	93	8,891	5,291	
	private	200	62	77	61	14,157	5,678	
Busan	total	144	40	36	68	8,940	4,083	
	national	4	2	-	2	303	132	
	public	62	16	11	35	4,008	2,362	
	private	78	22	25	31	4,629	1,589	
Daegu	total	92	23	18	51	6,966	2,951	
	national	1	-	-	1	71	29	
	public	42	7	4	31	3,262	1,850	
	private	49	16	14	19	3,633	1,072	
Incheon	total	122	39	35	48	7,798	4,283	
	national	1	1	-	-	47	17	
	public	89	27	23	39	5,770	3,560	
	private	32	11	12	9	1,981	706	
Gwangju	total	67	16	19	32	4,281	1,784	
	national	1	-	-	1	67	47	
	public	24	3	4	17	1,568	899	
	private	42	13	15	14	2,646	838	
Daejeon	total	62	16	12	34	4,144	1,866	
	national	-	-	-	-	-	-	
	public	34	4	3	27	2,400	1,347	
	private	28	12	9	7	1,744	519	
Ulsan	total	53	9	7	37	3,327	1,790	
	national	-	-	-	-	-	-	
	public	40	6	5	29	2,567	1,624	
	private	13	3	2	8	760	166	
Sejong	total	7	1	1	5	291	140	
	national	-	-	-	-	-	-	
	public	6	1	1	4	246	117	



```
(defn pipeline [dataset]
  (-> dataset
      (normalise-header (replace-words ["waman" "female"
                                        "femal" "female"
                                        "man" "male"
                                        "girl's" "female"
                                        "graduate" "graduates"
                                        "highschools" "high schools"]))
      (drop-rows 2)
```

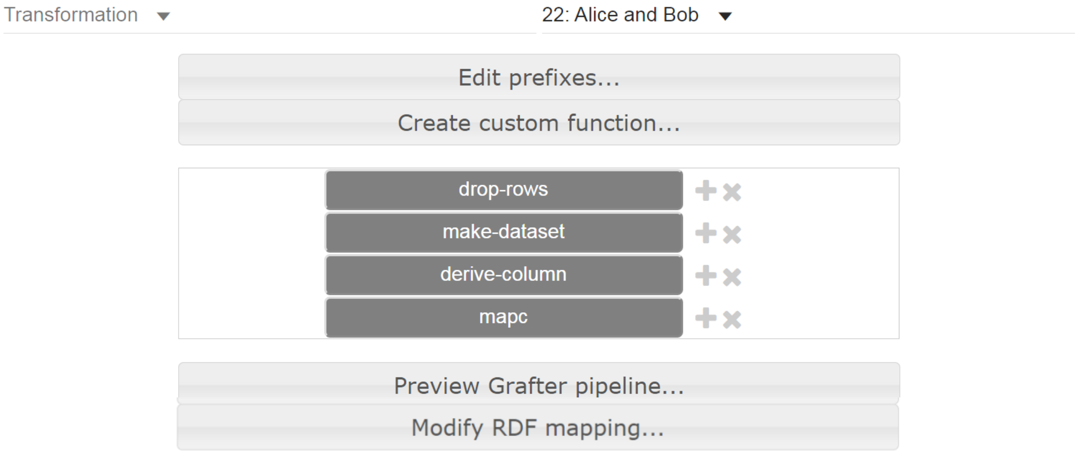
division	type	2013 # of high schools	2013 # of male high schools	2013 # of female high schools	2013 # of coed high schools	2013 # of teachers	2013 # of female teachers	20
Seoul	total	318	73	88	157	23,190	11,063	
	national	3	-	-	3	142	94	
	public	115	11	11	93	8,891	5,291	
	private	200	62	77	61	14,157	5,678	
Busan	total	144	40	36	68	8,940	4,083	
	national	4	2	-	2	303	132	
	public	62	16	11	35	4,008	2,362	
	private	78	22	25	31	4,629	1,589	
Daegu	total	92	23	18	51	6,966	2,951	
	national	1	-	-	1	71	29	
	public	42	7	4	31	3,262	1,850	
	private	49	16	14	19	3,633	1,072	
Incheon	total	122	39	35	48	7,798	4,283	
	national	1	1	-	-	47	17	
	public	89	27	23	39	5,770	3,560	
	private	32	11	12	9	1,981	706	
Gwangju	total	67	16	19	32	4,281	1,784	
	national	1	-	-	1	67	47	
	public	24	3	4	17	1,568	899	
	private	42	13	15	14	2,646	838	
Daejeon	total	62	16	12	34	4,144	1,866	
	national	-	-	-	-	-	-	
	public	34	4	3	27	2,400	1,347	
	private	28	12	9	7	1,744	519	
Ulsan	total	53	9	7	37	3,327	1,790	
	national	-	-	-	-	-	-	
	public	40	6	5	29	2,567	1,624	
	private	13	3	2	8	760	166	
Sejong	total	7	1	1	5	291	140	
	national	-	-	-	-	-	-	
	public	6	1	1	4	246	117	

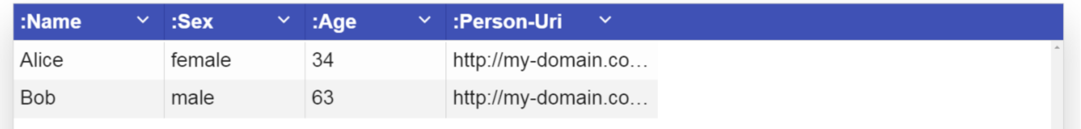
division	type	2013 # of high schoo	2013 # of male high scho	2013 # of female high scho	2013 # of coed high schools	2013 # of teachers	2013 # of female teachers	201
Seoul	total	318	73	88	157	23,190	11,063	
Seoul	national	3	-	-	3	142	94	27
Seoul	public	115	11	11	93	8,891	5,291	
Seoul	private	200	62	77	61	14,157	5,678	
Busan	total	144	40	36	68	8,940	4,083	
Busan	national	4	2	-	2	303	132	131
Busan	public	62	16	11	35	4,008	2,362	
Busan	private	78	22	25	31	4,629	1,589	
Daegu	total	92	23	18	51	6,966	2,951	
Daegu	national	1	-	-	1	71	29	3
Daegu	public	42	7	4	31	3,262	1,850	
Daegu	private	49	16	14	19	3,633	1,072	
Incheon	total	122	39	35	48	7,798	4,283	
Incheon	national	1	1	-	-	47	17	20
Incheon	public	89	27	23	39	5,770	3,560	
Incheon	private	32	11	12	9	1,981	706	
Gwangju	total	67	16	19	32	4,281	1,784	
Gwangju	national	1	-	-	1	67	47	4
Gwangju	public	24	3	4	17	1,568	899	
Gwangju	private	42	13	15	14	2,646	838	
Daejeon	total	62	16	12	34	4,144	1,866	
Daejeon	national	-	-	-	-	-	-	-
Daejeon	public	34	4	3	27	2,400	1,347	
Daejeon	private	28	12	9	7	1,744	519	
Ulsan	total	53	9	7	37	3,327	1,790	
Ulsan	national	-	-	-	-	-	-	-
Ulsan	public	40	6	5	29	2,567	1,624	
Ulsan	private	13	3	2	8	760	166	
Sejong	total	7	1	1	5	291	140	
Sejong	national	-	-	-	-	-	-	-
Sejong	public	6	1	1	4	246	117	

```
(defn pipeline [dataset]
  (-> dataset
      (normalise-header (replace-words ["waman" "female"
                                        "femal" "female"
                                        "man" "male"
                                        "girl's" "female"
                                        "graduate" "graduates"
                                        "highschools" "high schools"]))
      (drop-rows 2)
      (apply-columns {:division fill-when})
```

Grafterizer

- GUI tool for the Grafter suite; Open Source (EPL)
  - http://github.com/dapaas/grafterizer
- Specify tabular data transformations
  - Interactively preview results
  - Specialise transformations using custom functions
  - Use prefixes to form URIs

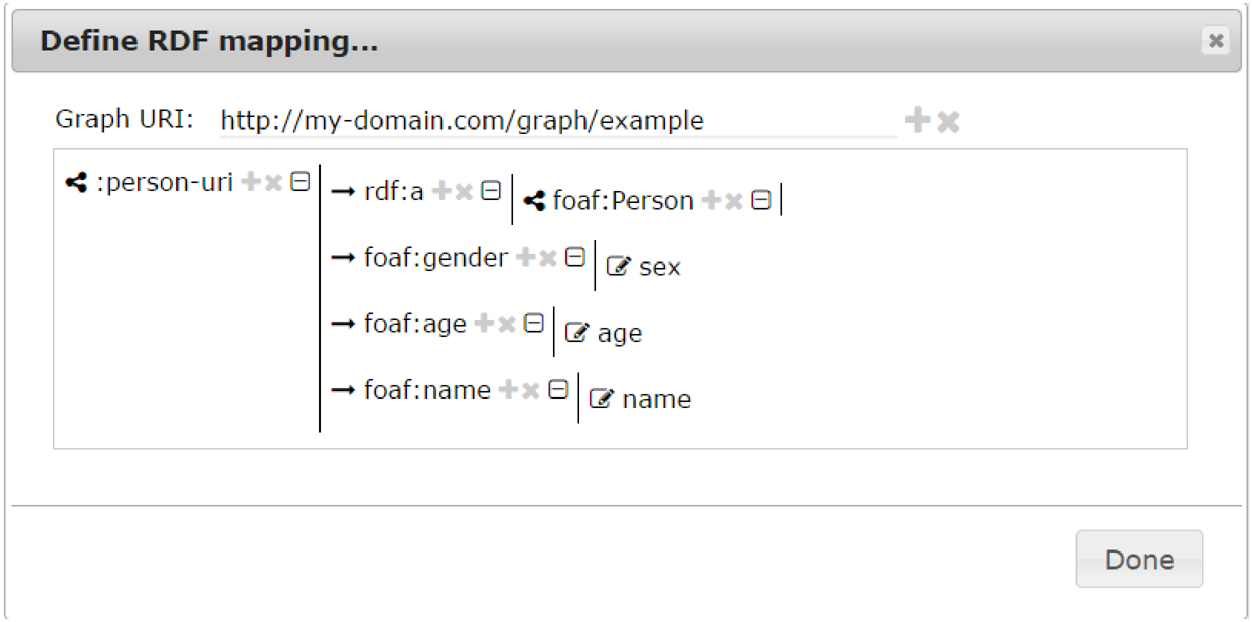




:Name	:Sex	:Age	:Person-Uri
Alice	female	34	http://my-domain.co...
Bob	male	63	http://my-domain.co...

Grafterizer (cont')

- Specify mappings from tabular data to RDF



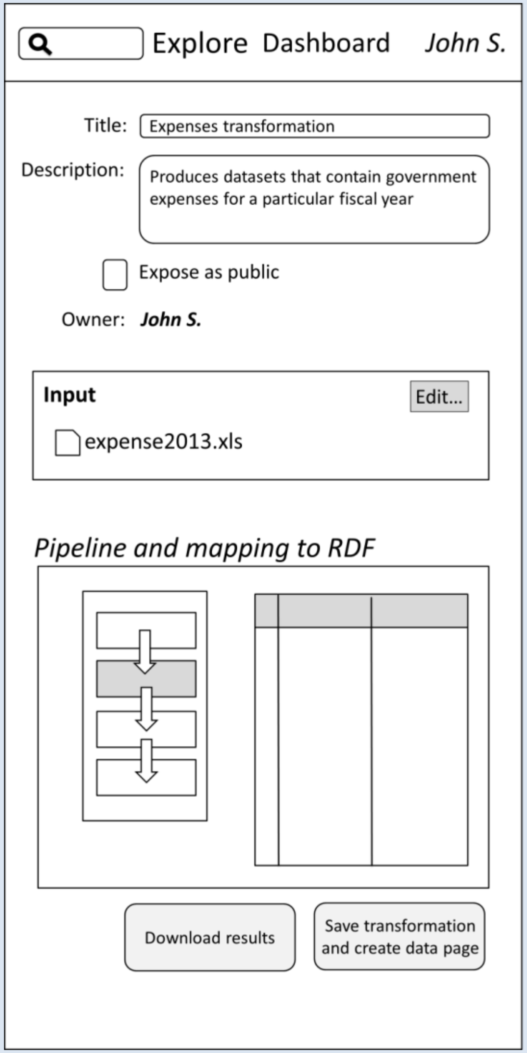
Define RDF mapping...

Graph URI: http://my-domain.com/graph/example

:person-uri
- rdf:a → foaf:Person
- foaf:gender → sex
- foaf:age → age
- foaf:name → name

Grafterizer concept

Transformation Page



Explore
Dashboard
John S.

Title: Expenses transformation

Description: Produces datasets that contain government expenses for a particular fiscal year

□ Expose as public

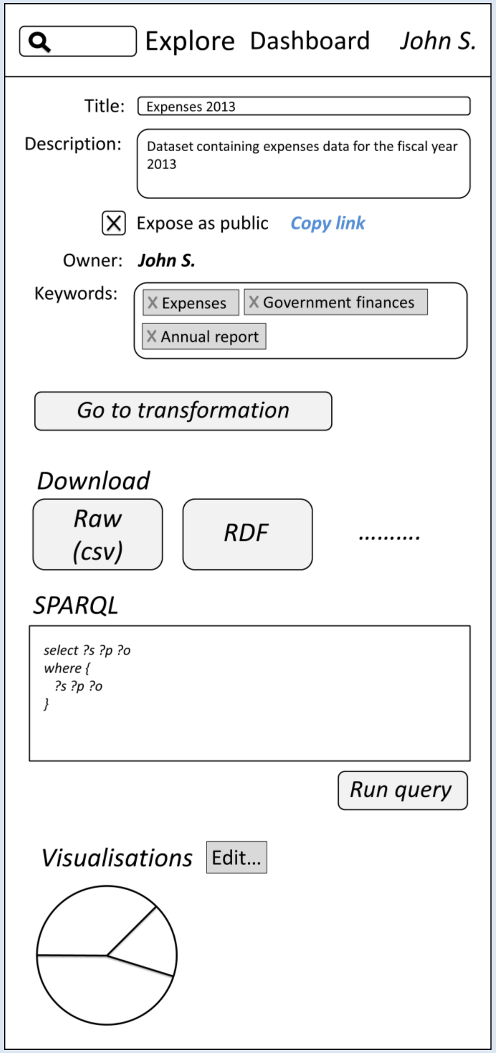
Owner: ***John S.***

Input

expense2013.xls

Edit...
Pipeline and mapping to RDF
Download results
Save transformation and create data page

Data Page



Title: Expenses 2013

Description: Dataset containing expenses data for the fiscal year 2013

☒ Expose as public *Copy link*

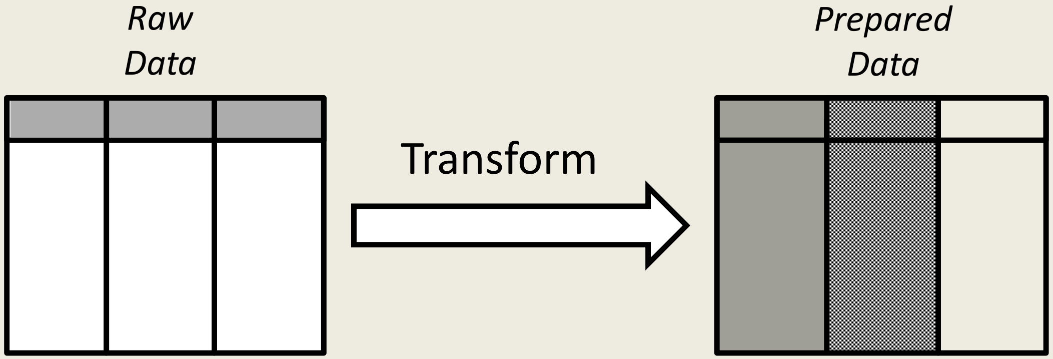
Owner: ***John S.***

Keywords: X Expenses X Government finances X Annual report

```
select ?s ?p ?o
where {
  ?s ?p ?o
}
```

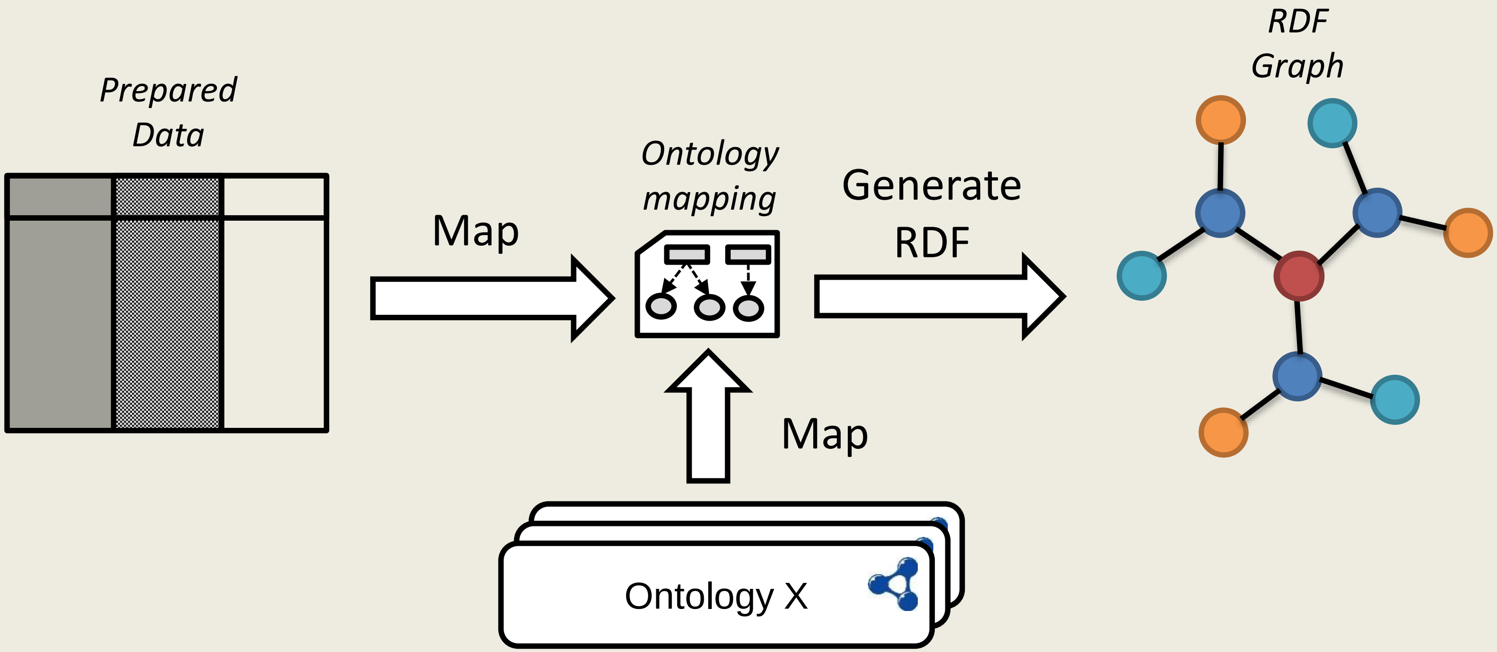
Use Case: Data Transformation

- Import raw tabular data
- Clean up and transform data using Grafterizer



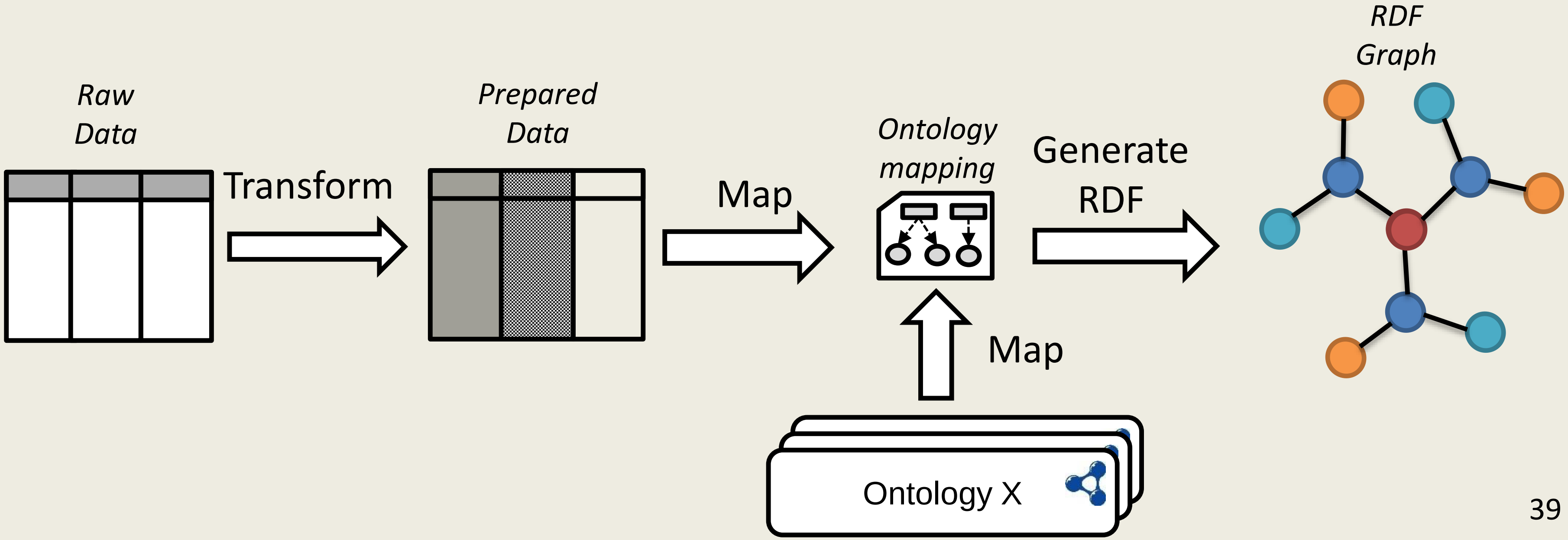
Use Case: Mapping to RDF

- Import prepared data
- Define ontology mapping using Grafterizer
- Generate RDF graph



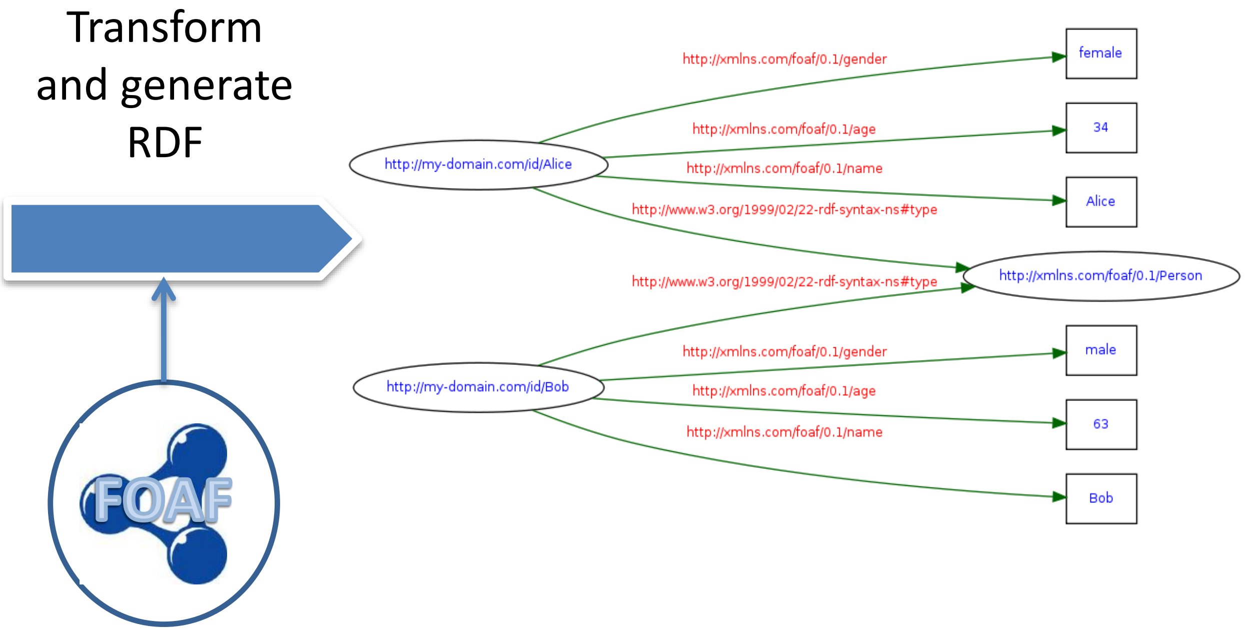
Use Case: Transformation and Mapping to RDF

- Import raw data
- Clean up and transform using Grafterizer
- Define ontology mapping using Grafterizer
- Generate RDF Graph



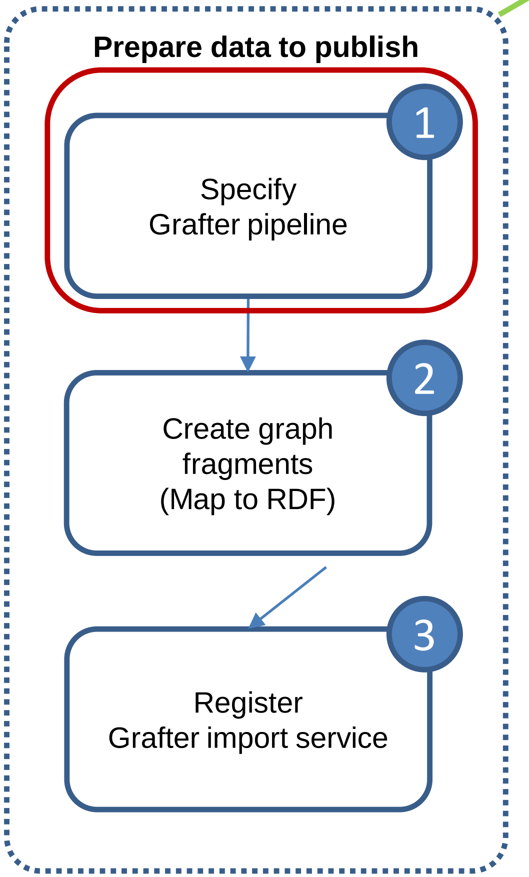
Example: Transformation and Mapping to RDF

Name	Sex	Age
Alice	f	"34"
Bob	m	"63"


Transform and generate RDF

FOAF
http://my-domain.com/id/Alice
http://xmlns.com/foaf/0.1/gender
female
http://xmlns.com/foaf/0.1/age
34
http://xmlns.com/foaf/0.1/name
Alice
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://xmlns.com/foaf/0.1/Person
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://my-domain.com/id/Bob
http://xmlns.com/foaf/0.1/gender
male
http://xmlns.com/foaf/0.1/age
63
http://xmlns.com/foaf/0.1/name
Bob

Example: Transformation and Mapping to RDF



Simple example

Example dataset input:

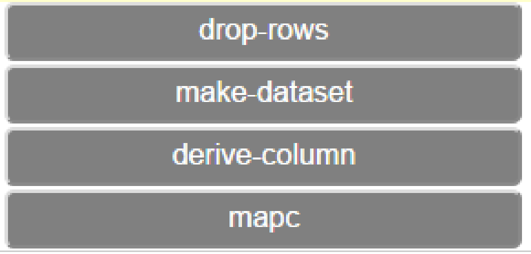
Name	Sex	Age
Alice	f	"34"
Bob	m	"63"

Example output: An RDF graph where

- Each row represents a foaf:Person
- 'Name', **as a URI**, represents the row node
- 'Sex' is **transformed to a full string** ('f' -> 'female'; 'm' -> 'male') and then mapped to foaf:gender
- 'Age' is mapped to foaf:age directly, after **parsing it as integer**

Pipeline (Data cleaning and transformation)

1. **Create a URI** based on the **'Name'** column
2. **Transform 'Sex' column** contents from single letter strings to full gender names
3. **Transform 'Age' column** contents to integers



drop-rows

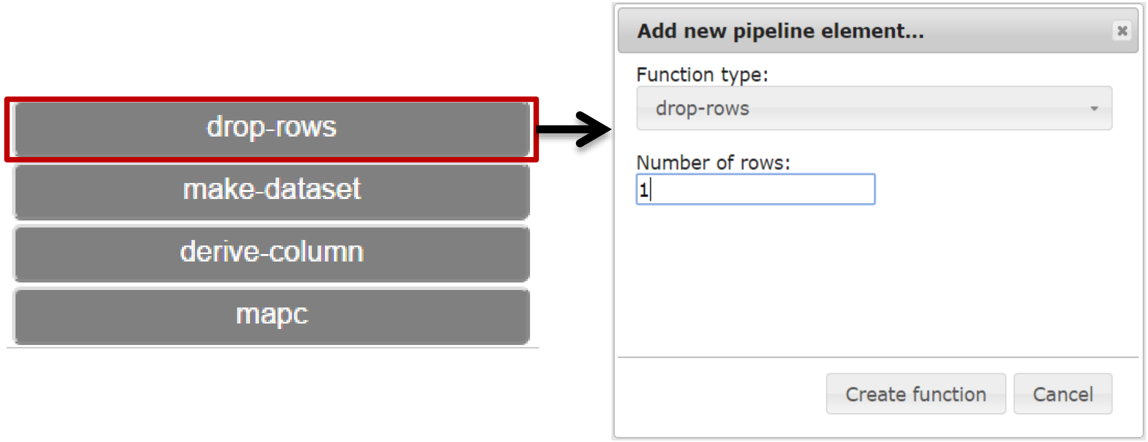
make-dataset

derive-column

mapc

Development process Grafterizer: Step 1 (pipeline)

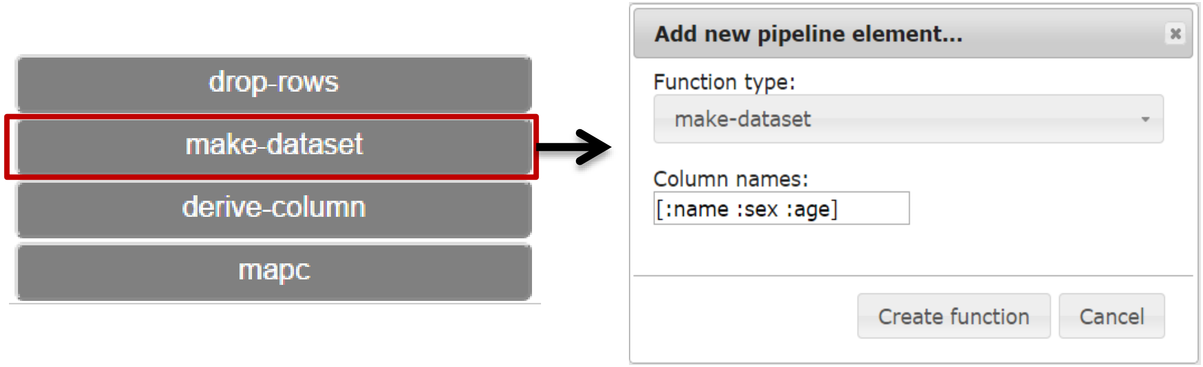
1. Removing the header row from the dataset



Function type: drop-rows

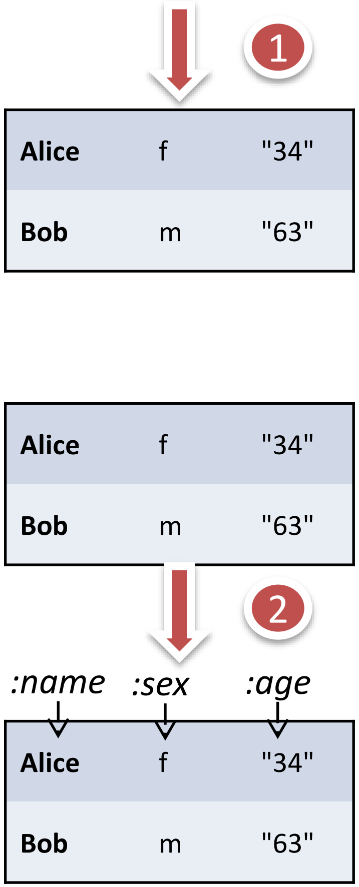
Number of rows: 1

2. Creating aliases - for referencing the columns in the rest of the pipeline



Function type: make-dataset

Column names: [:name :sex :age]



Alice	f	"34"
Bob	m	"63"

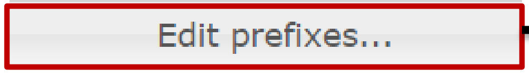
Alice	f	"34"
Bob	m	"63"

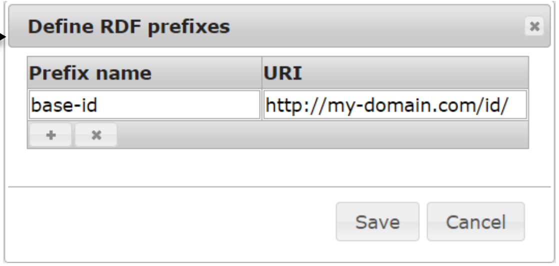
:name	*:sex*	*:age*
Alice	f	"34"
Bob	m	"63"

Development process Grafterizer: Step 1 (prefixes)

3. URI-ifying the name column

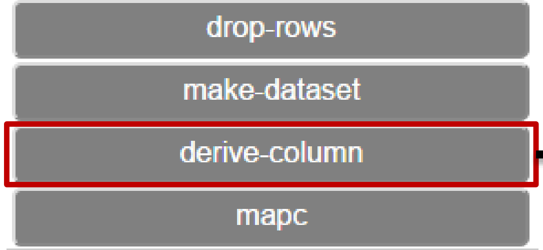
a) Creating the prefix definition

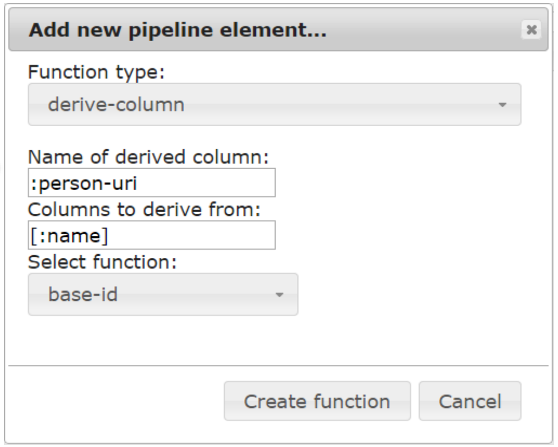



Define RDF prefixes

Prefix name	URI
base-id	http://my-domain.com/id/

b) Creating the pipeline element



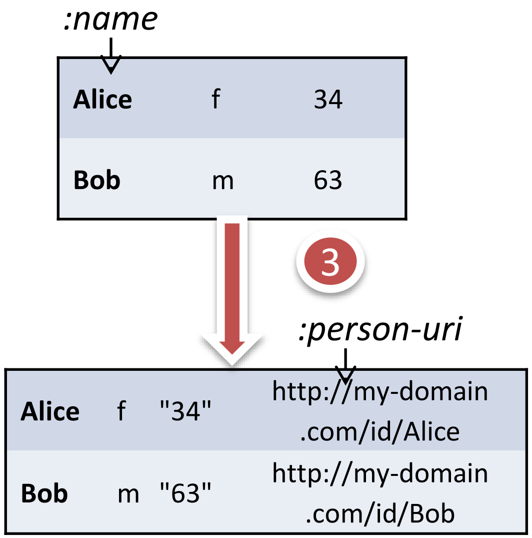

Add new pipeline element...

Function type: derive-column

Name of derived column: :person-uri

Columns to derive from: [:name]

Select function: base-id



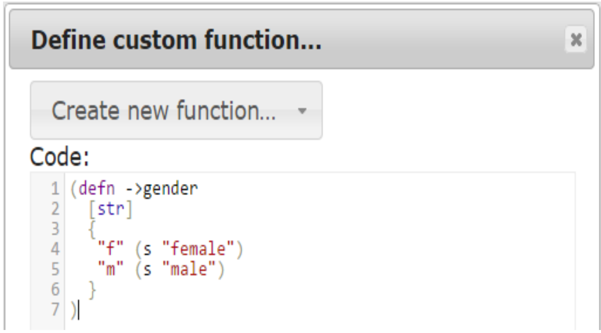
Alice	f	34
Bob	m	63

Alice	f	"34"	http://my-domain.com/id/Alice
Bob	m	"63"	http://my-domain.com/id/Bob

Development process Grafterizer: Step 1 (custom functions)

4. Apply transformations to the :age and :sex columns

a) Defining the custom transformations in Clojure

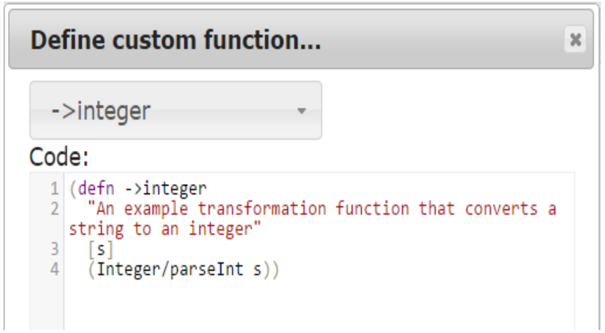


Define custom function...

Create new function...

Code:

```
(defn ->gender
  [str]
  {
   "f" (s "female")
   "m" (s "male")
  }
)
```



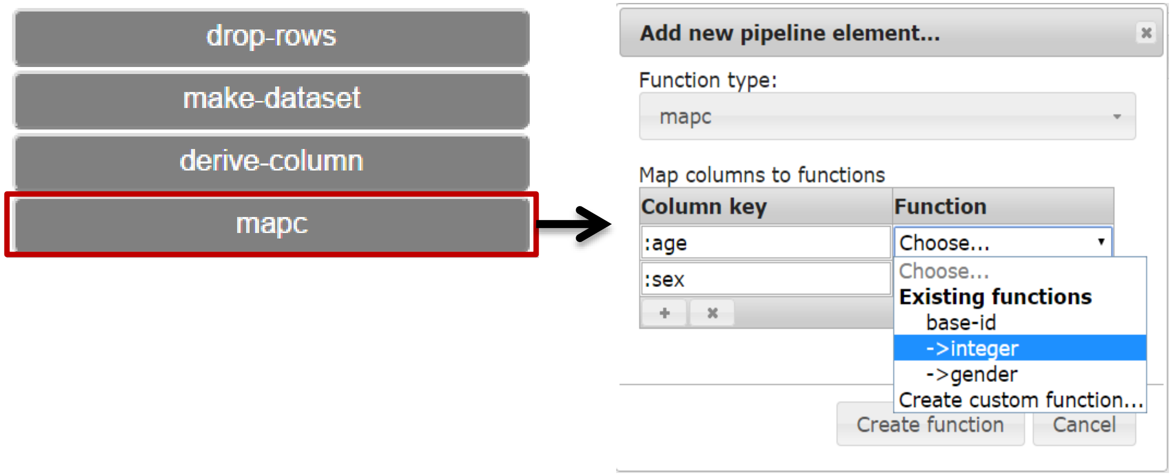
Define custom function...

->integer

Code:

```
(defn ->integer
  "An example transformation function that converts a string to an integer"
  [s]
  (Integer/parseInt s))
```

b) Applying the transformations to each of the columns

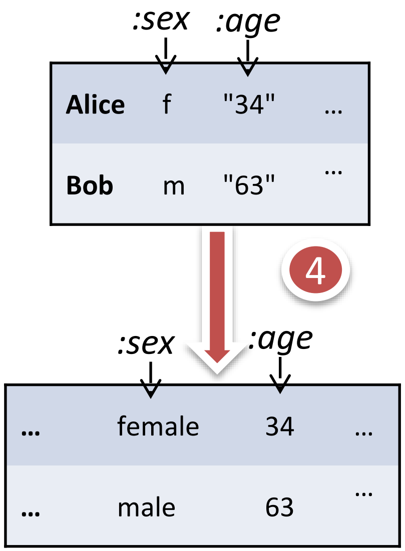


Add new pipeline element...

Function type: mapc

Map columns to functions

Column key	Function
:age	Choose...
:sex	



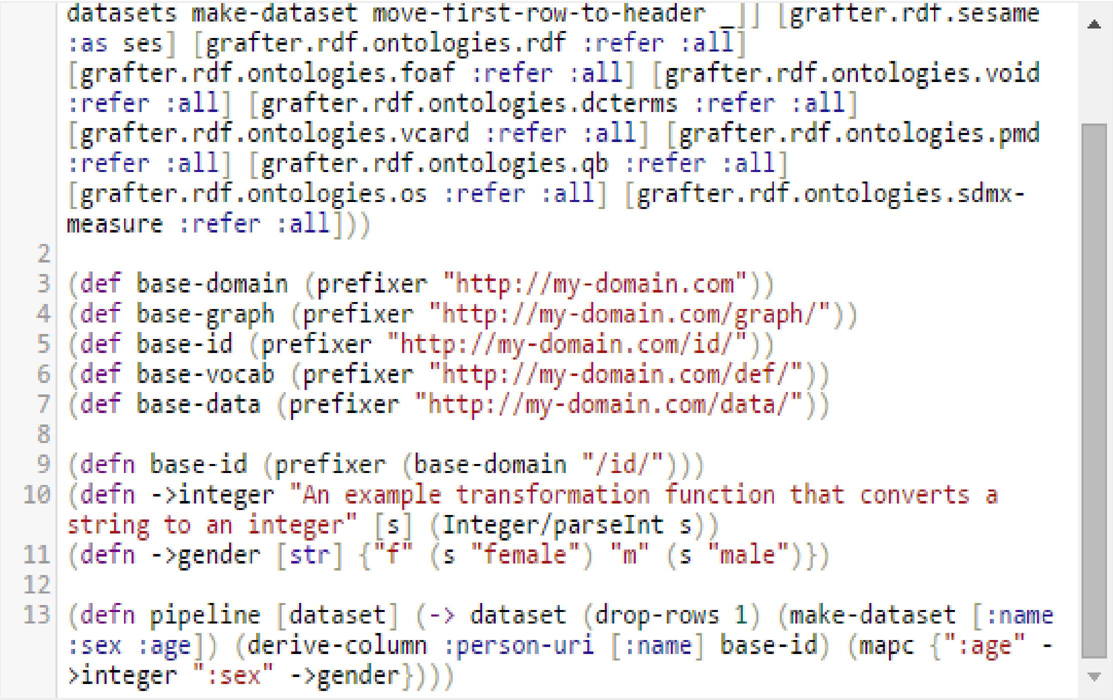
Alice	f	"34"	...
Bob	m	"63"	...

...	female	34	...
...	male	63	...

Development process Grafterizer: Step 1 (preview)

5. Preview Grafter pipeline

↓



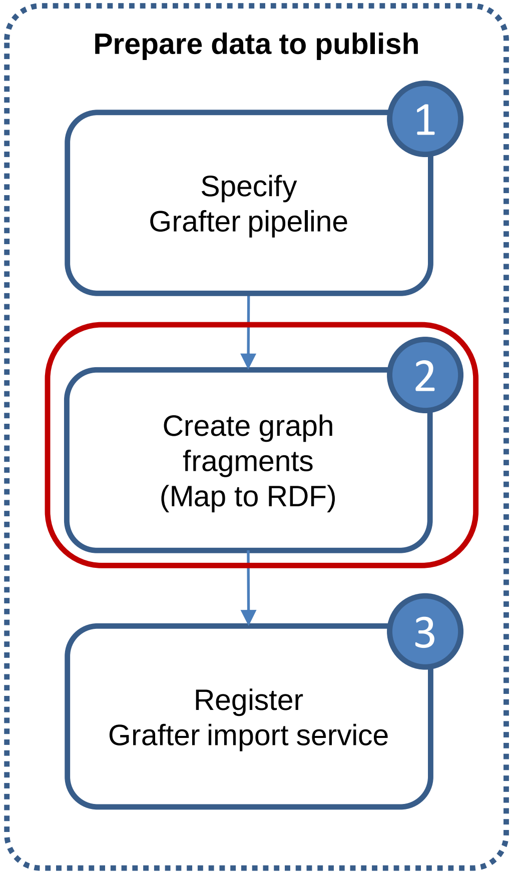
```
datasets make-dataset move-first-row-to-header _]] [grafter.rdf.sesame :as ses] [grafter.rdf.ontologies.rdf :refer :all] [grafter.rdf.ontologies.foaf :refer :all] [grafter.rdf.ontologies.void :refer :all] [grafter.rdf.ontologies.dcterms :refer :all] [grafter.rdf.ontologies.vcard :refer :all] [grafter.rdf.ontologies.pmd :refer :all] [grafter.rdf.ontologies.qb :refer :all] [grafter.rdf.ontologies.os :refer :all] [grafter.rdf.ontologies.sdmx-measure :refer :all]))

(def base-domain (prefixer "http://my-domain.com"))
(def base-graph (prefixer "http://my-domain.com/graph/"))
(def base-id (prefixer "http://my-domain.com/id/"))
(def base-vocab (prefixer "http://my-domain.com/def/"))
(def base-data (prefixer "http://my-domain.com/data/"))

(defn base-id (prefixer (base-domain "/id/")))
(defn ->integer "An example transformation function that converts a string to an integer" [s] (Integer/parseInt s))
(defn ->gender [str] {"f" (s "female") "m" (s "male")})

(defn pipeline [dataset] (-> dataset (drop-rows 1) (make-dataset [:name :sex :age]) (derive-column :person-uri [:name] base-id) (mapc {":age" ->integer ":sex" ->gender})))
```

Development process Grafterizer: Step 2 (RDF mapping)

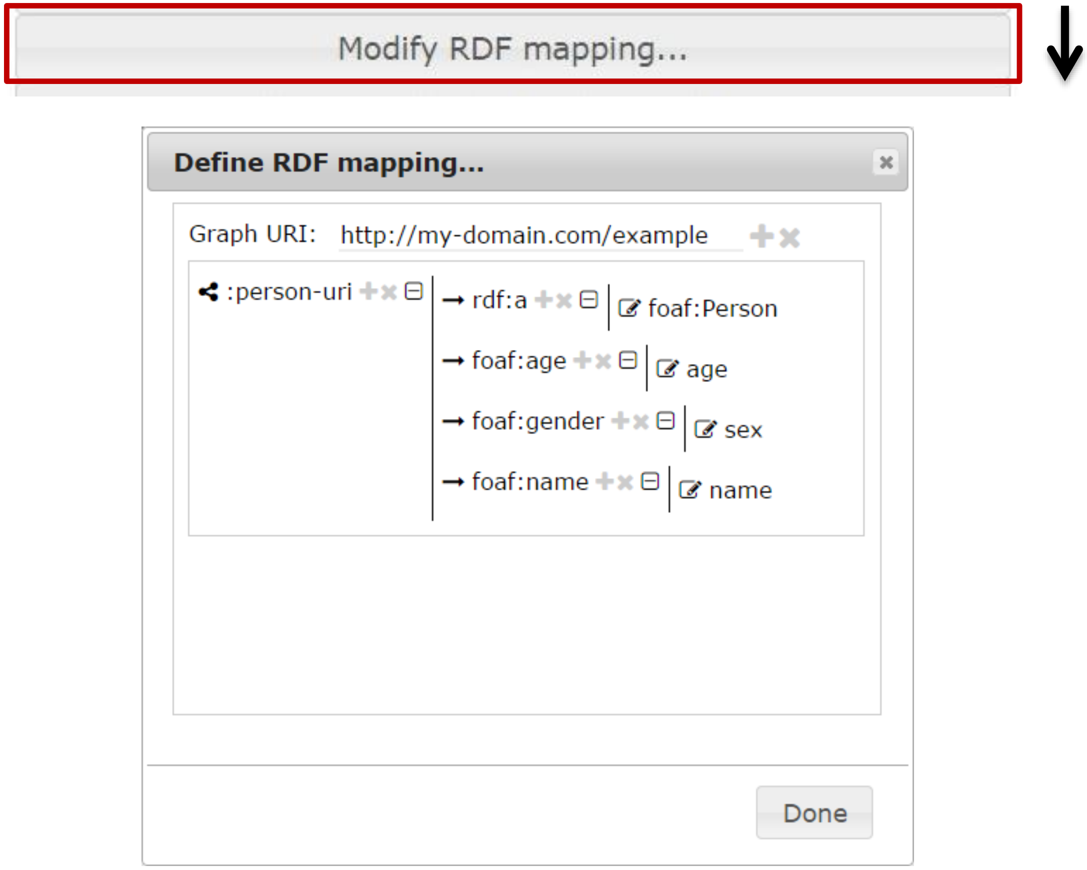


Prepare data to publish

1 Specify Grafter pipeline

2 Create graph fragments (Map to RDF)

3 Register Grafter import service



Modify RDF mapping...

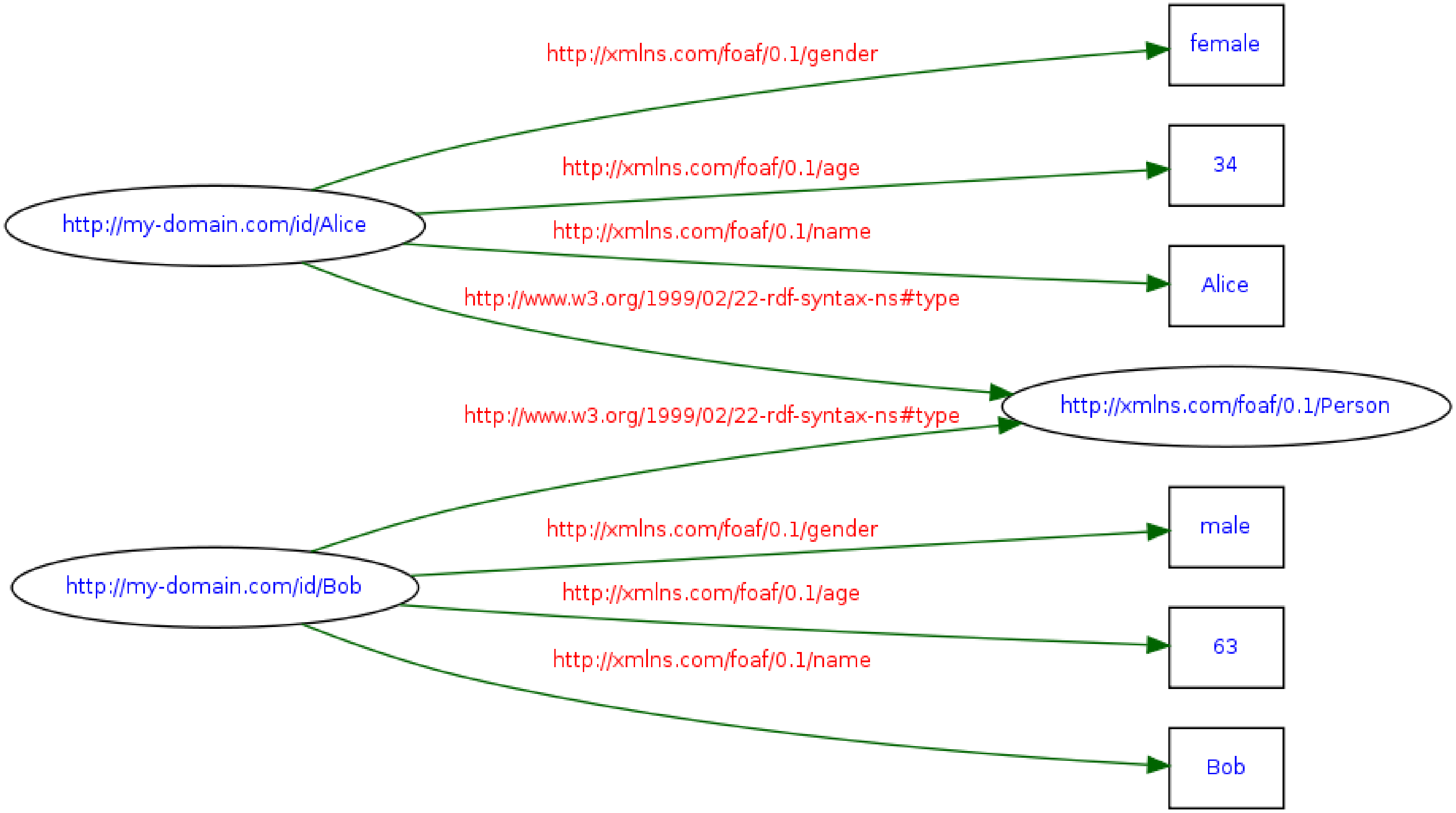
Define RDF mapping...

Graph URI: http://my-domain.com/example

:person-uri
- → rdf:a → foaf:Person
- → foaf:age → age
- → foaf:gender → sex
- → foaf:name → name

Done

Result of the process


http://xmlns.com/foaf/0.1/gender
female
http://xmlns.com/foaf/0.1/age
34
http://my-domain.com/id/Alice
http://xmlns.com/foaf/0.1/name
Alice
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://xmlns.com/foaf/0.1/Person
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://xmlns.com/foaf/0.1/gender
male
http://my-domain.com/id/Bob
http://xmlns.com/foaf/0.1/age
63
http://xmlns.com/foaf/0.1/name
Bob

DaPaaS RDF database-as-a-service

- Designed for live data services, instead of static datasets
  - A new RDF database can be operational within seconds
- Automated backups, operations, maintenance
- Based on an enterprise-grade RDF database
- Designed for scalability & availability, in the cloud
- Data import services (Grafter pipelines)

Related approaches for data cleaning and publication: WebKarma and OpenRefine

- Open-source software for data integration (support for mapping datasets to RDF)
- High-level functionality wrapped using GUI functions
  - E.g. importing, adding columns, modifying cells, etc.
  - More sophisticated GUI functionality includes: clustering, automatic reconciliation of the data, ontology mapping

+ Ready-to-use UI

+ Stable, tested

+ Support for many input formats out-of-the-box

- No programmatic/service APIs exposed

- Tight coupling hinders distribution

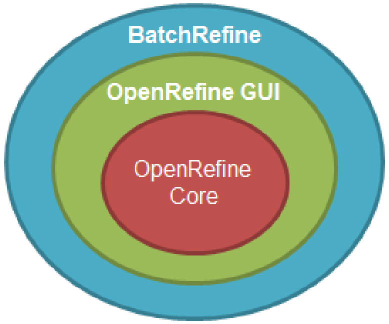
- No graphical DSL

Comparison with OpenRefine: Transformations over more than one dataset

OpenRefine

1. Defining the transformation
   a) Create new project and import dataset
   b) Define transformation through GUI
   c) Export JSON transformation
2. Transforming a new dataset (cannot be done in batch by default)
   a) Create new project and import dataset
   b) Import the JSON transformation
   c) Execute transformation and obtain result

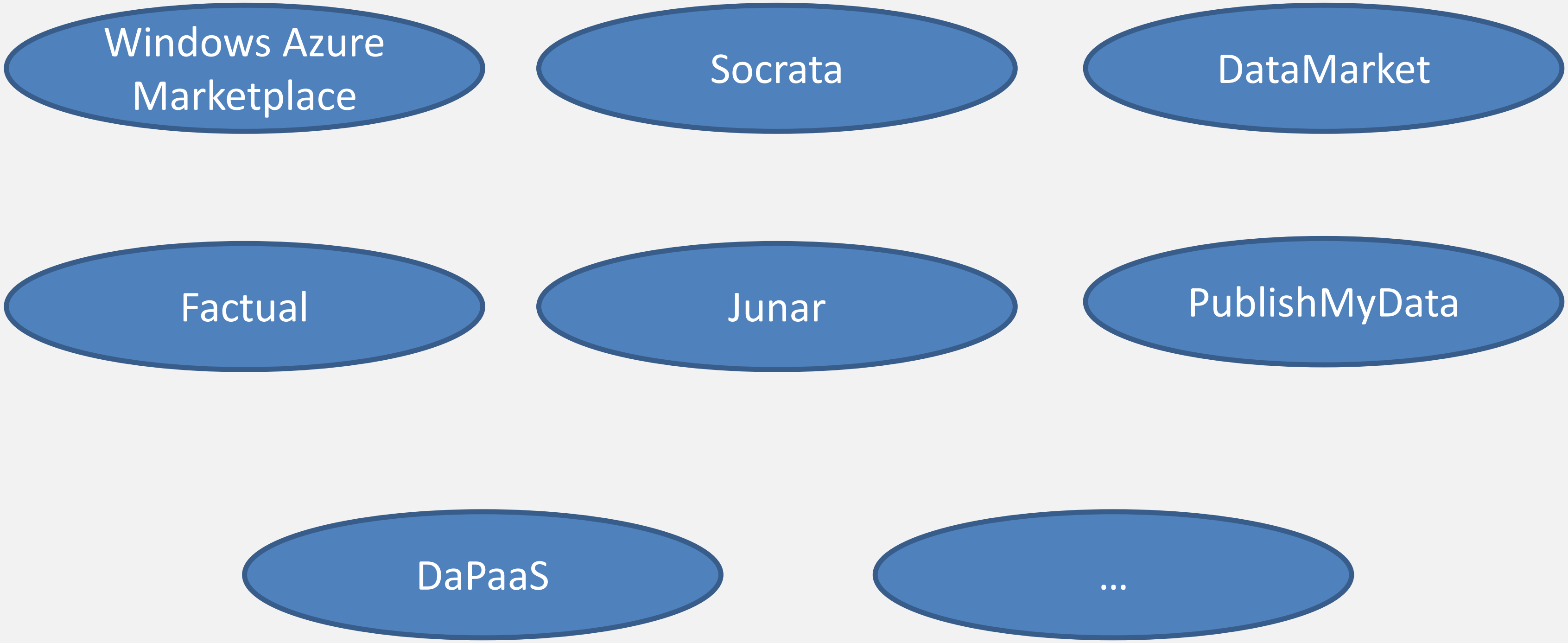
BatchRefine wraps the GUI APIs, instead of directly accessing the core



Grafter/Grafterizer (DaPaaS)

1. Defining the transformation
   a) Import dataset in Grafter GUI
   b) Define transformation through GUI
   c) Export and store **executable JAR** on the DaPaaS platform
2. Transforming a new dataset (in batch mode if necessary)
   a) Access the executable transformation through **REST service call** (dataset given as input parameter)

Relevant DaaS solutions



Windows Azure Marketplace

Socrata

DataMarket

Factual

Junar

PublishMyData

DaPaaS

…

Other relevant solutions

- **Comprehensive Knowledge Archive Network (CKAN)** (http://ckan.org/) – web-based open source data management system for the storage and distribution of open data; datahub (http://datahub.io/)

- **LOD2** (http://lod2.eu/) – research project aimed at providing an open source, integrated software stack for managing the lifecycle of Linked Data, from data extraction, enrichment, interlinking, to maintenance; not meant to be as-a-service solution

- **Project Open Data** (http://project-open-data.github.io/) – a set of open source tools, methodologies and use cases for publishing and utilising Open Data

- **COMSODE** (http://www.comsode.eu/) – research project aiming to create a publication platform for Open Data called Open Data Node

DaPaaS – targeted impacts

- A **reduction in the cost** for organisations (e.g. SMEs, public organizations, etc) which lack sufficient expertise and resources to publish open data

- A **reduction on the dependency** of open data publishers on generic Cloud platforms to build, deploy and maintain their open/linked data from scratch

- An **increase in the speed of publishing** new datasets and updating existing datasets

DaPaaS – targeted impacts (cont’)

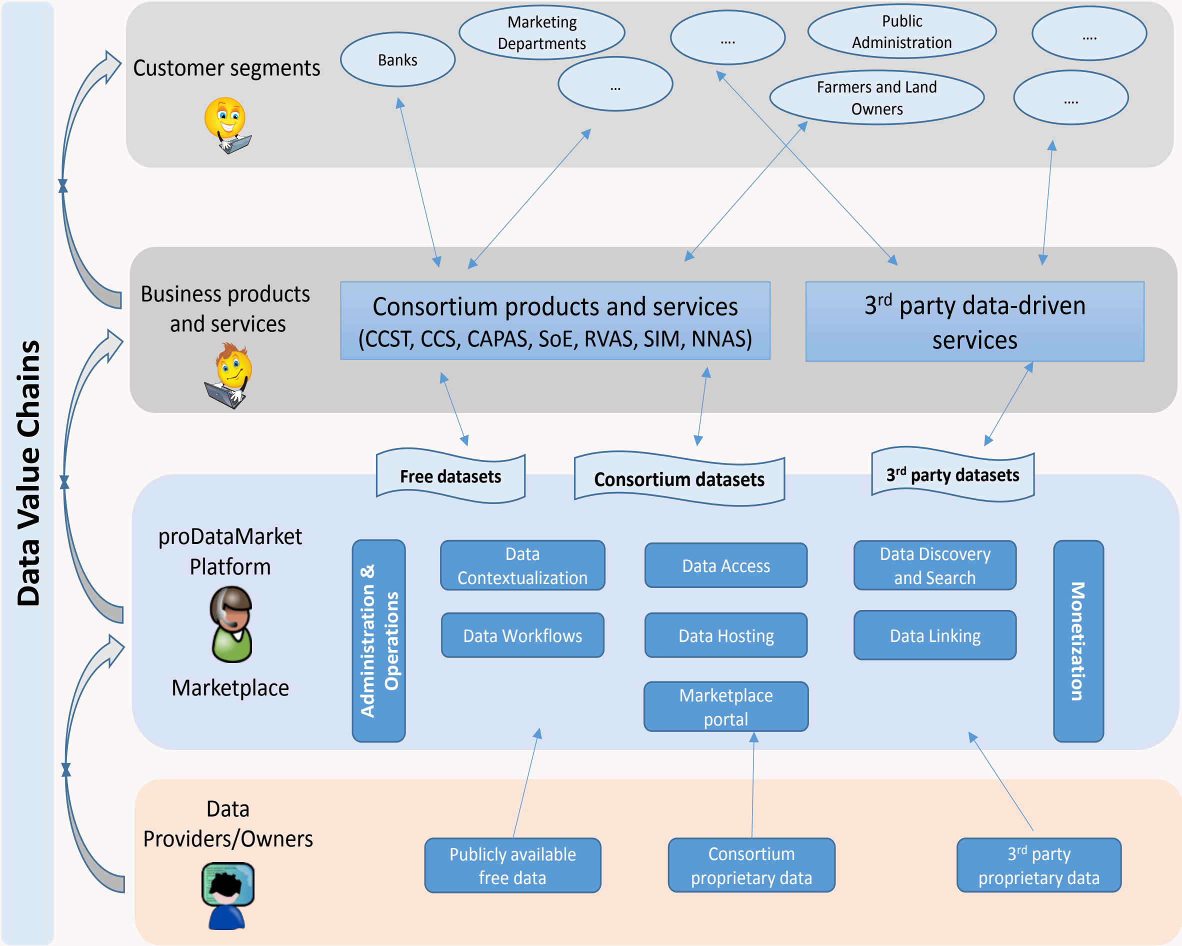
- A **reduction in the cost and complexity of developing** applications that use open data

- An **increase in the reuse of open data** by providing fast and seamless access to numerous open data sets to the applications hosted on the DaPaaS platform

Follow-up project: proDataMarket

http://prodatamarket.eu/

- How can we innovate (and make money) with (property-related) Open Data?
- H2020 Innovation action
- Duration: 2015-2017
- Budget: ~ 3.4M Euro



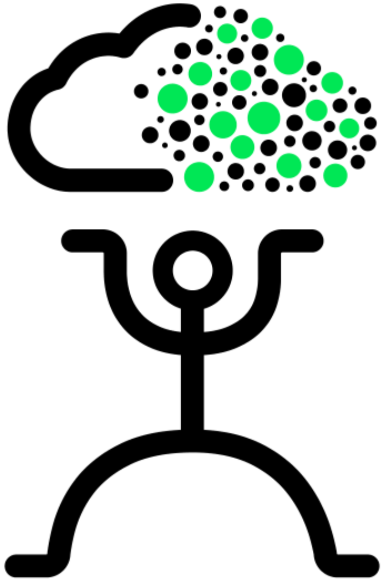
Summary

- Lots of open datasets, but very few actually used (e.g. low number of applications using them)

- Linked Data is a promising technology for Open Data, but difficult to use for publishers, developers, data workers

- DaPaaS – emerging solution (as-a-Service) for making Open (Linked) Data more accessible
  - Platform, portal, methodology, APIs
  - (Repeatable) Data Transformation is a core aspect of DaPaaS
  - Public release expected this year – stay tuned!

Thank you!



http://dapaas.eu

@dapaasproject

dapaas-platform@googlegroups.com

Contact: dumitru.roman@sintef.no

Event announcement

- "Data Labs" – Open Data Workshop/Tutorial
- When: July 2nd 2015
- Where: Oslo, Norway
- Organized by The ODI and SINTEF in the context of DaPaaS