DataGraft
Data-as-a-Service for Open Data
Opportunities for Publishing Property Data

https://datagraft.net

Dumitru Roman

dumitru.roman@sintef.no

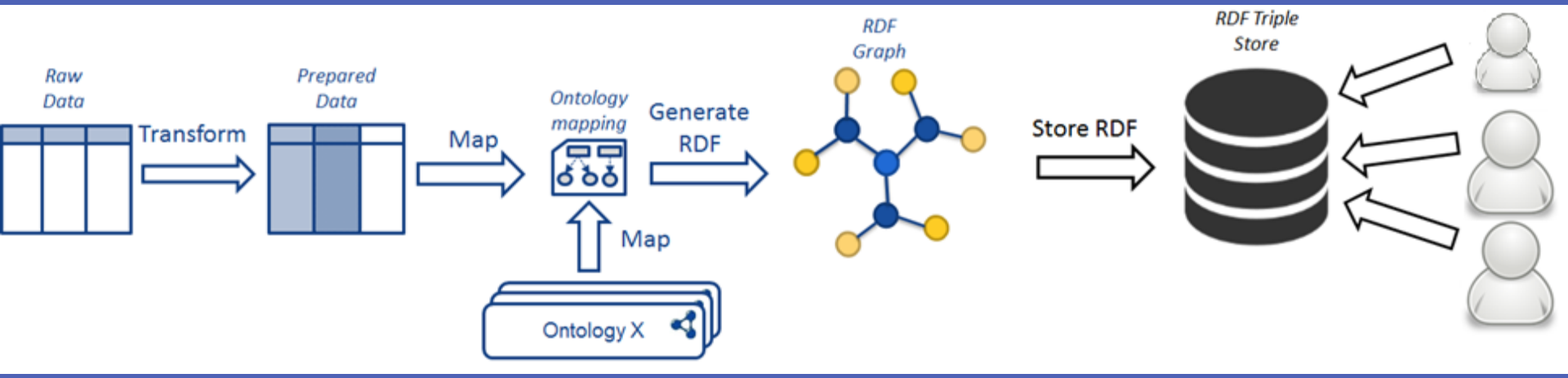
Outline

- What is DataGraft
- DataGraft in SmartOpenData
  - TRAGSA and ARPA Data Publishing
- DataGraft for Property Data

Developed to allow
data workers
to manage their data in a
simple, effective, and efficient way

Powerful
data transformation and
reliable data access capabilities

Data Transformation and RDF Publication Process



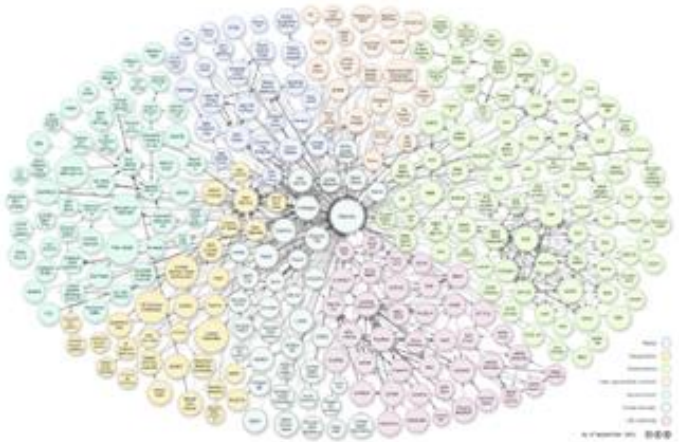
- Interactive design of transformations?
- Repeatable transformations?
- Reuse/share transformations (user-based access)?
- Cloud-based deployment of transformations?
- Self-serviced process?
- Data and Transformation as-a-Service?

DataGraft: Data-as-a-Service

For the Data Transformation and RDF Publication Process

Sharable/Repeatable/Reusable

Transformation



Tabular Data

Graph Data

DataGraft key feature: Flexible management and sharing of data and transformations

Interactively build, modify and share data transformations

Share transformations privately or publicly

Reuse transformations to repeatably clean and transform spreadsheet data

Fork, reuse and extend transformations built by other professionals from DataGraft's transformations catalog

Programmatically access transformations and the transformation catalogue

DataGraft key feature:
Reliable data hosting and querying services

Host data on DataGraft's reliable, cloud-based triplestore

Share data privately or publicly

Query data through your own SPARQL endpoint

Programmatically access the data catalogue

DataGraft beta

Search | Explore | Dashboard | Publish | Transform | SmartOpenDataTransformations

Data transformations / TRAGSA ChemicalCharacteristics SoilAcidity SoilPermeability-fork / TRAGSA3 soilAcidity soilPermeability 2

METADATA | PIPELINE | RDF MAPPING | CLOJURE | PREVIEWED DATA | ORIGINAL DATA

Edit prefixes | Edit utility functions

- make-dataset
- drop-rows
- derive-column
- derive-column
- derive-column
- derive-column
- derive-column
- derive-column
- mapc

idLitholo	idPermeab	IdAcidity	permeab-en	permeab-es	permeab-pt	acidity-en	acidity-es
101	3	2	High	Alta	Alta	Neutral	Neutro
102	3	2	High	Alta	Alta	Neutral	Neutro
10	1	3	Low	Baja	Baixa	Acid	Ácido
11	1	3	Low	Baja	Baixa	Acid	Ácido
12	1	3	Low	Baja	Baixa	Acid	Ácido
13	1	1	Low	Baja	Baixa	Basic	Básico
15	1	1	Low	Baja	Baixa	Basic	Básico
16	1	3	Low	Baja	Baixa	Acid	Ácido
210	1	1	Low	Baja	Baixa	Basic	Básico
21	1	1	Low	Baja	Baixa	Basic	Básico
22	2	2	Medium	Media	Média	Neutral	Neutro
2310	1	0	Low	Baja	Baixa	No data	No data
23	1	0	Low	Baja	Baixa	No data	No data
310	1	3	Low	Baja	Baixa	Acid	Ácido
33	2	3	Medium	Media	Média	Acid	Ácido

Automatic preview

Documentation | API | FAQ | Terms of use | Privacy policy | Cookie policy | Contact

Data transformations / TRAGSA ChemicalCharacteristics IdWorkUnit idLitholo

METADATA | PIPELINE | RDF MAPPING | CLOJURE

Title

TRAGSA ChemicalCharacteristics IdWorkUnit idLitholo

Description

Expose as public

Publisher: **smod**

Data transformations / TRAGSA Climatology Insolation / Aux 032000 Insolation 1

Fork

METADATA | PIPELINE | RDF MAPPING | CLOJURE

Edit prefixes

Edit utility functions

make-dataset

mapc

PREVIEWED DATA | ORIGINAL DATA

idInsol	Definition	Min	Max	idClustered	DefCl
900	Less than 900	0	900	1	Very Lov
1000	900 - 1000	900.1	1000	1	Very Lov
1100	1000 - 1100	1000.1	1100	1	Very Lov
1200	1100 - 1200	1100.1	1200	2	Low
1300	1200 - 1300	1200.1	1300	2	Low
1400	1300 - 1400	1300.1	1400	2	Low
1500	1400 - 1500	1400.1	1500	2	Low
1600	1500 - 1600	1500.1	1600	2	Low

Automatic preview

DataGraft beta

Search | Explore | Dashboard | Publish | Transform | SmartOpenDataTransformations

Data transformations / TRAGSA ChemicalCharacteristics SoilAcidity SoilPermeability-fork

METADATA | PIPELINE | RDF MAPPING | CLOJURE

Map the tabular data to RDF

Graph URI

http://data.smartopendata.eu/sp-pt-pilot/

(base-soil) *idLitholo*	rdf:a	(smod) Soil
	smod:soilPermea...	*permeab-en*
	smod:soilPermea...	*permeab-es*
	smod:soilPermea...	*permeab-pt*
	smod:soilAcidity	*acidity-en*
	smod:soilAcidity	*acidity-es*
	smod:soilAcidity	*acidity-pt*

```
(defpipe my-pipe "Pipeline to convert tabular data into a different tabular format." [data-file]
  (-> (read-dataset data-file)
      (-> (make-dataset move-first-row-to-header)
          (rename-columns (comp keyword string-as-keyword)))
      (drop-rows 1)
      (derive-column :permeab-en [:idPermeab] map_permeability_code_to_definition_en)
      (derive-column :permeab-es [:idPermeab] map_permeability_code_to_definition_es)
      (derive-column :permeab-pt [:idPermeab] map_permeability_code_to_definition_pt)
      (derive-column :acidity-en [:IdAcidity] map_acidity_code_to_definition_en)
      (derive-column :acidity-es [:IdAcidity] map_acidity_code_to_definition_es)
      (derive-column :acidity-pt [:IdAcidity] map_acidity_code_to_definition_pt)
      (mapc {:permeab-en string-literal
             :permeab-es string-literal
             :permeab-pt string-literal
             :acidity-en string-literal
             :acidity-es string-literal
             :acidity-pt string-literal})))
```

```
(def make-graph
  (graph-fn [{:keys
              [permeab-en permeab-es permeab-pt acidity-en acidity-es acidity-pt idLitholo]}]
   (graph "http://data.smartopendata.eu/sp-pt-pilot/"
          [(base-soil idLitholo)
           [rdf:a (smod "Soil")]
           [(smod "soilPermeabilityRate") permeab-en]
           [(smod "soilPermeabilityRate") permeab-es]
           [(smod "soilPermeabilityRate") permeab-pt]
           [(smod "soilAcidity") acidity-en]
           [(smod "soilAcidity") acidity-es]
           [(smod "soilAcidity") acidity-pt]])))
```

DataGraft beta

Search

Explore Dashboard Publish Transform SmartOpenDataTransformations

Explore / Dashboard

My data pages

Search

Data page	Date	Portal	Actions
TRAGSA Climatology Pluviometry	Y-day		
TRAGSA Climatology Temperature Annual Average	Y-day		
TRAGSA Climatology IdTempAA	Y-day		
TRAGSA Climatology Evapotranspiration	Y-day		
TRAGSA Climatology Temperature Annual Minimum	Y-day		
TRAGSA Climatology Temperature Annual Maximum	Y-day		
TRAGSA Climatology Humidity	Y-day		
TRAGSA Climatology Insolation	Y-day		
TRAGSA Climatology Radiation	Y-day		
TRAGSA Climatology RunOff	Y-day		
TRAGSA Climatology IdTempAA	Y-day		
AnimalSpecies	1 Sep		
Previewed datasets			

My transformations

Search

Transformation	Date	Actions
TRAGSA ChemicalCharacteristics SoilAcidity SoilPermeability	3 Sep	
TRAGSA WorkUnitLocation IdDistrict IdNuts	2 Sep	
TRAGSA WorkUnitLocation IdMuni IdDistrict	2 Sep	
TRAGSA WorkUnitLocation IdNeighbor IdMuni	2 Sep	
TRAGSA WorkUnitLocation IdWorkUnit	2 Sep	
TRAGSA WorkUnitLocation IdWorkUnit idCLC00	2 Sep	
TRAGSA WorkUnitLocation IdWorkUnit idCLC06	2 Sep	
TRAGSA WorkUnitLocation IdWorkUnit idCLC90	2 Sep	
TRAGSA WorkUnitLocation IdWorkUnit idLandSp	2 Sep	
TRAGSA WorkUnitLocation IdWorkUnit idForestry	2 Sep	
TRAGSA WorkUnitLocation IdWorkUnit idParcel	2 Sep	
TRAGSA WorkUnitLocation Parcels	2 Sep	
TRAGSA WorkUnitLocation Municipality	1 Sep	
TRAGSA WorkUnitLocation Neighborhood	1 Sep	
TRAGSA WorkUnitLocation Districts	1 Sep	
TRAGSA ChemicalCharacteristics SoilAcidity SoilPermeability	1 Sep	
TRAGSA ChemicalCharacteristics IdWorkUnit idLitholo	1 Sep	
TRAGSA Climatology Humidity	31 Aug	
TRAGSA Climatology Insolation	31 Aug	
TRAGSA Climatology Radiation	31 Aug	

← Previous Next →

DataGraft beta

Search Explore Sign In

Explore / Data Page

Preview ARPA Lakes Monitoring Stations

Data page properties

Name: ARPA Lakes Monitoring Stations

Description: Contains information about monitoring stations of lakes in 2014

Owner: sdn

Creation Date: 6 Sep 2015

Keyword: Lakes_dati2013_caricati2014,

SPARQL

Endpoint: https://rdf.datagraft.net/4830355550/db/repositories/1507022227_arpa-lakes-monitoring-stations

Query Query Builder

1:

EXECUTE

Table Results

EXPORT RDF EXPORT RAW VIEW PORTAL

SPARQL

Endpoint: https://rdf.datagraft.net/4831509243/db/repositories/1508276910_graft-computed-transformation-8

Query Query Builder

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX smod: <http://www.w3.org/2015/03/inspire/smod#>

SELECT DISTINCT ?soil

WHERE
{
?soil smod:soilPermeabilityRate ?soilPermeabilityRate .
?soil smod:soilAcidity ?soilAcidity .
?soil smod:soilAcidity "Acid" .
?soil smod:soilPermeabilityRate "High" .
}
```

EXECUTE

Table Results

Show 10 entries Search:

soil
http://data.smartopendata.eu/sp-pt-pilot/Soil/55
http://data.smartopendata.eu/sp-pt-pilot/Soil/64
http://data.smartopendata.eu/sp-pt-pilot/Soil/82
http://data.smartopendata.eu/sp-pt-pilot/Soil/35

Showing 1 to 4 of 4 entries Previous 1 Next

APIs

- Security and authentication
  - HTTPS
  - Authentication
  - CORS
- Model reference
- Usage
- Datasets Catalog
- RDF Repositories
- Transformations Catalog
- Grafter Transformations
  - Tabular transformation
  - Graft Transformation

Datasets Catalog

LIST ALL DATASETS

List of datasets catalog records using the DCAT vocabulary in RDF or JSON-LD.

Use the **showShared** header parameter to include the public datasets.

GET `/catalog/datasets/catalog` List user's datasets

Parameters Hide

showShared `y|n` (optional)
defaults to `n`

Accept `mimetype` (required)
`application/ld+json` or `application/rdf+xml`

Request Show

Response `200` Show

SEARCH DATASETS

Text search on the datasets metadata.

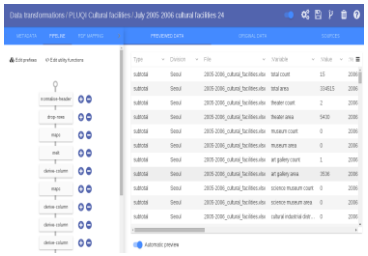
Use the **showShared** header parameter to include the public datasets.

GET `/catalog/datasets/search` Search user's datasets

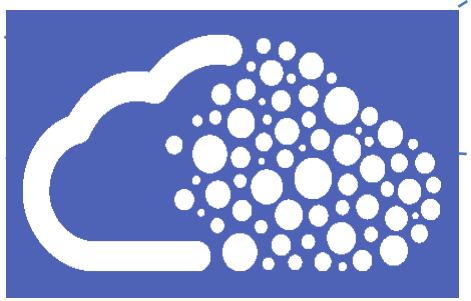
DataGraft Enablers

Grafter

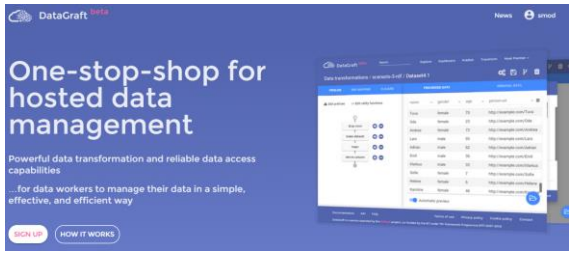
Grafterizer



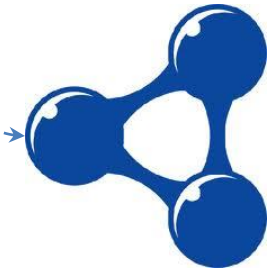
DataGraft



Data Portal



RDF DBaaS



DataGraft in SmOD: Use Cases

TRAGSA Pilot

- Number of transformations: 42
  - Created via reuse: 25
- Number of triples:
  - ~ 7.7M

ARPA Pilot

- Number of transformations: 5
  - Created via reuse: 2
- Number of triples:
  - ~ 14K

DataGraft in SmOD: Preliminary observations

- Positive aspects
  - Forking/reusing transformations helped us spend less time on creating new transformations
  - Possibility to edit parameters of each transformation step and change step order at any moment of creating the transformation made it easier to:
    - Create transformations in general
    - Correct mistakes made during transformation steps
    - Try the effects of transformation steps with different parameters
  - Custom code as utility functions provided flexibility in reuse of functions across transformations

- Cleaning data lacked some "nice to have" functionality, e.g. joining or sorting datasets
  - This was overcome with some preprocessing of the input files (e.g. 27 of 43 files needed some initial preprocessing in the TRAGSA pilot)

DataGraft for Property Data

Why property data?

One of the most valuable datasets managed by governments worldwide

Extensively used in various domains by private and public organizations

Some challenges in working with property data

- Difficult to access
- Cross-sectors
- Data is highly heterogeneous and possibly large
- Data quality
- Time-consuming integration
- Lack of innovation
- …

http://prodatamarket.eu

DataGraft – 1 package 2 audiences

DataGraft – targeted impacts

Reduction in costs
for organisations (e.g. SMEs, public organizations, etc.) which lack sufficient expertise and resources to publish open data

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

Increase in the speed of publishing
new datasets and updating existing datasets

Reduction in the cost and complexity of developing
applications that use open data

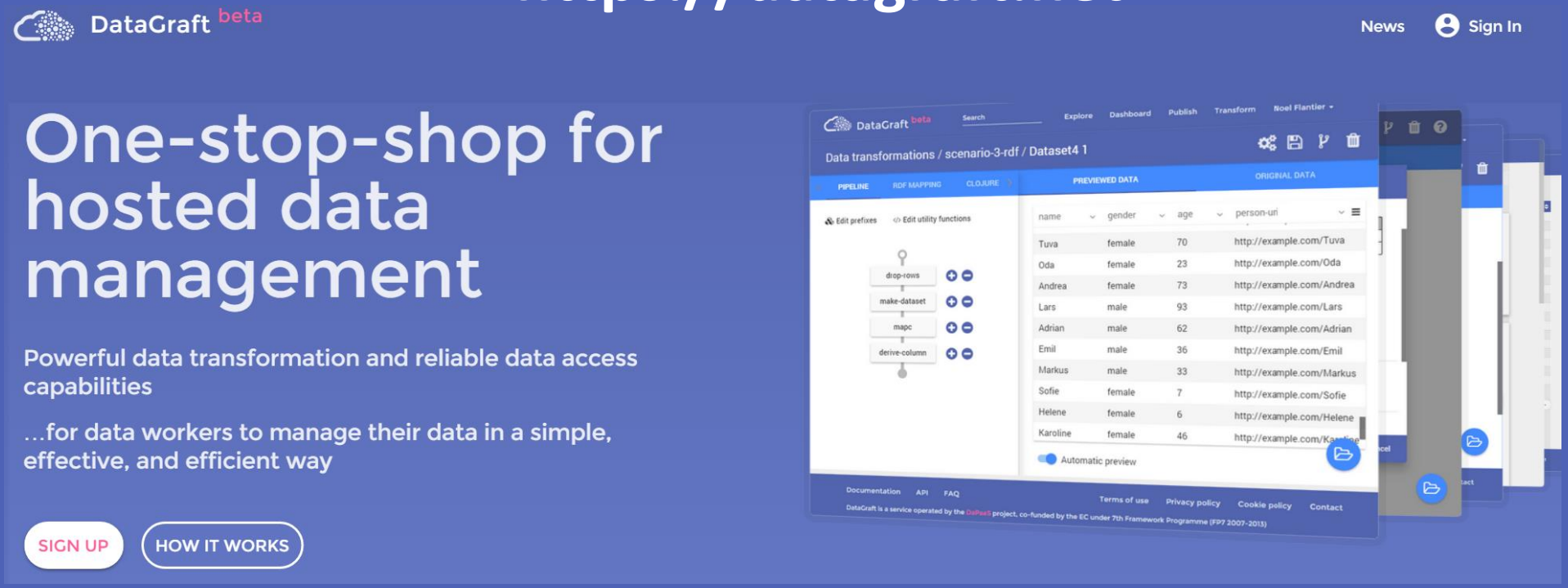
Increase in the reuse of open data
by providing reliable access to numerous open data sets to the applications hosted on DataGraft.net

Summary

- DataGraft – emerging solution (as-a-Service) for making Open (Linked) Data more accessible
  - Platform, portal, methodology, APIs
  - Developed/Operated by **DaPaaS**, with contributions from **SmOD**, **proDataMarket**, **OpenCube**
  - Successfully applied in SmOD for two pilot cases

- Key features:
  - Support for Sharable/Repeatable/Reusable Data Transformations
  - Reliable RDF Database-as-a-Service

https://datagraft.net



Thank you!

Contact: dumitru.roman@sintef.no