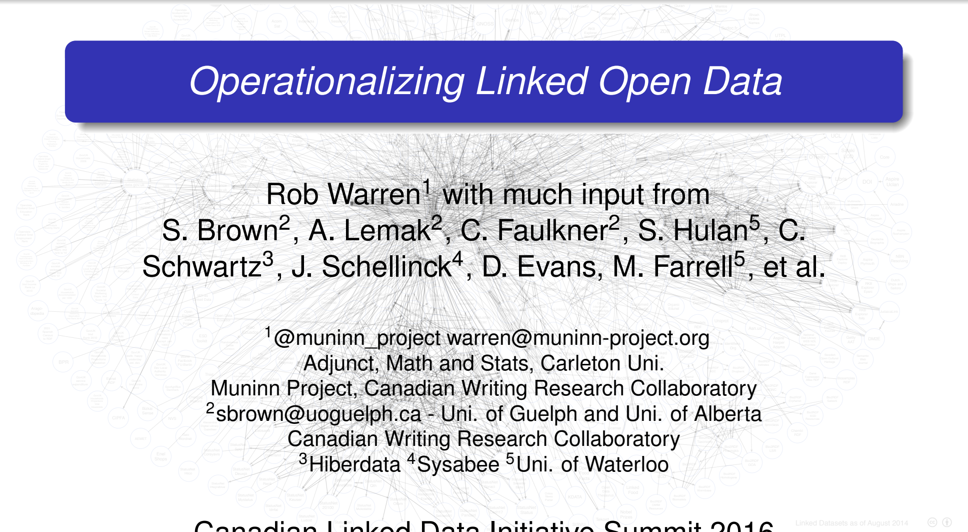


Linked Datasets as of August 2014

Operationalizing Linked Open Data

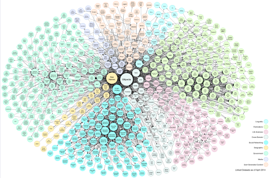
Rob Warren[1] with much input from
S. Brown[2], A. Lemak[2], C. Faulkner[2], S. Hulan[5], C. Schwartz[3], J. Schellinck[4], D. Evans, M. Farrell[5], et al.

[1]@muninn_project warren@muninn-project.org
Adjunct, Math and Stats, Carleton Uni.
Muninn Project, Canadian Writing Research Collaboratory
[2]sbrown@uoguelph.ca - Uni. of Guelph and Uni. of Alberta
Canadian Writing Research Collaboratory
[3]Hiberdata [4]Sysabee [5]Uni. of Waterloo

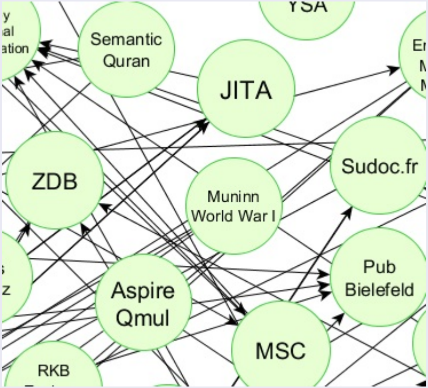
Canadian Linked Data Initiative Summit 2016
`https://github.com/rwarren2/cldisummit`

Who am I?

LOD Cloud 2014



Muninn WW1



Who am I?

CWRC

CWRC / CSÉC

UNIVERSITY OF ALBERTA

linked modernisms

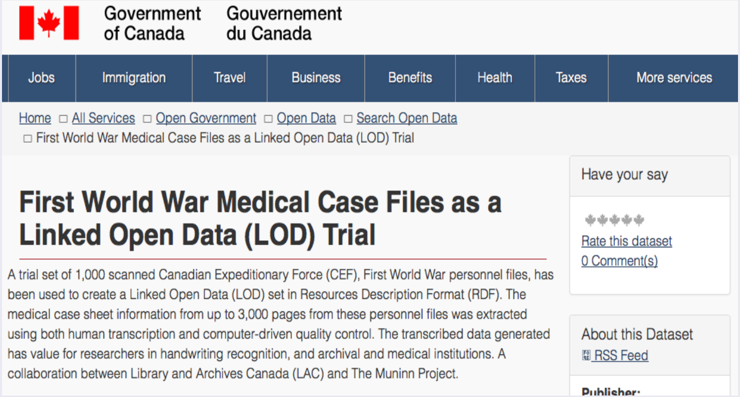
Canada Foundation for Innovation

UNIVERSITY of GUELPH

Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada

compute canada | calcul canada

First ★★★★★ data set on the Canada Open Data Portal



Government of Canada Gouvernement du Canada

Home □ All Services □ Open Government □ Open Data □ Search Open Data □ First World War Medical Case Files as a Linked Open Data (LOD) Trial

First World War Medical Case Files as a Linked Open Data (LOD) Trial

A trial set of 1,000 scanned Canadian Expeditionary Force (CEF), First World War personnel files, has been used to create a Linked Open Data (LOD) set in Resources Description Format (RDF). The medical case sheet information from up to 3,000 pages from these personnel files was extracted using both human transcription and computer-driven quality control. The transcribed data generated has value for researchers in handwriting recognition, and archival and medical institutions. A collaboration between Library and Archives Canada (LAC) and The Muninn Project.

Have your say

Rate this dataset
0 Comment(s)

About this Dataset
RSS Feed

Publisher:

Presentation Outline

1 Who am I?

2 **Why Linked (Open) Data?**

3 Field notes on vocabularies

4 Field notes on publishing data

5 Field notes on working with triples

The business value of LOD.

- Citations! If you can cite it, it exists!
- Externalize your costs to someone else.
- Document your data's idiosyncrasies.
- ~~Linked Data is just another fad.~~
- ~~It's already on my website.~~
- ~~People will steal my data.~~
- ~~There are errors is my data.~~

Observations:

1. There is a bigger market for the individual pieces of your publication than the whole of it.
2. There is a bigger market for your data with people that don't share your alphabet

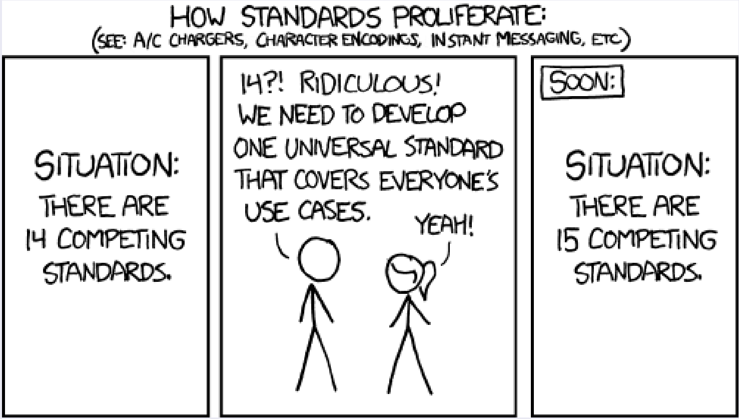
The propeller-head value of LOD.

- You have a machine readable URI to work with.
- You can support multiple serializations.
- You can still reference something, even if not “Open”.
- You can annotate the data to the *n*th degree.
- Easy provenance and tracking of changes.
- You get multiple languages and Unicode for free.

Observations:

1. Forces separation between the **data** and the **application**.
2. Your use cases for the data are never what people want out of the **application**.
3. *LOD engages with people by engaging their machines.*

Vocabularies: Use a standard. (Which one!?)


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

[a]

[a] https://xkcd.com/927/

Vocabulary use options:

1. Create your own.
2. Use one existing vocabulary.
3. Use multiple existing vocabularies.

The data consumer’s perspective:

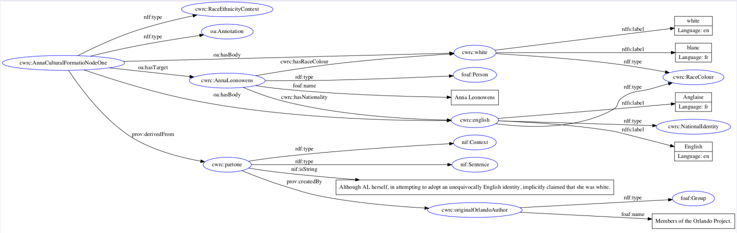
- Consumers want to know what to expect in vocabularies.
- Multiple vocabularies need relationships. (You build them).
- **The vast majority of data consumers cannot use ontology reasoning at query time.**

Case Studies:

Overview: CWRC (`http://www.cwrc.ca/`)

- Primarily Orlando TEI-style data.
- Schema definitions not ontologically sound.
- Custom ontology linked to other ontologies.
- Questions of ethnicity, race, skin colour alternate between vernacular and technical.

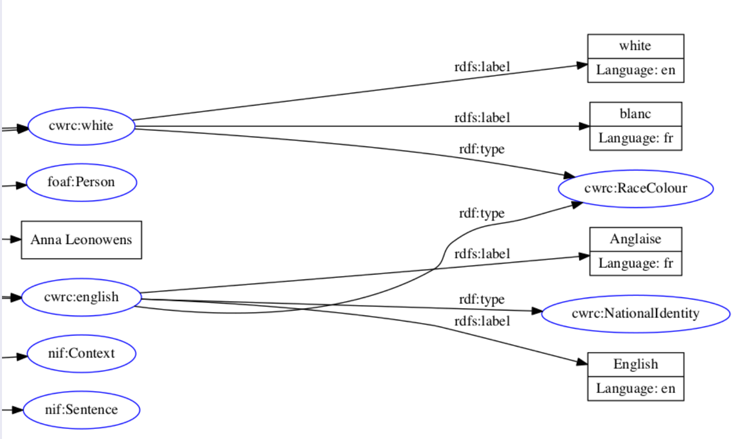
Case Studies: CWRC Entry



[a]

[a] Anna LeonOwens

Case Studies: CWRC Entry



[a]

[a] Anna LeonOwens

Outcomes

- The Ontology is a data explanation tool. Initially (wrongly) seen as a controlled vocabulary.
- Much time is being spent on teasing out the intent of the data as written.
- The process is very demanding of the scholars.
- The CWRC ontology in its final form will have paradoxes. Acceptable because it explains data that was not built within an *ontologically rational* framework.
- This is good enough for partial data exchange.
- Massive ancillary linkages to other dataset.

Case Studies:

Overview: Muninn (`http://rdf.muninn-project.org/`)

- Heterogeneous data sources: text, SQL, images, free form tabular.
- Erroneous, ambiguous and incomplete data.
- Multiple purpose built ontologies for specialized applications.
- Move to standardized ontologies as they become available. (re: Organization Ontology)
- No “single” truth, but you are free to decide for yourself.

Private Peat, by Harold R. Peat

I was sharing a box with a lad whom I heard the fellows call **Bob***.*

“You’re in the right direction-don’t turn round!”

Partial Information

```
<owl:oneOf rdf:parseType="Collection">
<owl:Thing rdf:about="Bob #1"/>
<owl:Thing rdf:about="Bob #2"/>
<owl:Thing rdf:about="Bob #3"/> ...
</owl:oneOf>
<rel:knowsByReputation
rdf:resource="The Mad Major"/>
```

Private Peat

Attestation Papers

DOB 1893-02-31 - February 31, 1893

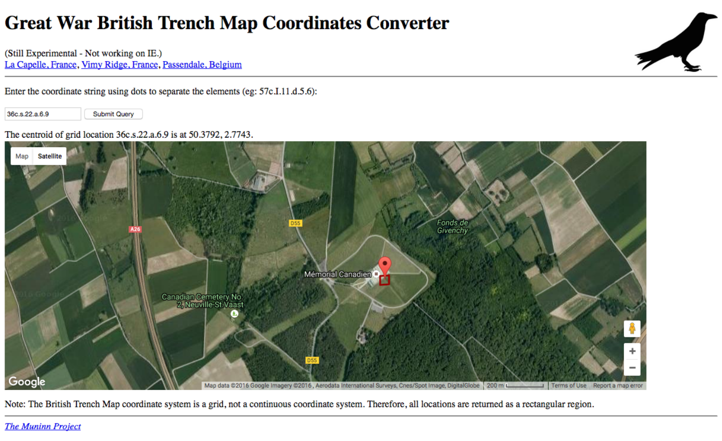
Partial Information

```
<owl:time rdf:about="Birth">
<time:hasDateTimeDescription>
<time:DateTimeDescription ...>
<time:year>1893</time:year>
<time:DateTimeDescription>
</time:hasDateTimeDescription>
<rdf:value>1893-02-31</rdf:value>
</owl:time>
```

Harry Baird

Case Studies Muninn:

British Trench Map Coordinate Translation App



Great War British Trench Map Coordinates Converter

(Still Experimental - Not working on IE.)
La Capelle, France, Vimy Ridge, France, Passendale, Belgium

Enter the coordinate string using dots to separate the elements (eg: 57c.I.11.d.5.6):

36c.s.22.a.6.9

The centroid of grid location 36c.s.22.a.6.9 is at 50.3792, 2.7743.

Map
Satellite
A26
D55
Fonds de Givenchy
Mémorial Canadien
Canadian Cemetery No. 2, Neuville-St. Vaast
D55
+
−
Google
Map data ©2016 Google Imagery ©2016 , Aerodata International Surveys, Cnes/Spot Image, DigitalGlobe
200 m
Terms of Use
Report a map error

Note: The British Trench Map coordinate system is a grid, not a continuous coordinate system. Therefore, all locations are returned as a rectangular region.

The Muninn Project

Field notes on Vocabularies: Conclusions

1. **The public interacts with Applications not Data, but Data is why we are here.**
2. Do not ever design vocabulary for the application.
3. Old data is never clean, sensical or well behaved. The ontology / vocabulary has to say so and work with it.
4. Reuse vocabularies and create new ones on a case by case basis.
5. Great resource at `https://lov.okfn.org/dataset/lov`

Publishing Linked Data:

Checklist:

- Dereferencable (URI’s for everything)?
- Content negotiation (**The data format is dead.**)?
- Public facing SPARQL server?
- Machine and Human readable vocabulary definition?
- Machine and Human readable data set definition?
- Production, in-house use of the SPARQL on day 1?

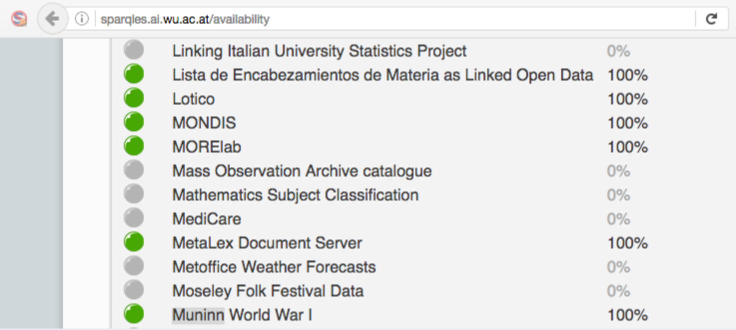
Important Note: People write bad programs.

“If builders built buildings the way programmers make programs, the first woodpecker to come along would destroy civilization.” - Gerald Weinberg

Corollary:

Get someone who knows public facing infrastructure to look things over for you.

SPARQL servers



sparqles.ai.wu.ac.at/availability

⚪	Linking Italian University Statistics Project	0%
🟢	Lista de Encabezamientos de Materia as Linked Open Data	100%
🟢	Lotico	100%
🟢	MONDIS	100%
🟢	MOREIab	100%
⚪	Mass Observation Archive catalogue	0%
⚪	Mathematics Subject Classification	0%
⚪	MediCare	0%
🟢	MetaLex Document Server	100%
⚪	Metoffice Weather Forecasts	0%
⚪	Moseley Folk Festival Data	0%
🟢	Muninn World War I	100%

SPARQL allows for custom retrieval queries over HTTP without having you involved.

An important note about SPARQL

Run SPARQL queries through a reverse HTTP proxy: ngix, polipo, etc.

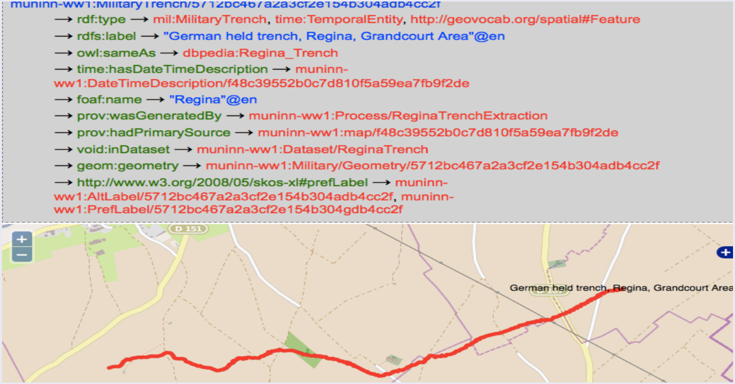
Why?

- Offending programmers can be safely ignored.
- Allows for light infrastructure abuse (auto-complete queries).
- Improves performance without heavy planning.

Tracking data in large data stores:

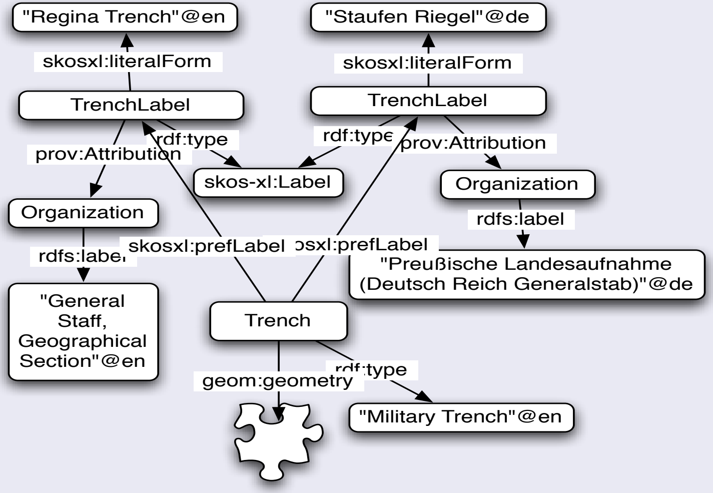
- Generate more data as a byproduct of operations:
  It is easier to delete old triples than to rebuild triples that should have existed.
- Tracking provenance of node is trivial; consider building it into your work flow.
- Data and meta-data are merging.
- The most awesome use of your data is a use case you have not thought of.

Dealing with contentious issues (1/2):



muninn-ww1:MilitaryTrench/5712bc467a2a3cf2e154b304adb4cc2f
→ rdf:type → mil:MilitaryTrench, time:TemporalEntity, http://geovocab.org/spatial#Feature
→ rdfs:label → "German held trench, Regina, Grandcourt Area"@en
→ owl:sameAs → dbpedia:Regina_Trench
→ time:hasDateTimeDescription → muninn-ww1:DateTimeDescription/f48c39552b0c7d810f5a59ea7fb9f2de
→ foaf:name → "Regina"@en
→ prov:wasGeneratedBy → muninn-ww1:Process/ReginaTrenchExtraction
→ prov:hadPrimarySource → muninn-ww1:map/f48c39552b0c7d810f5a59ea7fb9f2de
→ void:inDataset → muninn-ww1:Dataset/ReginaTrench
→ geom:geometry → muninn-ww1:Military/Geometry/5712bc467a2a3cf2e154b304adb4cc2f
→ http://www.w3.org/2008/05/skos-xl#prefLabel → muninn-ww1:AltLabel/5712bc467a2a3cf2e154b304adb4cc2f, muninn-ww1:PrefLabel/5712bc467a2a3cf2e154b304gdb4cc2f

Dealing with contentious issues(2/2):



"Regina Trench"@en

skosxl:literalForm

TrenchLabel

rdf:type

prov:Attribution

Organization

rdfs:label

"General Staff, Geographical Section"@en

"Staufen Riegel"@de

skosxl:literalForm

TrenchLabel

rdf:type

prov:Attribution

skos-xl:Label

Organization

rdfs:label

skosxl:prefLabel

osxl:prefLabel

"Preußische Landesaufnahme (Deutsch Reich Generalstab)"@de

Trench

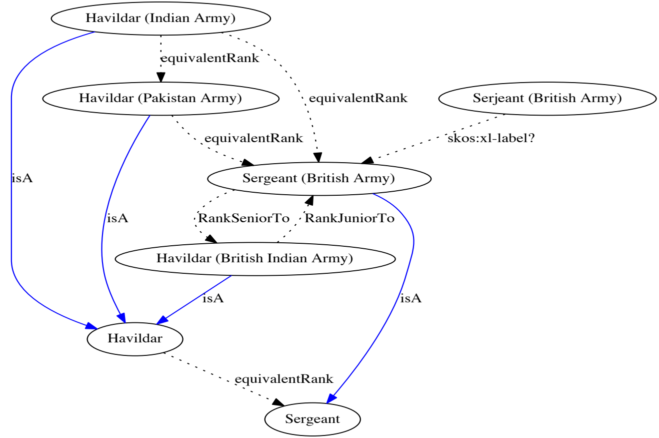
geom:geometry

rdf:type

"Military Trench"@en

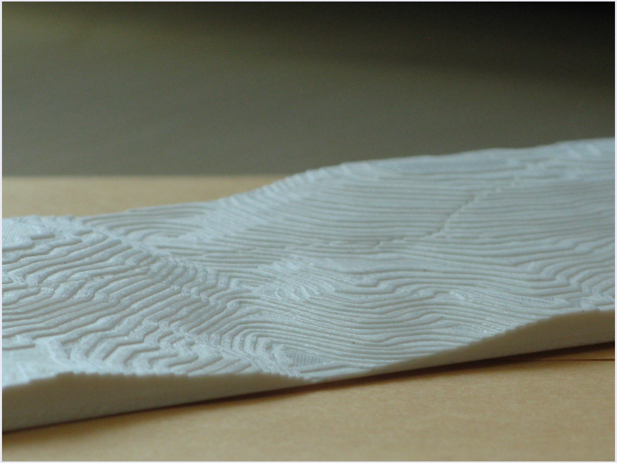
Important ontological note:

The *thing* and the *name of the thing* are not the same *thing*.


Havildar (Indian Army)
equivalentRank
Havildar (Pakistan Army)
equivalentRank
Serjeant (British Army)
equivalentRank
skos:xl-label?
Sergeant (British Army)
isA
isA
RankSeniorTo
RankJuniorTo
Havildar (British Indian Army)
isA
isA
Havildar
equivalentRank
Sergeant

Getting value out of low-value items:

Print your own Battlefield



Recap:

- Linked Open Data is about data, not applications.
- The *thing* and the *name of the thing* are not the same *thing*.
- The most awesome use of your data is a use case you have not thought of.
- Vocabulary use means something.
- LOD engages with people by engaging their machines.

Further information

- `http://www.cwrc.ca/`
- `http://www.muninn-project.org/`
- `https://www.youtube.com/watch?v=aJW16qFkGHU`

Questions?

CWRC / CSÉC

UNIVERSITY OF ALBERTA

linked modernisms

Canada Foundation for Innovation

UNIVERSITY of GUELPH

Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada
Canada

compute canada | calcul canada