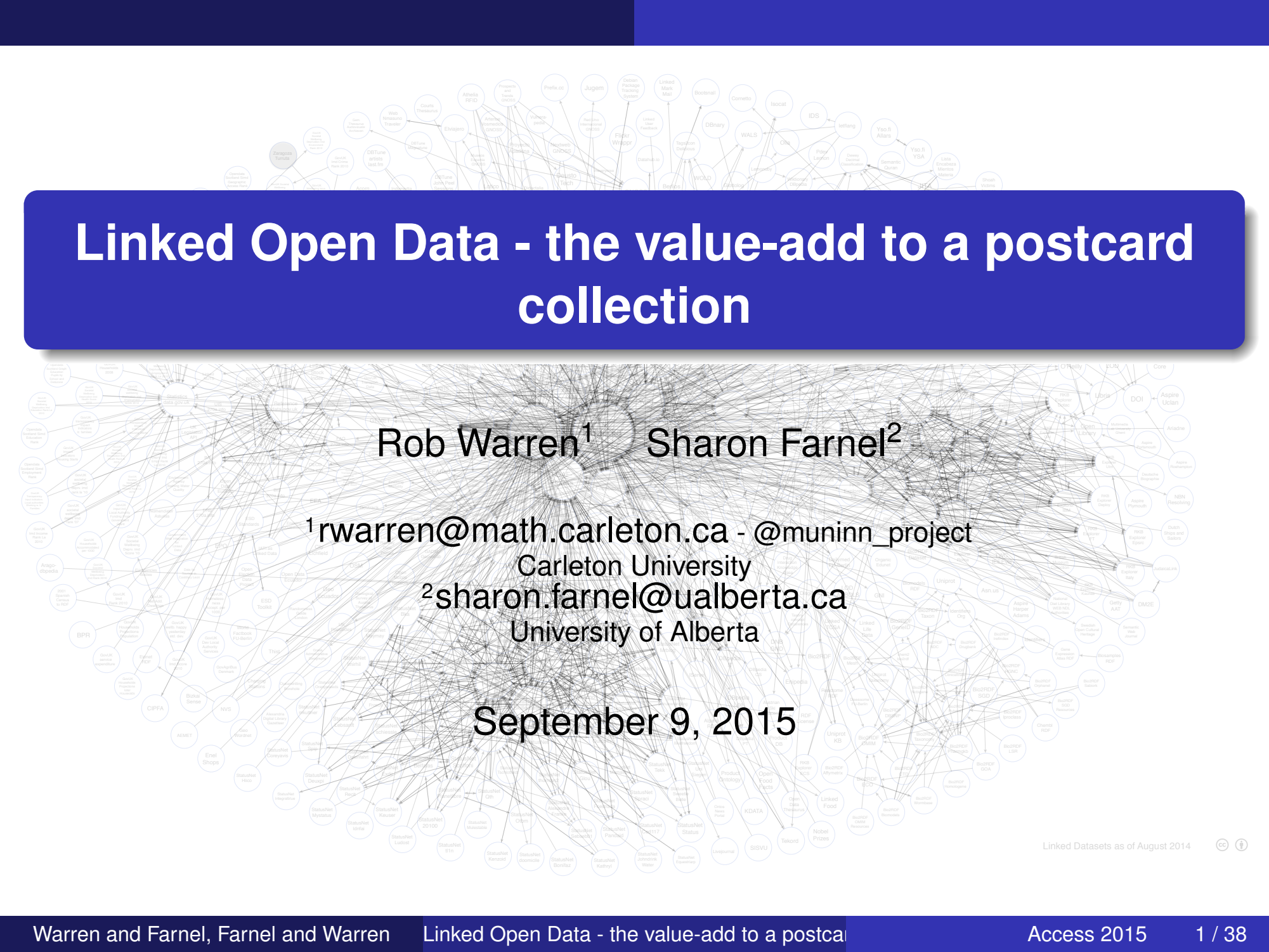
Linked Open Data - the value-add to a postcard collection

Rob Warren[1] Sharon Farnel[2]

[1]rwarren@math.carleton.ca - @muninn_project
Carleton University
[2]sharon.farnel@ualberta.ca
University of Alberta

September 9, 2015



Peel Prairie Postcards Collection

Prairie Postcards

Activities & sports 	Animals 	Buildings 
Business & industry 	Events 	Land & land use 
Natural phenomena 	Objects 	Organizations 
People 	Plants 	Vehicles 

Background

- Enable the documentation of decisions / processes using RDF and Prov ontology.
- Promote organizational memory without too much overhead.
- Clouds of clouds -> from collections of collections to custom organizations.
- How much can we offload to machines?

Project Budget

Sharon's Budget

Rob's Budget (Matching Funds)

Manual Meta-data creation

- 'Stub' records (place, ID, etc.) from CSV.
- In-house guidelines, use of authorities (e.g., LC).
- Largely one individual over 2 year period; other staff and student assistance prior to launch.
- Final QA before website launch; minor cleanup and enhancements ongoing.

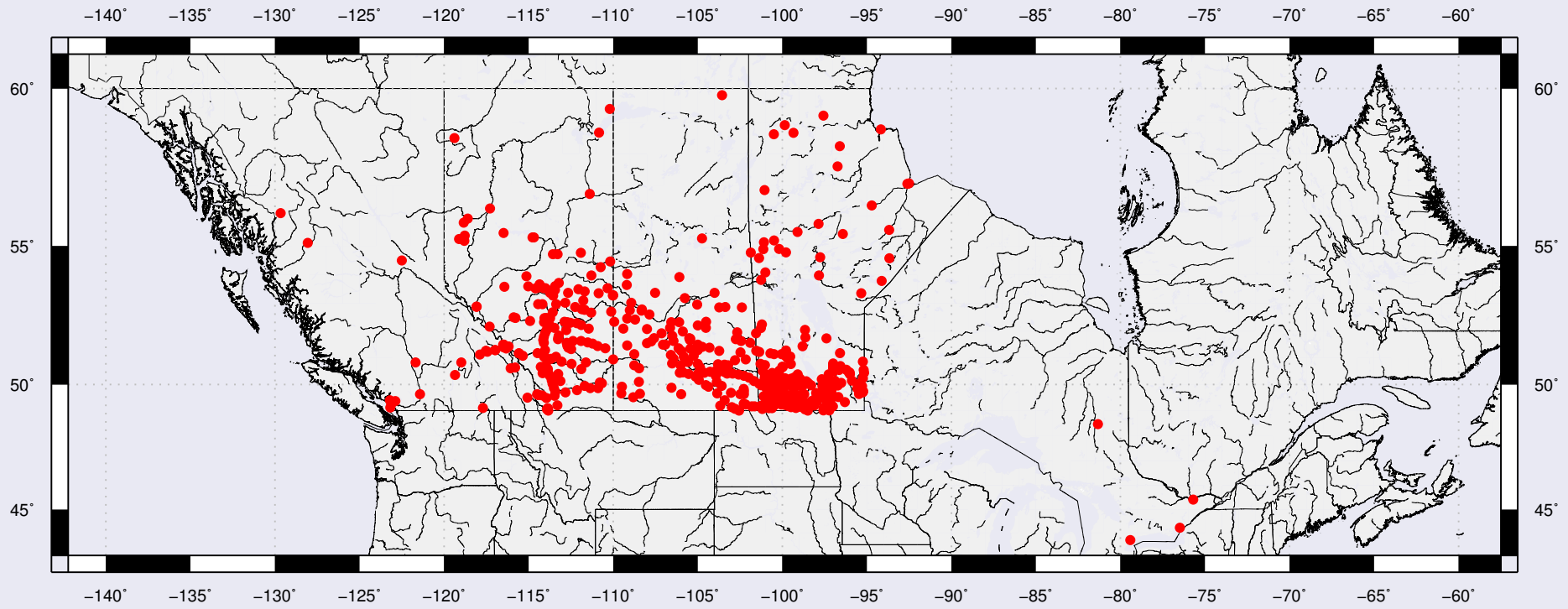
Tools and Processes

- Mods XML to rdf Mads via LC stylesheet (with minor edits).
- Additional RDF triples / linkages generated from raw XML Mods.
- Mostly perl, bash scripts and api's.

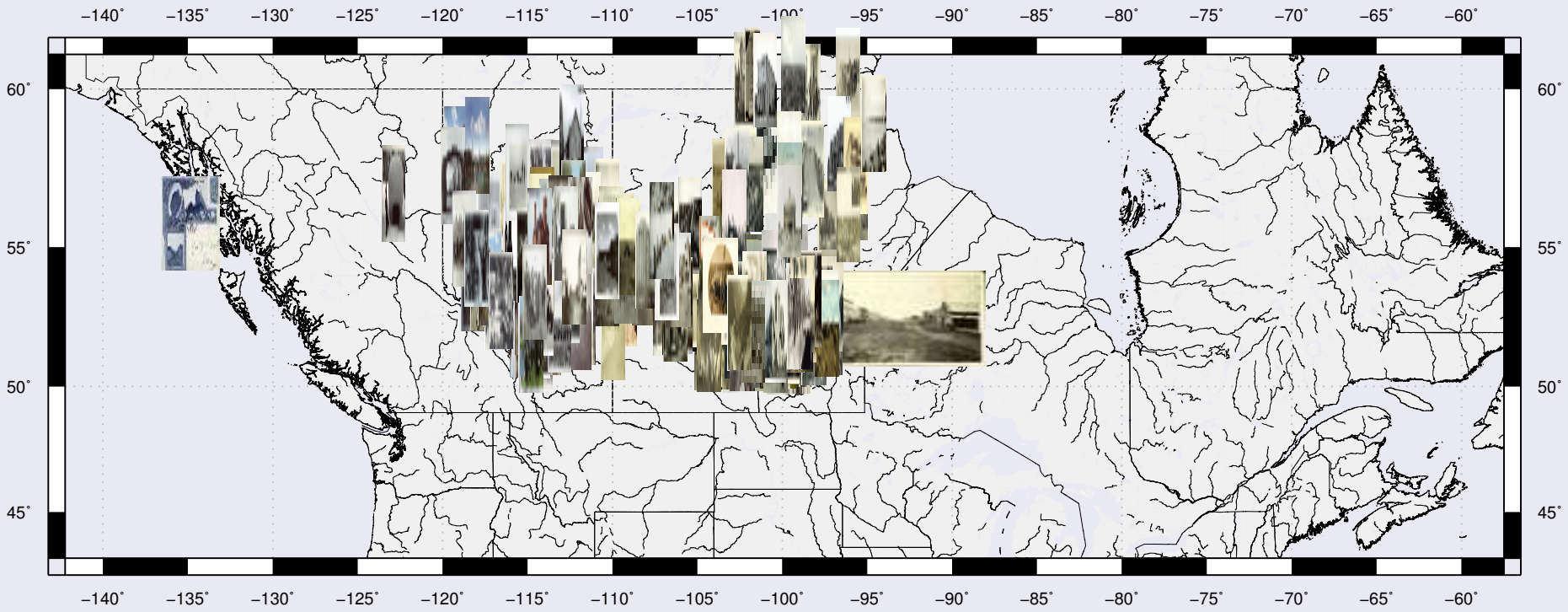
From strings to URIs

- Links to Geonames and Linked Geo Data (OSM).
- Links to LC Subject headings
- Try facial detection.
- Extract personal names and corporate names as foaf.
- Add full provenance information.
- ~~Extract colours from postcards: BW/Sepina?~~

URI's used


−140° −135° −130° −125° −120° −115° −110° −105° −100° −95° −90° −85° −80° −75° −70° −65° −60°
60°
55°
50°
45°

URI's used


−140° −135° −130° −125° −120° −115° −110° −105° −100° −95° −90° −85° −80° −75° −70° −65° −60°
60°
55°
50°
45°

Mads to geonames and linkgeodata

- Quality of location data ranges from great to plain wrong.
- Used geonames and linkedgeodata (OSM) api.
- Use OWL to deal with corner cases.
- Some data is likely unusable.

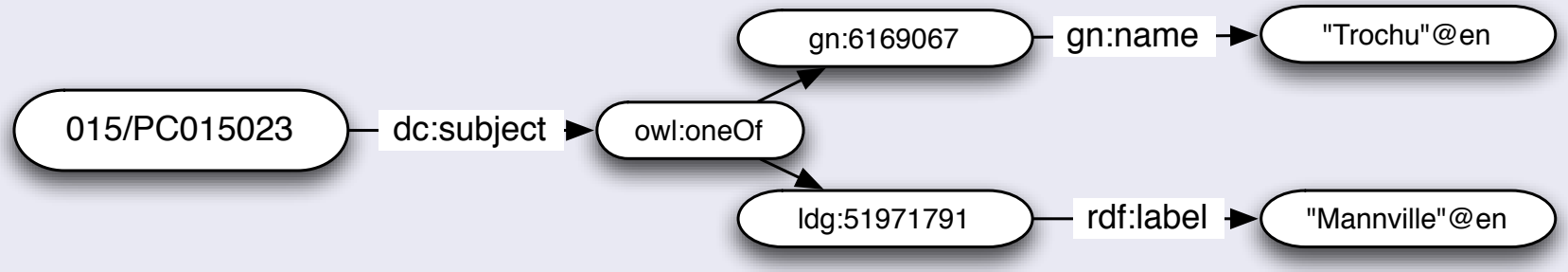
Data quality

People try very hard, but...

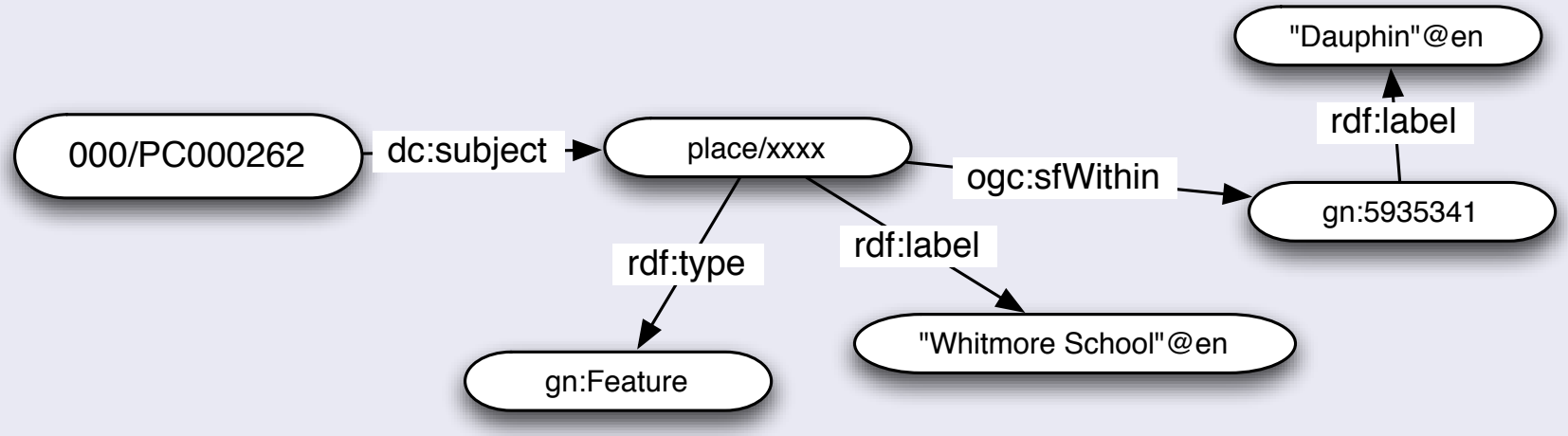
<city>Regina?</city>

??

<city>Trochu or Mannville</city>



<city>Whitmore School, Dauphin</city>

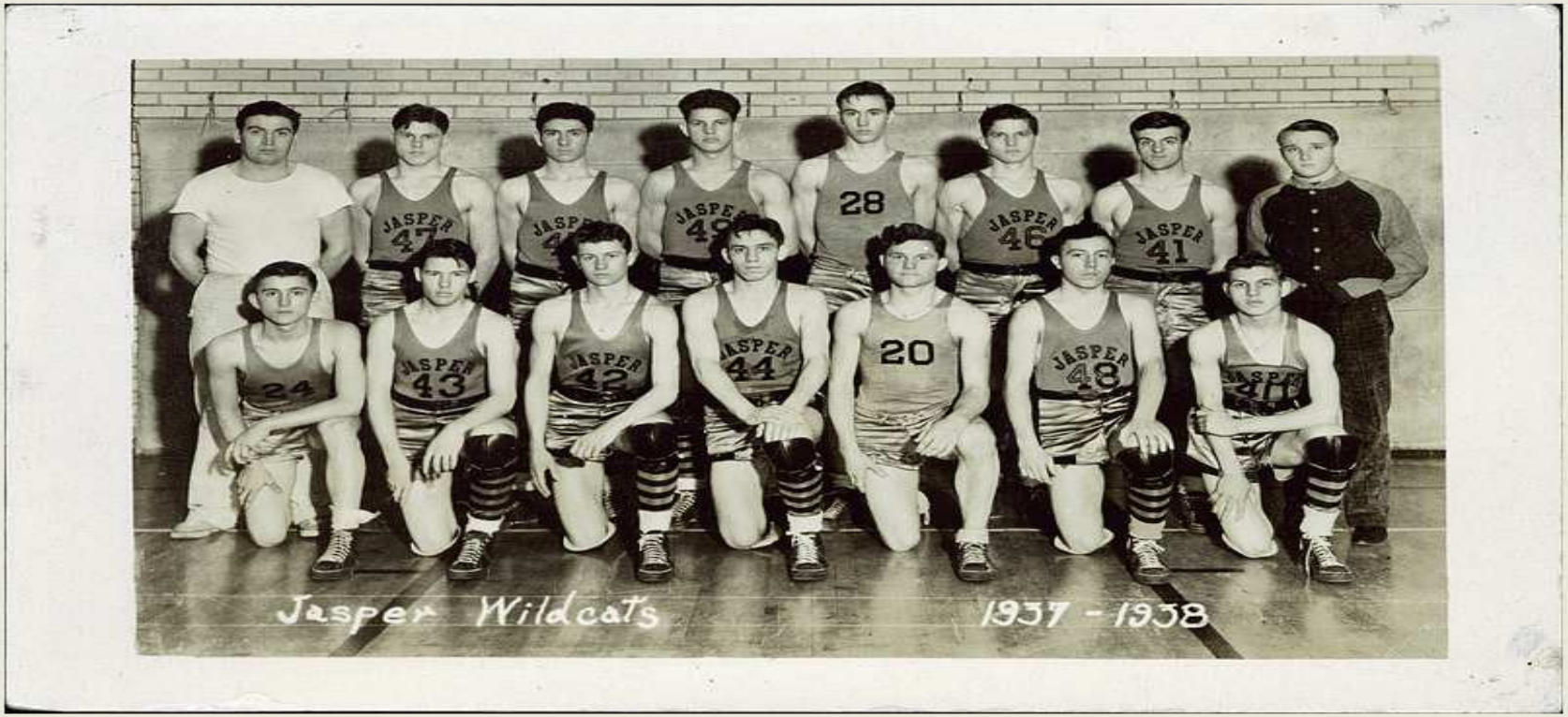


A running example

Postcard 8011

ShareThis

Jasper Wildcats, 1937-1938. [Jasper: c1937-1938.



Description: Group portrait of Wildcats basketball team players posing in two rows (standing and kneeling on their right knees) in front of a light colored background in a room.

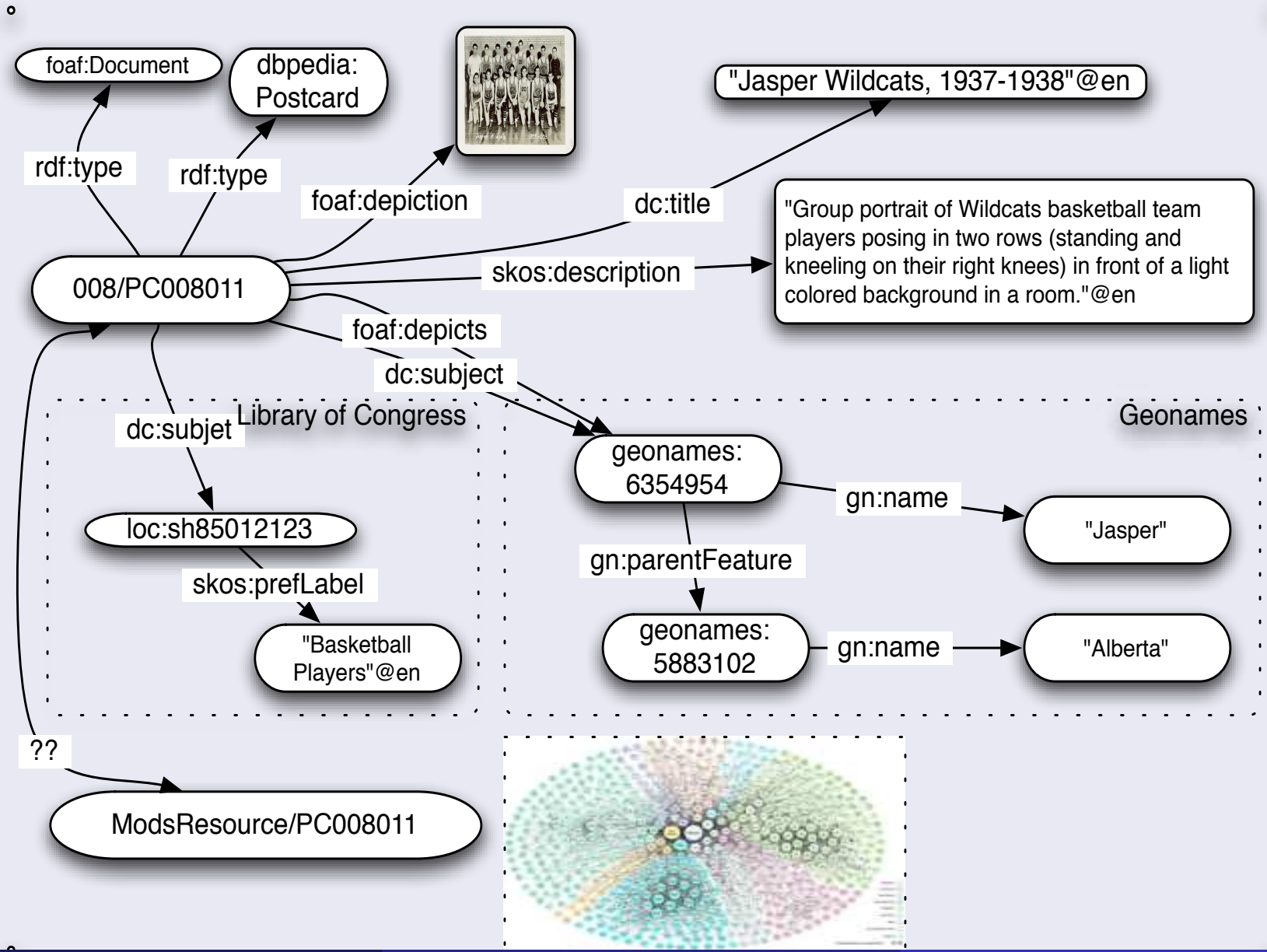
Physical description: 1 postcard : sepia ; 9 x 14 cm.

Language: English

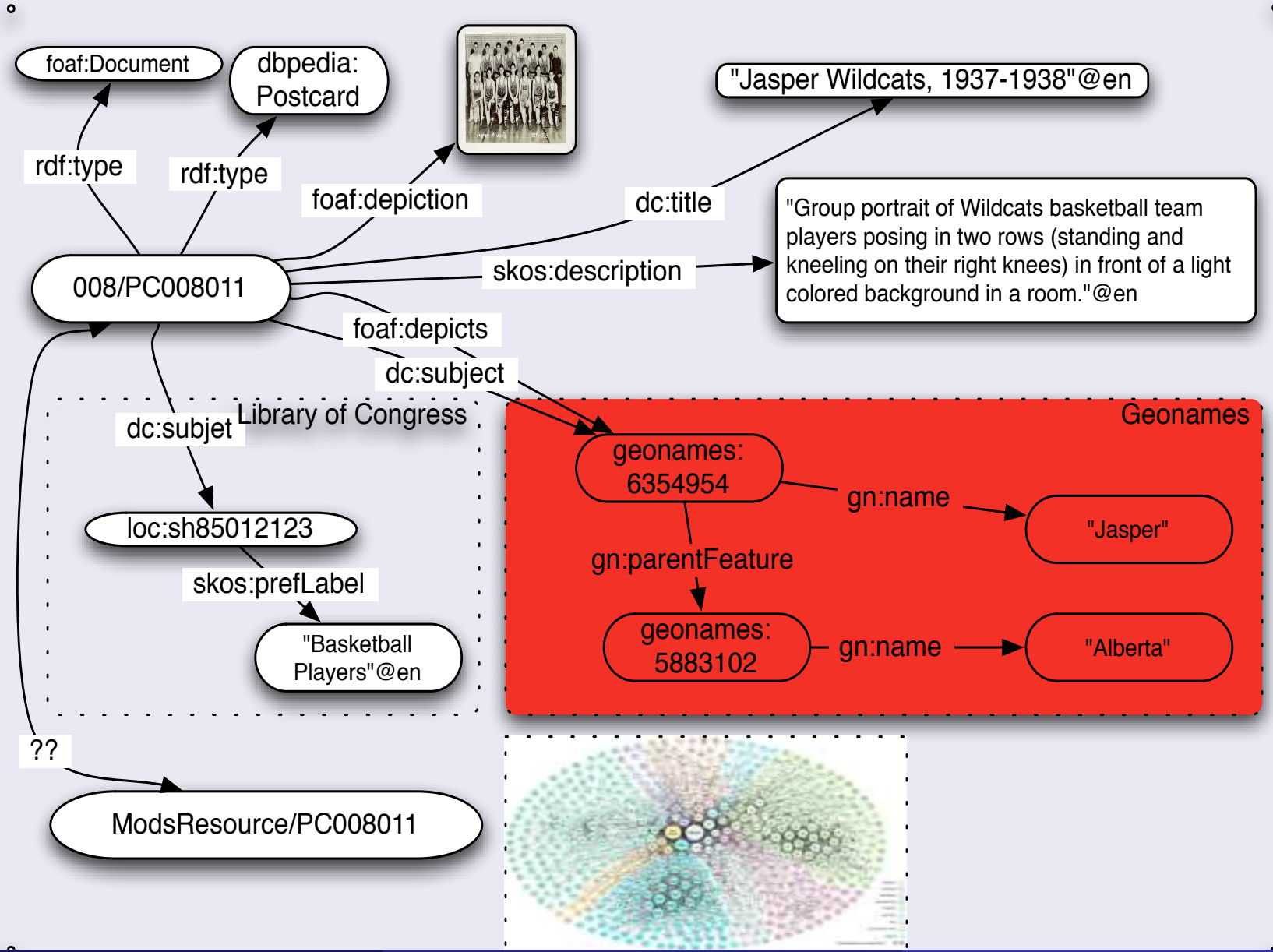
Subject headings:

- **Canada--Alberta--Jasper**
- **Basketball players**
- **Sport clothes**

Datasets / Process documents



Datasets / Process documents



Datasets / Process documents

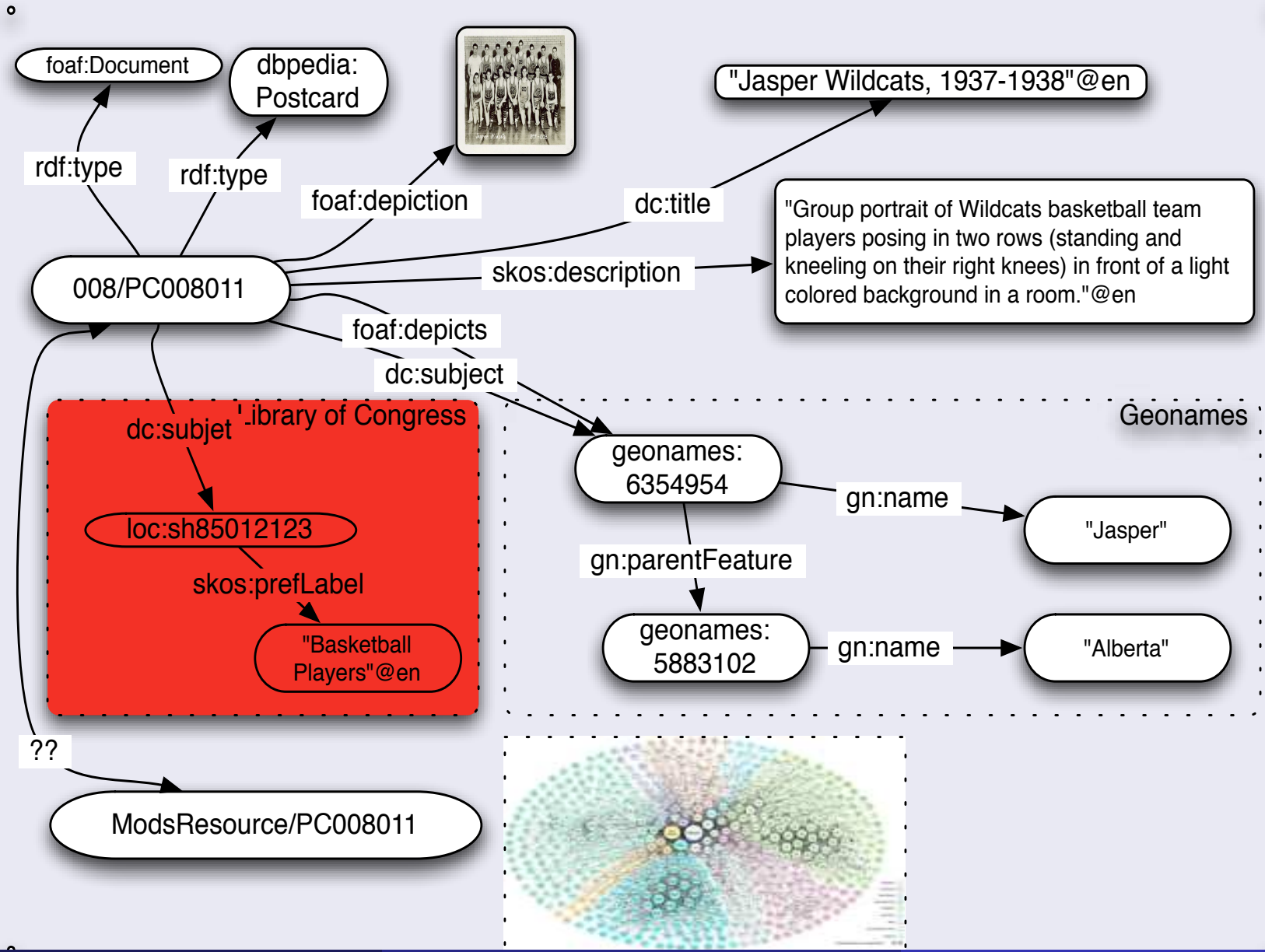


Image Analysis

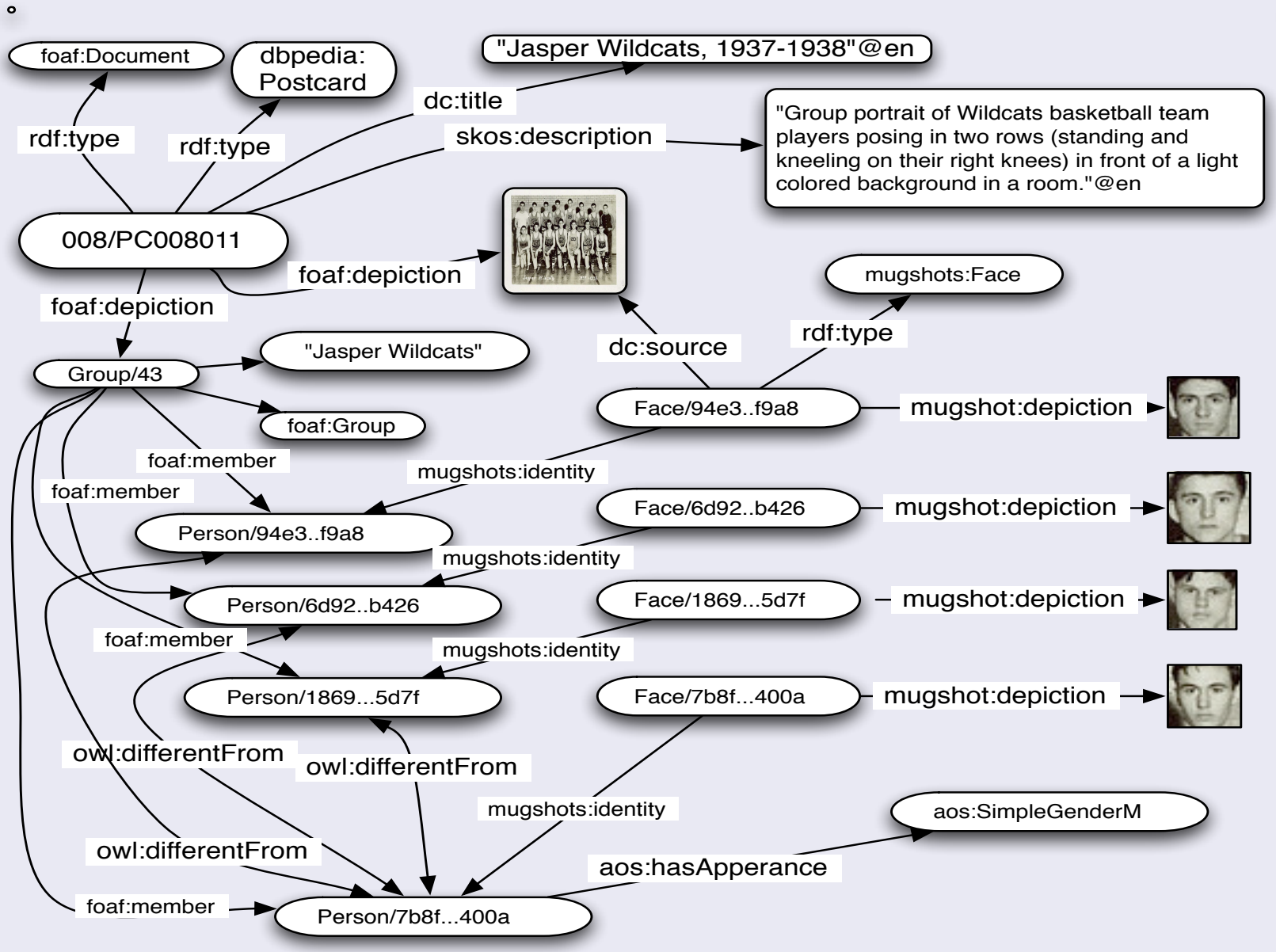


Image Analysis

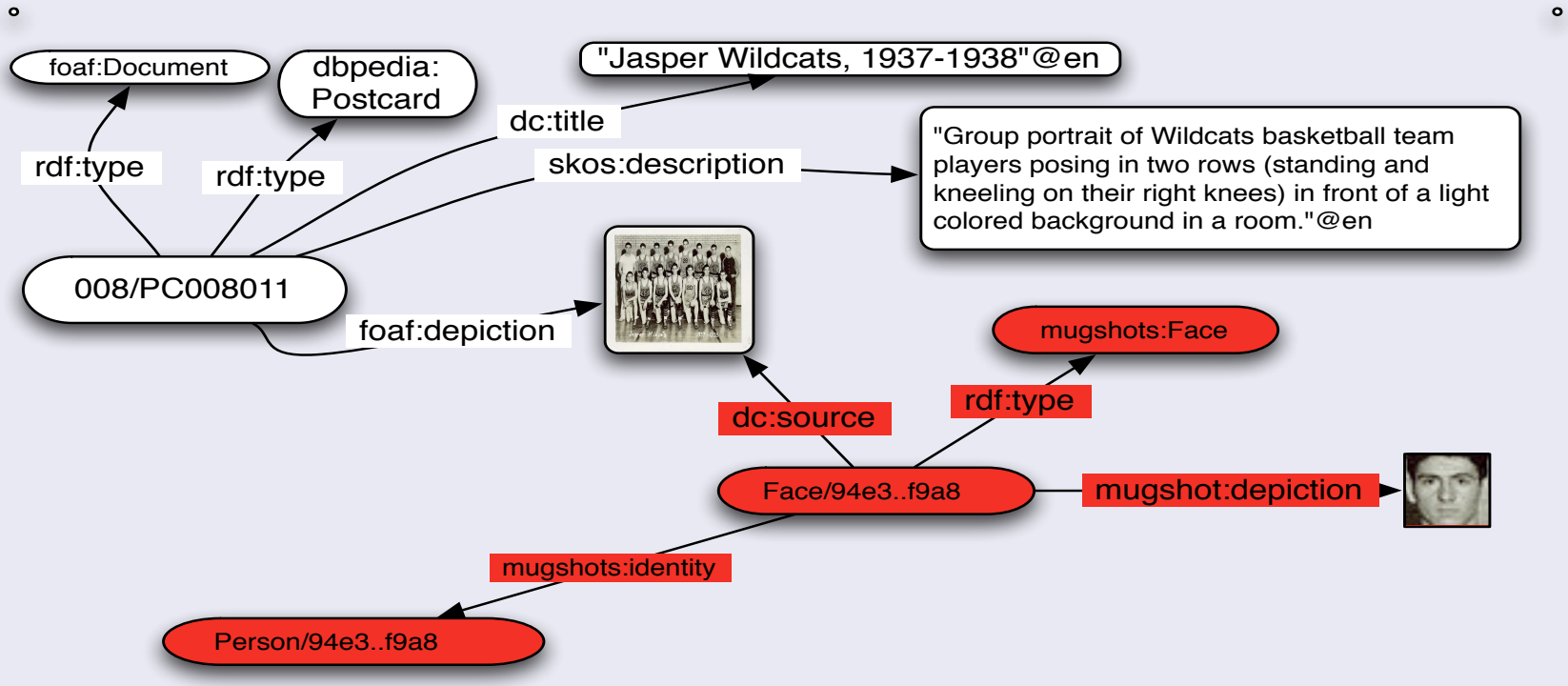


Image Analysis

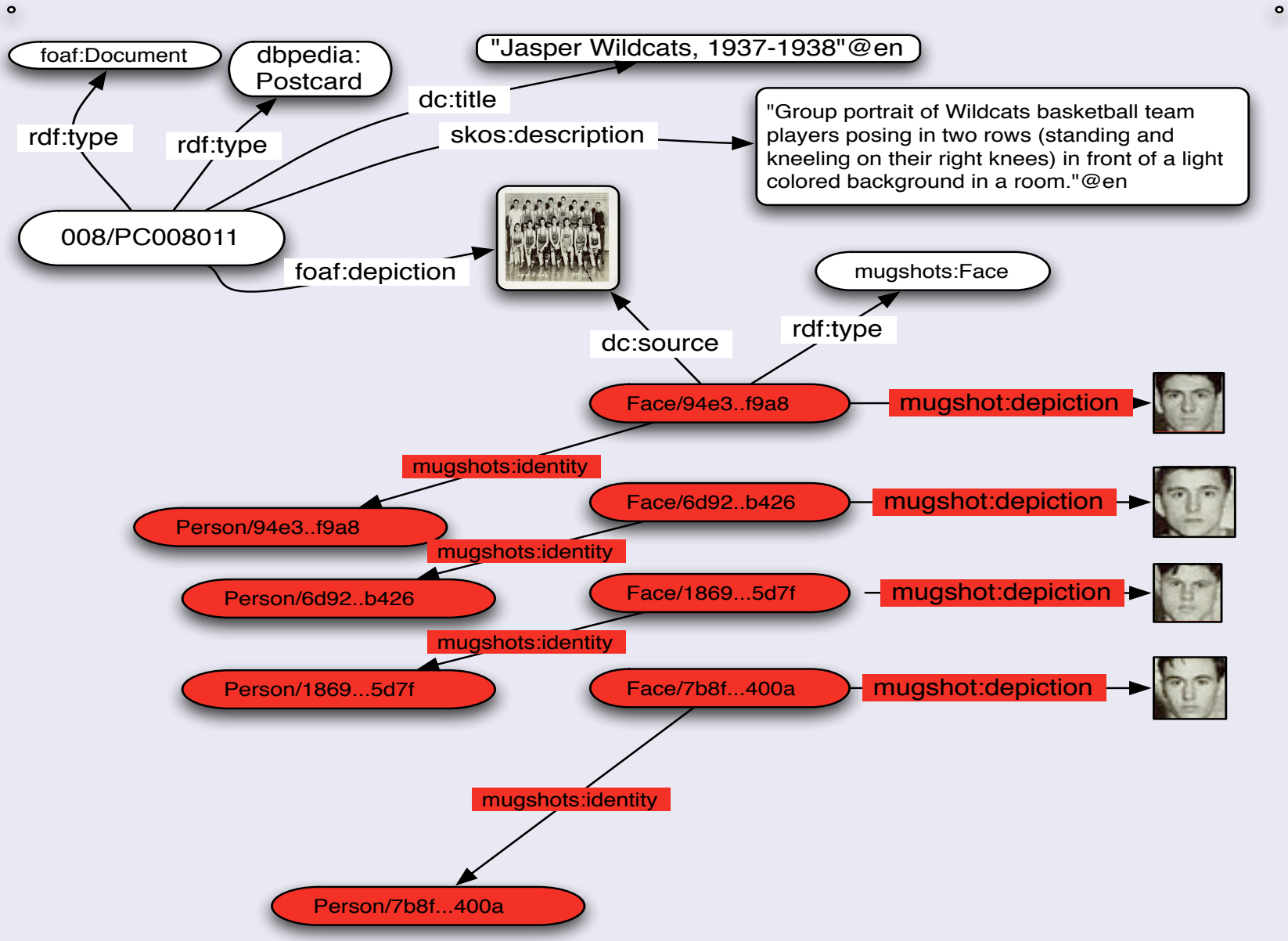


Image Analysis

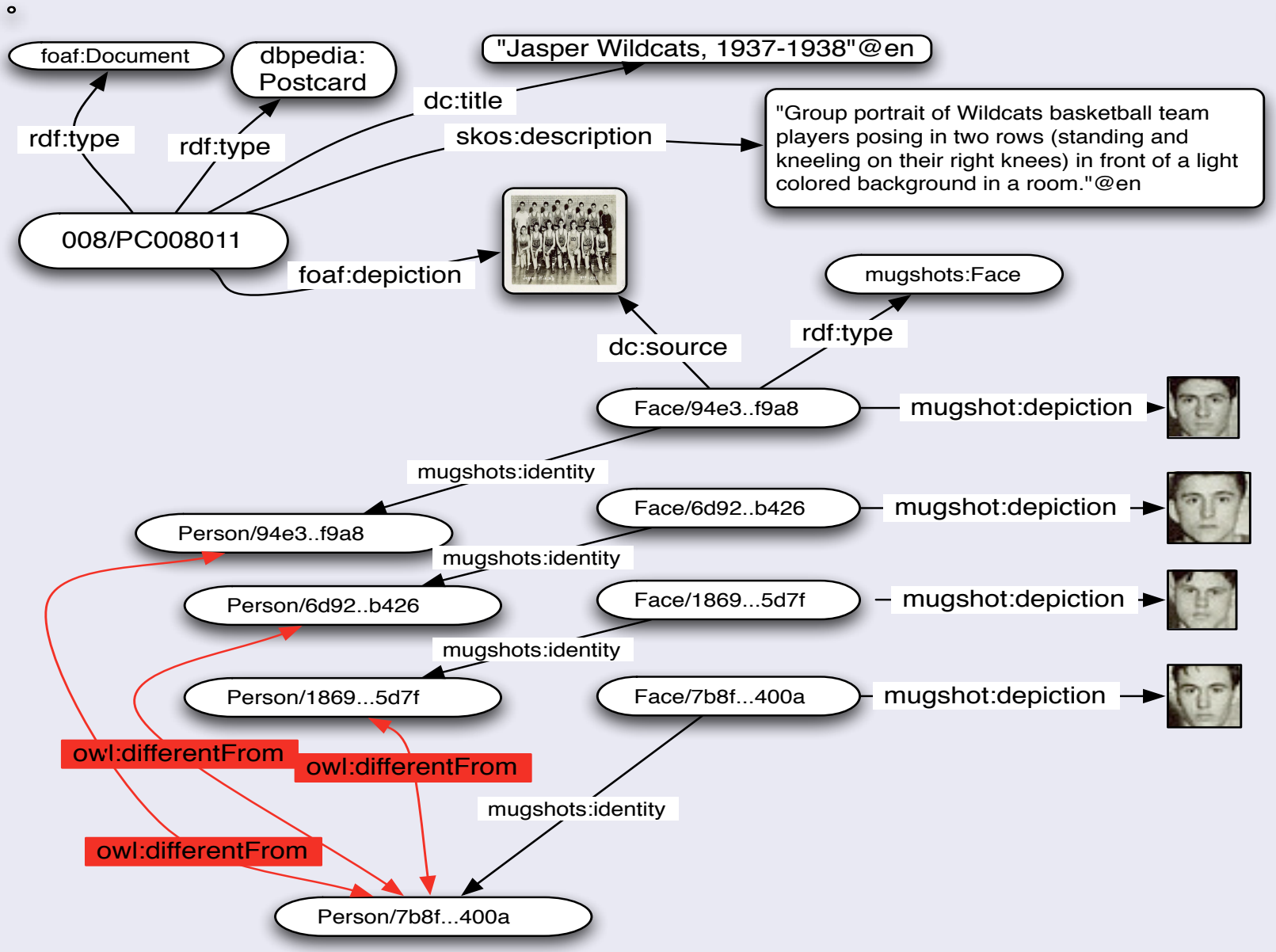


Image Analysis

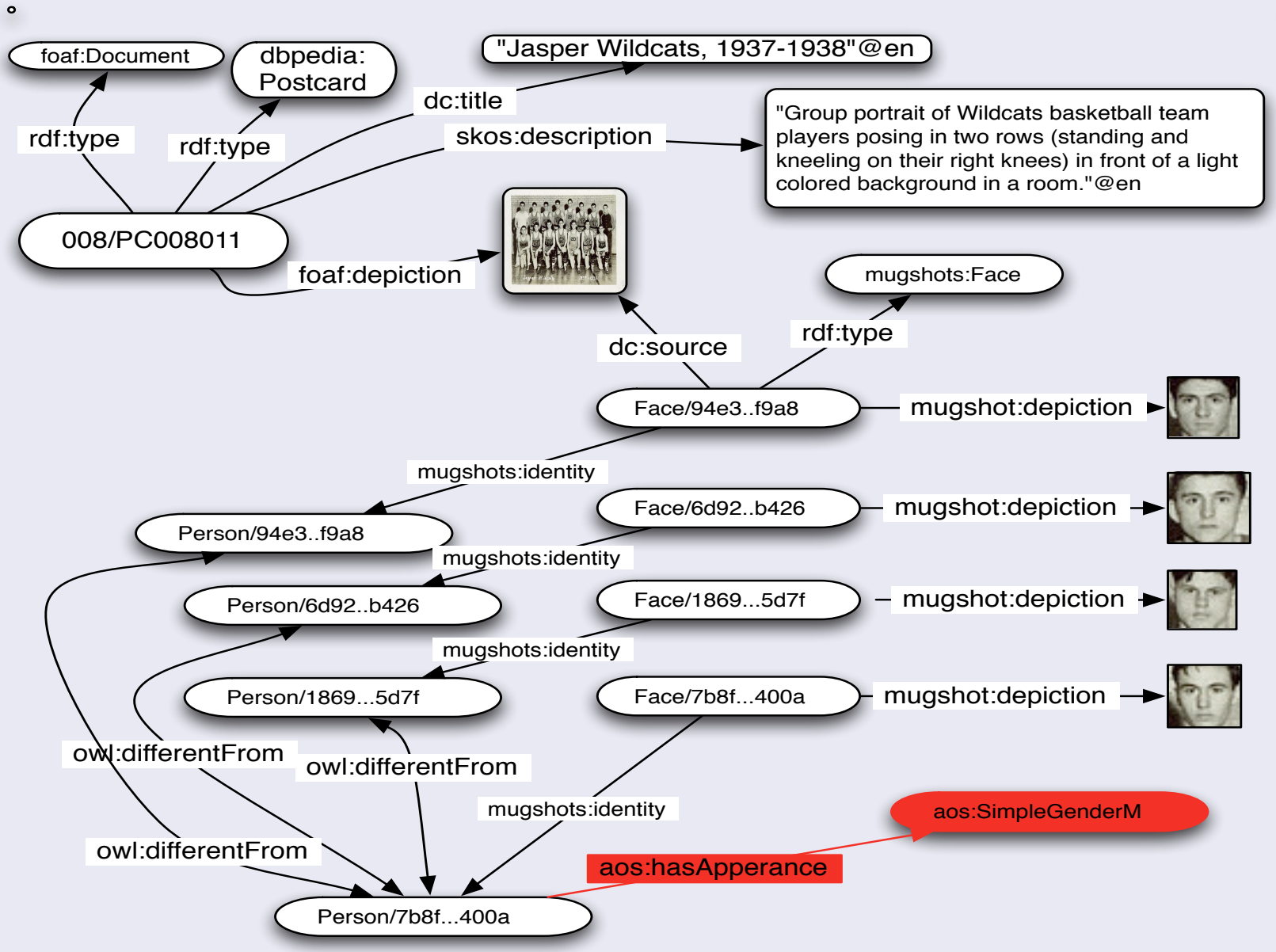


Image Analysis

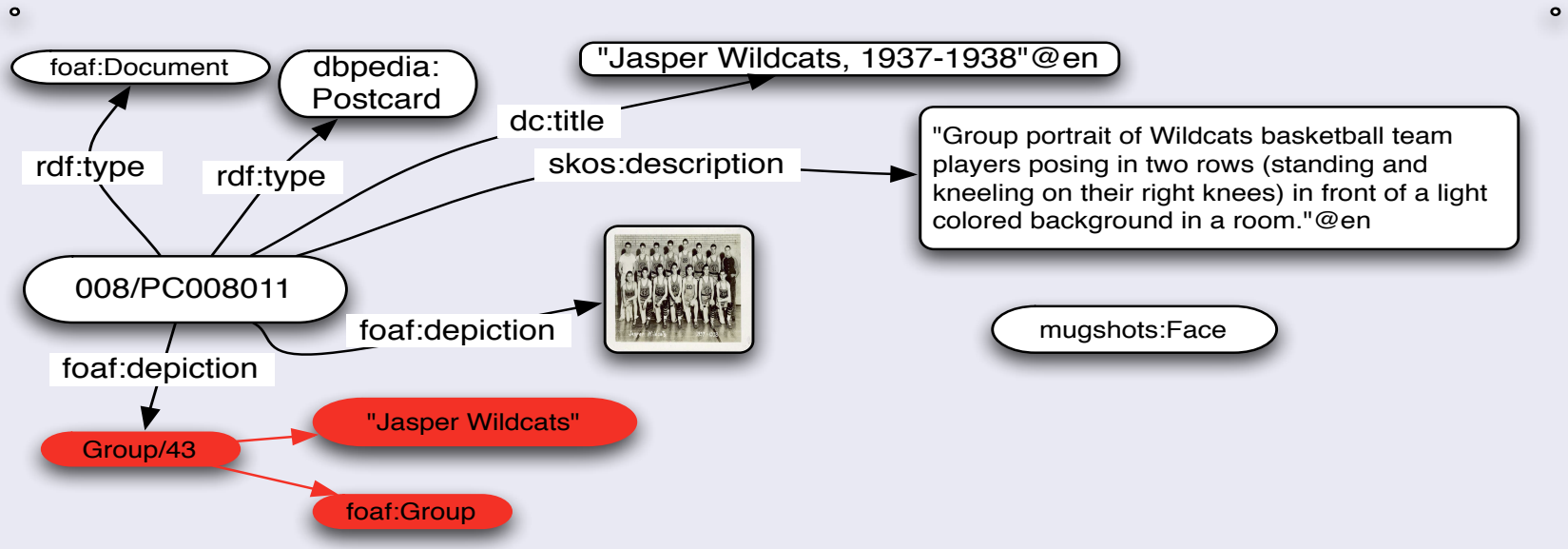


Image Analysis

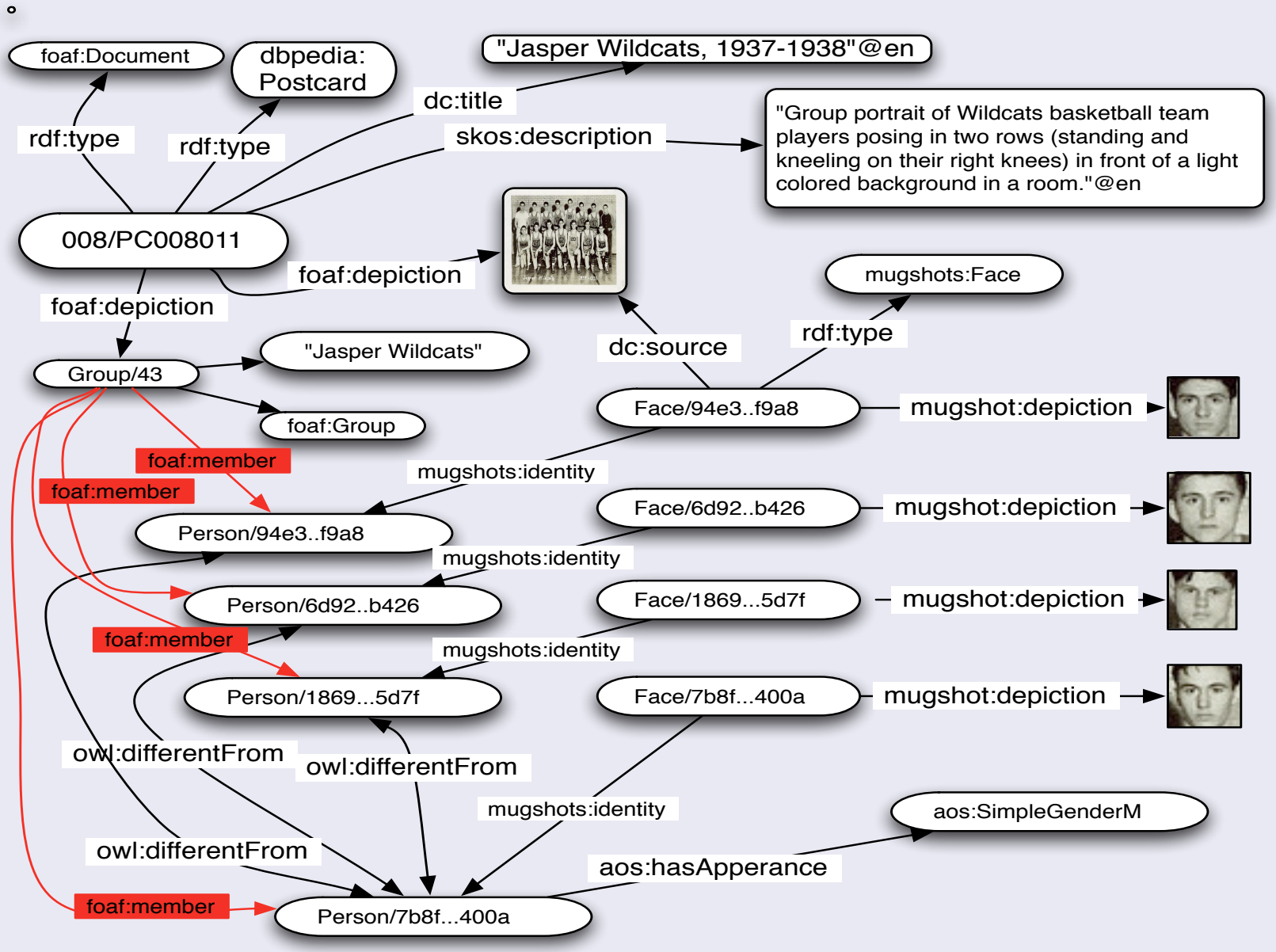
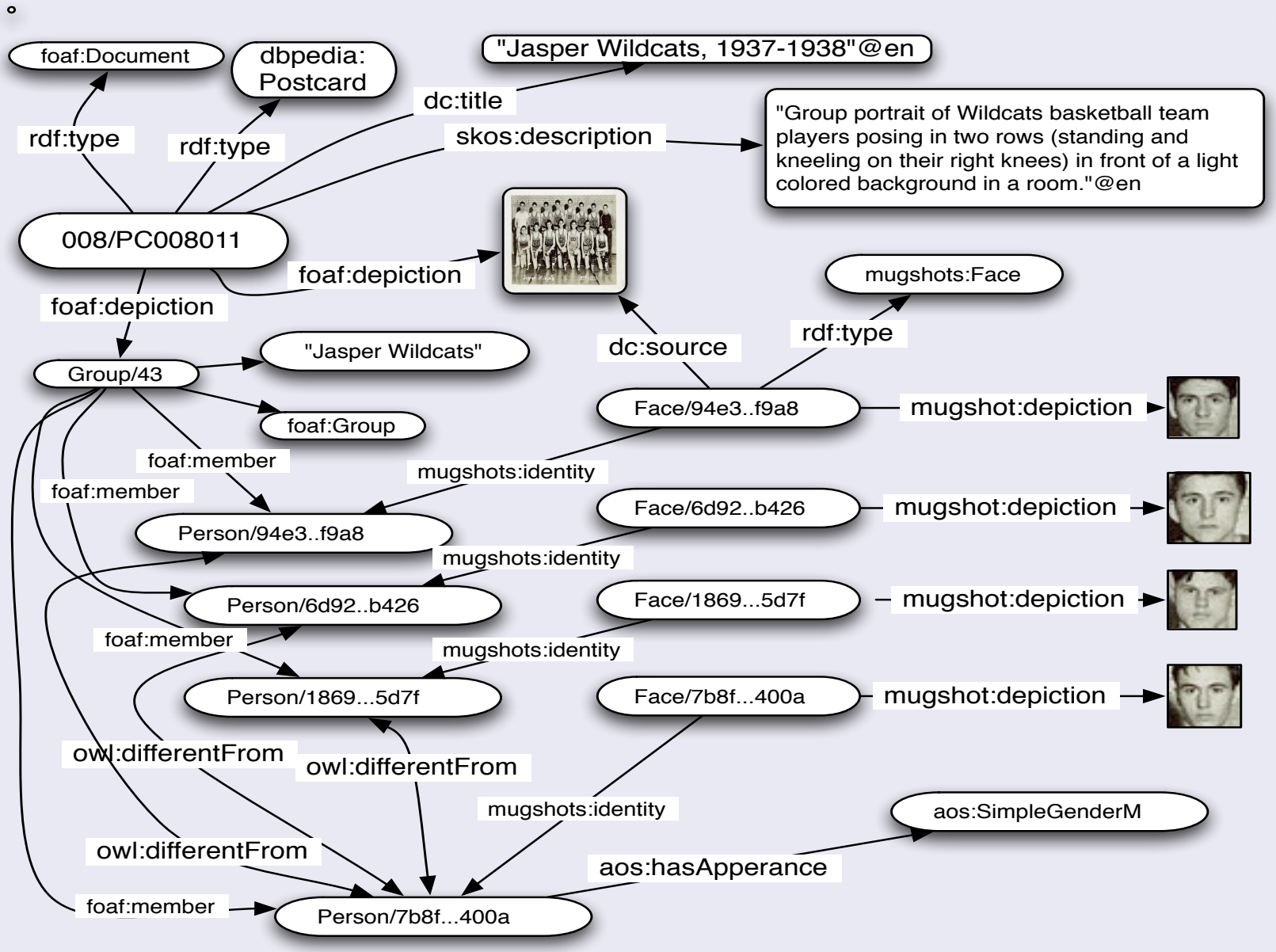
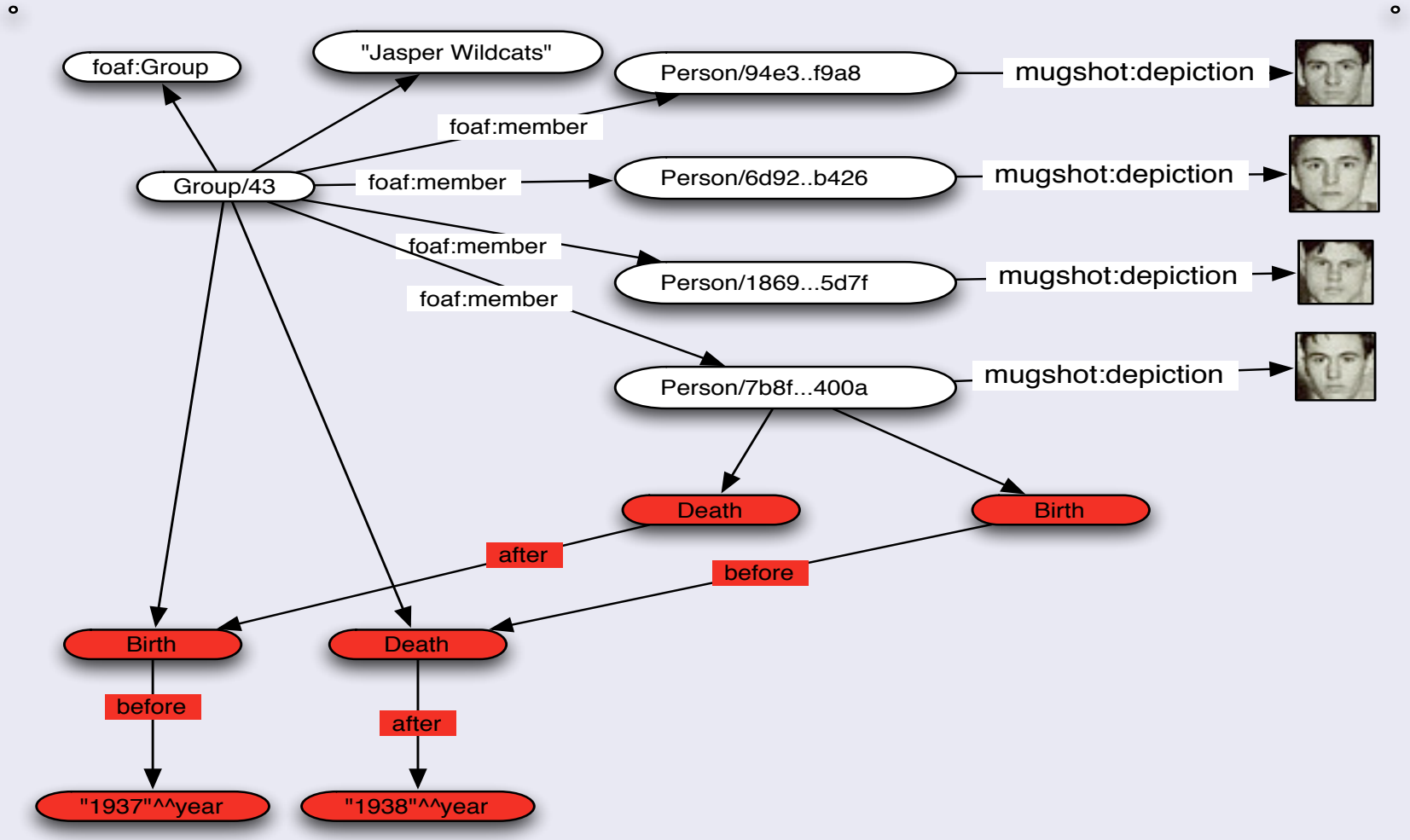


Image Analysis

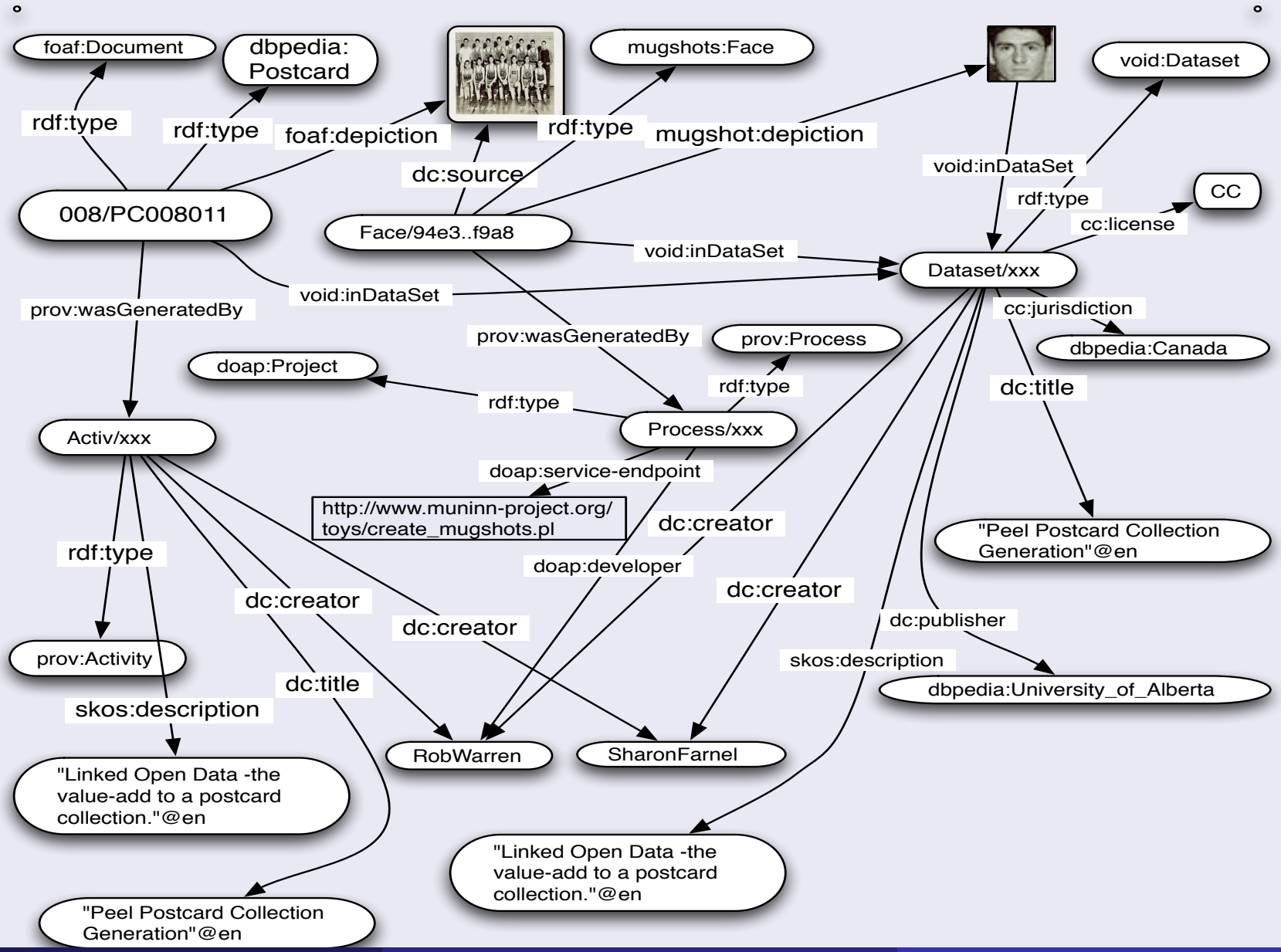


Support for transcription?

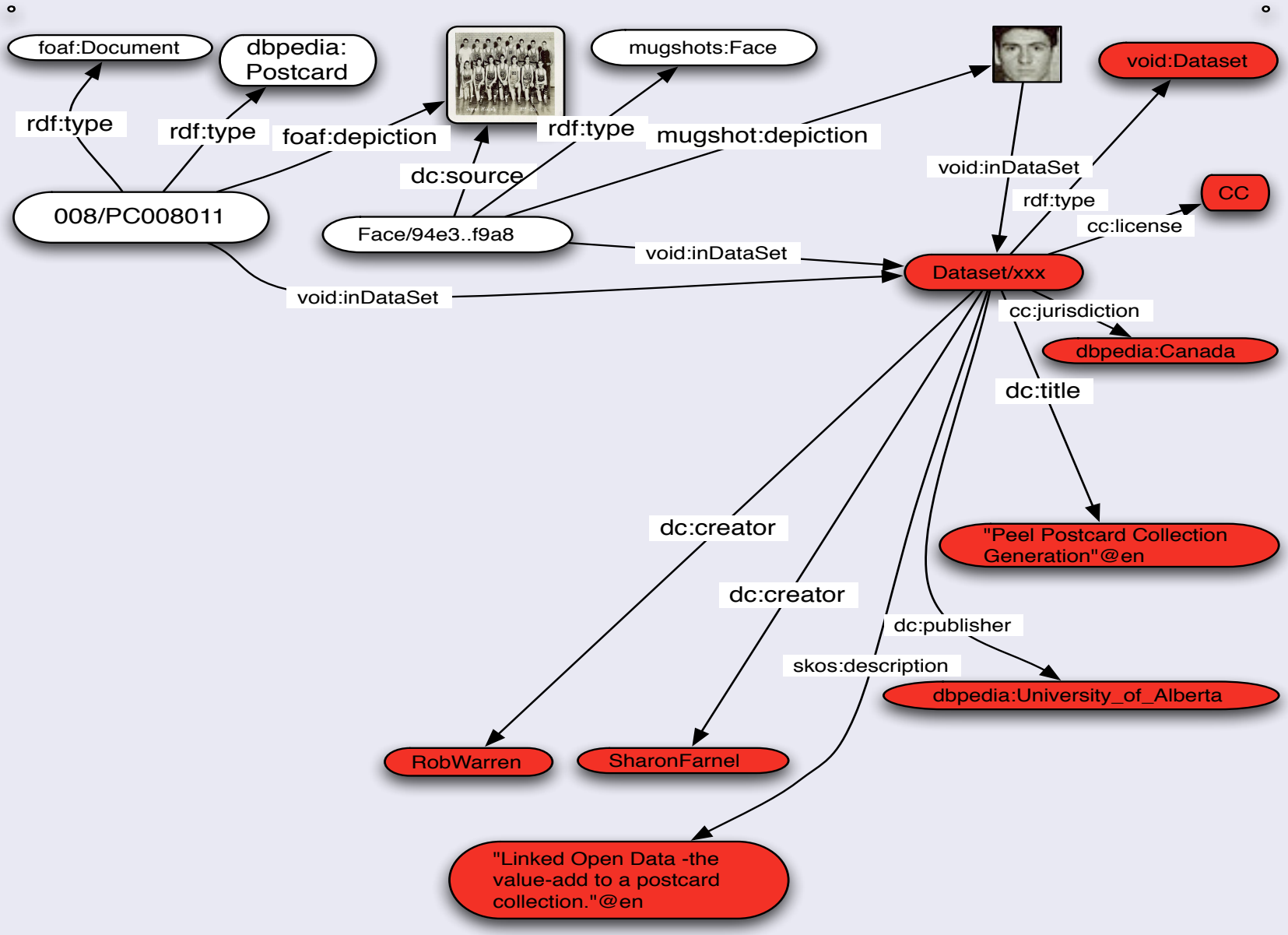
Image Analysis



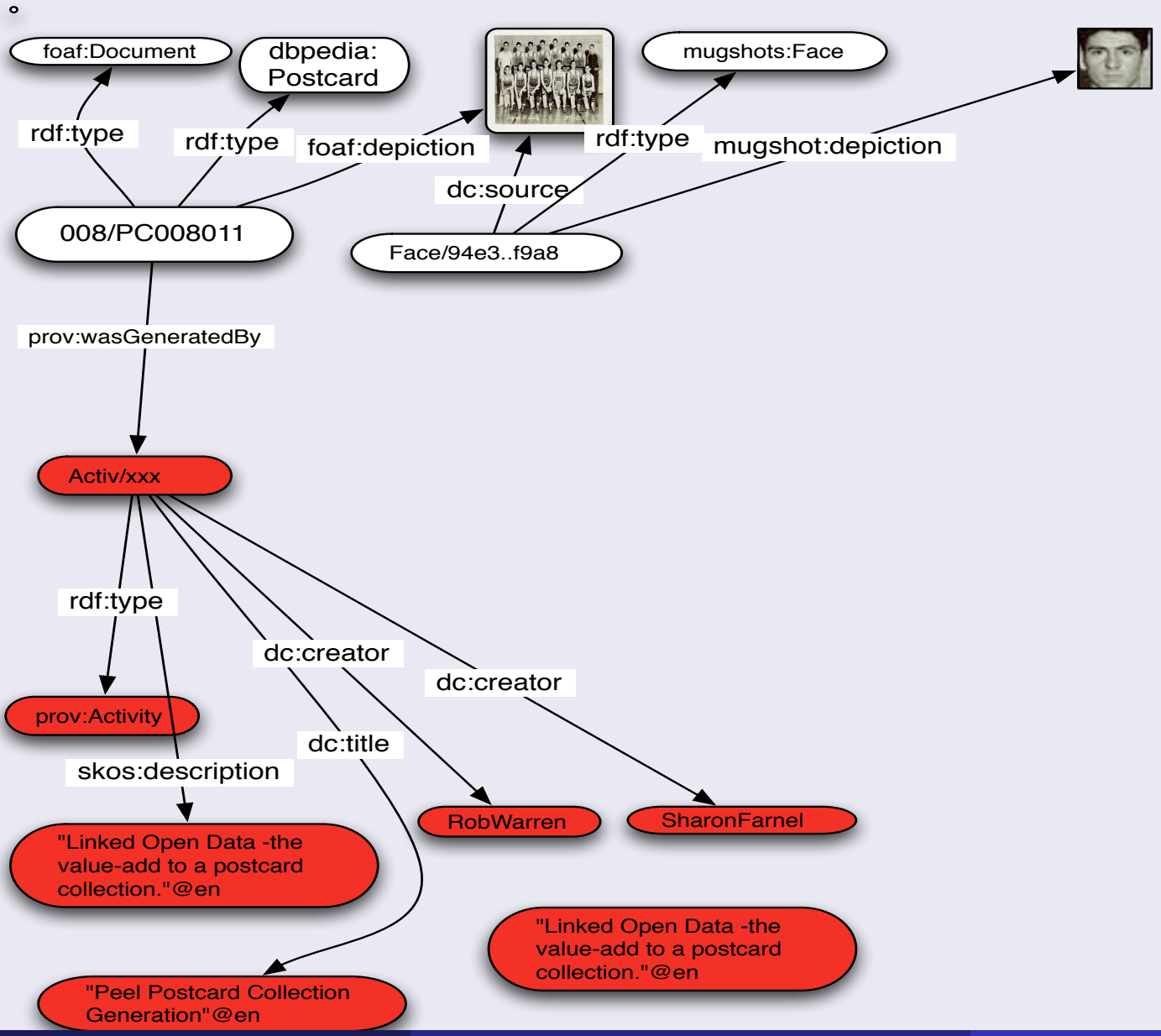
Datasets / Process documents



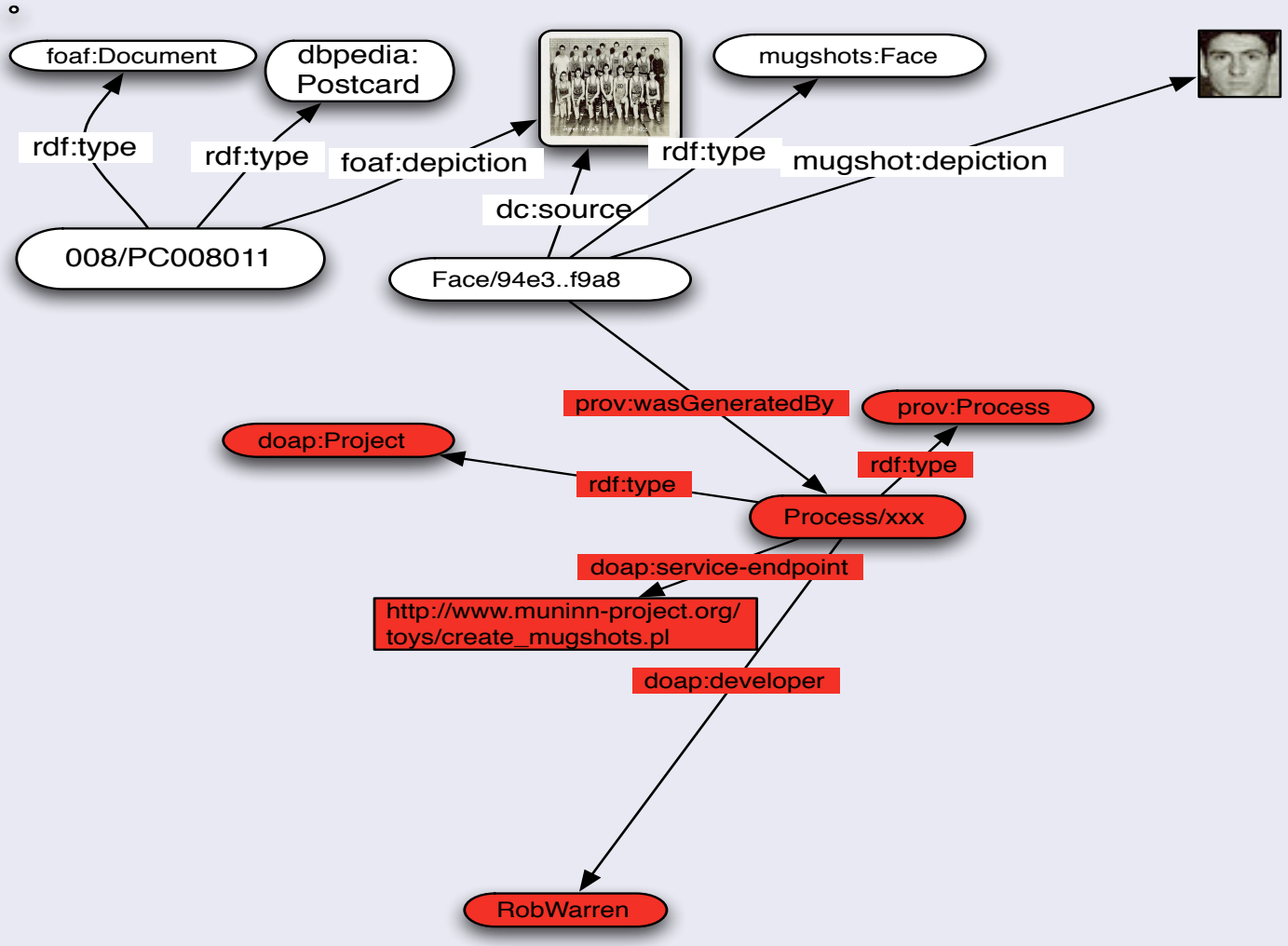
Datasets documentation



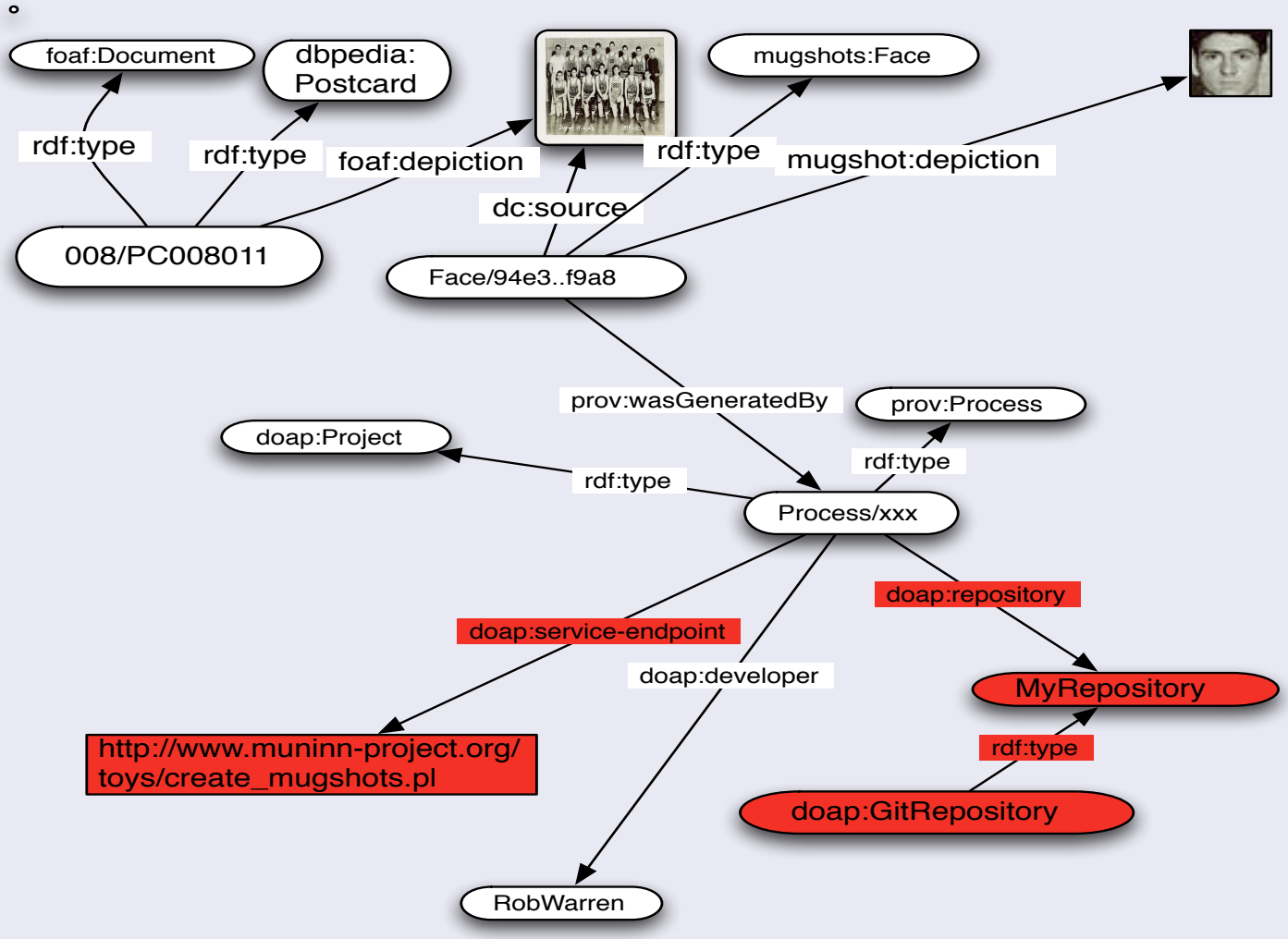
Datasets Creation



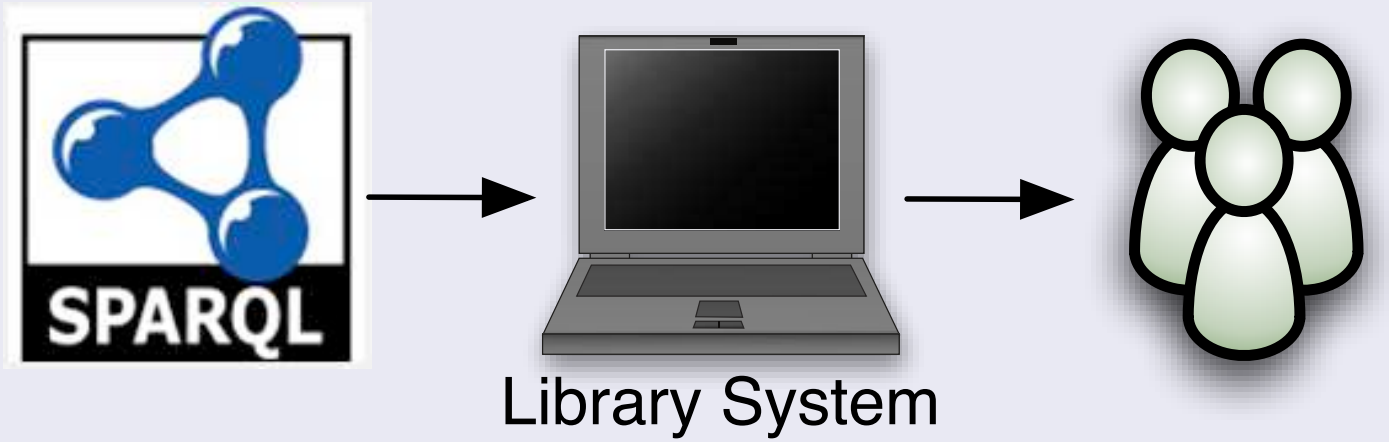
Datasets Process



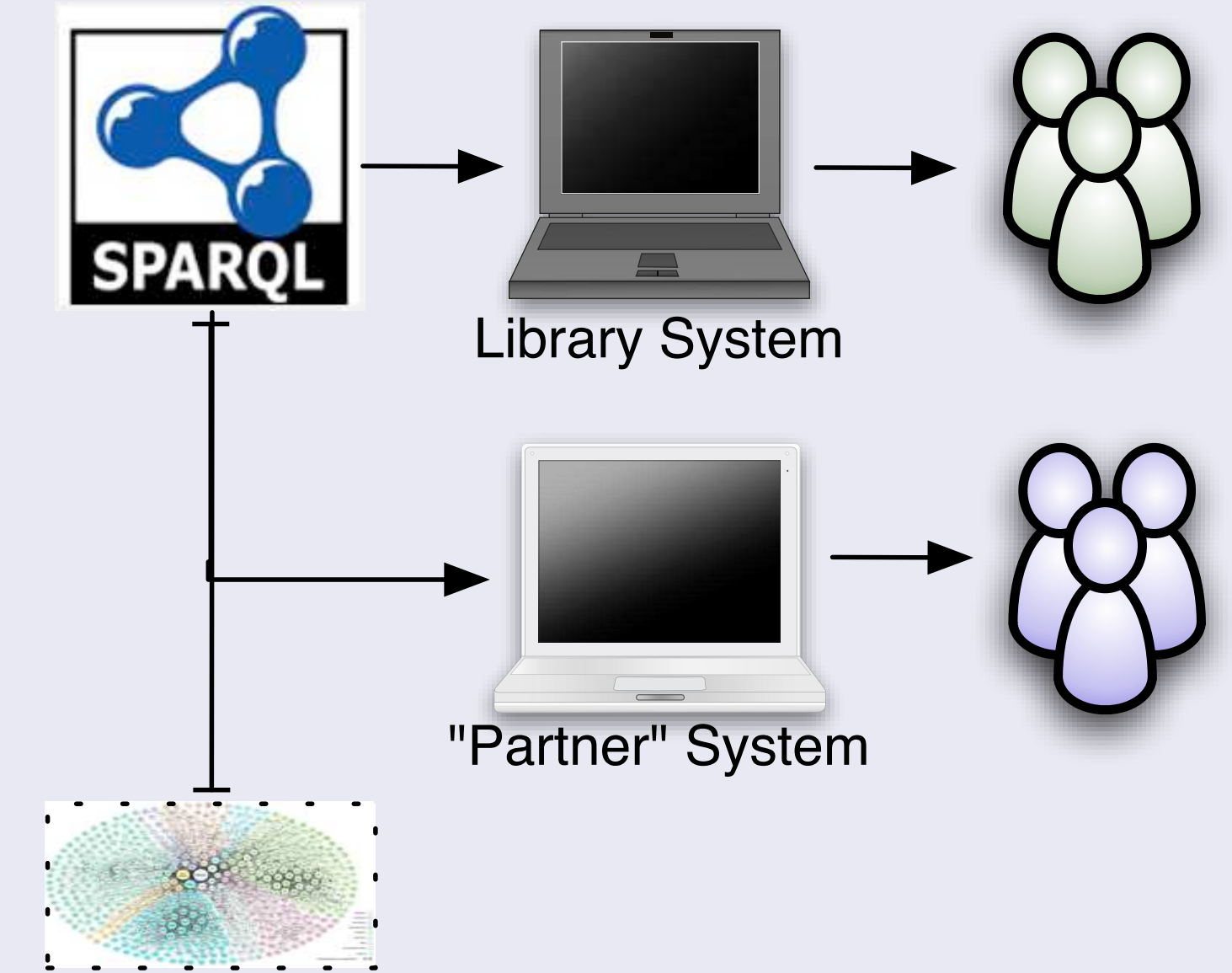
Datasets Process



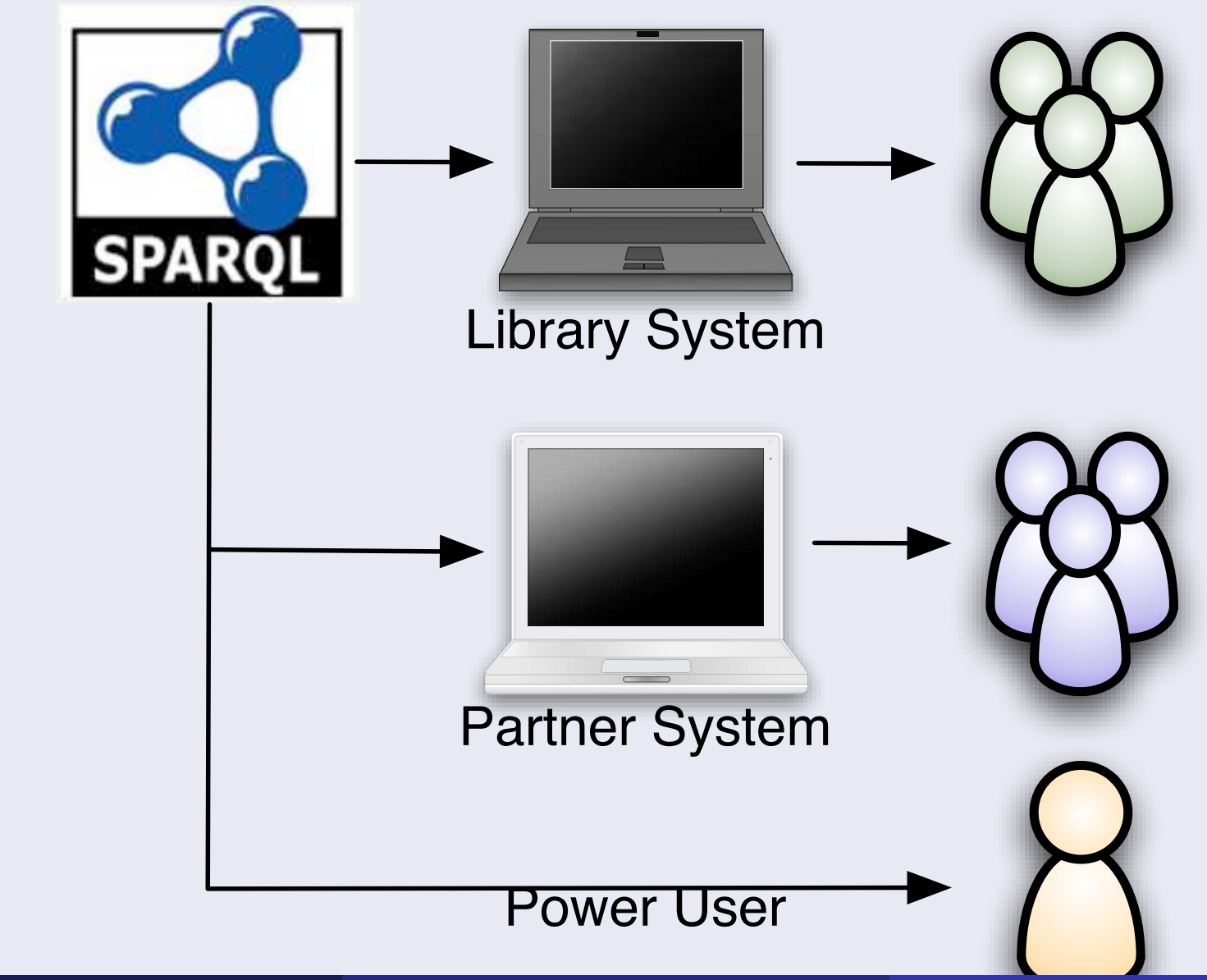
A possible future



A possible future

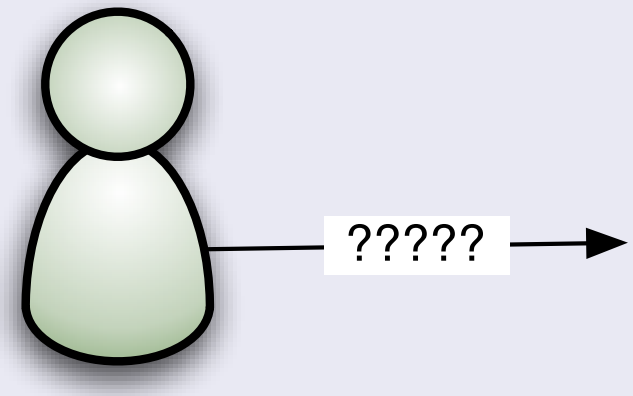


A possible future



Why is machine readable data important?

...but not for long!



slandora

slandora

slandora

Hydra

Hydra

Hydra

W3C SPARQL

W3C SPARQL

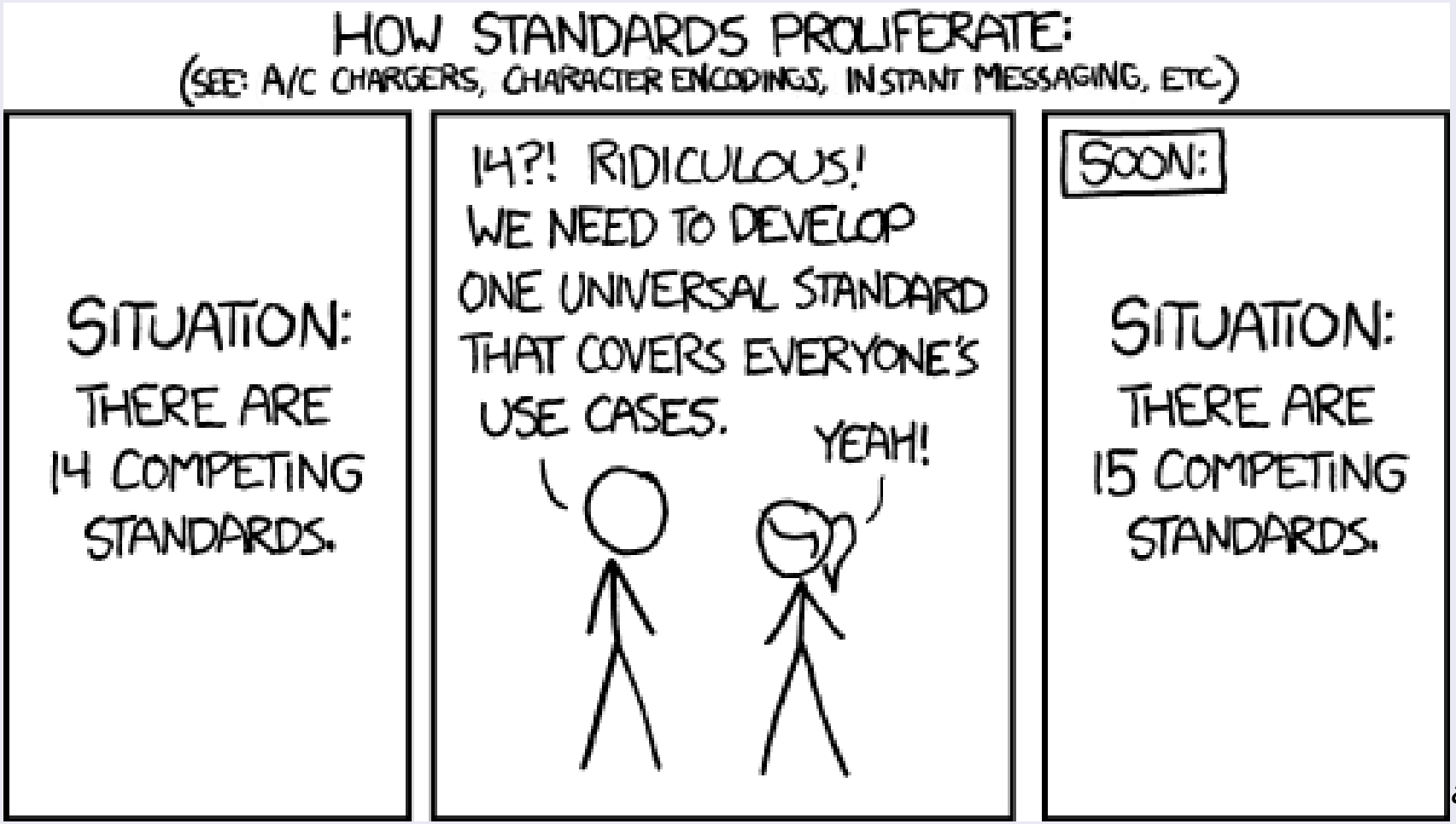
W3C SPARQL

The golden rule of Linked Open Data modelling:

The *thing* and the *name of the thing* aren't the same *thing*. [a]

[a] Slide shamelessly stolen from the Canadian Writing Research Collaboratory slide deck.

Linked Open Data and OWL let you pick and choose standards



[a] [b]

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

[b] Slide shamelessly stolen from the Canadian Writing Research Collaboratory slide deck.

Lessons Learned, Next Steps

- Play with conversion to RDFa or Turtle and endpoints.
- Enhance other image collections in a similar way.
- Text-based collections - named entity recognition?
- Excel is the enemy of data quality.
- The artifact might be less important than the (meta?)data.
- RDF / LOD allows you to publish data, document it and (sometimes) fix mistakes later on.
- Separate the data from the application.

Code and scripts at:
`https://github.com/muninn/PC-Access2015`