Understanding Data and Their Environment

Data Provenance

Professor Mark Elliot

The provenance of this talk

- Thanks to:
  - my colleagues Nuno and Stian
  - Prov primer
    - https://www.w3.org/TR/2013/NOTE-prov-primer-20130430/
  - Moreau and Groth (2014)
  - Zachary Ives
    - https://www.youtube.com/watch?v=wyt0Zhbd1T0

Outline

- What is provenance?
- Data provenance
  - What?
  - Why?
- Intro to modelling provenance
  - Using Prov

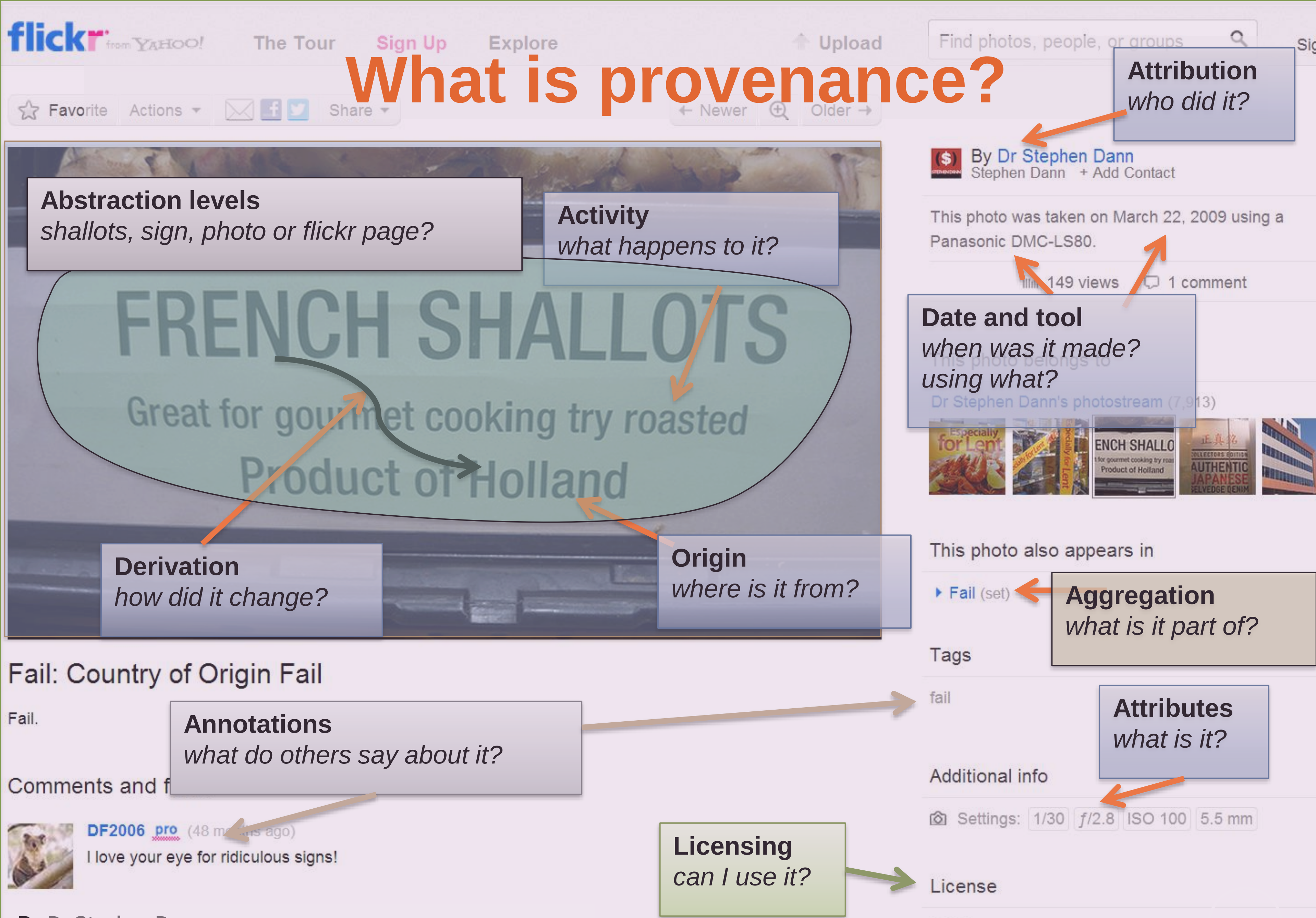
ILOs

- By the end of this session you should:
  - Have a top level understanding of what provenance (and particularly data provenance) is
  - Understand why it is important
  - Be able to sketch a basic provenance graph

Provenance

- French root
  - *Provenir – to come forth*
- Where a thing has come from
  - Initially applied to Art
  - Now applied to food, wine, architecture, historical documents and artefacts
    - All with slightly different meanings
  - But all imply some sort of record

What is provenance?



Attribution
who did it?

Abstraction levels
shallots, sign, photo or flickr page?

Activity
what happens to it?

Date and tool
when was it made?
using what?

Derivation
how did it change?

Origin
where is it from?

Aggregation
what is it part of?

Annotations
what do others say about it?

Attributes
what is it?

Licensing
can I use it?

By Dr Stephen Dann
Stephen Dann + Add Contact

This photo was taken on March 22, 2009 using a Panasonic DMC-LS80.

149 views 1 comment

This photo belongs to
Dr Stephen Dann's photostream (7,913)

This photo also appears in
Fail (set)

Tags
fail

Additional info
Settings: 1/30 ƒ/2.8 ISO 100 5.5 mm

License
Some rights reserved
Request to license Dr Stephen Dann's photos via Getty Images

Fail: Country of Origin Fail

Fail.

Comments and f

DF2006 pro (48 months ago)
I love your eye for ridiculous signs!

By Dr Stephen Dann
licensed under Creative Commons Attribution-ShareAlike 2.0 Generic

What is Data Provenance?

- Metadata of process
- A record
  - Who created the (data) object?
  - How the (data) object was created?
    - **Original** Data acquisition/capture

What is Data Provenance?

- **Original** Data acquisition/capture
  - Intentional Data
    - e.g. Surveys
  - Consequential Data
    - e.g. Administrative Data
  - Interactional data
    - e.g. Social media
  - Automatically generated data
    - e.g. Sensors

What is Data Provenance?

- Metadata of process
- A record
    - Who created the (data) object?
    - How the (data) object was created?
        - Original Data acquisition/capture
        - Processing/Analysis
        - Outputs (visualisation/reports/models)

Why Data Provenance?

- Trust
  - In data
  - In products
- Reproducibility
- Reusability
- Process management

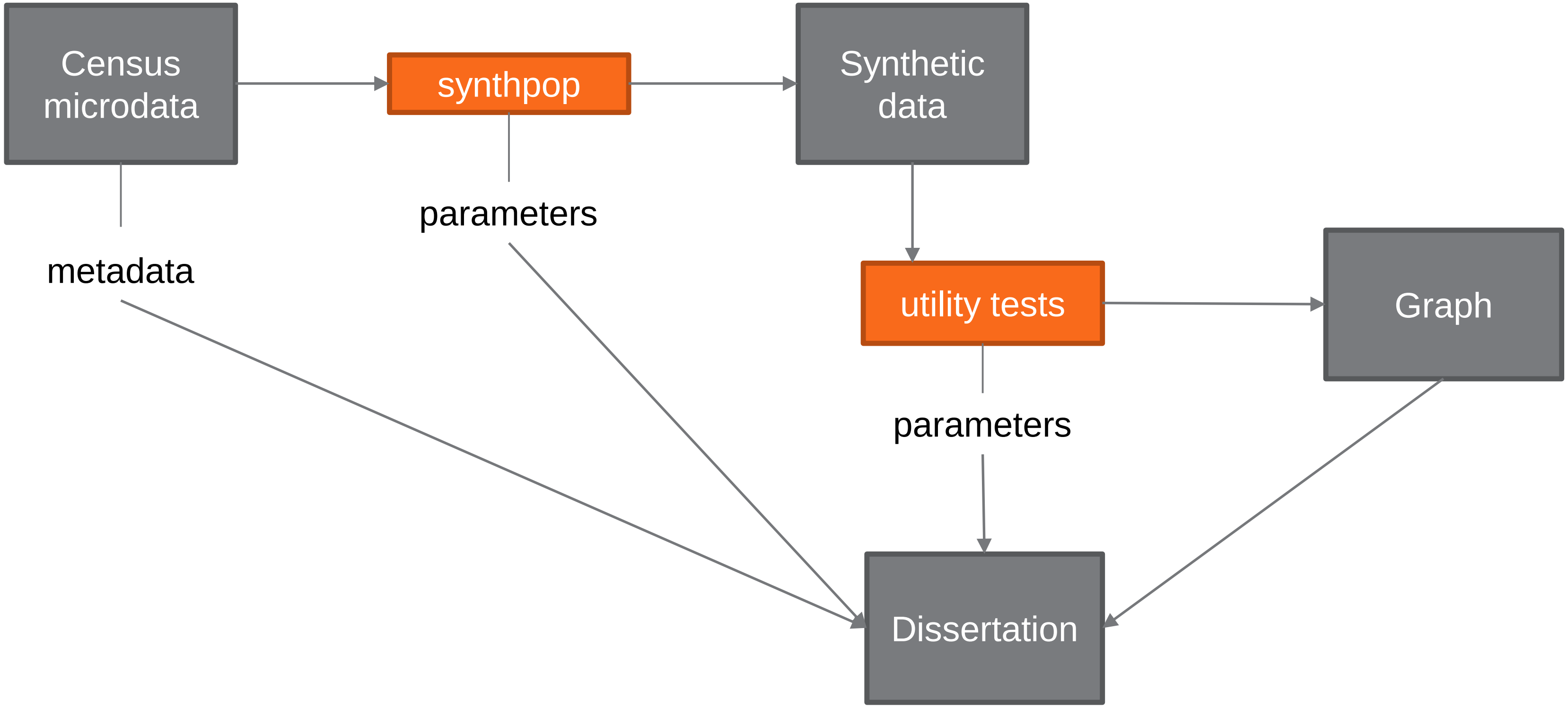
Raw data

- Provenance (aka paradata)
  - Tends to be domain specific
  - Describes the context of collection
    - Could Include
      - Agents
      - Study Design
      - Which instruments
      - Time frames/stamps
      - Environment
      - Parameters/settings
      - Reason for Collection

Derived data (products)

- Processes involved in creation
  - Parameters for those processes
- Inputs

An example of a process flow


Census microdata

synthpop

Synthetic data

metadata

parameters

utility tests

Graph

parameters

Dissertation

What are we trying to achieve 1

- Operationally
  - For each derived piece of data, product or output:
    - **How** was it created?
    - What were its **inputs**?
    - What were the **parameter settings** (if any)?

What are we trying to achieve 2

- Knowledge and Understanding
  - To be able to reason about
    - that information
    - and the data itself
  - To do that we need to have everything:
    - recorded in one place
    - captured according to some standard
    - appropriately connected to data(base) itself

Tracing and Logging

- Standard CS technique
  - See e.g. https://syrah.eecs.harvard.edu/pass

Tracing and Logging

- Useful but not as an end point
  - Logging systems are
    - not standardised and are subject to evolution
    - full of noise (irrelevant stuff)
  - We need causal descriptions not temporal ones.
    - Cause and time are related but not the same thing!

Key points so far

- Data Provenance is important for
    - User Trust
        - Reproducibility
        - Evaluation (of quality)
        - Harmonisation (for linkage etc.)

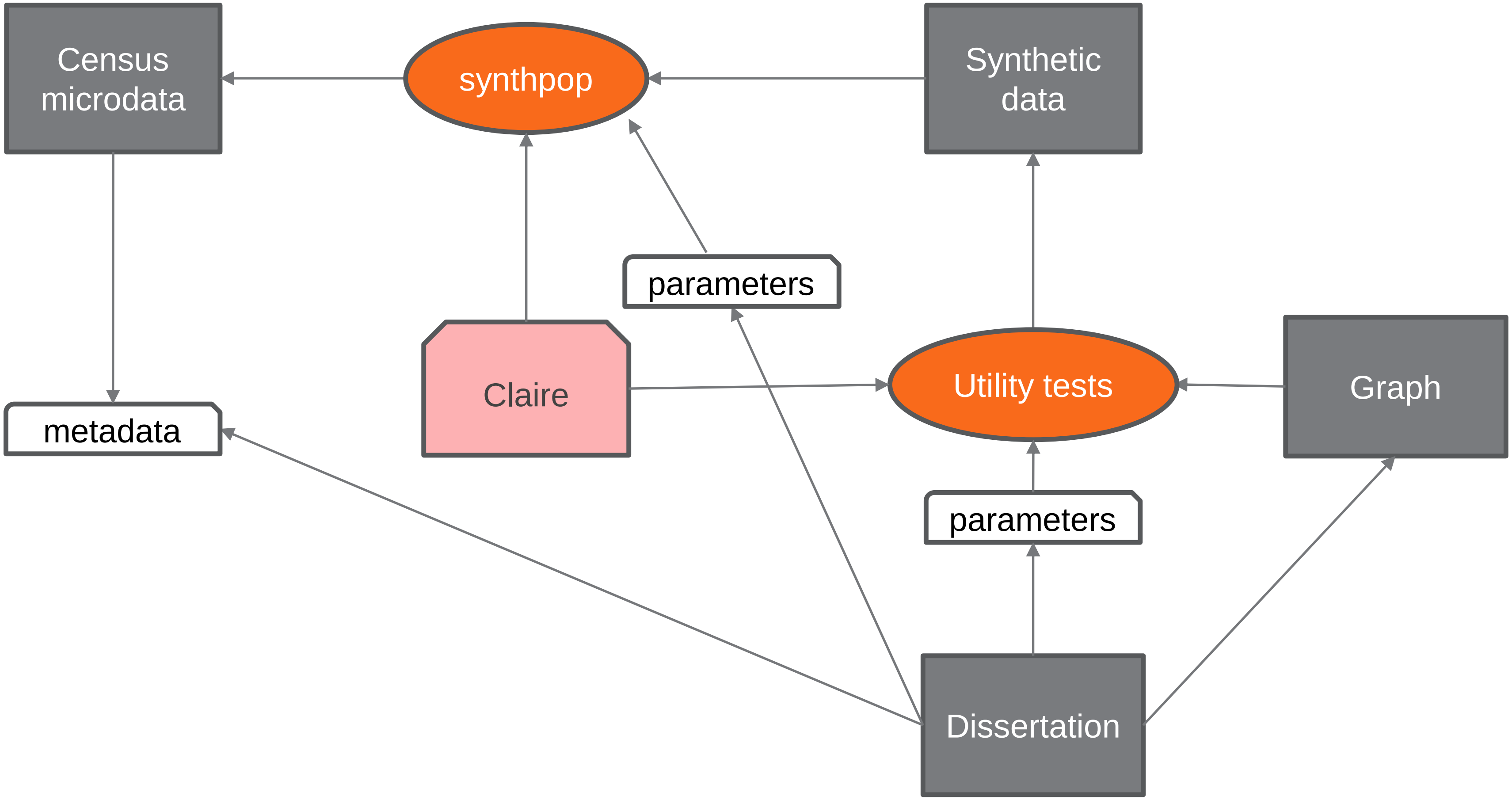
Key points so far

- Provenance of raw data captures:
  - Who
  - What
  - When
  - Why/What for
  - How

Summary 2

- On derived data (products) it captures:
  - Prior Processing steps
  - Inputs
  - Agents (users)

Modelling provenance (with PROV)


Census microdata
synthpop
Synthetic data
parameters
Claire
Utility tests
Graph
metadata
parameters
Dissertation

Exercise

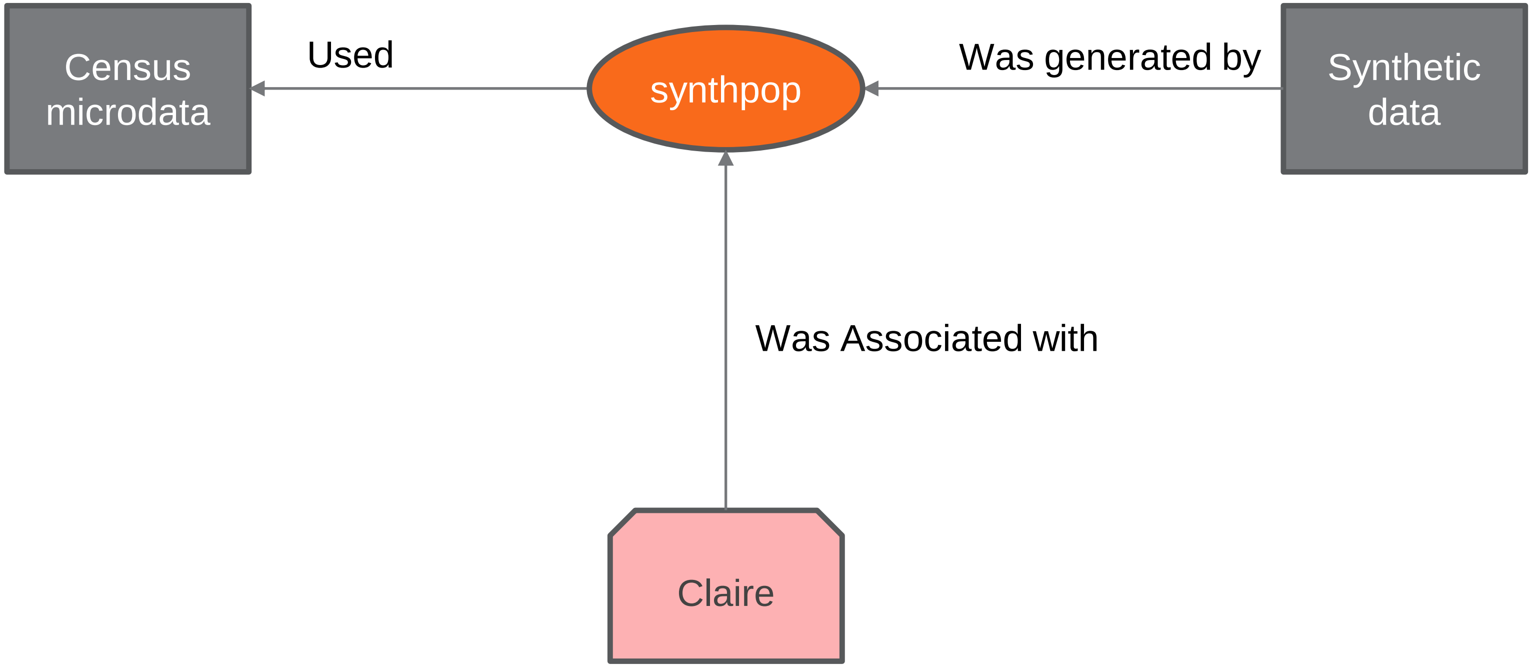
- In your groups
  - Have a go at producing a basis provenance graph for the dataset that you created

Prov DM

- W3C standard
- Core properties and relations defined
- Extensible

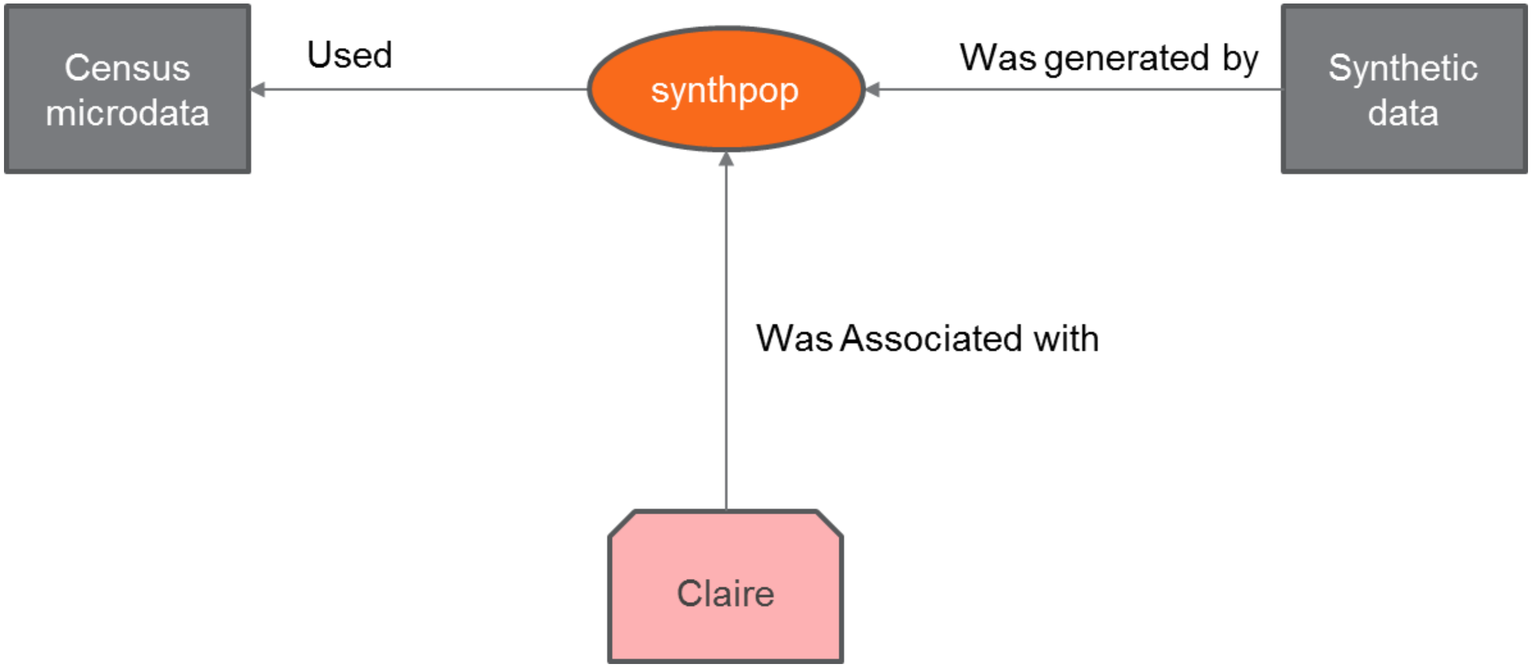
Relations

Concept	PROV-DM Label
Generation	WasGeneratedBy
Derivation	WasDerivedFrom
Association	WasAssociatedWith
Revision	WasRevisionOf
Usage	Used


Census microdata
Used
synthpop
Was generated by
Synthetic data
Was Associated with
Claire

Prov-DM Formats

- Prov-N
- XML
- Turtle



Census microdata

Used

synthpop

Was generated by

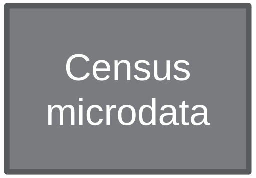
Synthetic data

Was Associated with

Claire

Prov-N

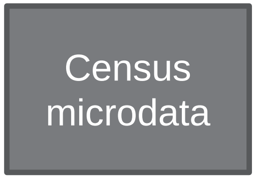
```
entity(exg:dataset1, [dcterms:title “UK Census
microdata 2011 teaching dataset”])
```



Census microdata

XML

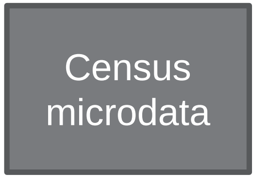
```
<prov:entity prov:id=“exg:dataset1”>
        <dct:title>UK Census microdata 2011 teaching dataset<\dct:title>
</prov:entity>
```



Census microdata

Turtle

```
exg:dataset1     a prov:Entity ; dcterms:title "UK
                 Census microdata 2011 teaching dataset" .
```



Census microdata

Summary

- Data Provenance allows users to understand and trust data
- Involves tracking what has happened to arrive at a particular data object
- There are standards for representing provenance
    - PROV is one Such Standard