UNIVERSITY OF Southampton

Pragmatic Linked Open Data and Preservation

Christopher Gutteridge
Linked Open Data Architect

UNIVERSITY OF Southampton

(or “Identifiers: the good, the bad and the ugly”)

Summary

UNIVERSITY OF Southampton

- With Linked Data you'll be dealing with other people's Identifiers
- You won't like them
- They will be too vague (or too specific)
- They will change

UNIVERSITY OF Southampton

Linked Open Data & RDF

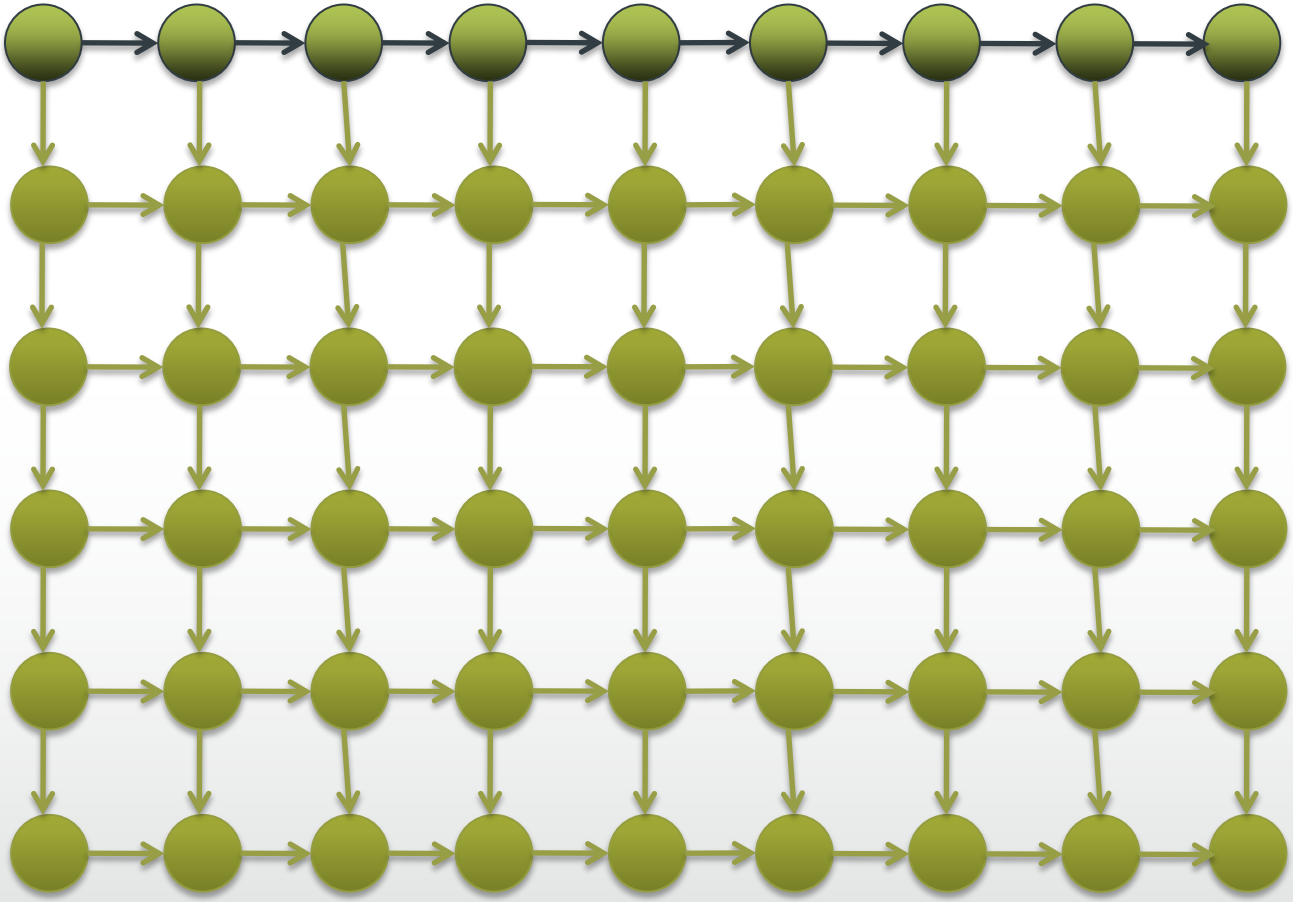
What is Linked Data?

UNIVERSITY OF Southampton

- **Linked data** – a technique for creating data which joins up with data from other data provider's unique identifiers.
- **Open data** - publishing data with a license which permits re-use, akin to creative commons for creative arts and to 'open source' software.
- **RDF** - a way of structuring data as a list of facts and globally unique identities for things. Easy to merge lists of facts from multiple sources.
  - *Person23 Lives-In City17*
  - *Person23 is-Named "Marvin Fenderson"*
  - *City17 is-Named "Southampton"*

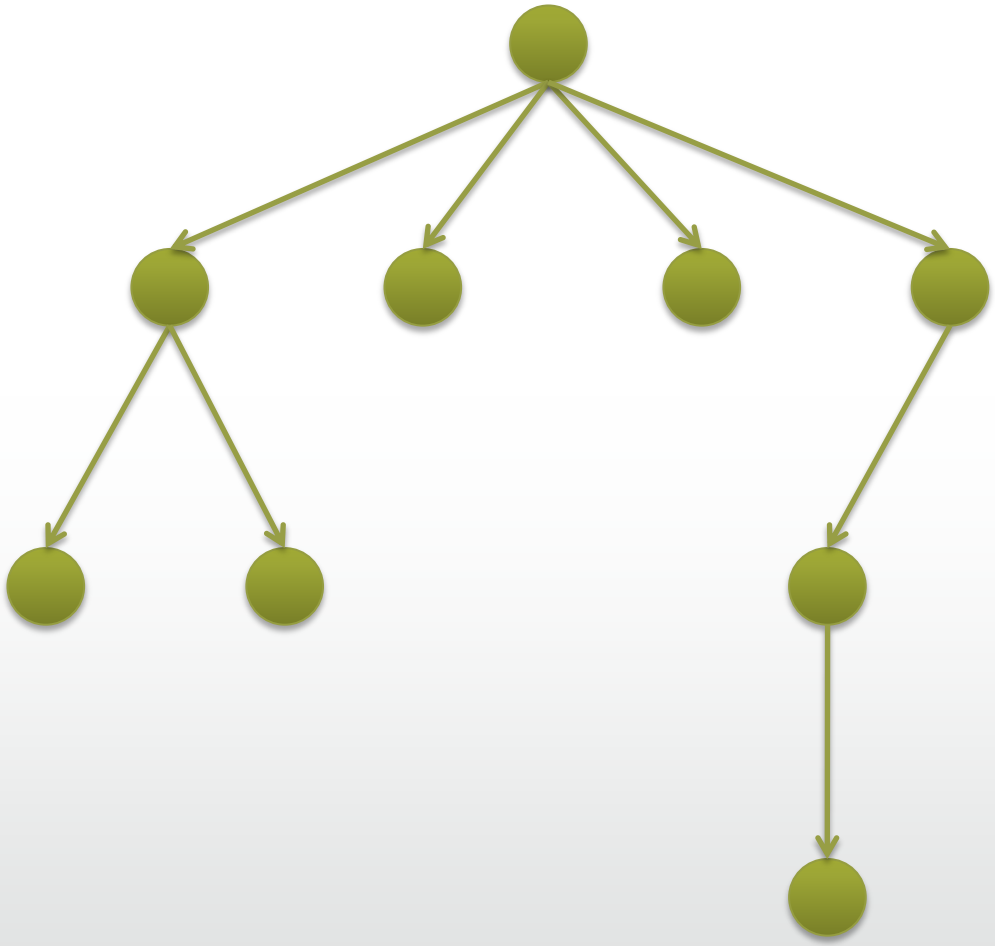
Tabular Data (CSV, Excel, SQL)

UNIVERSITY OF Southampton



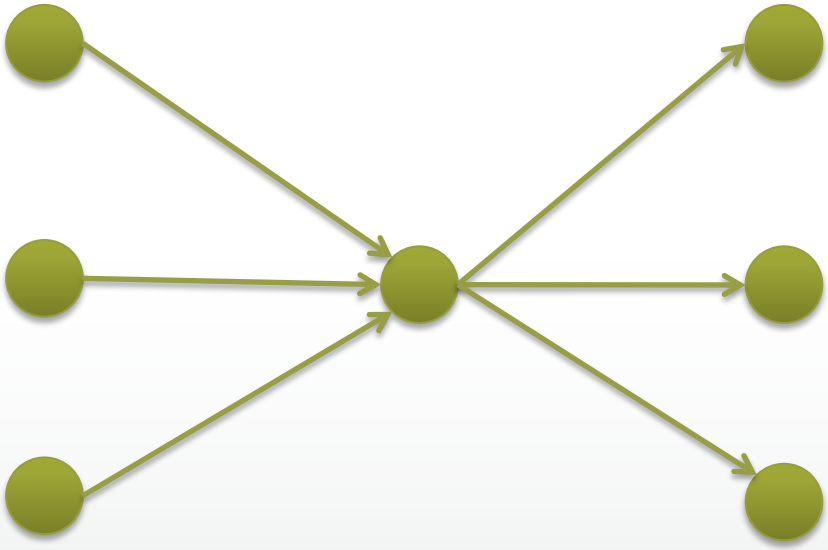
Tree Data (XML, JSON)

UNIVERSITY OF Southampton



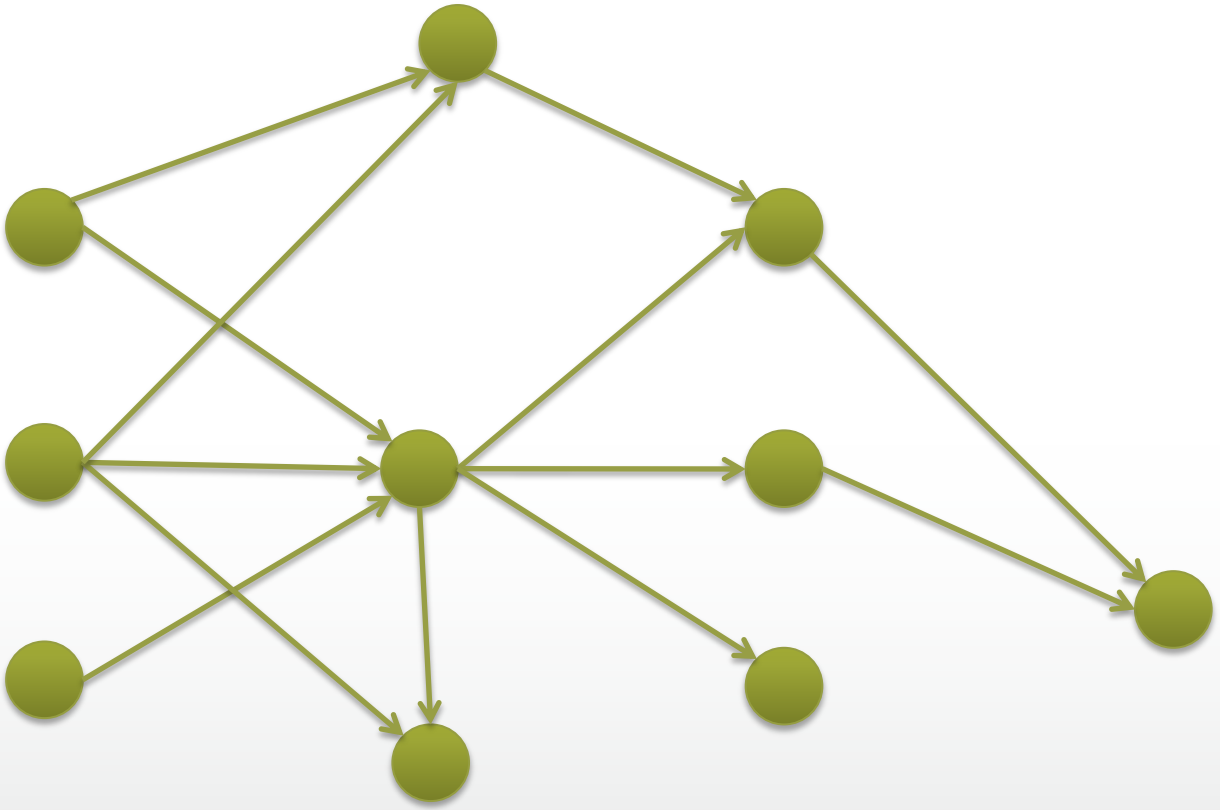
Graph Data (RDF)

UNIVERSITY OF Southampton



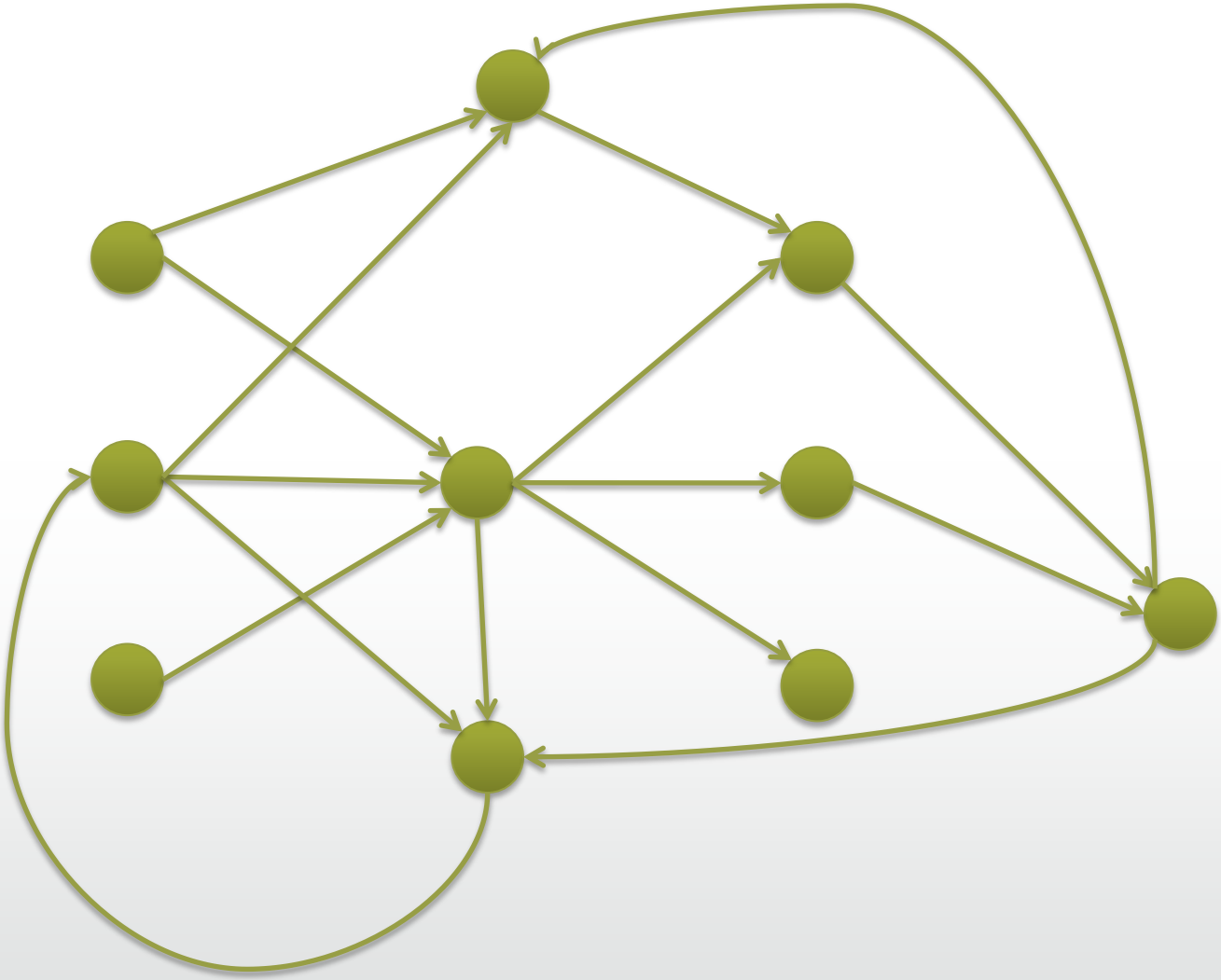
Graph Data (RDF)

UNIVERSITY OF Southampton



Graph Data (RDF)

UNIVERSITY OF Southampton



UNIVERSITY OF Southampton

data.southampton.ac.uk

UNIVERSITY OF Southampton

- Organisational Infrastructure data
- From all parts of the organisation
- Open Data (OGL)
- Linked Open Data (where we can)
- Emphasis on current information, not historical
- Contextual data for more valuable candidates for preservation (research output)

UNIVERSITY OF Southampton

The Good

UNIVERSITY OF Southampton

GETTING EVERYONE USING THE SAME IDENTIFIERS

Good; Surfacing Organisation IDs

UNIVERSITY OF Southampton

- Rooms
- Buildings
- Organisation Parts
- Research Facilities
- Equipment
- Research Output

URIs to identify everything

UNIVERSITY OF Southampton

URIs look like a web address, but (sometimes) identify a real world thing such as

Equipment
- *http://id.southampton.ac.uk/equipment/E0672*

Facilities
- *http://id.southampton.ac.uk/facility/F0041*

Campuses
- *http://id.southampton.ac.uk/site/1*

Organisations (and parts thereof)
- *http://id.southampton.ac.uk/*
- *http://id.southampton.ac.uk/org/FP*

Bad; Missing Identifiers

UNIVERSITY OF Southampton

- IDs for People are tricky

Bad; Missing Identifiers

UNIVERSITY OF Southampton

- No national URIs for things like
  - *JACS Codes*
  - *Academic sessions (years)*

Bad; Missing Identifiers

UNIVERSITY OF Southampton

- No national URIs for things like
  - *JACS Codes*
    - *HESA can help here*
  - *Academic sessions (years)*

Bad; Missing Identifiers

UNIVERSITY OF Southampton

- No national URIs for things like
  - *JACS Codes*
    - *HESA can help here*
  - *Academic sessions (years)*
    - *academic-session.data.ac.uk?*

University of Southampton

(the EPrints bit)

AN IDENTIFIER FOR EVERYTHING

UNIVERSITY OF Southampton

Each author and Editor of a work has

- Given Name (Marvin)
- Family Name (Fenderson)
- Optional Identifer

UNIVERSITY OF Southampton

Each author and Editor of a work has

- Given Name (Marvin)
- Family Name (Fenderson)
- Optional Identifer
  - *Semantics defined by repository*
  - *Can be as simple an email address*

UNIVERSITY OF Southampton

With ID:

http://eprints.myuni.ac.uk/id/person/<ID>

Without:

http://eprints.myuni.ac.uk/id/person/ext-<hash-of-eprintID-and-name>

UNIVERSITY OF Southampton

The Bad

UNIVERSITY OF Southampton

THE CAUTIONARY TALE OF BUILDING 53

Building numbers are assigned by University of Southampton Buildings and Estates.

We have no say in this (yet).

Building 53

Building 53

UNIVERSITY OF Southampton

Building 53

UNIVERSITY OF Southampton

UNIVERSITY OF Southampton

http://id.southampton.ac.uk/building/

53/2009

<building-number>/<year-of-construction>

UNIVERSITY OF Southampton

http://id.southampton.ac.uk/building/

53/2009

<building-number>/<year-of-construction>

…But that means every system referring to a building needs to know when it was built.

UNIVERSITY OF Southampton

http://id.southampton.ac.uk/building/

53

"the building identified as '53' by the University of Southampton when this data was compiled"

Pragmatic vs Perfect(?)

University of Southampton

Perfect Identifiers

- Will work for all time
- Are unambiguous
- Require no context
- Put effort on data owners
- Are (in some cases) a pipe dream
- Tend to be more hassle for “casual” comsumers

Pragmatic Identifiers

- Requires supporting data
- Put effort on data linkers
- Make preservation harder
- Are sometimes all you can hope for
- Easier for “casual” consumers

UNIVERSITY OF Southampton

The Curry

UNIVERSITY OF Southampton

Or…

IDENTIFIER AMBIGUITY AND THE SQUIFFY TUMMY

UNIVERSITY OF Southampton

http://id.southampton.ac.uk/products-and-services/

CurlyFries

Daily Menu Item URI

UNIVERSITY OF Southampton

http://id.southampton.ac.uk/products-and-services/

ChickenCurry/2012-07-19

University of Southampton

Today: 2012-07-18

Chicken Curry

Contains Dairy?

False

UNIVERSITY OF Southampton

Today: 2012-07-18

Chicken Curry

Contains Dairy?

False

2012-07-03

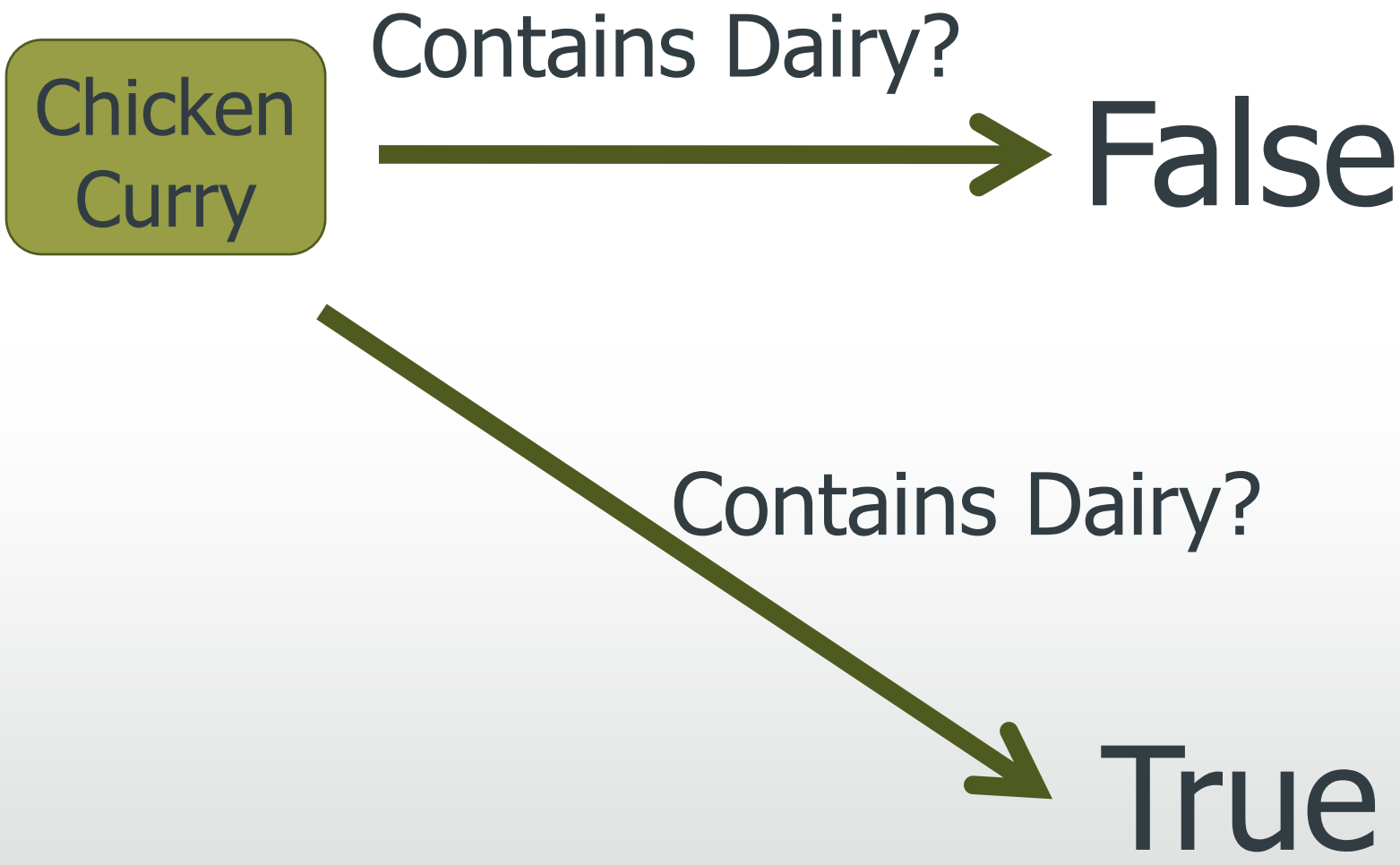
Chicken Curry

Contains Dairy?

True

The archived version…

University of Southampton



Summary

University of Southampton

- With Linked Data you'll be dealing with other people's Identifiers
- You won't like them
- They will be too vague (or too specific)
- They will change

Conclusions

University of Southampton

Create identifiers at the appropriate level of specificness

- *(don't over do it!)*

Don't expect other data providers to share your priorities

If data you wish to preserve links to open data, consider preserving some of the context from the 3rd party data

- *Labels*
- *Classes*
- *Whatnot*

UNIVERSITY OF Southampton

Christopher Gutteridge
cjg@ecs.soton.ac.uk
@cgutteridge

http://data.southampton.ac.uk/
http://data.ac.uk/