Engaging Content
Engaging People

Data Quality for Geospatial Linked Data at OSi

Prof. Declan O'Sullivan & Dr. Rob Brennan

(declan.osullivan@scss.tcd.ie, rob.brennan@scss.tcd.ie)

School of Computer Science and Statistics

ADAPT @ Trinity College Dublin, Ireland

The ADAPT Centre is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

Science Foundation Ireland (SFI) ADAPT

- SFI funded Research Centre from Jan 2015,
- SFI and industry funding of €49M over 6 years
- 29 academics across TCD, DCU, DIT, UCD

- Wholly or part funded industry-collaborative research

Geopspatial Linked Data Project co-funded with Ordnance Survey Ireland (OSi)

- **Goal**: Develop a semantic architecture and Linked Data platform for the OSi taking into account best practices and guidelines in the domain of geospatial information and industry and OSi's current technology stack.

- Started with the boundaries dataset, which was open and already available on data.gov.ie, but not as Linked Data.

What had to be done?

Features and *Geometries* with GeoSPARQL

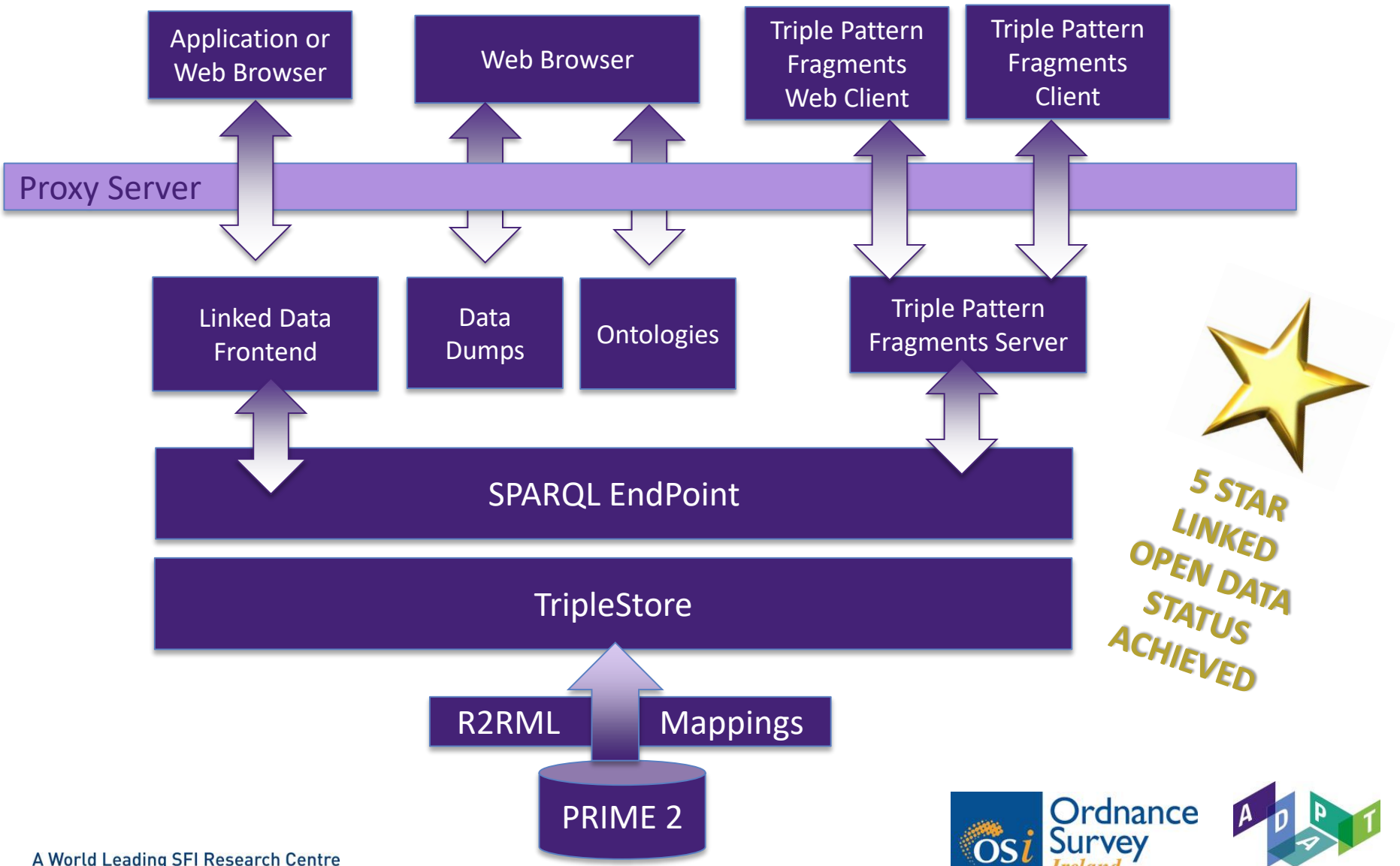
Ontologies developed and published

Modelling Provenance of *Statute Instruments* with **PROV-O**

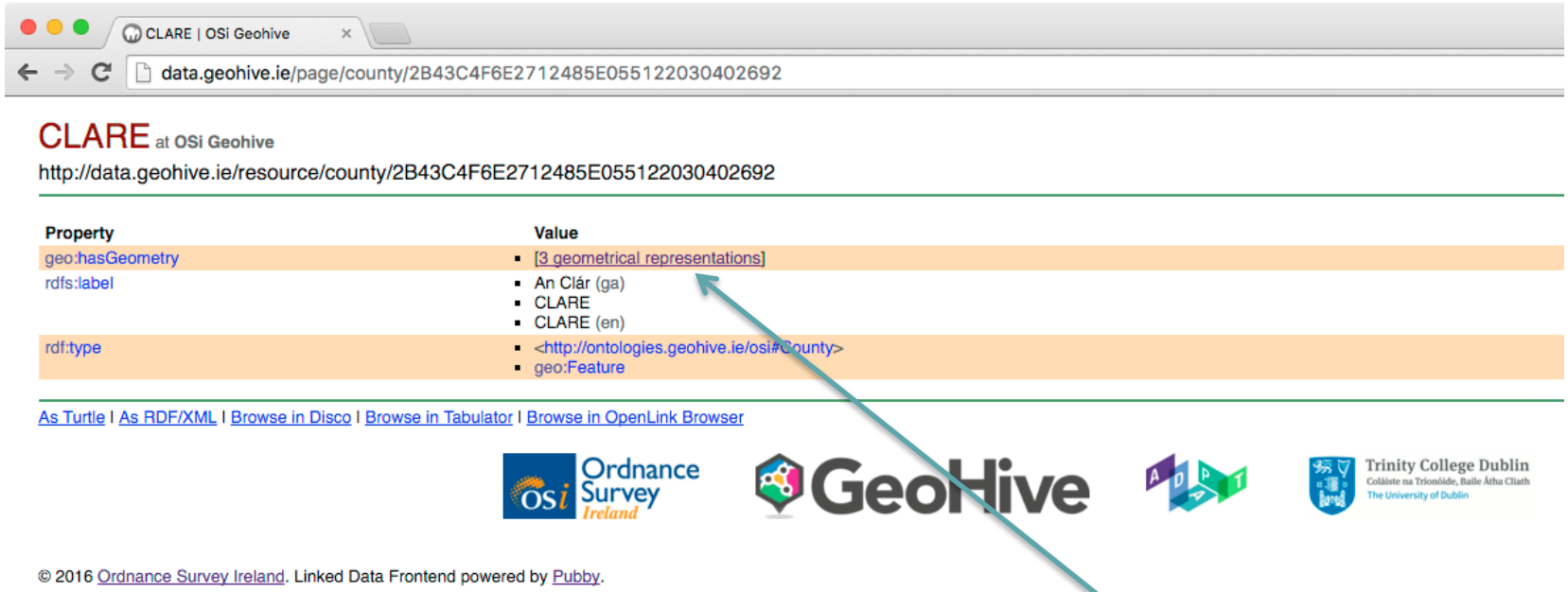
Workshops and agreement with DPER and CSO on **URI Strategy**

Uplift the Prime2 data into RDF with declarative R2RML mappings

Architecture of the LD Platform

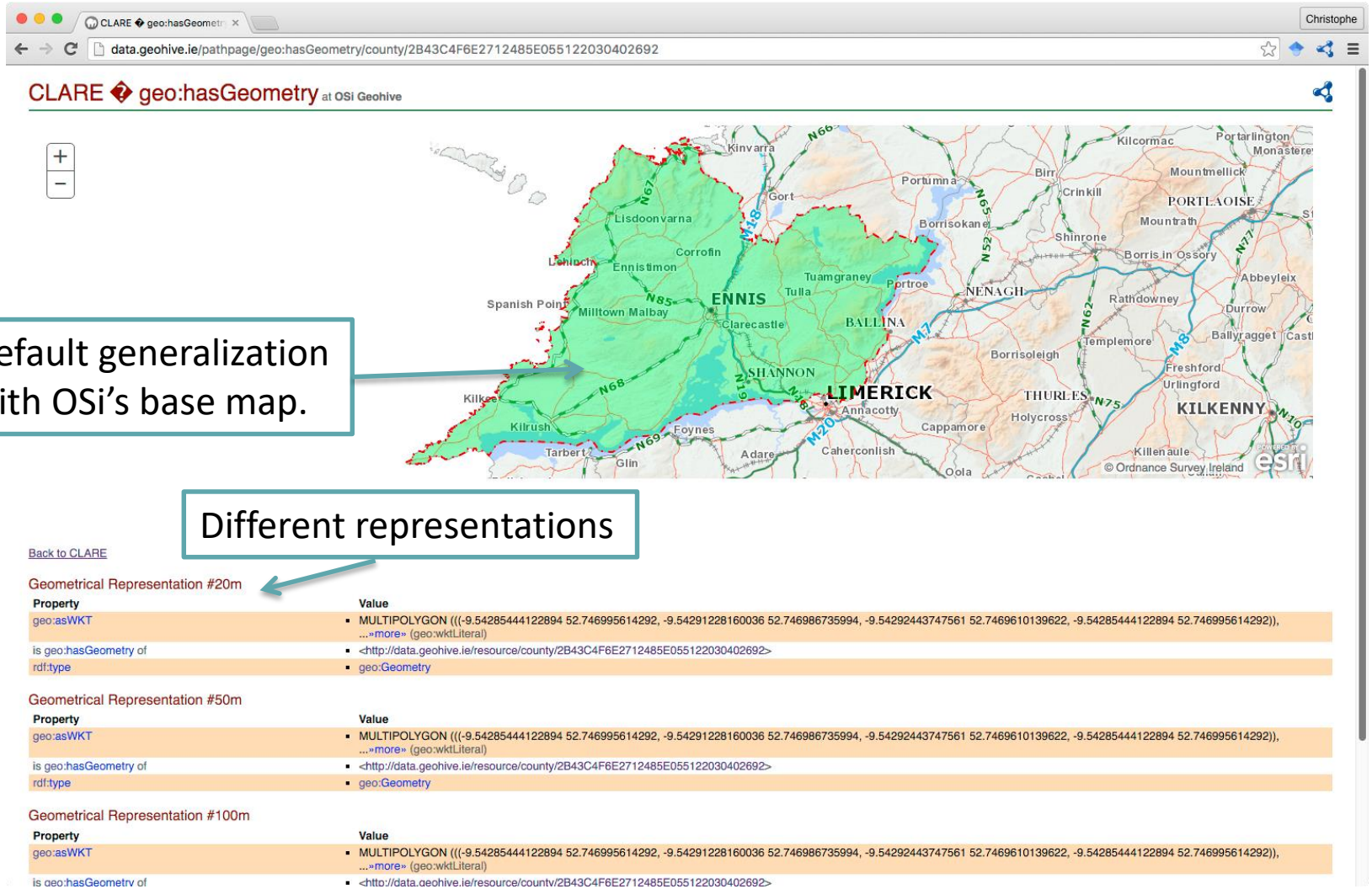


Implementation: Linked Data Frontend



Description of County Clare linking to its three representations

Implementation: Linked Data Frontend



Implementation: RDF view

2AE19629144F13A3E055000000000001-3

Format Tags

Quick Insert


```
<?xml version="1.0"?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:j.0="http://ontologies.geohive.ie/osi#"
    xmlns:ov="http://open.vocab.org/terms/"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#"
    xmlns:geo="http://www.opengis.net/ont/geosparql#"
    xmlns:osi="http://www.geohive.ie/ontology/osi#"
    xmlns:geof="http://www.opengis.net/def/function/geosparql/"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:foaf="http://xmlns.com/foaf/0.1/"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#">
  <geo:Feature rdf:about="http://data.geohive.ie/resource/county/2AE19629144F13A3E055000000000001">
    <rdfs:label xml:lang="en">DUBLIN</rdfs:label>
    <rdfs:label>DUBLIN</rdfs:label>
    <rdf:type rdf:resource="http://ontologies.geohive.ie/osi#County"/>
    <geo:hasGeometry>
      <geo:Geometry rdf:nodeID="A0">
        <geo:asWKT rdf:datatype="http://www.opengis.net/ont/geosparql#wktLiteral"
        >MULTIPOLYGON (((-6.17322835071853 53.4550587605824, -6.17324345299026 53.4550707210097, -6.17324216254192 53.4550537767041, -6
.17322835071853 53.4550587605824)), ((-6.17329096151098 53.4551163520743, -6.17331957312676 53.4551107316691, -6.17328786520046 53.4550
993801516, -6.17329096151098 53.4551163520743)), ((-6.17312028130295 53.4551606954401, -6.17315399560328 53.4551677544183, -6.173128751
27146 53.4551415904783, -6.17312028130295 53.4551606954401)), ((-6.08648660509308 53.3642989681213, -6.08650595979755 53.3643170664192,
-6.0865318305604 53.3642883500737, -6.08648660509308 53.3642989681213)), ((-6.08480359581157 53.3628327986365, -6.08481884316608 53.362
8137993202, -6.08478080293847 53.3627902145444, -6.08480359581157 53.3628327986365)), ((-6.08913429868248 53.2768485839977, -6.08918413
57233 53.2768222466741, -6.08914216545053 53.27681443712, -6.08913429868248 53.2768485839977)), ((-6.08118863931207 53.3610358713555, -
6.08119951194165 53.3610139348548, -6.0811088574775 53.3610056907902, -6.08118863931207 53.3610358713555)), ((-6.08819642115663 53.3652
```


Ordnance Survey Ireland

Science Foundation Ireland For what's next

Engaging Content
Engaging People

Assessing Linked Data quality

Objectives

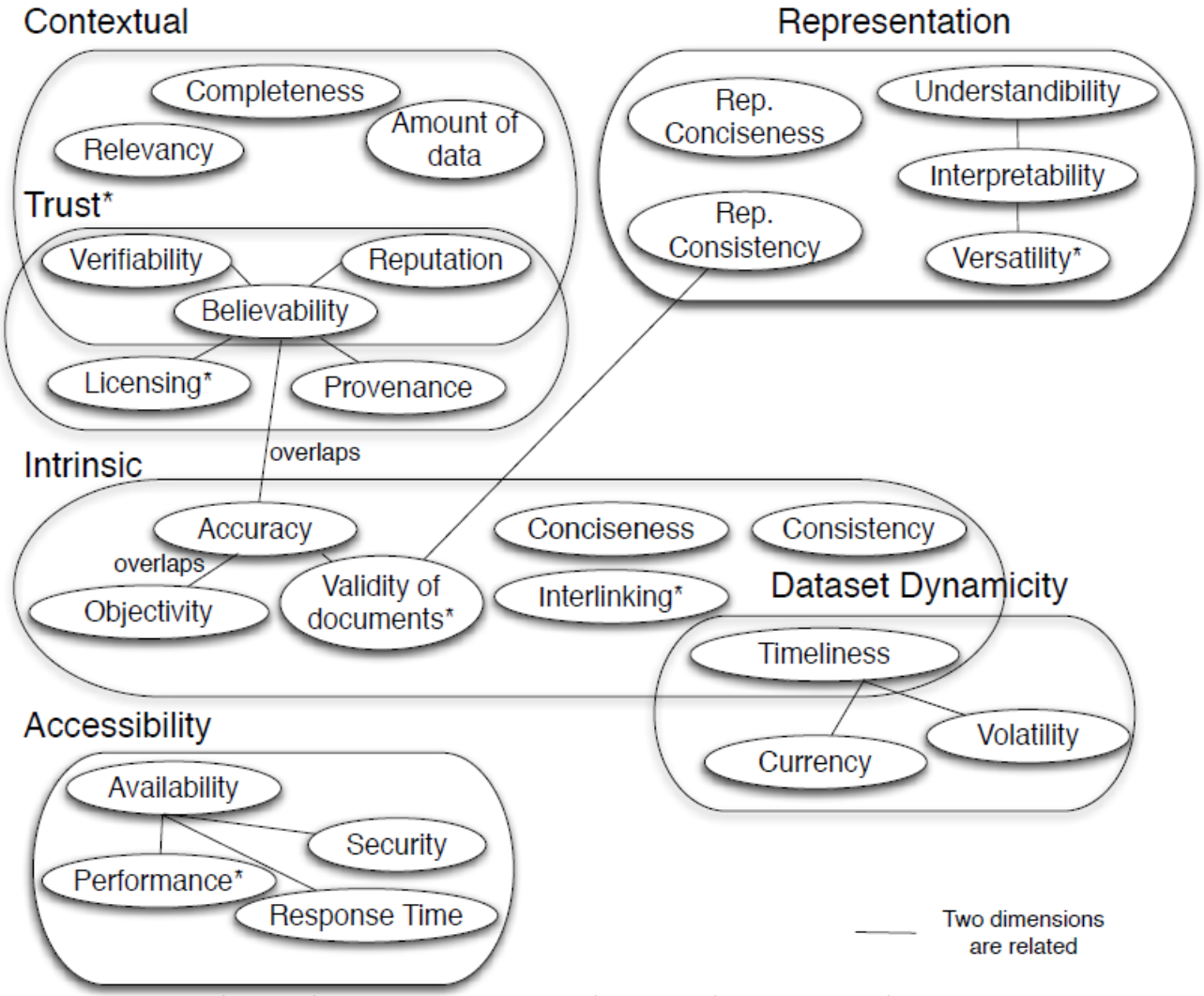
- Define a quality assurance process and tools for OSi Linked Data
  - Increase data quality
  - Build trust
  - Ease data consumption
  - Assist R2RML mapping maintenance
- Explore semantic data quality feedback for Prime2
  - OSi ontology as a source of validation rules [1]
- Contribute to OSi data quality and governance infrastructure
- **Not assessing survey accuracy**

[1] Kevin Feeney, Gavin Mendel-Gleason and Rob Brennan, Linked data schemata: fixing unsound foundations, *Semantic Web – Interoperability, Usability, Applicability*, 2017, p1435-2647

Steps

- Define quality metrics for data.geohive.ie
  - Examine how Linked Data quality is applicable
  - Use W3C quality metadata standards
- Deploy ADAPT's Luzzu quality metrics monitoring tool
- Generate and analyse metrics
  - Define thresholds
  - Address quality issues (data & process)

Linked Data Quality Dimensions



From: Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, Sören Auer, Quality assessment for Linked Data: A Survey, March 2015
Semantic Web 7(1):63-93, DOI: 10.3233/SW-150175

Linked Data Quality Assessment

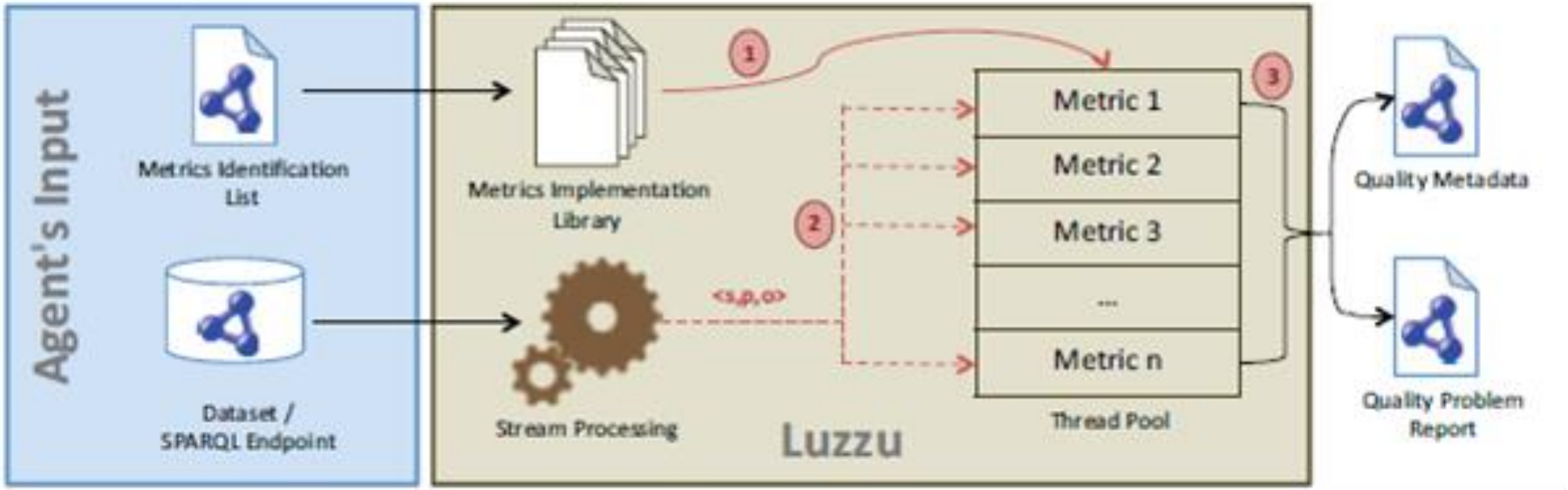
- Linked Data schemas
  - Theoretically provide basis for semantic quality assurance
  - Practically
    - Most Linked Data uses OWL/RDFS in ad hoc ways
    - Open world assumption blocks quality assessment
    - Knowledge models do not include traditional RDBMS integrity constraints
- Assessment Methods
  - Profiling /statistical methods
  - W3C SHACL constraint language
  - SPARQL-based assessment
  - Ontological inference

Luzzu Linked Data Quality Assessment Tool[1]

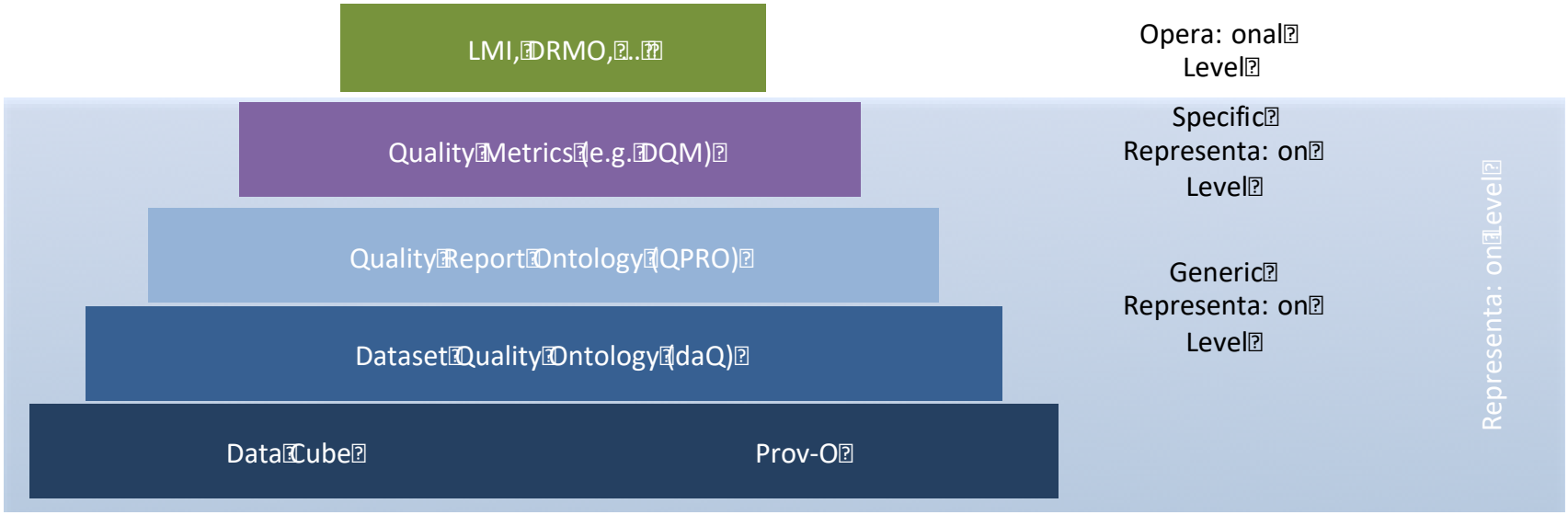
- Open source tool
- Stream-processing architecture for big data
- Produces linked data metrics and quality reports (encoded as Linked Data)
- 25 pre-defined Linked Data metrics
- Custom metrics via:
  - Declarative domain specific language for metrics
  - Java code plug-ins

1 http://theme-e.adaptcentre.ie/daq/daq.html
http://eis-bonn.github.io/Luzzu/

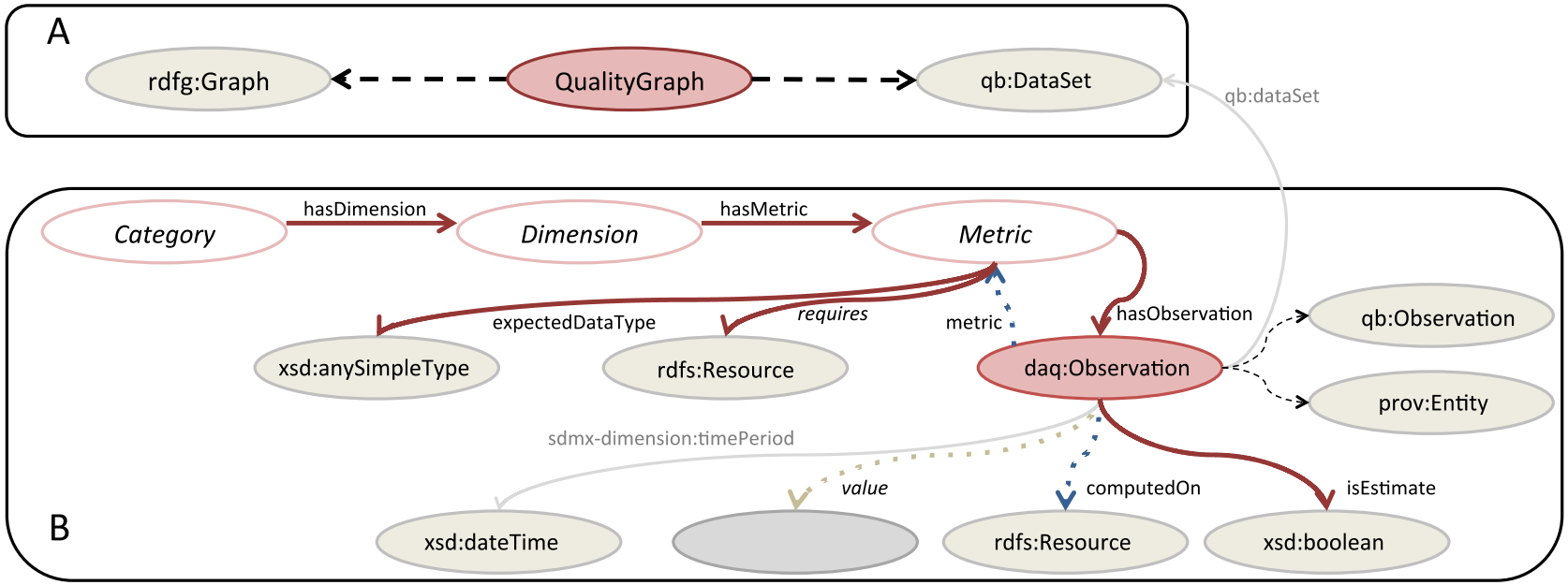
Luzzu Architecture



Luzzu Semantic Layers



Luzzu – Dataset Quality Vocabulary



Classes

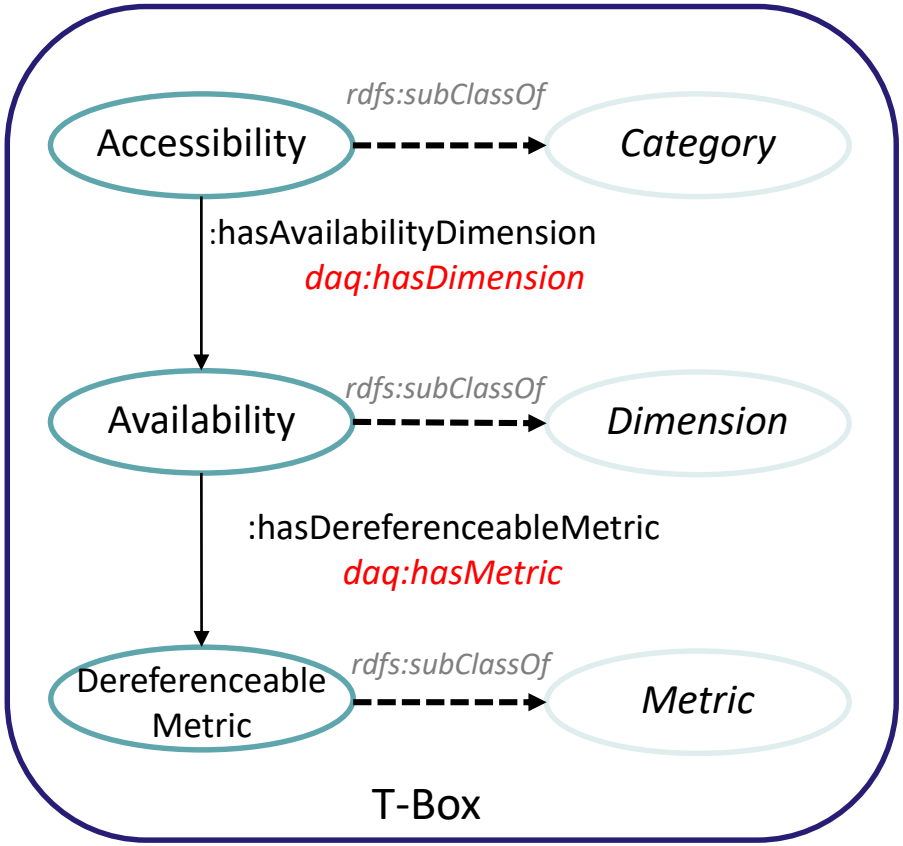
- rdfg:Graph — Concept in an existing ontology
- Dataset — Concept in proposed ontology
- *Category* — Abstract concept in proposed ontology (Not for direct use)
- Undefined object: Concept or Literal

Properties

- Subclass of
- hasObservation — Property in proposed ontology
- computedOn — Property is of type qb:DimensionProperty
- value — Property is of type qb:MeasureProperty
- qb:dataSet — Property in existing ontology

http://purl.org/eis/vocab/daq

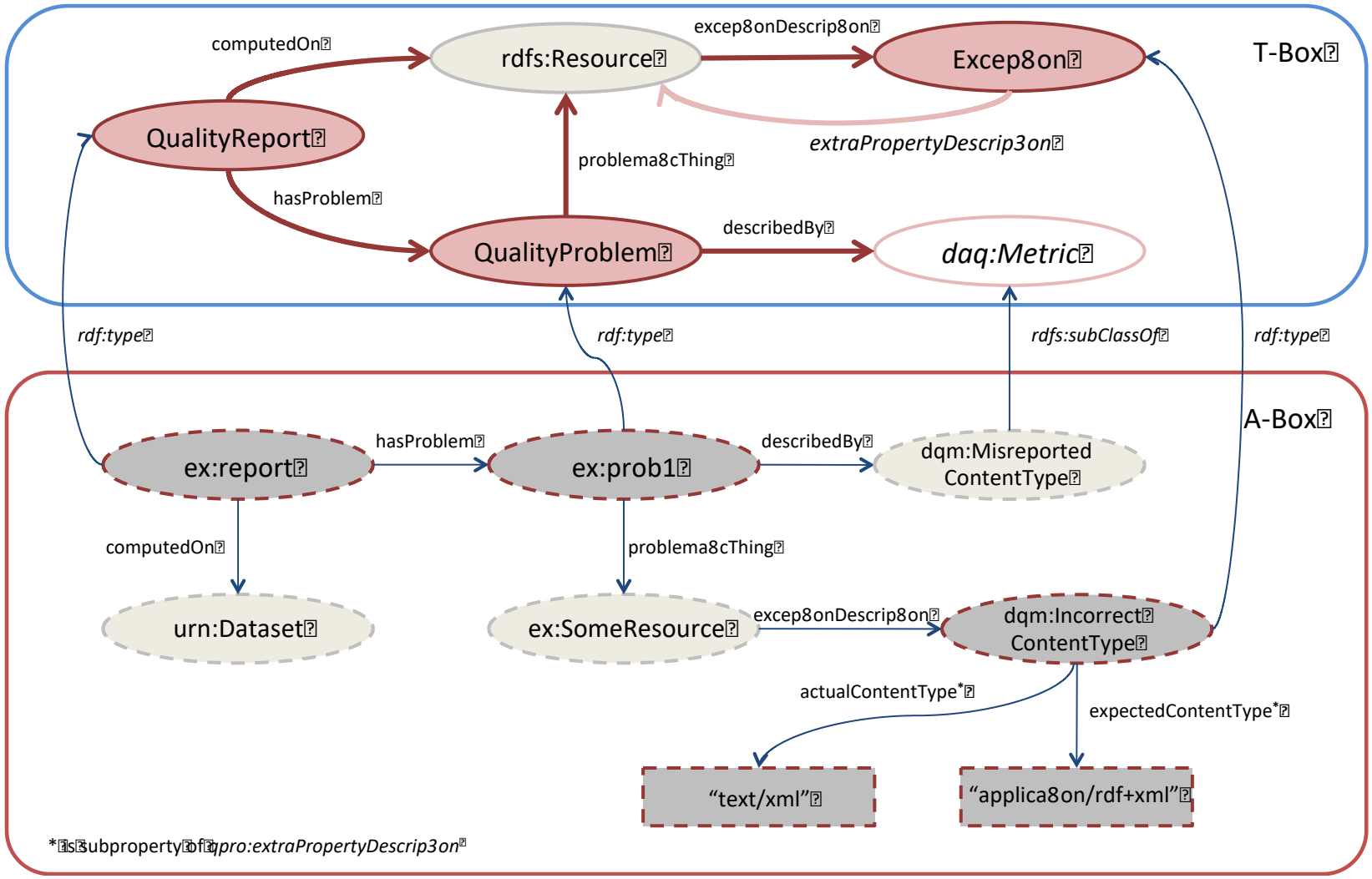
daQ – T-Box example for defining metrics in Luzzu



An Example:

https://goo.gl/zp1Y1N

Luzzu – Quality Problem Report



http://purl.org/eis/vocab/qpro

Example: https://goo.gl/mfKJu2

OSi Boundary Dataset Quality Metrics

- 19 metrics selected from Representational, Contextual, Intrinsic and Accessibility categories
- Example metrics:

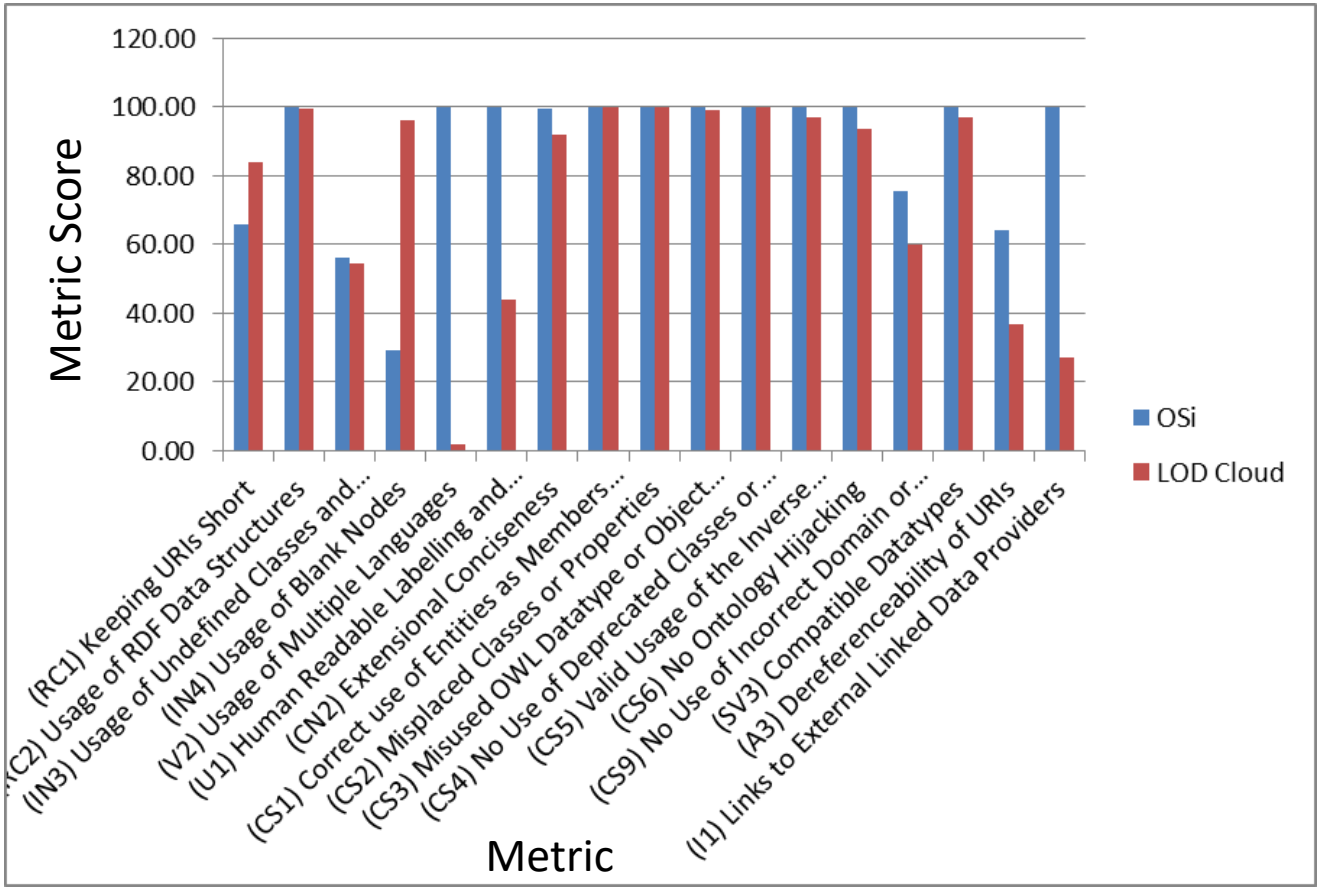
Code	Name	Description
Representational Category Metrics		
RC1	Keeping URIs Short	Observations on the length of URIs, best practice favours shorter URIs (<80 chars).
RC2	Minimal Usage of RDF Data Structures	Use of RDF features like reification, containers, and collections, is discouraged.
IN4	Usage of Blank Nodes	Best practice favours minimal usage.
IN3	Usage of Undefined Classes and Properties	Detecting the use of classes and properties without a formal definition, perhaps due to typos rather than omission.
V2	Usage of Multiple Languages	This metric checks the number of languages a dataset supports. Specifically, whether the data is evenly available in different languages.
Contextual Category Metrics		

Assessment Results

- Generally high scores for state of the art Linked Data metrics
- But 7 metrics fell significantly below 100%

But what does this mean?

- Comparison to a general survey of the quality of LOD [1]



[1] Jeremy Debattista, Christoph Lange, and Sören Auer, and Dominic Cortis, Evaluating the Quality of the LOD Cloud: An Empirical Investigation, Accepted, Semantic Web Journal, available at: http://www.semantic-web-journal.net/system/files/swj1757.pdf

Conclusions

- Luzzu Linked Data quality assessment tool provided useful insights
- Need a QA process for Linked Data release
- Goal-setting for OSi linked data quality is also important
  - Thresholds require business as well as technical input
- Custom metrics will be necessary for the above
- Follow-on actions to improve the quality of the OSi dataset(s):
  - Review URI schemes to ensure short URIs are achieved where possible
  - Increase the number of datasets interlinked to
  - Increase the amount of metadata published about the datasets
  - Fix validation errors detected e.g. class name typos in mappings

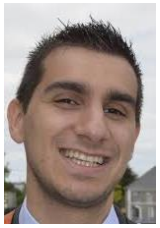
With thanks to the Team

Ordnance Survey Ireland (OSi)

Lorraine McNerney

Eamonn Clinton

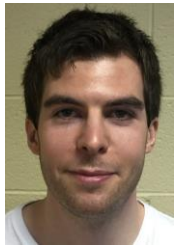
ADAPT @ Trinity College Dublin



Jeremy Debattista
Research Fellow

Christophe Debruyne
Research Fellow

Kris McGlinn
Research Fellow



Alan Meehan
Research Fellow

Darragh Blake
DLAB Research Engineer

Aoife Brady
DLAB Project Manager