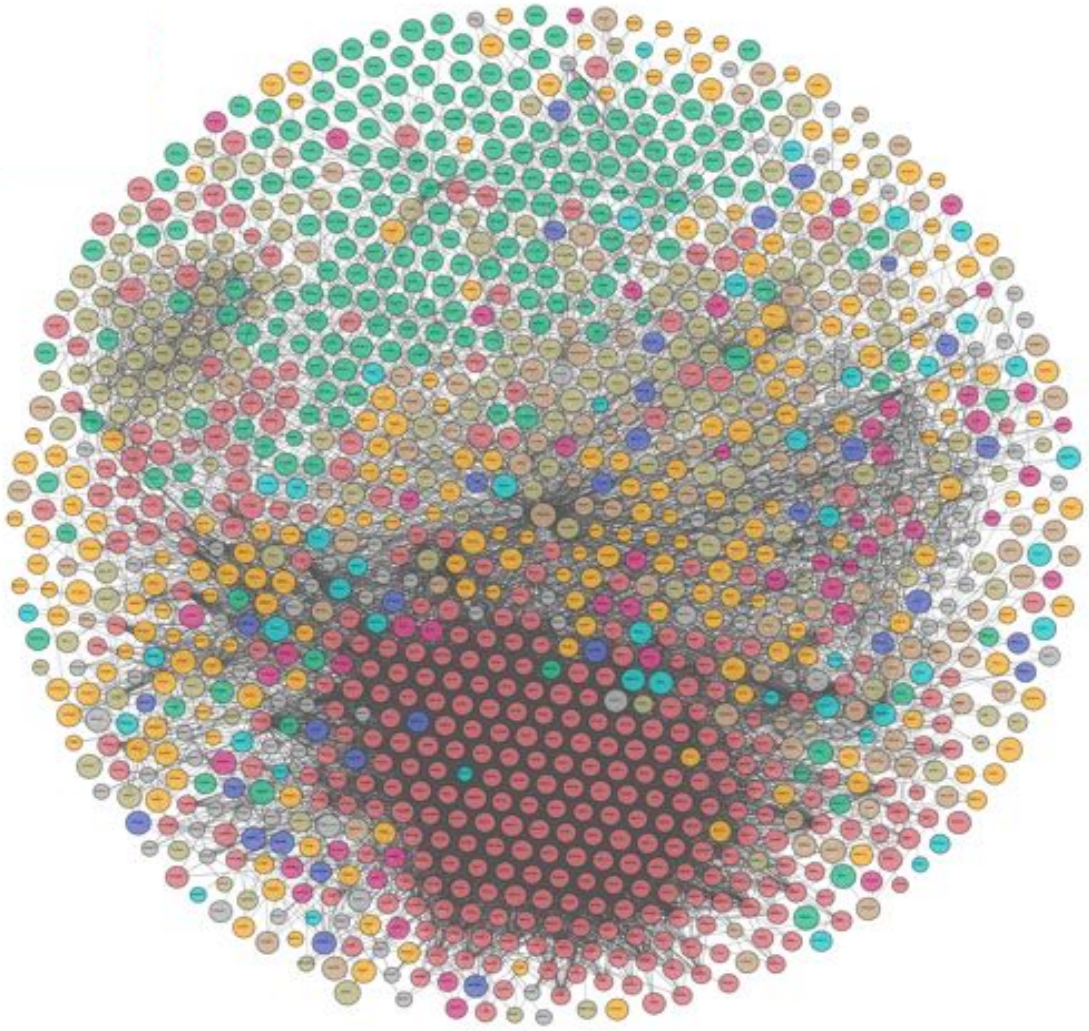
Is the LOD cloud at risk of becoming a museum for datasets? Looking ahead towards a fully collaborative and sustainable LOD cloud

Jeremy Debattista, Judie Attard, Rob Brennan, Declan O'Sullivan
ADAPT Centre, Trinity College Dublin, Ireland

This research has received funding from the Irish Research Council Government of Ireland Postdoctoral Fellowship award (GOIPD/2017/1204) and the ADAPT Centre for Digital Content Technology, funded under the SFI Research Centres Programme (Grant 13/RC/2106) and co-funded by the European Regional Development Fund.

The Current LOD Cloud



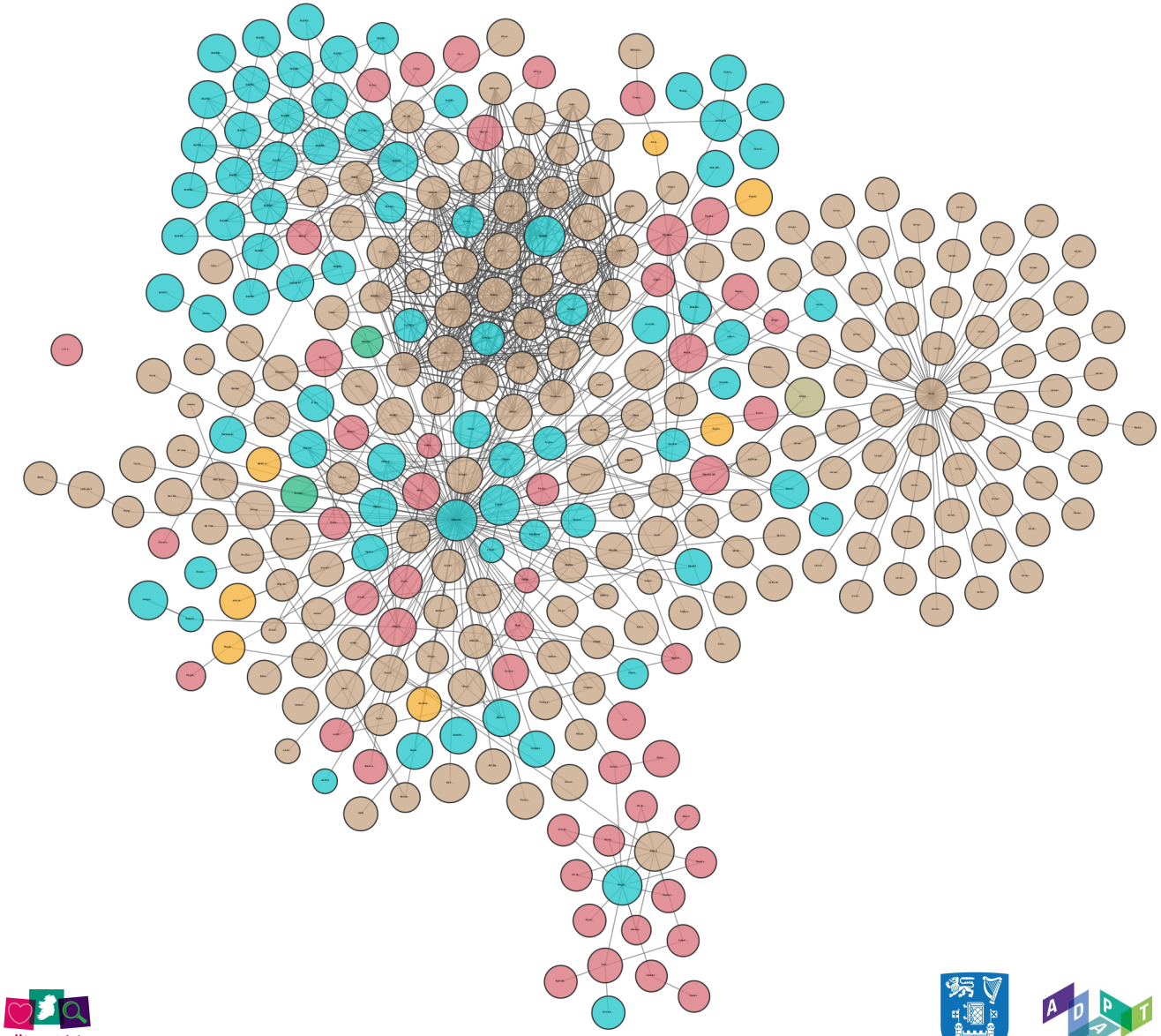
29/3/2019 - CC-BY http://lod-cloud.net

- 1,239 depicted datasets (increase of 5 from Jan' 19)
- Depicts datasets that have been published in Linked Data
- *Clustered catalog* of individual domain specific KGs demonstrating cohesion between interlinks and intralinks
- An image with embedded *metadata*

... but how much can we access?

Legend

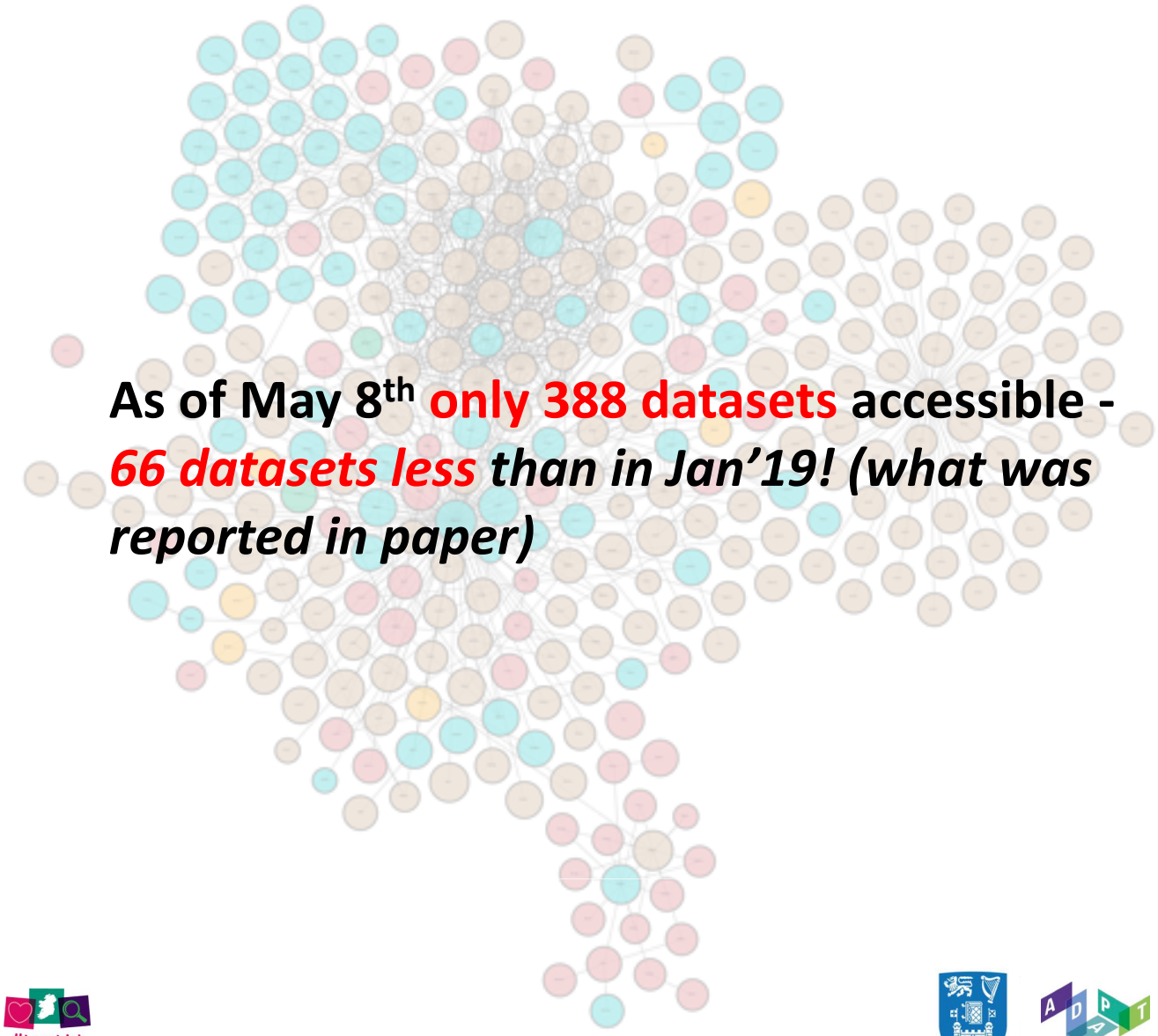
- Data Dump
- SPARQL
- voID
- Data Dump and SPARQL
- Data Dump and voID
- SPARQL and voID
- All Three Access Points



... but how much can we access?

Legend

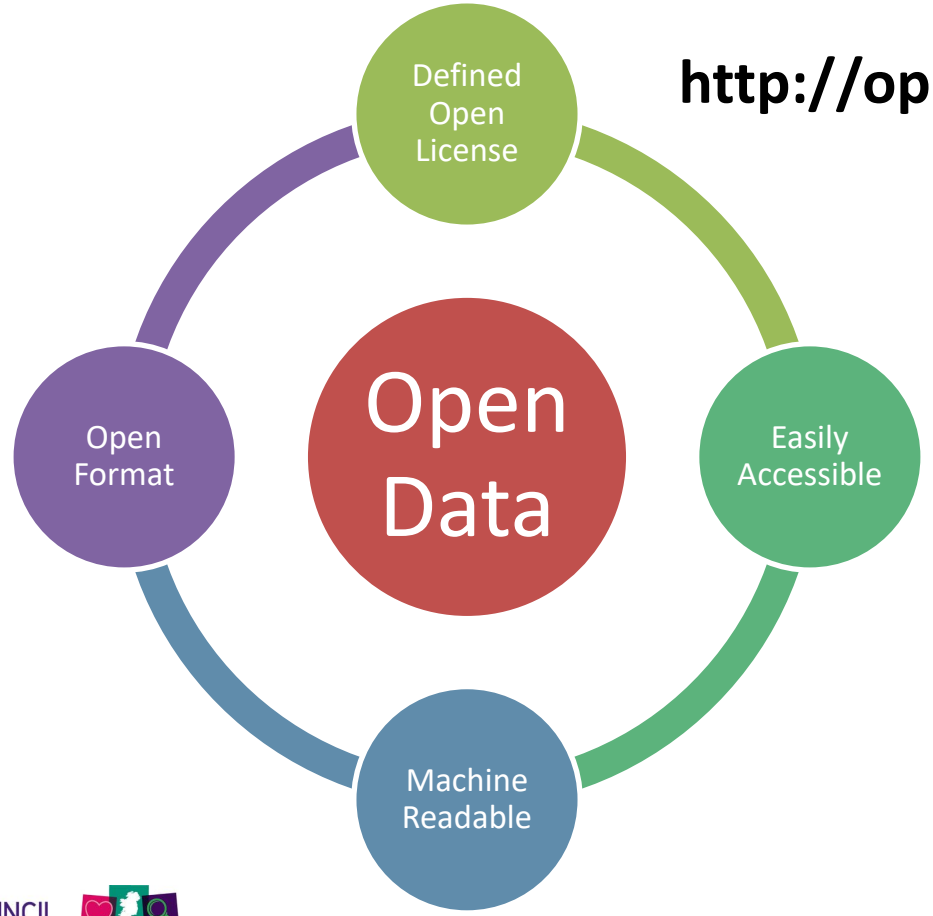
- Data Dump
- SPARQL
- voID
- Data Dump and SPARQL
- Data Dump and voID
- SPARQL and voID
- All Three Access Points



As of May 8th only 388 datasets accessible - *66 datasets less than in Jan'19! (what was reported in paper)*

The *Open Data* definition

Open Data should *be freely used, modified, and shared by anyone for any purpose*

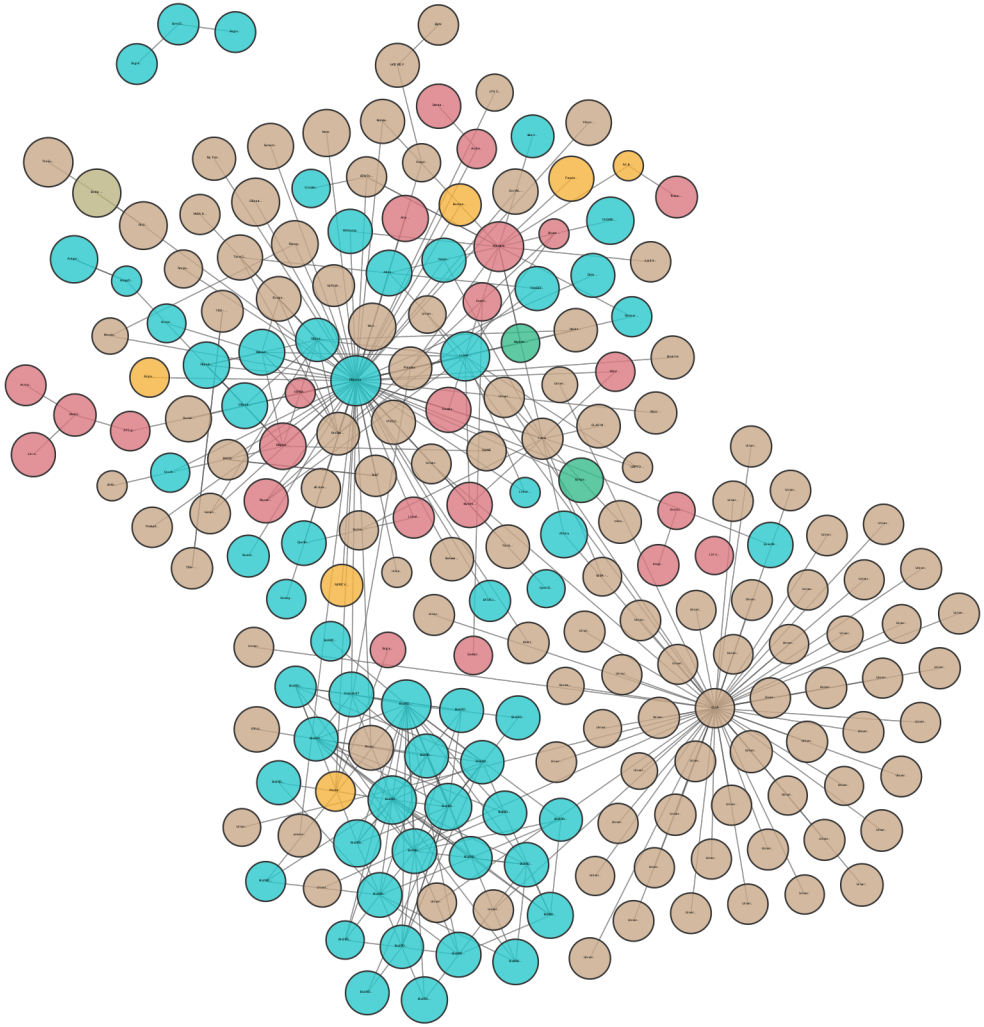


http://opendefinition.org/

... this is what the LOD Cloud should look like

Legend

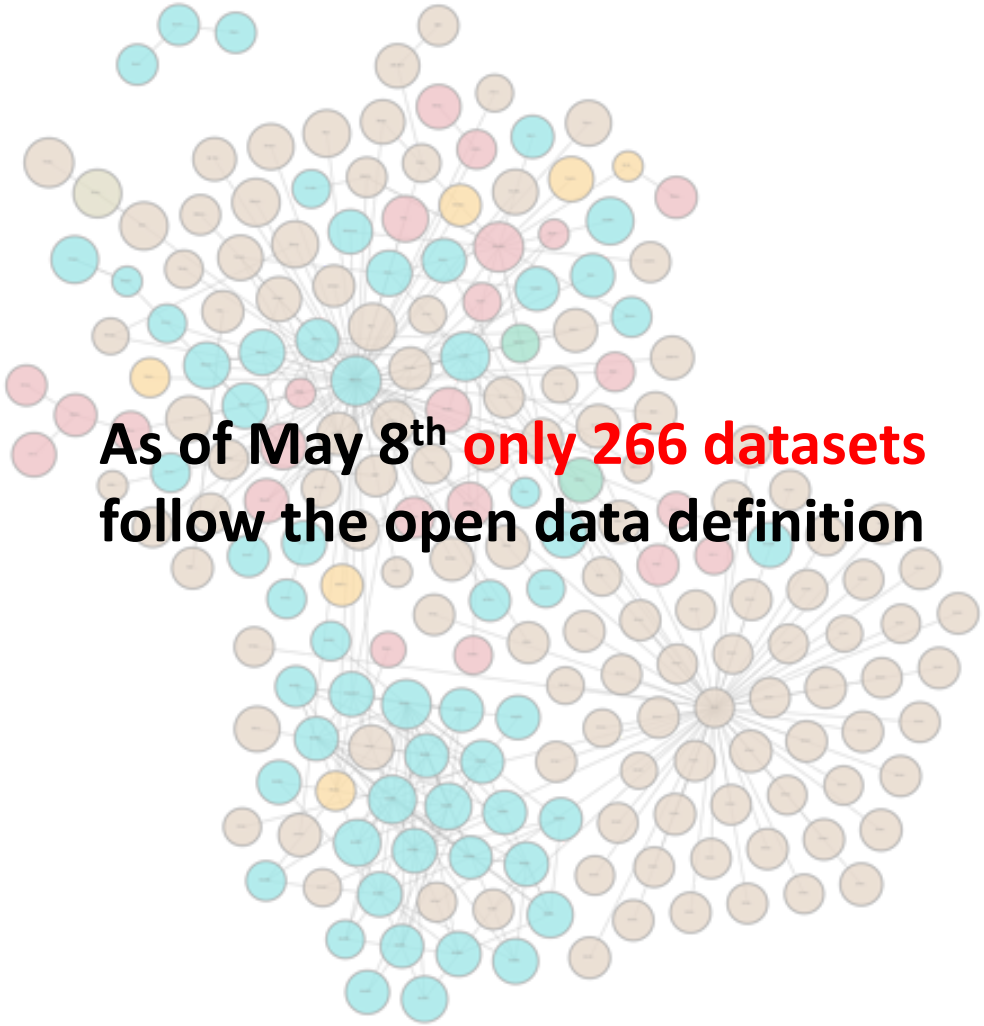
- Data Dump
- SPARQL
- voID
- Data Dump and SPARQL
- Data Dump and voID
- SPARQL and voID
- All Three Access Points



... this is what the LOD Cloud should look like

Legend

- Data Dump
- SPARQL
- voID
- Data Dump and SPARQL
- Data Dump and voID
- SPARQL and voID
- All Three Access Points



As of May 8th only 266 datasets follow the open data definition

Agenda

METADATA ANALYSIS

SUSTAINABLE STRATEGIES

Agenda

METADATA ANALYSIS

Metadata Analysis

Analysis:

- Identification of licenses for datasets in metadata
- Identification of the format/media types of available datasets
- Identification of dataset access points

Purpose:

- Discoverability and Openness of datasets in the LOD cloud

Not relevant for analysis:

- Size/Number of Triples in a dataset
- Number of external interlinks

NOTE: At this stage its only metadata analysis

Metadata Analysis – Gathering the Data

- LOD cloud provides a JSON file with all datasets: https://lod-cloud.net/lod-data.json
- Discrepancy between the JSON metadata and the voID metadata generated/provided by the LOD cloud
- Jupyter notebook available: https://github.com/jerdeb/lodexperiments

Metadata Analysis - Licenses

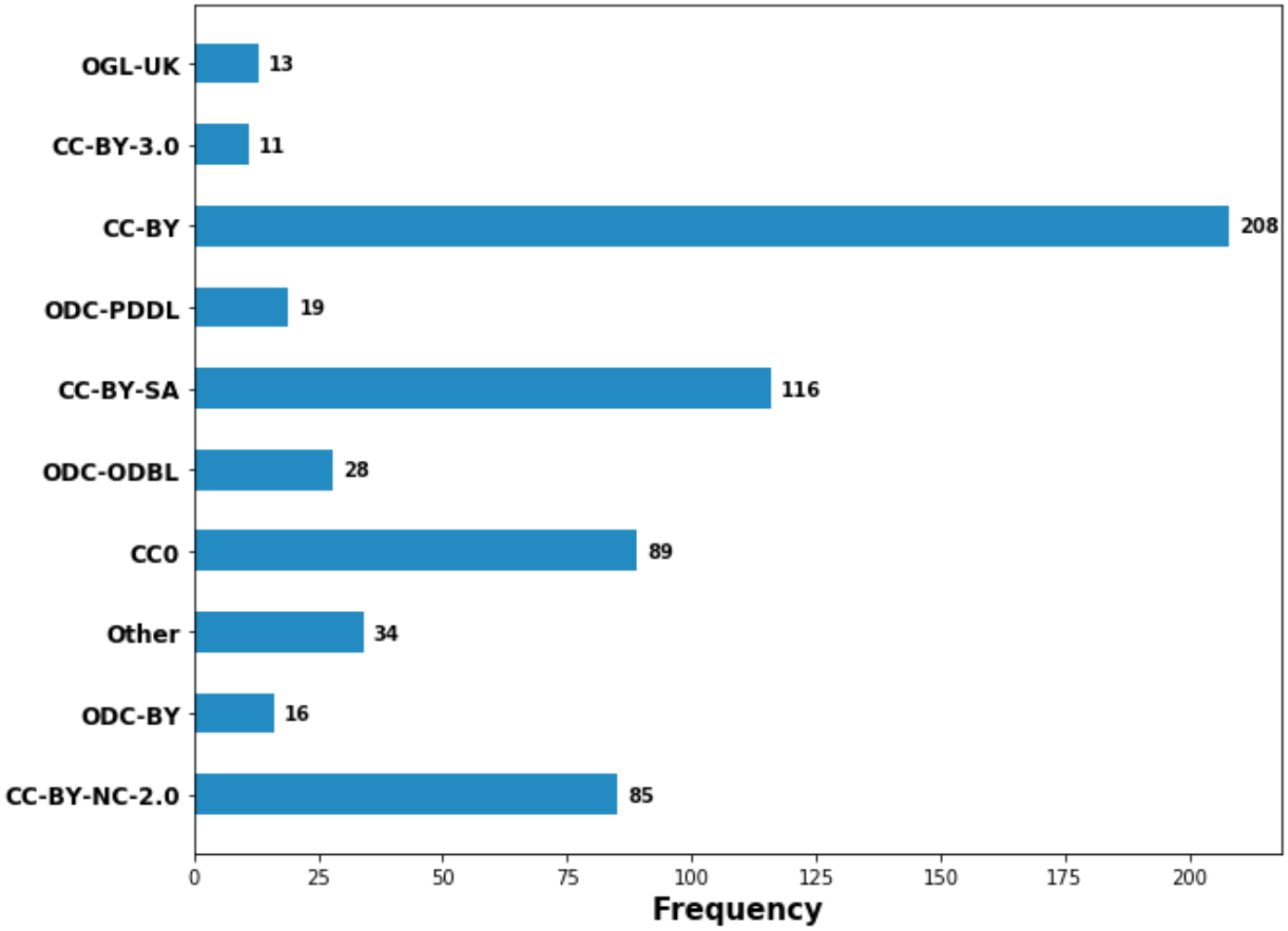
- JSON key: *license*
- Conformant Licenses: https://opendefinition.org/licenses/

Results:

- Number of dataset with a defined license: ***619 datasets*** (~ 45%, an increase of 5% from the observation done in 2015)
- Number of datasets with a conformant license: ***530 datasets***

- Regex: *license* or *copyright* and one of *under*, *grant* or *right*
  - 22 matches: 10 datasets with conformant licenses; 4 bad matches
  - 3 datasets with conflicting license between the description and the license field

Metadata Analysis - Licenses



Metadata Analysis – Media Types for Datasets

- JSON key: *media_type* – for each dataset distribution (download)
- Ideally using a registered Linked Data media type.
- `text/html` the most frequently used, but no RDFa embedded
- A large number of unregistered media types
- 596 distributions using `meta/void` and `meta/rdf-schema` but these are not registered

Media Type	Frequency
mapping/owl	26
meta/owl	27
text/plain	31
application/x-gzip	32
n-quads	32
None	91
application/x-ntriples	91
meta/sitemap	102
application/x-nquads	103
Others	109
application/rdf+xml	114
application/octet-stream	118
HTML	119
application/zip	137
text/turtle	252
meta/void	266
meta/rdf-schema	370
RDF	401
text/html	1107

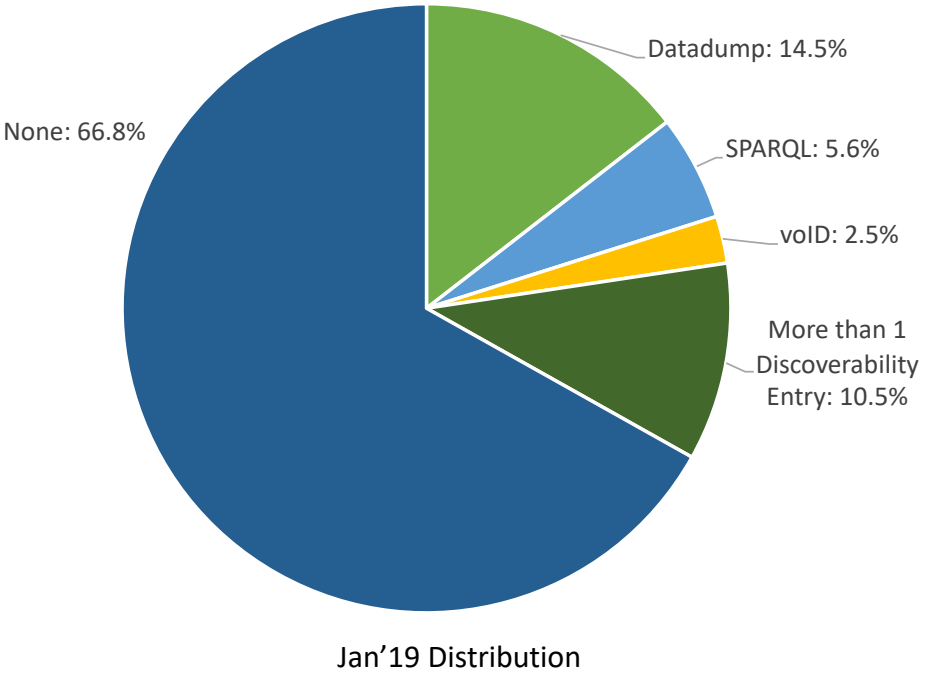
IRISH RESEARCH COUNCIL
An Chomhairle um Thaighde in Éirinn
#LoveIrish Research

Metadata Analysis - Accessibility

- *Potential* access point: data dump, SPARQL endpoint, voID description
- Set criteria for different access points:
  - 10 second timeout (for all)
  - *Data dumps* – distribution tagged with an set of pre-defined media types
  - *SPARQL* – return result for `ASK { ?s ?p ?o }`
  - *voID* – return **true** for `ASK { ?s a void:Dataset }` after loading metadata in memory

Metadata Analysis - Accessibility

- Jan'19: 33% of datasets have a discoverable data access point (454 datasets). 9% less than the observation of 2015
- Majority of dataset have a data dump distribution (199 datasets)
- May'19: only 388 datasets (28%) have an access point, with 226 having a data dump and 65 datasets with more than one discoverability entry point



Jan'19 Distribution

SUSTAINABLE STRATEGIES

The Sustainable LOD Cloud

C1. Publishers should own and maintain the datasets' metadata

C2. Lack of systematic and fine-granular metadata structures

C3. Invalid metadata descriptions

C4. Many dead and outdated datasets listed

C5. Lack of involvement of data consumers in the structure

The Sustainable LOD Cloud

C1. Publishers should own and maintain the datasets' metadata

- Adding dataset = filling google sheet form
- Dataset updated != metadata in LOD cloud updated

The Sustainable LOD Cloud

C2. Lack of systematic and fine-granular metadata structures

- No systematic structure in terms of properties, the property's values, and categorical values
- Attempts to leverage on DCAT and voID standards

C3. Invalid metadata descriptions

- Incorrect value for properties
- e.g DBpedia void:dataDump predicate incorrectly links to the DBpedia download page
- License predicate points to a human-readable page with no semantic description
- Incorrect media type values

The Sustainable LOD Cloud

C4. Many dead and outdated datasets listed

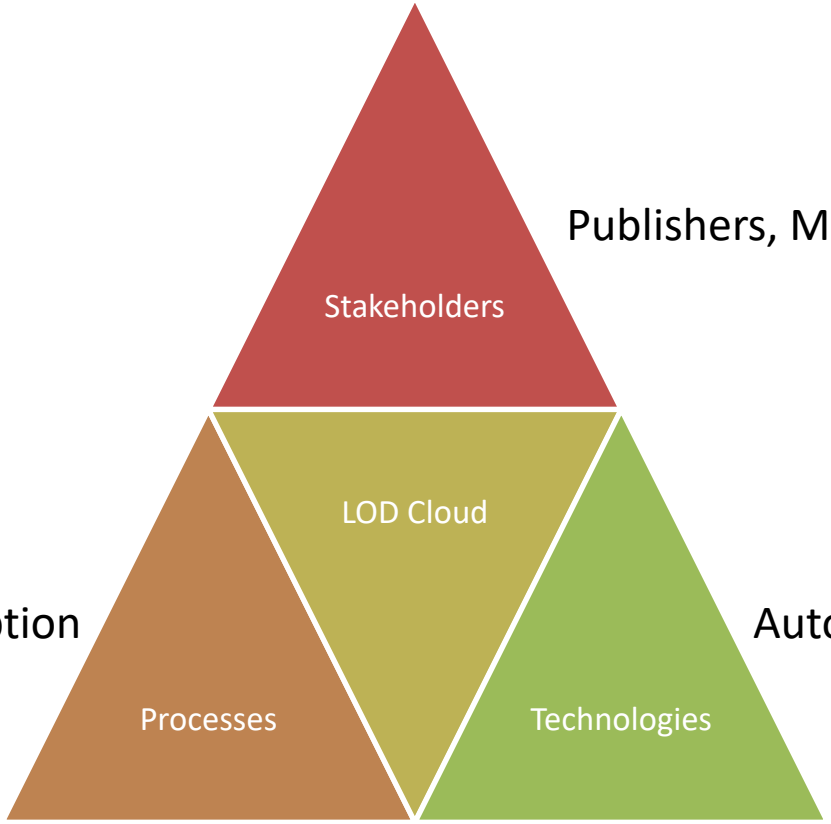
- Datasets not online (e.g. 270a.info datasets), yet still depicted
- Using LOD Laundromat as a preservation/archive tool

The Sustainable LOD Cloud

C5. Lack of involvement of data consumers in the structure

- Difficult to find the relevant dataset
- Parse JSON file or previously use datahub.io

Sustainable Strategies - Pillars



Publishers, Maintainers, Consumers

Publishing and Consumption

Automation and infrastructure

Service Operating Model (C1, C5)

- Federated model
- Culture of interaction between stakeholders
- *Publishers*: provide and maintain metadata, high quality datasets, uptime of endpoints
- *Maintainers*: ensure availability of services related to generation, cataloguing, and maintenance of the cloud, and availability of the cloud itself
- *Consumers*: comment and vote for different data sources

Sustainable Strategies

Identification of critical data elements (C2)

- Who? What? Where? How?
- One standard metadata model
- Glossary for values

Sustainable Strategies

Defining the key activities and control structures (C3, C4)

- Validation and correctness of candidate dataset's metadata
- Heartbeat checks of the availability of dataset distributions
- Prevention of abuse and spamming

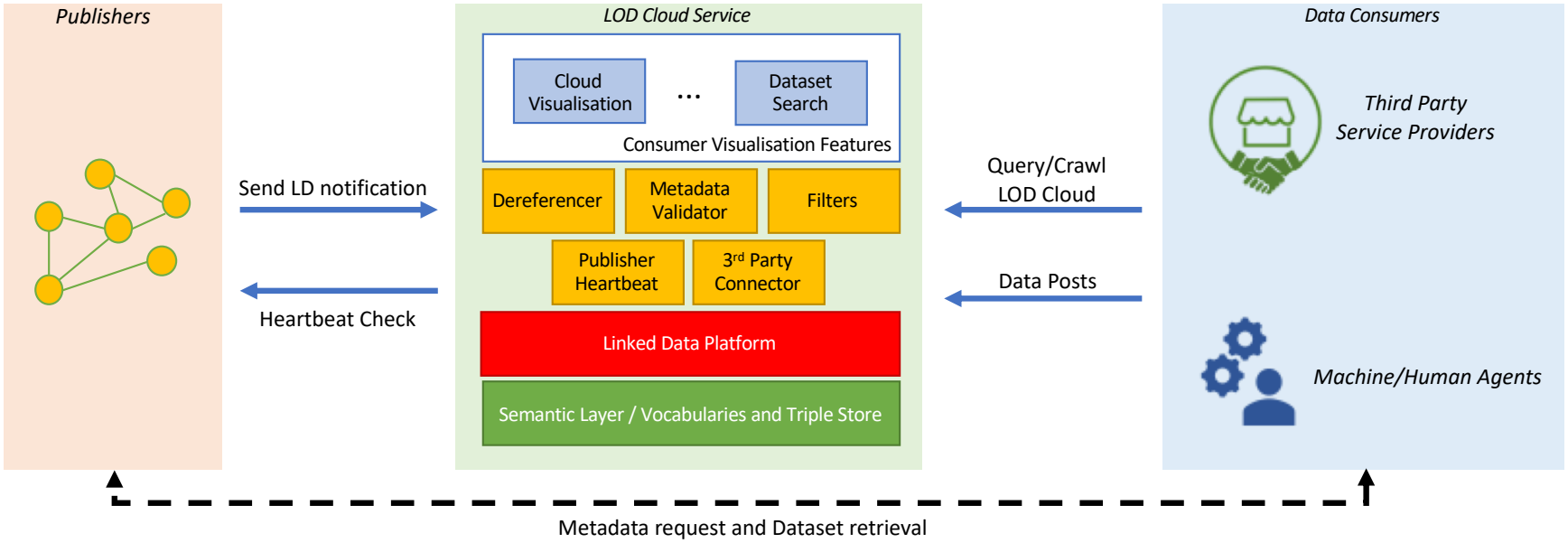
Capabilities for a sustainable architecture

DISCOVERY

UNDERSTANDABILITY

SOCIAL

Proposed architecture



Is the LOD cloud at risk of becoming a museum for datasets?

We need to strategically restructure the LOD cloud as a sustainable service with sound governance, rather than as an academic or research artefact.

jeremy.debattista@adaptcentre.ie

@jerdeb