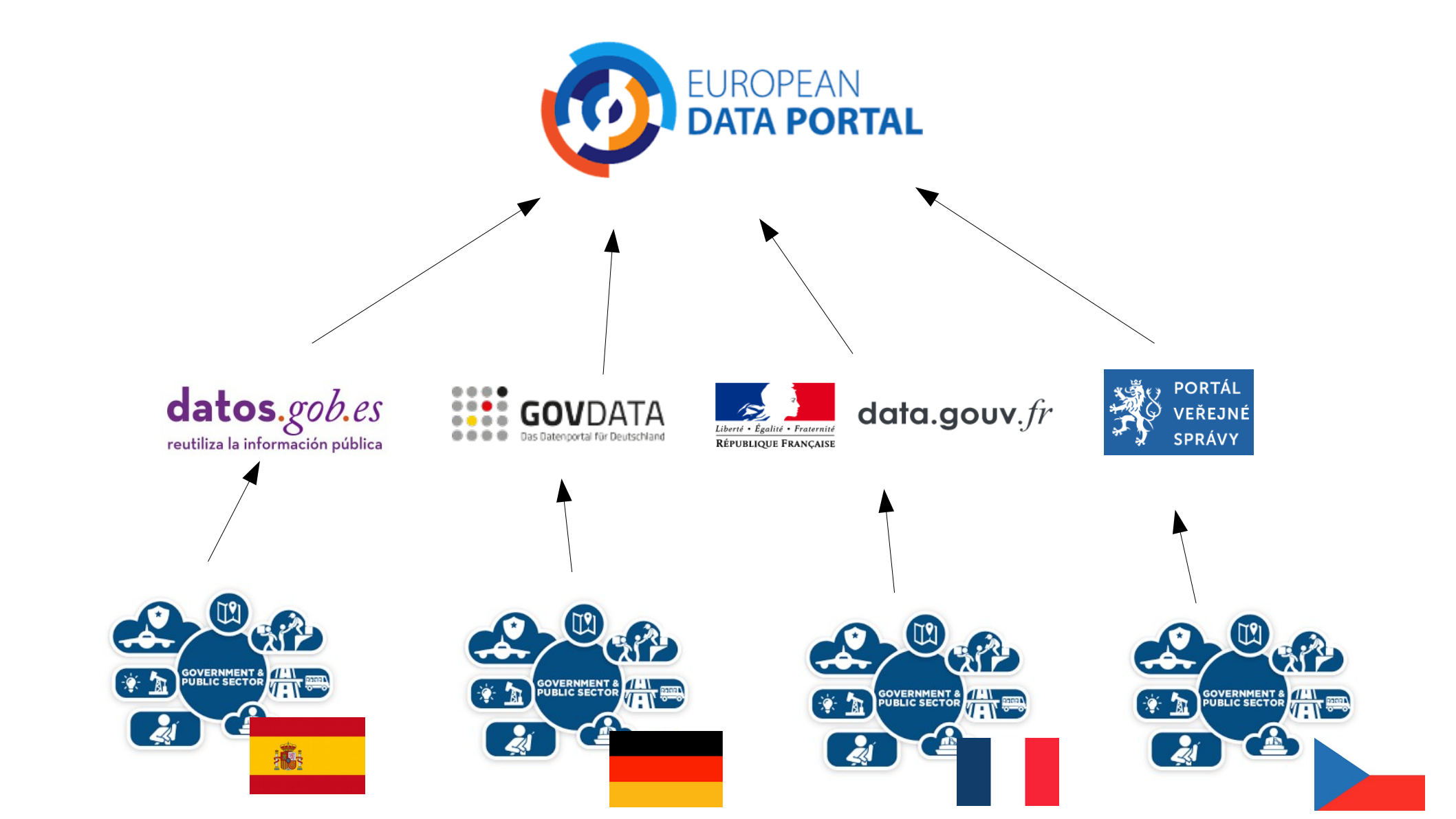
An Assessment of the Adoption and Quality of Linked Data in European Government Data (In-Use track)

Luis-Daniel Ibáñez [1], Ian Millard [2],
Hugh Glaser [2], Elena Simperl [1]

[1] University of Southampton
[2] Seme4 Ltd

European Data Portal

- European Commission initiative
- Harvest metadata of Public Sector Information of European public data portals.
- Index and search engine
- Co-locate tools and documentation


EUROPEAN
DATA PORTAL
datos.gob.es
reutiliza la información pública
GOVDATA
Das Datenportal für Deutschland
Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
data.gouv.fr
PORTÁL
VEŘEJNÉ
SPRÁVY
GOVERNMENT & PUBLIC SECTOR
GOVERNMENT & PUBLIC SECTOR
GOVERNMENT & PUBLIC SECTOR
GOVERNMENT & PUBLIC SECTOR

For more details...

- Industry talk tomorrow morning
  - “The European Data Portal: Scalable Harvesting and Management of Linked Open Data”
    Fabian Kirstein, Simon Dutkowski, Benjamin Dittwald, Manfred Hauswirth

Problems

- Dataset search
- Data Integration
  - A “federation of data lakes”
- Multi-linguality

Linked Data as a solution

- DCAT-AP
  - DCAT extension for public sector
- EU Vocabularies
  - Controlled vocabularies and ontologies
- General push for Linked Data technologies

EDP wants to answer

- Are publishers using Linked Data?
- Is the Linked Data they generate of enough quality to be queried and re-used?
- If not, how to improve?

For the SemWeb community

- Uptake and acceptance of our tools
  - Public Sector was an early adopter
- Identify challenges and areas of improvement

This paper

- Quantitative study of
  - Uptake of RDF as publishing format
  - Quality of Linked Datasets in EDP

Metrics - Uptake

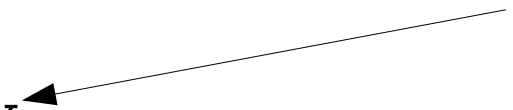
- Relative usage of RDF vs other formats
- Are they following recommendation when describing format in metadata?
  - How they deviate from it?

```
PREFIX
ex <https://datapublisher/#>
dcat < http://www.w3.org/ns/dcat#>
dct <http://purl.org/dc/terms/>


ex:anOpenDataset
        a dcat:Dataset ;
        dcat:distribution ex:aDistro .
ex:aDistro
        dct:format xxxxxxx ;
        dcat:MediaType xxxxxxxx
```

```
PREFIX
ex <https://datapublisher/#>
dcat < http://www.w3.org/ns/dcat#>
dct <http://purl.org/dc/terms/>
mrpo <http://publications.europa.eu/resource/authority/file-type/#>


ex:anOpenDataset
        a dcat:Dataset ;
        dcat:distribution Ex:aDistro .
ex:aDistro
        dct:format mrpo:xxx ;
```



What we want

```
SELECT ?format COUNT(?distribution) as ?numDistros
WHERE {
?distribution dct:format ?format .
}
GROUP BY ?format
```

Use of dct:Format

- 68% distributions include it (45% of datasets)
- From those including it
  - 45% OK
  - Varied errors like
    - Wrong suffix code (non-existent format)
    - A text literal

```
ex:aDistro
        dct:format ex:distro-UUID/format .
ex:distro-UUID/format
        a dcterms:IMT ;
        rdfs:label ‘PDF’ ;
        rdf:value ‘application/PDF’ .
```

Valid RDF but...

```
ex:aDistro
        dct:format ex:distro-UUID/format .
ex:distro-UUID/format
        a dcterms:IMT ;
        rdfs:label ‘PDF’ ;
        rdf:value ‘application/PDF’ .
```

Valid RDF but...

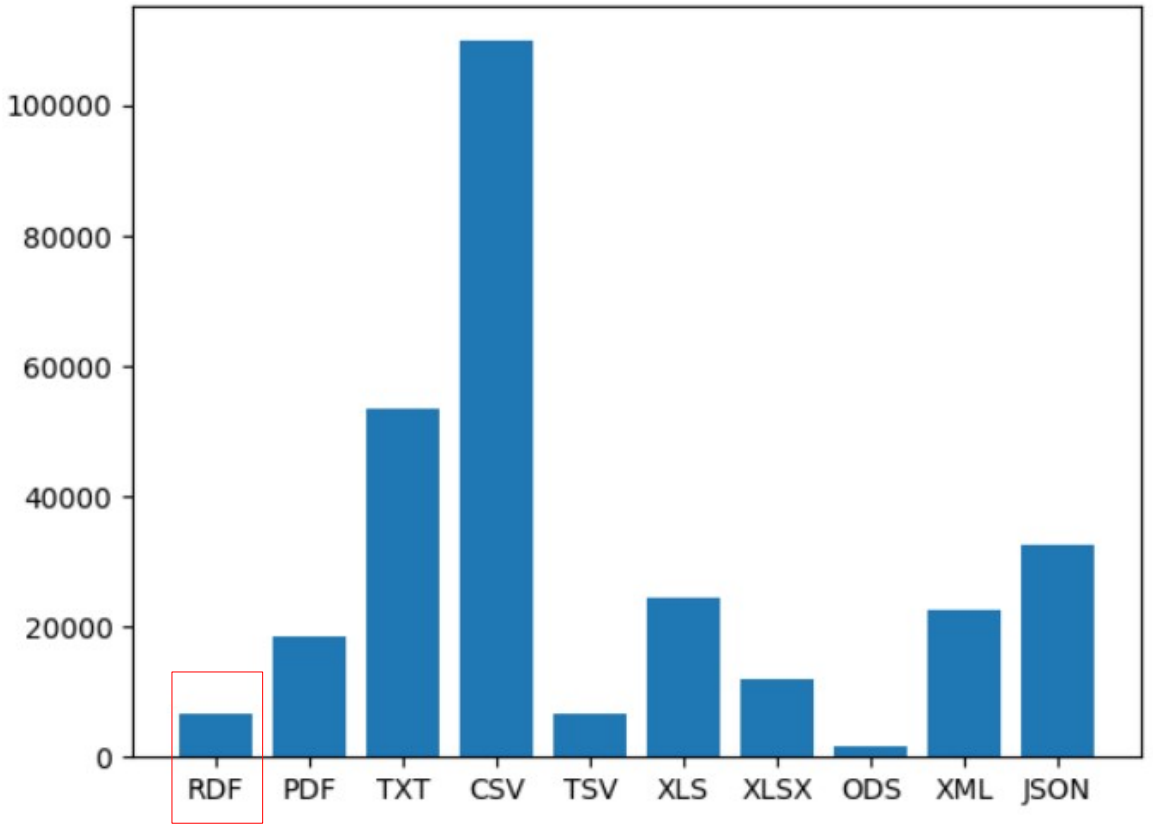
```
SELECT ?format COUNT(?distribution) as ?numDistros
WHERE {
?distribution dct:format ?format .
}
GROUP BY ?format
```

Won’t work...

ex:aDistro
dct:format b_node:123 .
b_node:123
rdfs:label ‘PDF’ ;

Same problem

Results - Uptake



Implications - Uptake

- Better procedures for interactive metadata required
  - EDP implemented SHACL rules and reports
- Massive number of tabular datasets
  - Help with transformation needed

Corpus

- RDF Datasets harvested by EDP
  - Could be downloaded and parsed valid.
- Main difference with previous work:
  - Not crawled, but what the EDP harvests

Quality Metrics

- Subset from previous general LOD quality studies [1,2,3]
  - Contextual
  - Representational
  - Accessibility

[1] Hogan, A. et al.: An empirical survey of Linked Data conformance. JWS 14-44 2012
[2] Debattista, J. et al.: Evaluating the quality of the LOD cloud: An empirical investigation. Semantic Web 9(6) 2018.
[3] Schmachtenberg, M. et al.: Adoption of the Linked Data Best Practices in Different Topical Domains. ISWC 2014.

Metrics – Contextual

- Provision of Provenance information
  - Analysis of dct:publisher usage

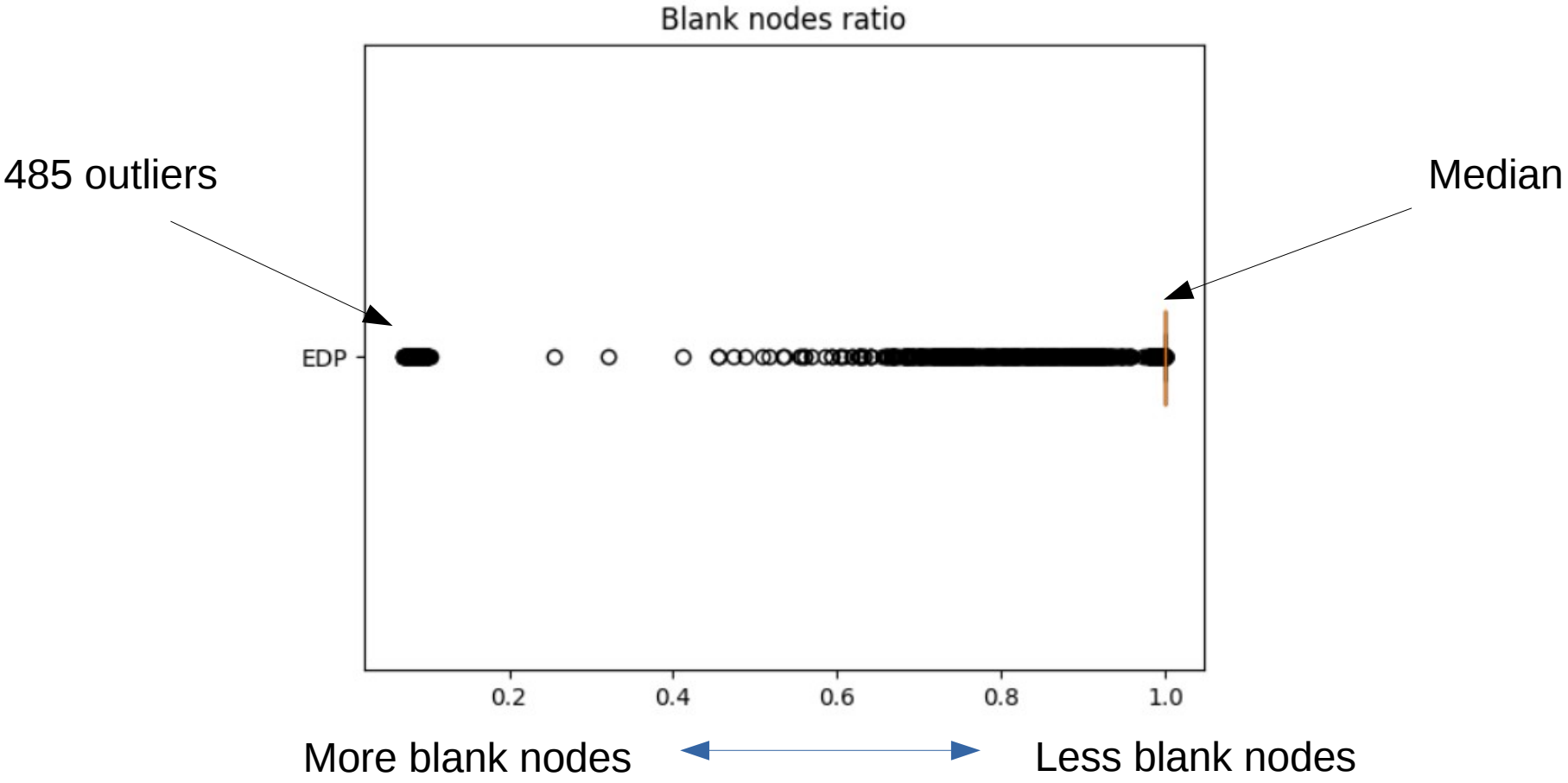
Results - Contextual

- 50.3% of all datasets have dct:publisher
- 42.8% of all RDF have dct:publisher
- Relatively low
  - But much than other studies for general LOD (16%)

Metrics – Representational

- Blank nodes
- Well-known vocabularies
- Proprietary vocabularies

Results – Blank node usage



Results – Well known vocabularies

- Lots of rdf, rdfs and dcterms
  - Mostly predicates
- Few foaf, wgs84
- Almost no rss, skos, bio
  - Different from general LOD Cloud

Results – Unknown vocabularies

Table 5. Top-10 not well-known vocabularies by dataset percentage

Vocabulary	% Datasets	# Hosts	# Preds	Deref-able?
socrata.com/rdf/terms	52.6%	4	1	No
opendata.aragon.es/def/Aragopedia	13.0%	1	52	No
w3.org/2000/10/swap/pim/usps#	2.7%	4	4	Yes
data.press.net/ontology/stuff/	2.1%	2	5	Yes
opendata.caceres.es/def/ontomunicipio	1.7%	2	139	HTML
purl.org/ctic/infraestructuras/	1.1%	1	5	No
opendata.unex.es/def/ontouniversidad	1.0%	1	63	HTML
dublincore.org/documents/dcmi-box/	0.7%	1	4	No
open.vocab.org/terms	0.6%	1	3	HTML
server1.avantic.net/opendata/vocab/raw/	0.5%	1	206	No

Realisation

- Two provinces of same country, different ontologies
  - And we can’t blame them, this two really embraced Linked Data!
- No apparent alignment among themselves or among (at least that has reached EDP)

Results – Unknown vocabularies

Table 5. Top-10 not well-known vocabularies by dataset percentage

Vocabulary	% Datasets	# Hosts	# Preds	Deref-able?
socrata.com/rdf/terms	52.6%	4	1	No
opendata.aragon.es/def/Aragopedia	13.0%	1	52	No
w3.org/2000/10/swap/pim/usps#	2.7%	4	4	Yes
data.press.net/ontology/stuff/	2.1%	2	5	Yes
opendata.caceres.es/def/ontomunicipio	1.7%	2	139	HTML
purl.org/ctic/infraestructuras/	1.1%	1	5	No
opendata.unex.es/def/ontouniversidad	1.0%	1	63	HTML
dublincore.org/documents/dcmi-box/	0.7%	1	4	No
open.vocab.org/terms	0.6%	1	3	HTML
server1.avantic.net/opendata/vocab/raw/	0.5%	1	206	No

Realisation

- A lot of RDF in our corpus is produced from CSV through portal provider “converters”
- Conversion generates many unique URIs
  - One namespace per csv
  - Linking nightmare

Metrics – Accessibility

- Links to external datasets
  - Counted Pay Level Domains other than the publisher in the datasets

Results - Links

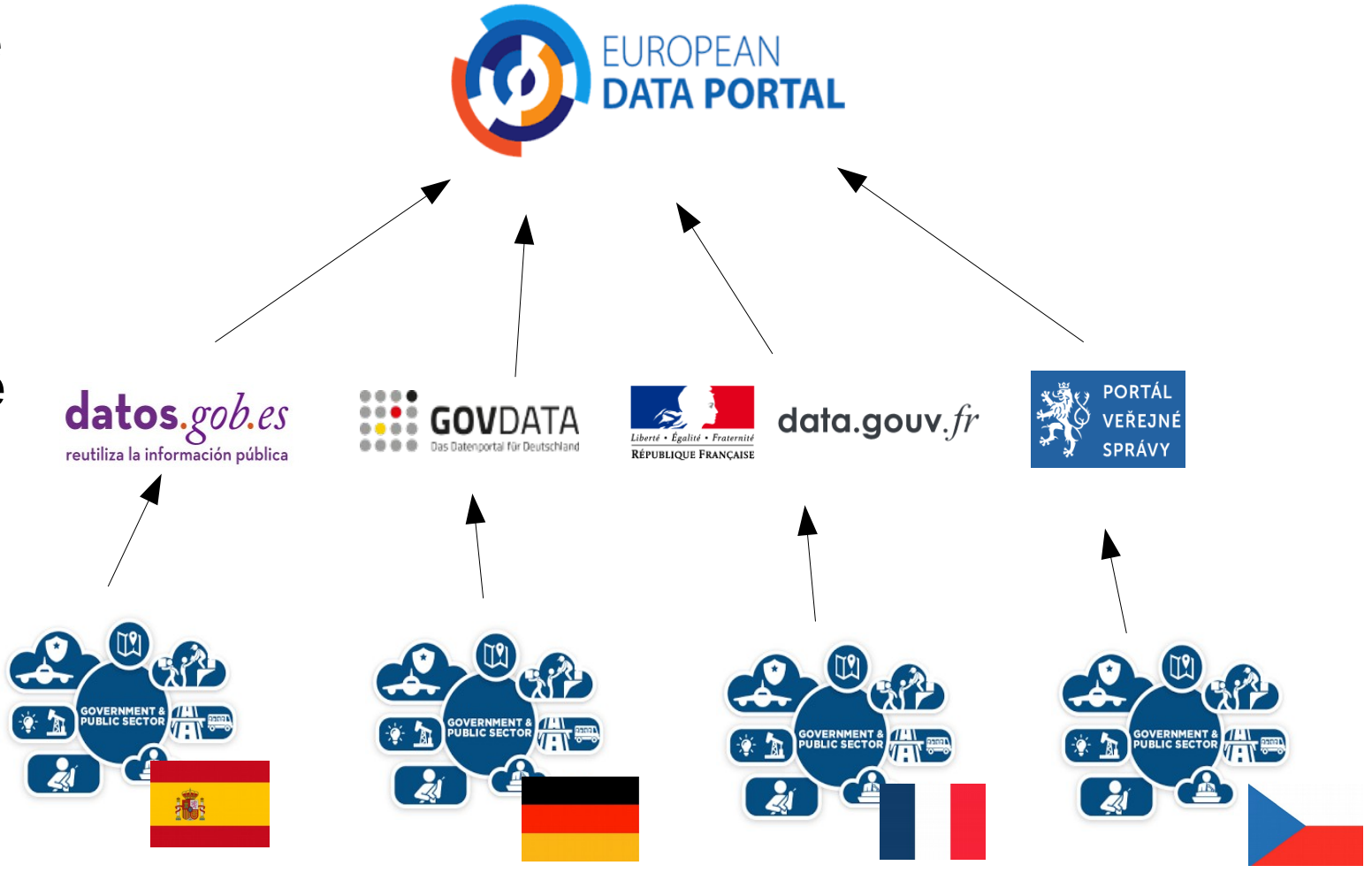
Domain	# Datasets	# (%) hosts
w3.org	2467	55 (74%)
es.dbpedia	769	4 (5.4%)
purl.org	577	35 (47.3%)
reference.data.gov.uk	504	12 (16.2%)
data.press.net	123	2 (2.7%)
murciaturistica.es	122	1 (1.35%)
geonames.org	119	4 (5.4%)
www.gijon.es	117	1 (1.35%)
schema.org	73	7 (9.5%)
dbpedia.org	43	11 (14.8%)
publications.europa.eu	4	4 (5.4%)

Summary

- Metadata is on good track
  - SHACL plus good comms/UI should do the trick
  - Suggestions after SHACL failures a need
- Data is not very good quality
  - Who takes responsibility of processing for linking?

Who runs the interlinking process?

Who maintains the mapping?



Summary

- Technology readiness gap?
  - Tools not yet there in full force
  - Or no one knows how to configure them
- Organisational challenge
  - Harder than a very large company

Next steps - Technical

- Keen to get our hands on the final versions of the Tabular to KG challenge contestants.
- Assess linking run by EDP

Next steps - Social

- Put data consumers in the loop
  - Validate with consumers
- Social coding == Social data-ing?
  - Patching and patch approval
  - Share re-uses and transformations?
  - Spam is a concern

Thank you!

l.d.ibanez@soton.ac.uk

www.europeandataportal.eu