



Linked Data and the Potential for Cultural Heritage

**Workshop:
European Cultural Heritage Online
Aggregation and Semantic Web
Stockholm, May 2012**

Gordon McKenna

**International Development Manager
Collections Trust, UK**

Collections
Trust

**Context –
The Linked Heritage Project**

<http://www.linkedheritage.org>

WP 2 – Selected Overview

Objectives:

- To explore the **state of the art in linked data**;
- To identify **appropriate models, processes and technologies** for the **deployment of linked data**;
- To consider how **linked data practices** can be applied to **cultural heritage**;
- To explore the **state of the art in persistent identifiers**.

Tasks and Deliverables:

- T2.1 – *Exploring cultural heritage information best practice*
 - D2.1 – ***Best practice report on cultural heritage linked data and metadata standards***
- T2.2 – *Resource identification [PIDs]*
 - D2.2 – ***State of the art report on persistent identifier standards and management tools***

1. Carry out **research** – What exists, survey
2. Make an **analysis** – Look for patterns and trends.
3. Give **simple advice** – practical and implementable
4. Reuse or create **tools** – Easy to use, audience relevant, adaptable open licence (e.g. Multilingual versions possible)
5. Identify **further needs** – Leading to further work

Partner Survey

Survey Method and Structure

- Aimed at **partners in Linked Heritage**
- Data collection – Online **SurveyMonkey** (supported by a RTF document)
- Sections:
 1. Participant information
 2. Metadata standards and use
 3. **Linked data use** and Europeana agreement



Familiar with the Linked Data Concept?

- Yes: 74.4%
- No: 25.6%

Used Linked Data?

- Yes: 15.4%
- No: 84.6%
- **Details:**
 - 4 – Dbpedia;
 - 3 – GeoNames;
 - 1 – Freebase;
 - 1 – IPTC;
 - 1 – SKOS;
 - 1 – [in-house];

Published Linked Data?

- Yes: 10.3%
- No: 89.7%
- **Details:**
 - <http://data.kunstkamera.ru/sparql;>
 - <http://data.kunstkamera.ru>
 - http://nektar.oszk.hu/wiki/Semantic_web
 - Thesaurus in SKOS



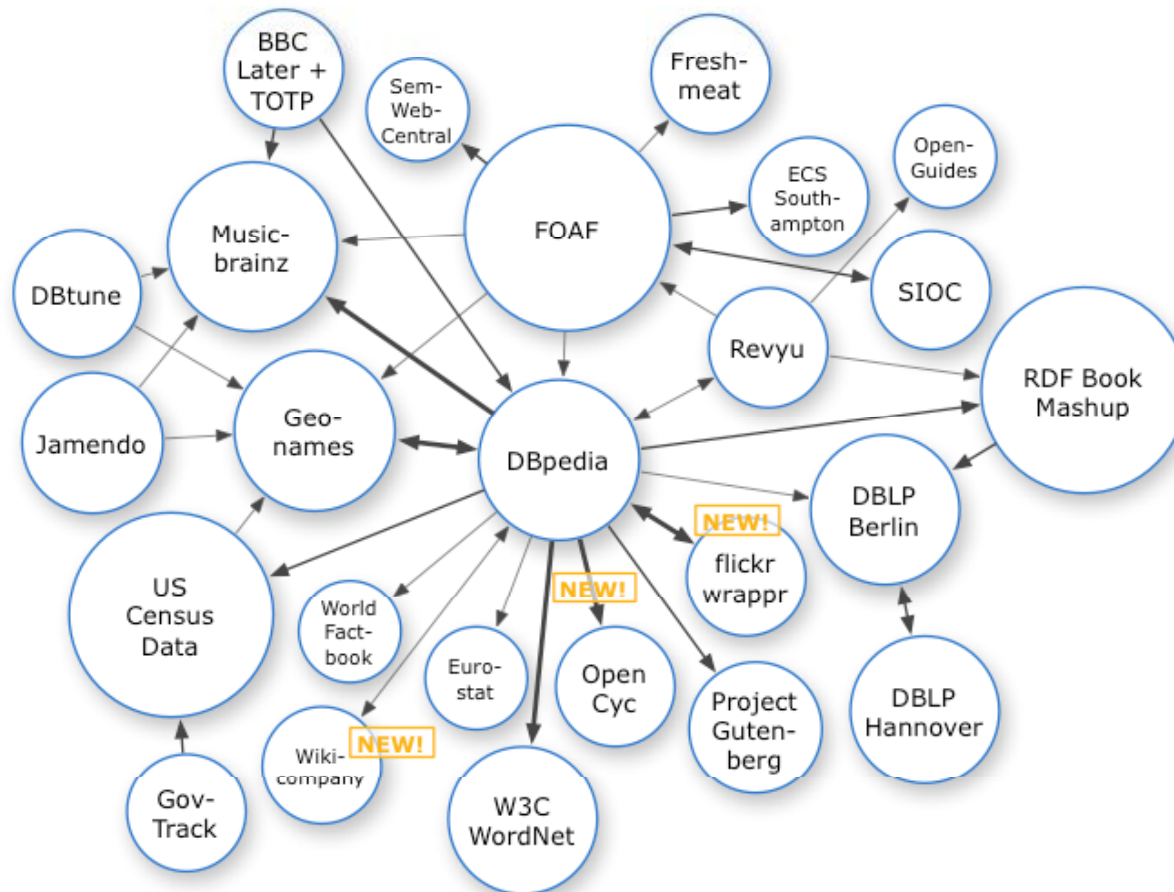
Know of Linked Data Projects?

- Yes: 38.5%
- No: 65.5%
- **Activity in:**
 - France
 - Germany
 - Israel
 - Italy
 - Russia
 - Spain
 - Sweden
 - United Kingdom

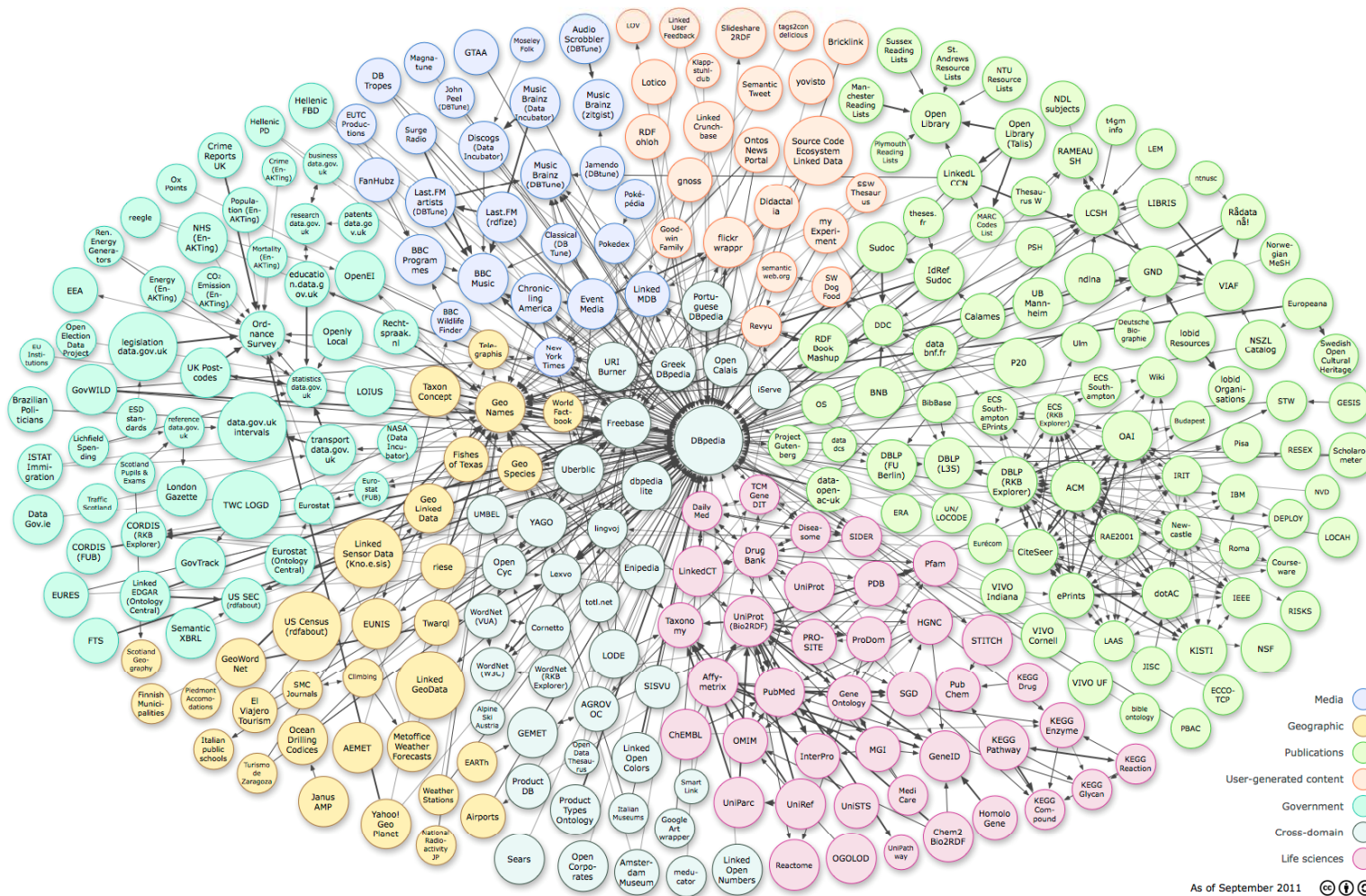
A market for basic information and guidance!

Research into the
Linked Open Data Cloud

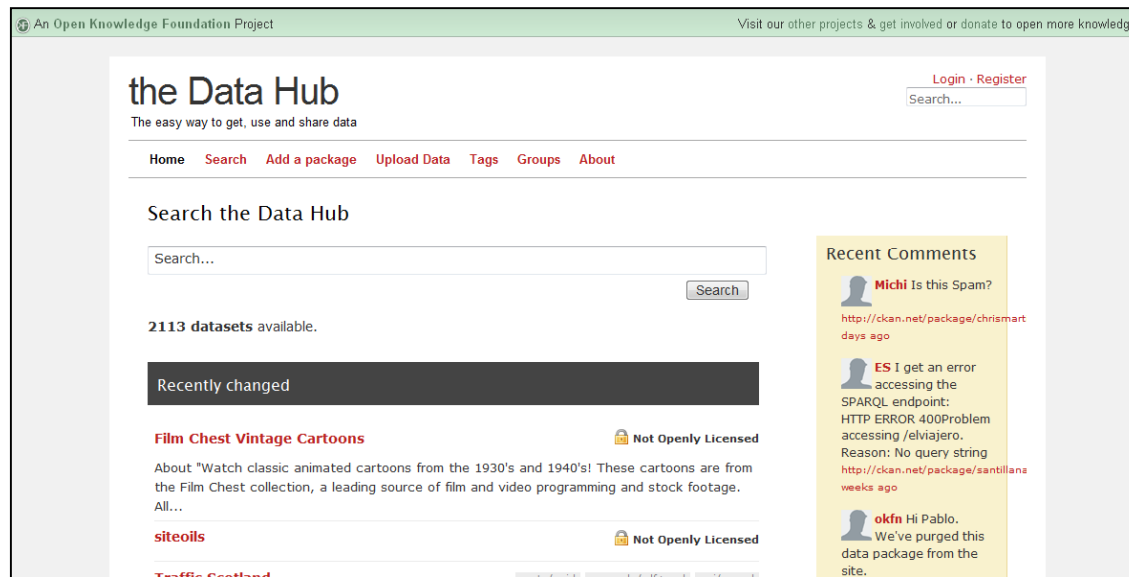
Linking Open Data Cloud - 2007



<http://linkeddata.org>



- Part of **CKAN** – **C**omprehensive **K**nowledge **A**rchive **N**etwork)
- Registry of open [and not open] knowledge
 - **Packages** [The circles in the Cloud]
 - **Projects** (and a few closed ones).



<http://thedatahub.org>

Is the LOD Cloud Open?

‘Open’ = commercial use

311 packages:

- Yes 42.6%
- No **57.4%**

c38 billion triples:

- Yes 30.9%
- No **69.1%**

Open Licences Used

	<i>Packages</i>	<i>Triples</i>
• CC BY	28.8%	45.8%
• CC BY-SA	18.2%	10.2%
• PDDL	10.6%	0.2%
• CC0	9.1%	2.9%
• UK Crown Copyright with data.gov.uk rights	7.6%	27.4%
• Other (Public Domain)	6.8%	7.0%
• Other (Open)	5.3%	5.0%
• Other (Attribution)	3.0%	0.4%
• UK Open Government Licence (OGL)	3.0%	0.1%
• GNU FDL	3.0%	<0.1%
• ODbL	2.3%	0.9%
• GNU GPL	0.8%	<0.1%
• New and Simplified BSD licences	0.8%	0.1%

Not Open Licences Used (or Not)

	<i>Packages</i>	<i>Triples</i>
• [not given]	69.1%	89.4%
• None	14.6%	0.3%
• CC BY-NC	7.3%	5.8%
• Other (Not Open)	6.7%	<0.1%
• Other (Non-Commercial)	0.6%	3.9%
• CC BY-SA	0.6%	<0.1%

Top Ten Open Packages (Triples)

1. LinkedGeoData	3.00 billion
2. UK Legislation	1.90 billion
3. Linked Sensor Data (Kno.e.sis)	1.73 billion
4. data.gov.uk Time Intervals	1.00 billion
5. DBpedia	1.00 billion
6. Open Library data mirror in the Talis Platform	0.54 billion
7. The Open Library	0.40 billion
8. Freebase	0.34 billion
9. transport.data.gov.uk	0.33 billion
10. Data Incubator: MusicBrainz	0.18 billion



Top Ten Not Open Packages (Triples)

1. TWC: Linking Open Government Data	9.80 billion
2. Data.gov	6.40 billion
3. Source Code Ecosystem Linked Data	1.50 billion
4. 2000 U.S. Census in RDF (rdfabout.com)	1.00 billion
5. PubMed	0.80 billion
6. DBTune.org MySpace RDF Service	0.66 billion
7. UniParc	0.63 billion
8. DBTune.org AudioScrobbler RDF Service	0.60 billion
9. Linking Italian University Statistics Project	0.59 billion
10. UniProt UniRef	0.49 billion

Top Packages Linked To By Packages

	<i>Packages</i>	<i>Links (million)</i>
1. DBpedia	158	31.53
2. GeoNames Semantic Web	38	9.35
3. [none]	34	0
4. DBLP Computer Science Bibliography (RKBExplorer)	27	1.34
5. Association for Computing Machinery (ACM) (RKBExplorer)	26	1.49
6. ePrints3 Institutional Archive Collection (RKBExplorer)	26	0.28
7. Freebase	25	10.45
8. CiteSeer (Research Index) (RKBExplorer)	24	0.80
9. School of Electronics and Computer Science, University of Southampton (RKBExplorer)	24	0.04
10. ReSIST Project Wiki (RKBExplorer)	24	<0.01 [408]

Top Packages Linked To By Links

	<i>Packages</i>	<i>Links (million)</i>
1. UniProtKB Taxonomy	6	46.63
2. MARC Codes List	3	42.41
3. QDOS	1	40.00
4. UniProtKB	10	31.14
5. DBpedia	158	31.53
6. Ordnance Survey Linked Data [UK]	16	29.72
7. UniParc	1	27.53
8. IdRef: Sudoc authority data	3	20.04
9. Sudoc bibliographic data	1	20.00
10. flickr™ wrappr	4	16.36

Cultural Packages in the Cloud

	<i>Triples (million)</i>
• VIAF: The Virtual International Authority File	200.0
• Europeana Linked Open Data	185.0
• British National Bibliography (BNB)	80.2
• Hungarian National Library (NSZL) catalog	19.3
• Amsterdam Museum as Linked Open Data in the Europeana Data Model	5.0
• Library of Congress Subject Headings	4.2
• Swedish Open Cultural Heritage	3.4
• Calames	2.0
• RAMEAU subject headings (STITCH)	1.6
• data.bnf.fr - Bibliothèque nationale de France	1.4
• National Diet Library of Japan subject headings	1.3
• Gemeenschappelijke Thesaurus Audiovisuele Archieven	1.0
• Gemeinsame Normdatei (GND)	0.6
• Archives Hub Linked Data	0.4
• Thesaurus for Graphic Materials (t4gm.info)	0.1
• Italian Museums (LinkedOpenData.it)	<0.1
• Thesaurus W for Local Archives	<0.1
• MARC Codes List Open Data	<0.1

Cultural Heritage – Licences Used

Open licences

	<i>Number</i>
• CC0	2
• Other (Public Domain)	1
• Other (Open)	1
• ODbL	1

Not open licences

	<i>Number</i>
• [not given]	9
• CC BY-SA	3
• Other (non-commercial)	1



Cultural Heritage – Formats Used

	<i>Packages</i>
• Simple Knowledge Organization System	11
• Dublin Core	7
• Friend of a Friend	3
• Basic Geo	1
• Bibliographic Ontology	1
• DBpedia	1
• Music Ontology	1
• Object Reuse and Exchange	1
• vCard	1
• XML Schema	1

Cultural Heritage – Links

	<i>Packages</i>	<i>Links</i>
1. DBpedia	5	82,308
2. Library of Congress Subject Headings	4	108,135
3. VIAF: The Virtual International Authority File	2	1,820,684
4. GeoNames Semantic Web	2	510,658
5. Dewey Decimal Classification (DDC)	2	200,543
6. RAMEAU subject headings (STITCH)	2	83,530
7. Swedish Open Cultural Heritage	1	100,489
8. Gemeinsame Normdatei (GND)	1	20,000
9. IdRef: Sudoc authority data	1	10,000
10. [DCMI Type Vocabulary – not in The Cloud]	1	10,000
11. UK Postcodes	1	5,000
12. AGROVOC	1	700
13. Hungarian National Library (NSZL) catalog	1	136
14. [none]	1	0

Open Data – **Licensing?**

- **Must have one** ⇒ Before publishing make a decision?
- What kind of licence **can you give** (CC useable?)?
- What kind of **3rd party use** do you want to allow?

Linkable Data – **Publishing?**

- Use **Persistent Identifiers**;
- Select '**standard**' data formats;
- Carefully **choose** what you are publishing

Linking Data – **Which package(s)** do you link to?

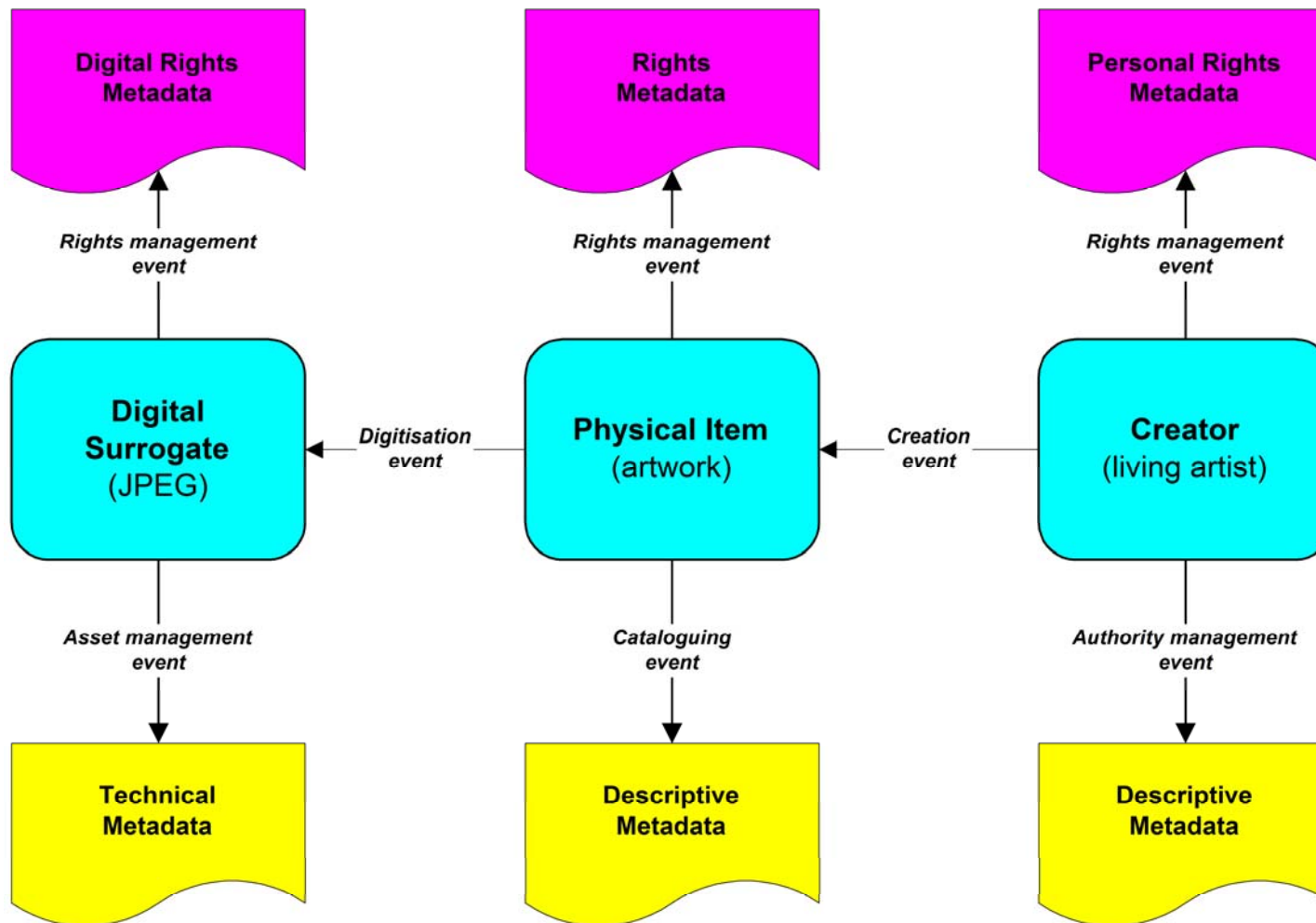
- **Trusted** source?
- Presence of **PIDs** and **maintained** resource?

Linked Culture Cloud – **shared resource?**

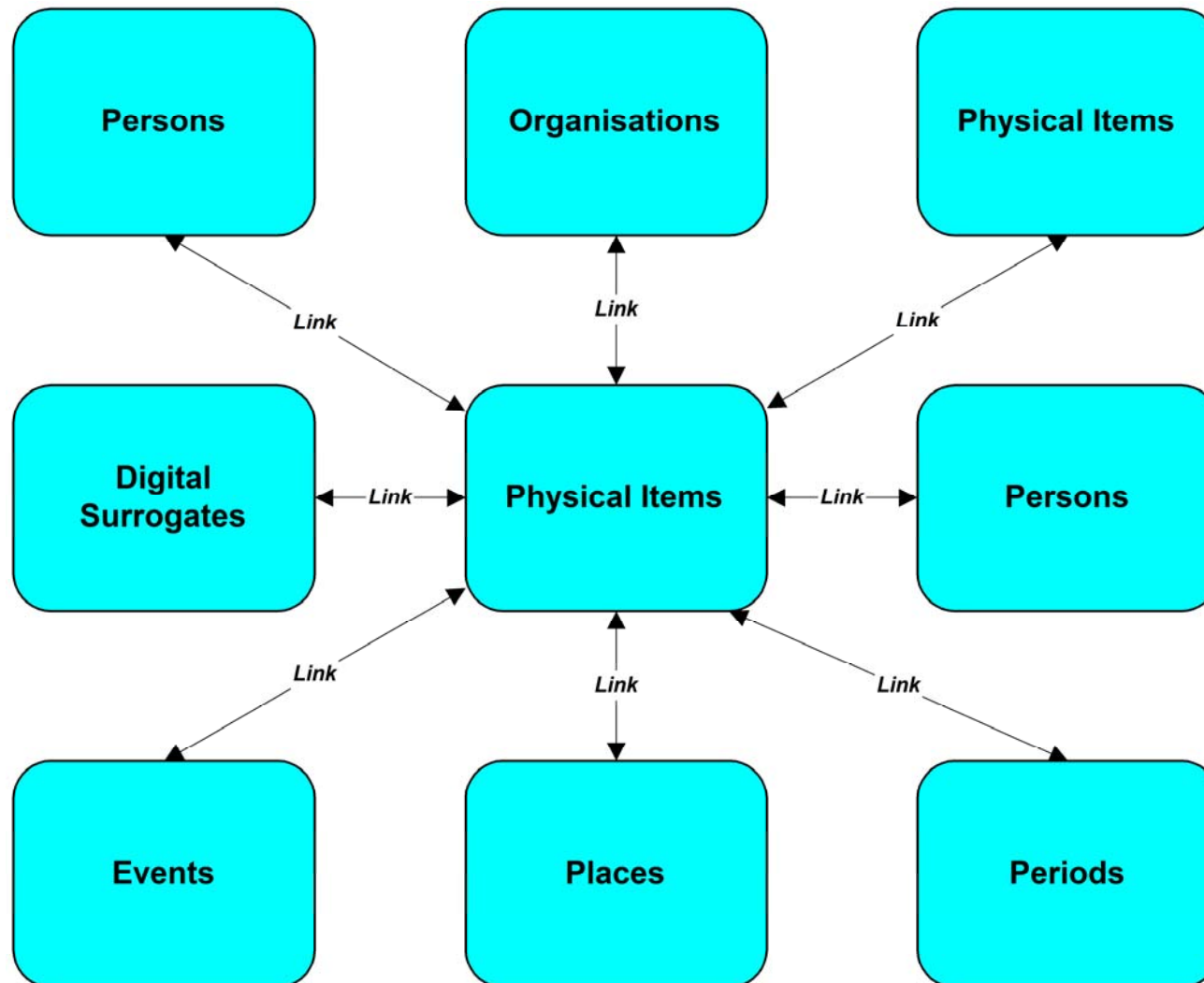
- **Sub-set** of the LOD Cloud / CKAN;
- Information **relevant** for cultural institutions
- **Feed into** general LOD Cloud

Persistent Identifiers?
It is the Links!

Metadata – A ‘Simple’ Example



Links – A ‘Simple’ Example (Honest!)



So it is easy?

What the User Sees:

The Reality:



The Semantic Web

The Work

Linked Data Principles

Tim Berners-Lee 2007 –

<http://www.w3.org/DesignIssues/LinkedData.html>

1. Use **URIs as names** for things;
2. Use **HTTP URIs** so that **people can look up** those names;
3. When someone looks up a URI, provide **useful RDF** information;
4. Include RDF statements that **link to other URIs** so that they can discover related things.

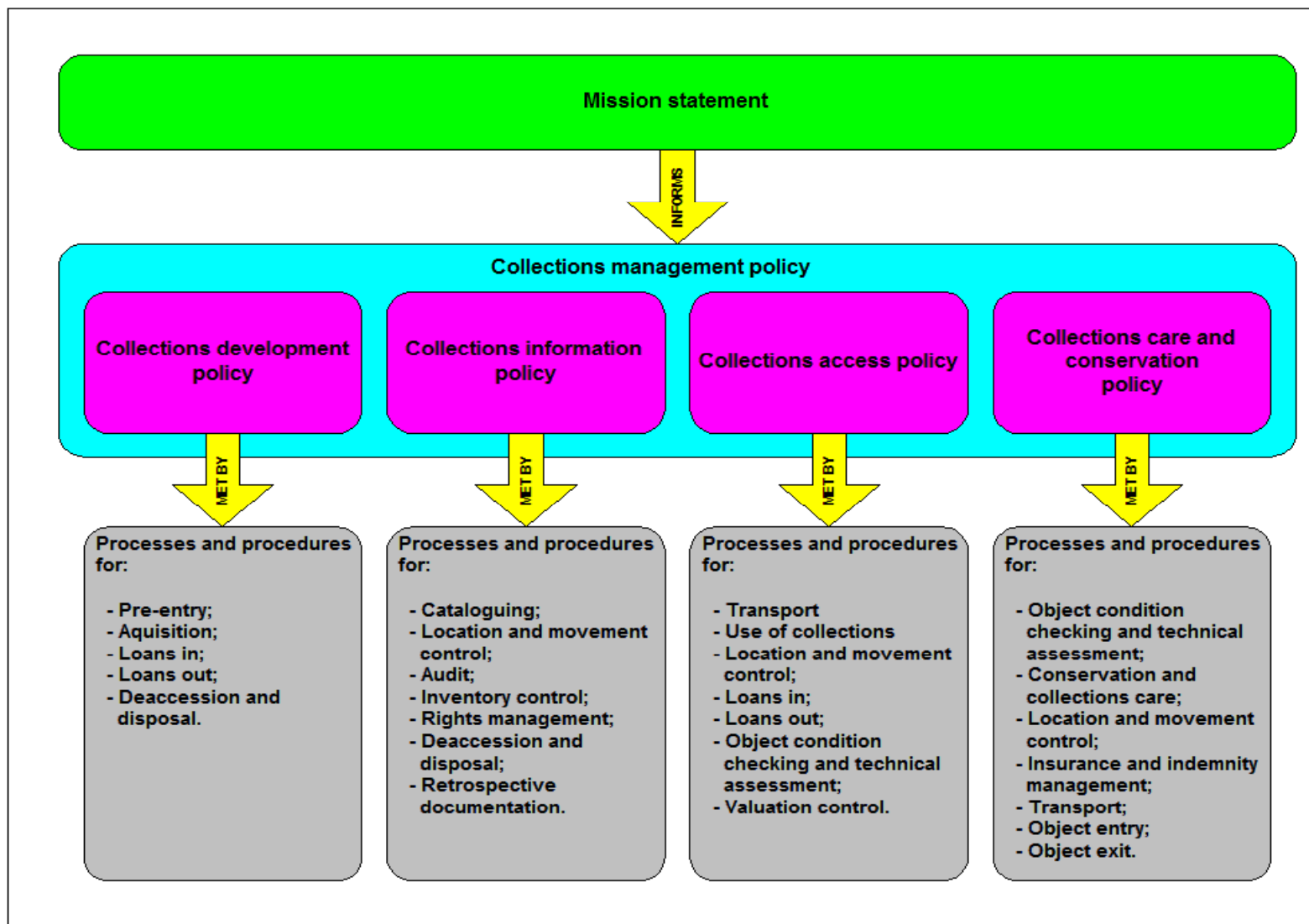
ATHENA Survey on PIDs

- PIDS were used in **75%** of the countries surveyed;
- Organisations are **aware of the basic reasons** for using PIDs;
- PIDs may have a **restricted use** in an organisation;
- **URIs** in 75% of the countries surveyed
- PID systems (ARKs, PURLs, DOIs) much **less used**, often **not maintained**, **project** related
- Organisations asked for **more information**

Advice – PIDs in Organisations

- **Uniqueness environment** – How unique;
- **Persistent** – How persistent, how maintained;
- **Resolvable** – Do PIDs lead to something;
- **Cost effective** – Use free or very low cost systems;
- **Supported by policy** – PIDs part of written policy;
- **Managed** – Embedded processes and procedures.

Policy Framework – PAS 197





Thank you!

gordon@collectionstrust.org.uk