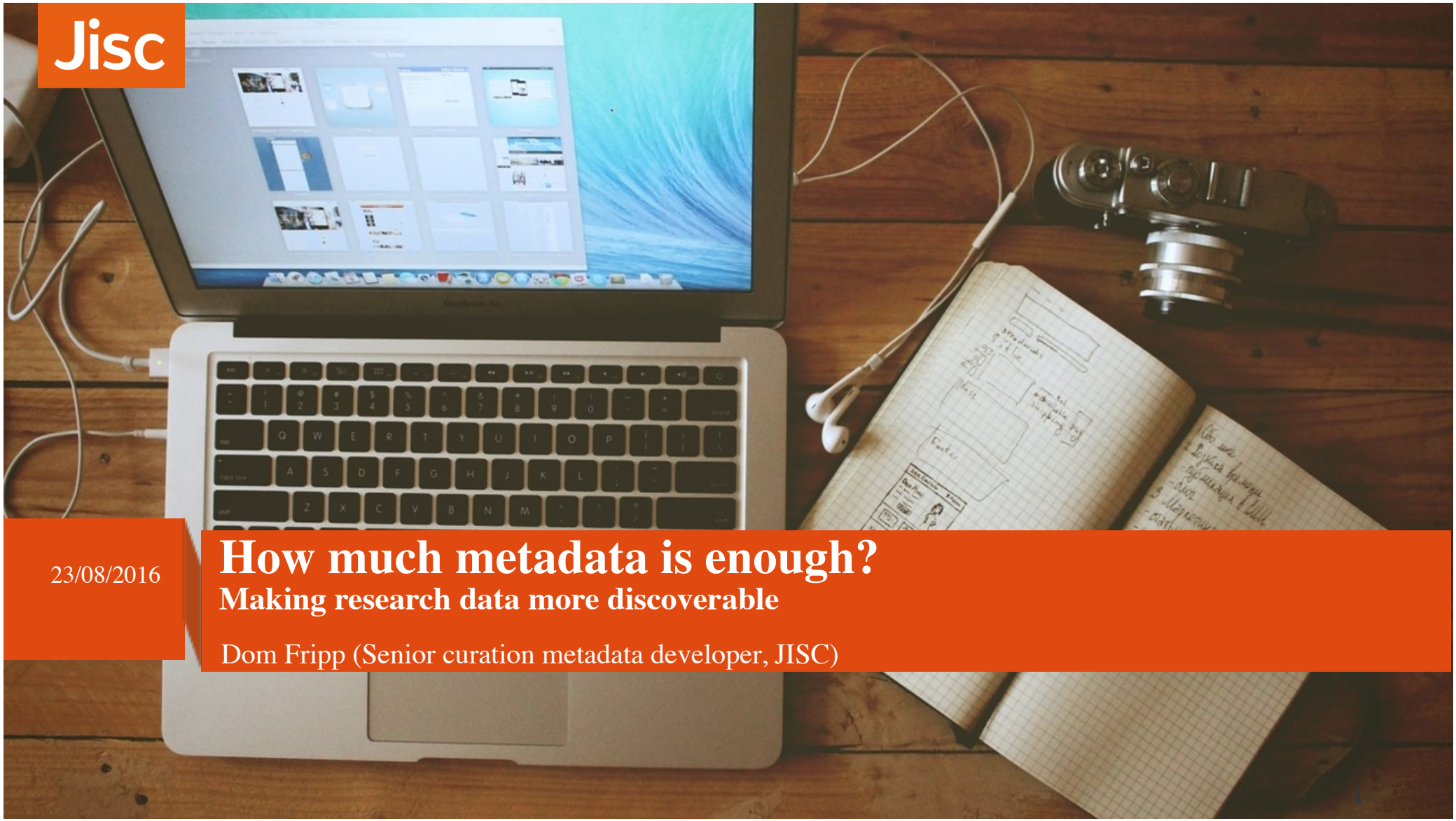


23/08/2016

How much metadata is enough?

Making research data more discoverable

Dom Fripp (Senior curation metadata developer, JISC)

About me

@Domicus

#JiscRDM

dom.fripp@jisc.ac.uk

- MSc. Information and Library Science 2010
- Research Data Management 2013
- Jisc project work since January 2016
  - UK Research Data Discovery Service
  - Research Data Shared Service

Jisc

What we'll cover

- Why does it matter?
- Raiders of the FAIR principles
- The UK Research Data Discovery Service
- Developing a core metadata profile
  - How was it created?
  - What does it do?
- Current challenges and opportunities
- Summary

Why does it matter?

- **Mandated by funders**
  - Discoverable
  - Defined access period
  - Use of identifiers
- **Research Impact**
  - Citable/usable data(sets)
- **Reproducibility**
  - Emulation
  - Environmental capture
- **Applicable to arts and digital humanities**
  - Scans / images
  - Text mining
  - Performance
  - Multimedia
  - Consent
  - Physical objects

Raiders of the FAIR principles

Copyright Disclaimer Under Section 107 of the Copyright Act 1976, allowance is made for "fair use" for purposes such as criticism, comment, news reporting, teaching, scholarship, and research. Fair use is a use permitted by copyright statute that might otherwise be infringing. Non-profit, educational or personal use tips the balance in favor of fair use.


“Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process.”
Wilkinson et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship SCIENTIFIC DATA | 3:160018 | DOI: 10.1038/sdata.2016.18


Findable

Jisc

Findable

- 1. To be **Findable** any Data Object should be uniquely and persistently identifiable
  1.1. The same Data Object should be re-findable at any point in time, thus Data Objects should be **persistent**, with emphasis on their metadata
  1.2. A Data Object should minimally contain basic machine actionable metadata that allows it to be distinguished from other Data Objects
  1.3. Identifiers for any concept used in Data Objects should therefore be **Unique** and **Persistent**


Accessible

Jisc

Accessible

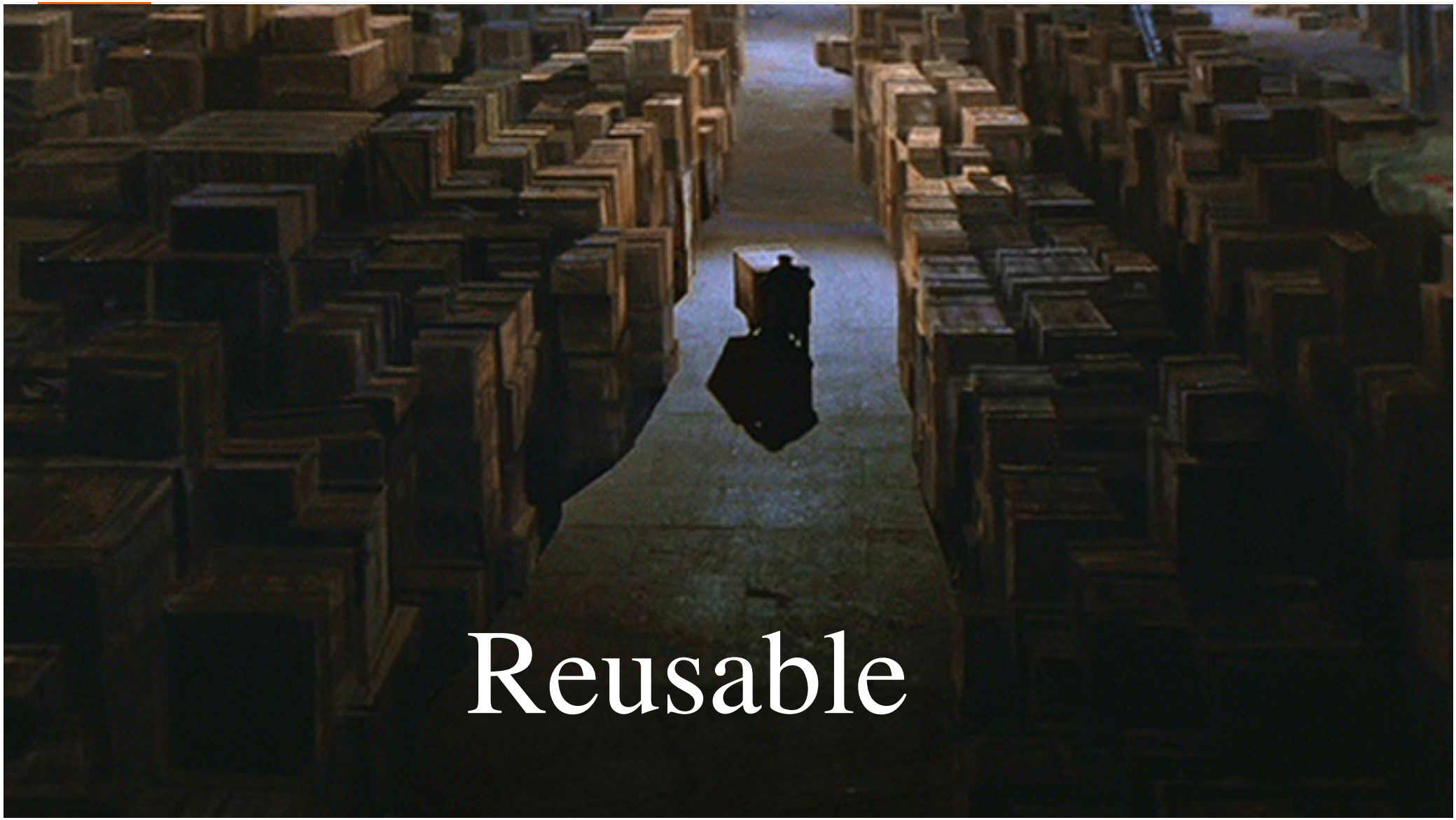
- 2. Data is **Accessible** in that it can be always obtained by machines and humans
  2.1 Upon appropriate authorization
  2.2 Through a well-defined protocol
  2.3 Thus, machines and humans alike will be able to judge the actual accessibility of each Data Object.


Interoperable

Jisc

Interoperability

- 3. Data Objects can be **Interoperable** only if:
  3.1. (Meta) data is machine-actionable
  3.2. (Meta) data formats utilize shared vocabularies and/or ontologies
  3.3 (Meta) data within the Data Object should thus be both syntactically parseable and semantically machine-accessible


Reusable

Jisc

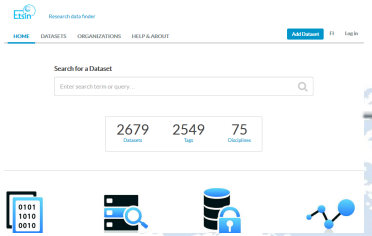
Reusable

- 4. For Data Objects to be **Re-usable** additional criteria are:
  4.1 Data Objects should be compliant with **principles 1-3**
  4.2 (Meta) data should be sufficiently well-described and rich that it can be automatically (or with minimal human effort) linked or integrated, like-with-like, with other data sources.
  4.3 Published Data Objects should refer to their sources with rich enough metadata and provenance to enable proper citation.

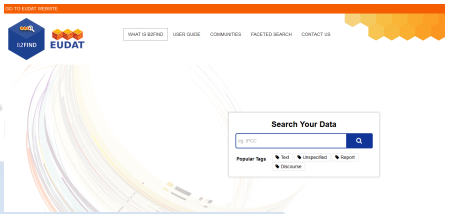
UK Research Data Discovery Service

Developing a core metadata profile

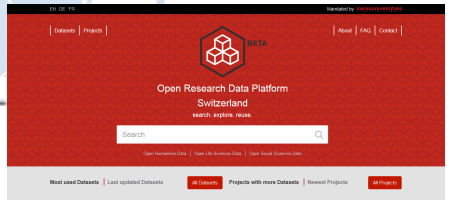
Global landscape



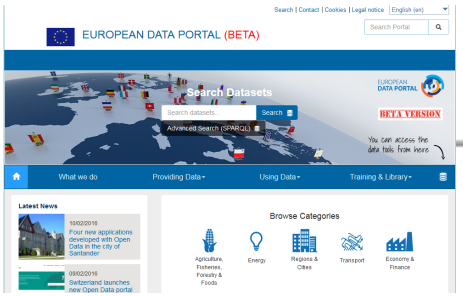
http://etsin.avointiede.fi



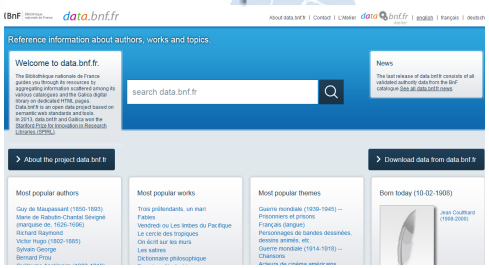
http://b2find.eudat.eu/



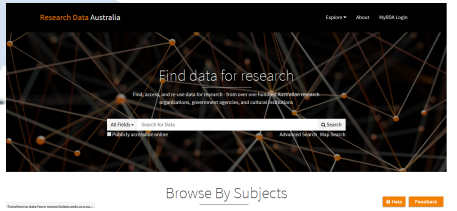
http://www.openresearchdata.ch/



http://www.europeandataportal.eu/



http://data.bnf.fr/



https://researchdata.ands.org.au/

Participating organisations

Pilots

- University of Hull
- University of St Andrews
- University of Glasgow
- Oxford Brookes University
- University of Edinburgh
- University of Oxford
- University of

Data Centres

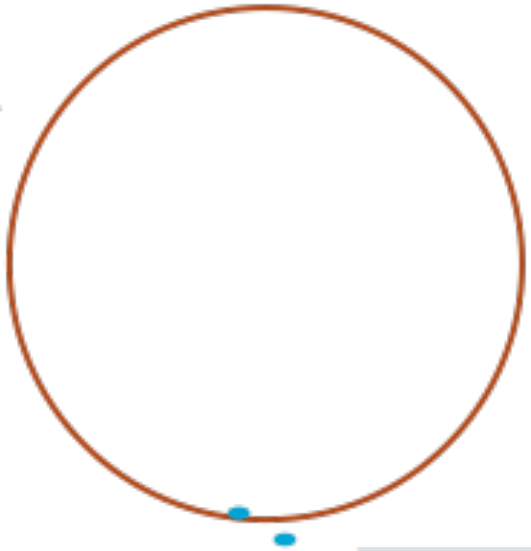
- Archaeology Data Centre
- Cambridge Crystallographic Data Centre
- ISIS/ICAT - STFC
- UK Data Service
- Visual Arts Data Centre
- NERC

Who's it for? (user stories)

» MoSCoW prioritisation

Researcher
- » Discover datasets
- » Discover related objects / resources
- » Find data across disciplines by location
- » Find exemplar data to inspire my research
- » Targeted search for topical data
- » Visual search for data
- » Find linked open data
- » Understand metadata quality
- » Understand data quality
- » Show research impact

Data repository
- » Show repository impact
- » Metadata rights respected
- » Show licence and rights of data
- » Index to external services
- » Force refresh of registry content



Project / research manager
- » Reporting to funders
- » Find research outputs of my institution

Machine
- » Harvestable registry
- » Show relationships between resources

System manager
- » No duplicate records
- » Harvest datasets
- » Update platform software

Funder
- » Return on investment

http://rdds.jiscinvolve.org/wp/2015/05/08/initial-workshop/

Jisc

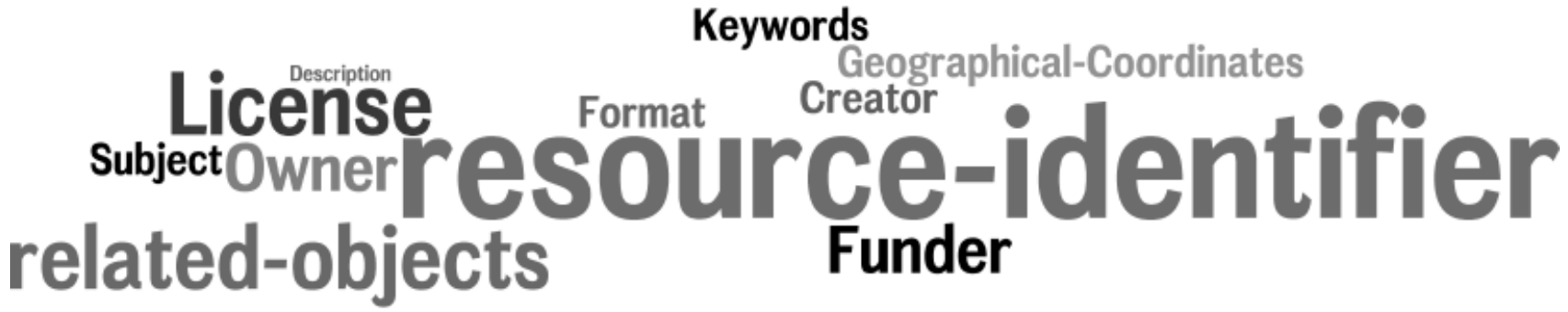
Metadata aims

To produce a draft of a metadata profile that fulfils the following criteria:

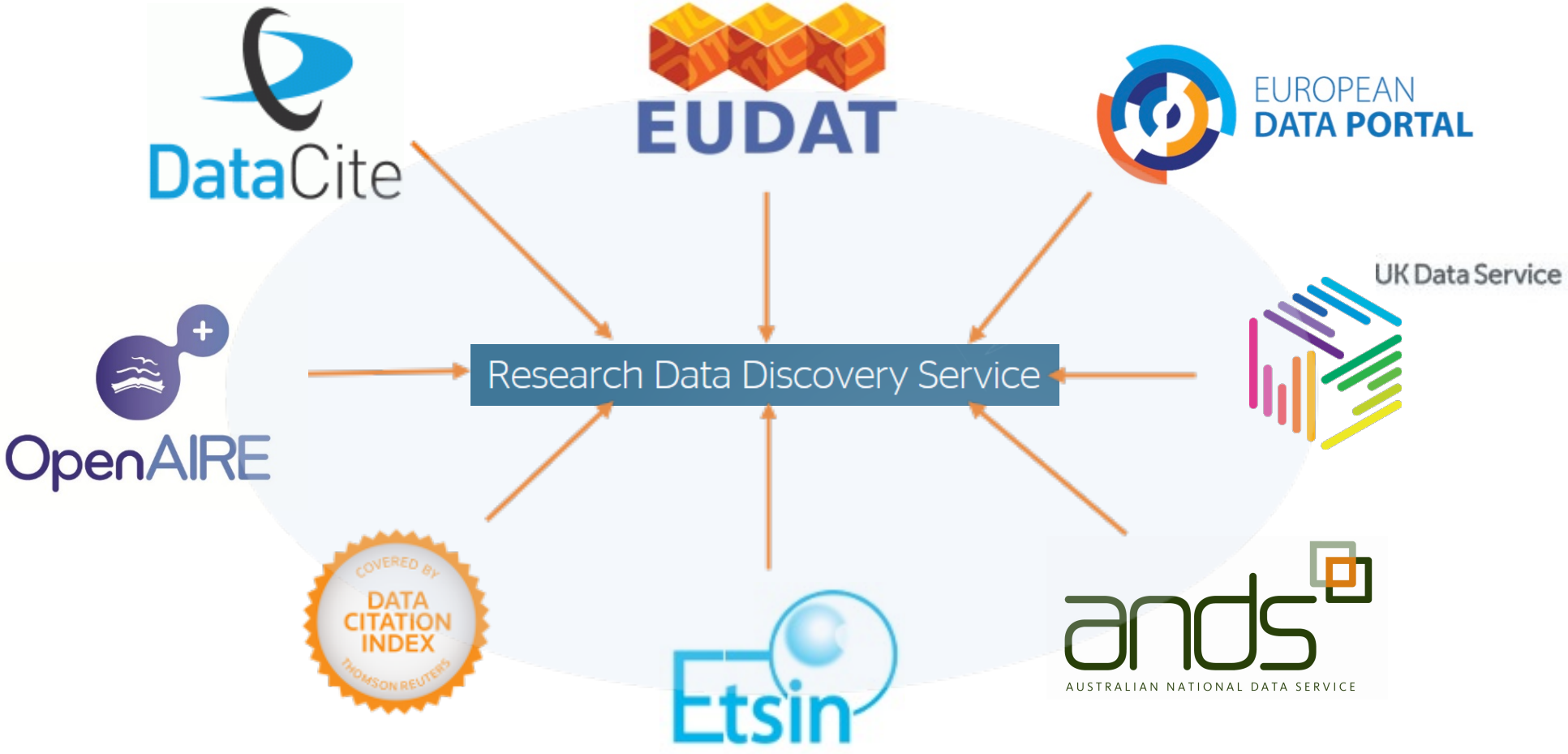
- Meet user requirements (based on the evidence provided by user cases and correlation with common fields in use across the research data domain).
- Simple enough (in core form) to map onto the CKAN instance that underpins the portal.
- Based on existing schemas (and good practice) in use across the research data domain.
- Flexible enough to develop along with the service and user needs.

Fields matching user requirements

Admin Harvest / Ingest Browse Search Reporting Usability Policy



Existing services & schemas

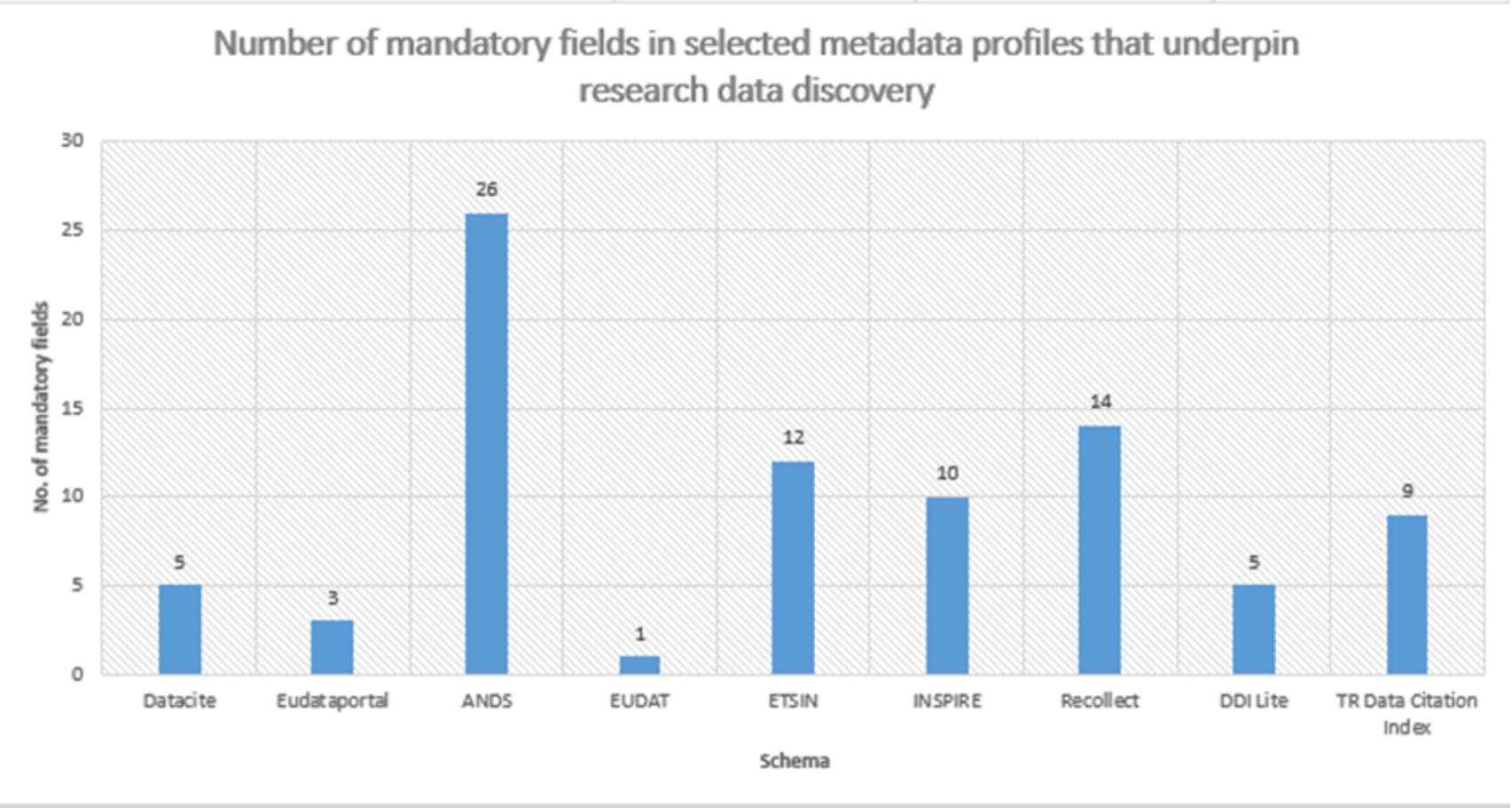


Rationale

» What is mandatory?

» What is common?

Rationale


Number of mandatory fields in selected metadata profiles that underpin research data discovery
No. of mandatory fields
30
25
20
15
10
5
0
5
3
26
1
12
10
14
5
9
Datacite
Eudataportal
ANDS
EUDAT
ETSIN
INSPIRE
Recollect
DDI Lite
TR Data Citation Index
Schema

Core metadata profile

http://tinyurl.com/ha49dj3

ID	Field	Obligation	Occurrence	Description	Content	Example / CV value
1	**Creator**	Mandatory	Repeatable	The name of the primary data creator	free text	e.g. last name, first name ; corporate name
1.1	Creator identifier	Optional	Repeatable	Identifiers associated with creator	http:URI	e.g. http://orcid.org/0000-0001-5352-4666
1.2	Creator identifier scheme	Mandatory if 1.1 is present	Non-repeatable	The scheme of the identifier in 1.1	free text	e.g. ORCiD
1.3	Creator affiliation	Optional	Repeatable	The institution / project / research group that the creator belonged to at the time of publication	free text	e.g. University of London
2	**Title**	Mandatory	Repeatable	The first title of the dataset	free text	e.g. Aerial survey data of Brent Knoll camp
2.1	title type	Optional	Repeatable	To specify additional titles associated with the dataset	Controlled Vocabulary	AlternativeTitle Subtitle TranslatedTitle
3	**Description**	Optional	Repeatable	Text summary explaining the dataset	free text	
4	**subject**	Optional	Repeatable	To provide filtering and associated research	free text	e.g. Archaeology
4.1	subject scheme	Optional	Non-repeatable	describe the standard used for text in subject field.	free text	e.g. Library of Congress Subject Heading
4.2	subject scheme identifier	Optional	Non-repeatable	where applicable	http URI	e.g. http://id.loc.gov/authorities/subjects/sh85006507
5	**keywords**	Recommended	Repeatable	Words and terms associated with the dataset	free text	e.g. drone, 3D map

Colour code

Pink - From user requirements collected during the project

Blue - common research data metadata fields

Grey - associated fields

Core metadata profile

- Findable
  - Persistent identifiers accommodated
- Accessible
  - Persistent identifiers accommodated
  - Access information / protocols harvested
- Interoperable
  - Crosswalks to other schemas
  - Vocabularies & standards used where possible
- Reusable
  - License information harvested
  - Rich descriptive metadata also harvested

ID	Field	Obligation	Occurrence	Description	Content	Example / CV value
1	**Creator**	Mandatory	Repeatable	The name of the primary data creator	free text	e.g. last name, first name ; corporate name
1.1	Creator identifier	Optional	Repeatable	Identifiers associated with creator	http:URI	e.g. http://orcid.org/0000-0001-5352-4666
1.2	Creator identifier scheme	Mandatory if 1.1 is present	Non-repeatable	The scheme of the identifier in 1.1	free text	e.g. ORCID
1.3	Creator affiliation	Optional	Repeatable	The institution / project / research group that the creator belonged to at the time of publication	free text	e.g. University of London
2	**Title**	Mandatory	Repeatable	The first title of the dataset	free text	e.g. Aerial survey data of Brent Knoll camp
2.1	title type	Optional	Repeatable	To specify additional titles associated with the dataset	Controlled Vocabulary	AlternativeTitle Subtitle TranslatedTitle
3	**Description**	Optional	Repeatable	Text summary explaining the dataset	free text	
4	**subject**	Optional	Repeatable	To provide filtering and associated research	free text	e.g. Archaeology
4.1	subject scheme	Optional	Non-repeatable	describe the standard used for text in subject field.	free text	e.g. Library of Congress Subject Heading
4.2	subject scheme identifier	Optional	Non-repeatable	where applicable	http URI	e.g. http://id.loc.gov/authorities/subjects/sh85006507
5	**keywords**	Recommended	Repeatable	Words and terms associated with the dataset	free text	e.g. drone, 3D map
6	**GeoLocation**	Optional	Repeatable	Geographical info where applicable	free text	e.g. Brent Knoll, Somerset Levels, Somerset
6.1	geolocation point	Optional	Non-repeatable		coordinates	e.g. 51°15'14"N 2°58'46"W
6.2	geolocationbox	Optional	Non-repeatable		coordinates	
6.3	geolocationplace	Optional	Non-repeatable		free text	

Schema cardinality (v1.1)

- **Mandatory**
  - Creator
  - Title
  - Unique resource identifier
  - Date
  - Contact
- **Recommended**
  - Keywords
  - Publisher
  - Project
  - Rights
- **Optional**
  - Description
  - Subject
  - Geolocation
  - Format
  - Resource Type
  - Language
  - Related Resource
  - Contributor

OAI-PMH

DDI 2.5

GEMINI 2.2

Dublin Core

ADS

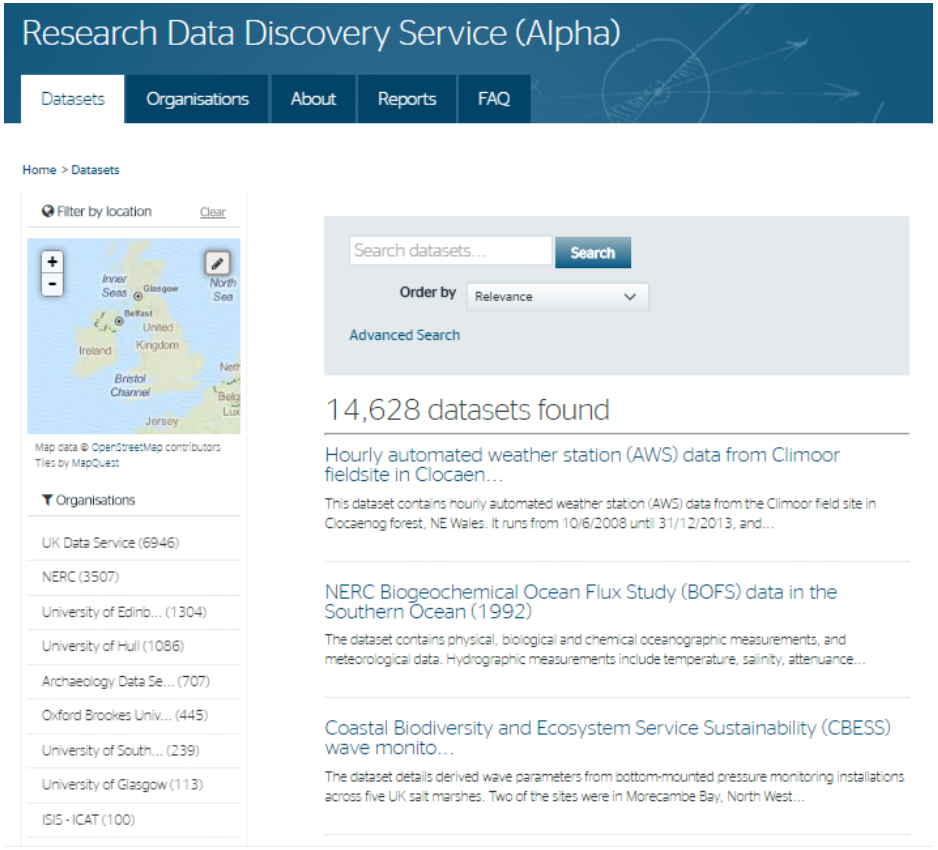
MODS

Crosswalks across schemas

ID	Field	Card.	Description	Content	Mapping						
					DC	MODS	DDI2.5	ADS	GEMINI2	Datacite	OpenAire /C4D CERIF
1	Creator	M, R	The name of the primary data creator	free text	Creator	<name><namePart><role><roleTerm type=text"> creator	AuthEnty: ddi:codeBook/ddi:stdyDscr/ddi:citation/ddi:titlStmt/ddi:AuthEnty/text()	Creators	/gmd:MD_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:pointOfContact/gmd:CI_ResponsibleParty[gmd:role/gmd:CI_RoleCode/@codeListValue='author']/gmd:individualName/gco:CharacterString	creatorName	cfResProd.cfPers_ResProd [relationship : Creator]
1.1	Creator identifier	O, R	Identifiers associated with creator	http:URI						nameIdentifier	cfPers.cfFedId. cfFedId
1.2	Creator identifier scheme	M If 1.1	Scheme of the identifier in 1.1	free text						nameIdentifierScheme	cfPers.cfFedId. cfFedId_Class
1.3.	Creator affiliation	O, R	The institution / project / research group that the creator belonged to at the time of publication	free text			AuthEnty/affiliation	<actor type='creator'><organisation>	/gmd:MD_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:pointOfContact/gmd:CI_ResponsibleParty[gmd:role/gmd:CI_RoleCode/@codeListValue='author']/gmd:organisationName/gco:CharacterString	affiliation	cfPers.cfPers_OrgUnit
2	**Title**	**M, R**	The first title of the dataset	free text	Title	TitleInfo	Titl: 'ddi:codeBook/ddi:stdyDscr/ddi:citation/ddi:titlStmt/ddi:titl/text()	Title	Title	Title	cfResProd.cfName
2.1	title type	O, R	To specify additional titles associated with the dataset	titleType controlled vocabulary						titleType	
3	**Description**	**O, R**	Text summary explaining the dataset	free text	description	abstract	Abstract: 'ddi:codeBook/ddi:stdyDscr/ddi:stdyInfo/ddi:abstract/text()	description	abstract	Description	cfResProd.cfDesc

Latest developments

- Formalise schema development process
- Additional user requirements (e.g. metadata licenses)
- Richer mappings from schemas in use at repositories
- OAI-PMH format options

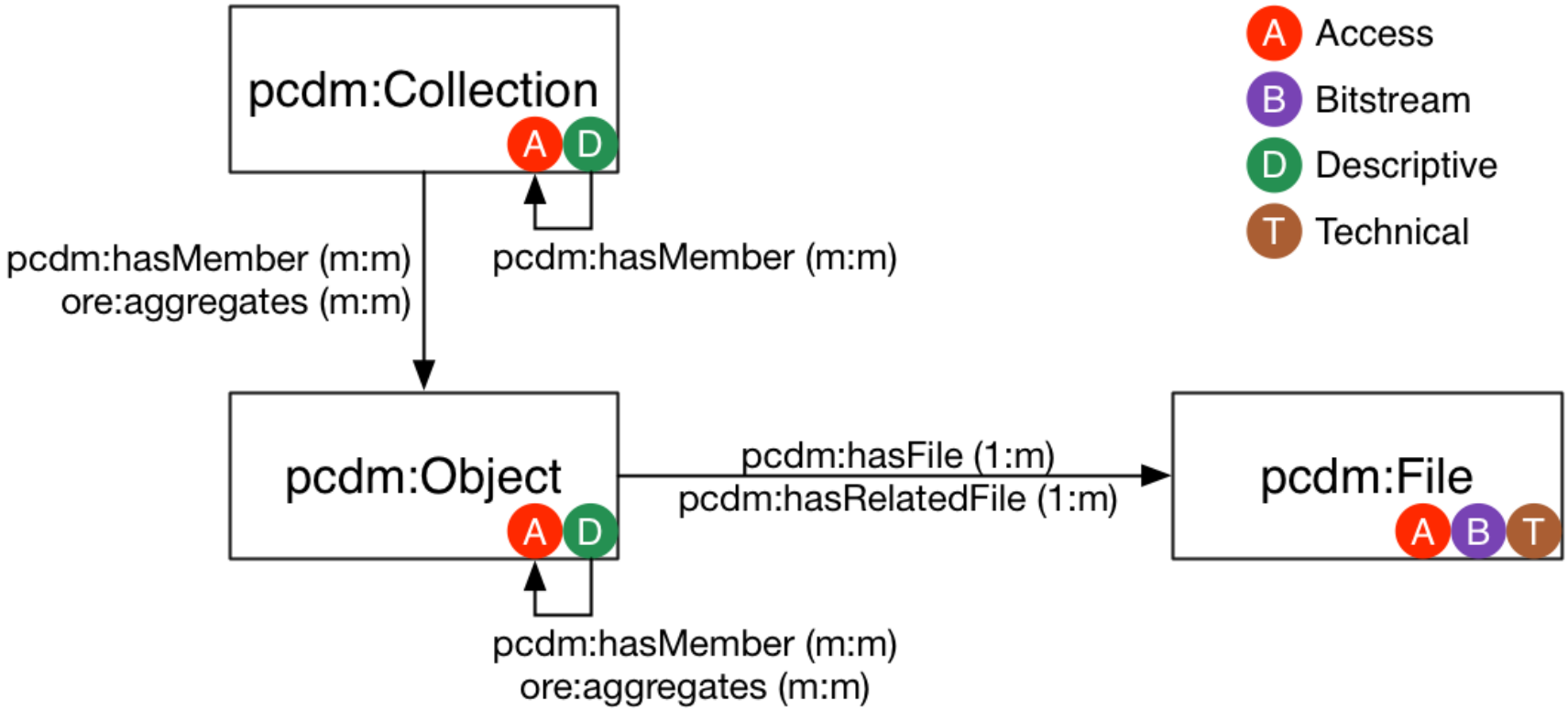


Research Data Shared Service

- Metadata micro-services
- Thresholds for acceptance
- Less than minimum rejected
- Fields weighted to score a bronze, silver or gold rating.
- Encourages metadata completeness
- Automated field population

File level and dataset level metadata

» Portland Common Data Model



Summary

How much is enough?

- It depends…
- No one size fits all solution
- The FAIR principles offer guidelines
- Researchers don't like entering metadata
- Automation can help enrich metadata records
- International scale community consensus is essential

Check out

» General

How much metadata is enough? (Dom Fripp)

https://rdds.jiscinvolve.org/wp/2016/03/18/how-much-metadata-is-enough/

FAIR article in Nature

http://www.nature.com/articles/sdata201618

PCDM

https://github.com/duraspace/**pcdm**/wiki

» Jisc Discovery Service

http://www.slideshare.net/JiscRDM

https://researchdata.jiscinvolve.org/wp/

» JiscRDM

http://www.slideshare.net/JiscRDM

https://researchdata.jiscinvolve.org/wp/

https://research-data-network.readme.io/

» Schemas

DDI 2.5 (social sciences)

http://www.ddialliance.org/Specification/DDI-Codebook/2.5/

GEMINI 2.2 (geographical)

https://www.agi.org.uk/about/resources/category/81-gemini?download=18:gemini-2-2 (PDF)