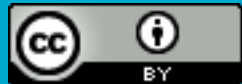


OzNome 5-star Tool:

A Rating System for making data FAIR and Trustable

Simon Cox, Jonathan Yu
20 October 2017

LAND AND WATER
www.csiro.au





OzNome – “A connected Australia”

CSIRO-led initiative to enhance and connect information infrastructures across Australia.

<https://research.csiro.au/oznome/>



Tools, products, services



Methods, approaches, practices

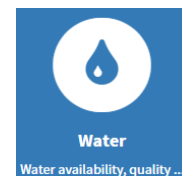


Infrastructure

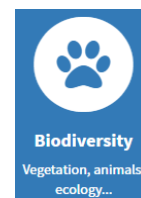


Initial focus on enhancing and connecting **environmental information infrastructures** in Australia, starting with CSIRO L&W.

<https://research.csiro.au/oznome/oznome-land-water/>



+



Motivation

- Environmental data comes from many sources
- Solving big problems the data to be connected
- In order to be connectable, the data should be FAIR

How to make information 'connectable'?

- Follow the FAIR principles
- Assessment tool with recommendations on improvements

Box 2 | The FAIR Guiding Principles

To be Findable:

- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
 - A1.1 the protocol is open, free, and universally implementable
 - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

To be Reusable:

- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
 - R1.1. (meta)data are released with a clear and accessible data usage license
 - R1.2. (meta)data are associated with detailed provenance
 - R1.3. (meta)data meet domain-relevant community standards

Findable
Accessible
Interoperable
Reusable



Findable:

- Citeable via a stable, persistent web identifier
- Described with appropriate metadata
- Indexed in a well known system, e.g. google search, catalog search

Accessible:

- Available on the web? Via standardised web service?
- Curated with a commitment that this data will be available long term
- Indexed in a well known system

Interoperable/Reusable:

- Common formats
- Discoverable, community-endorsed schema or data model
- Unambiguous definitions for all elements (e.g. column definitions, units of measure) linked to accessible (standard) definitions
- Linked to other data using external identifiers (e.g. URIs)
- Licenced

Data rating criteria

1. [published](#)
2. [hosted](#)
3. [curated](#)
4. [updated, maintained](#)
5. [licensed](#)
6. [citeable](#)
7. [described](#)
8. [findable](#)
9. [loadable](#)
10. [useable](#)
11. [comprehensible](#)
12. [connected, linked](#)
13. [assessable](#)
14. [trusted](#)

Key word	Levels	
published	<ul style="list-style-type: none"> a. No external access b. External access, non-web protocol (e.g. physical media distribution) c. Published via the web 	<i>Implicitly FAIR?</i>
hosted	<ul style="list-style-type: none"> a. not on web b. files on web-server c. repository with web interface d. web service - local API e. RESTful web service - OpenAPI/Swagger f. standard web API (SPARQL, OGC WMS/WFS/WCS/SOS/WPS, ...) 	<i>FAIR - Accessible</i>
curated	<ul style="list-style-type: none"> a. once-off dump, no ongoing commitment b. best effort c. institutional repository d. certified repository 	
updated, maintained	<ul style="list-style-type: none"> a. one-time dataset b. part of data series, occasional/irregular update c. part of data series with regular updates 	<i>More than FAIR!</i>

Is it intended to be published? How? How often?

Key word	Levels	<i>FAIR – Findable, Reusable</i>	
licensed	<ul style="list-style-type: none"> a. no licence b. licence described in text c. standard licence (e.g. Creative Commons) 		
citeable	<ul style="list-style-type: none"> a. Not citeable b. Local identifier (may change) c. Web identifier (transient URL or query) d. Persistent web identifier (PURL, DOI, handle, ARK, etc) 		
described	<ul style="list-style-type: none"> a. no metadata b. text description (abstract) and keywords c. basic metadata (e.g. Dublin Core) d. specialized metadata (e.g. Darwin Core, ISO 19115, scientific data profile of schema.org) e. rich metadata using (standard) RDF vocabularies (e.g. DCAT, ADMS, PROV, GeoDCAT, OMV, VoID) 		
findable	<ul style="list-style-type: none"> a. not indexed b. indexed in a local, organizational catalogue c. metadata harvested or pushed into a community (e.g. Research Data Australia, Re3Data) or jurisdictional catalogue d. visible in general-purpose indexes (Google, Bing) e. highly ranked in general-purpose indexes 		

Indexed? Identified? Licensed?

FAIR - Interoperable

Key word	Levels
loadable	<ul style="list-style-type: none">a. bespoke file formatb. standard data-format, denoted by a MIME-type (CSV, JSON, XML, netCDF, etc)c. choice from multiple standard formats
useable	<ul style="list-style-type: none">a. implicit schema, not formalizedb. explicit schema, formalized in DDL, XSD, data-package, RDFS/OWL, JSON-Schema or similarc. community schema, available from a (standard) location
comprehensible	<ul style="list-style-type: none">a. local field labelsb. field labels linked to text explanationsc. standard labels (e.g. CF Conventions, UCUM units)d. some field names linked to standard, externally managed vocabulariese. all field names linked to standard, externally managed vocabularies
connected, linked	<ul style="list-style-type: none">a. no linksb. in-bound links from a catalogue or landing pagec. out-bound links to related data

Format, structure, semantics, links

netCDF metadata example

☐ **e0_avg: Grid**

time: latitude: longitude:

```
_FillValue: -999.0  
name: e0_avg  
long_name: Potential evapotranspiration (atmospheric demand):  
averaged across both HRUs (mm)  
units: mm  
standard_name: e0_avg  
_ChunkSizes: 75, 1, 50
```

netCDF metadata example - interoperable

☐ **e0_avg**: Grid

time: latitude: longitude:

```
_FillValue: -999.0
name: e0_avg
long_name: Potential evapotranspiration (atmospheric demand): averaged across both HRUs (mm)
units: mm
standard_name: e0_avg
scaledQuantityKind_id: http://registry.it.csiro.au/sandbox/csiro/oznome/AWRA-L/potential-
evapotranspiration
substanceOrTaxon_id: http://registry.it.csiro.au/sandbox/soil/soil-object/soil
unit_id: http://registry.it.csiro.au/def/qudt/1.1/qudt-unit/Millimeter
featureOfInterest_id: http://registry.it.csiro.au/sandbox/csiro/oznome/feature-type/critical-zone
```

Key word	Levels	
assessable	<ul style="list-style-type: none"> a. No quality or lineage information b. Lineage statement in text c. Formal provenance trace (W3C PROV-O or similar) 	<i>FAIR – Reusable</i>
trusted	<ul style="list-style-type: none"> a. no information about usage b. usage statistics available. c. Clearly endorsed by reputable organization or framework 	<i>More than FAIR?</i>

Quality, provenance, trusted?

5-star assessment tool

<http://oznome.csiro.au/5star/>



5 ★ OZNOME DATA

OzNome proposes a 5-star scheme for assessing the social, technical and informational attributes of data. The 5-star scheme aims to help users know whether some data or a service is 'OzNomic'? Here, we give examples and explain costs and benefits that come along with it.

Findable	★★★★★
Accessible	★★★★★
Interoperable	★★★★★
Reusable	★★★★★
Trusted	★★★★★

Self-assessment tool (version 1)

The following questionnaire provides you with a tool to assess whether your dataset meets the 5 ★ OZNOME data criteria. After answering the questions, the tool displays a chart summarising your data according to the scheme.

Questionnaire	Results										
<p>Tell us about your data</p> <p>... publication and indexing</p> <p>1. * Dataset identity</p> <p>Dataset name or title <input type="text"/></p> <p>URL <input type="text"/></p> <p>2. * Published - is the data accessible to users other than the creator or owner?</p> <p><input type="radio"/> No</p> <p><input type="radio"/> By individual arrangement</p> <p><input type="radio"/> File download</p> <p><input type="radio"/> Institutional or community repository</p> <p><input type="radio"/> Bespoke web service (informal API)</p> <p><input type="radio"/> Bespoke web service (OpenAPI/Swagger)</p> <p><input type="radio"/> Standard web service API (e.g. OGC)</p> <p>3. Citeable - denoted using a formal identifier</p> <p><input type="radio"/> Not citeable</p> <p><input type="radio"/> Local identifier</p> <p><input type="radio"/> Web address (URL - not guaranteed stable)</p>	<table><tbody><tr><td>Findable</td><td>★★★★★</td></tr><tr><td>Accessible</td><td>★★★★★</td></tr><tr><td>Interoperable</td><td>★★★★★</td></tr><tr><td>Reusable</td><td>★★★★★</td></tr><tr><td>Trusted</td><td>★★★★★</td></tr></tbody></table>	Findable	★★★★★	Accessible	★★★★★	Interoperable	★★★★★	Reusable	★★★★★	Trusted	★★★★★
Findable	★★★★★										
Accessible	★★★★★										
Interoperable	★★★★★										
Reusable	★★★★★										
Trusted	★★★★★										

Interoperable?

> 10 years of information standards work in CSIRO

Accessible ★★★★★
Interoperable ★★★★★
Reusable ★★★★★
Trusted ★★★★★

Self-assessment tool (version 1)

The following questionnaire provides you with a tool to assess whether your dataset meets the 5 ★ OZNAME data criteria. After answering the questions, the tool displays a chart summarising your data according to the scheme.

Questionnaire	Results
<p>Tell us about your data</p> <p>... linked and useable</p> <p>6. Loadable - represented using a common or community-endorsed (i.e. standard) format</p> <ul style="list-style-type: none"><input type="radio"/> bespoke format (text, binary)<input type="radio"/> one standard format, denoted by a MIME-type<input type="radio"/> multiple standard formats <p>7. Useable - structured using a discoverable, community-endorsed (standard?) schema or data model</p> <ul style="list-style-type: none"><input type="radio"/> no formal schema<input type="radio"/> explicit schema or data model, formalized in DDL, XSD, DDI, RDFS, JSON-Schema, data-package or similar<input type="radio"/> community-shared schema or data model, available from a standard location <p>8. Comprehensible - supported with unambiguous definitions for all internal elements</p> <ul style="list-style-type: none"><input type="radio"/> local field codes or labels<input type="radio"/> labels with full text explanations<input type="radio"/> community standard labels (e.g. CF Conventions, UCUM units)<input type="radio"/> some fields linked to externally managed definitions<input type="radio"/> all fields linked to standard, externally managed definitions <p>9. Linked - to other data and definitions using public identifiers (e.g. URIs)</p> <ul style="list-style-type: none"><input type="radio"/> no links<input type="radio"/> in-bound links from a catalogue or landing-page<input type="radio"/> out-bound links to related data and definitions <p>10. Licensed - conditions for re-use are available and clearly expressed</p> <ul style="list-style-type: none"><input type="radio"/> no license<input type="radio"/> license described in text<input type="radio"/> link to a standard license (e.g. Creative Commons) <p>Previous Next</p>	<p>Findable ★★★★★ Accessible ★★★★★ Interoperable ★★★★★ Reusable ★★★★★ Trusted ★★★★★</p>

OzNome maturity estimation

ASRIS



oznome data rating 4.21 stars

Findable via ANDS/RDA

Accessible - Available as web service

Interoperable/Reusable:
web services, standard schema.
Standard vocabularies.

Trusted: Reliable operationalised

OzNome maturity estimation

AWRA-L



oznome data rating 3.53 stars

Findable via Google search

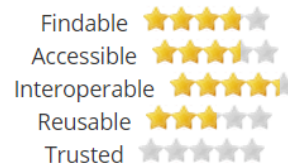
Accessible - publish limited set from 2005

Available as web service

Interoperable/Reusable: some web services, reference definitions as text

Trusted: Reliable operationalised updates of 2005 data

AWRA-L CSIRO Cache



oznome data rating 2.90 stars

Findable/Accessible:

1911-2016 dataset test deployments and aggregates
via connected infrastructure (internal CSIRO)
Enhanced connectivity via web services

Interoperable/Reusable: web services, reference definitions as Linked Data and externally hosted observable properties vocabulary definitions

Trusted: Not operational and no trusted repository

Oznome data assessment criteria

<https://confluence.csiro.au/display/OZNOME/Data+ratings>

Key word	Matching <u>FAIR</u> Principle
published	<i>Implicitly FAIR</i>
hosted	A1 - A2
curated	<i>More than FAIR</i>
updated, maintained	<i>More than FAIR</i>
licensed	R1.1
citeable	F1
described	R1 , F2 , F3
findable	R1 , F2 , F3
loadable	I1
useable	I2 , R1.3
comprehensible	I2
connected, linked	I3
assessable	R1.2
trusted	<i>More than FAIR</i>

Summary & conclusions

- Augment FAIR principles
 - curated, updated, maintained, trusted
- Add specific guidance on maturity within each criterion
 - Tuned to geospatial/environmental data
- Form-based tool for self-assessment

Links and references

- FAIR principles <https://www.force11.org/group/fairgroup/fairprinciples>
- FAIR principles and metrics for evaluation
<https://www.slideshare.net/micheldumontier/fair-principles-and-metrics-for-evaluation>
- OzNome data assessment criteria -
<https://confluence.csiro.au/display/OZNOME/Data+ratings>
- 5-star tool - <http://oznome.csiro.au/5star/>

Thank you

Land and Water

Simon J D Cox
Research Scientist

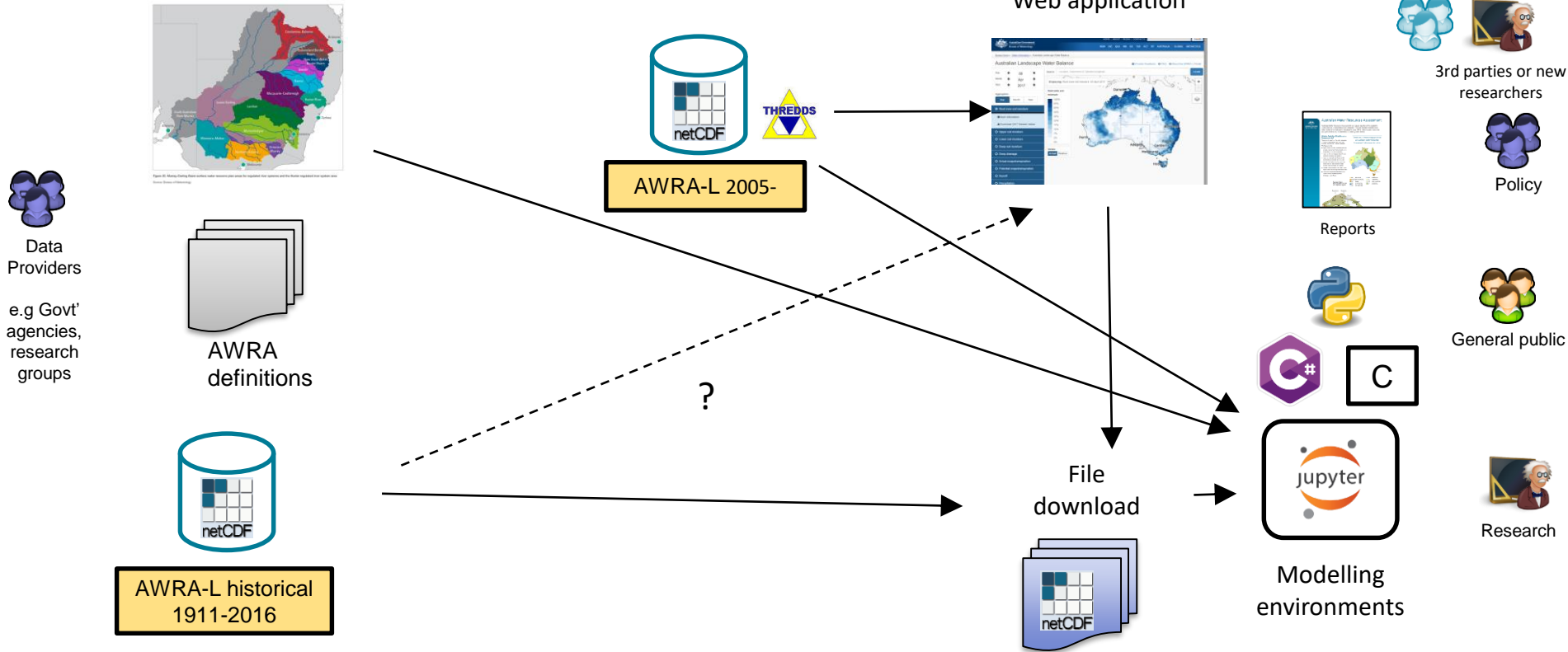
t +61 2 9545 2365
e simon.cox@csiro.au
w people.csiro.au/Simon-Cox

Land and Water

Jonathan Yu
Research Scientist

t +61 3 9545 2457
e jonathan.yu@csiro.au
w people.csiro.au/Jonathan-Yu

Geofabric features




Fundamental Questions

- In what ways can we **assess** the FAIRness of a digital resource?
- To what degree can we **automate** this assessment?
- Must we treat each type of digital resource **differently**?
- **Who will use the metrics**? The producers, the funders, or the users?
- Can one resource be **more FAIR** than another?
- Will/should FAIRness assessments **impact funding** decisions?
- Should only one **organization** define these metrics? Or can **anybody** make their own metrics? What happens if a digital resources scores well against one set of metrics, but not another?

AWRA-L Draft Vocabularies

Contents (tree view)

deep drainage Deep drainage is an estimate of the water that drains from the bott...

 **evapotranspiration** Evapotranspiration is an estimate of the total evapotranspiration f...

actual evapotranspiration Actual Evapotranspiration is an estimate of the total evapotranspir...

potential evapotranspiration The potential evapotranspiration in AWRA-L is calculated on a 0.05 ...

precipitation Daily precipitation grids are produced by the Bureau from approxima...

runoff Runoff represents a modelled estimate expected from a small unimpai...

 **soil moisture** Soil Moisture estimate represents the percentage of available water...

deep soil moisture Deep Soil Moisture represents the percentage of available water con...

 **root zone soil moisture** Root Zone Soil Moisture is the sum of water in the AWRA-L Upper and...

License: <https://creativecommons.org/licenses/by/4.0/>

Referenceable metadata for implicit observable properties

Online AWRA vocabulary register (draft)

<http://registry.it.csiro.au/sandbox/csiro/oznome/AWRA-L>

Mappings from AWRA data to online draft AWRA vocabularies

<https://confluence.csiro.au/display/OFW/AWRA-L+Vocabulary+mappings>

AWRA Draft Vocabulary example

Entry: potential evapotranspiration

URI: <http://registry.it.csiro.au/sandbox/csiro/oznome/AWRA-L/potential-evapotranspiration>

The potential evapotranspiration in AWRA-L is calculated on a 0.05 degree (approximately 5 x 5 km) national grid using the Penman (1948) equation. Potential evapotranspiration provides an upper limit on evaporation and transpiration processes from the soil and vegetation and depends solely on the available energy at the surface. The daily gridded climate datasets used to produce this estimate include downward solar irradiance, and maximum and minimum air temperature produced by the Bureau of Meteorology (Jones et al., 2009) and windspeed at 2 m which is input as a spatially-gridded long-term average (McVicar et al., 2008).

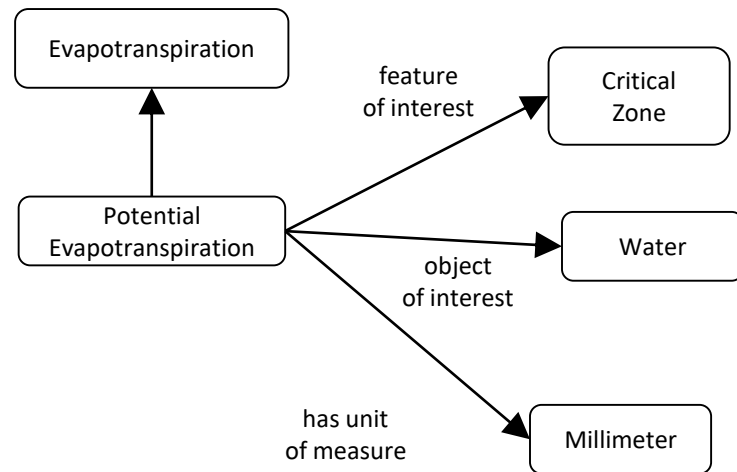
Definition

broader	evapotranspiration
description	The potential evapotranspiration in AWRA-L is calculated on a 0.05 degree (approximately 5 x 5 km) national grid using the Penman (1948) equation. Potential evapotranspiration provides an upper limit on evaporation and transpiration processes from the soil and vegetation and depends solely on the available energy at the surface. The daily gridded climate datasets used to produce this estimate include downward solar irradiance, and maximum and minimum air temperature produced by the Bureau of Meteorology (Jones et al., 2009) and windspeed at 2 m which is input as a spatially-gridded long-term average (McVicar et al., 2008).
feature of interest	critical zone
generalization	evapotranspiration
label	potential evapotranspiration
object of interest	water
pref label	potential evapotranspiration
source	http://www.bom.gov.au/water/landscape/ 58518bc790ff7
type	scaled quantity kind Concept quantity kind mechanics quantity kind

Links

- Has broader concept
- evapotranspiration
- Feature of interest
- critical zone
- Object of interest
- water
- Has more general quantity kind
- evapotranspiration
- Has unit of measure
- millimeter

<http://registry.it.csiro.au/sandbox/csiro/oznome/AWRA-L/potential-evapotranspiration>



Key word	Levels	<i>FAIR - Interoperable</i>	
loadable	a. bespoke file format b. standard data-format, denoted by a MIME-type (CSV, JSON, XML, netCDF, etc) c. choice from multiple standard formats		

Key word	Levels	<i>FAIR - Interoperable</i>	
loadable	<ul style="list-style-type: none"> a. bespoke file format b. standard data-format, denoted by a MIME-type (CSV, JSON, XML, netCDF, etc) c. choice from multiple standard formats 		
useable	<ul style="list-style-type: none"> a. implicit schema, not formalized b. explicit schema, formalized in DDL, XSD, data-package, RDFS/OWL, JSON-Schema or similar c. community schema, available from a (standard) location 		

FAIR - Interoperable

Key word	Levels
loadable	<ul style="list-style-type: none">a. bespoke file formatb. standard data-format, denoted by a MIME-type (CSV, JSON, XML, netCDF, etc)c. choice from multiple standard formats
useable	<ul style="list-style-type: none">a. implicit schema, not formalizedb. explicit schema, formalized in DDL, XSD, data-package, RDFS/OWL, JSON-Schema or similarc. community schema, available from a (standard) location
comprehensible	<ul style="list-style-type: none">a. local field labelsb. field labels linked to text explanationsc. standard labels (e.g. CF Conventions, UCUM units)d. some field names linked to standard, externally managed vocabulariese. all field names linked to standard, externally managed vocabularies

Box 2 | The FAIR Guiding Principles

To be Findable:

- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
 - A1.1 the protocol is open, free, and universally implementable
 - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

To be Reusable:

- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
 - R1.1. (meta)data are released with a clear and accessible data usage license
 - R1.2. (meta)data are associated with detailed provenance
 - R1.3. (meta)data meet domain-relevant community standards

37 repositories

DANS-EASY	https://easy.dans.knaw.nl/ui/home
EUDAT-B2Share	https://b2share.eudat.eu/
Zenodo	https://zenodo.org
PseudoBase	http://www.ekevanbatenburg.nl/PKBASE/PKB.HTML
OpenML	http://www.openml.org/
Profiles-Registry	http://www.profilesregistry.nl/
Mendeley-Data	https://data.mendeley.com/
4TU.Centre for Research Data	http://data.4tu.nl/
CancerData.org	https://www.cancerdata.org
DHS Data Access	http://www.dhsdata.nl
WorldClim	http://worldclim.org/
World Data Centre for Soil	http://www.isric.org/
Infrared Space Observatory	http://www.cosmos.esa.int/web/iso/access-the-archive
Longitudinal Aging Study Amsterdam	http://www.lasa-vu.nl/index.htm
Southeast Asian Climate Assessment & Dataset	http://saca-bmkg.knmi.nl/
TRAILS	https://www.trails.nl/
ICOS Carbon Portal	https://www.icos-cp.eu/node/1
CESSDA	http://cessda.net/
SeaDataNet	http://www.seadatanet.org/
LISS	https://www.lissdata.nl/lissdata/
ORGIDS / RodRep	http://www.orgids.com/ / http://www.rodrep.com/
earth2observe	http://www.earth2observe.eu/
EDGAR	http://edgar.jrc.ec.europa.eu/

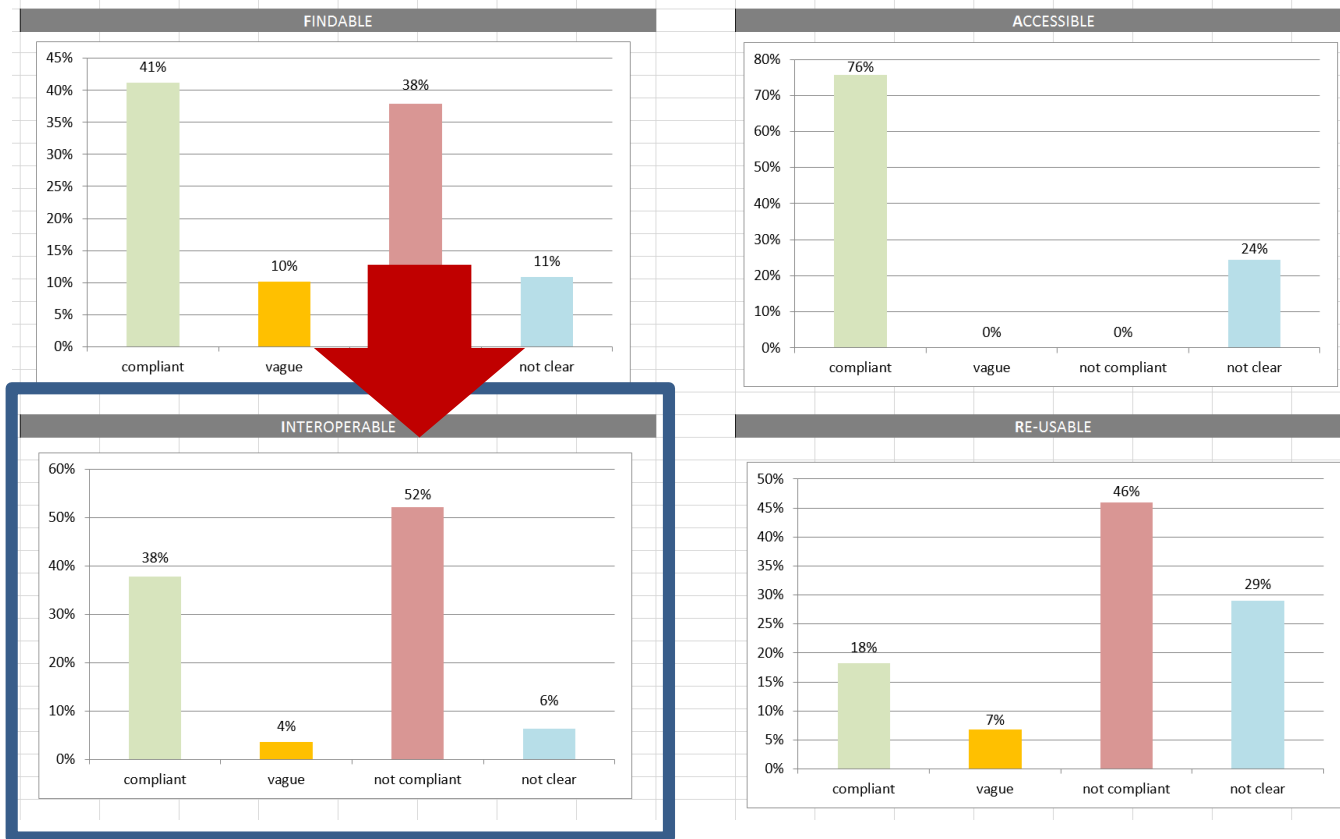
KNMI	https://data.knmi.nl/datasets
STITCH	http://stitch.embl.de/
ECA&D	http://www.ecad.eu/
Europeana	http://www.europeana.eu/portal/en
Mycobank	http://www.mycobank.org/
AlgaeBase	http://www.algaebase.org/
Amsterdam Cohort Studies	https://www.amsterdamcohortstudies.org/acsc/index.asp
ICTWSS	http://uva-aiaa.net/en/ictwss
Share ERIC	http://www.share-project.org/
LOVD3	http://databases.lovd.nl/whole_genome/genes
CARIBIC	http://www.caribic-atmospheric.com/
EIDA	http://www.orfeus-eu.org/data/eida/
Sound and Vision	http://www.beeldengeluid.nl/en
Figshare	https://figshare.com/

Scoring the resources

General Overview (GO)	DANS-EASY	UDAT-B2Shar	Zenodo	PseudoBase	OpenML	Profiles-Registry	Mendeley-Data	4TU
	https://easy	https://b2share	https://zenodo	http://www.eke	http://www.openml	http://www.profiles-registry	https://data.mendeley	http://data.4tu
FINDABLE								
(meta)data are assigned a globally unique and eternally persistent identifier								
data are described with rich metadata								
(meta)data are registered or indexed in a searchable resource (able to google data-objects)								
metadata specify the data identifier								
ACCESSIBLE								
(meta)data are retrievable by their identifier using a standardized communications protocol								
the protocol is open, free, and universally implementable								
the protocol allows for an authentication and authorization procedure, where necessary								
metadata are accessible, even when the data are no longer available								
Interoperable								
(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation*								
(meta)data use vocabularies that follow FAIR principles								
(meta)data include qualified references to other (meta)data								
Re-usable								
meta(data) have a plurality of accurate and relevant attributes								
(meta)data are released with a clear and accessible data usage license								
(meta)data are associated with their provenance								
(meta)data meet domain-relevant community standards								

	complies completely
	just about / maybe not
	fails to comply
	unclear

Overall evaluation



Unpacking data



Nitrogen
Newtons
North
Noon
Neutral
November
Moles
No!

One symbol, many meanings

Geofabric features

