Elaboration of the FAIR principles and metrics

Michel Dumontier

Box 2 | The FAIR Guiding Principles

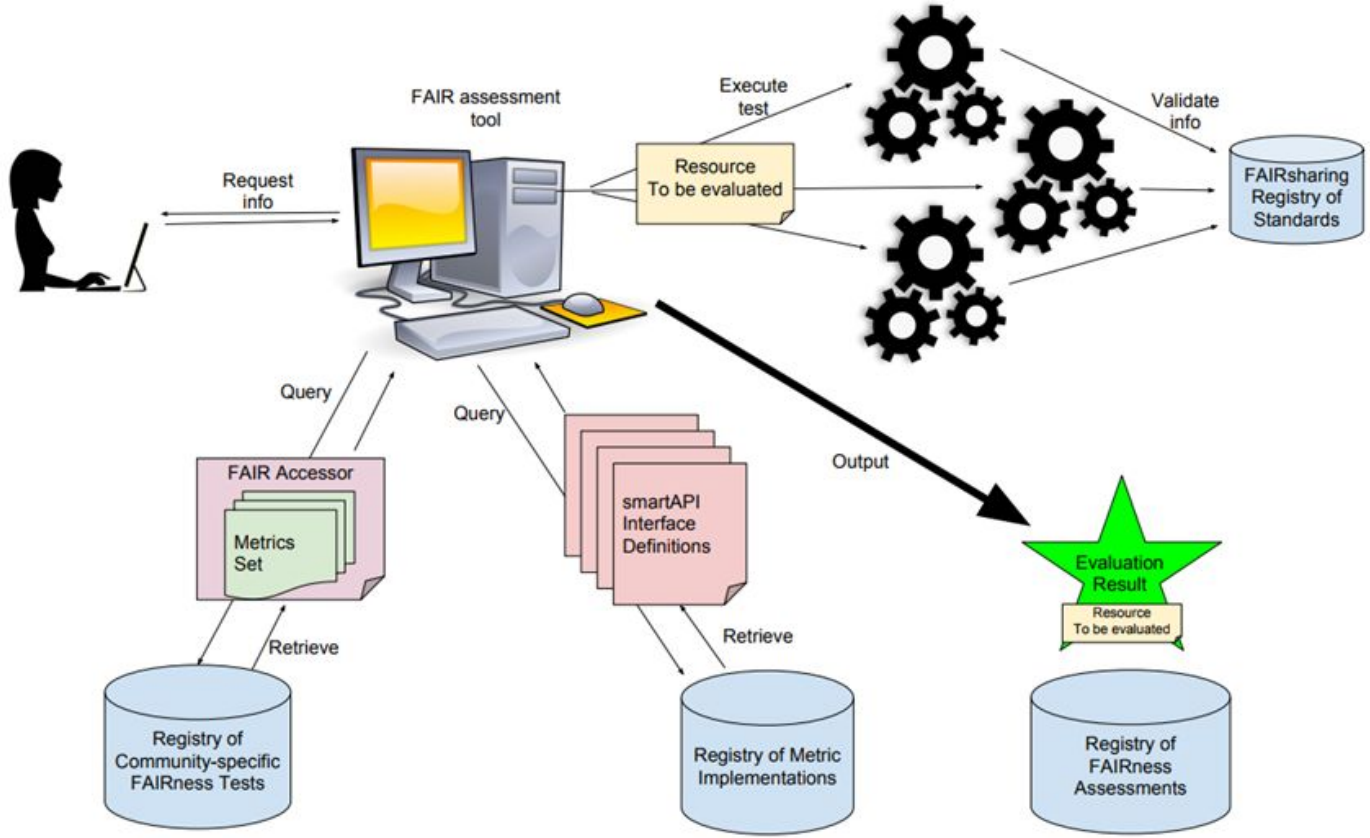
To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

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

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



https://www.nature.com/articles/sdata2018118

F1: (meta) data are assigned **globally unique** and **persistent identifiers**

The **uniqueness of an identifier** is a necessary condition to **unambiguously refer that resource,** and that resource alone. Otherwise, an identifier shared by multiple resources will confound efforts to describe that resource, or to use the identifier to retrieve it.

A **persistent identifier** is one where once assigned, an identifier denotes the **same referent indefinitely**, and is never re-assigned to another entity. This ensures that the identifier is can be used in the future to specifically find the content that it refers to.

Example Identifiers

3336842

gi:3336842

doi:10.1038/sdata.2016.18

doi:10.4121/uuid:5146dd06-98e4-426c-9ae5-dc8fa65c549f

dg.4503/00e6cfa9-a183-42f6-bb44-b70347106bbe

https://orcid.org/0000-0001-8888-635X

https://raw.githubusercontent.com/FAIRDataInitiative/FAIR-principles/master/fair.trustyuri.trig#np.RA4FsMT1XSZMh-JgNmAuOVQ3qyTzkaMldT_KxR1HSdoUA

Globally unique and persistent identifiers

Obtain **globally unique and persistent identifiers** from a **software or service** that uses algorithms that can detect changes in the content.

- Persistent URLs: http://www.purlz.org
- Identifiers.org: http://identifiers.org
- Digital Object Identifier: http://www.doi.org
- Archival Resource Key (ARK): http://n2t.net/e/ark_ids.html
- Cross-Ref (for funding agencies): https://www.crossref.org/services/funder-registry/
- Global research identifier Identifiers: https://www.grid.ac

Globally unique and persistent identifiers owing to content fingerprint

- Data GUIDs https://dataguids.org
- Trusty URIs: http://trustyuri.net/

F1 Metrics

1) provide a **URL** to **a document that describes the identifier scheme.** FAIRsharing lists different identifier schemes
https://fairsharing.org/standards/?q=&selected_facets=type_exact:identifier%20schema

FAIRsharing.org
standards, databases, policies

2) provide a **URL** to **a document that details the identifier management plan**. This includes aspects related to **what happens in the event of IT disaster**, **backwards compatiblity** in the event of changes in identifier scheme, and the **transfer of authority.**

DMP ONLINE

F2: Data are described with rich metadata

Metadata are **structured descriptions of a resource**. They include elements such as identifier, title, description, creator, version, provenance, license, etc.

Metadata play an important role in **enabling users to find** a resource of interest. For instance, indexing metadata in a **search system** can enable users to find resources using keywords or attribute-based filters.

Rich metadata are simply metadata with a plurality of attributes that could be useful for users to find (F1) and reuse (R) the resource of interest. **Be generous** with you descriptions, you never know what people will search for!

W3C Interest Group Note

Dataset Descriptions: HCLS Community Profile

W3C Interest Group Note 14 May 2015

This version:
http://www.w3.org/TR/2015/NOTE-hcls-dataset-20150514/
Latest version:
http://www.w3.org/TR/hcls-dataset/
Editors:
Alasdair J.G. Gray, Heriot-Watt University, UK <A.J.G.Gray@hw.ac.uk>
Joachim Baran, Stanford University, USA <kim@codamono.com>
M. Scott Marshall, MAASTRO Clinic, The Netherlands <m.scott.marshall@maastro.nl>
Michel Dumontier, Stanford University, USA <michel.dumontier@stanford.edu>

```
###Summary Level (Complete)
:chembl
    rdf:type dctypes:Dataset;
    dct:title "ChEMBL"@en ;
    dct:alternative "ChEMBLdb"@en ;
    dct:description "ChEMBL is a database of bioactive compounds, their quantitative properties and
    bioactivities (binding constants, pharmacology and ADMET, etc). The data is abstracted and curated
    from the primary scientific literature."@en ;
    dct:publisher :ebi ;
    foaf:page <http://www.ebi.ac.uk/chembl/> ;
    schemaorg:logo <http://www.ebi.ac.uk/rdf/sites/ebi.ac.uk.rdf/files/resize/images/rdf/chembl_service_logo-146x48.gif> ;
    dct:license <http://creativecommons.org/licenses/by-sa/3.0/> ;
    dct:rights """The data in ChEMBL is covered by the Creative Commons By Attribution. Under the -BY clause,
    we request attribution for subsequent use of ChEMBL. For publications using ChEMBL data, the primary
    current citation is:

1. A. Gaulton, L. Bellis, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, R. Akhtar, A.P. Bento,
B. Al-Lazikani, D. Michalovich, & J.P. Overington (2012) 'ChEMBL: A Large-scale Bioactivity Database For Chemical
Biology and Drug Discovery' Nucl. Acids Res. Database Issue. 40 D1100-1107 DOI:10.1093/nar/gkr777 PMID:21948594

If ChEMBL is incorporated into other works, we ask that the ChEMBL IDs are preserved, and that the release
number of ChEMBL is clearly displayed."""@en ;
    dcat:theme ncit:C48807 ; #chemical
    dcat:keyword "assay"^^xsd:string, "chemical"^^xsd:string ;
    dct:references <http://dx.doi.org/10.1093/bioinformatics/btt765> ;
    rdfs:seeAlso <http://en.wikipedia.org/wiki/ChEMBL> ;
    cito:citesAsAuthority <http://nar.oxfordjournals.org/content/40/D1/D1100> ;
    dct:hasPart :chembl17_rdf_molecule_dataset, :chembl17_rdf_target_dataset ;
#Identifiers
    idot:preferredPrefix "chembl" ;
    idot:alternatePrefix "chembldb" ;
#Provenance and Change
    pav:hasCurrentVersion :chembl17 ;
    dct:accrualPeriodicity freq:quarterly;
#Availability/Distributions
    dcat:accessURL <ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb> ;
    void:sparqlEndpoint <https://www.ebi.ac.uk/rdf/services/chembl/sparql>;
.
```

Rich metadata

Specifications

- HCLS Community Profile
- Dublin Core
- DICOM image metadata
- ISA

Tools

- CEDAR workbench
- DTL metadata editor
- ISA tools

F2 Metrics

1) Provide a **URL** to your **machine-readable metadata document**.
2) Provide a **URL** to the **standard** that the metadata document **uses**.

F3: Metadata clearly and explicitly include the identifier of the data it describes

Metadata are intended to provide information about a digital resource. However, digital resources are often separate from their metadata (they are in different files and in different formats). Since F1 specifies that metadata and data must have different identifiers, it is important that **metadata contain the resource identifier** so that the resource can be **exactly** accessed by its identifier (A1).

```
<metadata ...>
  <dc:identifier xsi:type="dcterms:URI">http://www.ukoln.ac.uk/</dc:identifier>
```

F3 metrics

1) Provide the **URL** for the **metadata document**
2) Provide the **resource identifier** that should be found in the metadata document

F4: (meta)data are registered or indexed in a searchable resource

Rich metadata are key to understand the nature, provenance, and accessibility of the digital resource. However, to be truly findable by a person or machine that is not familiar with the resource requires that the **metadata are indexed in a web-accessible (and FAIR) database.**

F4 examples

https://fairsharing.org/search/?q=http://www.w3.org/TR/hcls-dataset/

https://fairsharing.org/search/?q=hcls+dataset+description

https://www.google.co.uk/search?q=https://www.w3.org/TR/hcls-dataset/

https://www.google.co.uk/search?q=hcls+dataset+description

F4 metrics

Up to 4 URLs, in i) a specialized repository and ii) a web search engine using a) resource identifier and b) other descriptor. Results should be in a standard machine readable format.

A1: (meta)data are retrievable by their identifier using a standardised communication protocol.

A1.1: The protocol is open, free and universally implementable

Digital resources and their metadata should be retrievable throguh standardised communication protocols. **Open, free, and standardised communication** protocols to ensure the possibility of access while eliminating a monetary tariff and additonal effort to gain authorized access to a digital resource.

A1 examples

TCP/IP, as the standard to enable client/server interactions on the internet

HTTP as the standard for client/server interactions to interact with web content

https://www.w3.org/Protocols/

Skype, as a proprietary non-standardized protocol
https://en.wikipedia.org/wiki/Skype_protocol

A1: (meta)data are retrievable by their identifier using a standardised communication protocol.

A1.2: The protocol allows for an authentication and authorisation when required

Some digital resources contain **sensitive data**, and require **additional measures** (such as institutional review board approval) to be followed before access can be granted. **The ‘A’ in FAIR does not imply that the resource must be ‘Open’ or ‘Free’, but it does require that the exact conditions and the process to access restricted data are transparent and made public.** Any additional authentication and authorization procedures must be specified. Therefore, prior to the release of a restricted digital resource, publishers must take steps to clarify eligibility and process.

A1.2 examples

Institutional review board (IRB) approval

Agreement to terms and conditions

Registration

Payment of fees

Use of HTTPS for secure communication

Use of keys to encrypt and decrypt content

A1 metrics

1) Provide a URL to the communication protocol.
2) Provide a URL to a document that describes restricted access protocol, if any

A2: Metadata should be accessible even when the data is no longer available

Making all digital resources ever created available for all of time is an unsolved problem.

Either by design or by accident, digital resources may become lost or inaccessible. Given that any digital resource may have been used and are referenced by others, it is extremely important that auditors have, in the very least, access to the metadata in order to best understand their nature and their provenance. Therefore, **metadata should persist even when the digital resources they describe are not available**.

A2 metrics

1) A URL to a web-accessible document that describes how resources and their metadata will persist in the longest time possible (e.g. data management plan).

I1: (meta)data use a formal, accessible, shared and broadly applicable language for knowledge representation

Consumers of digital resources spend far too much time trying and money to make sense of archaic formats that are poorly documented. Moreover, the **lack of formal specifications mean that machines cannot readily process the content.** With the exception of media languages (e.g. jpg), several languages (e.g. rdf, json-ld, json, xml) have now been standardized to allow structured content to be created, along with a syntax and semantics that makes them directly interpretable by people and machines.

I1 examples

text file

CSV file

JSON

XML

JSON-LD

RDF

Formal specification

To satisfy this principle, the format of the digital resource must be specified in Backus–Naur Form (BNF) or variant thereof (e.g. EBNF, ABNF, etc), or be from one a registry of valid digital media formats or knowledge representation languages.

Excerpt from OWL EBNF

ontology ::= '*Ontology(*' [***ontologyID***] { ***directive*** } '*)*'
directive ::= '*Annotation(*' ***ontologyPropertyID ontologyID*** '*)*'
| '*Annotation(*' ***annotationPropertyID URIreference*** '*)*'
| '*Annotation(*' ***annotationPropertyID dataLiteral*** '*)*'
| '*Annotation(*' ***annotationPropertyID individual*** '*)*'
| ***axiom***
| ***fact***

https://www.w3.org/TR/owl-semantics/syntax.html

I1 metrics

FAIRsharing.org
standards, databases, policies

1) A URL to a **machine-readable specification** of the format used to represent the **digital resource**
2) A URL to a **machine-readable specification** of the format used to represent the **metadata**

I2: (meta)data use vocabularies that follow the FAIR principles

Shared specifications of knowledge, such as vocabularies, ontologies, and data models offer **reusable concepts and data structures** that foster interoperability at both a social and technological level. However, these specifications should also be FAIR so that they can themselves be findable, accessible, interoperable, and reusable.

I2 example

proteinX isExpressedIn amoeba

use NCBI taxonomy and their identifiers to specify species for an experiment

proteinX isExpressedIN ncbitaxon:5775

I2 metrics

FAIRsharing.org
standards, databases, policies

1) Provide a URL to the vocabularies used
2) Provide a URL to their FAIR assessment

I3: (meta)data include qualified references to other (meta)data

FAIR data or metadata typically do not exist in a silo - they can and should be connected to other digital resources.

I3 asks that these connections or references are “qualified”, that is to say that **the nature of a relationship to another resource is clearly indicated**.

For instance, subsequent versions of metadata or digtial resources can be linked to prior versions using a named relation such as “prior version”. Data items, such as a named city "e.g. Maastricht" should be linked to cities in a global repository such as Wikidata or GeoNames.

I3 Metrics

Connection to other resources can be specified in a formal document called a LinkSet (https://www.w3.org/TR/void/). To satisfy this metric, provide a URL to the LinkSet document for the digital resource.

R1: meta(data) are richly described with a plurality of accurate and relevant attributes

A plurality of accurate and relevant attributes are needed not *only* to find (F2) a resource of interest, but also to determine whether the resource is a) appropriate for the new intended use and b) allowed. This comes down to:

i) is there a license and are the terms satisfactory

ii) is there detailed provenance to understand how the resource was generated

iii) does it meet the community standard in terms of quality and availability

R1.1: (meta)data are released with a clear and accessible data usage license

Digital **resources and their metadata** must have terms of use, or license. The lack of a license should default to that no rights are granted, thereby detering lawful use. Note that the combination restrictive license conditions may ultimately preclude the use of any one resource.

https://en.wikipedia.org/wiki/Comparison_of_free_and_open-source_software_licenses

License	Author	Latest version	Publication date	Linking	Distribution	Modification	Patent grant	Private use	Sublicensing	TM grant
Academic Free License[10]	Lawrence E. Rosen	3.0	2002	Permissive	Permissive	Permissive	Yes	Yes	Permissive	No
Affero General Public License	Affero Inc	2.0	2007	Copyleft[11]	Copyleft except for the GNU AGPL[11]	Copyleft[11]	?	Yes[11]	?	?
Apache License	Apache Software Foundation	2.0	2004	Permissive[12]	Permissive[12]	Permissive[12]	Yes[12]	Yes[12]	Permissive[12]	No[12]
Apple Public Source License	Apple Computer	2.0	August 6, 2003	Permissive	?	Limited	?	?	?	?
Artistic License	Larry Wall	2.0	2000	With restrictions	With restrictions	With restrictions	No	Permissive	With restrictions	No
Beerware	Poul-Henning Kamp	42	1987	Permissive	Permissive	Permissive	No	Permissive	Permissive	No
BSD License	Regents of the University of California	3.0	?	Permissive[13]	Permissive[13]	Permissive[13]	Manually[13]	Yes[13]	Permissive[13]	Manually[13]
Boost Software License	?	1.0	August 17, 2003	Permissive	?	Permissive	?	?	?	?
Creative Commons Zero	Creative Commons	1.0	2009	Public Domain[14][15]	Public Domain	Public Domain	No	Public Domain	Public Domain	No
CC-BY	Creative Commons	4.0	2002	Permissive[16]	Permissive	Permissive	No	Yes	Permissive	?
CC-BY-SA	Creative Commons	4.0	2002	Copyleft[16]	Copyleft	Copyleft	No	Yes	No	?
CeCILL	CEA / CNRS / INRIA	2.1	June 21, 2013	Permissive	Permissive	Permissive	No	Permissive	With restrictions	No
Common Development and Distribution License	Sun Microsystems	1.0	December 1, 2004	Permissive	?	Limited	?	?	?	?
Common Public License	IBM	1.0	May 2001	Permissive	?	Copyleft	?	?	?	?
Cryptix General License	Cryptix Foundation	N/A	1995	Permissive	Permissive	Permissive	Manually	Yes	?	Manually
Eclipse Public License	Eclipse Foundation	1.0	February 2004	Limited[17]	Limited[17]	Limited[17]	Yes[17]	Yes[17]	Limited[17]	Manually[17]
Educational Community License	Indiana University[18]	1.0	2007	Permissive	?	Permissive	?	?	?	?
European Union Public Licence	European Commission	1.2	May 2017	Copylefted, with an explicit compatibility list[19]	Copylefted, with an explicit compatibility list[19]	Copylefted, with an explicit compatibility list[19]	Yes[20]	Yes[20]	Copylefted, with an explicit compatibility list[19]	No[20]
GNU Affero General Public License	Free Software Foundation	3.0	2007	GNU GPLv3 only[21]	Copylefted[22]	Copylefted[22]	Yes[23]	Copylefted[23]	Copylefted[22]	Yes[23]
GNU General Public License	Free Software Foundation	3.0	June 2007	GPLv3 compatible only[24][25]	Copylefted[22]	Copylefted[22]	Yes[26]	Yes[26]	Copylefted[22]	Yes[26]
GNU Lesser General Public License	Free Software Foundation	3.0	June 2007	With restrictions[27]	Copylefted[22]	Copylefted[22]	Yes[28]	Yes	Copylefted[22]	Yes[28]
IBM Public License	IBM	1.0	August 1999	Copylefted	?	Copylefted	?	?	?	?
ISC license	Internet Systems Consortium	N/A	June 2003	Permissive	Permissive	Permissive	?	?	?	?
LaTeX Project Public License	LaTeX project	1.3c	?	Permissive	?	Permissive	?	?	?	?
Microsoft Public License	Microsoft	N/A	?	Permissive	Permissive	Permissive	No	Permissive	?	No
MIT license / X11 license	MIT	N/A	1988	Permissive[29]	Permissive[29]	Permissive[29]	Manually[29]	Yes[29]	Permissive[29]	Manually[29]
Mozilla Public License	Mozilla Foundation	2.0	January 3, 2012	Permissive[30]	Copylefted[30]	Copylefted[30]	Yes[30]	Yes[30]	Copylefted[30]	No[30]
Netscape Public License	Netscape	1.1	?	Limited	?	Limited	?	?	?	?
Open Software License[10]	Lawrence Rosen	3.0	2005	Permissive	Copylefted	Copylefted	Yes	Yes	Copylefted	?
OpenSSL license	OpenSSL Project	N/A	?	Permissive	?	Permissive	?	?	?	?
Python Software Foundation License	Python Software Foundation	2	?	Permissive	?	Permissive	?	?	?	?
Q Public License	Trolltech	?	?	Limited	?	Limited	?	?	?	?
Sleepycat License	Sleepycat Software	N/A	1996	Permissive	With restrictions	Permissive	No	Yes	No	No
Unlicense	unlicense.org	1	December 2010	Permissive/Public domain	Permissive/Public domain	Permissive/Public domain	?	Permissive/Public domain	Permissive/Public domain	?
W3C Software Notice and License	W3C	20021231	December 31, 2002	Permissive	?	Permissive	?	?	?	?
Do What The Fuck You Want To Public License (WTFPL)	Banlu Kemiyatorn, Sam Hocevar	2	December 2004	Permissive/Public domain	Permissive/Public domain	Permissive/Public domain	No	Yes	Yes	No
XCore Open Source License also separate "Hardware License Agreement"	XMOS	?	February 2011	Permissive	Permissive	Permissive	Manually	Yes	Permissive	?
XFree86 1.1 License	The XFree86 Project, Inc	?	?	Permissive	?	Permissive	?	?	?	?
zlib/libpng license	Jean-Loup Gailly and Mark Adler	?	?	Permissive	?	Permissive	?	?	?	?

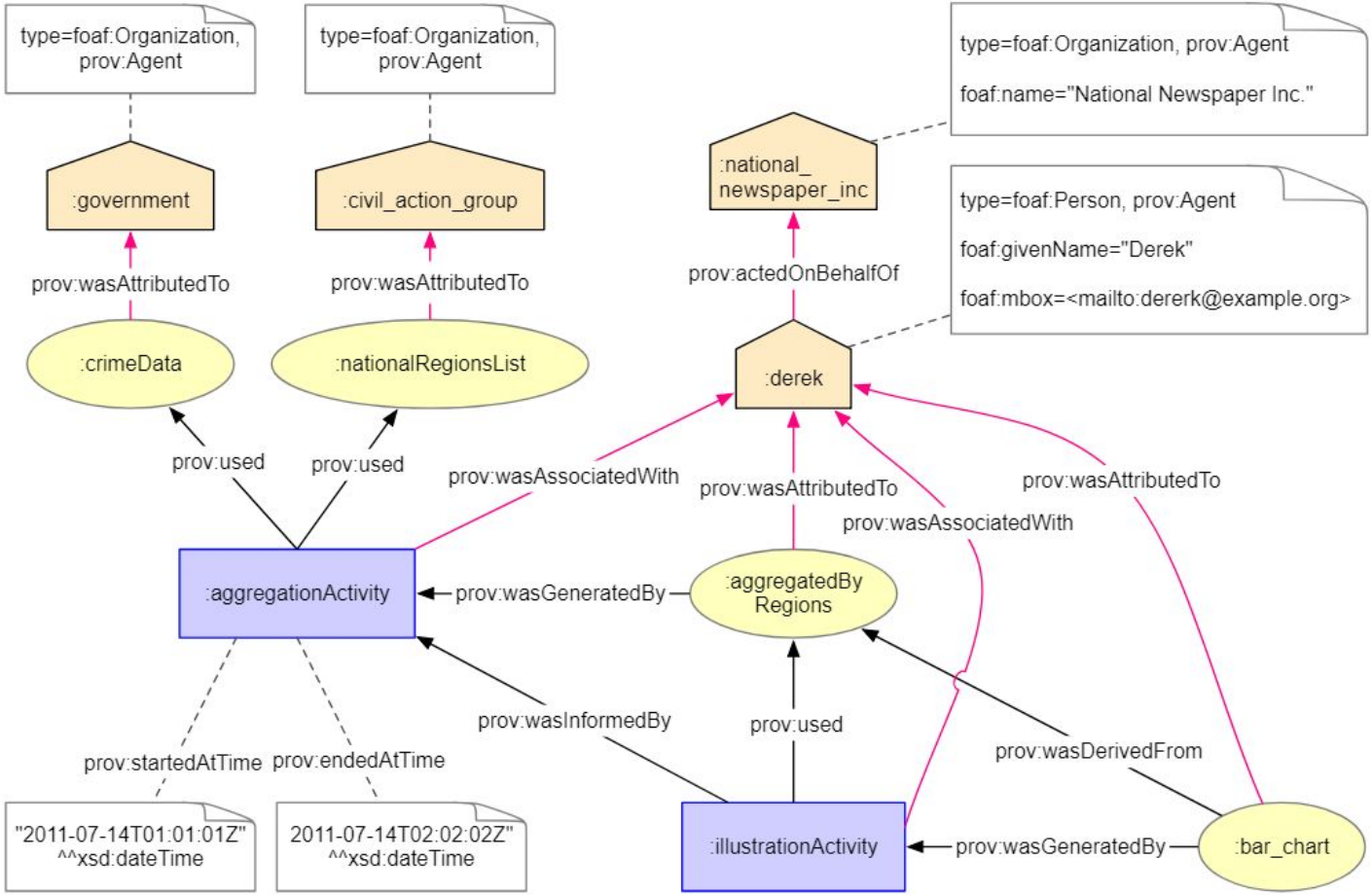
R1.1 metrics

1) A URL to the (machine readable) license/terms of use for the digital resource
2) A URL to the (machine readable) license/terms of use for the metadata

R1.2: (meta)data are associated with detailed provenance

Provenance is a trail of information about how an item came into existence and how it was handled since then. Detailed provenance can help people and machines assess whether a resource meets a reuse criteria. Resource metadata should provide detailed provenance. Provenance vocabularies and data models include PROV and DCAT.

Some **questions** to answer: Where did the digital resource come from? How has it been processed? Who to cite if you reuse the data? How the author wishes to be acknowledged? Does it contain content from someone else? Ideally the workflow is described in a machine-readable format.



https://www.w3.org/TR/2013/REC-prov-o-20130430/

R1.2 Metrics

FAIRsharing.org
standards, databases, policies

1) Provide the URL of the vocabulary used to describe the provenance of the digital resource.

R1.3: (meta)data meet domain-relevant community standards

It is easier to reuse digital resources when they are made available with a standardized format that conforms to a data model with lots of tooling support. Data should be easy to search and aggregate and reuse shared vocabularies.

Community standards or best practices should be followed where they exist, unless there is good reason not to do so. Many communities have developed minimal reporting standards (MIAME, MIAPE, etc.) while others have standardized extensive data models that are machine readable and can be automatically validated.

FAIRsharing.org
standards, databases, policies

R1.3 examples

FAIRsharing.org
standards, databases, policies

http://schema.datacite.org [for general purpose, not domain-specific]
http://dublincore.org/specifications [for general purpose, not domain-specific]
https://www.ncbi.nlm.nih.gov/geo/info/MIAME.html

http://cds.u-strasbg.fr/doc/catstd.htx [astrophysics]
https://www.iso.org/standard/53798.html [geographic information and services]
http://cfconventions.org [climate and forecast]
http://www.iucr.org/resources/cif [crystallographic information]
http://www.nexusformat.org [neutron, x-ray, and muon experiment data]
http://www.ddialliance.org/Specification [social, behavioral, and economic sciences]
https://sdmx.org [statistical data]
https://knb.ecoinformatics.org/#tools/eml [ecology]

R1.3 Metrics

1) A URL to a result of automated validation for the digital resource
2) A certificate of compliance to the standard by an independent party of a validation to the community standard

Summary

Lots to think about!

Metrics require published evidence of action

Metrics will change in the future, towards automated machine processing

FAIR will drive standardization and interoperability across communities

The end result is digital resources that are easier to find and reuse.