FAIR Metrics ALL

Mark D. Wilkinson, Susanna-Assunta Sansone,
Erik Schultes, Peter Doorn,
Luiz Olavo Bonino da Silva Santos, Michel Dumontier

January 11, 2018

FIELD	*DESCRIPTION*
Metric Identifier	FM-F1A: `https://purl.org/fair-metrics/FM_F1A`
Metric Name	Identifier Uniqueness
To which principle does it apply?	F1
What is being measured?	Whether there is a scheme to uniquely identify the digital resource.
Why should we measure it?	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. Examples of identifier schemes include, but are not limited to URN, IRI, DOI, Handle, trustyURI, LSID, etc. For an in-depth understanding of the issues around identifiers, please see http://dx.plos.org/10.1371/journal.pbio.2001414
What must be provided?	URL to a registered identifier scheme.
How do we measure it?	An identifier scheme is valid if and only if it is described in a repository that can register and present such identifier schemes (e.g. fairsharing.org). Information about the identifier scheme must be presented with a machine-readable document containing the FM1 attribute with the URL to where the scheme is described. see specification for implementation.
What is a valid result?	Present or Absent
For which digital resource(s) is this relevant?	All

FIELD	*DESCRIPTION*
Metric Identifier	FM-F1B: `https://purl.org/fair-metrics/FM_F1B`
Metric Name	Identifier persistence
To which principle does it apply?	F1
What is being measured?	Whether there is a policy that describes what the provider will do in the event an identifier scheme becomes deprecated.
Why should we measure it?	The change to an identifier scheme will have widespread implications for resource lookup, linking, and data sharing. Providers of digital resources must ensure that they have a policy to manage changes in their identifier scheme, with a specific emphasis on maintaining/redirecting previously generated identifiers.
What must be provided?	A URL that resolves to a document containing the relevant policy.
How do we measure it?	Use an HTTP GET on URL provided.
What is a valid result?	Present (a 200,202,203 or 206 HTTP response after resolving all and any prior redirects. e.g. 301 -> 302 -> 200 OK.) or Absent (any other HTTP code)
For which digital resource(s) is this relevant?	All
Comments	A first version of this metric would focus on just checking a URL that resolves to a document. We can't verify that document. A second version would indicate how to structure the data policy document with a particular section (similar to how the CC licenses now have a formal structure in RDF). A third version would insist that that document and section is signed by an approved organization and made available in an appropriate repository.

FIELD	*DESCRIPTION*
Metric Identifier	FM-F2: `https://purl.org/fair-metrics/FM_F2`
Metric Name	Machine-readability of metadata
To which principle does it apply?	F2 - Data are described with rich metadata
What is being measured?	The availability of machine-readable metadata that describes a digital resource.
Why should we measure it?	Richness of metadata can refer to many different aspects. One aspect is that the machine readability of metadata makes it possible to optimize their discovery. For instance, Web search engines suggest the use of particular structured metadata elements to optimize search. Thus, the machine-readability aspect can help people and machines find a digital resource of interest. Here, we focus on metadata being sufficiently rich in this sense - that the metadata document and the metadata elements are machine readable. Otherwise, it will also be difficult to understand what the digital resource is and what information is being provided about it.
What must be provided?	A URL to a document that contains machine-readable metadata for the digital resource. Furthermore, the file format must be specified.
How do we measure it?	HTTP GET on the metadata URL. A response of [a 200,202,203 or 206 HTTP response after resolving all and any prior redirects. e.g. 301 -> 302 -> 200 OK.] indicates that there is indeed a document. The second URL should resolve to the record of a registered file format (e.g. DCAT, DICOM, schema.org etc.) in a registry like FAIRsharing.
What is a valid result?	Machine-readable or Machine-not-readable
For which digital resource(s) is this relevant?	All

Comments	A first version of this metric would focus on just checking a URL that resolves to a document. We can't verify that document. A second version would indicate how to structure the data policy document with a particular section (similar to how the CC licenses now have a formal structure in RDF). A third version would insist that that document and section is signed by an approved organization and made available in an appropriate repository.

FIELD	*DESCRIPTION*
Metric Identifier	FM-F3: `https://purl.org/fair-metrics/FM_F3`
Metric Name	Resource Identifier in Metadata
To which principle does it apply?	F3 - metadata clearly and explicitly include the identifier of the data it describes
What is being measured?	Whether the metadata document contains the globally unique and persistent identifier for the digital resource.
Why should we measure it?	The discovery of digital object should be possible from its metadata. For this to happen, the metadata must explicitly contain the identifier for the digital resource it describes. A metadata document should also not result in ambiguity about the digital object it is describing. This can be assured if the metadata document explicitly refers to the digital object by its IRI.
What must be provided?	The URL of the metadata and the IRI of the digital resource it describes.
How do we measure it?	Parsing the metadata for the given digital resource IRI.
What is a valid result?	Present or absent
For which digital resource(s) is this relevant?	All
Examples of their application across types of digital resource	None
Comments	None

FIELD	*DESCRIPTION*
Metric Identifier	FM-F4: `https://purl.org/fair-metrics/FM_F4`
Metric Name	Indexed in a searchable resource
To which principle does it apply?	F4 - (meta)data are registered or indexed in a searchable resource
What is being measured?	The degree to which the digital resource can be found using web-based search engines.
Why should we measure it?	Most people use a search engine to initiate a search for a particular digital resource of interest. If the resource or its metadata are not indexed by web search engines, then this would substantially diminish an individual's ability to find and reuse it. Thus, the ability to discover the resource should be tested using i) its identifier, ii) other text-based metadata.
What must be provided?	The persistent identifier of the resource and one or more URLs that give search results of different search engines.
How do we measure it?	We perform an HTTP GET on the URLs provided and attempt to to find the persistent identifier in the page that is returned. A second step is to follow each of the top 10 hits and examine the resulting documents for presence of the identifier.
What is a valid result?	true -> the persistent identifier was found in the search results.
For which digital resource(s) is this relevant?	All

Examples of their application across types of digital resource	- my Zenodo Deposit for polyA (https://doi.org/10.5281/zenodo.47641) Test Query: 10.5281/zenodo.47641 orthology GOOGLE: Pass (#1 hit); BING: Fail (no hits); Yahoo: Fail (no hits); Baidu: Pass (#1 hit) Test Query: "protein domain orthology RNA Processing" Google: Pass (Hit #13); BING: Fail (not in top 40); Yahoo: Fail: (Not in top 40); Baidu: Pass (#1 Hit) - myExperiment Workflow (http://www.myexperiment.org/workflows/2969.html) Test Query: "workflow common identifiers EMC ontology" GOOGLE: Pass (#2 and #5 hit); BING: Fail (not in top 40, though OTHER workflows were found in top 10!); Yahoo: Fail (not in top 40, though other workflows found in top 10); Baidu: Pass (5/10 pages contained a link to the workflow, but the workflow itself was not discovered) - Jupyter notebook on GitHub (https://github.com/VidhyasreeRamu/GlobalClimateChange/blob/master/GlobalWarmingAnalysis.ipynb) Test Query: "github python climate change earth surface temperature" Google: Fail (not in top 40; other similar Jupyter notebooks found in github); Bing: Fail (not in top 40... but MANY links to Microsoft Surface! LOL!); Yahoo: Fail (not in top 40); Baidu: Fail (not even a github hit in top 40!)
Comments	None

FIELD	*DESCRIPTION*
Metric Identifier	FM-A1.1: `https://purl.org/fair-metrics/FM_A1.1`
Metric Name	Access Protocol
To which principle does it apply?	A1.1 - the protocol is open, free, and universally implementable
What is being measured?	The nature and use limitations of the access protocol.
Why should we measure it?	Access to a resource may be limited by the specified communication protocol. In particular, we are worried about access to technical specifications and any costs associated with implementing the protocol. Protocols that are closed source or that have royalties associated with them could prevent users from being able to obtain the resource.
What must be provided?	i) A URL to the description of the protocol ii) true/false as to whether the protocol is open source iii) true/false as to whether the protocol is (royalty) free
How do we measure it?	Do an HTTP get on the URL to see if it returns a valid document. Ideally, we would have a universal database of communication protocols from which we can check this URL. We also check whether questions 2 and 3 are true or false.
What is a valid result?	The HTTP GET on the URL should return a 200,202,203 or 206 HTTP response after resolving all and any prior redirects. e.g. 301 -> 302 -> 200 OK. The other two should be false
For which digital resource(s) is this relevant?	All
Examples of their application across types of digital resource	None
Comments	None

FIELD	*DESCRIPTION*
Metric Identifier	FM-A1.2: `https://purl.org/fair-metrics/FM_A1.2`
Metric Name	Access authorization
To which principle does it apply?	A1.2 - the protocol allows for an authentication and authorization procedure, where necessary
What is being measured?	Specification of a protocol to access restricted content.
Why should we measure it?	Not all content can be made available without restriction. For instance, access and distribution of personal health data may be restricted by law or by organizational policy. In such cases, it is important that the protocol by which such content can be accessed is fully specified. Ideally, electronic content can be obtained first by applying for access. Once the requester is formally authorized to access the content, they may receive it in some electronic means, for instance by obtaining an download URL, or through a more sophisticated transaction mechanism (e.g. authenticate, authorize), or by any other means. The goal should be to reduce the time it takes for valid requests to be fulfilled.
What must be provided?	i) true/false concerning whether authorization is needed ii) a description of the process to obtain access to restricted content.
How do we measure it?	computational validation of the data provided
What is a valid result?	a valid answer contains a true or false for the first question. if true, a non-empty textual description for the process is provided.
For which digital resource(s) is this relevant?	All
Examples of their application across types of digital resource	None
Comments	None

FIELD	*DESCRIPTION*
Metric Identifier	FM-A2: `https://purl.org/fair-metrics/FM_A2`
Metric Name	Metadata Longevity
To which principle does it apply?	A2 - metadata are accessible, even when the data are no longer available
What is being measured?	The existence of metadata even in the absence/removal of data
Why should we measure it?	Cross-references to data from third-party's FAIR data and metadata will naturally degrade over time, and become "stale links". In such cases, it is important for FAIR providers to continue to provide descriptors of what the data was to assist in the continued interpretation of those third-party data. As per FAIR Principle F3, this metadata remains discoverable, even in the absence of the data, because it contains an explicit reference to the IRI of the data.
What must be provided?	URL to a formal metadata longevity plan
How do we measure it?	Resolve the URL
What is a valid result?	- Successful resolution - Returns a document that represents a plan or policy of some kind - Preferably certified (e.g. DSA)
For which digital resource(s) is this relevant?	All metadata
Examples of their application across types of digital resource	None
Comments	None

FIELD	*DESCRIPTION*
Metric Identifier	FM-I1: `https://purl.org/fair-metrics/FM_I1`
Metric Name	Use a Knowledge Representation Language
To which principle does it apply?	I1 - (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
What is being measured?	use of a formal, accessible, shared, and broadly applicable language for knowledge representation.
Why should we measure it?	The unambiguous communication of knowledge and meaning (what symbols are, and how they relate to one another) necessitates the use of languages that are capable of representing these concepts in a machine-readable manner.
What must be provided?	URL to the specification of the language
How do we measure it?	- The language must have a BNF (or other specification language) - The URL resolves (accessible) - The document has an IANA media-type (i.e. it is sufficiently widely-accepted and shared that it has been registered) - The language can be arbitrarily extended (e.g. PDBml can be used to represent knowledge, but only about proteins)
What is a valid result?	BNF found
For which digital resource(s) is this relevant?	All
Examples of their application across types of digital resource	None

Comments	michel: there must be a syntax and associated semantics for that language. This is sufficient mark: there needs to be some identity or denotation in the language; ('vanilla') xml and json are not FAIR, so should fail this test *** can you (i) identify elements and (ii) make statements about them, and iii) is there a formally defined interpretation for that -> HTML fails; PDF fails shared -> that there are many users of the language . acknowledged within your community -> hard to prove. . could we use google to query for your filetype (can't discriminate between different models) -> has a media type –> This SHOULD be stated as a IANA code [IANA-MT] standardization of at least this listing process is a good measure of "sharedness" broadly applicable . that the language is extensible to a domain of interest . you can define your own elements in accordance with the semantics of the language gff3 is not in the IANA list -> what steps would the community need to execute to be listed here? cases like GFF, PDB are not broadly applicable biopax -> is defined vnd.biopax.rdf+xml and built on rdf -> allows users to create new elements and relate them jpg -> widely used, registered, but primarily for image content pdf -> registered, enables users to create their own dictionary.

FIELD	*DESCRIPTION*
Metric Identifier	FM-I2: `https://purl.org/fair-metrics/FM_I2`
Metric Name	Use FAIR Vocabularies
To which principle does it apply?	I2 - (meta)data use vocabularies that follow FAIR principles
What is being measured?	The metadata values and qualified relations should themselves be FAIR, for example, terms from open, community-accepted vocabularies published in an appropriate knowledge-exchange format.
Why should we measure it?	It is not possible to unambiguously interpret metadata represented as simple keywords or other non-qualified symbols. For interoperability, it must be possible to identify data that can be integrated like-with-like. This requires that the data, and the provenance descriptors of the data, should (where reasonable) use vocabularies and terminologies that are, themselves, FAIR.
What must be provided?	IRIs representing the vocabularies used for (meta)data
How do we measure it?	Resolve UUIDs, check FAIRness of the returned document(s)
What is a valid result?	Successful resolution; document is amenable to machine-parsing and identification of terms within it.
For which digital resource(s) is this relevant?	All
Examples of their application across types of digital resource	None
Comments	None

FIELD	*DESCRIPTION*
Metric Identifier	FM-I3: `https://purl.org/fair-metrics/FM_I3`
Metric Name	Use Qualified References
To which principle does it apply?	I3 - (meta)data include qualified references to other (meta)data
What is being measured?	Relationships within (meta)data, and between local and third-party data, have explicit and 'useful' semantic meaning
Why should we measure it?	One of the reasons that HTML is not suitable for machine-readable knowledge representation is that the hyperlinks between one document and another do not explain the nature of the relationship - it is "unqualified". For Interoperability, the relationships within and between data must be more semantically rich than "is (somehow) related to". Numerous ontologies include richer relationships that can be used for this purpose, at various levels of domain-specificity. For example, the use of skos for terminologies (e.g. exact matches), or the use of SIO for genomics (e.g. "has phenotype" for the relationship between a variant and its phenotypic consequences). Similarly, dbxrefs must be predicated with a meaningful relationship what is the nature of the cross-reference? Finally, data silos thwart interoperability. Thus, we should reasonably expect that some of the references/relations point outwards to other resources, owned by third-parties; this is one of the requirements for 5 star linked data.
What must be provided?	Linksets (in the formal sense) representing part or all of your resource
How do we measure it?	The linksets must have qualified references At least one of the links must be in a different Web domain (or the equivalent of a different namespace for non-URI identifiers)

What is a valid result?	- References are qualified - Qualities are beyond “Xref” or “is related to” - One of the cross-references points outwards to a distinct namespace
For which digital resource(s) is this relevant?	All
Examples of their application across types of digital resource	None
Comments	

FIELD	*DESCRIPTION*
Metric Identifier	FM-R1.1: `https://purl.org/fair-metrics/FM_R1.1`
Metric Name	Accessible Usage License
To which principle does it apply?	R1.1 - (meta)data are released with a clear and accessible data usage license
What is being measured?	The existence of a license document, for BOTH (independently) the data and its associated metadata, and the ability to retrieve those documents
Why should we measure it?	A core aspect of data reusability is the ability to determine, unambiguously and with relative ease, the conditions under which you are allowed to reuse the (meta)data. Thus, FAIR data providers must make these terms openly available. This applies both to data (e.g. for the purpose of third-party integration with other data) and for metadata (e.g. for the purpose of third-party indexing or other administrative metrics)
What must be provided?	IRI of the license (e.g. its URL) for the data license and for the metadata license
How do we measure it?	Resolve the IRI(s) using its associated resolution protocol
What is a valid result?	A document containing the license information
For which digital resource(s) is this relevant?	All
Examples of their application across types of digital resource	None
Comments	

FIELD	*DESCRIPTION*
Metric Identifier	FM-R1.2: `https://purl.org/fair-metrics/FM_R1.2`
Metric Name	Detailed Provenance
To which principle does it apply?	R1.2 - (meta)data are associated with detailed provenance
What is being measured?	That there is provenance information associated with the data, covering at least two primary types of provenance information: - Who/what/When produced the data (i.e. for citation) - Why/How was the data produced (i.e. to understand context and relevance of the data)
Why should we measure it?	Reusability is not only a technical issue; data can be discovered, retrieved, and even be machine-readable, but still not be reusable in any rational way. Reusability goes beyond "can I reuse this data?" to other important questions such as "may I reuse this data?", "should I reuse this data", and "who should I credit if I decide to use it?"
What must be provided?	Two URLs (IRIs). One of these URLs points to one of the vocabularies used to describe citational provenance (e.g. dublin core). The second points to one of the vocabularies (likely domain-specific) that is used to describe contextual provenance (e.g. EDAM)
How do we measure it?	We resolve the URLs/IRIs according to their associated protocols.
What is a valid result?	IRI 1 should resolve to a recognized citation provenance standard such as Dublin Core. IRI 2 should resolve to some vocabulary that itself passes basic tests of FAIRness
For which digital resource(s) is this relevant?	All
Examples of their application across types of digital resource	None

Comments	Many data formats have fields specifically for Provenance information. -> could fairsharing curate these 4 fields? for every format and vocabulary? Some formats do not have these fields. For example, although gff can have arbitrary headers, the standard itself does not provide specific fields to capture detailed provenance. It therefore would

FIELD	*DESCRIPTION*
Metric Identifier	FM-R1.3: `https://purl.org/fair-metrics/FM_R1.3`
Metric Name	Meets Community Standards
To which principle does it apply?	R1.3 - (meta)data meet domain-relevant community standards
What is being measured?	Certification, from a recognized body, of the resource meeting community standards.
Why should we measure it?	Various communities have recognized that maximizing the usability of their data requires them to adopt a set of guidelines for metadata (often in the form of "minimal information about..." models). Non-compliance with these standards will often render a dataset 'reuseless' because critical information about its context or provenance is missing. However, adherence to community standards does more than just improve reusability of the data. The software used by the community for analysis and visualization often depends on the (meta)data having certain fields; thus, non-compliance with standards may result in the data being unreadable by its associated tools. As such, data should be (individually) certified as being compliant, likely through some automated process (e.g. submitting the data to the community's online validation service)
What must be provided?	A certification saying that the resource is compliant
How do we measure it?	Validate the electronic signature as coming from a community authority (e.g. a verisign signature)
What is a valid result?	Successful signature validation
For which digital resource(s) is this relevant?	All
Examples of their application across types of digital resource	None

Comments	Such certification services may not exist, but this principle serves to encourage the community to create both the standard(s) and the verification services for those standards. A potentially useful side-effect of this is that it might provide an opportunity for content-verification - e.g. the certification service provides a hash of the data, which can be used to validate that it has not been edited at a later date.