Features of a FAIR Vocabulary

Fuqi Xu[1,*][0000−0002−5923−3859], Nick Juty[2,*][0000−0002−2036−8350], Carole Goble[2][0000−0003−1219−2137], Simon Jupp[3][0000−0002−0643−3144], Helen Parkinson[1][0000−0003−3035−4195], and Mélanie Courtot[1,†][0000−0002−9551−6370]

[1] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom
mcourtot@gmail.com
[2] University of Manchester, , Manchester M13 9PL, United Kingdom
[3] SciBite, BioData Innovation Centre, Wellcome Genome Campus, Hinxton, Cambridge CB10 1DR, United Kingdom

Abstract. The FAIR Principles explicitly require the use of FAIR vocabularies, but what precisely constitutes a FAIR vocabulary remains unclear. Here we provide definitions for FAIR vocabularies, examine the application of the FAIR Principles to vocabularies, align their requirements with the Open Biomedical Ontologies (OBO) Principles, and propose FAIR Vocabulary Features (FVFs). We also design assessment approaches for FAIR vocabularies by mapping the FVFs with existing FAIR assessment indicators. Finally, we demonstrate how FVFs can be used for evaluating and improving vocabularies using exemplar biomedical vocabularies.

Keywords: FAIR principles · Vocabulary · Ontology · Assessment

1 Introduction

The Findable, Accessible, Interoperable and Reusable (FAIR) Principles [40] have gained traction in the biomedical community since their publication in 2016, with many groups attempting to improve their data quality, develop FAIR capable data resources, and design generic FAIR assessment tools for biomedical data[6] [16] [39]. Due to the heterogeneous nature and broad scope of biomedical data, from molecules to human studies via interdisciplinary analysis, stringent requirements for its FAIRness need to be met to ensure its usefulness toward benefiting human health. While assessing the FAIR level of datasets and data resources [11], we noted a futile cycle with respect to the 'Interoperable' FAIR Principle, *"I2 - (Meta)data use vocabularies that follow FAIR principles"*. To comply with that principle, datasets need to use FAIR vocabularies, which themselves need to be FAIR. FAIR vocabularies promote the exchange of biomedical

*These authors contributed equally to this paper.
†To whom correspondence should be addressed.

data, which are usually generated, annotated and used by different groups of researchers. FAIR vocabularies promote biomedical data FAIRness throughout the data life cycle, during the data generation, curation, and distribution processes, and support data exchange and integration across data resources.

Multiple efforts have been made to develop standards for FAIR vocabularies. The FAIRsFAIR recommendations provide guidance[24] on FAIR semantic artefacts, as well as supporting vocabulary search engines and repositories. Garijo and Poveda-Villalon[22] discussed detailed requirements of ontology URI and versioning strategies, as well as the formatting of the ontologies. Ten simple rules[15] for converting print-based or other forms of legacy vocabularies to FAIR vocabularies have also been proposed.

Researchers have also developed approaches to assess the FAIR level of digital objects both manually and automatically; FAIRsharing hosts FAIR indicators for automated tests in their FAIR Maturity Evaluation Service[39], FAIR metrics in the F-UJI Automated FAIR Data Assessment Tool[4], and the Research Data Alliance(RDA) Data Maturity Model Specification and Guidelines[5], also known as the RDA indicators. Among them, the RDA indicators are a set of representative and descriptive indicators to evaluate the FAIR level of data and have been used in many projects and with many types of data. Some automated assessments of the FAIRness of vocabulary have been developed to measure the FAIR level of public, machine-readable vocabularies, such as FOOPS![21] To the best of our knowledge, there have not yet been quantifiable FAIR assessment approaches developed to measure the FAIR level of different formats of vocabularies objectively.

Therefore, in this paper, we distinguish the concepts of FAIR data, FAIR data resources, and FAIR vocabularies, and propose a set of general FAIR Vocabulary Features (FVFs) as a set of satisfiable features for vocabularies. We also adapted the RDA indicators to measure the FAIR level of vocabularies. Further, we provide example assessments based on selected ontologies available from the EMBL-EBI Ontology Lookup Service (OLS)[25] and other vocabulary resources.

2 FAIR Data and FAIR Vocabulary

In the execution of this work, we note the distinction between FAIR data and FAIR capable data resources. In our analysis, data can be FAIR, to a greater or lesser extent, and data resources and data vocabularies are capable of supporting FAIRness (FAIR capable) at different levels. Data vocabularies are designed to support FAIR data, and they can also be considered as FAIR data resources. The orthogonality of these concepts is an important context for this work when determining the features of a FAIR vocabulary. We must also determine whether a vocabulary itself is 1) FAIR in terms of its application to FAIR data 2) FAIR in the context of FAIR capable resources 3) FAIR in the context of other vocabularies. A FAIR vocabulary has a set of FAIR features and have a list of associated FAIR indicators. It is usable for annotation, analysis and presentation of data, and is deployable in the context of a FAIR capable data resource or tools. It also

serves 'aggregation' use cases where data originates from different domain, and enables data interoperability where different vocabularies are used.

Vocabularies come in different forms, such as lists, thesaurus, taxonomies, and ontologies; each at different levels of semantic maturity and FAIR requirements. The International Classification of Diseases (ICD-11)[31] is a large taxonomy of disease and is the global standard for diagnostic information, disease definitions and synonyms. The Gene Ontology (GO)[12] is a well established and highly regarded and utilised biomedical resource. It contains over 43000 terms and has been cross referenced in other classification systems, such as UniProt[13], HAMAP[32], and InterPro[9]. GO is also a reference OBO Foundry ontology[34] and has been reused in many other resources. The Experimental Factor Ontology (EFO)[27], on the other hand, is an application ontology built for communities like the Open Targets[29] for describing experimental variables.

3 Existing Vocabulary Standards

In determining features of FAIR vocabularies, we considered previous standardisation work by the Open Biomedical Ontology (OBO) community to determine whether the OBO Principles[18] addressed elements of vocabulary FAIRness. The OBO Principles aim to coordinate the development of biomedical ontologies, which focus specifically on ontologies, covering both the development of ontologies and ontology themselves. Despite the OBO principles predating the FAIR principles, a comparison of the two aided us in defining FVFs. Table 1 summarises the key points of the OBO Foundry principles, and assesses their suitability as FVFs. We also noted that not all the OBO principles possess the same level of maturity or granularity, and therefore some were unmappable and excluded from the comparison. As a result, this analysis did not include OBO Principle 1, 4, 6, 9 - 12 and 20. The rationale for suitability as FVFs is discussed in detail below and further in Supplemental Table 1.

4 FAIR Vocabulary Features

Based on the analysis of OBO foundry practices and our previous experience working with and developing ontologies, we propose eleven features for FAIR vocabulary in Table 2, covering requirements for identifiers, access protocols, knowledge representation, etc. Supplementary Table 2 shows the relationship among FAIR Vocabulary Features, the FAIR principles and requirements for FAIR vocabularies.

Table 2also provides examples for each FAIR feature, but does not exhaustively cover all current practices across the various vocabularies; each feature is represented in different formats and at varying FAIRness levels amongst those vocabularies. For example, for *FVF-6: versioning and persistent vocabularies*, of all ontologies indexed and updated in OLS, 59.3%[§] of vocabularies use a date format of *"yyyy-mm-dd"* in the *"versionIRI"*, such as *"http://purl.obolibrary.org/*

[§]See details in Supplementary Material 3

Table 1. An analysis of the OBO Foundry principles as putative FAIR Vocabulary Features

ID	**OBO Principle Summary**	**Suitable as FAIR Vocabulary Feature?**
Principle 1: Open	The ontology MUST be openly available to be used by all without any constraint other than (a) its origin must be acknowledged and (b) it is not to be altered and subsequently redistributed in altered form under the original name or with the same identifiers.	No
Principle 2: Common Format	The ontology is made available in a common formal language in an accepted concrete syntax.	Yes
Principle 3: URI/Identifier Space	Each class and relation (property) in the ontology must have a unique URI identifier.	Yes
Principle 4: Versioning	The ontology provider has documented procedures for versioning the ontology, and different versions of ontology are marked, stored, and officially released.	No
Principle 5: Scope	The scope of an ontology is the extent of the domain or subject matter it intends to cover. The ontology must have a clearly specified scope and content that adheres to that scope.	Yes
Principle 6: Textual Definitions	The ontology has textual definitions for the majority of its classes and for top level terms in particular	No
Principle 7: Relations	Relations should be reused from the Relations Ontology (RO).	Yes
Principle 8: Documentation	The owners of the ontology should strive to provide as much documentation as possible. The documentation should detail the different processes specific to an ontology life cycle and target various audiences (users or developers).	Yes
Principle 9: Documented Plurality of Users	The ontology developers should document that the ontology is used by multiple independent people or organizations.	No
Principle 10: Commitment to Collaboration	OBO Foundry ontology development, in common with many other standards-oriented scientific activities, should be carried out in a collaborative fashion.	Yes
Principle 11: Locus of Authority	There should be a person who is responsible for communications between the community and the ontology developers, for communicating with the Foundry on all Foundry-related matters, for mediating discussions involving maintenance in the light of scientific advance, and for ensuring that all user feedback is addressed.	No
Principe 12: Naming Conventions	Naming conventions are used	No
Principle 16: Maintenance	The ontology needs to reflect changes in scientific consensus to remain accurate over time.	Yes
Principle 20: Responsiveness	Ontology developers MUST offer channels for community participation and SHOULD be responsive to requests.	No

Table 2. FAIR Vocabulary Feature details

ID	Features	Description	Examples
FVF-1	Vocabulary and constituted terms are assigned globally unique and persistent identifiers.	Vocabulary itself and its constituent terms should have identifiers that are globally unique and persistent to ensure that each item can be identified unambiguously over time.	Examples of globally unique and persistent identifiers are PURL[19], identifiers.org[41], and w3id.org[37]. The OBO foundry provides identifier policy[2] for biomedical ontologies and requires using PURLs as with standard prefixes, such as http://purl.obolibrary.org/obo/GO_0000022.
FVF-2	Vocabulary and constituted terms have rich metadata.	Vocabulary itself and its constituent terms should have sufficient metadata to support discovery by both humans and machines.	Metadata of the vocabulary should provide information about the creation date, creator and editor, version, licence, target domain and short descriptions. Metadata of its terms should describe term editing history, definition source, and other metadata.
FVF-3	Vocabulary and constituted terms can be accessed using the identifiers, preferably by both humans and machines.	The URIs of vocabulary itself and its constituent terms can be dereferenced by both humans and machines.	http://www.ebi.ac.uk/efo/EFO_0000311 resolves to term "Cancer" in the Experimental Factor Ontology, which can be accessed by both humans using ontology browsers and machines through the OLS API.
FVF-4	Vocabulary and constituted terms are registered or indexed in a searchable engine or a resource.	Vocabulary itself and its constituent terms are registered in vocabulary archives or other vocabulary management systems and indexed by local or/and global search engines.	EMBL-EBI Ontology Lookup Service and NCBI BioPortal[38]are two popular public vocabulary archives. Property *X-Robots-Tag:index* in vocabularies allows them to be indexed by search engines.
FVF-5	Vocabulary and constituted terms are retrievable using a standardised communications protocol, preferably open, free and universally implementable protocols, such as HTTPS, HTTP, FTP. The protocol should also allow identifying the accessor and grant access based on the accessor privilege, when necessary.	Vocabulary itself and its constituent terms are retrievable using a standardised communications protocol, preferably open, free and universally implementable protocols, such as HTTPS, HTTP, FTP. The protocol should also allow identifying the accessor and grant access based on the accessor privilege, when necessary.	Most public ontologies can be accessed using the HTTP or HTTPS protocol. For example, EFO uses HTTP protocol. The Unified Medical Language System[10] uses HTTPS protocol and only allows access by authenticated users.
FVF-6	Vocabulary and constituted terms are persistent over time and are appropriately versioned.	Changes in the vocabulary are reflected in different versions. Vocabularies and their terms are versioned, and each unaltered version of the vocabulary can be identified and retrieved in perpetuity. Vocabulary metadata is available even when the vocabulary is no longer available.	Changes in EFO are included in each release and identified with versioned IRI, such as, http://www.ebi.ac.uk/efo/releases/v3.31.0/efo.owl, which resolves to the versioned vocabulary. OBO foundry also provides guidelines[20] for ontology versioning and how different versions of the vocabularies should be labelled, stored and published.
FVF-7	Vocabulary and constituted terms use a formal, accessible and broadly applicable, and preferably machine-understandable language for knowledge representation.	Vocabulary itself and its constituent terms use a formal, accessible and broadly applicable, and preferably machine-understandable language for knowledge representation.	OWL-based vocabularies can be serialised using RDF-XML, or relational databases e.g. ChEBI[23] can be converted into OWL[7]
FVF-8	Vocabulary and constituted terms use qualified references to other vocabularies.	Vocabulary reuse terms from other vocabularies when applicable, provide adequate metadata about external terms, and follow vocabulary cross-reference standards.	EFO reuses human anatomy terms such as "liver" from UBERON[28] (UBERON_0002107) and linked to the original UBERON term. Property *Xref* indicates a cross-reference relationship between two vocabulary terms. MIREOT[14] defines a methodology and minimum information requirements for importing external terms into an extant ontology.
FVF-9	Vocabulary and constituted terms are described with a plurality of accurate and relevant attributes.	Vocabulary terms include sufficient attributes, such as labels, synonyms, definitions, examples of usage, and cross-references, to support the interpretation and reuse of the vocabulary terms.	The OBO flat-file format specification[17] provides a list of recommended mandatory and optional attributes. Each vocabulary term must have an ID and a name. The recommended attributes include definition, synonym, Xref, relationship, and etc.
FVF-10	Vocabularies are released with a standard data usage licence, preferably a machine-readable licence.	The vocabulary includes information about how the vocabulary can be reused.	Common public data usage licences are CC-BY[1] and MIT[3]. For example, Gene Ontology uses Creative Commons Attribution 4.0 Unported License. SNOMED CT™ affiliate license agreement.
FVF-11	Vocabularies meet domain-relevant community standards.	Vocabularies cover essential terms for the specific domain, reflect knowledge of this domain and can be used in existing data standards and data models.	Community standards, such as minimum information requirements and data models can be found in FAIRsharing[33]. The Plant Phenotyping Experiment Ontology (PPEO)[8] implements the Minimum Information about Plant Phenotyping Experiment(MIAPPE)[26]standards and covers essential attributes to describe a MIAPPE-compliant phenotype dataset.

-obo/scdo/releases/2021-04-15/scdo.owl". 2.51% of vocabularies use semantic versioning (x.x.x) such as *"http://www.ebi.ac.uk/efo/releases/v3.34.0/efo.owl"* or other forms of numeric versioning, such as *http://www.orpha.net/version3.2*. 31.66% of vocabularies do not provide valid machine-readable versioned IRIs. For *FVF-1: identifiers*, 74% of vocabularies use OBO-format PURLs, identifier.org, w3id.org identifiers, as well as other domain-specific identifiers. For *FVF-5: accessible using standard protocols*, of all 199 selected ontologies, only one ontology used the HTTPS protocol; the rest use HTTP protocols.

5 FAIR Vocabulary Feature Indicators

FAIR vocabulary Features outline general characteristics of a FAIR vocabulary, however, those features need to be objectively quantified to be useful in vocabulary selection, development and assessment. Hence, we propose aligning FVFs with FAIR indicators to enable computation of a discrete FAIR score, with the aim to offer an objective quantitative evaluation of vocabularies and to guide subsequent improvements.

We mapped the RDA indicators to FAIR Vocabulary Features, filtered out indicators that do not apply to vocabularies (see details in Supplementary Table 4, specified the digital object which the indicator refers to, and identified within each indicator the relevant standards used in corresponding domains. It is worth noting that when mapping the RDA indicators on datasets, *metadata* refers to the metadata to which the vocabulary can be applied, while in the context of vocabularies, *metadata* and *data* refer to the description of the vocabulary and the vocabulary information. Therefore, we combined the indicators evaluating data and metadata in the mapping, wherever possible. The FVFs, associated with selected indicators, can be used as indicators for FAIR Vocabulary as shown in Table 3.

6 Assessment against Indicators for FAIR Vocabulary Features

We tested the FAIR Vocabulary Features and corresponding indicators on three representative vocabularies, GO, EFO and ICD-11, as shown in Table 4. For each FVF, three compliance levels are assigned; if a vocabulary meets the requirements of all indicators, *full compliance* is achieved. Otherwise, depending on the scoring within each FVF, *partial compliance* or *no compliance* results are given. The percentages of *full compliance*, *partial compliance* and *no compliance* features are also calculated. Supplementary Table 5-7 provide the assessment details.

From the assessment results, both the Gene Ontology and Experimental Factor Ontology are vocabularies of high FAIR level, with over 80% FVFs fulfilled. The Gene Ontology only partially complies with '*FVF-6: Vocabularies and their terms are persistent over time and are appropriately versioned*', with a *Fail* in

Table 3. Indicators for FAIR Vocabulary Features. Alignment between the FAIR Vocabulary Features and RDA Data Maturity level indicators

FAIR vocabulary Feature	**RDA indicator ID**	**Indicator**
FVF-1: Vocabulary and their terms are assigned globally unique and persistent identifiers.	RDA-F1-01M	Metadata is identified by a persistent identifier
	RDA-F1-01D	Data is identified by a persistent identifier
	RDA-F1-02M	Metadata is identified by a globally unique identifier
	RDA-F1-02D	Data is identified by a globally unique identifier
FVF-2: Vocabularies and their terms have rich metadata.	RDA-F2-01M	Rich metadata is provided to allow discovery
FVF-3: Vocabularies and their terms can be accessed using the identifiers, preferably by both human and machine.	RDA-A1-01M	Metadata contains information to enable the user to get access to the data
	RDA-A1-02M	Metadata can be accessed manually (i.e. with human intervention)
	RDA-A1-02D	Data can be accessed manually (i.e. with human intervention)
	RDA-A1-03M	Metadata identifier resolves to a metadata record
	RDA-A1-03D	Data identifier resolves to a digital object
	RDA-A1-05D	Data can be accessed automatically (i.e. by a computer program)
FVF-4: Vocabularies and their terms are registered or indexed in a searchable engine or a resource.	RDA-F4-01M	Metadata is offered in such a way that it can be harvested and indexed
FVF-5: Vocabularies and their terms are retrievable using a standardised communications protocol, preferably open, free and universally implementable protocols. and allows for authentication and authorisation, where necessary.	RDA-A1-04M	Metadata is accessed through standardised protocol
	RDA-A1-04D	Data is accessible through standardised protocol
	RDA-A1.1-01M	Metadata is accessible through a free access protocol
	RDA-A1.1-01D	Data is accessible through a free access protocol
	RDA-A1.2-01D	Data is accessible through an access protocol that supports authentication and authorisation
FVF-6: Vocabularies and their terms are persistent over time and are appropriately versioned.	RDA-A2-01M	Metadata is guaranteed to remain available after data is no longer available
	RDA-R1.2-01M	Metadata includes provenance information according to community-specific standards
	RDA-R1.2-02M	Metadata includes provenance information according to a cross-community language
FVF-7: Vocabularies and their terms use a formal, accessible and broadly applicable, and preferably machine-understandable language for knowledge representation.	RDA-I1-01M	Metadata uses knowledge representation expressed in standardised format
	RDA-I1-01D	Data uses knowledge representation expressed in standardised format
	RDA-I1-02M	Metadata uses machine-understandable knowledge representation
	RDA-I1-02D	Data uses machine-understandable knowledge representation
FVF-8: Vocabularies and terms use qualified references to other vocabularies.	RDA-I3-02D	Data includes qualified references to other data
	RDA-I3-03M	Metadata includes qualified references to other metadata
FVF-9: Vocabularies and terms are described with a plurality of accurate and relevant attributes.	RDA-R1-01M	Plurality of accurate and relevant attributes are provided to allow reuse
FVF-10: Vocabularies are released with a standard data usage licence, preferably machine-readable licence.	RDA-R1.1-01M	Metadata includes information about the licence under which the data can be reused
	RDA-R1.1-02M	Metadata refers to a standard reuse licence
	RDA-R1.1-03M	Metadata refers to a machine-understandable reuse licence
FVF-11: Vocabularies meet domain relevant community standards.	RDA-R1.3-01M	Metadata complies with a community standard
	RDA-R1.3-01D	Data complies with a community standard
	RDA-R1.3-02M	Metadata is expressed in compliance with a machine-understandable community standard
	RDA-R1.3-02D	Data is expressed in compliance with a machine-understandable community standard

Table 4. FAIR vocabulary feature applied, assessment results of Gene ontology, Experimental factor Ontology and ICD-11

	Vocabulary		
FAIR vocabulary Feature	**Gene Ontology**	**Experimental Factor Ontology**	**ICD-11**
FVF-1: Vocabulary and their terms are assigned globally unique and persistent identifiers.	Full Compliance	Full Compliance	Partial Compliance
FVF-2: Vocabularies and their terms have rich metadata.	Full Compliance	No Compliance	Full Compliance
FVF-3: Vocabularies and their terms can be accessed using the identifiers, preferably by both human and machine.	Full Compliance	Full Compliance	Partial Compliance
FVF-4: Vocabularies and their terms are registered or indexed in a searchable engine or a resource.	Full Compliance	Full Compliance	No Compliance
FVF-5: Vocabularies and their terms are retrievable using a standardised communications protocol, preferably open, free and universally implementable protocols. and allows for authentication and authorisation, where necessary.	Full Compliance	Full Compliance	Full Compliance
FVF-6: Vocabularies and their terms are persistent over time and are appropriately versioned.	Partial Compliance	Partial Compliance	Partial Compliance
FVF-7: Vocabularies and their terms use a formal, accessible and broadly applicable, and preferably machine-understandable language for knowledge representation.	Full Compliance	Full Compliance	No Compliance
FVF-8: Vocabularies and terms use qualified references to other vocabularies.	Full Compliance	Full Compliance	Partial Compliance
FVF-9: Vocabularies and terms are described with a plurality of accurate and relevant attributes.	Full Compliance	Full Compliance	No Compliance
FVF-10: Vocabularies are released with a standard data usage licence, preferably machine-readable licence.	Full Compliance	Full Compliance	Full Compliance
FVF-11: Vocabularies meet domain relevant community standards.	Full Compliance	Full Compliance	No Compliance
FAIR Vocabulary Feature summary			
FVF, full compliance	90.91%	81.82%	27.27%
FVF, partial compliance	9.09%	9.09%	36.36%
FVF, no compliance	0.00%	9.09%	36.36%

'*Indicator RDA-R1.2-02M: Metadata includes provenance information according to a cross-community language*'. '*FVF-2: Vocabularies and their terms have rich metadata*' was not complied with since no general description of the ontology is provided in the released artefact. Compared with the ontologies, the taxonomy, ICD-11, fully complies with 18.18% FVFs, and partially complies with 36.36% FVFs. This is because ICD-11 neither refers to other vocabularies, nor adheres to other community standards, such as vocabulary formats. ICD-11 was selected for this evaluation as it already offers significant FAIR improvements over ICD-10[30], such as providing a standard licence.

7 Discussion

The FAIR Vocabulary Features we propose integrate multiple FAIR vocabulary requirements and can be used as FAIR vocabulary standards to guide the development and maintenance of vocabularies. Each FVF is associated with indicators enabling its quantifiable, objective assessment. Those indicators can be connected to related standards and extended to existing or emerging standards in other domains. For example, *FVF-8: cross-referencing* other vocabularies can be linked to the ontology cross-reference standards, MIREOT. We focused on how FVFs can be applied to ontologies, and demonstrated the potential for using them across other forms of vocabularies; for example, with the ICD-11, assessment would inform authors on the means to enrich their resources. Because of our expertise and requirements, this manuscript focuses on the biomedical domain; however, we anticipate this framework could be reused in other domains.

Integrating the FAIR vocabulary features with FAIR indicators makes it possible to assess the FAIR level of vocabularies, identify progressive ontology development use cases, and improve those vocabularies. We selected the RDA indicators as it has proven to be useful in many datasets and has been referenced by other assessment approaches in FAIRassist.org; yet, FVFs could alternatively be aligned to other FAIR-principle based indicators which would similarly reflect the guiding principles proposed by Wilkinson et al. Besides manual assessment, quantifiable formal indicators are also amenable to becoming machine actionable. Some efforts already exist, and reusing shared indicators will make it possible to perform automated FAIR vocabulary assessments.

The indicators were proposed to objectively measure the FAIR level of ontologies, yet this score does not reflect an absolute FAIR level for the vocabulary. Indeed, depending on the purpose and requirements of the vocabulary, some FVFs can be more or less important than other features. For example, for internal vocabularies which are used and shared within an institution, having global identifiers (FVF-1) is not a mandatory requirement. Instead of comparing the FAIR score of different vocabularies to find the 'FAIRer' one, we propose the FAIR score should be used to measure and guide the evolution of FAIR vocabularies by successively comparing the FAIR levels of iteratively developed versions. For example, compared to ICD-10, its successor, ICD-11 has incorpo-

rated many features to make it FAIRer, such as providing APIs for easier access, having a machine-readable license, etc.

From the assessment results of the two ontologies and ICD-11, ontology-based vocabularies follow stricter semantics and therefore fared better in the scoring of FAIR features. For example, many ontology-related standards have been established, including formats, such as OWL, guidelines such as the OBO principles, minimum information standards, such as MIBBI[36], and mechanisms for cross-references or incorporating external ontologies, such as MIREOT. This naturally reflects in a high score for compliance with community standards, which is a core part of FAIR Vocabulary Features and which improves the interoperability and reusability of a vocabulary.

The FAIR Vocabulary features and assessments provide insights on how to improve vocabularies. For example, based on the EFO assessments, the FAIR level of EFO could easily be improved by adding a description of the aim and function of EFO. This way, different vocabulary management services can harvest the information.

Acknowledgements

This work is funded by the IMI-FAIRplus project (Grant number 802750) and the European Molecular Biology Laboratory - European Bioinformatics Institute core funds.

8 References

[1] Creative commons — attribution 4.0 international — CC BY 4.0, https://creativecommons.org/licenses/by/4.0/
[2] ID policy, http://www.obofoundry.org/id-policy
[3] The MIT license | open source initiative, https://opensource.org/licenses/MIT
[4] F-UJI automated FAIR data assessment tool (2020), https://www.fairsfair.eu/f-uji-automated-fair-data-assessment-tool
[5] FAIR data maturity model: specification and guidelines - draft (2020), https://www.rd-alliance.org/group/fair-data-maturity-model-wg/outcomes/fair-data-maturity-model-specification-and-guidelines
[6] FAIRassist.org (2021), https://fairassist.org/#!/
[7] Antoniou, G., van Harmelen, F.: Web ontology language: OWL. pp. 67–92. International Handbooks on Information Systems, Springer (2004). https://doi.org/10.1007/978-3-540-24750-0_4
[8] Arnaud, E., Cooper, L., Shrestha, R., et al.: Towards a reference plant trait ontology for modeling knowledge of plant traits and phenotypes (2012), http://wrap.warwick.ac.uk/59831/
[9] Blum, M., Chang, H.Y., Chuguransky, S., et al.: The InterPro protein families and domains database: 20 years on **49**, D344–d354 (2021). https://doi.org/10.1093/nar/gkaa977

[10] Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology **32**, D267–270 (2004). https://doi.org/10.1093/nar/gkh061
[11] Burdett, T., Xu, F., Courtot, M., et al.: FAIRplus: D3.2 IMI FAIR metrics publication . https://doi.org/10.5281/zenodo.4428633
[12] Consortium, T.G.O.: The gene ontology resource: 20 years and still GOing strong **47**, D330–d338 (2019). https://doi.org/10.1093/nar/gky1055
[13] Consortium, T.U.: UniProt: the universal protein knowledgebase in 2021 **49**, D480–d489 (2021). https://doi.org/10.1093/nar/gkaa1100
[14] Courtot, M., Gibson, F., Lister, A., et al.: MIREOT: the minimum information to reference an external ontology term pp. 1–1 (2009). https://doi.org/10.1038/npre.2009.3576.1
[15] Cox, S.J.D., Gonzalez-Beltran, A.N., Magagna, B., et al.: Ten simple rules for making a vocabulary FAIR **17**(6), e1009041 (2021). https://doi.org/10.1371/journal.pcbi.1009041
[16] Drysdale, R., Cook, C.E., Petryszak, R., et al.: The ELIXIR core data resources: fundamental infrastructure for the life sciences **36**(8), 2636–2642 (2020). https://doi.org/10.1093/bioinformatics/btz959
[17] Foundry, T.O.: The OBO flat file format specification, version 1.2, `https://owlcollab.github.io/oboformat/doc/GO.format.obo-1%5F2.html`
[18] Foundry, T.O.: OBO foundry principles, overview, `http://www.obofoundry.org/principles/fp-000-summary.html`
[19] Foundry, T.O.: PURL administration, `https://purl.prod.archive.org/`
[20] Foundry, T.O.: Versioning (principle 4), `http://www.obofoundry.org/principles/fp-004-versioning.html`
[21] Garijo, D., Corcho, O., Poveda-Villalon, M.: FOOPS!: An ontology pitfall scanner for the FAIR principles p. 4 (2021), `http://ceur-ws.org/Vol-2980/paper321.pdf`
[22] Garijo, D., Poveda-Villalón, M.: Best practices for implementing FAIR vocabularies and ontologies on the web (2020), `http://arxiv.org/abs/2003.13084`
[23] Hastings, J., Owen, G., Dekker, A., et al.: ChEBI in 2016: Improved services and an expanding collection of metabolites **44**, D1214–1219 (2016). https://doi.org/10.1093/nar/gkv1031
[24] Hugo, W., Le Franc, Y., Coen, G., et al.: D2.5 FAIR semantics recommendations second iteration (2020). https://doi.org/10.5281/zenodo.4314321
[25] Jupp, S., Burdett, T., Malone, J., et al.: A new ontology lookup service at EMBL-EBI p. 2 (2015), `http://ceur-ws.org/Vol-1546/paper%5F29.pdf`
[26] Krajewski, P., Chen, D., Ćwiek, H., et al.: Towards recommendations for metadata and data handling in plant phenotyping **66**(18), 5417–5427 (2015). https://doi.org/10.1093/jxb/erv271
[27] Malone, J., Holloway, E., Adamusiak, T., et al.: Modeling sample variables with an experimental factor ontology **26**(8), 1112–1118 (2010). https://doi.org/10.1093/bioinformatics/btq099
[28] Mungall, C.J., Torniai, C., Gkoutos, G.V., et al.: Uberon, an integrative multi-species anatomy ontology **13**(1), R5 (2012). https://doi.org/10.1186/gb-2012-13-1-r5

[29] Ochoa, D., Hercules, A., Carmona, M., et al.: Open targets platform: supporting systematic drug–target identification and prioritisation **49**, D1302–d1310 (2021). https://doi.org/10.1093/nar/gkaa1027
[30] Organization, W.H.: International classification of diseases for mortality and morbidity statistics (10h revision) (2010), `https://www.who.int/classifications/icd/ICD10Volume2_en_2010.pdf`
[31] Organization, W.H.: International classification of diseases for mortality and morbidity statistics (11th revision) (2021), `https://icd.who.int/browse11/l-m/en`
[32] Pedruzzi, I., Rivoire, C., Auchincloss, A.H., et al.: HAMAP in 2015: updates to the protein family classification and annotation system **43**, D1064–d1070 (2015). https://doi.org/10.1093/nar/gku1002
[33] Sansone, S.A., McQuilton, P., Rocca-Serra, P., et al.: FAIRsharing as a community approach to standards, repositories and policies **37**(4), 358–367 (2019). https://doi.org/10.1038/s4158701900808
[34] Smith, B., Ashburner, M., Rosse, C., et al.: The OBO foundry: coordinated evolution of ontologies to support biomedical data integration **25**(11), 1251–1255 (2007). https://doi.org/10.1038/nbt1346
[35] Snomed: SNOMED home page, `https://www.snomed.org/`
[36] Taylor, C.F., Field, D., Sansone, S.A., et al.: Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project **26**(8), 889–896 (2008). https://doi.org/https://doi.org/110.1038/nbt.1411
[37] W3id: w3id.org - permanent identifiers for the web, `https://w3id.org/`
[38] Whetzel, P.L., Noy, N.F., Shah, N.H., et al.: BioPortal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications **39**, W541–w545 (2011). https://doi.org/10.1093/nar/gkr469
[39] Wilkinson, M.: The FAIR maturity evaluation service (2021), `https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/#!/`
[40] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., et al.: The FAIR guiding principles for scientific data management and stewardship **3**(1), 160018 (2016). https://doi.org/10.1038/sdata.2016.18
[41] Wimalaratne, S.M., Juty, N., Kunze, J., et al.: Uniform resolution of compact identifiers for biomedical data **5**, 180029 (2018). https://doi.org/10.1038/sdata.2018.29

Supplementary Table 1: The suitability of OBO principles used as FAIR Vocabulary Features

OBO-1: The ontology MUST be openly available to be used by all without any constraint other than (a) its origin must be acknowledged and (b) it is not to be altered and subsequently redistributed in altered form under the original name or with the same identifiers.	**Suitable as FAIR Vocabulary Feature? No**
While it is highly desirable to have open source semantic artefacts, as it is to have open source code and ultimately open research. This is not a prerequisite for vocabularies to be FAIR, nor is it well aligned with the FAIR principles which require licensing information to be provided but do not mandate it be open source.	
OBO-2: The ontology is made available in a common formal language in an accepted concrete syntax.	**Suitable as FAIR Vocabulary Feature? Yes**
Common formats define minimum standards for accessing an ontology, support using ontologies in a FAIR capable resource, and are the foundation of interoperable vocabulary. Many common formal languages have been proposed during the development history of ontologies. OWL format is currently used as a W3C standard.	
OBO-3: Each class and relation (property) in the ontology must have a unique URI identifier.	**Suitable as FAIR Vocabulary Feature? Yes**
Identifiability is a core FAIR principle, and in order to be fulfilled, all elements of the ontology, such as classes and relationships, should be clearly and uniquely identified.	
OBO-4: The ontology provider has documented procedures for versioning the ontology, and different versions of ontology are marked, stored, and officially released.	**Suitable as FAIR Vocabulary Feature? No**
Proper versioning improves the findability and reusability of the vocabulary. We proposed a corresponding FVF to cover this aspect. Yet, this principle evaluates the development, especially the documentation process, of the ontology rather than FAIR vocabulary.	
OBO-5: The scope of an ontology is the extent of the domain or subject matter it intends to cover. The ontology must have a clearly specified scope and content that adheres to that scope.	**Suitable as FAIR Vocabulary Feature? Yes**
Users must be able to determine which ontologies meet their needs in order to implement these in FAIR capable resources and to be interoperable with other vocabularies. A clear specification of this aids in FAIR implementation; vocabularies do not exist in isolation.	

OBO-6: The ontology has textual definitions for the majority of its classes and for top level terms in particular	**Suitable as FAIR Vocabulary Feature? No**
While textual definitions provide human-readable content and are generally desirable, this is not essential as an FVF as ontology content has definitions in terms of logical axioms (e.g., position in the hierarchy) and term labels.	
OBO-7: Relations should be reused from the Relations Ontology (RO).	**Suitable as FAIR Vocabulary Feature? Yes**
Relation standards promote interoperability across different ontologies. This principle focuses on the relationships within and across ontologies. It can be adapted and used in a broader range of FAIR vocabularies.	
OBO-8: The owners of the ontology should strive to provide as much documentation as possible. The documentation should detail the different processes specific to an ontology life cycle and target various audiences (users or developers).	**Suitable as FAIR Vocabulary Feature? Yes**
Rich metadata of the vocabulary, such as the purpose and status of the ontology, promotes the reuse of the ontology.	
OBO-9: The ontology developers should document that the ontology is used by multiple independent people or organizations.	**Suitable as FAIR Vocabulary Feature? No**
Usage of vocabularies depends on the content and there are strategies for interoperating ontologies in what is anyway a crowded semantic space. Evidence that an ontology is highly used - when measurable - assumes a level of maturity that is unlikely for some starting communities, and would not reflect fairness of the resource.	
OBO-10: OBO Foundry ontology development, in common with many other standards-oriented scientific activities, should be carried out in a collaborative fashion.	**Suitable as FAIR Vocabulary Feature? Yes**
Vocabularies should be implementable in FAIR data resources and should reflect community needs. Responsiveness to community needs comes via collaboration and development should therefore be collaborative.	
OBO-11: There should be a person who is responsible for communications between the community and the ontology developers, for communicating with the Foundry on all Foundry-related matters, for mediating discussions involving maintenance in the light of scientific advance, and for ensuring that all user feedback is addressed.	**Suitable as FAIR Vocabulary Feature? No**
Having a designed contact is important for the community to provide feedback and ensures someone is having an ongoing editorial responsibility for the ontology. While it does pertain to sustainability of the resource and its possible evolution, it doesn't	

directly speak to its level of fairness.	
OBO-12: Naming conventions are used	**Suitable as FAIR Vocabulary Feature? No**
Consistency of naming is a best practice feature of ontologies but does not detract from deployment in support of the FAIR principles.	
OBO-16: The ontology needs to reflect changes in scientific consensus to remain accurate over time.	**Suitable as FAIR Vocabulary Feature? Yes**
Vocabularies codify knowledge. To fulfil one of their primary functions, they must evolve. Dead ontologies fail to support FAIR capable resources to interoperate.	
OBO-20: Ontology developers MUST offer channels for community participation and SHOULD be responsive to requests.	**Suitable as FAIR Vocabulary Feature? No**
Having communication channels to collect and respond to community requirements supports the maintenance and evolution of the ontology. However, it is not directly linked to the FAIRness of the vocabulary.	

Supplementary Table 2: FAIR Vocabulary Features mapped to FAIR principles and FAIR vocabulary requirements

	Findability	Accessibility	Interoperability	Reusability
FAIR in terms of application to FAIR data.			FVF-11	FVF-2 FVF-6 FVF-9 FVF-10 FVF-11
FAIR in terms of serving as a FAIR data resource.	FVF-1 FVF-4 FVF-6	FVF-3 FVF-5 FVF-10	FVF-7	FVF-2 FVF-6 FVF-7 FVF-9
FAIR in the context of interacting with other vocabularies.			FVF-8	

Supplementary Materials 3: VersionIRI analysis

We fetched ontologies indexed in the OLS repository and selected those that are successfully loaded and up-to-date. OLS contains 266 biomedical ontologies by the time we access the database (https://www.ebi.ac.uk/ols/api/ontologies). We filtered out ontologies which could not be indexed automatically (without a valid loaded timestamp), and removed inactive ontologies based on the date information in the versionIRI section. 200 ontologies are selected based on these criteria.

We recognise the limitations of the ontology selection approaches. The filtering relys on the metadata collected by OLS instead of the ontology itself, and therefore might not correctly reflect the ontology status. We filted out some inactive ontologies based on the loading time (only ontologies with a loading timestamp after 2019-01-01 are choosen) and date information in the versioneIRI (ontologies with date before 2019-01-01 in the verionIRI are removed). But these criterias does not ensure all vocabularies selected are up-to-date. For example, for ontologies using semantic versioning format where no date information is provided in the versionIRI, or some update information are collected in other metadata fields such as 'annotation' 'editor comments', etc.

Despite the constraints of the analysis, it still provides enough information to showcase the status of current vocabularies. A complete list of selected ontologies are provided in the table below.

Supplementary Table 4: RDA data maturity indicators that are not mapped to FAIR Vocabulary Features

ID	Indicator
RDA-F3-01M	Metadata includes the identifier for the data
RDA-I2-01M	Metadata uses FAIR-compliant vocabularies
RDA-I2-01D	Data uses FAIR-compliant vocabularies
RDA-I3-01M	Metadata includes references to other metadata
RDA-I3-01D	Data includes references to other data
RDA-I3-02M	Metadata includes references to other data
RDA-I3-04M	Metadata include qualified references to other data

Supplementary Table 5: FAIR assessment results of Gene ontology

RDF FAIR indicators version	v0.05
Project name	FAIR assessment Gene Ontology
Assessment date	2021-08-02
Dataset version	Release 2021-07-02
Dataset link	https://github.com/geneontology/go-ontology and http://geneontology.org/

FAIR vocabulary feature summary	
FVF, full compliance	90.91%
FVF, partial compliance	9.09%
FVF, no compliance	0.00%

FAIR vocabulary Feature	RDA indicator ID	Indicator	Assessment - RDA	Assessment - FVF	Assessment details
FVF-1: Vocabulary and their terms are assigned globally unique and persistent identifiers.	RDA-F1-01M	Metadata is identified by a persistent identifier	1	Full Compliance	Metadata are provided in http://geneontology.org/docs/ontology-documentation/. It can also be found in the OBO foundry repository https://github.com/OBOFoundry/OBOFoundry.github.io/edit/master/ontology/go.md. But they are not standard persistent identifiers.
	RDA-F1-01D	Data is identified by a persistent identifier	1		Gene ontology uses PURL identifiers http://purl.obolibrary.org/obo/go.owl
	RDA-F1-02M	Metadata is identified by a globally unique identifier	1		http://geneontology.org/docs/ontology-documentation/ is globally unique identifier.
	RDA-F1-02D	Data is identified by a globally unique identifier	1		http://purl.obolibrary.org/obo/go.owl is globally unique identifier.

FVF-2: Vocabularies and their terms have rich metadata.	RDA-F2-01M	Rich metadata is provided to allow discovery	1	Full Compliance	Descriptive text is provided in http://geneontology.org. Rich metadata for indexing and reuse is provided in https://github.com/OBOFoundry/OBOFoundry.github.io/edit/master/ontology/go.md.
FVF-3: Vocabularies and their terms can be accessed using the identifiers, preferably by both human and machine.	RDA-A1-01M	Metadata contains information to enable the user to get access to the data	1	Full Compliance	The metadata includes data download links http://geneontology.org/docs/download-ontology/.
	RDA-A1-02M	Metadata can be accessed manually (i.e. with human intervention)	1		The metadata can be accessed from the gene ontology website.
	RDA-A1-02D	Data can be accessed manually (i.e. with human intervention)	1		Data can be downloaded from http://geneontology.org/docs/download-ontology/
	RDA-A1-03M	Metadata identifier resolves to a metadata record	1		http://geneontology.org/docs/ontology-documentation// is resolvable and directs to the metadata.
	RDA-A1-03D	Data identifier resolves to a digital object	1		The vocabulary identifier http://purl.obolibrary.org/obo/go.owl resolves to the ontology source files. Identifiers such as http://purl.obolibrary.org/obo/GO_0098743 resolves to ontology terms.
	RDA-A1-05D	Data can be accessed automatically (i.e. by a computer program)	1		Data can be downloaded using command line tools, such as curl, wget, etc.
FVF-4: Vocabularies and their terms are registered or indexed in a searchable engine or a resource.	RDA-F4-01M	Metadata is offered in such a way that it can be harvested and indexed	1	Full Compliance	Gene ontology has been indexed by EMBL OLS, BioPortal and other semantic repositories. Also it is indexed in Google search.

FVF-5: Vocabularies and their terms are retrievable using a standardised communications protocol, preferably open, free and universally implementable protocols. and allows for authentication and authorisation, where necessary.	RDA-A1-04M	Metadata is accessed through standardised protocol	1	Full Compliance	The metadata can be accessed through the HTTP protocol.
	RDA-A1-04D	Data is accessible through standardised protocol	1		Data can be accessed through HTTP protocol.
	RDA-A1.1-01M	Metadata is accessible through a free access protocol	1		HTTP is a free access protocol.
	RDA-A1.1-01D	Data is accessible through a free access protocol	1		HTTP is a free access protocol.
	RDA-A1.2-01D	Data is accessible through an access protocol that supports authentication and authorisation	NA		HTTP HTTP allows access control. But authentication and authorisation are not required by Gene Ontology.
FVF-6: Vocabularies and their terms are persistent over time and are appropriately versioned.	RDA-A2-01M	Metadata is guaranteed to remain available after data is no longer available	1	Partial Compliance	Metadata and data can be found in version controlled repositories on Github.
	RDA-R1.2-01M	Metadata includes provenance information according to community-specific standards	1		The metadata includes links to access different snapshots of the ontology. The snapshots are in owl/obo format and has PURL identifiers.
	RDA-R1.2-02M	Metadata includes provenance information according to a cross-community language	0		
FVF-7: Vocabularies and their terms use a formal, accessible and broadly applicable, and preferably machine-understandable language for knowledge representation.	RDA-I1-01M	Metadata uses knowledge representation expressed in standardised format	1	Full Compliance	The metadata is provided in a standard format and can be harvested by major vocabulary serivices, such as OLS and BioPortal.
	RDA-I1-01D	Data uses knowledge representation expressed in standardised format	1		The data uses OWL and OBO standards.
	RDA-I1-02M	Metadata uses machine-understandable knowledge	1		Basic metadata is provided in OWL.

		representation			
	RDA-I1-02D	Data uses machine-understandable knowledge representation	1		The GO data uses OWL and OBO formats, which are machine-readable community formats.
FVF-8: Vocabularies and terms use qualified references to other vocabularies.	RDA-I3-02D	Data includes qualified references to other data	1	Full Compliance	The Gene ontology cross reference policy is here: http://geneontology.org/docs/download-mappings/ Data from other vocabularies are provided as 'xref'/
	RDA-I3-03M	Metadata includes qualified references to other metadata	1		The Gene ontology cross reference policy is provided here: http://geneontology.org/docs/download-mappings/
FVF-9: Vocabularies and terms are described with a plurality of accurate and relevant attributes.	RDA-R1-01M	Plurality of accurate and relevant attributes are provided to allow reuse	1	Full Compliance	Gene Ontology includes sufficient term attributes. http://geneontology.org/docs/GO-term-elements
FVF-10: Vocabularies are released with a standard data usage licence, preferably machine-readable licence.	RDA-R1.1-01M	Metadata includes information about the licence under which the data can be reused	1	Full Compliance	Gene Ontology Consortium data and data products are licensed under the Creative Commons Attribution 4.0 Unported License.
	RDA-R1.1-02M	Metadata refers to a standard reuse licence	1		
	RDA-R1.1-03M	Metadata refers to a machine-understandable reuse licence	1		
FVF-11: Vocabularies meet domain relevant community standards.	RDA-R1.3-01M	Metadata complies with a community standard	1	Full Compliance	
	RDA-R1.3-01D	Data complies with a community standard	1		
	RDA-R1.3-02M	Metadata is expressed in compliance with a machine-understandable community standard	1		

	RDA-R1.3-02D	Data is expressed in compliance with a machine-understandable community standard	1		

Supplementary Table 6: FAIR assessment results of Experimental Factor Ontology

RDF FAIR indicators version	v0.05
Project name	EFO assessment
Assessment date	2021-08-02
Dataset version	3.32.0
Dataset link	http://www.ebi.ac.uk/efo/releases/v3.32.0/efo.owl

FAIR vocabulary feature summary	
FVF, full compliance	81.82%
FVF, partial compliance	9.09%
FVF, no compliance	9.09%

FAIR vocabulary Feature	**RDA indicator ID**	**Indicator**	**Assessment - RDA**	**Assessment - FVF***	**Assessment details**
FVF-1: Vocabulary and their terms are assigned globally unique and persistent identifiers.	RDA-F1-01M	Metadata is identified by a persistent identifier	1	Full Compliance	Both the data and metadata use identifier: http://www.ebi.ac.uk/efo/efo.owl
	RDA-F1-01D	Data is identified by a persistent identifier	1		
	RDA-F1-02M	Metadata is identified by a globally unique identifier	1		
	RDA-F1-02D	Data is identified by a globally unique identifier	1		
FVF-2: Vocabularies and their terms have rich metadata.	RDA-F2-01M	Rich metadata is provided to allow discovery	0	No Compliance	Description of EFO has been provided in ontology browsers, such as OLS and BioPortal. However, the descriptions are not included in the EFO source file.
FVF-3: Vocabularies and their terms can be accessed using the	RDA-A1-01M	Metadata contains information to enable the user to get access	1	Full Compliance	EFO and its terms has unique identifiers and can be accessed.

identifiers, preferably by both human and machine.		to the data			
	RDA-A1-02M	Metadata can be accessed manually (i.e. with human intervention)	1		
	RDA-A1-02D	Data can be accessed manually (i.e. with human intervention)	1		
	RDA-A1-03M	Metadata identifier resolves to a metadata record	1		
	RDA-A1-03D	Data identifier resolves to a digital object	1		
	RDA-A1-05D	Data can be accessed automatically (i.e. by a computer program)	1		
FVF-4: Vocabularies and their terms are registered or indexed in a searchable engine or a resource.	RDA-F4-01M	Metadata is offered in such a way that it can be harvested and indexed	1	Full Compliance	The metadata is provided in OWL format and has been harvested by both OLS and BioPortal
FVF-5: Vocabularies and their terms are retrievable using a standardised communications protocol, preferably open, free and universally implementable protocols. and allows for authentication and authorisation, where necessary.	RDA-A1-04M	Metadata is accessed through standardised protocol	1		EFO can be accessed using HTTP protocol, and it is an open-acess ontology.
	RDA-A1-04D	Data is accessible through standardised protocol	1		
	RDA-A1.1-01M	Metadata is accessible through a free access protocol	1		
	RDA-A1.1-01D	Data is accessible through a free access protocol	1		
	RDA-A1.2-01D	Data is accessible through an access protocol that supports authentication and authorisation	NA	Full Compliance	
FVF-6: Vocabularies and their terms are persistent over time and are appropriately versioned.	RDA-A2-01M	Metadata is guaranteed to remain available after data is no longer available	1	Partial Compliance	EFO follows vocabulary release guidelines, and its versioned copies can be found on Github. But it doesn't strictly follows cross community language standards, such as rdfs, xmls

					standards.
	RDA-R1.2-01M	Metadata includes provenance information according to community-specific standards	1		
	RDA-R1.2-02M	Metadata includes provenance information according to a cross-community language	0		
FVF-7: Vocabularies and their terms use a formal, accessible and broadly applicable, and preferably machine-understandable language for knowledge representation.	RDA-I1-01M	Metadata uses knowledge representation expressed in standardised format	1	Full Compliance	EFO can be downloaded in OWL and OBO, which are standardised format and machine-understandable.
	RDA-I1-01D	Data uses knowledge representation expressed in standardised format	1		
	RDA-I1-02M	Metadata uses machine-understandable knowledge representation	1		
	RDA-I1-02D	Data uses machine-understandable knowledge representation	1		
FVF-8: Vocabularies and terms use qualified references to other vocabularies.	RDA-I3-02D	Data includes qualified references to other data	1	Full Compliance	EFO reuses terms from other vocabularies and provides suffient reference, such as source of the external term.
	RDA-I3-03M	Metadata includes qualified references to other metadata	1		
FVF-9: Vocabularies and terms are described with a plurality of accurate and relevant attributes.	RDA-R1-01M	Plurality of accurate and relevant attributes are provided to allow reuse	1	Full Compliance	
FVF-10: Vocabularies are released with a standard data usage licence, preferably machine-readable	RDA-R1.1-01M	Metadata includes information about the licence under which the data can be reused	1	Full Compliance	

licence.	RDA-R1.1-02M	Metadata refers to a standard reuse licence	1		
	RDA-R1.1-03M	Metadata refers to a machine-understandable reuse licence	1		
FVF-11: Vocabularies meet domain relevant community standards.	RDA-R1.3-01M	Metadata complies with a community standard	1	Full Compliance	EFO uses the standard OWL format, complies with OBO principles and imports terms following the MIREOT standards.
	RDA-R1.3-01D	Data complies with a community standard	1		
	RDA-R1.3-02M	Metadata is expressed in compliance with a machine-understandable community standard	1		
	RDA-R1.3-02D	Data is expressed in compliance with a machine-understandable community standard	1		

Supplementary Table 7: FAIR assessment results of ICD-11

Project name	ICD-11 FAIR assessment
Assessment date	2021-08-02
Dataset version	05/2021
Dataset link	ICD-11 browser and ICD11 print version:https://icd.who.int/en print version:https://icd.who.int/browse11/Downloads/Download?fileName=print_en.zip

FAIR vocabulary feature summary	
FVF, full compliance	27.27%
FVF, partial compliance	36.36%
FVF, no compliance	36.36%

FAIR vocabulary Feature	**RDA indicator ID**	**Indicator**	**Assessment - RDA**	**Assessment - FVF**	**Assessment details**
FVF-1: Vocabulary and their terms are assigned globally unique and persistent identifiers.	RDA-F1-01M	Metadata is identified by a persistent identifier	0	Partial Compliance	No metadata identifier.
	RDA-F1-01D	Data is identified by a persistent identifier	1		Example data identifier: 2C25.1 Small cell carcinoma of bronchus or lung.
	RDA-F1-02M	Metadata is identified by a globally unique identifier	0		
	RDA-F1-02D	Data is identified by a globally unique identifier	0		The identifiers in ICD-11 has been through several iterations. Currently, ICD-11 provides identifiers such as (1C60-1C62.Z), a more persistent identifier system, http://id.who.int/icd/entity/911707612 is still under development. This assessment is based on the 1C60-1C62.Z system

FVF-2: Vocabularies and their terms have rich metadata.	RDA-F2-01M	Rich metadata is provided to allow discovery	1	Full Compliance	
FVF-3: Vocabularies and their terms can be accessed using the identifiers, preferably by both human and machine.	RDA-A1-01M	Metadata contains information to enable the user to get access to the data	1	Partial Compliance	ICD-11 has provided API, web browser, and pdf documents for human and machine access.
	RDA-A1-02M	Metadata can be accessed manually (i.e. with human intervention)	1		
	RDA-A1-02D	Data can be accessed manually (i.e. with human intervention)	1		
	RDA-A1-03M	Metadata identifier resolves to a metadata record	0		
	RDA-A1-03D	Data identifier resolves to a digital object	1		
	RDA-A1-05D	Data can be accessed automatically (i.e. by a computer program)	1		
FVF-4: Vocabularies and their terms are registered or indexed in a searchable engine or a resource.	RDA-F4-01M	Metadata is offered in such a way that it can be harvested and indexed	0	No Compliance	ICD-11 metadata is provided in word documents and can not be directly indexed by vocabulary databases.
FVF-5: Vocabularies and their terms are retrievable using a standardised communications protocol, preferably open, free and universally implementable protocols. and allows for authentication and authorisation, where necessary.	RDA-A1-04M	Metadata is accessed through standardised protocol	1	Full Compliance	
	RDA-A1-04D	Data is accessible through standardised protocol	1		ICD-11 uses HTTPS protocol.https://icd.who.int/browse11/l-m/en#http%3a%2f%2fid.who.int%2ficd%2fentity%2f911707612
	RDA-A1.1-01M	Metadata is accessible through a free access protocol	1		
	RDA-A1.1-01D	Data is accessible through a free access protocol	1		
	RDA-A1.2-01D	Data is accessible through an access protocol that supports authentication and authorisation	NA		

FVF-6: Vocabularies and their terms are persistent over time and are appropriately versioned.	RDA-A2-01M	Metadata is guaranteed to remain available after data is no longer available	1	Partial Compliance	Previous versions of ICD-11 can access at https://icd.who.int/browse11/l-m/en/releases. However, the versioning style does not follow common community standards.
	RDA-R1.2-01M	Metadata includes provenance information according to community-specific standards	1		
	RDA-R1.2-02M	Metadata includes provenance information according to a cross-community language	0		
FVF-7: Vocabularies and their terms use a formal, accessible and broadly applicable, and preferably machine-understandable language for knowledge representation.	RDA-I1-01M	Metadata uses knowledge representation expressed in standardised format	0	No Compliance	ICD-11 is published mainly as a pdf document and doesn't use standard vocabulary formats.
	RDA-I1-01D	Data uses knowledge representation expressed in standardised format	0		
	RDA-I1-02M	Metadata uses machine-understandable knowledge representation	0		
	RDA-I1-02D	Data uses machine-understandable knowledge representation	0		
FVF-8: Vocabularies and terms use qualified references to other vocabularies.	RDA-I3-02D	Data includes qualified references to other data	0	Partial Compliance	Terms in ICD-11 doesn't refer to other terms.
	RDA-I3-03M	Metadata includes qualified references to other metadata	1		The ICD-11 description refers to other projects and publications.
FVF-9: Vocabularies and terms are described with a plurality of accurate and relevant attributes.	RDA-R1-01M	Plurality of accurate and relevant attributes are provided to allow reuse	0	No Compliance	ICD-11 contains only a minimum description of each disease.

FVF-10: Vocabularies are released with a standard data usage licence, preferably machine-readable licence.	RDA-R1.1-01M	Metadata includes information about the licence under which the data can be reused	1	Full Compliance	ICD11 provides licensing documentation. https://icd.who.int/en/docs/ICD11-license.pdf https://icd.who.int/browse11. Licensed under Creative Commons Attribution-NoDerivatives 3.0 IGO licence (CC BY-ND 3.0 IGO).
	RDA-R1.1-02M	Metadata refers to a standard reuse licence	1		
	RDA-R1.1-03M	Metadata refers to a machine-understandable reuse licence	1		
FVF-11: Vocabularies meet domain relevant community standards.	RDA-R1.3-01M	Metadata complies with a community standard	0	No Compliance	
	RDA-R1.3-01D	Data complies with a community standard	0		
	RDA-R1.3-02M	Metadata is expressed in compliance with a machine-understandable community standard	0		
	RDA-R1.3-02D	Data is expressed in compliance with a machine-understandable community standard	0		