FAIR Characterization of PubChem

Leonardo Guerreiro Azevedo, Julio Tesolin, Gabriel Banaggia
IBM Research – Brazil
lga@br.ibm.com, {julio.tesolin,gbanaggia}@ibm.com

Abstract

This document presents the answers of the FAIR characterization questionnaire for PubChem Compound dataset.

1 The Characterization

This document present the use the FAIR characterization questionnaire to outline Compound data and metadata from PubChem. Information comes from the "Perfluorooctanoic acid" compound [4], PubChem's homepage, and PubChemRDF to answer the questions. PubChemRDF [1] translates some data modules of the PubChem repository (like information about Substances and Compounds) into Resource Description Framework (RDF) format. RDF is a directed, labeled graph data format where datasets are represented as triples (subject, predicate, object)[1] [7]. Its goal is to allow researchers to use semantic technologies (like RDF triple stores[2] and SPARQL[3] query engines) to query and analyze PubChem data.

PubChem data is stored in a relational database, and PubChem provides service interfaces to access them. This data storage type restricts several aspects related to FAIR. PubChemRDF, in turn, satisfies the Interoperability and Reusability principles much more consistently. We use both repositories in the characterization, and prioritize the responses from PubChemRDF due to the use of Semantic Web concepts, which are a better fit for the FAIR principles.

The questionnaire answers are presented as follows. Table 1 presents the context characterization. Tables 2, 3, and 4 depict the responses for questions about Findability principles. Table 5 shows the answers for Accessibility principles. Tables 6, 7, and 8 present the responses for Interoperability principles. Finally, Table 9, and Table 10 present the answers for Reusability principles.

Table 1: Questionnaire answers for the context of the FAIRness assessment of some PubChem compound.

Q.ID	Questions
Q1	What is your community? Chemistry[4].
Q2	Which digital object will be evaluated in this assessment? This assessment is for the data and metadata about Compounds, and we are using the compound "Perfluorooctanoic acid" [4] to perform the assessment.

[1] Like the triple (LeonardoDaVinci, created, TheMonalisa)

[2] *Triplestores*, or RDF stores, are the matter of choice for storing and querying RDF data, like Jena and AllegroGraph.

[3] SPARQL is the query language for RDF (Resource Description Framework) defined by the W3C [5].

Table 2: Questionnaire answers to assess F1 considering some PubChem compounds.

Princ.	Q.id	Questions
F1	Q3	What is the main identifier of the data (*i.e.*, data is understood as any digital object)? The main data identifier is the Compound ID (CID).
	Q4	Are there other attributes able to identify the data? What are them? Yes, there are. They are presented in Section 2 (“Names and Identifiers”) of the digital object split in “Computed Descriptors” and “Other Identifiers”. Examples of computed ones are: IUPAC Name, InChI, InChIKey, Canonical SMILES, and Molecular Formula.
	Q5	Is the data identifier (ID) globally unique or is it only unique in the database domain or for a specific context? Yes, it is. Considering the digital object identifier computed is the URL https://pubchem.ncbi.nlm.nih.gov/compound/9554, it is unique.
	Q6	Is the data ID persistent? We did not find an answer for this question in the documentation. It requires to ask a specialist.
	Q7	Is the data ID resolvable, *e.g.*, to a landing page? Yes, it is. The URL identifier is resolvable in a Browser, and its schema is stored in identifiers.org as pubchem.compound:9554 which is resolvable to a web page.
	Q8	Are there metadata describing the data? Yes, there are.
	Q9	Do the metadata have a distinct ID from the data? No, data and metadata have the same identifier.

References

[1] Gang Fu, Colin Batchelor, Michel Dumontier, Janna Hastings, Egon Willighagen, and Evan Bolton. PubChemRDF: towards the semantic annotation of PubChem compound and substance databases. *Journal of Cheminformatics*, 7(1):34, July 2015.

[2] Valeria Granata. Materials cloud fip. https://archive.materialscloud.org/static/documents/fip/Materials%20Cloud%20V1.1.pdf, 2020.

[3] PubChem. Downloading pubchem data. https://pubchem.ncbi.nlm.nih.gov/docs/downloads. Accessed in 2023–03-02.

[4] PubChem. Perfluorooctanoic acid (cid 9554). https://pubchem.ncbi.nlm.nih.gov/compound/9554. Accessed in 2023–03-02.

[5] W3C. Sparql 1.1 overview. https://www.w3.org/TR/sparql11-overview/, 2013. Accessed in 2023–05-05.

[6] wikidata. Open access. https://www.wikidata.org/wiki/Q232932. Accessed in 2023–02-28.

[7] Nurul Syeilla Syazhween Zulkefli, Nurazzah Abd Rahman, Zainab Abu Bakar, Sharifalillah Nordin, Tengku Mohd Tengku Sembok, and Noor Hasimah Ibrahim Teo. Evaluation of triple indices in retrieving web documents. In *2013 International Conference on Advanced Computer Science Applications and Technologies*, pages 525–529, December 2013.

Table 3: Questionnaire answers to assess F2 considering some PubChem compounds.

Princ.	Q.id	Questions
F2	Q10	Which metadata schemas, if any, are used to describe the data? The set of standardized ontologies used by PubChem to define the domain-specific knowledge are[5]: Chemical Entities of Biological Interest (ChEBI), CHEMical INFormation ontology (CHEMINF), Protein Ontology (PRO), Gene Ontology (GO), Semanticscience Integrated Ontology (SIO), Basic Formal Ontology (BFO), Ontology for Biomedical Investigations (OBI), Information Artifact Ontology (IAO), BioAssay Ontology (BAO), Units of Measurement (UO), Citation Typing Ontology (CiTO), FRBR-aligned Bibliographic Ontology (FaBiO), Dublin Core Metadata Initiative (DCMI) Terms, Simple Knowledge Organization System (SKOS), BioPAX, National Drug File-Reference Terminology (NDF-RT), and National Center Institute thesaurus (NCIt). All of the biomedical ontologies, such as ChEBI, CHEMINF, PRO, GO, BFO, SIO, and BAO, are interfaced by the NIH Roadmap National Center for Biomedical Ontology (NCBO) through its BioPortal, and comply with an evolving set of shared principles established by the Open Biomedical Ontologies (OBO) foundry. Adoption of these core ontologies helps to ensure that the mapping of chemical and biological information is compatible across multiple Semantic Web resources.
	Q11	What kinds of metadata (*e.g.*, descriptive, administrative and structural) are used to describe the data? Considering the PubChem CID 9554 web page, the kinds of metadata are descriptive, administrative, and structural.
	Q12	Which of these metadata schemas are domain specific and which are domain-agnostic? Considering the names of the schemas, the following ones are domain-agnostic: Semanticscience Integrated Ontology (SIO), Basic Formal Ontology (BFO), Information Artifact Ontology (IAO), Units of Measurement (UO), Citation Typing Ontology (CiTO), FRBR-aligned Bibliographic Ontology (FaBiO), Dublin Core Metadata Initiative (DCMI) Terms, Simple Knowledge Organization System (SKOS), and National Center Institute thesaurus (NCIt). Considering the names of the schemas, the following ones are domain-specific: Chemical Entities of Biological Interest (ChEBI), CHEMical INFormation ontology (CHEMINF), Protein Ontology (PRO), Gene Ontology (GO), Ontology for Biomedical Investigations (OBI), BioAssay Ontology (BAO), BioPAX, and National Drug File-Reference Terminology (NDF-RT). Further investigation with a domain expert would help to classify this ontology in a better way.

Table 4: Questionnaire answers to assess F3 and F4 considering some PubChem compounds.

Princ.	Q.id	Questions
F3	Q13	Does the metadata include the identifier of the data it describes? Yes, it does. This can be checked by analyzing the web page content, *e.g.*, using cURL.
	Q14	What is the technology that links metadata to the data (and vice-versa)? Data and metadata are presented as a single object, and there are some metadata properties that contain the digital object ID.
	Q15	How are the metadata and data linked? Data and metadata are presented as a single object.
F4	Q16	Is metadata registered or indexed in a searchable resource? We believe that the metadata is indexed in a searchable tool, *e.g.*, we can reach a page of a PubChem's compound by searching on Google. However, a confirmation with a PubChem specialist is required.
	Q17	Which searchable resource is used to register or index the metadata? We did not find an answer for this question in the documentation. It requires to ask a specialist.
	Q18	Which is the standardized mechanism or service used to provision the metadata? The technologies used are PubChem's search tool, PubChem's web service interface, and download of PubChem's RDF database.
	Q19	How is the metadata available or indexed? (*E.g.*, as a static web page, in a database, JSON returned from an API call) The metadata is available as a web page, making a service call to the PubChem web service interface, and downloading the PubChem RDF database[6].
	Q20	Is data registered or indexed in a searchable resource? We believe that the metadata is indexed in a searchable tool, *e.g.*, we can reach a page of a PubChem's compound by searching on Google. However, a confirmation with a PubChem specialist is required.
	Q21	Which searchable resource is used to register or index the data? We did not find an answer for this question in the documentation. It requires to ask a specialist.
	Q22	Which is the standardized mechanism or service used to provision the data? The technologies used are PubChem's search tool, web service interface, and download of PubChem RDF database.
	Q23	How is the data available or indexed? (*E.g.*, as a static web page, in a database, JSON returned from an API call) The data is available as a web page, making a service call to the PubChem web service interface, and downloading the PubChem RDF database[7].

Table 5: Questionnaire answers to assess Accessibility considering the PubChem's FIP [2].

Princ.	Q.id	Questions
A1.1	Q24	Which communication protocols are used to access the metadata? HTTPS, and FTP (in the case of PubChemRDF download).
	Q25	Is the protocol used to access the metadata standardized, open, free, and universally implementable? Yes, it is.
	Q26	Which communication protocols are used to access the data? HTTPS, and FTP (in the case of PubChemRDF download).
	Q27	Is the protocol used to access the data standardized, open, free, and universally implementable? Yes, it is.
A1.2	Q28	What are the security mechanisms used for metadata access, such as ones used for authentication and authorization, and access conditions and access levels? Open access [6]
	Q29	What are the security mechanisms used for data access, such as ones used for authentication and authorization, and access conditions and access levels? Open access [6]
	Q30	What security information is provided in the metadata that allows one to access the data manually or through a client application? We did not find an answer for this question in the documentation. It requires to ask a specialist.
A2	Q31	Are data and metadata independently stored? No, they are not. They are stored together in the same object.
	Q32	What is the metadata longevity plan? We did not find an answer for this question in the documentation. It requires to ask a specialist.
	Q33	What is the data longevity plan, if any? We did not find an answer for this question in the documentation. It requires to ask a specialist.

Table 6: Questionnaire answers to assess Interoperability I1 considering some PubChem compounds.

Princ.	Q.id	Questions
I1	Q34	What is the knowledge representation used for metadata? *E.g.*, Relational, Document, Key Value, Graph, Object, Hierarchical, Network. PubChem data is stored in a relational database. PubchemRDF is part of PubChem data in RDF.
	Q35	Is the knowledge representation used for metadata formal, accessible, shared, and broadly applicable? Since that Semantic Web is the main tool for interoperability, we consider the relational representation of PubChem's meta is formal and broadly applicable although brings several issues concerning access and sharing. On the other hand, PubChemRDF fits all requirements.
	Q36	In what formats the knowledge representation used for metadata is provided? *E.g.*, eXtensible Markup Language (XML), Turtle (TTL), JSON, JSON-LD, CSV, BLOB, CLOB. The PubChem documentation presents the use of XML to access data through programmatic services, but it is not clear if other formats are used. PubChemRDF data is provided in Turtle.
	Q37	Are the formats used for knowledge representation of metadata formal, accessible, shared, and broadly applicable? Yes, they are.
	Q38	What is the knowledge representation used for data? *E.g.*, Relational, Document, Key Value, Graph, Object, Hierarchical, Network. PubChem data is stored in a relational database. PubchemRDF is part of PubChem data in RDF.
	Q39	Is the knowledge representation used for data formal, accessible, shared, and broadly applicable? Since that Semantic Web is the main tool for interoperability, we consider the relational representation of PubChem's meta is formal and broadly applicable although brings several issues concerning access and sharing. On the other hand, PubChemRDF fits all requirements.
	Q40	In what formats the knowledge representation used for data is provided? *E.g.*, eXtensible Markup Language (XML), Turtle (TTL), JSON, JSON-LD, CSV, BLOB, CLOB. The PubChem documentation presents the use of XML to access data through programmatic services, but it is not clear if other formats are used. PubChemRDF data is provided in Turtle.
	Q41	Are the formats used for knowledge representation of data formal, accessible, shared, and broadly applicable? Yes, they are.

Table 7: Questionnaire answers to assess Interoperability I2 and I3 considering some PubChem compounds.

Princ.	Q.id	Questions
I2	Q42	Which structured vocabularies are used for metadata? *Structured vocabularies range from simple taxonomies or thesauri (e.g., in SKOS) to ontologies in OWL available in public accessible registry.* The set of structure vocabularies are the following ontologies: Chemical Entities of Biological Interest (ChEBI), CHEMical INFormation ontology (CHEMINF), Protein Ontology (PRO), Gene Ontology (GO), Semanticscience Integrated Ontology (SIO), Basic Formal Ontology (BFO), Ontology for Biomedical Investigations (OBI), Information Artifact Ontology (IAO), BioAssay Ontology (BAO), Units of Measurement (UO), Citation Typing Ontology (CiTO), FRBR-aligned Bibliographic Ontology (FaBiO), Dublin Core Metadata Initiative (DCMI) Terms, Simple Knowledge Organization System (SKOS), BioPAX, National Drug File-Reference Terminology (NDF-RT), and National Center Institute thesaurus (NCIt). All of the biomedical ontologies, such as ChEBI, CHEMINF, PRO, GO, BFO, SIO, and BAO, are interfaced by the NIH Roadmap National Center for Biomedical Ontology (NCBO) through its BioPortal, and comply with an evolving set of shared principles established by the Open Biomedical Ontologies (OBO) foundry. Adoption of these core ontologies helps to ensure that the mapping of chemical and biological information is compatible across multiple Semantic Web resources.
	Q43	Are the used vocabularies for metadata FAIR? We did not find an answer for this question in the documentation. It requires to ask a specialist.
	Q44	Which structured vocabularies are used for data? *Structured vocabularies range from simple taxonomies or thesauri (e.g., in SKOS) to ontologies in OWL available in public accessible registry.* The set of structure vocabularies are the following ontologies: Chemical Entities of Biological Interest (ChEBI), CHEMical INFormation ontology (CHEMINF), Protein Ontology (PRO), Gene Ontology (GO), Semanticscience Integrated Ontology (SIO), Basic Formal Ontology (BFO), Ontology for Biomedical Investigations (OBI), Information Artifact Ontology (IAO), BioAssay Ontology (BAO), Units of Measurement (UO), Citation Typing Ontology (CiTO), FRBR-aligned Bibliographic Ontology (FaBiO), Dublin Core Metadata Initiative (DCMI) Terms, Simple Knowledge Organization System (SKOS), BioPAX, National Drug File-Reference Terminology (NDF-RT), and National Center Institute thesaurus (NCIt). All of the biomedical ontologies, such as ChEBI, CHEMINF, PRO, GO, BFO, SIO, and BAO, are interfaced by the NIH Roadmap National Center for Biomedical Ontology (NCBO) through its BioPortal, and comply with an evolving set of shared principles established by the Open Biomedical Ontologies (OBO) foundry. Adoption of these core ontologies helps to ensure that the mapping of chemical and biological information is compatible across multiple Semantic Web resources.
	Q45	Are the used vocabularies for data FAIR? We did not find an answer for this question in the documentation. It requires to ask a specialist.

Table 8: Questionnaire answers to assess Interoperability I4 and I3 considering some PubChem compounds.

Princ.	Q.id	Questions
I3	Q46	Which qualified references the metadata include to other data or metadata? *Qualified references means any external metadata used to enrich the information.* PubChem use references, *e.g.*, to cite the source of information and to reference other ontologies that contains the meaning of used predicates. However, we believe there are other references which needs an investigation with a PubChem expert.
	Q47	Which qualified references the data include to other data or metadata? *Qualified references means any external metadata used to enrich the information.* PubChem use references, *e.g.*, to cite the source of information and to reference other ontologies that contains the meaning of used predicates. However, we believe there are other references which needs an investigation with a PubChem expert.

Table 9: Questionnaire answers to assess Reusability considering the some PubChem compounds.

Princ.	Q.id	Questions
R1	Q48	What are the relevant metadata attributes? The metadata attributes are presented in the PubChem Compound page or the PubChem RDF database. As the data and metadata are presented together, we need a specialist to distinguish them.
	Q49	What is the required accuracy of each metadata attribute, if any? We need a specialist to answer this question.
	Q50	What are the relevant data attributes? The data attributes are presented in the PubChem Compound page or the PubChem RDF database. As the data and metadata are presented together, we need a specialist to distinguish them.
	Q51	What is the required accuracy of each attribute, if any? We need a specialist to answer this question.
R1.1	Q52	Which usage license is used for your metadata? PubChem is a open access database; however, there are exceptions where licensing agreements prevent data contributors from allowing bulk downloads of some data sets [3]. So, PubChem data and metadata may have more than one distinct license, *e.g.*, in PubChem compound web page, it is presented the license of each data sources that contributed to the compound data and metadata (see Section 19 - Information Sources) [4].
	Q53	Is the metadata usage license clear? *Clear means that if it is easy to find the license under which the metadata is released.* Yes, it is.
	Q54	Is the metadata usage license accessible? *Accessible license means that the license has no (or few) restrictions to reuse the metadata.* As there are more than one license for the (meta)data, the reuse restrictions vary.
	Q55	Which usage license is used for your data? PubChem is a open access database; however, there are exceptions where licensing agreements prevent data contributors from allowing bulk downloads of some data sets [3]. So, PubChem data and metadata may have more than one distinct license, *e.g.*, in PubChem compound web page, it is presented the license of each data sources that contributed to the compound data and metadata (see Section 19 - Information Sources) [4].
	Q56	Is the data usage license clear? *Clear means that if it is easy to find the license under which the data is released.* Yes, it is.
	Q57	Is the data usage license accessible? As there are more than one license for the (meta)data, the reuse restrictions vary.

Table 10: Questionnaire answers to assess Reusability considering some PubChem compounds.

Princ.	Q.id	Questions
R1.2	Q58	Which metadata schemas do you use for describing provenance of the metadata? We did not find an answer for this question in the documentation. It requires to ask a specialist.
	Q59	Which metadata schemas do you use for describing provenance of the data? We did not find an answer for this question in the documentation. It requires to ask a specialist.
	Q60	What are the attributes used for data provenance? We did not find an answer for this question in the documentation. It requires to ask a specialist.
R1.3	Q61	What are the domain-relevant community standards for metadata? *Domain-relevant community standards mean minimum information standards, well-established and sustainable file formats, common types for information, use of template and standardized vocabularies and ontologies etc.* We did not find an answer for this question in the documentation. It requires to ask a specialist.
	Q62	Do the metadata under assessment meet these domain-relevant community standards? We did not find an answer for this question in the documentation. It requires to ask a specialist.
	Q63	What are the domain-relevant community standards for data? *Domain-relevant community standards mean minimum information standards, well-established and sustainable file formats, common types for information, use of template and standardized vocabularies and ontologies etc.* We did not find an answer for this question in the documentation. It requires to ask a specialist.
	Q64	Do the data under assessment meet these domain-relevant community standards? We did not find an answer for this question in the documentation. It requires to ask a specialist.