Chemical Twins: A Discrete Semantic Representation of Important Chemical Information

Stuart J. Chalk

Department of Chemistry, University of North Florida

schalk@unf.edu

National Science Foundation
WHERE DISCOVERIES BEGIN

Grant #1835643

ACS Meeting – Spring 2022

Outline

* Semantic Data
* What are Digital Twins?
* The SciData Framework
* Chemical Metadata in SciData
* FAIR Digital Objects
* Conclusion

"FAIR Data needs FAIR Chemicals"

https://kidsfirstdrc.org/news/fair-data/

Semantic Data

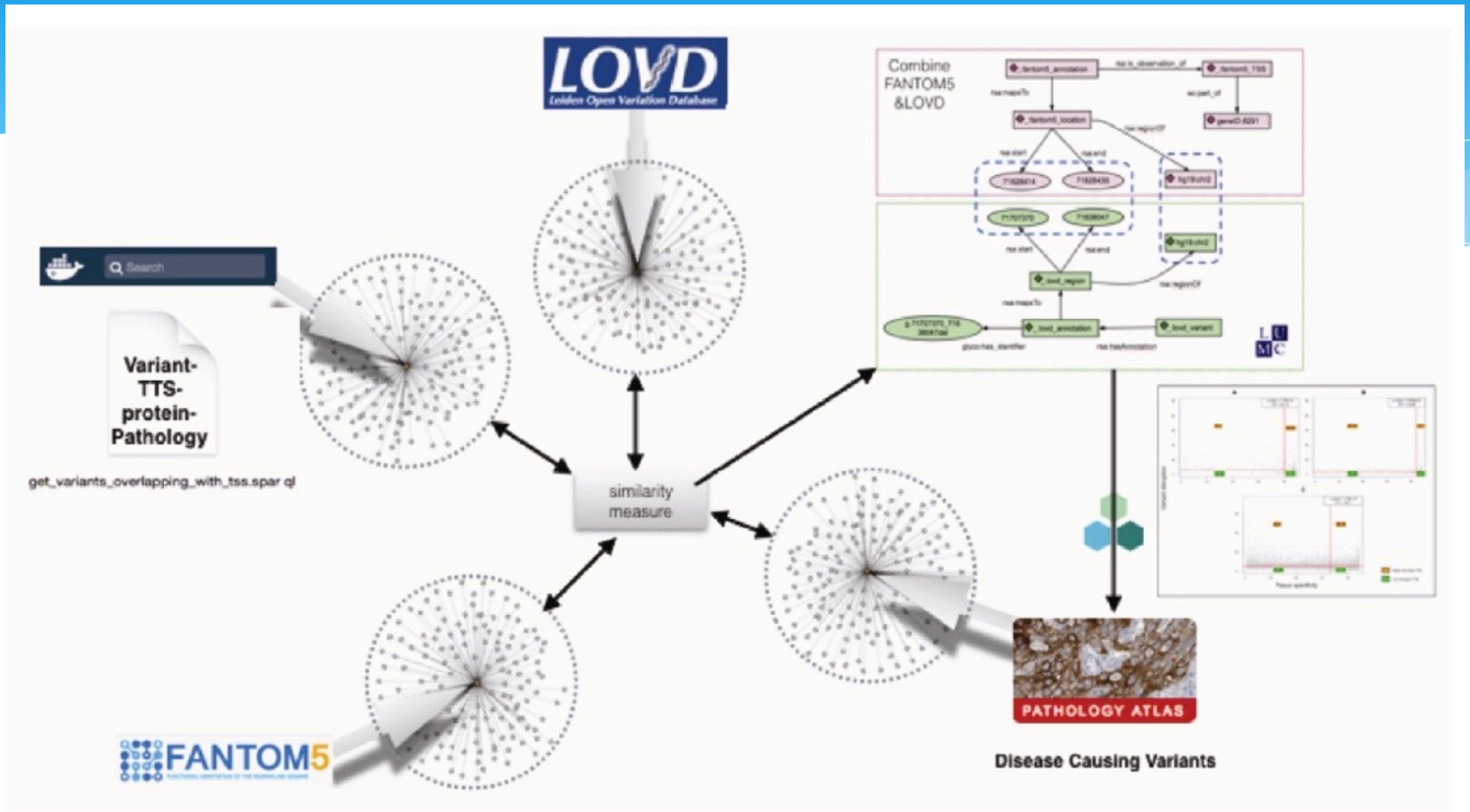
* Sir Tim Berners Lee coined the term "Semantic Web" in 2001 https://www.w3.org/2013/data/
* Resource Description Framework (RDF) "triples" are discrete statements of known information in a "Subject" - "Predicate" - "Object" (SPO) format "Subject" - "Predicate" - "Object" – "Graph" ("quad")
* RDF encodings include RDF/XML, Turtle, and JavaScript Object Notation for Linked Data (JSON-LD)

Digital Twins

* *"A digital twin is a virtual representation of an object or system that spans its lifecycle, is updated from real-time data, and uses simulation, machine learning and reasoning to help decision-making."*
* A semantic digital twin can be thought of as a statement of explicitly known facts (assertations in RDF) about a "thing".

https://www.ibm.com/blogs/internet-of-things/iot-cheat-sheet-digital-twin/

Knowlets



Barend Mons; FAIR Science for Social Machines: Let's Share Metadata Knowlets in the Internet of FAIR Data and Services. *Data Intelligence* 2019; 1 (1): 22–42. doi: https://doi.org/10.1162/dint_a_00002

Digital Twins of Chemical Substances

* Create a representation of known* explicit statements about the abstract concept of a chemical substance
  * Identifiers
  * Descriptors
  * Molecular Graph – atoms and bonds
* Requirements: standardized, semantic, shareable
* Store as a JSON-LD file in the SciData framework

* What is 'known' in this context needs to be debated

SciData Framework

* JSON-LD file with different sections of data
  * General metadata
  * Methodology
  * System
  * Dataset
* Context section in JSON-LD provides semantic annotation of the file contents (metadata and data)

SciData Framework

* Any label starting with "@" is a keyword in JSON-LD
* The "@graph" section is the chemical substance data
* JSON-LD can be converted to RDF using software (RDFlib) https://github.com/RDFLib/


```
{
    "@context": [
        "https://stuchalk.github.io/scidata/contexts/scidata.jsonld",
        "https://stuchalk.github.io/scidata/contexts/chemtwin.jsonld",
        {
            "sdo": "https://stuchalk.github.io/scidata/ontology/scidata.owl#",
            "obo": "http://purl.obolibrary.org/obo/",
            "ss": "https://semanticscience.org/resource/",
            "atc": "http://purl.bioontology.org/ontology/ATC/",
            "w3i": "https://w3id.org/skgo/modsci#"
        },
        {"@base": "https://example.com/"}
    ],
    "@id": "https://example.com/1",
    "generatedAt": "2021-07-07 09:44:04.208975",
    "version": 2,
    "@graph": {
        "@id": "example/1",
        "@type": "sdo:scidataFramework",
        "uid": "",
        "title": "",
        "authors": [ [7 lines]
        "description": "",
        "publisher": "",
        "version": "1.1",
        "keywords": [],
        "permalink": "",
        "toc": [],
        "ids": [],
        "scidata": {
            "@id": "scidata",
            "@type": "sdo:scientificData",
            "discipline": "w3i:Chemistry",
            "methodology": {},
            "system": {},
            "dataset": {}
        },
        "sources": [],
        "rights": []
    }
}
```

Chemical Twin Concept

- Finding aid
- Chemical Metadata
  - General
  - Identifiers
  - Descriptors
  - Molgraph


```
{
  "@context": [
    "https://stuchalk.github.io/scidata/contexts/scidata.jsonld",
    "https://stuchalk.github.io/scidata/contexts/chemtwin.jsonld",
    { [6 lines]
    {"@base": "https://example.com/chemtwin_MYMOFIZGZYHOMD-UHFFFAOYSA-N/"}
  ],
  "@id": "https://example.com/facet/00008979",
  "generatedAt": "2021-07-07 09:44:04.208975",
  "version": 2,
  "@graph": {
    "@id": "https://example.com/chemtwin_MYMOFIZGZYHOMD-UHFFFAOYSA-N/",
    "@type": "sdo:scidataFramework",
    "uid": "chemtwin_MYMOFIZGZYHOMD-UHFFFAOYSA-N",
    "title": "Chemical Substance SciData JSON-LD file for molecular oxygen",
    "authors": [ [7 lines]
    "description": "Metadata, identifiers, descriptors and molecular graph about a chemical substance",
    "publisher": "Chalk Research Laboratory, University of North Florida",
    "version": "1.1",
    "keywords": ["chemical twin","digital twin","chemical compound"],
    "permalink": "https://example.com/chemtwin/MYMOFIZGZYHOMD-UHFFFAOYSA-N/2",
    "toc": ["sdo:scientificData","sdo:system","sdo:compound","dc:source","dc:rights"],
    "ids": ["obo:IAO_0000578","ss:CHEMINF_000123","ss:CHEMINF_000022","obo:NCIT_C1940","obo:CHEBI_25555","ss:SIO_011118",
    ],
    "scidata": {
      "@id": "scidata",
      "@type": "sdo:scientificData",
      "discipline": "w3i:Chemistry",
      "system": {
        "@id": "system/",
        "@type": "sdo:system",
        "facets": [
          {
            "@id": "substance/1/",
            "@type": "sdo:substance",
            "iupacname": "molecular oxygen",
            "formula": "O2",
            "molweight": 31.999,
            "monoisotopicmass": 31.9898,
            "identifiers": { [16 lines]
            "descriptors": { [28 lines]
            "molgraph": { [42 lines]
          }
        ]
      }
    },
    "sources": [ [7 lines]
    "rights": [ [7 lines]
  }
}
```

Chemical Twin Concept

* Identifiers
* Descriptors
* Molgraph


```
"molgraph": {
    "@id": "molgraph/",
    "@type": "ss:CHEMINF_000022",
    "elements": [
        {
            "@id": "element/1/",
            "@type": "obo:NCIT_C1940",
            "name": "Oxygen",
            "element": "obo:CHEBI_25805"
        }
    ],
    "atoms": [
        {
            "@id": "atom/1/",
            "@type": "obo:CHEBI_33250",
            "element": "element/1/",
            "doublebonds": 1
        },
        {
            "@id": "atom/2/",
            "@type": "obo:CHEBI_33250",
            "element": "element/1/",
            "doublebonds": 1
        }
    ],
    "bonds": [
        {
            "@id": "bond/1/",
            "@type": "ss:SIO_011118",
            "order": "2",
            "atoms": [
                "atom/1/",
                "atom/2/"
            ]
        }
    ]
}
```

JSON-LD Context Files

- Context Files
  - SciData
  - Chemical Twin
    - Identifiers
    - Descriptors
    - Molgraph


```
{
  "@context": [
    {
      "@vocab": https://www.w3.org/2001/XMLSchema#,
      ss: https://semanticscience.org/ontology/sio.owl#,
      obo: http://purl.obolibrary.org/obo/,
      atoms: {
        "@id": "obo:CHEBI_33250"
      },
      bonds: {
        "@id": "ss:CHEMINF_000063"
      },
      elements: {
        "@id": "obo:CHEBI_24431"
      },
      atom: {
        "@id": "ss:SIO_010037",
        "@type": "@id"
      },
      bond: {
        "@id": "ss:SIO_011118",
        "@type": "@id"
      },
      charge: {
        "@id": "ss:CHEMINF_000120",
        "@type": "integer"
      },
      bonded: {
        "@id": "ss:SIO_000132",
        "@type": "@id"
      },
      element: {
        "@id": "obo:CHEBI_33250",
        "@type": "@id"
      },
```

Searching RDF

query | add data | edit | info

SPARQL Query

To try out some SPARQL queries against the selected dataset, enter your query here.

Example Queries

Selection of triples | Selection of classes

Prefixes

rdf | rdfs | owl | xsd

Content Type (SELECT): JSON

Content Type (GRAPH): Turtle

```
1  SELECT ?graph ?object
2  WHERE {
3    GRAPH ?graph {
4      ?subject <http://www.wikidata.org/prop/direct/P235> ?object
5      }
6  }
7  LIMIT 100
```

Table | Response | 22 results in 0.024 seconds | Simple view ☐ | Ellipse ☑ | Filter query results | Page size: 50

	graph	object
1	<https://scidata.unf.edu/facet/00002661>	"WPYMKLBDIGXBTP-UHFFFAOYSA-N"^^<https://www.w3.org/2001/XMLSchema#string>
2	<https://scidata.unf.edu/facet/00002677>	"JVTAAEKCZFNVCJ-UHFFFAOYSA-N"^^<https://www.w3.org/2001/XMLSchema#string>
3	<https://scidata.unf.edu/facet/00002680>	"YGSDEFSMJLZEOE-UHFFFAOYSA-N"^^<https://www.w3.org/2001/XMLSchema#string>
4	<https://scidata.unf.edu/facet/00002740>	"HEFNNWSXXWATRW-UHFFFAOYSA-N"^^<https://www.w3.org/2001/XMLSchema#string>
5	<https://scidata.unf.edu/facet/00002954>	"SLXKOJJOQWFEFD-UHFFFAOYSA-N"^^<https://www.w3.org/2001/XMLSchema#string>

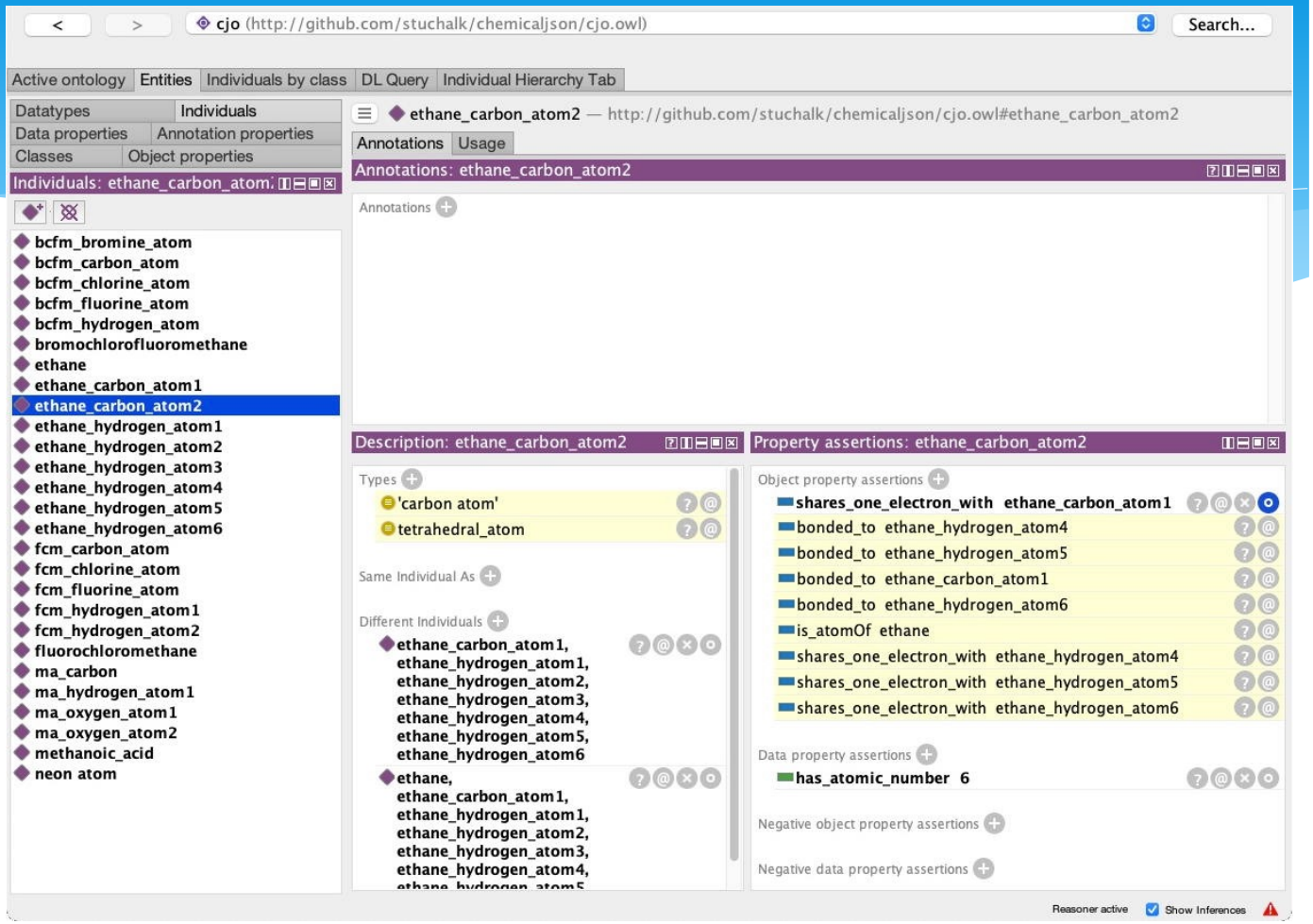
Analyzing Triples

* 10471 compounds 20 elements
* 7.5 million triples (~791 triples/file)
* ‘Classification’ - ChemOnt Ontology
* ‘Chemical entity’ – individual elements

	Predicate	Count	Definition
2	<http://purl.allotrope.org/ontologies/property#AFX_0001154>	161324	classification
3	<http://purl.obolibrary.org/obo/CHEBI_24431>	46403	chemical entity
4	<http://purl.obolibrary.org/obo/CHEBI_33250>	2162237	atom
5	<http://purl.obolibrary.org/obo/IAO_0000129>	20982	version number
6	<http://purl.obolibrary.org/obo/IAO_0000578>	10491	centrally registered identifier
7	<http://purl.obolibrary.org/obo/IAO_0000590>	10491	written name
13	<http://www.wikidata.org/prop/direct/P1578>	563	Gmelin number
14	<http://www.wikidata.org/prop/direct/P1579>	1573	Reaxys registry number
15	<http://www.wikidata.org/prop/direct/P2017>	10453	isomeric SMILES
16	<http://www.wikidata.org/prop/direct/P231>	3255	CAS Registry Number
17	<http://www.wikidata.org/prop/direct/P233>	9893	canonical SMILES
18	<http://www.wikidata.org/prop/direct/P234>	10491	InChI
19	<http://www.wikidata.org/prop/direct/P235>	10491	InChIKey
25	<https://bioportal.bioontology.org/ontologies/EDAM#data_3103>	4779	ATC Code
26	<https://purl.org/dc/terms/description>	10491	description
27	<https://purl.org/dc/terms/hasPart>	52455	file has a part
28	<https://purl.org/dc/terms/identifier>	111150	identifier
29	<https://purl.org/dc/terms/license>	10491	usage license
30	<https://purl.org/dc/terms/publisher>	10491	publisher
31	<https://purl.org/dc/terms/rightsHolder>	10491	rights holder
32	<https://purl.org/dc/terms/subject>	31473	subject
33	<https://purl.org/dc/terms/title>	20982	title
34	<https://schema.org/url>	20982	url
49	<https://semanticscience.org/ontology/sio.owl#CHEMINF_000254>	10456	rotatable bond count
50	<https://semanticscience.org/ontology/sio.owl#CHEMINF_000280>	10471	covalent unit count
51	<https://semanticscience.org/ontology/sio.owl#CHEMINF_000300>	10471	heavy atom count
52	<https://semanticscience.org/ontology/sio.owl#CHEMINF_000301>	10471	isotope atom count
53	<https://semanticscience.org/ontology/sio.owl#CHEMINF_000381>	9232	aromatic cycle count
54	<https://stuchalk.github.io/scidata/ontology/scidata.owl#doublebondcount>	146901	chemical bond of order 2
61	<https://stuchalk.github.io/scidata/ontology/scidata.owl#scientificDiscipline	10491	scientific discipline 'chemistry'
62	<https://stuchalk.github.io/scidata/ontology/scidata.owl#singlebondcount>	500622	chemical bond of order 1
63	<https://stuchalk.github.io/scidata/ontology/scidata.owl#triplebondcount>	1778	chemical bond of order 3
64	<https://w3id.org/reproduceme#ORCID>	10491	ORCID
65		7561885	

Semantic Reasoning

* Build an OWL ontology to reason (infer) other chemical information

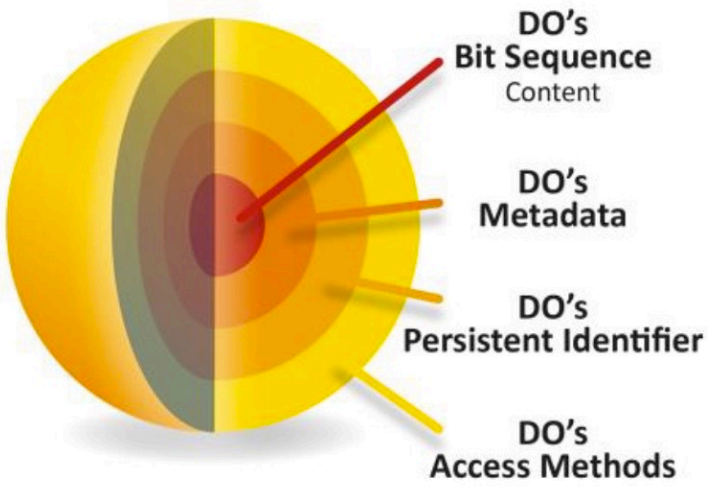


"Open semantic chemical structures: Ideas on the use of JSON-LD for representation of chemical entities", Stuart J. Chalk, paper presented at the 254th ACS Meeting in Washington, DC August 2017

FAIR Digital Objects

* FAIR Digital Objects (FDOs)

  "A FAIR digital object is a unit composed of data and/or metadata regulated by structures or schemas, and with an assigned globally unique and persistent identifier (PID), which is findable, accessible, interoperable and reusable both by humans and computers for the reliable interpretation and processing of the data represented by the object."

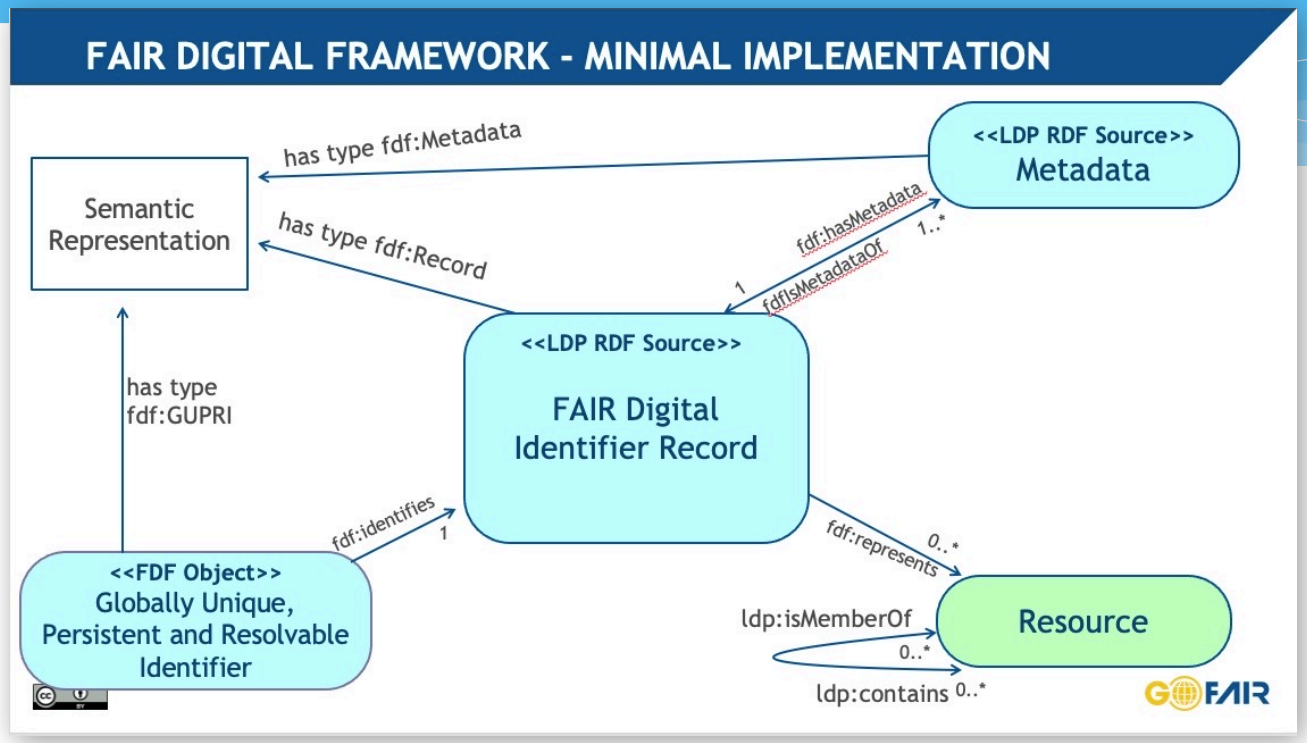


FDO Base Definition https://fairdo.org/library/

https://datashare.rzg.mpg.de/s/RTeYZGe3QMgEciH/download?path=%2FFDO%20Public%20Documents%2FSpecification-Docs&files=Machine%20Actionability%20for%20FDOs-1-1.pdf

https://www.dona.net/digitalobjectarchitecture

FAIR Digital Objects



https://github.com/GEDE-RDA-Europe/GEDE/blob/master/FAIR%20Digital%20Objects/Paris-FDO-workshop/GEDE_Paris_Session%201_Bonino.pptx

Chemical Twins as FAIR Digital Objects

* Cordra DO Software
  https://www.cordra.org/
  * Implements FDO architecture
  * Provides "operations' interface
  * Open source
* Python REST Client – CordraPy
  https://github.com/usnistgov/CordraPy

More Info … Give Us Your Thoughts

* ChemTwin GitHub Repository
  https://github.com/stuchalk/ChemTwins-dev

* Email: schalk@unf.edu
* Skype: stuartchalk
* LinkedIn: https://www.linkedin.com/in/stuchalk
* ORCID: http://orcid.org/0000-0002-0703-7776

National Science Foundation
WHERE DISCOVERIES BEGIN

Grant #1835643