GOing FAIR & DOing FAIR

GEDE Workshop on Digital Objects
26.9.2018
APCO, Rue Montoyer 47, 1000 Bruxelles, Belgium

Erik Schultes, PhD
International Science Coordinator
GO FAIR International Support and Coordination Office, Leiden
erik.schultes@go-fair.org
go-fair.org

The **Digital Object framework** is an abstraction layer striving towards technology-independent, future-proof, and increasingly automated operations between data, software, and compute resources.

The **15 FAIR Principles** are high-level specifications for the automated Findability, Accession, semantic Interoperation, and Re-use of data, software and services.

Re: relation between FAIR Principles and DOs

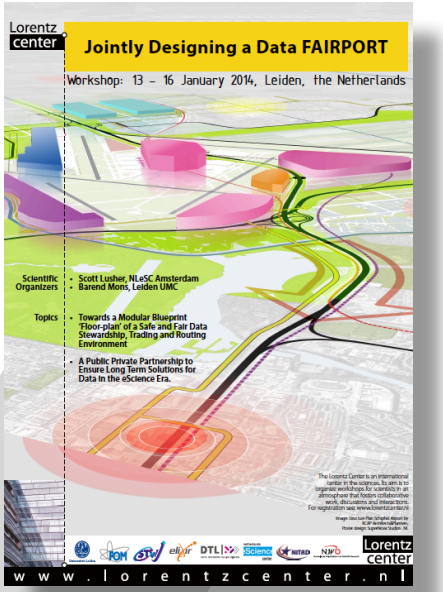
Modified from Larry Lannom
17 September 2018

A Framework for Distributed Digital Object Services

Robert Kahn
Corporation for National Research Initiatives

Robert Wilensky
University of California at Berkeley

May 13, 1995
cnri.dlib/tn95-01



2012: Multiple national research funding bodies conclude that an international effort is needed to build social and technical bridges needed for open sharing and re-use of data

2014: Initial set of RDA Working Groups focusing on core infrastructures form the Data Fabric group

Research Data Alliance (RDA)

2014-2107: Data Fabric adopts a model and framework for developing the technical components needed across all research data management. Abstraction to an object level with each object persistently identified is seen as key to future development

RDA Data Fabric

2017: A set of large research infrastructure providers come together in the C2CAMP initiative to operationalize the digital object model

2018: C2CAMP member organizations begin demonstration and production efforts structuring, curating and interoperating across collections of Digital Objects.

C2CAMP Digital Object Architecture

FORCE11 Community

Scholars, librarians, archivists, publishers and research funders changing scholarly communication through the use of information technology.

FAIR Principles

The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data (2016), https://www.nature.com/articles/sdata201618

GO FAIR - GO BUILD Implementation Networks

2017: A consortium committed to defining and creating materials and tools as elements of the Internet of FAIR Data and Services.

Synergy with GO FAIR effort is clear and C2CAMP becomes a GO Fair Implementation Network

Digital Object approach provides technical solutions needed for each of the FAIR principles:

- *Findable* requires persistent identification
- *Accessible* (over time) requires access metadata bound to the object instead of its transitory computing environment
- *Interoperable* requires implementation details to be abstracted away, with objects identified independently of owner/location and defined by function, type, and content, not current tech detail
- *Reusable* requires all of the above

Implementing the FAIR Principles With Digital Objects as Fundamental Technical Driver

GO FAIR

A Framework for Distributed Digital Object Services

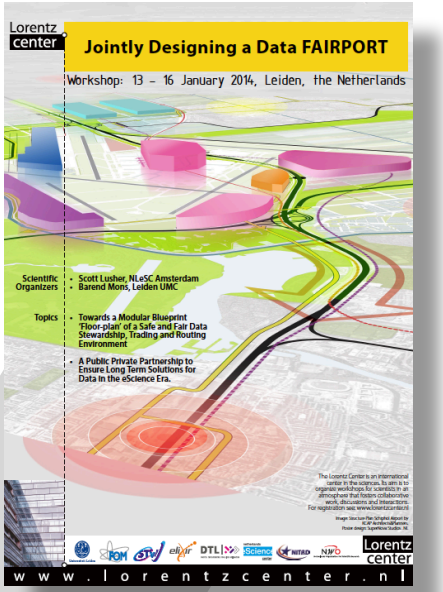
Robert Kahn
Corporation for National Research Initiatives

Robert Wilensky
University of California at Berkeley

May 13, 1995
cnri.dlib/tn95-01

Re: relation between FAIR Principles and DOs

Modified from Larry Lannom
17 September 2018



2012: Multiple national research funding bodies conclude that an international effort is needed to build social and technical bridges needed for open sharing and re-use of data

Theoretical

Empirical

2014: Initial set of RDA Working Groups focusing on core infrastructures form the Data Fabric group

Research Data Alliance (RDA)

2014-2107: Data Fabric adopts a model and framework for developing the technical components needed across all research data management. Abstraction to an object level with each object persistently identified is seen as key to future development

RDA Data Fabric

2017: A set of large research infrastructure providers come together in the C2CAMP initiative to operationalize the digital object model

2018: C2CAMP member organizations begin demonstration and production efforts structuring, curating and interoperating across collections of Digital Objects.

C2CAMP Digital Object Architecture

FORCE11 Community

Scholars, librarians, archivists, publishers and research funders changing scholarly communication through the use of information technology.

FAIR Principles

The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data (2016), https://www.nature.com/articles/sdata201618

GO FAIR - GO BUILD Implementation Networks

2017: A consortium committed to defining and creating materials and tools as elements of the Internet of FAIR Data and Services.

Synergy with GO FAIR effort is clear and C2CAMP becomes a GO Fair Implementation Network

Digital Object approach provides technical solutions needed for each of the FAIR principles:

- *Findable* requires persistent identification
- *Accessible* (over time) requires access metadata bound to the object instead of its transitory computing environment
- *Interoperable* requires implementation details to be abstracted away, with objects identified independently of owner/location and defined by function, type, and content, not current tech detail
- *Reusable* requires all of the above

Implementing the FAIR Principles With Digital Objects as Fundamental Technical Driver

GO FAIR

FAIR and GO FAIR

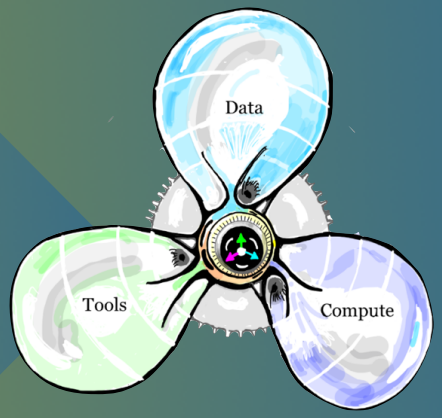
Lorentz

EOSC

FAIRdICT

GO FAIR

IFDS

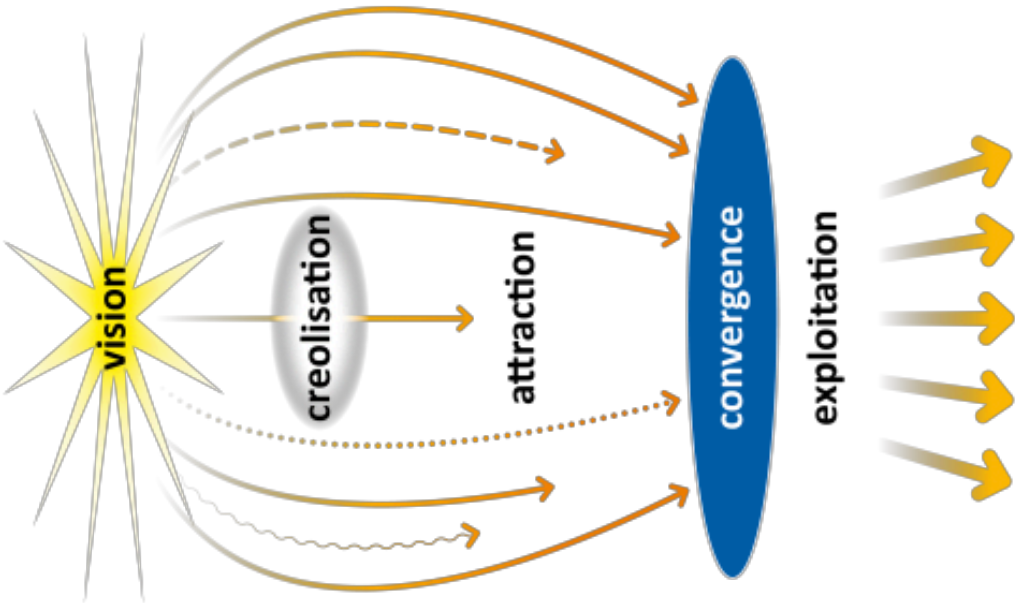


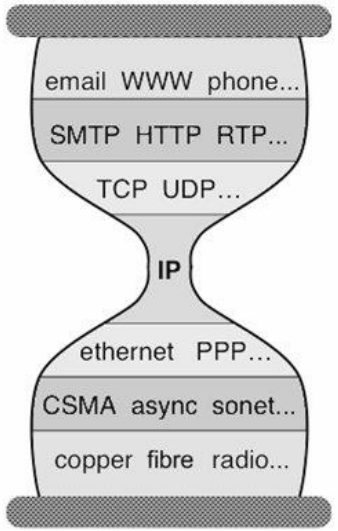
Birth	Infancy		Adolescence		Maturity
2014	2015	2016	2017	2018…	

FAIR and GO FAIR

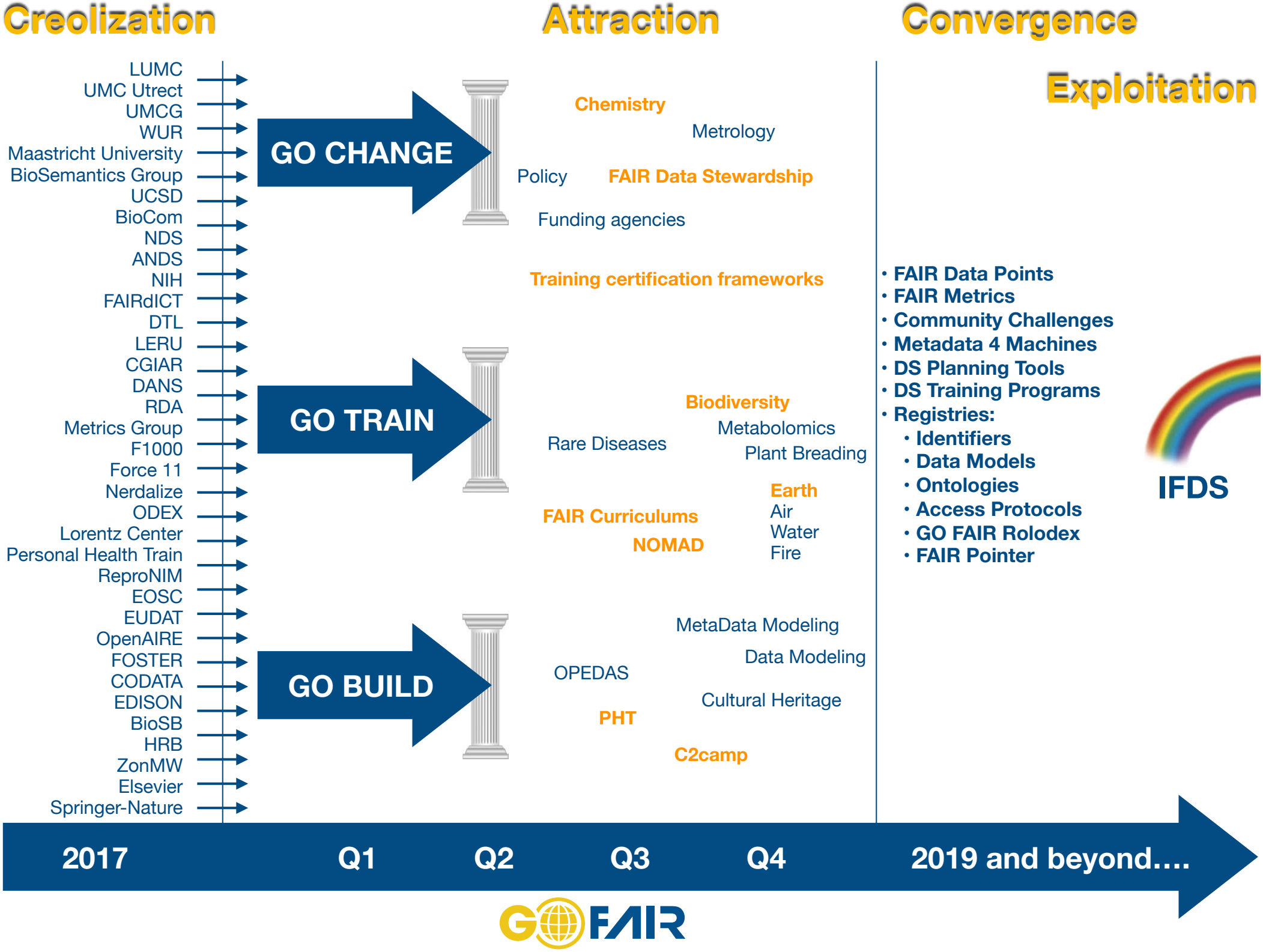
Common Patterns in Revolutionary Infrastructures and Data

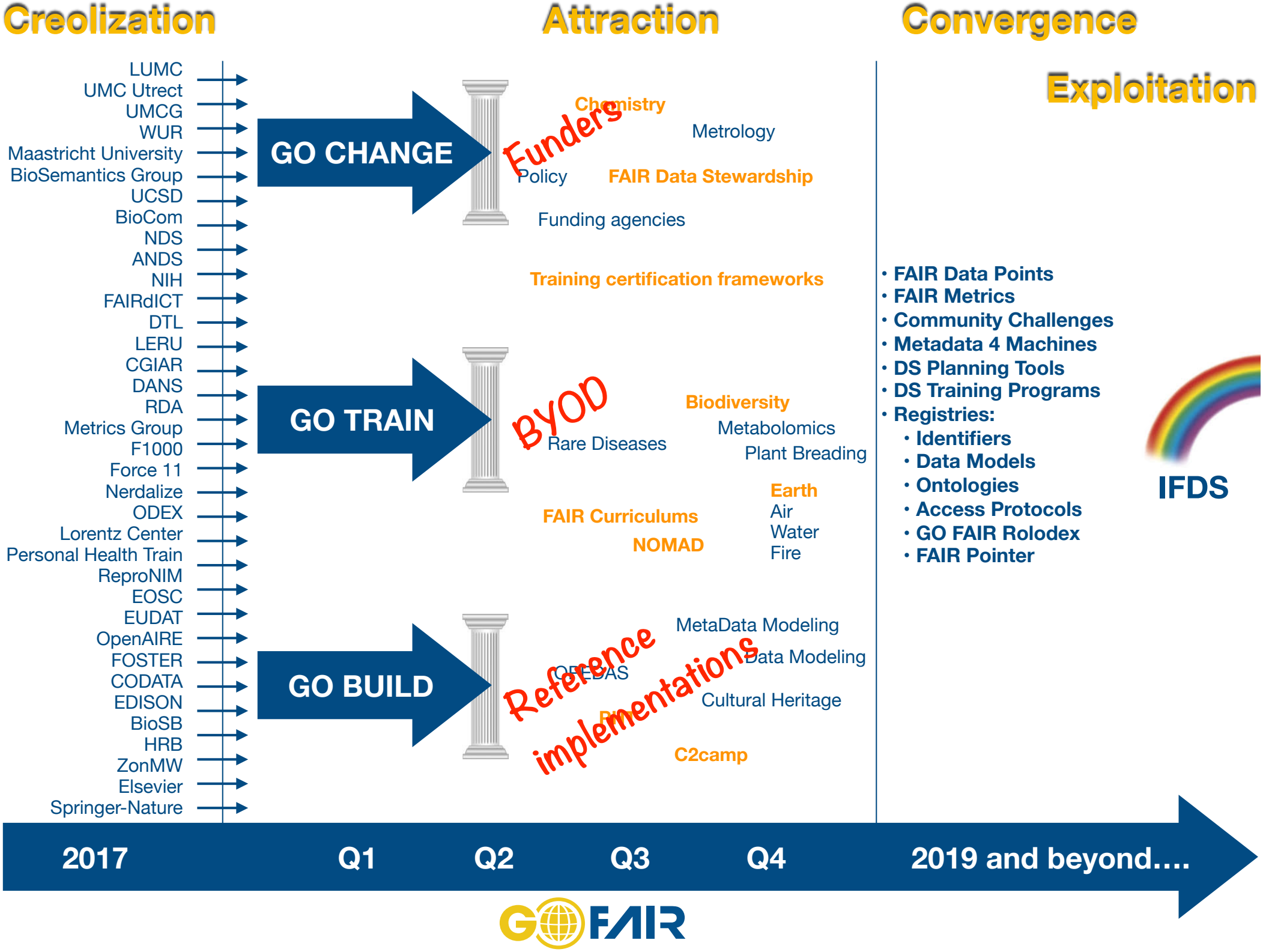
Peter Wittenburg, Max Planck Computing and Data Facility
George Strawn, US National Academy of Sciences
February 2018
https://www.rd-alliance.org/sites/default/files/Common_Patterns_in_Revolutionising_Infrastructures-final.pdf





- Minimal standards
- Voluntary
- Critical Mass


Creolization
Attraction
Convergence
Exploitation
LUMC
UMC Utrect
UMCG
WUR
Maastricht University
BioSemantics Group
UCSD
BioCom
NDS
ANDS
NIH
FAIRdICT
DTL
LERU
CGIAR
DANS
RDA
Metrics Group
F1000
Force 11
Nerdalize
ODEX
Lorentz Center
Personal Health Train
ReproNIM
EOSC
EUDAT
OpenAIRE
FOSTER
CODATA
EDISON
BioSB
HRB
ZonMW
Elsevier
Springer-Nature
GO CHANGE
Chemistry
Metrology
Policy
FAIR Data Stewardship
Funding agencies
Training certification frameworks
GO TRAIN
Biodiversity
Metabolomics
Rare Diseases
Plant Breading
Earth
Air
Water
Fire
FAIR Curriculums
NOMAD
GO BUILD
MetaData Modeling
Data Modeling
OPEDAS
Cultural Heritage
PHT
C2camp
• FAIR Data Points
• FAIR Metrics
• Community Challenges
• Metadata 4 Machines
• DS Planning Tools
• DS Training Programs
• Registries:
• Identifiers
• Data Models
• Ontologies
• Access Protocols
• GO FAIR Rolodex
• FAIR Pointer
IFDS
2017
Q1
Q2
Q3
Q4
2019 and beyond….
GO FAIR


Creolization
Attraction
Convergence
Exploitation
LUMC
UMC Utrect
UMCG
WUR
Maastricht University
BioSemantics Group
UCSD
BioCom
NDS
ANDS
NIH
FAIRdICT
DTL
LERU
CGIAR
DANS
RDA
Metrics Group
F1000
Force 11
Nerdalize
ODEX
Lorentz Center
Personal Health Train
ReproNIM
EOSC
EUDAT
OpenAIRE
FOSTER
CODATA
EDISON
BioSB
HRB
ZonMW
Elsevier
Springer-Nature
GO CHANGE
GO TRAIN
GO BUILD
Funders
Chemistry
Metrology
Policy
FAIR Data Stewardship
Funding agencies
Training certification frameworks
BYOD
Biodiversity
Metabolomics
Rare Diseases
Plant Breading
Earth
Air
Water
Fire
FAIR Curriculums
NOMAD
Reference implementations
MetaData Modeling
Data Modeling
OPENDAS
Cultural Heritage
C2camp
• FAIR Data Points
• FAIR Metrics
• Community Challenges
• Metadata 4 Machines
• DS Planning Tools
• DS Training Programs
• Registries:
• Identifiers
• Data Models
• Ontologies
• Access Protocols
• GO FAIR Rolodex
• FAIR Pointer
IFDS
2017
Q1
Q2
Q3
Q4
2019 and beyond….
GO FAIR

“Data and services that are
findable,
accessible,
interoperable,
re-usable
both for machines and for people.”

The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data (2016), https://www.nature.com/articles/sdata201618

GO FAIR

“Data and services that are
findable,
accessible,
interoperable,
re-usable
both for machines and for people.”

The FAIR Guiding Principles for scientific data management and stewardship,
Scientific Data (2016), https://www.nature.com/articles/sdata201618

GO FAIR

15 FAIR Principles

Findable:

F1 (meta)data are assigned a globally unique and persistent identifier;

F2 data are described with rich metadata;

F3 metadata clearly and explicitly include the identifier of the data it describes;

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

Accessible:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol;

- **A1.1 the protocol is open, free, and universally implementable;**
- **A1.2 the protocol allows for an authentication and authorization procedure, where necessary;**

A2 metadata are accessible, even when the data are no longer available;

Interoperable:

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

I2 (meta)data use vocabularies that follow FAIR principles;

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

Reusable:

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

- **R1.1 (meta)data are released with a clear and accessible data usage license;**
- **R1.2 (meta)data are associated with detailed provenance;**
- **R1.3 (meta)data meet domain-relevant community standards;**

The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data (2016), https://www.nature.com/articles/sdata201618

GO FAIR

15 FAIR Principles & DOA

Findable:

PID

F1 (meta)data are assigned a globally unique and persistent identifier;

Type

F2 data are described with rich metadata;

Binding

F3 metadata clearly and explicitly include the identifier of the data it describes;

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

Type Registries

Accessible:

DOIP

A1 (meta)data are retrievable by their identifier using a standardized communications protocol;

A1.1 the protocol is open, free, and universally implementable;

A1.2 the protocol allows for an authentication and authorization procedure, where necessary;

A2 metadata are accessible, even when the data are no longer available;

Binding

Interoperable:

Meaningful Messeging

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

I2 (meta)data use vocabularies that follow FAIR principles;

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

Abstraction: Future Proofing

Reusable:

Type

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

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

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

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

The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data (2016), https://www.nature.com/articles/sdata201618

GO FAIR

relation between FAIR Principles and DOs

Modified from Peter Wittenburg, 30 August 2018

To be Findable:	**In all circumstance the DO concept ensures that DOs can either be found by using the PID or by using the metadata. Both give humans and machines access to all entities of the DO.**
F1. (meta)data are assigned a globally unique and eternally persistent identifier.	This is inherent to the DO model. All DOs have PIDs including metadata objects, collections and other digital entities.
F2. data are described with rich metadata.	The DO model only guarantees that there is place for different types of metadata and that there is a clear and stable relationship, i.e. even machines can find the metadata types
F3. (meta)data are registered or indexed in a searchable resource.	The DO concept makes clear that metadata have identities, i.e. they can be exposed and harvested without breaking the relationship to the DO and thus its bit sequence for example which is of great importance.
F4. metadata specify the data identifier.	The DO concept states that the PID of the DO is to be found in the metadata and that the the PID record should include the PID of the metadata. Double referencing ensures that it does not matter whether clients first have a PID or first come from metadata search.

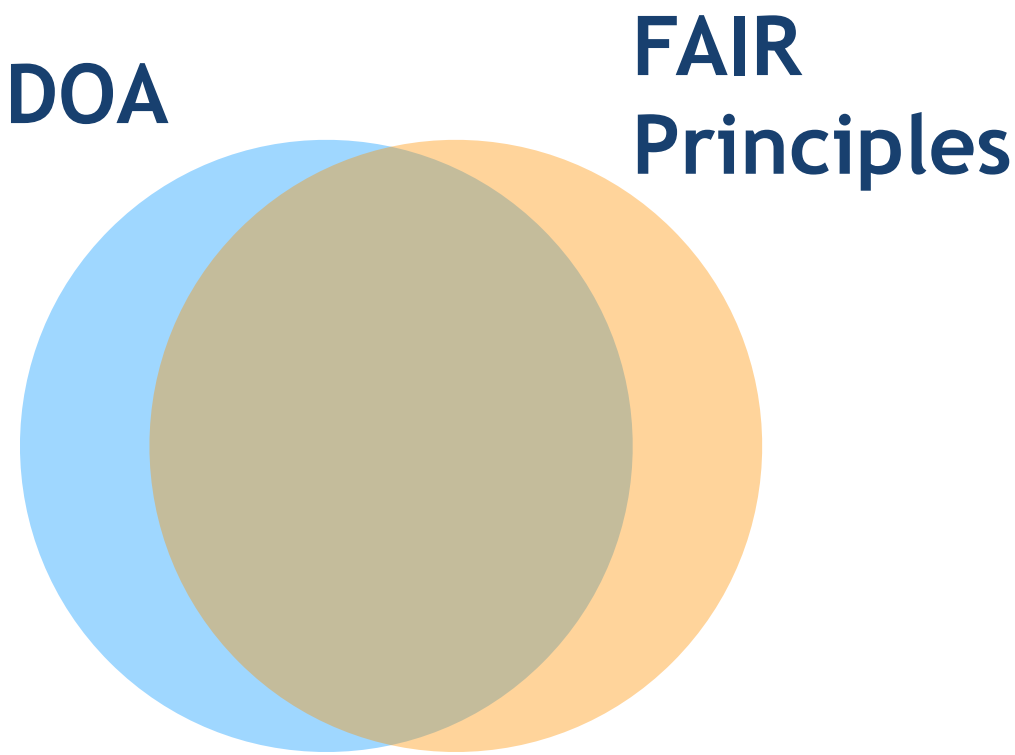
To be Accessible:	**The DO concept enables to build an infrastructure that makes data/metadata accessible and it supports all requirements.**
A1 (meta)data are retrievable by their identifier using a standardized communication protocol.	The model states that data and metadata have PIDs. The Handle System has a standard protocol to resolve a PID to its attributes. In addition, the DOIP will guarantee that there is standardised way to access all entities of a DO independent of its implementation (files, clouds, databases, etc.)
A1.1 the protocol is open, free, and universally implementable.	The PID protocol is open, free and universally implementable. The DOIP protocol is new and is currently in discussion, but it will fulfil the criteria.
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.	Security mechanisms are being supported. The Handle protocol requires certificate based authentication to prevent misuse by unauthorised persons. The security features of DOIP are currently under discussion. The PID record, however, allows including pointers to access permission records, to block chain entries including smart contracts and transactions clearly linked with an identified DO.
A2 metadata are accessible, even when the data are no longer available.	The DO concept allows deleting the bit sequences but maintain the other entities. PID and metadata should be modified in a way that users will be informed about this event.

To be Interoperable:	**DOs take care of interoperability at the level of data management, access etc., however, we can only say that they facilitate interoperability for example at semantic level.**
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	This is partly out of the scope for DOs, however, DOs facilitate syntactic and semantic interoperability since it allows humans and machines to find schemas and concept definitions in a stable way. There is a small comment that data and metadata goes beyond knowledge presentation and heterogeneity at all levels is huge - but this is a gap in the FAIR principles :) The major costs in industry in data integration are on the level of a bad data organisation and model differences.
I2. (meta)data use vocabularies that follow FAIR principles.	DOs support the FAIR principles in so far as they are as vocabs are DOs as other digital entities and thus can be referenced etc using the same basic mechanisms and the binding
I3. (meta)data include qualified references to other (meta)data.	Here DOs help to achieve the goal of stable references through its mechanisms at the level of data organisation etc. Of course it is the task of other actors (humans, machines) to get the references to other DOs based on content etc.

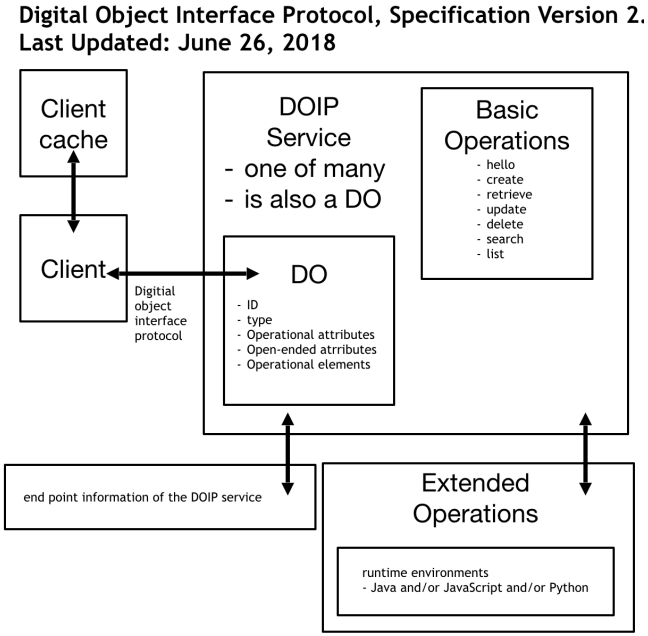
To be Re-usable:	**At some level DOs enable smart re-usability. Some aspects DOs only facilitate.**
R1. meta(data) have a plurality of accurate and relevant attributes.	
R1.1. (meta)data are released with a clear and accessible data usage license.	DOs mechanisms can be tuned in so far that the attributes for example point to a blockchain entry including smart contracts which are basically actionable licenses and as we know one can use this blockchain entry also to record all transactions. It's again the binding facility of DOs that can get this working in a stable and safe way. I discussed this with industry and for them this combination with clear identity and safe blockchain entries is attractive
R1.2. (meta)data are associated with their provenance.	Well this can also be accomplished using the binding mechanism. Some communities add provenance to their metadata either directly or indirectly through a pointer in metadata. One could also include an attribute in the PID record to get this done. PS: most people I know of do not store provenance of metadata records. There is another dimension: since (descriptive) metadata should be open, it is being copied, enriched, modified etc. by interested people for whatever purpose. This all is out of control. So we can only look at the "authorised" metadata (don't know a good term here).
R1.3. (meta)data meet domain-relevant community standards.	this is out of scope of DOs of course except that it helps again in stable referencing etc.

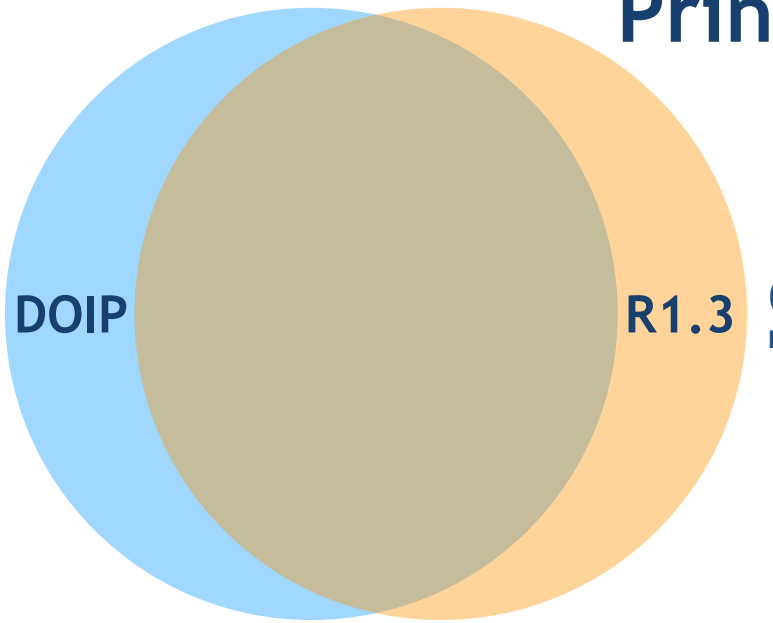
GO FAIR

All Digital Objects are FAIR (FAIR Objects?)



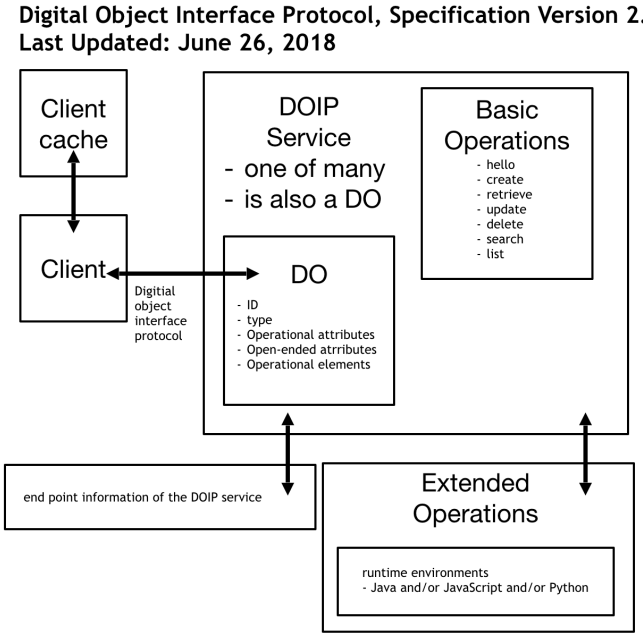
All Digital Objects are FAIR (FAIR Objects?)

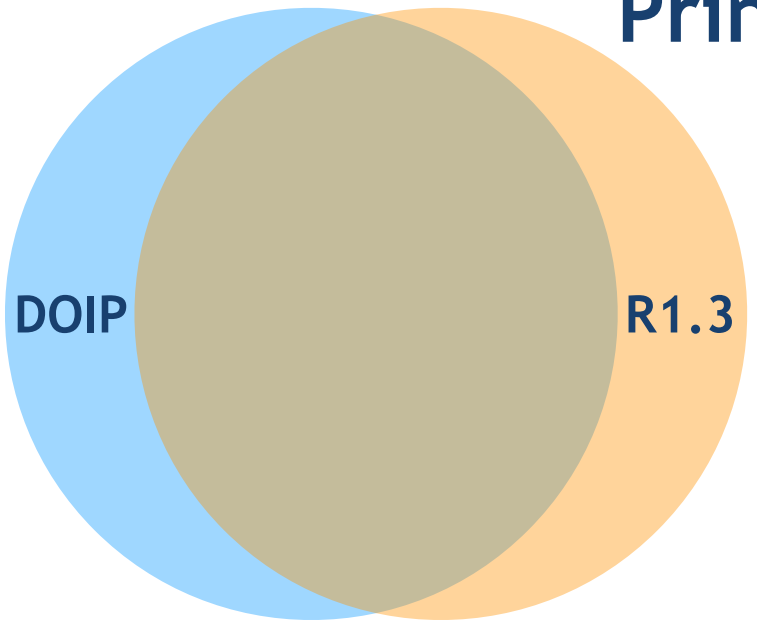




(meta)data meet domain-relevant community standards;

All Digital Objects are FAIR (FAIR Objects?)

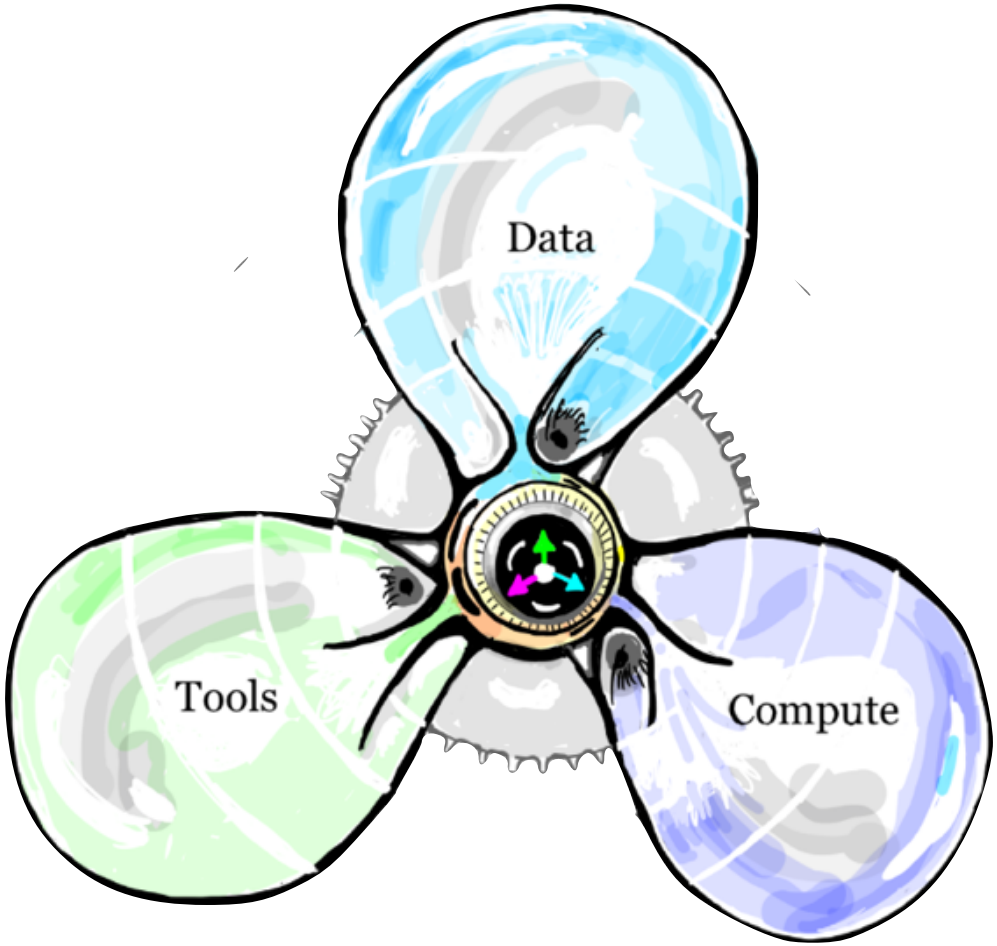




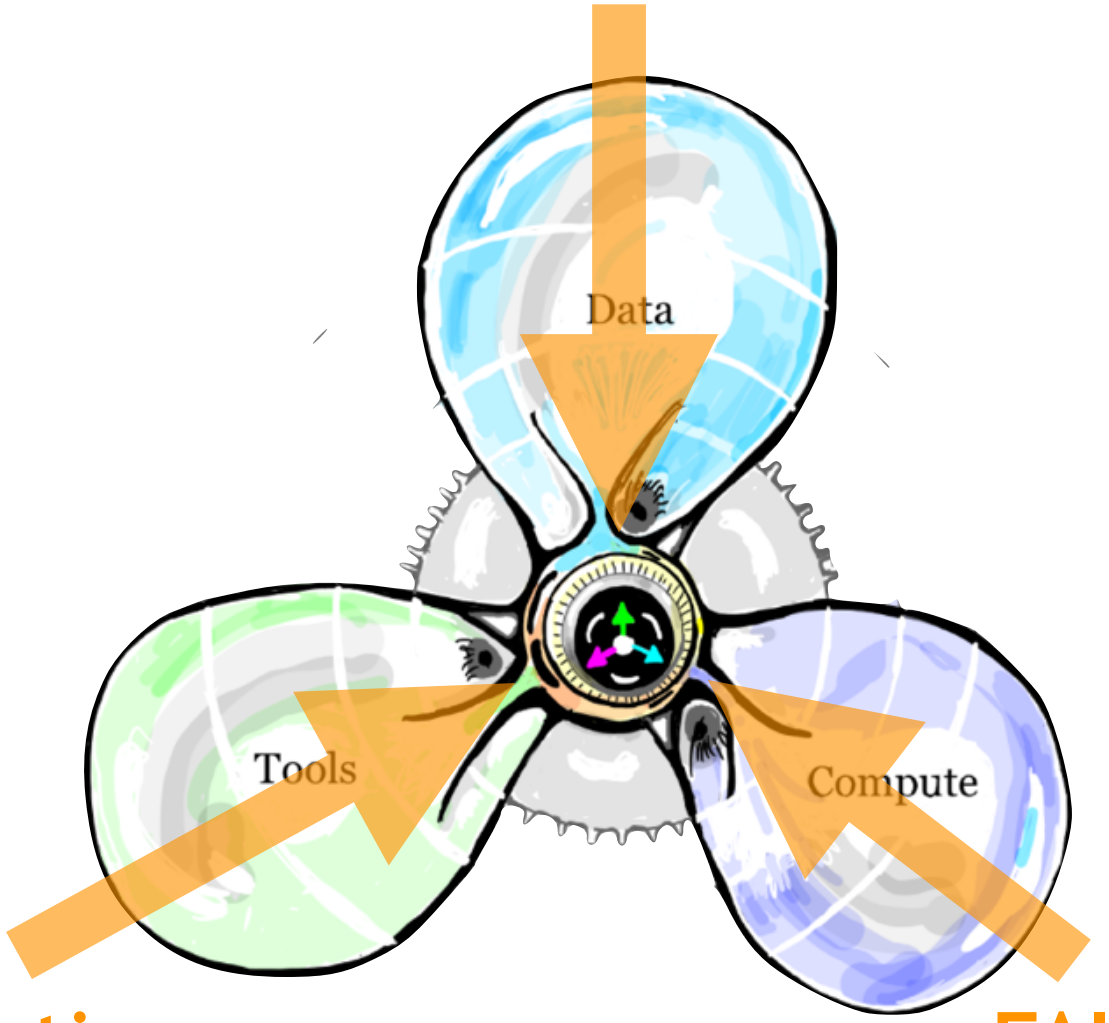
(meta)data meet domain-relevant community standards;

- FAIR Metrics
- Community Challenges
- Metadata 4 Machines
- DS Planning Tools
- DS Training Programs

GEDE Digital Object Topic Group


Data
Tools
Compute

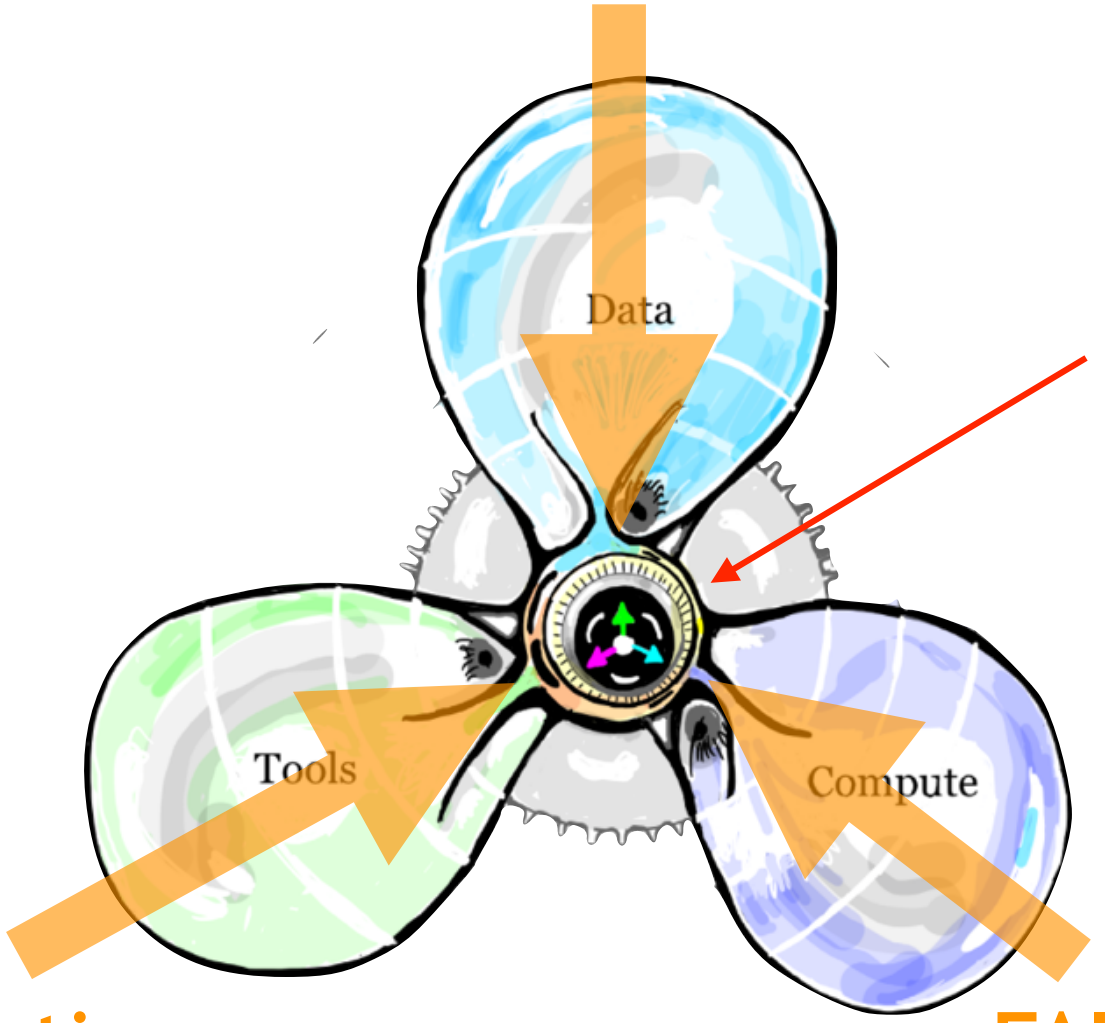
FAIRification “GOing FAIR”



FAIRification

FAIRification

FAIRification “GOing FAIR”



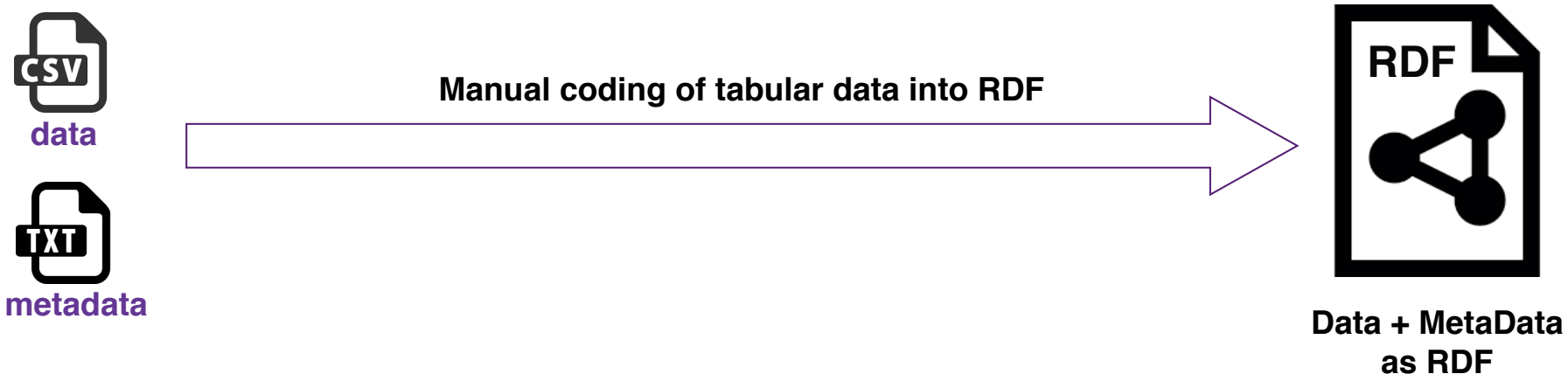
DOIP
“DOing FAIR”

FAIRification

FAIRification

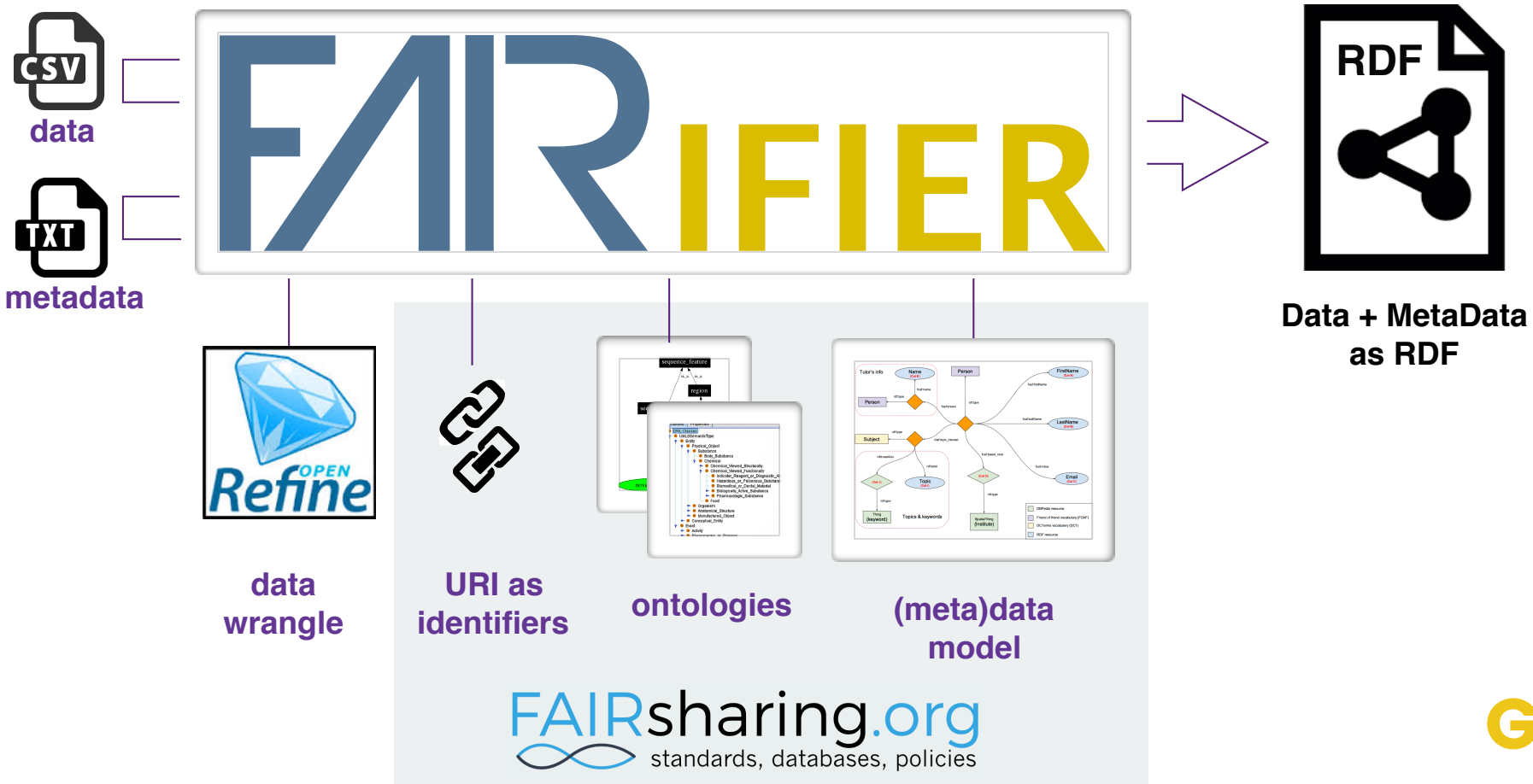
A Reference Implementation of the 15 Principles to show how Data can be made FAIR for machines

- FAIR-dICT project,DTL: https://www.dtls.nl/fair-data/fair-dict/
- Interoperability and FAIRness through a novel combination of Web technologies, https://peerj.com/articles/cs-110/



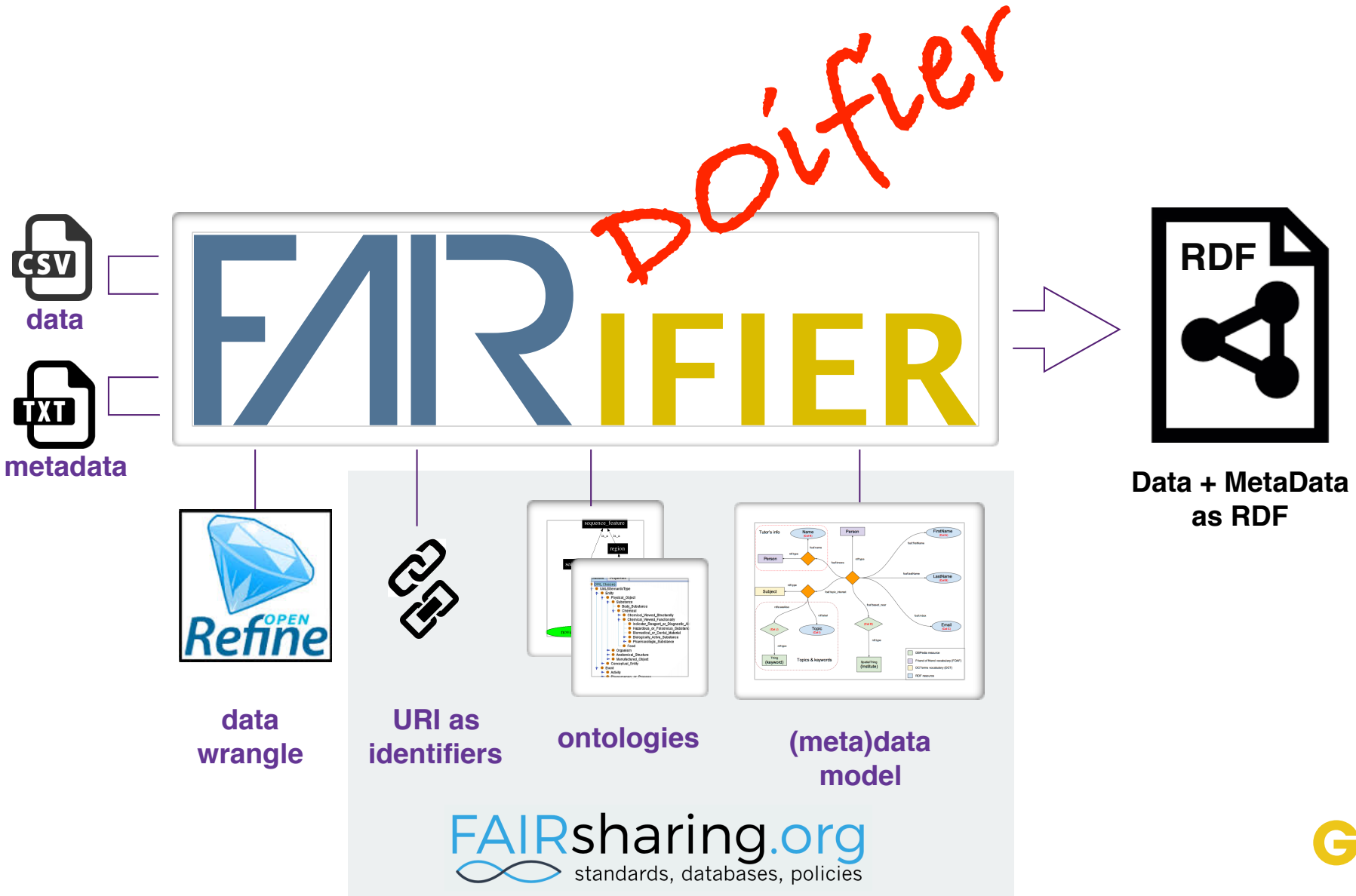
A Reference Implementation of the 15 Principles to show how Data can be made FAIR for machines

- FAIR-dICT project,DTL: https://www.dtls.nl/fair-data/fair-dict/
- Interoperability and FAIRness through a novel combination of Web technologies, https://peerj.com/articles/cs-110/



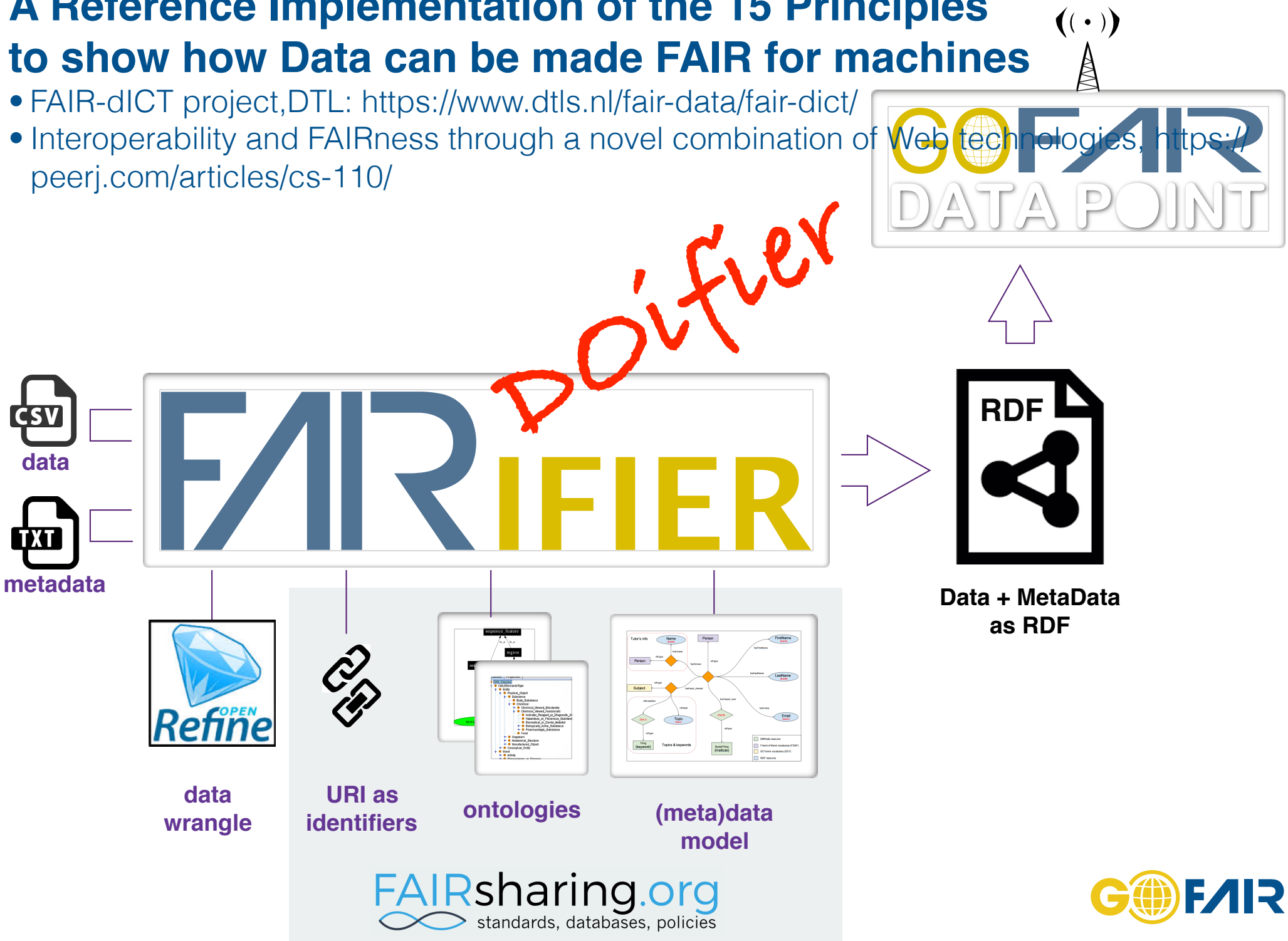
A Reference Implementation of the 15 Principles to show how Data can be made FAIR for machines

- FAIR-dICT project,DTL: https://www.dtls.nl/fair-data/fair-dict/
- Interoperability and FAIRness through a novel combination of Web technologies, https://peerj.com/articles/cs-110/



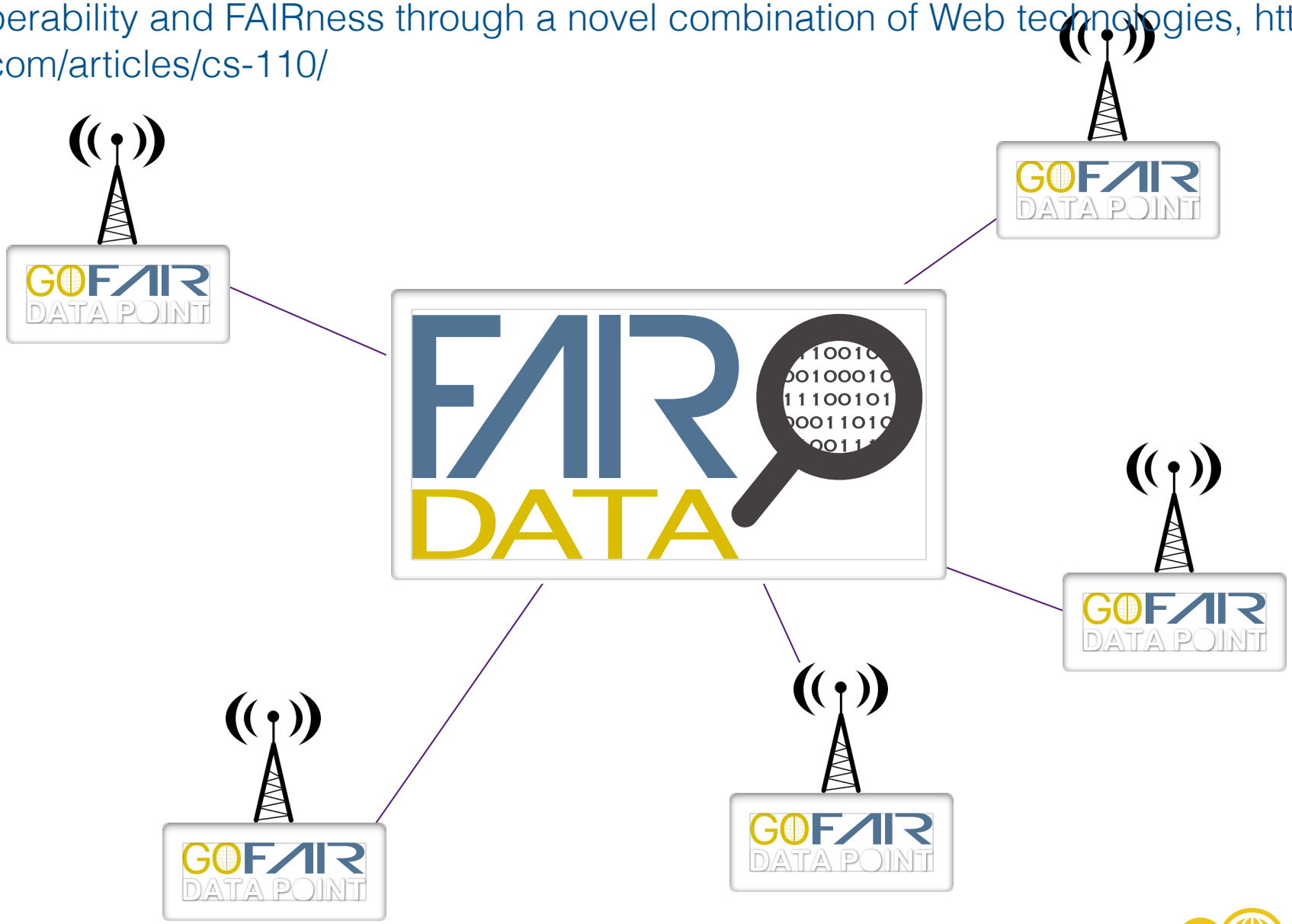
A Reference Implementation of the 15 Principles to show how Data can be made FAIR for machines

- FAIR-dICT project,DTL: https://www.dtls.nl/fair-data/fair-dict/
- Interoperability and FAIRness through a novel combination of Web technologies, https://peerj.com/articles/cs-110/



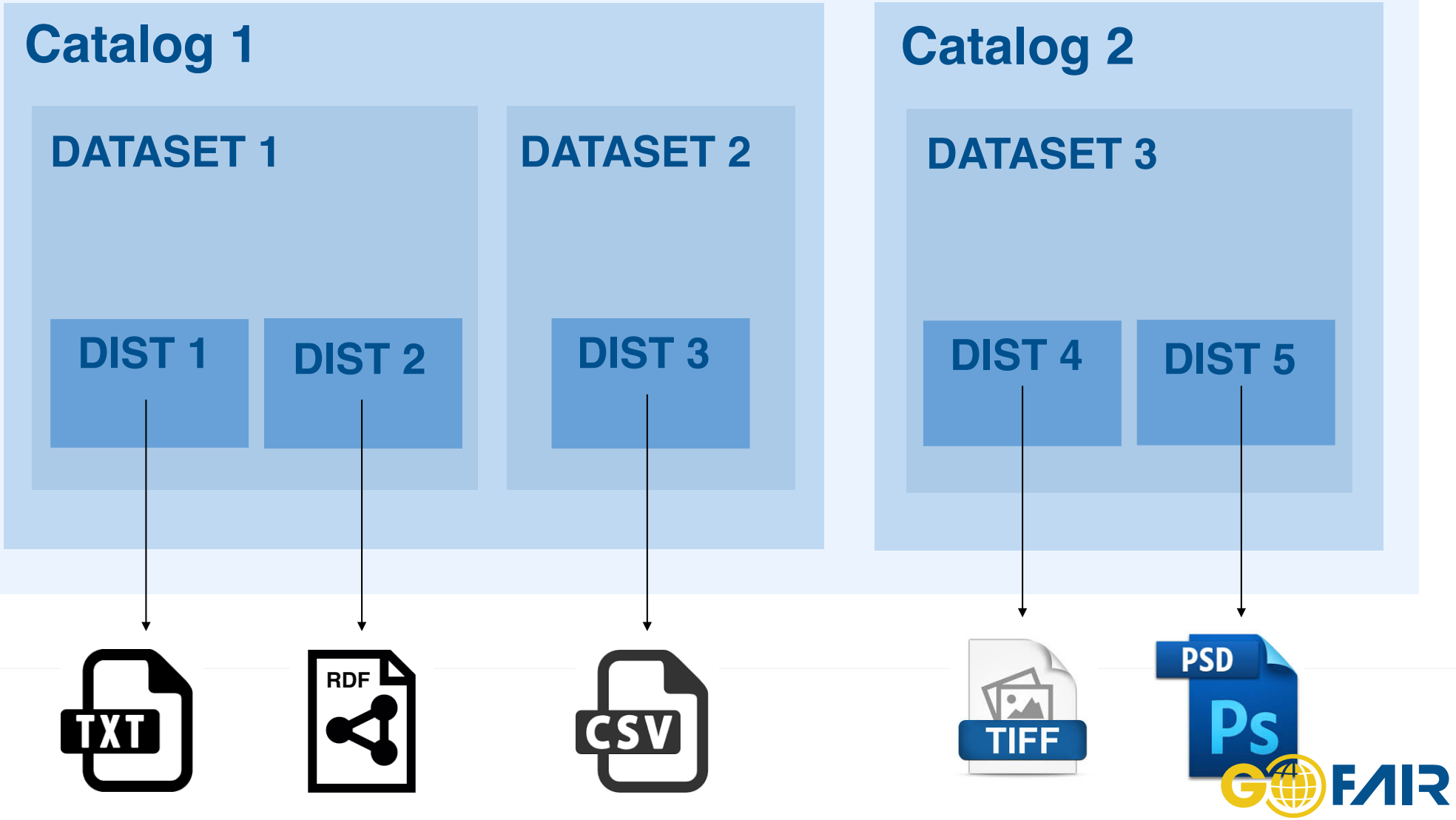
A Reference Implementation of the 15 Principles to show how Data can be made FAIR for machines

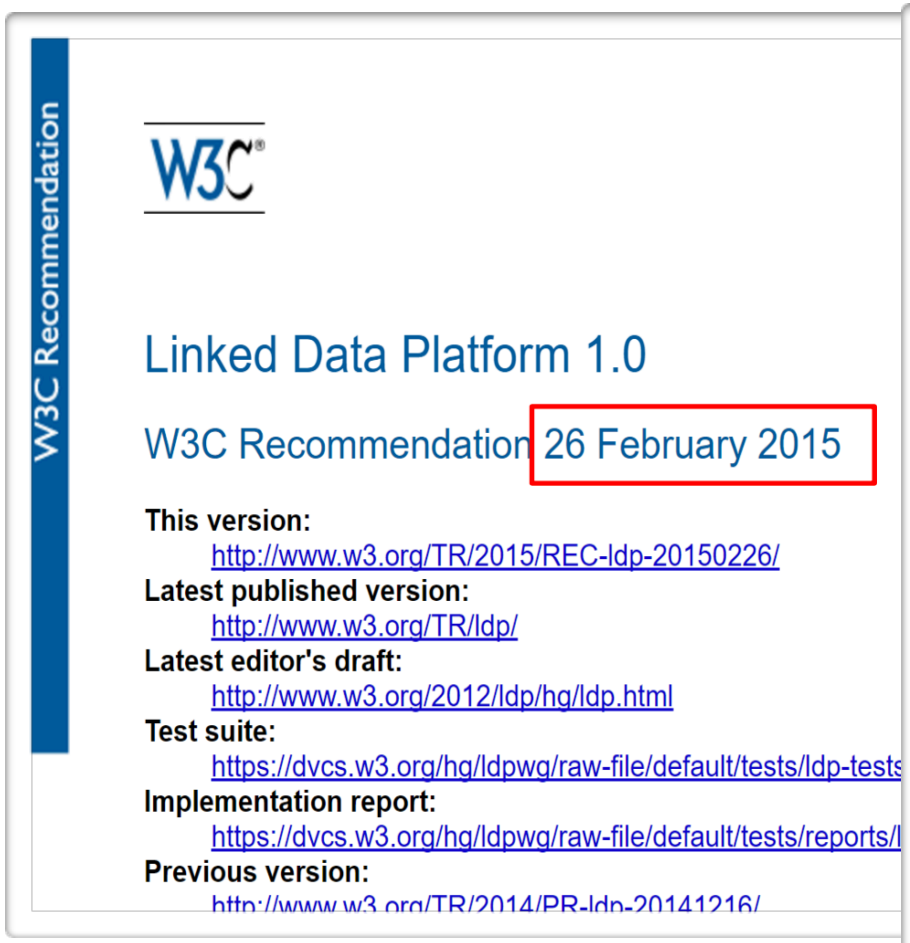
- FAIR-dICT project,DTL: https://www.dtls.nl/fair-data/fair-dict/
- Interoperability and FAIRness through a novel combination of Web technologies, https://peerj.com/articles/cs-110/

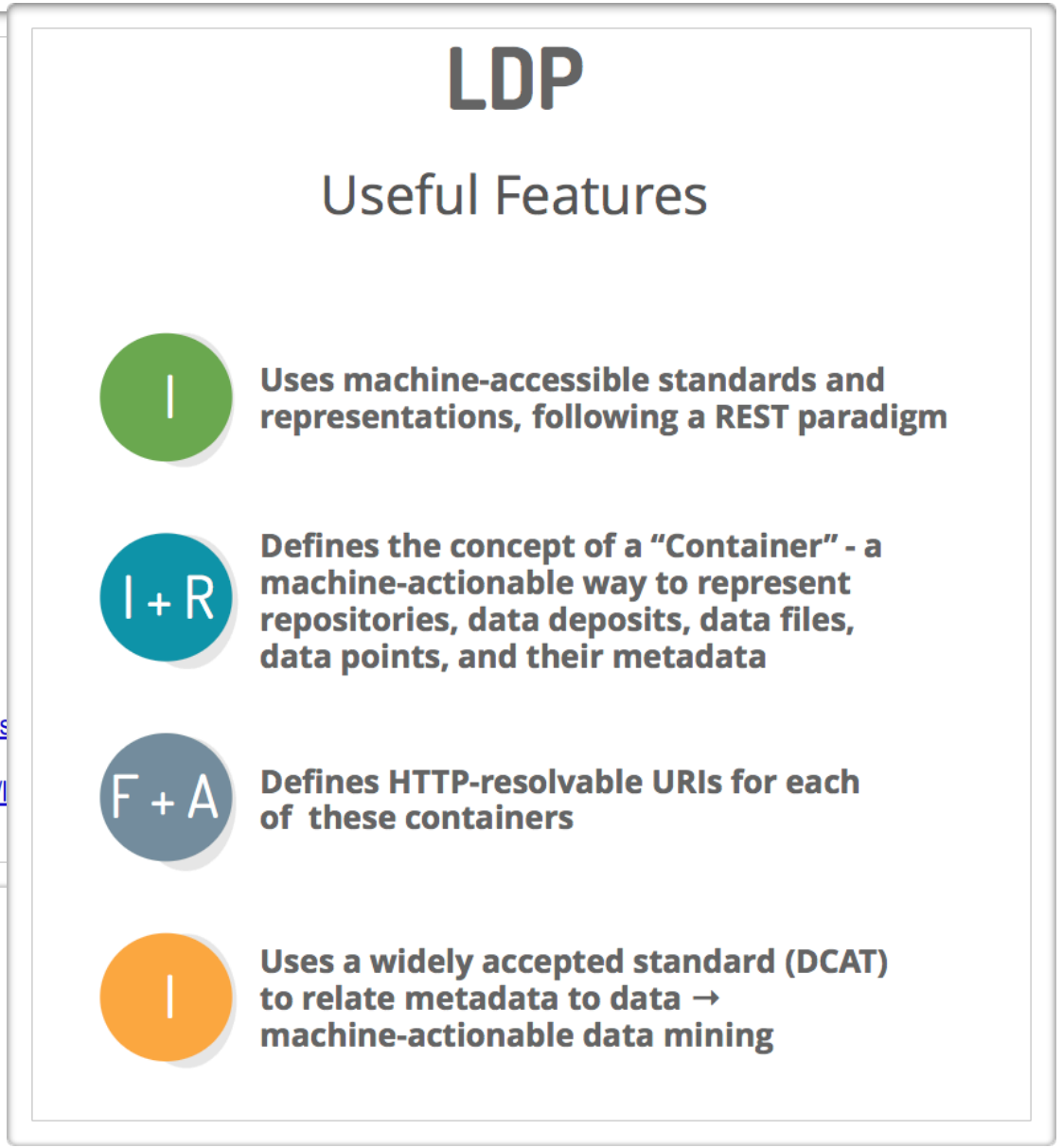


FAIR Data Point

```
GET <URL>
http://www.w3.org/TR/vocab-dcat/
```







Mark D. Wilkinson
CBGP-UPM/INIA, Madrid

W3C Recommendation

W3C

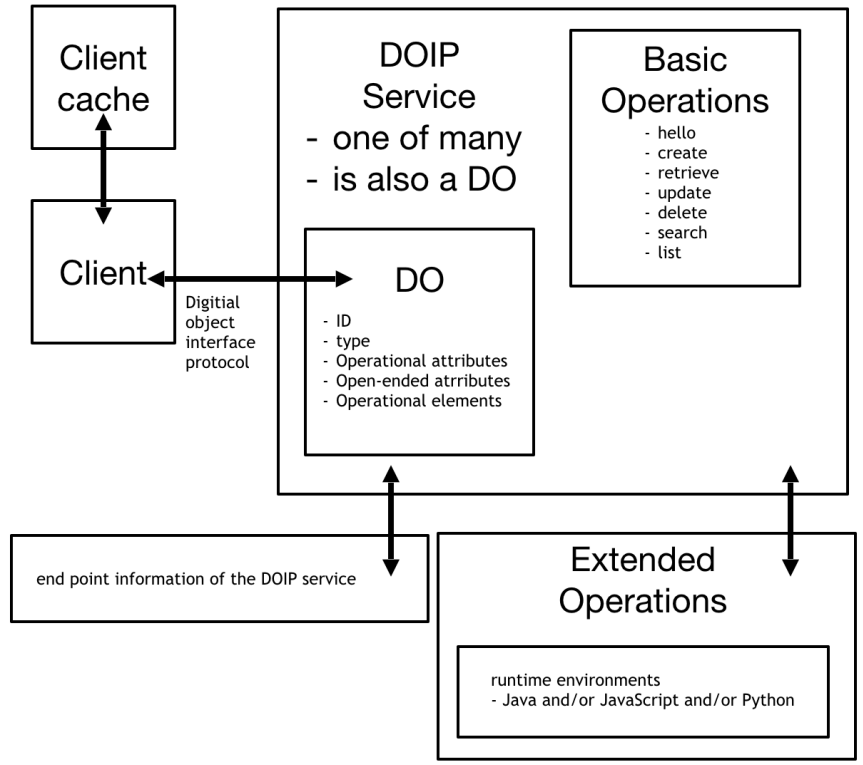
Linked Data Platform 1.0

W3C Recommendation 26 February 2015

This version:

http://www.w3.org/TR/2015/REC-ldp-20150226/

Digital Object Interface Protocol, Specification Version 2
Last Updated: June 26, 2018



LDP

Useful Features

- I — **Uses machine-accessible standards and representations, following a REST paradigm**
- I + R — **Defines the concept of a "Container" - a machine-actionable way to represent repositories, data deposits, data files, data points, and their metadata**
- + A — **Defines HTTP-resolvable URIs for each of these containers**
- I — **Uses a widely accepted standard (DCAT) to relate metadata to data → machine-actionable data mining**

Mark D. Wilkinson
CBGP-UPM/INIA, Madrid

GO FAIR

Find me all known low molecular weight inhibitors of the Human p65 Protein. Separate the list based on those that were found in curated databases, from those that were found in self-deposited data archives. Keep track of the license and citation information for each one. If data is relevant, but not public, please provide the contact information for the person I need so I can request the data.

Mark D. Wilkinson
CBGP-UPM/INIA, Madrid

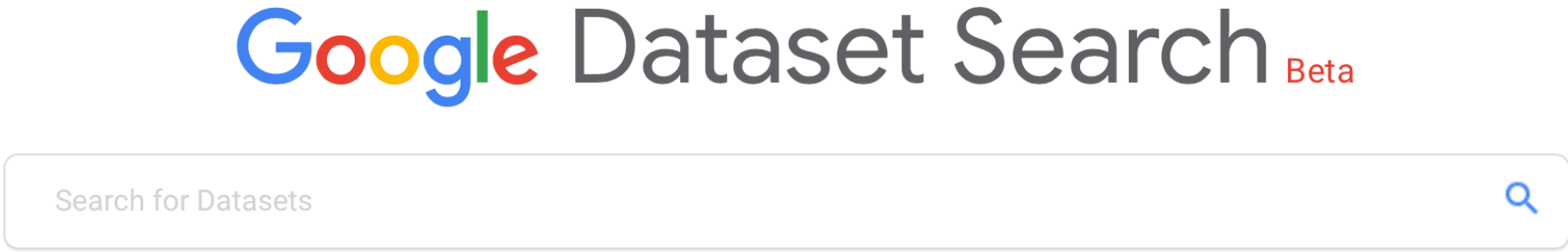
GO FAIR

FAIR metadata

http://136.243.4.200:8087/fdp

Title	FDP of biosemantics group	
Metadata ID	fdp	
Description	This is a prototype FDP for hosting research and student projects datasets	
Issued	2017-05-23T09:43:15.57Z	
Modified	2018-08-20T13:09:55	
Version	1.0E0	
License	http://rdflicense.appspot.com/rdflicense/cc-by-nc-nd3.0	
Access Rights	This resource has no access restriction	
Specification	http://rdf.biosemantics.org/fdp/shex/fdpMetadata	
Language	http://id.loc.gov/vocabulary/iso639-1/en	
Publisher	Biosemantic group	
Metrics	Type	https://purl.org/fair-metrics/FM_F1A
	Value	https://www.ietf.org/rfc/rfc3986.txt
	Type	https://purl.org/fair-metrics/FM_A1.1
	Value	https://www.wikidata.org/wiki/Q8777
Catalogs	http://136.243.4.200:8087/fdp/catalog/Transcriptomics	
	http://136.243.4.200:8087/fdp/catalog/multiomics	
	http://136.243.4.200:8087/fdp/catalog/textmining	
	http://136.243.4.200:8087/fdp/catalog/Biosamples	
	http://136.243.4.200:8087/fdp/catalog/Patient_Registries_1.0_998ccbcf-8714-426a-a28e-9335a86adb19	
	http://136.243.4.200:8087/fdp/catalog/SCA3_HD_multi-omics_blood_data	
Institution Country	http://lexvo.org/id/iso3166/NL	
Download RDF	ttl rdf+xml jsonld	

September 5, 2018



Google Dataset Search

Biosemantics Group LUMC

About

Gene disease association (LUMC)

semlab1.liacs.nl

Dataset provided by

Biosemantics group leiden

License

Attribution-NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

Available download formats from providers

HTML

Description

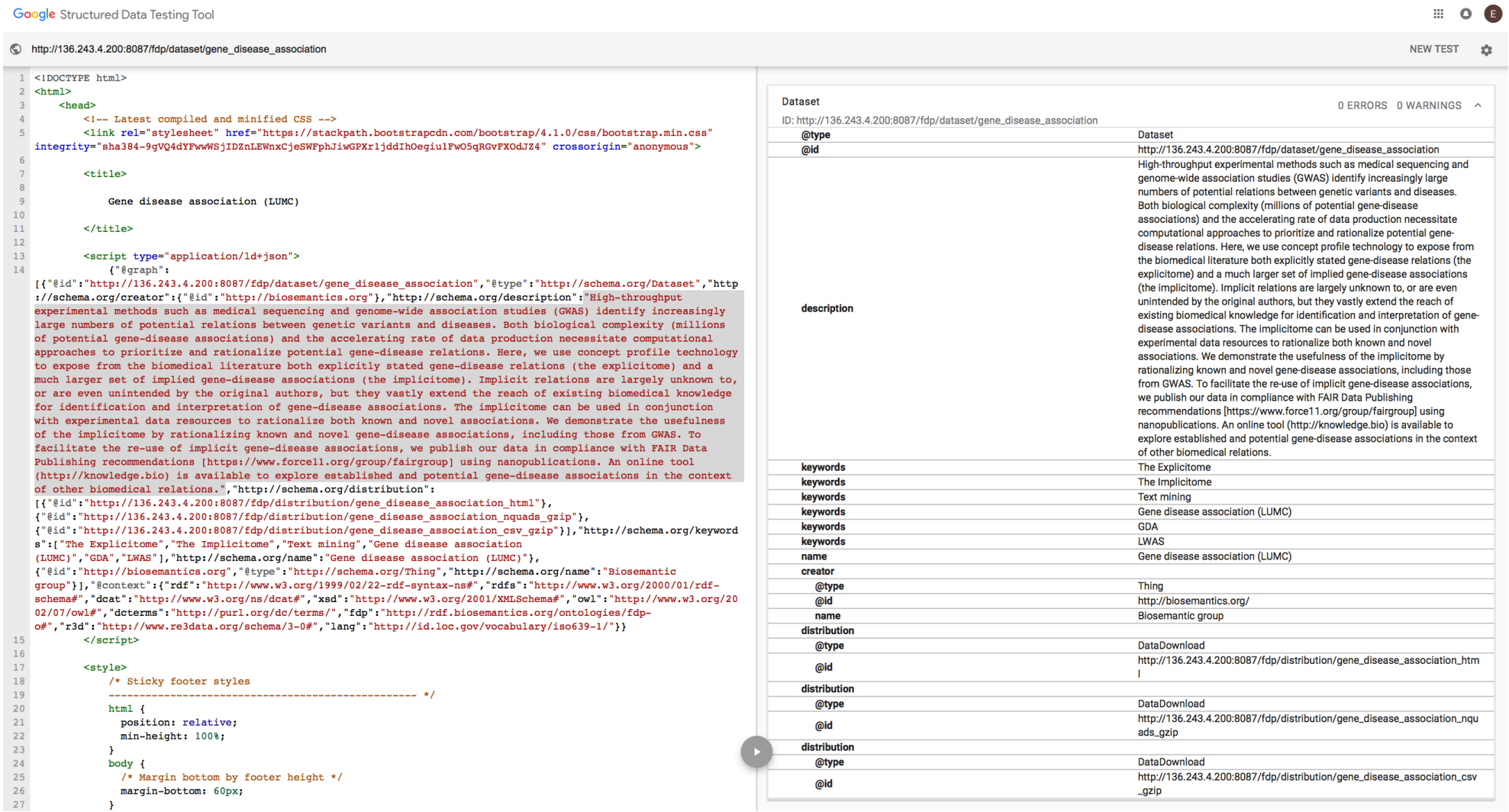
Gene disease association dataset from LUMC

L

Gene disease association (LUMC)

semlab1.liacs.nl

GO FAIR


Google Structured Data Testing Tool
http://136.243.4.200:8087/fdp/dataset/gene_disease_association
NEW TEST
<!DOCTYPE html>
<html>
<head>
<!-- Latest compiled and minified CSS -->
<link rel="stylesheet" href="https://stackpath.bootstrapcdn.com/bootstrap/4.1.0/css/bootstrap.min.css"
integrity="sha384-9gVQ4dYFwwWSjIDZnLEWnxCjeSWFphJiwGPXr1jddIhOegiu1FwO5qRGvFXOdJZ4" crossorigin="anonymous">
<title>
Gene disease association (LUMC)
</title>
<script type="application/ld+json">
{"@graph":
[{"@id":"http://136.243.4.200:8087/fdp/dataset/gene_disease_association","@type":"http://schema.org/Dataset","http://schema.org/creator":{"@id":"http://biosemantics.org"},"http://schema.org/description":"High-throughput experimental methods such as medical sequencing and genome-wide association studies (GWAS) identify increasingly large numbers of potential relations between genetic variants and diseases. Both biological complexity (millions of potential gene-disease associations) and the accelerating rate of data production necessitate computational approaches to prioritize and rationalize potential gene-disease relations. Here, we use concept profile technology to expose from the biomedical literature both explicitly stated gene-disease relations (the explicitome) and a much larger set of implied gene-disease associations (the implicitome). Implicit relations are largely unknown to, or are even unintended by the original authors, but they vastly extend the reach of existing biomedical knowledge for identification and interpretation of gene-disease associations. The implicitome can be used in conjunction with experimental data resources to rationalize both known and novel associations. We demonstrate the usefulness of the implicitome by rationalizing known and novel gene-disease associations, including those from GWAS. To facilitate the re-use of implicit gene-disease associations, we publish our data in compliance with FAIR Data Publishing recommendations [https://www.force11.org/group/fairgroup] using nanopublications. An online tool (http://knowledge.bio) is available to explore established and potential gene-disease associations in the context of other biomedical relations.","http://schema.org/distribution":
[{"@id":"http://136.243.4.200:8087/fdp/distribution/gene_disease_association_html"},
{"@id":"http://136.243.4.200:8087/fdp/distribution/gene_disease_association_nquads_gzip"},
{"@id":"http://136.243.4.200:8087/fdp/distribution/gene_disease_association_csv_gzip"}],"http://schema.org/keywords":["The Explicitome","The Implicitome","Text mining","Gene disease association (LUMC)","GDA","LWAS"],"http://schema.org/name":"Gene disease association (LUMC)"},
{"@id":"http://biosemantics.org","@type":"http://schema.org/Thing","http://schema.org/name":"Biosemantic group"}],"@context":{"rdf":"http://www.w3.org/1999/02/22-rdf-syntax-ns#","rdfs":"http://www.w3.org/2000/01/rdf-schema#","dcat":"http://www.w3.org/ns/dcat#","xsd":"http://www.w3.org/2001/XMLSchema#","owl":"http://www.w3.org/2002/07/owl#","dcterms":"http://purl.org/dc/terms/","fdp":"http://rdf.biosemantics.org/ontologies/fdp-o#","r3d":"http://www.re3data.org/schema/3-0#","lang":"http://id.loc.gov/vocabulary/iso639-1/"}}
</script>
<style>
/* Sticky footer styles
-- */
html {
position: relative;
min-height: 100%;
}
body {
/* Margin bottom by footer height */
margin-bottom: 60px;
}
Dataset
0 ERRORS 0 WARNINGS
ID: http://136.243.4.200:8087/fdp/dataset/gene_disease_association
@type Dataset
@id http://136.243.4.200:8087/fdp/dataset/gene_disease_association
description High-throughput experimental methods such as medical sequencing and genome-wide association studies (GWAS) identify increasingly large numbers of potential relations between genetic variants and diseases. Both biological complexity (millions of potential gene-disease associations) and the accelerating rate of data production necessitate computational approaches to prioritize and rationalize potential gene-disease relations. Here, we use concept profile technology to expose from the biomedical literature both explicitly stated gene-disease relations (the explicitome) and a much larger set of implied gene-disease associations (the implicitome). Implicit relations are largely unknown to, or are even unintended by the original authors, but they vastly extend the reach of existing biomedical knowledge for identification and interpretation of gene-disease associations. The implicitome can be used in conjunction with experimental data resources to rationalize both known and novel associations. We demonstrate the usefulness of the implicitome by rationalizing known and novel gene-disease associations, including those from GWAS. To facilitate the re-use of implicit gene-disease associations, we publish our data in compliance with FAIR Data Publishing recommendations [https://www.force11.org/group/fairgroup] using nanopublications. An online tool (http://knowledge.bio) is available to explore established and potential gene-disease associations in the context of other biomedical relations.
keywords The Explicitome
keywords The Implicitome
keywords Text mining
keywords Gene disease association (LUMC)
keywords GDA
keywords LWAS
name Gene disease association (LUMC)
creator
@type Thing
@id http://biosemantics.org/
name Biosemantic group
distribution
@type DataDownload
@id http://136.243.4.200:8087/fdp/distribution/gene_disease_association_html
distribution
@type DataDownload
@id http://136.243.4.200:8087/fdp/distribution/gene_disease_association_nquads_gzip
distribution
@type DataDownload
@id http://136.243.4.200:8087/fdp/distribution/gene_disease_association_csv_gzip

Measuring FAIRness

http://fairmetrics.org

FAIR Metrics

The FAIR Metrics Group took-on the challenge of designing a framework for evaluating "FAIRness".

Discoverability and reusability are not abstract concepts, but imply concrete behaviors and properties that must hold true for the fulfillment of the FAIR objectives. Given this, it must therefore be possible to precisely define a measurable set of properties and behaviors that assess FAIRness. Over the short 1 month lifespan of the FAIR Metrics Working Group, we have created a cogent framework for developing FAIR metrics manifested as a simple form with 8 questions that structures fruitful conversations about proposed metrics.

Our approach recognizes that the diversity in opinion must play a key role in crafting fair and effective FAIR guidelines. Communities must not only understand what is meant by FAIR, but must also be able to monitor the FAIRness of their digital resources, in a realistic, but quantitative manner. We recognize that what is considered FAIR in one community may be quite different from FAIRness in another community - different community norms and practices make this a certainty! As such, our approach focuses on the mechanism by which metrics can be created by community members themselves, rather than attempting to create a set of one-size-fits-all metrics to apply to every resource.

With a mechanism in-place to design metrics, we now open the process of generating metrics to community participation. We have created several exemplar metrics that we think will be broadly applicable; however, additional metrics may be designed and published through ***our open submission process***, or simply shared within your community through your normal communication channels.

Our proposed FAIR Metrics can be found here.

We have selected an approach to publishing FAIR Metrics that is, itself, FAIR. This takes the form of a FAIR Accessor (a kind of Linked Data Platform Container), which describes a subset of metrics, the community to which they are applicable, other relevant metadata, and links to each

FAIR Metrics GitHub

FAIR Metrics Paper

Metrics Process

Metrics Authoring Framework

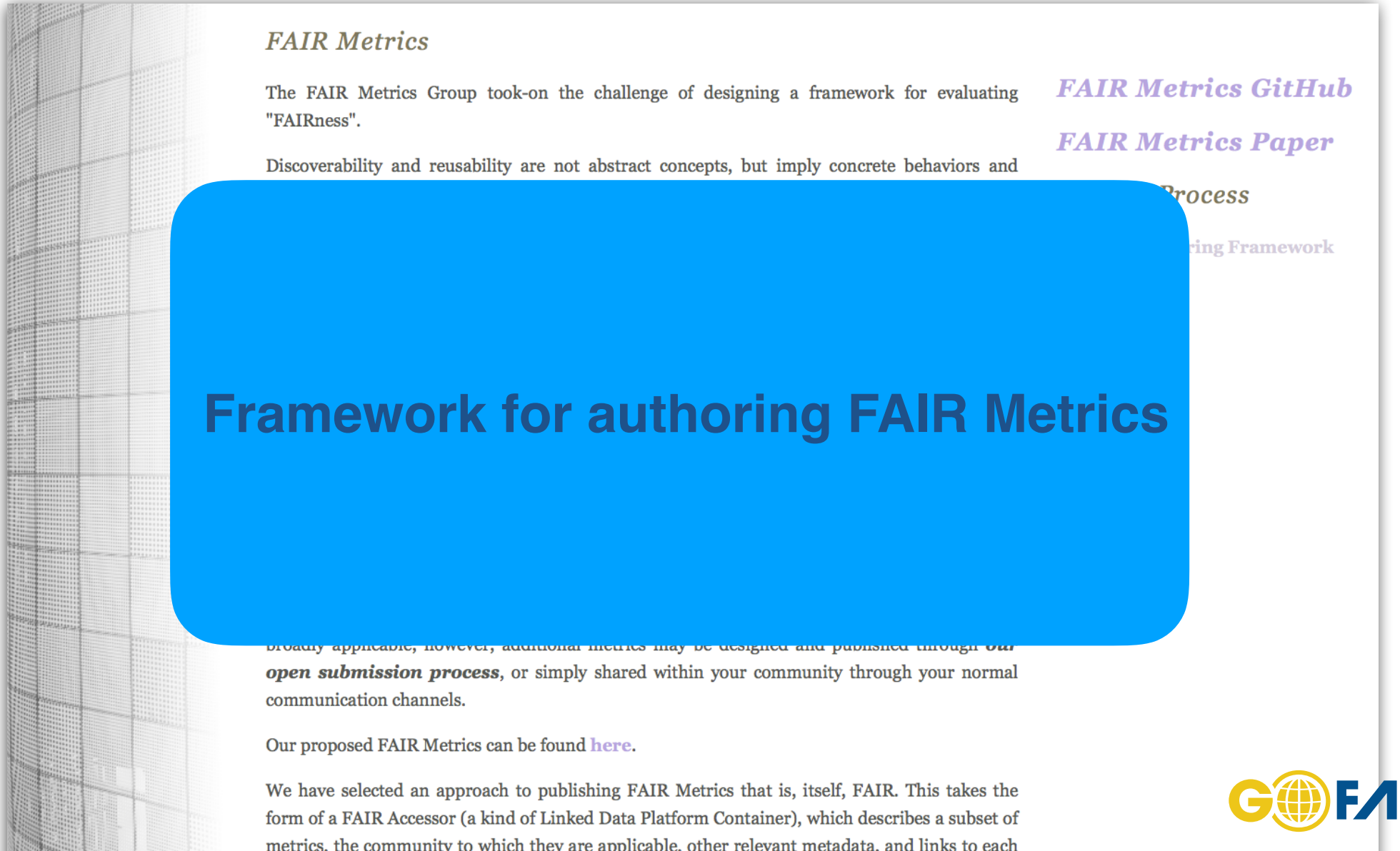
Metrics Form

About Us

GO FAIR

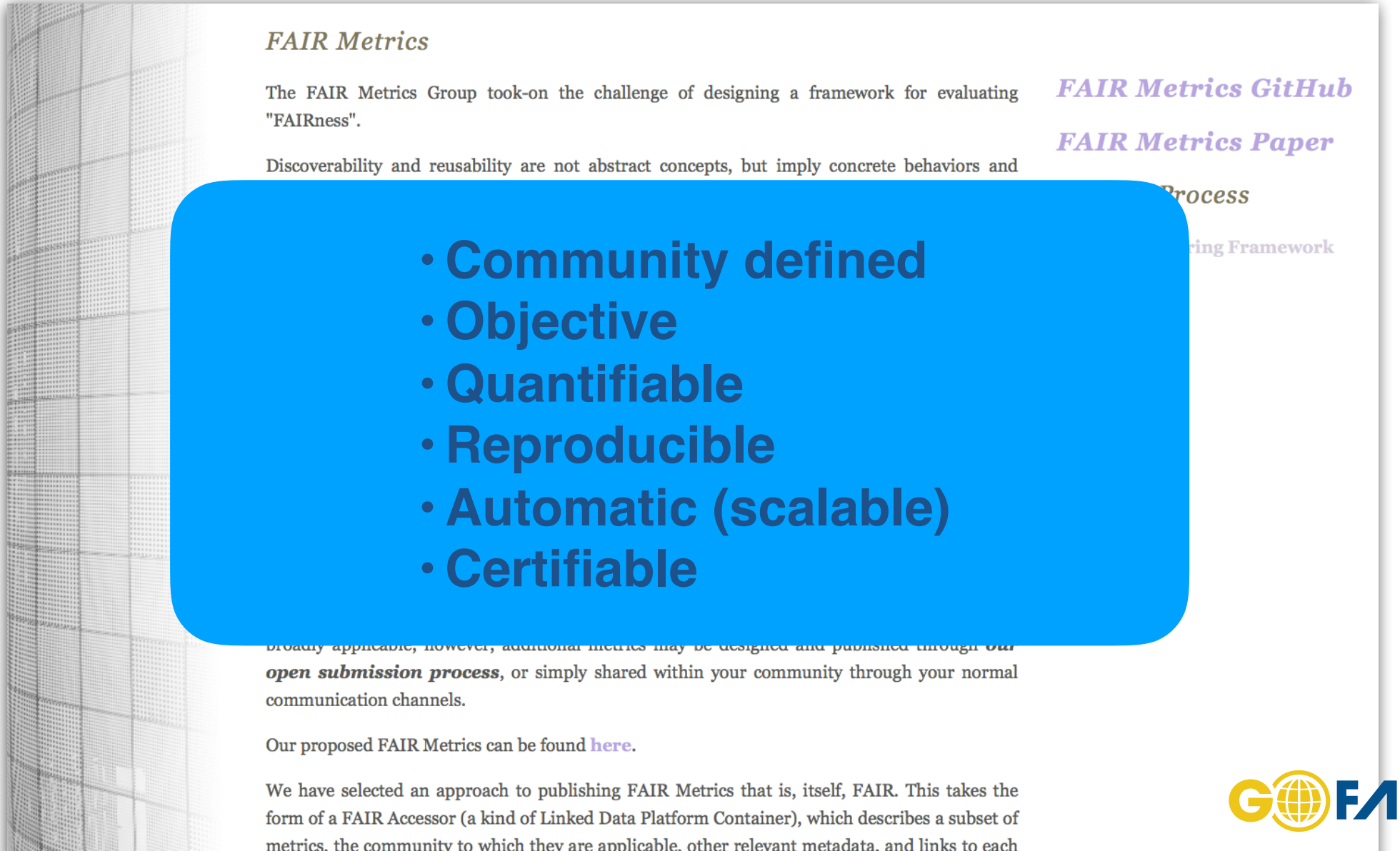
Measuring FAIRness

http://fairmetrics.org



Measuring FAIRness

http://fairmetrics.org



Measuring FAIRness

http://fairmetrics.org

FAIR Metrics

The FAIR Metrics Group took-on the challenge of designing a framework for evaluating

FAIR Metrics GitHub

www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

Comment: A design framework and exemplar metrics for FAIRness

Mark D. Wilkinson[1], Susanna-Assunta Sansone[2], Erik Schultes[3], Peter Doorn[4], Luiz Olavo Bonino da Silva Santos[5,6] & Michel Dumontier[7]

Received: 28 November 2017
Accepted: 9 May 2018
Published: 26 June 2018

The FAIR Principles[1] (https://doi.org/10.25504/FAIRsharing.WWI10U) provide guidelines for the publication of digital resources such as datasets, code, workflows, and research objects, in a manner that makes them Findable, Accessible, Interoperable, and Reusable (FAIR). The Principles have rapidly been adopted by publishers, funders, and pan-disciplinary infrastructure programmes and societies. The Principles are aspirational, in that they do not strictly define how to achieve a state of "FAIRness", but rather they describe a continuum of features, attributes, and behaviors that will move a digital resource closer to that goal. This ambiguity has led to a wide range of interpretations of FAIRness, with some resources even claiming to already "be FAIR"! The increasing number of such statements, the emergence

Measuring FAIRness

http://fairmetrics.org

FM	Question	Dataverse	Dryad	Nano-pub	Zenodo	Yale ISPS	Figshare	Broad's SCP	SeaData Net's CDI	Wikidata
IRI Exists	1	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI
F1A	2	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI
F1B	3	IRI	IRI	IRI	NRP	none	IRI	IRI	IRI	IRI
F2A	4A	IRI	IRI	IRI	IRI	none	none	IRI	IRI	IRI
F2A	4B	IRI	none	IRI	IRI	"Multiple"	none	IRI	IRI	IRI
F3	5A	IRI	IRI	IRI	IRI	none	NRP	IRI	IRI	IRI
F3	5B	IRI	IRI	IRI	IRI	IRI	IRI	IRI	none	IRI
F4	6A	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI
F4	6B	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI
A1.1	7A	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI
A1.1	7B	true	true	true	true	true	true	true	true	true
A1.1	7C	true	true	true	true	true	true	true	true	true
A1.2	8A	false	false	false	false	false	false	false	true	false
A1.2	8B	N/A	N/A	N/A	N/A	NRP	NRP	NRP	link	N/A
A2	9	IRI	IRI	none	IRI	none	IRI	none	IRI	NRP
I1	10	IRI	IRI	IRI	IRI	none	none	NRP	IRI	IRI
I2	11	IRI	IRI	IRI	none	none	none	IRI	IRI	IRI
I3	12	NRP	IRI	IRI	none	none	none	NRP	NRP	IRI
R1.1	13	IRI	IRI	IRI	IRI	IRI	IRI	NRP	IRI	IRI
R1.2	14A	IRI	IRI	IRI	IRI	none	none		NRP	NRP
R1.2	14B		none		none	none	none			
R1.3	15	NRP			none	none	none	NRP		

Measuring FAIRness

http://fairmetrics.org

Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

HOME | ABOUT | SUBMIT | ALERTS / RSS | CHANNELS

Search

Advanced Search

New Results

Previous Next

Evaluating FAIR-Compliance Through an Objective, Automated, Community-Governed Framework

Posted September 16, 2018.

Download PDF

Share

Citation Tools

Email

Tweet Like 0

G+

Mark D Wilkinson, Michel Dumontier, Susanna-Assunta Sansone, Luiz Olavo Bonino da Silva Santos, Mario Prieto, Julian Gautier, Peter McQuilton, Derek Murphy, Merce Crosas, Erik Schultes

doi: https://doi.org/10.1101/418376

This article is a preprint and has not been peer-reviewed [what does this mean?].

Abstract | Info/History | Metrics | Preview PDF

Subject Area

Scientific Communication and Education

14 Core FAIR Metrics

Findable:

FM-F1A FM-F1B

F1 (meta)data are assigned a globally **unique** and **persistent** identifier;

FM-F2

F2 data are described with **rich metadata**;

FM-F3

F3 metadata clearly and explicitly include the **identifier of the data** it describes;

FM-F4

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

Accessible:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol;

A1.1 the protocol is **open, free, and universally implementable**; FM-A1.1

FM-A1.2

A1.2 the protocol allows for an **authentication and authorization** procedure, where necessary;

A2 metadata are accessible, **even when the data are no longer available**; FM-A2

Interoperable:

I1 (meta)data use a formal, accessible, shared, and broadly applicable **language for knowledge representation**. FM-I1

FM-I2

I2 (meta)data use **vocabularies that follow FAIR principles**;

I3 (meta)data include **qualified references** to other (meta)data; FM-I3

Reusable:

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

R1.1 (meta)data are released with a clear and **accessible data usage license**; FM-R1.1

R1.2 (meta)data are associated with **detailed provenance**; FM-R1.2

R1.3 (meta)data meet domain-relevant **community standards**; FM-R1.3

Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016)

http://fairmetrics.org

https://github.com/FAIRMetrics/Metrics/blob/master/ALL.pdf

GO FAIR

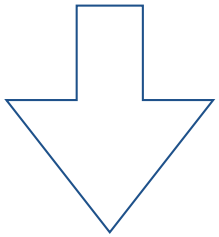
The FAIR Metrics Template

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

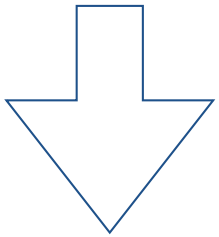
The FAIR Metrics Upgrades

Example: FM-F1B, Identifier Persistence

v1.0 **checks** for HTTP 200 return



v2.0 **validates** a standard RDF persistence policy



v3.0 **scores** multiple parameters of persistence policy

14 Core FAIR Metrics
21 Questions
22 Community Challenges

FAIR Principle F1: (meta) data are assigned globally unique and persistent identifiers.
Fundamental requirement for accurate and reproducible machine actionability. Examples: Universally unique identifier (UUID): https://en.wikipedia.org/wiki/Universally_unique_identifier ; Digital Object Identifier (DOI): http://www.doi.org

FAIR Metric F1A:
Question 1: *Provide an URL to a registered scheme that defines the globally-unique structure of the identifier(s) for your digital resource.*

FAIR Metric F1B:
Question 2: *Provide an URL to a document that defines the persistence policy of your identifier(s).*

Community challenge:
(1) What are your required (or preferred) identifier registration services ?
(2) What is your minimal persistence policy?
(3) Can you make your persistence policy machine-readable?

GO FAIR

FAIR Metrics Community Challenges

Find

(1) What are your required (or preferred) identifier registration services ?
(2) What is your minimal persistence policy?
(3) Can you make your persistence policy machine-readable?
(4) Can you define a minimal set of metadata for your community?
(5) Can you make your metadata machine-readable?
(6) Can you define the metadata model that explicitly links data and metadata?
(7) Can you make this metadata model machine-readable?
(8) What is the required (preferred) search engine for your community ?

Access

(9) What is the required (preferred) communication protocol for your community ?
(10) What is your required (preferred) protocols for restricting access to data ?
(11) Can you make this protocol machine-readable?
(12) What is your minimal persistence policy for metadata?
(13) Can you make this persistence policy machine-readable?

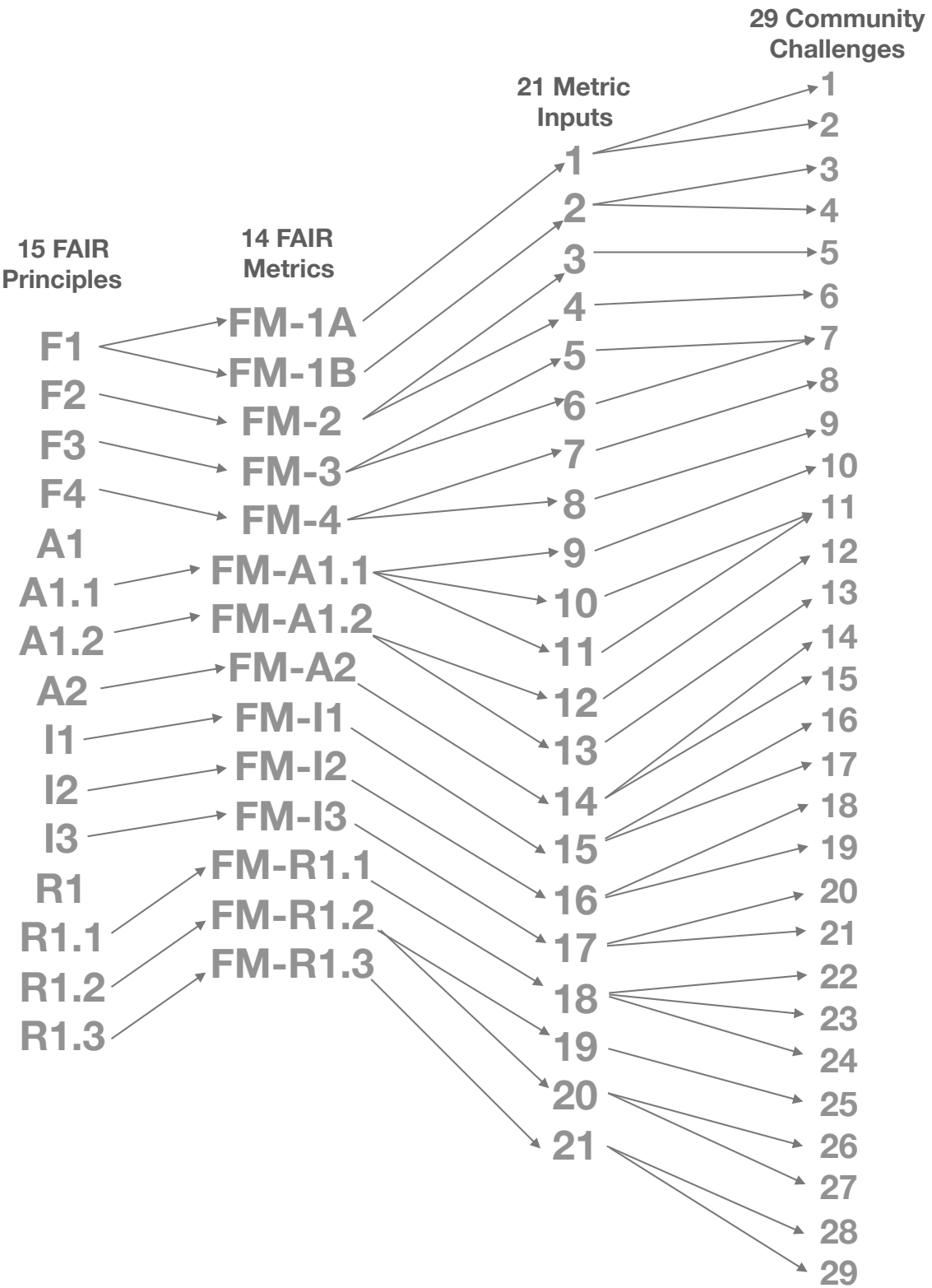
Interoperate

(14) What is your required (preferred) standards in knowledge representation ?
(15) What are your required (preferred) vocabularies ?
(16) What is your required LinkSet ?

Reuse

(17) What is your required (preferred) usage license framework?
(18) Can you make these usage licenses machine-readable?
(19) What is your required (preferred) provenance metadata descriptions?
(20) Can you make this provenance metadata machine-readable?
(21) What are your certification criteria for data & metadata?
(22) What is your machine-actionable validation-certification system ?

GO FAIR



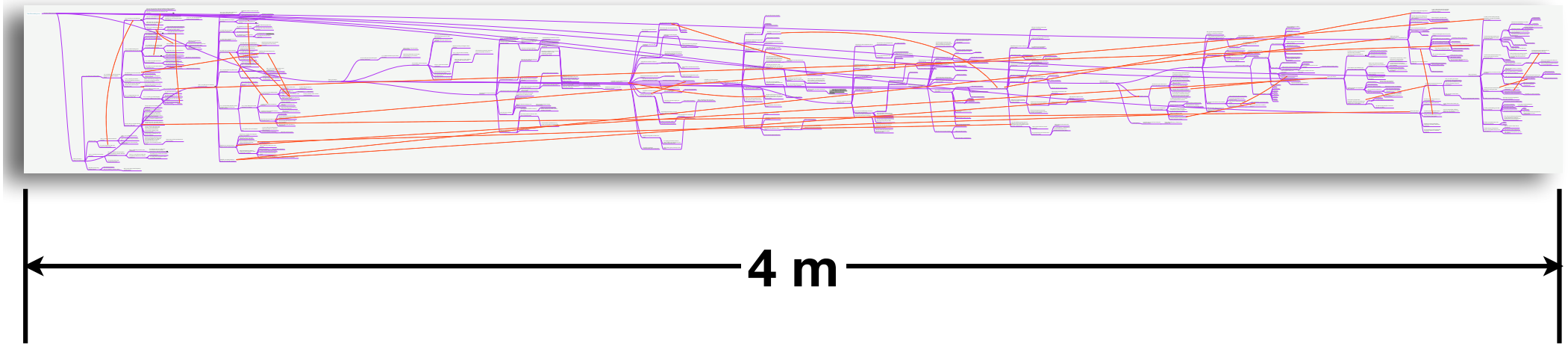
The 29 Community Challenges are 29 ways to create convergence within and between communites

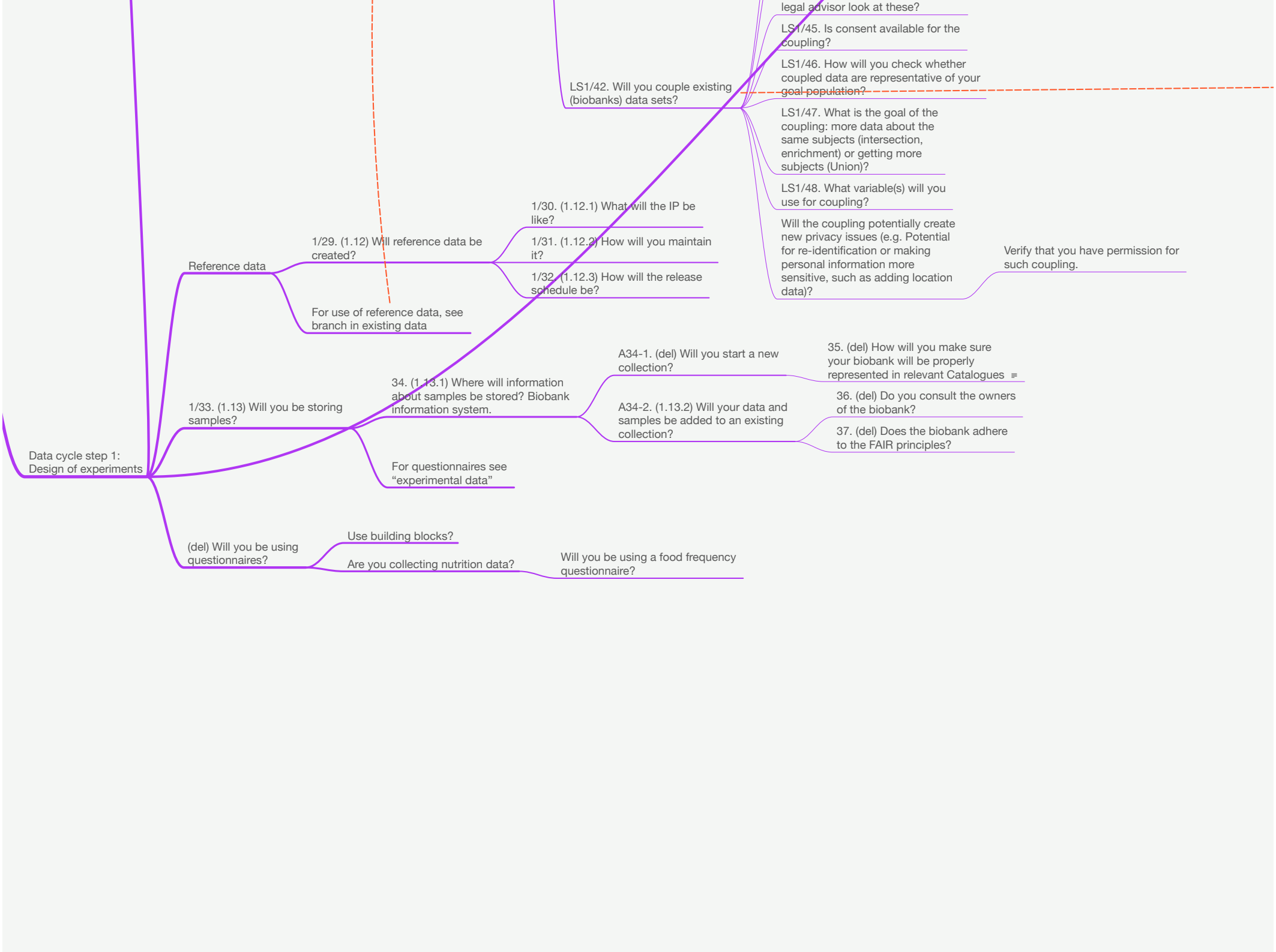
FAIR Data Stewardship

Rob Hooft

DTL

DUTCH TECHCENTRE FOR LIFE SCIENCES




Data cycle step 1: Design of experiments
Reference data
1/29. (1.12) Will reference data be created?
1/30. (1.12.1) What will the IP be like?
1/31. (1.12.2) How will you maintain it?
1/32. (1.12.3) How will the release schedule be?
For use of reference data, see branch in existing data
1/33. (1.13) Will you be storing samples?
34. (1.13.1) Where will information about samples be stored? Biobank information system.
A34-1. (del) Will you start a new collection?
35. (del) How will you make sure your biobank will be properly represented in relevant Catalogues
A34-2. (1.13.2) Will your data and samples be added to an existing collection?
36. (del) Do you consult the owners of the biobank?
37. (del) Does the biobank adhere to the FAIR principles?
For questionnaires see “experimental data”
(del) Will you be using questionnaires?
Use building blocks?
Are you collecting nutrition data?
Will you be using a food frequency questionnaire?
LS1/42. Will you couple existing (biobanks) data sets?
legal advisor look at these?
LS1/45. Is consent available for the coupling?
LS1/46. How will you check whether coupled data are representative of your goal population?
LS1/47. What is the goal of the coupling: more data about the same subjects (intersection, enrichment) or getting more subjects (Union)?
LS1/48. What variable(s) will you use for coupling?
Will the coupling potentially create new privacy issues (e.g. Potential for re-identification or making personal information more sensitive, such as adding location data)?
Verify that you have permission for such coupling.

ELIXIR

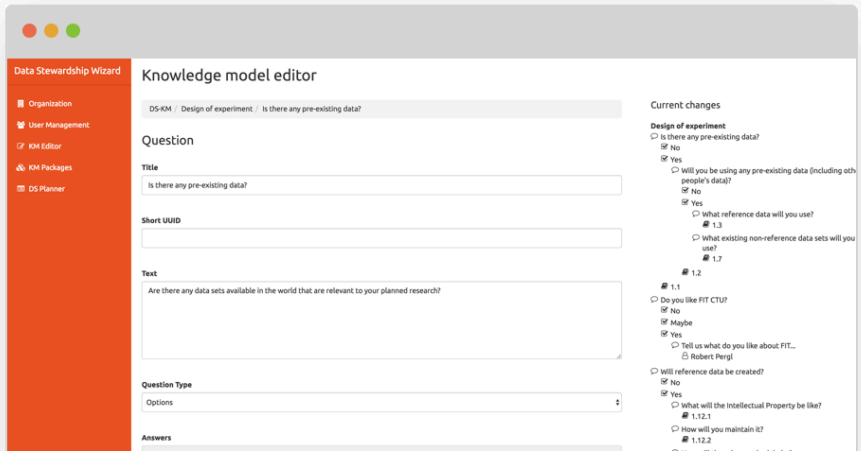
Data Stewardship Wizard

https://dsw.fairdata.solutions

Go to App

Smart Data Management Plans for FAIR Open Science

For serious researchers and data stewards



Easily create comprehensive data management plans

Our smart questionnaire will effortlessly guide you through the vast knowledge of data stewardship by asking you **questions**, offering **hints**, **multimedia contents**, **external resources** and **community help**.

This is data stewardship done seriously for the project success, not just to make your funder happy!

Current Phase

Before Submitting the Proposal

Design of experiment

Before you decide to embark on any new study, it is nowadays good practice to conside
data generation part of your study as limited as possible. It is not because we can gene
data that we always need to do so. Creating data with public money is bringing with it t
those data well and (if potentially useful) make them available for re-use by others.

Is there any pre-existing data?

Are there any data sets available in the world that are relevant to your planned research?

Desirable: *Before Submitting the DMP*
Data Stewardship for Open Science: *atq*

- ○ No
- ○ Yes

Will reference data be created?

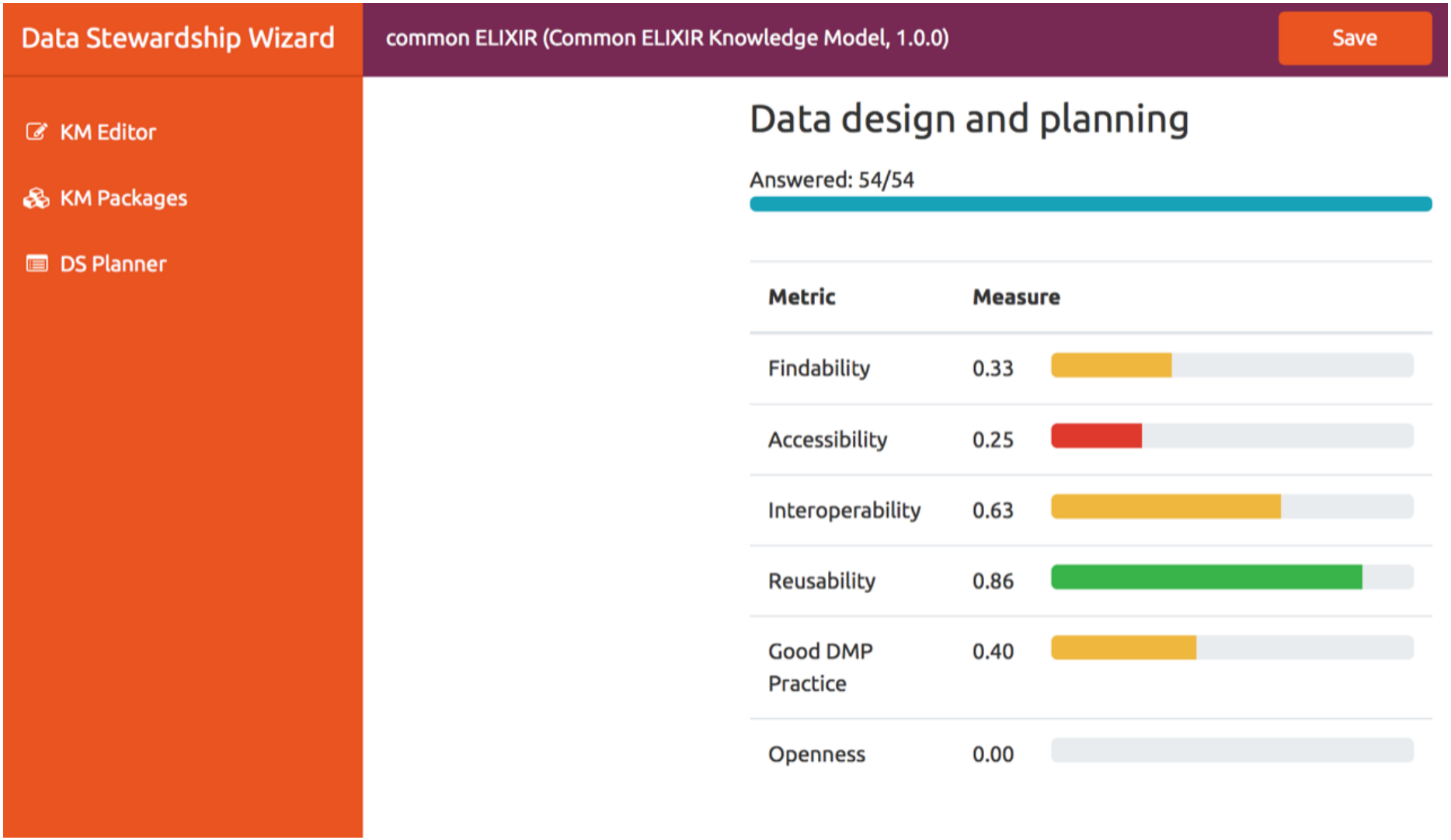
Will any of the data that you will be creating form a reference data set for future research

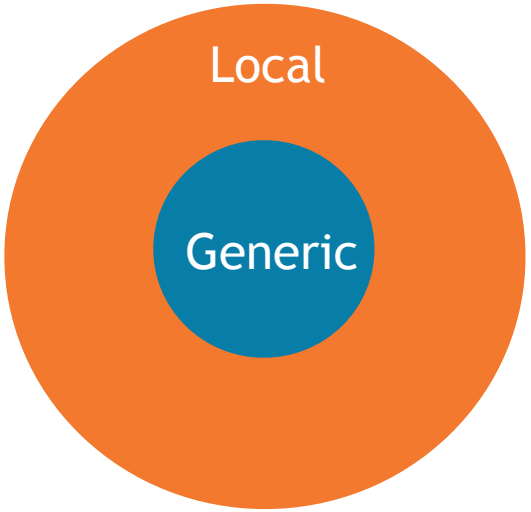
Desirable: *Before Submitting the DMP*
Data Stewardship for Open Science: *rbz*

- ○ No
- ○ Yes

Will you be storing samples?

Desirable: *Before Submitting the DMP*
Data Stewardship for Open Science: *kuz*

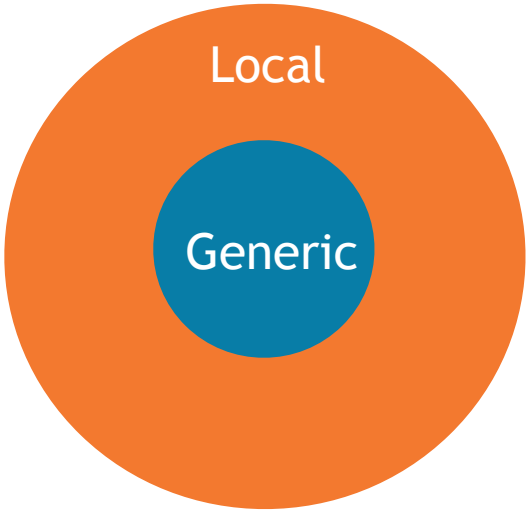

Data Stewardship Wizard
common ELIXIR (Common ELIXIR Knowledge Model, 1.0.0)
Save
KM Editor
KM Packages
DS Planner
Data design and planning
Answered: 54/54
Metric
Measure
Findability
0.33
Accessibility
0.25
Interoperability
0.63
Reusability
0.86
Good DMP Practice
0.40
Openness
0.00



ELIXIR Data Stewardship Knowledge Model

https://github.com/DataStewardshipWizard/ds-km

Data in the tool is highly configurable. It is separated in Generic and Local: So we can separate life science specific (like “human health data”), ELIXIR specific (like names of experts), and even for local institutes we can adapt it.

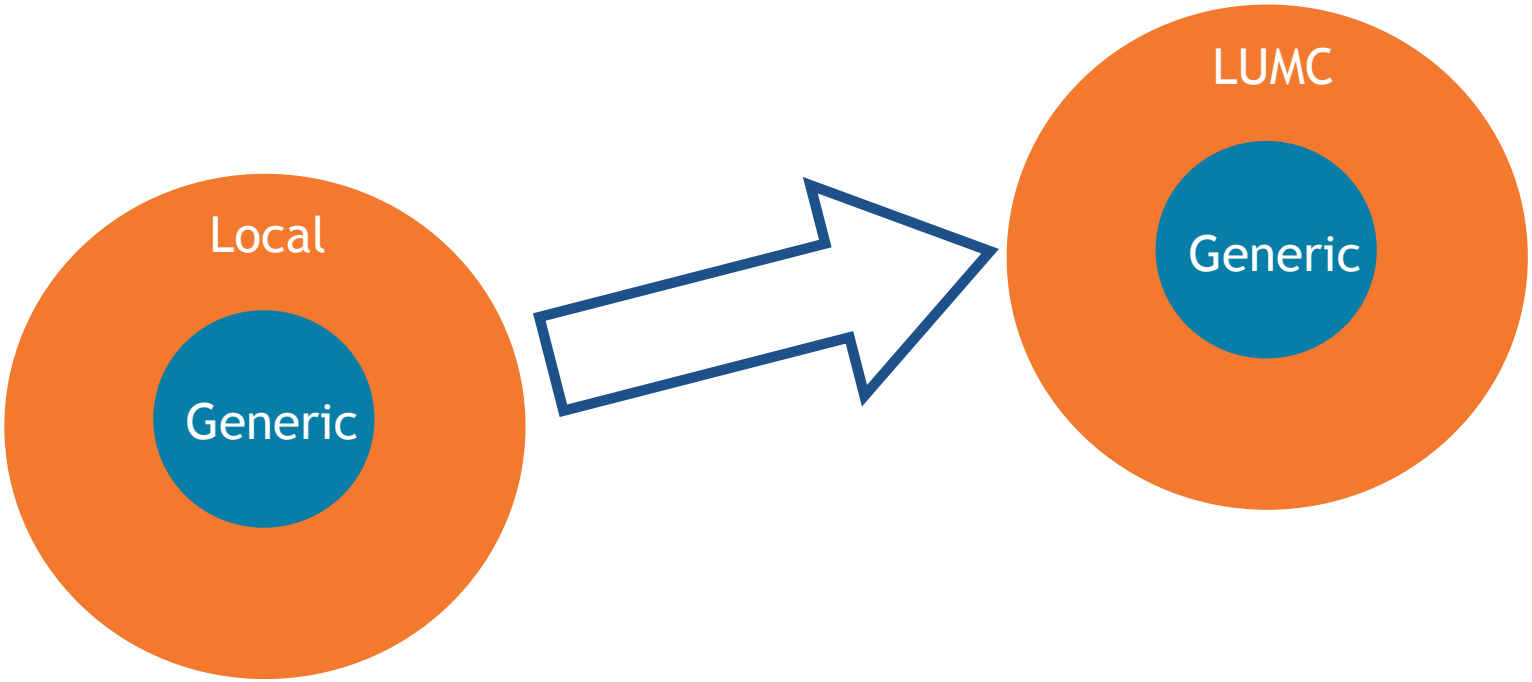


Can be localized by:
- Research Community
- Organization
- FAIR metrics
- Funding requirments

ELIXIR Data Stewardship Knowledge Model

https://github.com/DataStewardshipWizard/ds-km

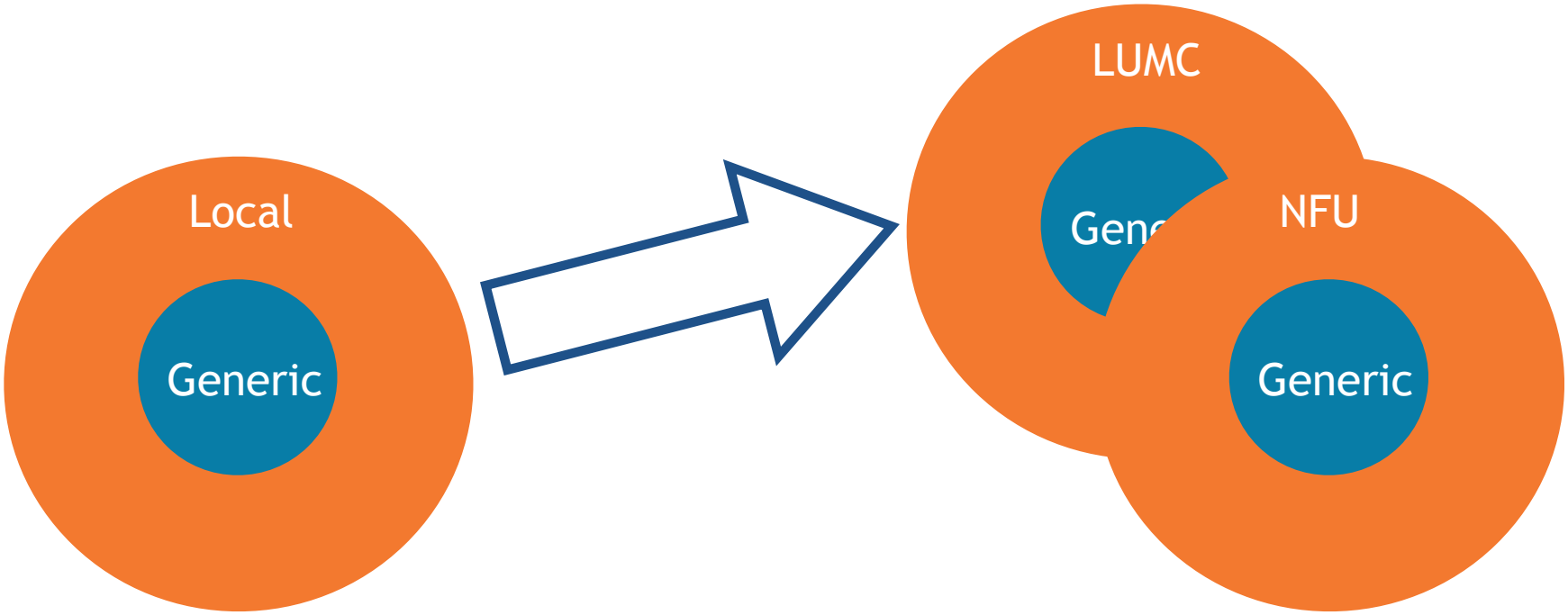
Data in the tool is highly configurable. It is separated in Generic and Local: So we can separate life science specific (like “human health data”), ELIXIR specific (like names of experts), and even for local institutes we can adapt it.



ELIXIR Data Stewardship Knowledge Model

https://github.com/DataStewardshipWizard/ds-km

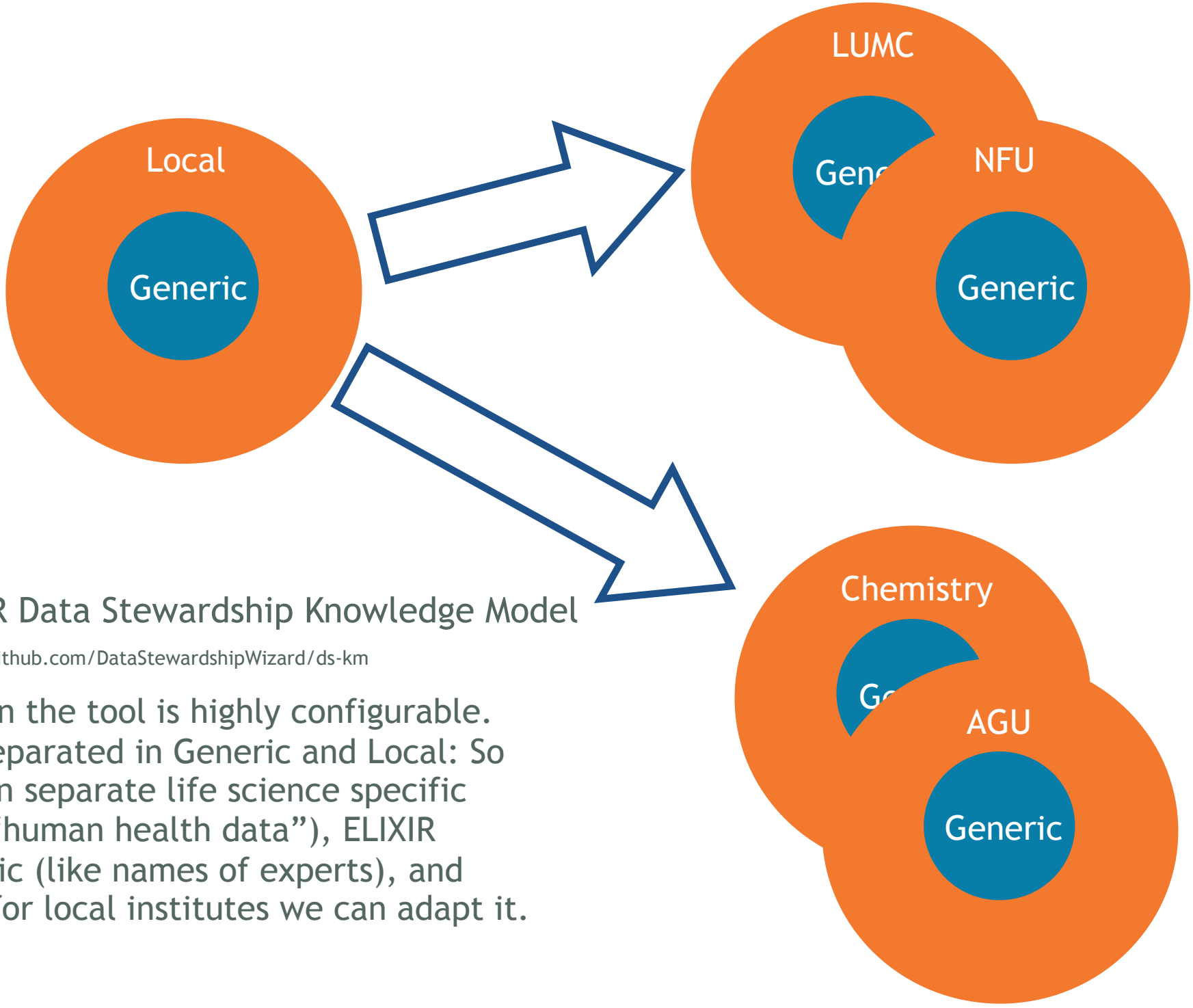
Data in the tool is highly configurable. It is separated in Generic and Local: So we can separate life science specific (like “human health data”), ELIXIR specific (like names of experts), and even for local institutes we can adapt it.



ELIXIR Data Stewardship Knowledge Model

https://github.com/DataStewardshipWizard/ds-km

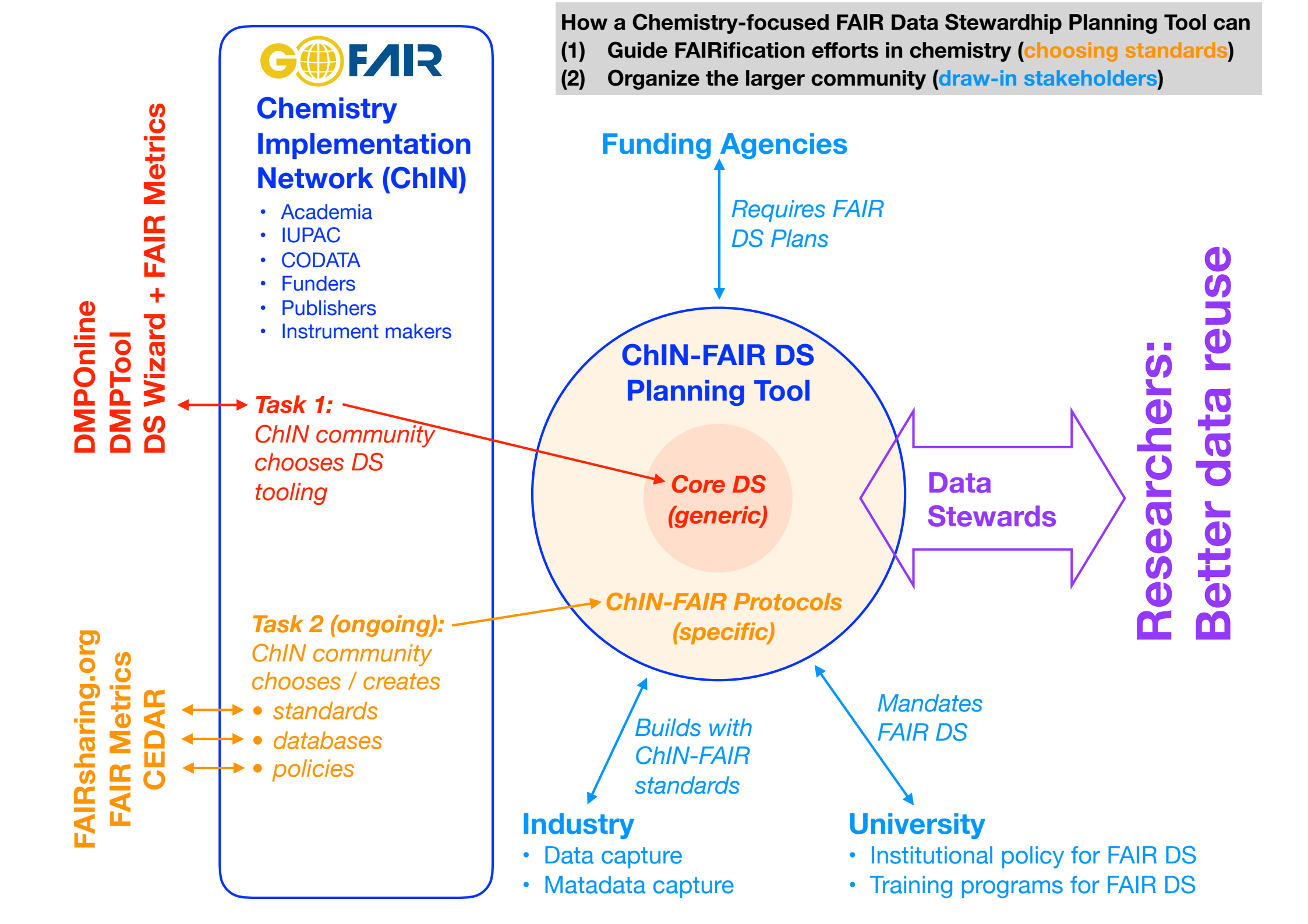
Data in the tool is highly configurable. It is separated in Generic and Local: So we can separate life science specific (like “human health data”), ELIXIR specific (like names of experts), and even for local institutes we can adapt it.



ELIXIR Data Stewardship Knowledge Model

https://github.com/DataStewardshipWizard/ds-km

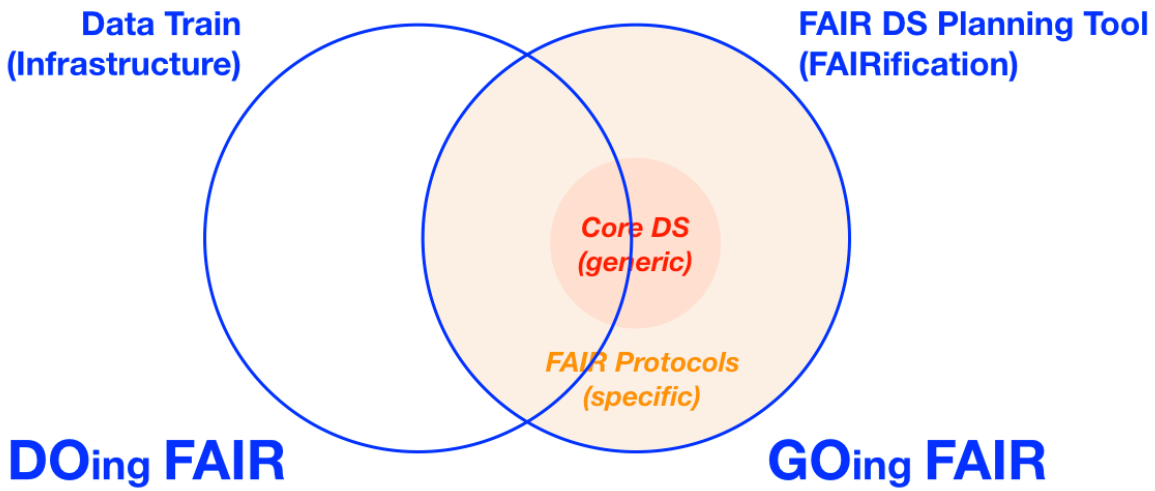
Data in the tool is highly configurable. It is separated in Generic and Local: So we can separate life science specific (like “human health data”), ELIXIR specific (like names of experts), and even for local institutes we can adapt it.


How a Chemistry-focused FAIR Data Stewardship Planning Tool can
(1) Guide FAIRification efforts in chemistry (choosing standards)
(2) Organize the larger community (draw-in stakeholders)
GO FAIR
Chemistry Implementation Network (ChIN)
• Academia
• IUPAC
• CODATA
• Funders
• Publishers
• Instrument makers
DMPOnline
DMPTool
DS Wizard + FAIR Metrics
Task 1:
ChIN community chooses DS tooling
FAIRsharing.org
FAIR Metrics
CEDAR
Task 2 (ongoing):
ChIN community chooses / creates
• standards
• databases
• policies
Funding Agencies
Requires FAIR DS Plans
ChIN-FAIR DS Planning Tool
Core DS (generic)
ChIN-FAIR Protocols (specific)
Data Stewards
Researchers: Better data reuse
Builds with ChIN-FAIR standards
Mandates FAIR DS
Industry
• Data capture
• Matadata capture
University
• Institutional policy for FAIR DS
• Training programs for FAIR DS

go-fair.org

What is the DOA ?

- DOs are not just passive, but they are active, i.e. a DO has a type and there are functions assigned to types which are the basic step towards automatism
-


Data Train
(Infrastructure)
FAIR DS Planning Tool
(FAIRification)
Core DS
(generic)
FAIR Protocols
(specific)
DOing FAIR
GOing FAIR

Convergence
Social versus techncial
DOHIP = DO Human Interface Protocol
DOFIP = DO FAIR Interface Protocol

Semantically informed (meaningful) rouing functions |

Kahn Meaningful exchange protocol for any data type

- Handels for Digital Objects to resolve identities into useful "state" information
- DOIP

Schwardmann The DOIP describes a world of digital objects and operations provided b repositories, referenced by identifiers and enhanced by type metadata, which is made interoperable again by type definitions in data type registries.

Kahn Meaningless exchange protocol for any data type

- Handels for Digital Objects to resolve identities into useful "state" information
- exchange of in general meaningless messages between Internet devices with an IP address
- DOIP

Introduction of Dr. Robert Kahn

Dr. Kahn was not only one of the two designers of TCP/IP and thus at the source of our current days Internet, he also wrote in 1995 the first paper on Digital Objects together with Robert Wilensky. This early paper was revised in 2006. Since TCP/IP only describes the exchange of in general meaningless messages between Internet devices with an IP address, there was a need to exchange meaningful entities. FTP was an early protocol to exchange files and HTTP was another very successful protocol to exchange Web information. Kahn & Wilensky realised the need to define a protocol that is generic and thus includes files, web pages and other possible entities which they called Digital Objects. Kahn and his team at CNRI then started to design and develop components that could realise such a world of DOs. The most well-known component is the Handle System which allows anyone now to assign Handles to Digital Objects and to resolve identities into useful "state" information. With the setup of the DONA Foundation the Handle System can now be seen as a common good not owned by one person or company anymore. Other components have been design and partly developed such as the Digital Object Interface Protocol which may indeed change our practices.
It should be mentioned that Dr. Kahn got many awards for his work amongst which is the Turing award.

Schwardmann The DOIP describes a world of digital objects and operations provided by repositories, referenced by identifiers and enhanced by type metadata, which is made interoperable again by type definitions in **data type registries**.

Ulrich Schwardmann (ePIC): Objects, types, collections and operations in DOIP

The DOIP describes a world of digital objects and operations provided by repositories, referenced by identifiers and enhanced by type metadata, which is made interoperable again by type definitions in data type registries. In the world of digital objects one always can see valuable applications like collection repositories to structure these objects in a generic way and to build user specific views and work benches on such structures by a digital object browser. The next step is to include also operations into this picture. Since with REST services operations are already ubiquous in the HTTP world, the question is, how this huge amount of existing technology can be adapted into the DOIP world. The talk will also try to show a possible bridge here.

DOIP
- DOs
- Operations (repositories)
- PIDs
- Type metadata
- Data type registries

Schultes DO versus FAIR

Erik Schultess (GO FAIR): GOing FAIR & DOing FAIR

The Digital Object framework is an abstraction layer striving towards technology-independent, future-proof, and increasingly automated operations between data, software, and compute resources. The 15 FAIR Principles are high-level specifications for the automated Findability, Accession, semantic Interoperation, and Re-use of data, software and services. Although the DO Framework and the FAIR Principles have very different origins, they nonetheless share overlapping and complementary features. In their current states of development, the DOIP specifies elementary “informatics" operations on DOs, while the FAIR Principles provide some specification for the technical, domain specific, and provenance metadata that are necessary to inform DOIP operations. I will describe GO FAIR, a bottom initiative coordinating a very large and diverse stakeholder community actively building implementations (including “Metadata for Machines”) that demonstrate the FAIR Principles in practice. I will describe how ongoing GO FAIR activities, including the C2camp Implementation Network, could play a role in accelerating the adoption and application of DOs in an emerging data infrastructure.

DO: The Digital Object framework is an abstraction layer striving towards technolog
independent, future-proof, and increasingly automated operations between data,
software, and compute resources.

FAIR: The 15 FAIR Principles are high-level specifications for the automated Findabil
Accession, semantic Interoperation, and Re-use of data, software and services.

Xxx

Twan Gosen, Dieter van Uytvanck (CLARIN): Digitals Objects as direct input into the CLARIN Language Resource Switchboard

Numerous repositories offer data and associated metadata, following the CMDI specification, within the CLARIN infrastructure. The CMDI metadata is harvested and collected in a central repository: the Virtual Language Observatory (VLO). After finding a relevant piece of information, the data object and metadata object can be provided, either as bit streams or by identifiers, to the Language Resource Switchboard (LRS), to get a suggestion of tools available to operate on the data object. In this context, digital objects are a natural fit, **provided the available metadata matches our needs, to enhance the LRS**. By providing a digital object identifier to the switchboard, the switchboard can utilize the DO protocol to obtain all relevant information about the digital about, such as mime type and language. Any of the suggested tools can in turn utilize the DO protocol to obtain a bitstream to the actual object data itself in order to run the tools processing pipeline.

Margareta Hellström (ICOS): How a Digital Object Architecture could help ICOS streamline data service provisioning

The ICOS Carbon Portal (CP) manages, curates and disseminates both greenhouse gas observational data (measured by its own station networks), as well as outputs of advanced atmospheric and ecosystem models (provided by external parties). The CP assigns Handle-based PIDs (from ePIC or DataCite) to all data objects it manages. All relevant metadata are kept in the CP catalogue , which is based on semantic web & open linked data (LOD) concepts. We are now actively looking into the best way to support (automated) workflows and processing of ICOS data in cloud environments (like EGI federated cloud). Important aspects include optimizing access to, and linking of, all relevant metadata from various sources (the ICOS catalog, databases at PID registries, and others), including descriptive information (context of acquisition), data contents (variable types), and data processing basics (model version). It remains to be seen to what degree ICOS can adapt the DO Network approach, but **we are willing to be one of the test cases** – perhaps with a special focus on combining data type definitions stored across both LOD-based and PID-based registries.

Carlo Maria Zwölf (VAMDC): The DO Case in Virtual Atomic and Molecular Data Centre

The VAMDC e-infrastructure federates into an interoperable way ~30 heterogeneous and independent atomic and molecular databases. All the outputs produced by this infrastructure are identified by a resolvable PID (Persistent Identifier) and are formatted using a rigorous XSD schema (XML Schema for Atoms Molecules and Solids; This schema is a computer model for all the processes and physics contained in VAMDC).
The data extracted from VAMDC are indeed a good approximation of what is designed by "Digital Object".
In our presentation we will describe the main issues we experienced:

- In using for scientific purposes generic repositories (e.g. Zenodo, Eudat) for storing VAMDC-DO.
- In automatic handling (e.g. comparison, cross-matching) of VAMDC-DO.

XxX

world's biodiversity; it's past, present and future and its influence on global challenges in environment and society. In June 2018 the European Strategy Forum on Research Infrastructures (ESFRI) accepted the importance of this resource and included the Distributed System of Scientific Collections (DiSSCo) into the ESFRI Roadmap 2018 as a priority research infrastructure to commence operations in 2025.

With an expected 30-year lifespan, DiSSCo aims at digital transformation of today's slow, expensive, inefficient and limited system where the need to physically visit collections and the absence of linkages to relevant information represent significant impediments. Our architecture inserts a **'Digital Specimen Object Layer'** unifying natural science collections into a single data-driven European virtual Collection offering wider, more flexible, 'FAIR' access for a range of biodiversity science and policy applications. This is expected to lead to faster insights for lower cost. Acting as surrogates for the physical specimens in collections, DiSSCo places Digital Specimens at the heart of an interconnected graph of diverse and dispersed data classes that can include imagery, taxonomy, relevant scientific literature, genetic sequence and trait data, agricultural, toxicology and ecosystem data and much more.

Immediate concerns for consideration in the present workshop include: i) the social aspects of 'selling' the benefits of the digital object approach; ii) achieving balance between fast presentation of informative registry records and the need to fetch and unpack comprehensive object content from a repository; and iii) extensibility of dynamic Digital Specimens to support new information types whilst maintaining backwards compatibility for older systems.

Alex Hardisty
Director of Informatics Projects, School of Computer Science and Informatics, Cardiff University

11th September 2018.

Xxx

- Also beyond data processing, the diversity of consumers for ENES data products and data services is increasing. This includes users largely unfamiliar with the data generation and refinement process, which presuppose knowledge of limitations and assumptions not obvious to those unfamiliar with Earth system modelling. ENES therefore has a strong motivation to enter discussions with **new user communities**, particularly in the area of **social sciences, public administration, planning and policy making.**

The ENES community looks towards initiatives such as C2CAMP to address these points, starting with individual prototypes and small-scale solutions. There is however now a good opportunity to reach out to other communities and work towards a larger, overarching architecture and operationally capable solutions, based on the concepts and models the C2CAMP participants have discussed and matured in the past. Central design principles are adherence to agreed interfaces, for example based on specifications from RDA, IETF or W3C; finding the balance between complexity and feasibility, particularly regarding **metadata and semantics**; and building long-term sustainable services for data processing and **provenance tracking**.

Xxx

Dear all,

yesterday I had a very productive interaction with Erik also about the joint paper. I cc to Larry as well although Larry is not involved in the DAMDID paper.
Let me briefly make a few comments about the DAMDID paper:

- I read it again and I think it is an excellent paper. Erik elaborates on the way how TCP/IP made it and this is a nuance that we did not express. As far as I remember the way Erik puts is absolutely correct, but George knows better. So it adds new aspects to our paper in addition to all the other statements about FAIR and GO FAIR etc. . I just have one question about the text – see below.
- I should not appear as co-author this time for two main reasons: 1) I am reviewer of proposals where I believe Barend is involved, so I should not be too close to GO FAIR office etc. 2) As a co-organiser of the DAMDID workshop, I should not appear in too many papers. I will present something on DO and I will be mentioned by Simon as contributor to the FAIR EG document. That's sufficient.
- But if George would appear as co-author it would be great showing how close we synch.
- I have one small point wrt to the text. On page 2 you start saying that by the early 2000's etc. people see the need for general purpose infrastructures which is correct. Then you immediately jump to the Semantic Web. This too fast for me. Sem Web is certainly one important pillar, but there is all the work on data/research/cyber infrastructures etc. which started in some communities also in 2000 and which led to the ESFRI infrastructures in Europe.

A few other points:
- The more I think about our current work I see more and more clear that joining the ideas behind FAIR and DOs is really adding to get a critical pivot point. I agreed with Erik to go a bit deeper into this relationship. A short doc will come today perhaps and we might want to add a sub-chapter in chapter 5 called "FAIR and DOs".
- Erik is discussing

Xxx

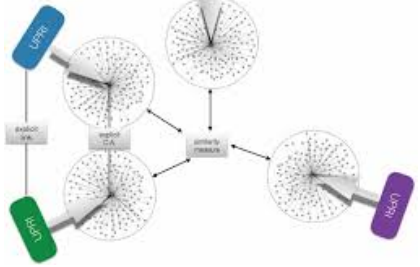
Digital Object View

The notion of Digital Object is increasingly often accepted and a repository itself can be s described by different types of metadata. RDA DFT has defined a digital collection in a recursive way a uch a complex collection. This way of looking at repositories will facilitate automation.

FAIR metadata

Title	FDP of biosemantics group
Metadata ID	fdp
Description	This is a prototype FDP for hosting research and student projects datasets
Issued	2017-05-23T09:43:15.57Z
Modified	2018-08-20T13:09:55
Version	1.0E0
License	http://rdflicense.appspot.com/rdflicense/cc-by-nc-nd3.0
Access Rights	This resource has no access restriction
Specification	http://rdf.biosemantics.org/fdp/shex/fdpMetadata
Language	http://id.loc.gov/vocabulary/iso639-1/en
Publisher	Biosemantic group
Metrics	Type https://purl.org/fair-metrics/FM_F1A Value https://www.ietf.org/rfc/rfc3986.txt Type https://purl.org/fair-metrics/FM_A1.1 Value https://www.wikidata.org/wiki/Q8777
Catalogs	http://136.243.4.200:8087/fdp/catalog/Transcriptomics http://136.243.4.200:8087/fdp/catalog/multiomics http://136.243.4.200:8087/fdp/catalog/textmining http://136.243.4.200:8087/fdp/catalog/Biosamples http://136.243.4.200:8087/fdp/catalog/Patient_Registries_1.0_998ccbcf-8714-426a-a28e-9335a86adb19 http://136.243.4.200:8087/fdp/catalog/SCA3_HD_multi-omics_blood_data
Institution Country	http://lexvo.org/id/iso3166/NL
Download RDF	ttl rdf+xml jsonld

Xxx



Statements

General

1. Metadata statements are assertions about Digital Objects (which can be anything digitally represented including collections of digital entities) and are part of the DO to make it findal interpretable and re-usable.
2. Metadata assertions are made by different actors (humans, machines) with different roles for different purposes at different times.
3. There are quite a number of different types of (metadata) statements about the DO all being summarised under the term "metadata":
    - type information to facilitate automatic processing,
    - descriptive information to facilitate searches, scientific collection building, inferencing etc.,
    - scientific information which goes much deeper than the usual descriptive information to enable deep scientific analysis,
    - system/state information to help managing the DOs,
    - provenance information to cover creation information,
    - context information to cover the context of DOs emergence,
    - access rights information to indicate who is allowed to access the content,
    - license information to indicate the terms under which re-usage may happen,
    - transaction information to cover events of re-usage,
    - annotations on content to cover information added to parts of the DOs content,
    - etc.
4. Many of these types are not yet clearly defined and agreed. However, there is an urgent need for convergence on classification.

Digital Objects, Schemas, Categories

5. Metadata objects should be Digital Objects themselves, i.e. they should have a PID.
6. Metadata should include the PID of the DO it is associated with and the DO's PID record should point to the DO of the metadata object. Bi-directional links will ensure that machines c
7. Much metadata still exist as undocumented spreadsheets or databases with privately defined categories as DataONE has shown.
8. Metadata assertions are being represented using various technologies (XML objects, relational databases, RDF assertions, etc.) dependent on the context of creation.
9. The used "metadata categories" need to be defined in open registries to facilitate re-use and interoperability in particular by machines requiring referential and functional integrity.
10. Schemas that are being used to describe the structure of XML objects or database tables need to be registered in open registries to facilitate re-use and interoperability.
11. An increasing amount of schemas is being used and more reuse should be promoted by using existing such as RDA Metadata Directory and schema.org.
12. The used explicit relationship types need to be defined in open registries and should reuse existing standards such as SKOS, OWL, etc.
13. Metadata schemas in some communities include shallow hierarchies to allow bundling.

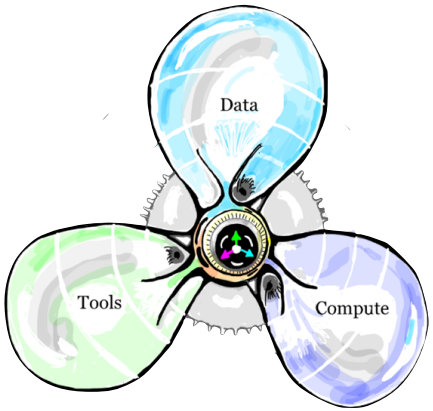
Exporting & Reuse

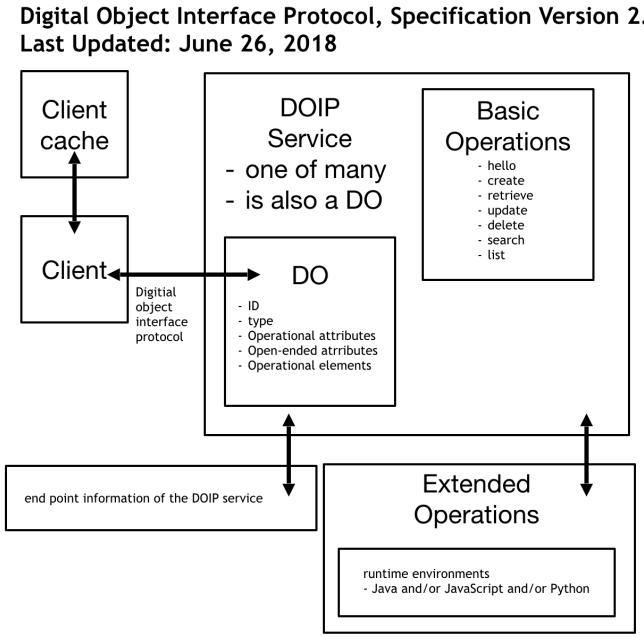
14. Metadata is being used for many different fields of application such as for Data Management Plans, for publishing data, for searching by occasional users, for collection building by sci scientific analysis, for the orchestration of workflows etc.
15. The interpretation of the term "rich" in relation to metadata as is used in the FAIR principles is very much dependent on the field of application. What may be sufficiently rich for gen all sufficient for workflow orchestration etc.
16. It is not yet clear whether we can design a component based system that allows users to reuse components in an incremental way which would not require to generate metadata for c again and again.
17. Metadata of most types should be open and thus will be re-used and changed in various ways to satisfy different purposes. Copied metadata objects therefore live their own life.
18. It is widely agreed that for example for indexing supporting fast searches and similar operations metadata should be offered for harvesting using standard protocols such as OAI/PMH c
19. Independent of the chosen representation technology metadata should be offered as RDF triples with explicit relational semantics to allow further semantic processing.
20. Knowlets (see image) are excellent ways to point to core concepts and to represent their semantic relationships in a specific semantic space and thus extract relevance, meaningfuln relationships, etc. They are utterly dynamic constructs due to continuous changes leading to complex management tasks
21. Not all metadata types/assertions are relevant for constructing useful Knowlets. Format specifications for example can hardly be exploited.

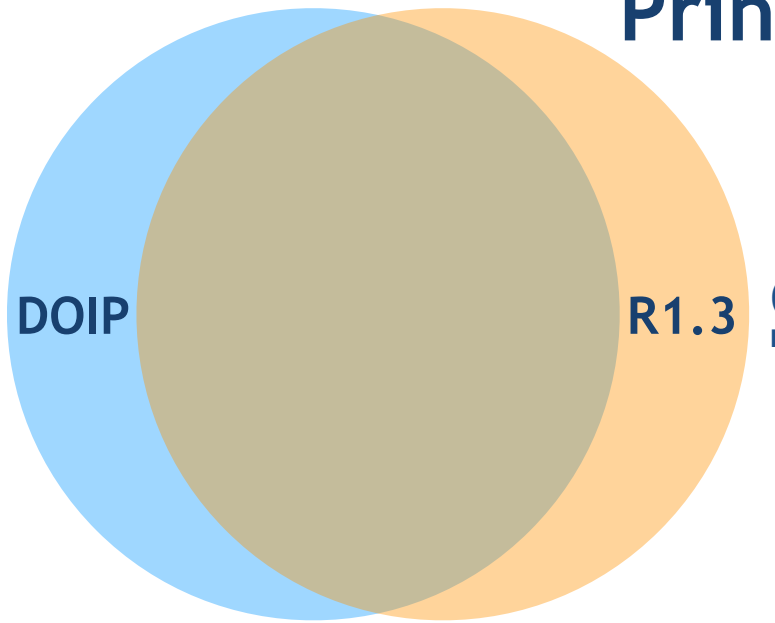
Xxx

a switchboard solution linking specific data types to tools that have proven their usefulness for these types. DOs are exactly the type of solution allowing realising such a switchboard elegantly since they are typed and have other relevant metadata that can automatically be retrieved and these types can be associated with specific tools using the same basic mechanisms. Using these methods would enable a "computer-naive" researcher to create his/her workflows only applying domain knowledge that then can automatically process data of specific types in the intended way.

RDA, FORCE11, C2CAMP, GO FAIR, at PIDapalooza







(meta)data meet domain-relevant community standards;

"hard" part of DOA is the domain level metadata

DOing FAIR (meaningful exchange)

Conclude that metadata is GO FAIR… the DS tool and M4M as convergers

GOing FAIR (converging on metadata)

FAIR Principles

globally unique and persistent identifier

rich, community agreed upon, presistent metadata

broadly applicable language for knowledge representation

standardized communications protocol

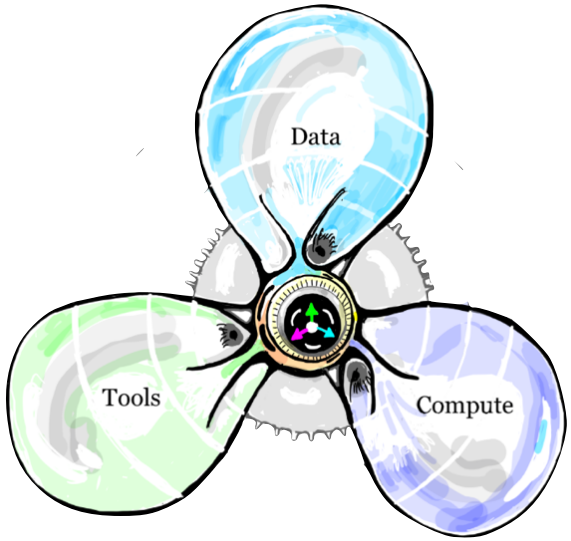
DOA

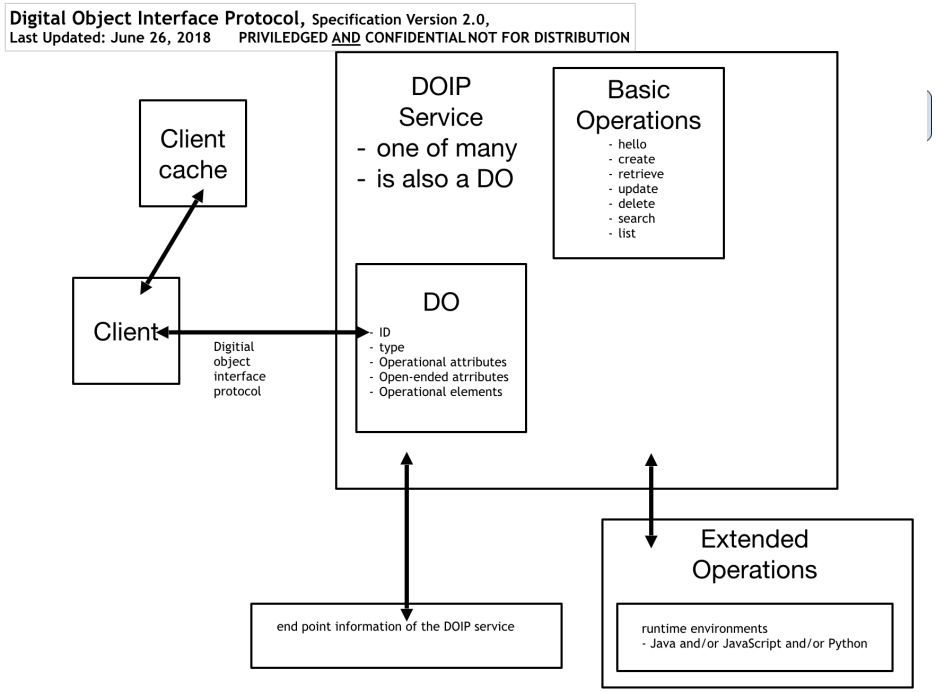
PID

Metadata: typing

Metadata: semanticlly meaningful

DOIP





FAIR

F, A, I & R by machines

GO FAIR

Convergence

- Minimal standards / Maximum freedome to innova
- Voluntary (but aware of attractors, no vendor lock
- Working implementations (decisions)

DOA

DOs, DOA, DOIP

DOing FAIR

What is the realtionship between

FAIR Data
&
Digital Object Architecture

Data Infrastructre:
Vision
Creolization
Attraction
Convergence
Exploitation

	Internet	WWW	IFDS
Objects	networks	Webpages	Data
Identifers	Internet adresses	URLs	PID
exchange function	TCP/IP	HTTP	DOIP

Questions from Margreet:

1. Registers for (pre)clinical studies (such as the one she is presenting) must be FAIR > therefore you showed Nicoline the assessment of the 9 example repositories. *She can assess her register as well.*

FM	Question	Dataverse	Dryad	Nano-pub	Zenodo	Yale ISPS	Figshare	Broad's SCP	SeaData Net's CDI	Wikidata
IRI Exists	1	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI
F1A	2	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI
F1B	3	IRI	IRI	IRI	NRP	none	IRI	IRI	IRI	IRI
F2A	4A	IRI	IRI	IRI	IRI	none	none	IRI	IRI	IRI
F2A	4B	IRI	none	IRI	IRI	"Multiple"	none	IRI	IRI	IRI
F3	5A	IRI	IRI	IRI	IRI	none	NRP	IRI	IRI	IRI
F3	5B	IRI	IRI	IRI	IRI	IRI	IRI	IRI	none	IRI
F4	6A	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI
F4	6B	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI
A1.1	7A	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI
A1.1	7B	true	true	true	true	true	true	true	true	true
A1.1	7C	true	true	true	true	true	true	true	true	true
A1.2	8A	false	false	false	false	false	false	false	true	false
A1.2	8B	N/A	N/A	N/A	N/A	NRP	NRP	NRP	link	N/A
A2	9	IRI	IRI	none	IRI	none	IRI	none	IRI	NRP
I1	10	IRI	IRI	IRI	IRI	none	none	NRP	IRI	IRI
I2	11	IRI	IRI	IRI	none	none	none	IRI	IRI	IRI
I3	12	NRP	IRI	IRI	none	none	none	NRP	NRP	IRI
R1.1	13	IRI	IRI	IRI	IRI	IRI	IRI	NRP	IRI	IRI
R1.2	14A	IRI	IRI	IRI	IRI	none	none		NRP	NRP
R1.2	14B		none		none	none	none			
R1.3	15	NRP			none	none	none	NRP		

Response from Erik:

Indeed, 9 digital resources answered the 21 questions in a questionnaire prompting the necessary inputs to the 14 FAIR Metrics:
https://www.nature.com/articles/sdata2018118

Here is a link to those 21 questions:
https://docs.google.com/document/d/1EerbZVvT-C6LblcVRrVP4B7H-reD3jU8LNEmBZogTZo/edit?usp=sharing

We can follow the examples, and do the same for the preclinicaltrials.eu resource.

Questions from Margreet:

2. A database generated or reused in a (pre)clinical study must be FAIR > we can discuss the possible answers to the corresponding community challenges. In the case of Nicoline, the community is formed by (pre)clinical researchers. Preregistration is a specific feature of the community of (pre)clinical researchers. *How can we fit in preregistration as a FAIR metric? You suggested F1A, F1B. I think F4 would be one as well.*

Response from Erik:

We can create machine readable metadata that describes explicilty the preregistration. We can also create a new FAIR Metric that validates (cdertifies) the preregistration (see template below).

Metric Identifier	FM-CT1 (FAIR Metric Clinical Trail 1)
Metric Name	Preregistration
To which principle does it apply?	R1.2 (meta)data are associated with detailed provenance
What is being measured?	The existence of clinical trail preregistration
Why should we measure it?	Preregistration is important for Increased transparency and reduced risk of bias and help avoid duplication.
What must be provided?	A URL to the preclinical registration document
How do we measure it?	Use HTTP GET on URL provided.
What is a valid result?	Now, HTTP 200; Later, validted RDF file
For which digital resource(s) is this relevant?	preclinicaltrails.eu

Questions from Margreet:

3. ZonMw wants to use FAIRmetrics for monitoring datamanagement in its projects. ZonMw wants to make discipline or community specific metrics. Therefore, I would like to discuss on sept 21: *what metric(s) should ZonMw apply to check whether a (pre)clinical researcher has preregistered its study and database in a FAIR (pre)clinical register?*

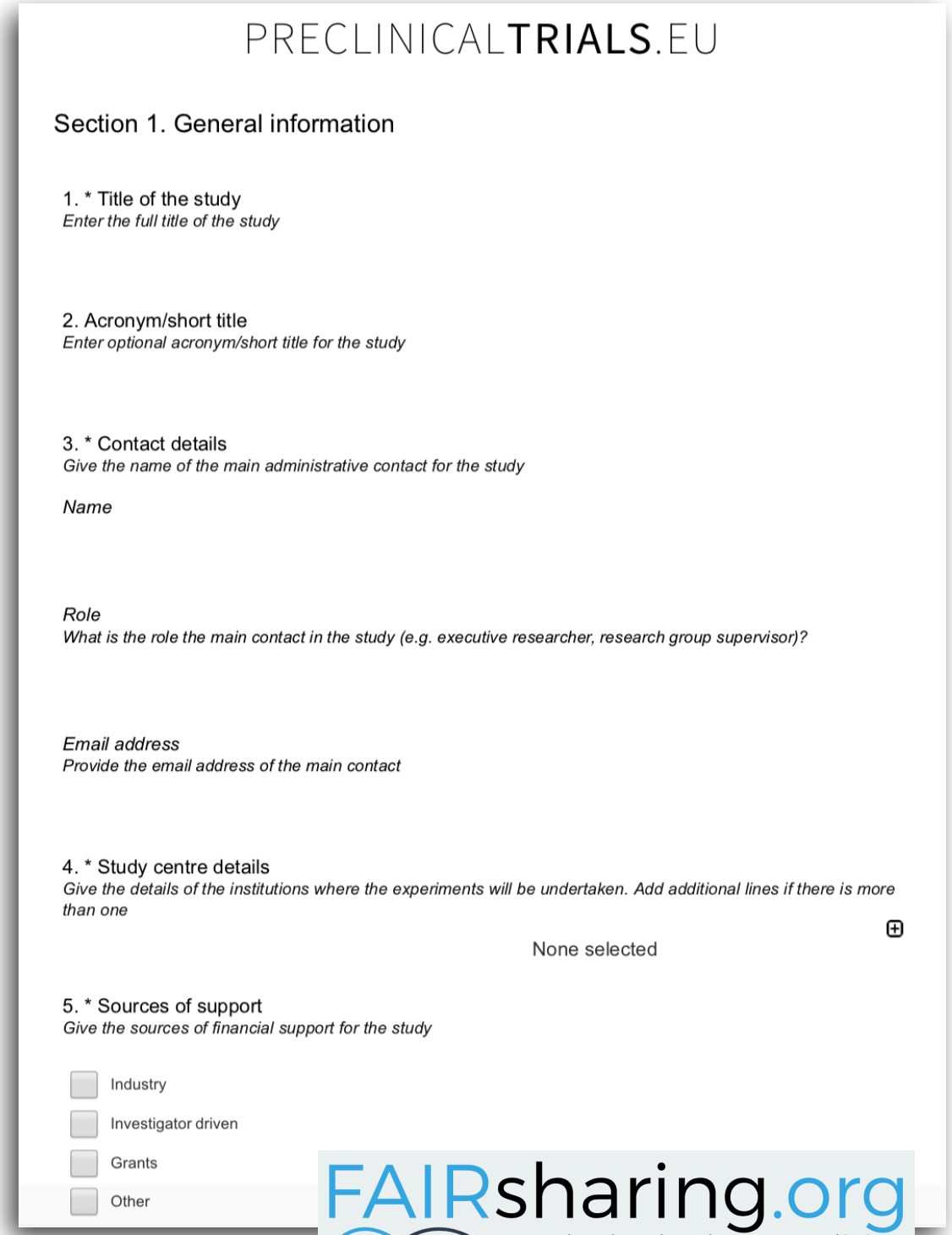
Would it be sufficient to make a community specific version of the FAIRmetrics F1A, F1B, F4?

Or should we design a new metric?

The community must use the FAIR Metrics Framework (sheet 8 in your slide set) to make a community specific metric > is it to your opinion feasible to show/discuss this framework in the workshop?

Response from Erik:

Yes… great use case, and worth discussing in the workshop.

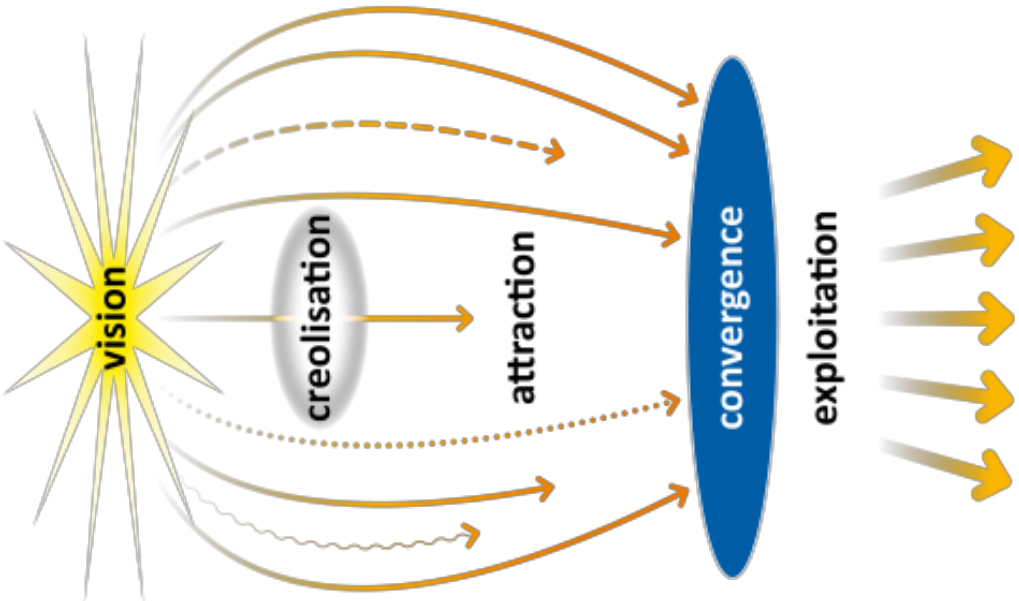


FAIRsharing.org
standards, databases, policies

FAIR and GO FAIR

Common Patterns in Revolutionary Infrastructures and Data

Peter Wittenburg, Max Planck Computing and Data Facility
George Strawn, US National Academy of Sciences
February 2018
https://www.rd-alliance.org/sites/default/files/Common_Patterns_in_Revolutionising_Infrastructures-final.pdf



	Internet	WWW	IFDS
Objects	networks	webpages	data
Identifers	Internet adresses	URLs	PID
exchange function	TCP/IP	HTTP	DOIP