



# A Portal to (Meta)data

**Bruna dos Santos Vieira**

FAIR Data Team

Data Stewards training  
RI, 20 June 2023

Health-

---

## The Vision

To have a smart infrastructure for research and innovation where users easily **find** resources and **reuse** them



---

## Requirements

Infra must require that resources are described according to **international standards** (rich enough and interoperable):



- **DCAT** (EU Health Data Space > DCAT Health, DCAT AP for Portals)

And that users provide as much as possible **semantic enriched metadata** (unambiguous, machine-interpretable) for their resources of interest



What does this icon mean?





What does this icon mean?  
Where can these icon be found?

Battery level





All mobile electronic devices





All mobile electronic devices





All mobile electronic devices



All mobile electronic devices that take photographs







All mobile electronic devices



All mobile electronic devices that take photographs





### All mobile electronic devices



### All mobile electronic devices that take photographs



DSRL camera



Smartphone / mirrorless cameras



# Its the same for research metadata

All datasets have titles and descriptions



Some datasets (with title and description) share more properties in a specific domain



A subdomain shares even more specific properties





## HRI Core Metadata Schema

- Health-RI will supply a generic schema for users to increase findability of resources of interest (e.g. datasets)
- Based on international standards (e.g. DCAT, DCAT-AP portals) and domain metadata requirements (e.g. specific for OMICs data)



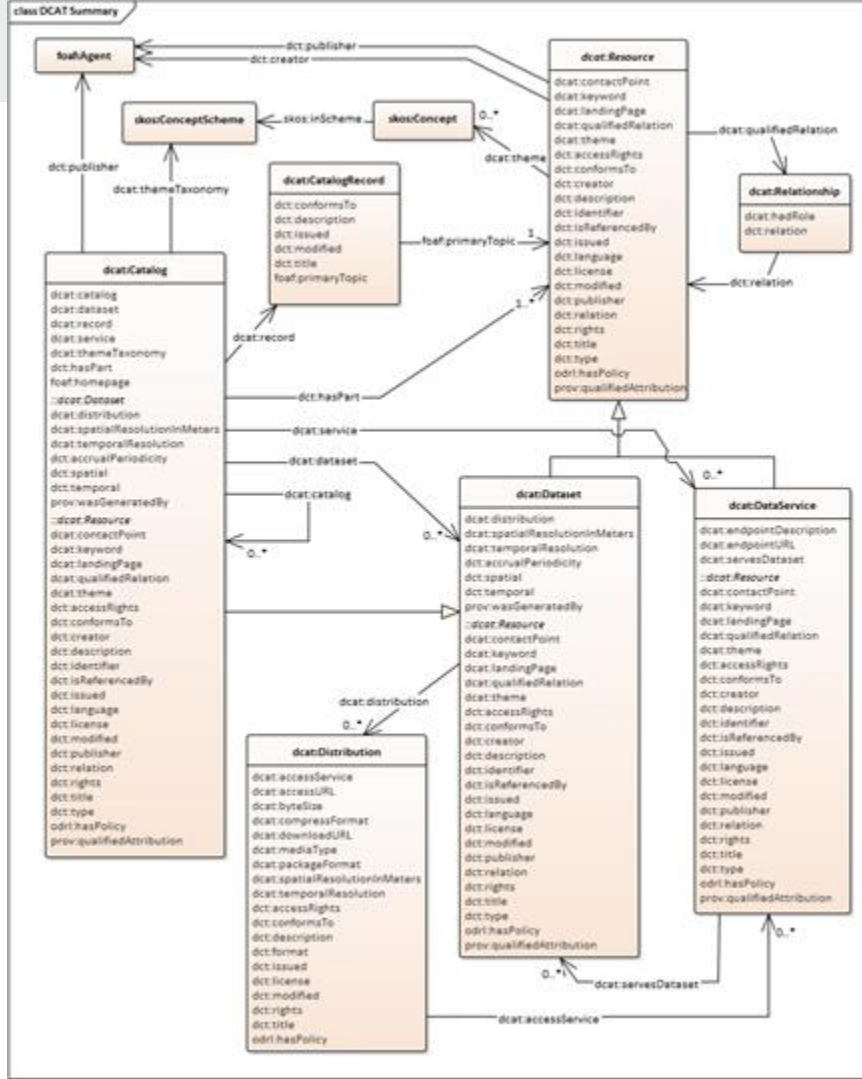
## Structure of metadata schemas

- Classes (e.g. dataset)
- Classes have Properties (e.g. dataset's title)
- Properties should point to an instance (e.g. "*The Human Genome Dataset*")

Person > Person's name > Bruna



# DCAT



<https://www.w3.org/TR/vocab-dcat-2/>

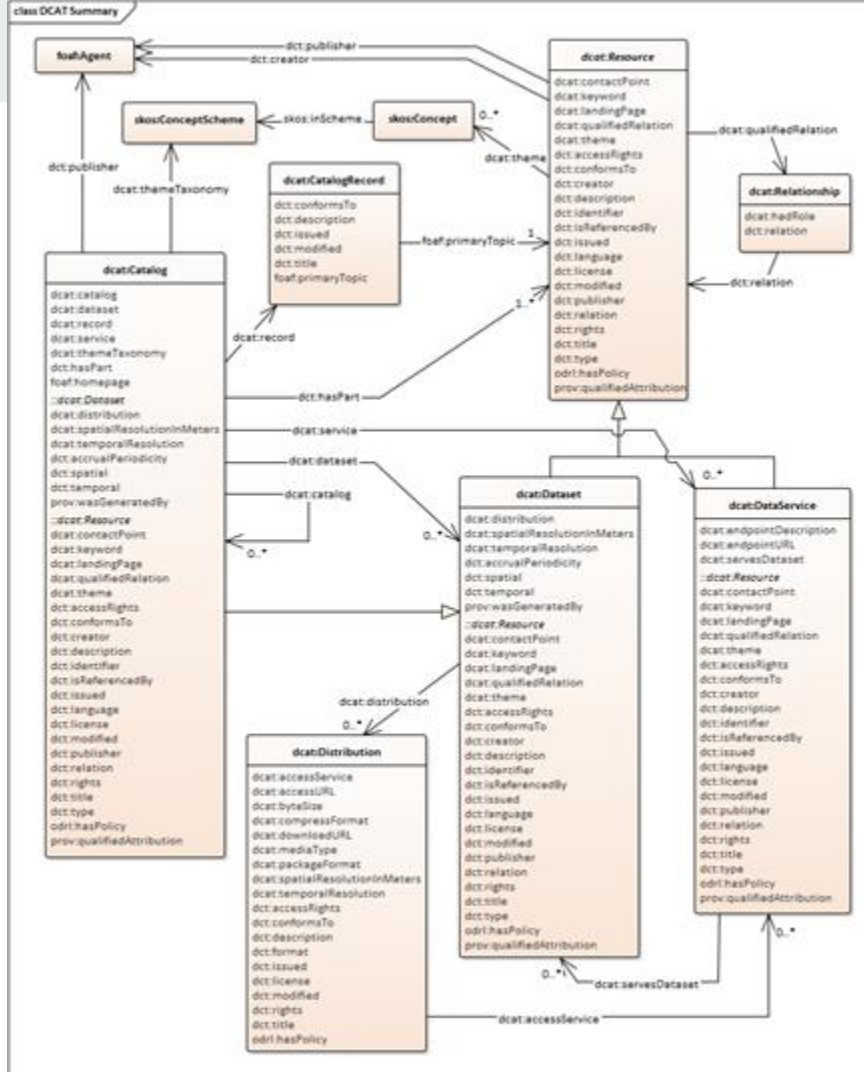
# DCAT

Main classes:

- Catalog
- Dataset
- Distribution

New classes:

- data service
- relationship
- catalogRecord



And it borrows from:  
**foaf:Agent**  
**skos:Concept**,  
**ConceptScheme**

<https://www.w3.org/TR/vocab-dcat-2/>



DCAT Dataset class inherits :: properties of the resource super-class

<b>dcat:Dataset</b>
dcat:distribution
dcat:spatialResolutionInMeters
dcat:temporalResolution
dct:accrualPeriodicity
dct:spatial
dct:temporal
prov:wasGeneratedBy
<b>::dcat:Resource</b>
dcat:contactPoint
dcat:keyword
dcat:landingPage
dcat:qualifiedRelation
dcat:theme
dct:accessRights
dct:conformsTo
dct:creator
dct:description
dct:identifier
dct:isReferencedBy
dct:issued
dct:language
dct:license
dct:modified
dct:publisher
dct:relation
dct:rights
dct:title
dct:type
odrl:hasPolicy
prov:qualifiedAttribution

...and adds extra dataset-specific properties







## DCAT-AP Portal

- Extension of DCAT
- Defined mandatory fields for EU Portals



# Comparison

Structure DCAT AP:


- Mandatory or optional
- Cardinality
- No (copy) inheritance
- Borrows properties from other schemas
- Borrows classes from dcat
- Specific for data portals

<b>«mandatory» dcat:Dataset</b>	<b>dcat:Dataset</b>
<b>«optional»</b> adms:versionNotes [0..*] dcat:qualifiedRelation [0..n] dcat:spatialResolutionInMeters [0..n] dcat:temporalResolution [0..n] dct:creator [0..1] dct:identifier [1..*] dct:isReferencedBy [0..n] dct:issued [0..1] dct:modified [0..1] owl:versionInfo [0..1] prov:qualifiedAttribution [0..n] prov:wasGeneratedBy [0..n] <b>«mandatory»</b> dct:description [1..*] dct:title [1..*]	dcat:distribution dcat:spatialResolutionInMeters dcat:temporalResolution dct:accrualPeriodicity dct:spatial dct:temporal prov:wasGeneratedBy ::dcat:Resource dcat:contactPoint dcat:keyword dcat:landingPage dcat:qualifiedRelation dcat:theme dct:accessRights dct:conformsTo dct:creator dct:description dct:identifier dct:isReferencedBy dct:issued dct:language dct:license dct:modified dct:publisher dct:relation dct:rights dct:title dct:type odrl:hasPolicy prov:qualifiedAttribution

Structure DCAT

- No mandatory or optional
- No cardinality
- Copies inheritance
- Borrows properties from other schemas
- Borrows classes from dcat
- generic

# Defining HRI Metadata Schemas

- Technical Metadata Team (TMT  ) will support the technical part of building a metadata schema (Luiz, Kees, Bruna)
- Only domain experts (working groups and portal teams) can define the metadata **properties** (content).
- TMT add main properties from DCAT, DCAT AP portals, and supplied schemas from the nodes
- **Who will tell, apart from DCAT, which are the other obligatory fields?**

Once domain schemas are out for review:

- Community will review for relevance
- Portal groups will review taking into account usability, visuals etc
- TMT final check and embed the obligatory fields in the core schema

# The core metadata schema



Some datasets (with title and description) share more properties in a specific domain



A subdomain shares even more specific properties





# Defining Core

Currently we have:

- Identified EU standards (DCAT, DCAT AP Portals, DCAT AP Health)
- Collected some NL Nodes and Health-RI metadata schemas (RUMC, AUMC, Princess Maxima, Covid Portal)
- Mapped all of the above [here](#)

For HRI portal release 0.9:

- Minimal Core: DCAT-AP Portals mandatory fields

For later (HRI portal release 2.0):

- Plan to release HRI Core metadata schema answering:  
**What apart from DCAT AP Portals mandatory fields should be in the HRI core?**

# The Petals metadata schema

All datasets have titles and descriptions



Some datasets (with title and description) share more properties in a specific domain



A subdomain shares even more specific properties



Petals





## Defining Petals - example Omics Domain

- Domains (e.g. Omics group) will specialize the generic schema into their needs and properties (e.g. add omics metadata such as ISA tab elements, or the extensions made in the FAIR data cube, FAIR Genome and X-Omics projects)
- Feedback / Result from Domain groups expected to be shared via Github [health-ri-metadata/Leaves/Omics/](#)
- Request Omics group lead to add you to the HRI Metadata repo [health-ri-metadata/Leaves/Omics/](#)



# What metadata should you prioritise

## FAIR

- Important to **find** your dataset (e.g. diagnosis, sample size, subjects (people,demo))
- Increase **accessibility** (which protocol was used e.g. a form sent to the medical ethical committee)
- Increase **interoperability** (which vocabulary, coding language was used in your data)
- Increase **reusability** (consent/license, provenance, standards used for coding your data, study protocols as a quality standard, pointer to the data)

## Other

- Important for the “visuals” of the portal (e.g. Logo URL, Landing Page URL)
- Important for your **domain** (e.g. tnm for an onco/ cancer dataset, profiling for omics)



## Agreeing on properties - example

- Explain all classes and properties (or at least all dataset's properties)
- Vote for:
  - mandatory/ optional
  - cardinalidade (min, max)
  - Identify non-technical items which may be important for portal (e.g. resource logo URL, resource landing page)
- Results of voting will define the minimal schema for first release



# Metadata Schemas and Portal Releases

- Published HRI Core Metadata Schema (*w obligatory fields of DCAT AP + what apart from DCAT-AP do we need?*)
- Published Domain Schemas (Leaves) (*w obligatory fields per domain*)
- Documentation for users to follow on “how to describe their resource”

## Releases

- DCAT-AP Portals mandatory fields will be required for portal **0.9** release
- HRI Core Metadata Schema (what apart from DCAT-AP do we need?) and leaves expected to be released for portal release **2.0**

# The Sunflower

## Core:

**Minimal** DCAT AP Portals

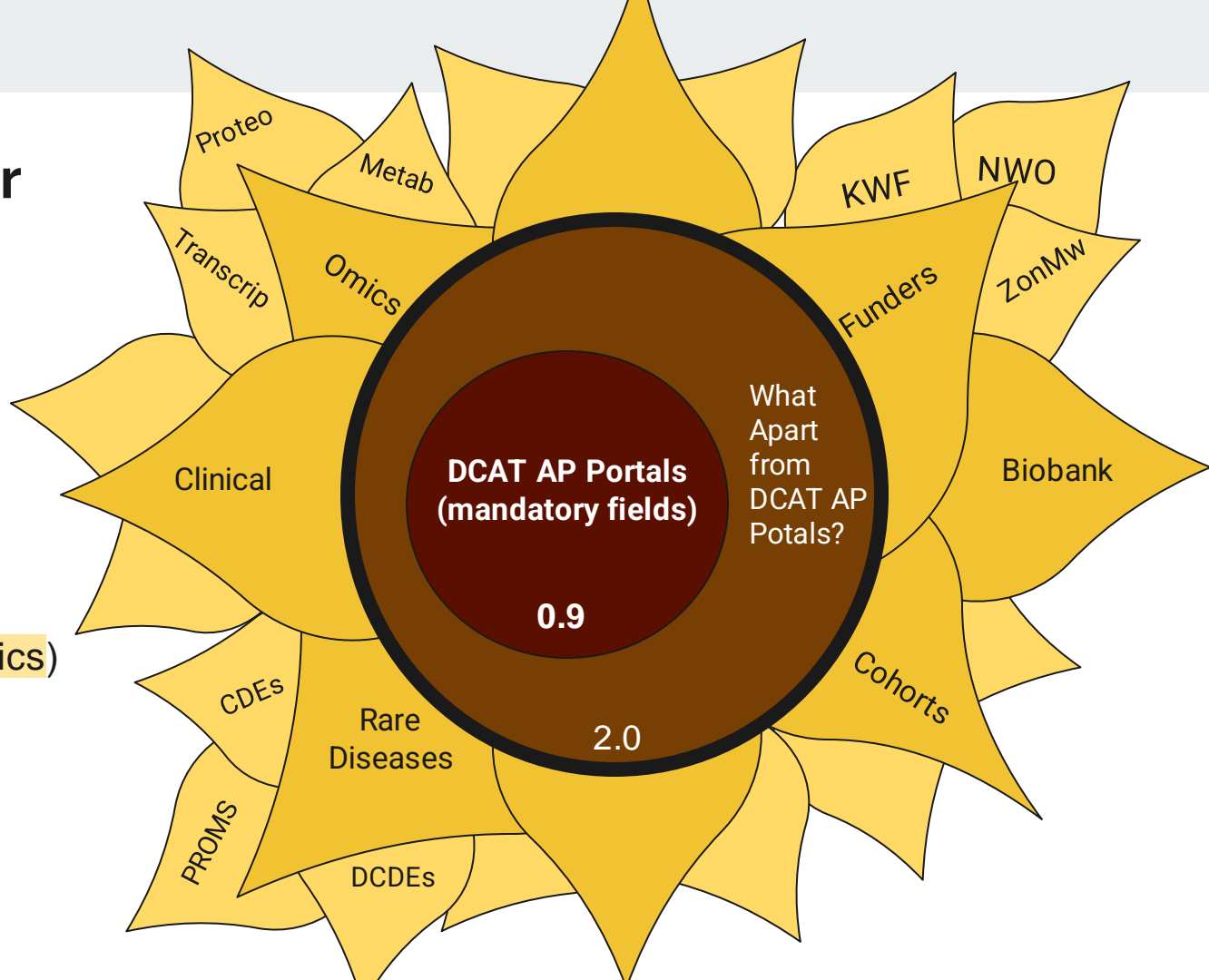
**Extended** Health-RI

## Leaves (Petals?):

-Domain: *Your Omics*

-sub-domains (transcriptomics)

<https://github.com/Health-RI/health-ri-metadata/>





# Acknowledgements

- Health-ri Hub and Nodes experts
  - Julia
  - Kees
  - Luiz
  - Mijke
  - Marianne
  - Jeroen
  - Rita
  - Sander