Data codification, standardisation, and disease identification

Ronald Cornet – Amsterdam UMC
Professor in Medical Informatics

Co-funded by the European Union

10.5281/zenodo.14927480

ERDERA has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement N°101156595.
Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or any other granting authority, who cannot be held responsible for them.

About me

Full Professor of Medical Informatics – Reusable Health Data at Amsterdam UMC

Principal Educator on FAIR Data

Principal Investigator on Reusable Health Data

Data codification, standardisation, and disease identification

Debunking buzzwords

1. Data standardization – Representing data in a standardized way
2. Data codification – Using codes for standardization
3. Disease identification – Using codes for standardized representation of diseases

Data standardization

Real-world example (partially synthetic data)

- **Free-text description**

I am a full professor from Amsterdam UMC, the Netherlands, whose research focuses on reusable health data, notably in the domain of rare diseases, involved in projects on RDs at large, rare hematologic diseases, and neuromuscular diseases

- **Somewhat “structured”**

	A	B	C	D	E
1	**Participant**	**Profile**	**Country**	**Organisation**	**Disease**
2	1	Senior researcher	Netherlands	Amsterdam UMC	All RDs, Hematologic diseases, Neuromuscular diseases
3	2	Patient advocate	Belgium	Endo-ERN	Rare bone diseases

Structured does not equal **Standardized**

- **Standards** need to be applied to be able to scale up
- E.g.,
  - How do we identify participants? E.g., incremental number? Social security number?
  - What is **relevant information** about participants?
  - What does each of these **data elements** mean?
  - Which are **allowed values** for each of these data elements?

	A	B	C	D	E
1	**Participant**	**Profile**	**Country**	**Organisation**	**Disease**
2	1	Senior researcher	Netherlands	Amsterdam UMC	All RDs, Hematologic diseases, Neuromuscular diseases
3	2	Patient advocate	Belgium	Endo-ERN	Rare bone diseases

Schema / Information model / Data dictionary

Specification of data **structure**

- Example
  Set of common data elements for Rare Diseases Registration
  Specific for rare disease registries

EUROPEAN PLATFORM ON RARE DISEASE REGISTRATION (EU RD Platform)

SET OF COMMON DATA ELEMENTS FOR RARE DISEASES REGISTRATION

GROUP	ELEMENT N°	ELEMENT NAME	ELEMENT DESCRIPTION	CODING	COMMENT
1. Pseudonym	1.1.	Pseudonym	Patient's pseudonym	• String	https://eu-rd-platform.jrc.ec.europa.eu/spider_
2. Personal information	2.1.	Date of birth	Patient's date of birth	• Date (dd/mm/yyyy)	
	2.2.	Sex	Patient's sex at birth	• Female • Male • Undetermined • Foetus (Unknown)	
3. Patient Status	3.1.	Patient's status	Patient alive or dead	• Alive • Dead • Lost in follow-up • Opted-out	If dead then answer question 3.2
	3.2.	Date of death	Patient's date of death	• Date (dd/mm/yyyy)	
4. Care pathway	4.1.	First contact with specialised centre	Date of first contact with specialised centre	• Date (dd/mm/yyyy)	
5. Disease history	5.1.	Age at onset	Age at which symptoms/signs first appeared	• Antenatal • At birth • Date (dd/mm/yyyy) • Undetermined	
	5.2.	Age at diagnosis	Age at which diagnosis was made	• Antenatal • At birth • Date (dd/mm/yyyy) • Undetermined	
6. Diagnosis	6.1.	Diagnosis of the rare disease	Diagnosis retained by the specialised centre	Orpha code (strongly recommended – see link) / Alpha code/ ICD-9 code/ ICD-9-CM code / ICD-10 code	http://www.orphadata.org/cgi-bin/inc/product1.inc.php
	6.2.	Genetic diagnosis	Genetic diagnosis retained by the specialised centre	International classification of mutations (HGVS) (strongly recommended – see link) / HGNC / OMIM code	http://www.hgvs.org
	6.3	Undiagnosed case	How the undiagnosed case is defined	• Phenotype (HPO) • Genotype (HGVS)	
7. Research	7.1.	Agreement to be contacted for research purposes	Patient's permission exists for being contacted for research purposes	• YES • NO	
	7.2.	Consent to the reuse of data	Patient's consent exists for his/her data to be reused for other research purposes	• YES • NO	
	7.3.	Biological sample	Patient's biological sample available for research	• YES • NO	If YES answer question 7.4
	7.4.	Link to a biobank	Biological sample stored in a biobank	• YES (if appropriate use link) • NO	https://directory.bbmri-eric.eu
8. Disability	8.1.	Classification of functioning/disability	Patient's disability profile according to International Classification of Functioning and Disability (ICF)	• Disability profile / Score	http://www.who.int/classifications/icf/whodasii/en/

Data codification

For **structure** AND **content**

- Coding helps to uniquely identify data elements and values
- Independent of languages or synonyms
- E.g., osallistujan tunniste, معرف المشارك, participant identifier
- Code:
  - Code system: LOINC ; Code: 82787-3
  - https://loinc.org/82787-3/
  - Shorthand: **LOINC:82787-3**

LOINC CODE **82787-3**

LONG COMMON NAME **Participant identifier**

Fully-Specified Name

Component	Participant identifier
Property	ID
Time	Pt
System	^Patient
Scale	Nom
Method	

Additional Names

Long Common Name	Participant identifier
Short Name	Participant id

Source: https://loinc.org/82787-3/

	A	B	C	D	E
1	**Participant**	**Profile**	**Country**	**Organisation**	**Disease**
	https://loinc.org/82787-3	**http://www.w3.org/2006/vcard/ns#role**		**https://loinc.org/66477-1/**	**https://schema.org/affiliation**	**http://xmlns.com/foaf/spec/#term_topic_interest**
	Participant	**Role**		**Country of current residence**	**Affiliation**	**Topic of Interest**
	1	http://purl.obolibrary.org/obo/NCIT_C19495 | Senior researcher		ISO 3166-2:NL | Netherlands	https://ror.org/05grdyy37 | Amsterdam UMC	*All RDs, Hematologic diseases, Neuromuscular diseases*
	2	http://purl.obolibrary.org/obo/NCIT_C93178 | Patient advocate		ISO 3166-2:BE | Belgium	https://ror.org/05s4nk876 | Endo-ERN	*Rare bone diseases*

Using codes

Not one code system to rule them all

- Use of a multitude of coding systems
  - LOINC – Observations
  - VCARD – for describing People and Organizations
  - schema.org – for structured data on the Internet
  - FOAF – Friend of a Friend
  - NCI Thesaurus – Reference terminology for many NCI and other systems
  - ISO 3166-2 – Country codes
  - ROR – Research Organization Registry

https://loinc.org/82787-3	http://www.w3.org/2006/vcard/ns#role	https://loinc.org/66477-1/	https://schema.org/affiliation	http://xmlns.com/foaf/spec/#term_topic_interest
Participant	Role	Country of current residence	Affiliation	Topic of Interest
1	http://purl.obolibrary.org/obo/NCIT_C19495 | Senior researcher	ISO 3166-2:NL | Netherlands	https://ror.org/05grdyy37 | Amsterdam UMC	*All RDs, Hematologic diseases, Neuromuscular diseases*
2	http://purl.obolibrary.org/obo/NCIT_C93178 | Patient advocate	ISO 3166-2:BE | Belgium	https://ror.org/05s4nk876 | Endo-ERN	*Rare bone diseases*

Using codes

Not one code system to rule them all

- Use of a multitude of coding systems
- Selecting the right system requires training and tools
  - Ontology Lookup Service (OLS)
  - NCBO BioPortal
  - UMLS – Metathesaurus
  - Linked Open Vocabularies

https://loinc.org/82787-3	http://www.w3.org/2006/vcard/ns#role	https://loinc.org/66477-1/	https://schema.org/affiliation	http://xmlns.com/foaf/spec/#term_topic_interest
Participant	Role	Country of current residence	Affiliation	Topic of Interest
1	http://purl.obolibrary.org/obo/NCIT_C19495 | Senior researcher	ISO 3166-2:NL | Netherlands	https://ror.org/05grdyy37 | Amsterdam UMC	*All RDs, Hematologic diseases, Neuromuscular diseases*
2	http://purl.obolibrary.org/obo/NCIT_C93178 | Patient advocate	ISO 3166-2:BE | Belgium	https://ror.org/05s4nk876 | Endo-ERN	*Rare bone diseases*

Using codes

Not one code system to rule them all

- Use of a multitude of coding systems
- Selecting the right system requires training and tools
- Coding systems offer:
  - Unique **identification**
  - Synonyms and **descriptions** in multiple languages
  - (Means for) Mappings among coding systems, e.g., ISO3166-2:NL = SNOMED:223672001
  - Any additional information
  - Web-based linking
  - Plus …

https://loinc.org/82787-3	http://www.w3.org/2006/vcard/ns#role	https://loinc.org/66477-1/	https://schema.org/affiliation	http://xmlns.com/foaf/spec/#term_topic_interest
Participant	Role	Country of current residence	Affiliation	Topic of Interest
1	http://purl.obolibrary.org/obo/NCIT_C19495 | Senior researcher	ISO 3166-2:NL | Netherlands	https://ror.org/05grdyy37 | Amsterdam UMC	*All RDs, Hematologic diseases, Neuromuscular diseases*
2	http://purl.obolibrary.org/obo/NCIT_C93178 | Patient advocate	ISO 3166-2:BE | Belgium	https://ror.org/05s4nk876 | Endo-ERN	*Rare bone diseases*

Disease identification

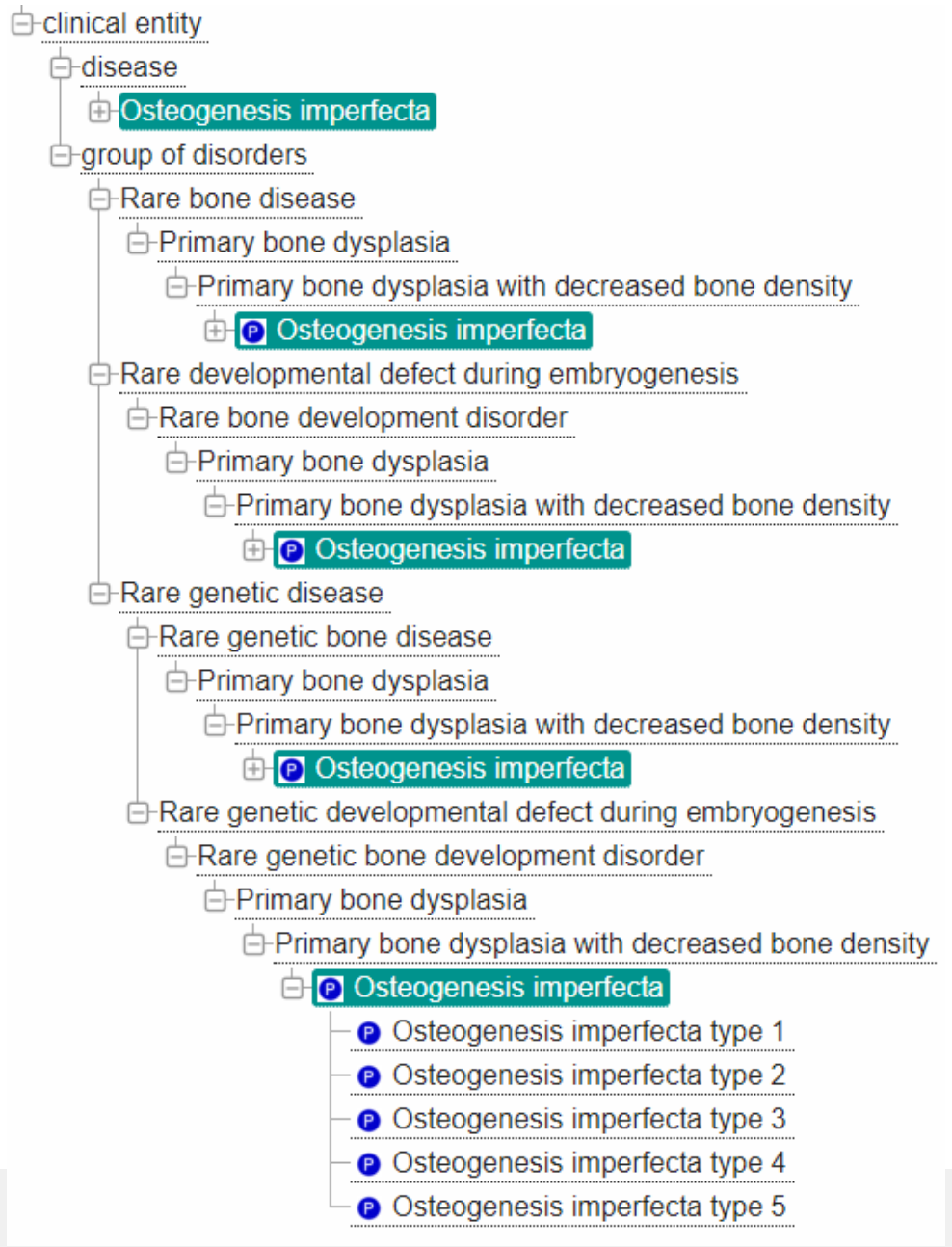
Store and retrieve disease data at arbitrary level of detail

To **record** data at varying detail

- OI
- OI type 5

To **select** data at varying detail

- All patients with a type of OI
- All patients with a type of Rare bone disease



Disease identificat

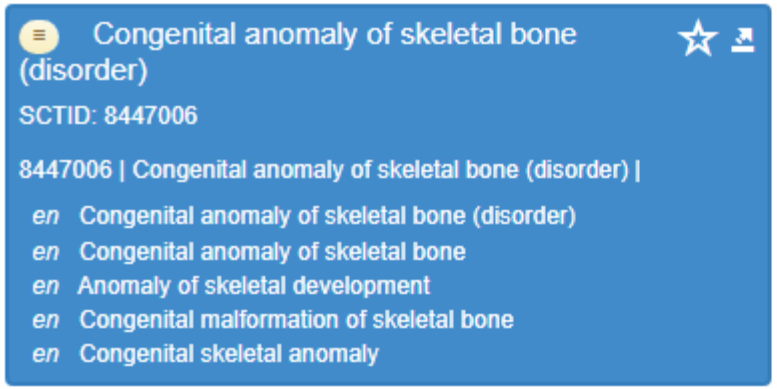
Store and retrieve disease data

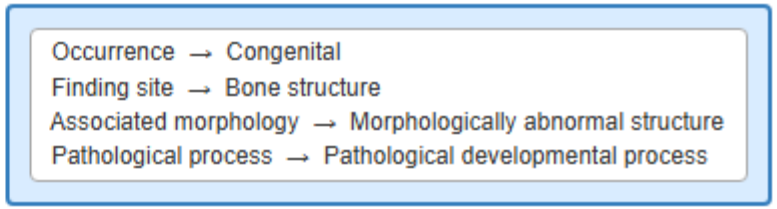
To **record** data at varying detail

- OI
- OI type 5

To **select** data at varying detail

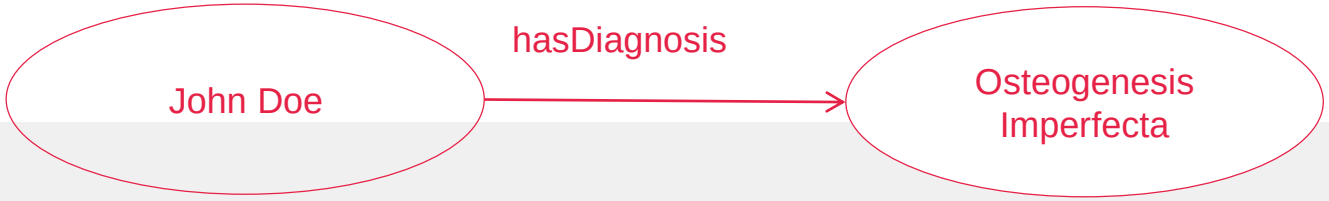
- All patients with a type of OI
- All patients with a type of Rare bone





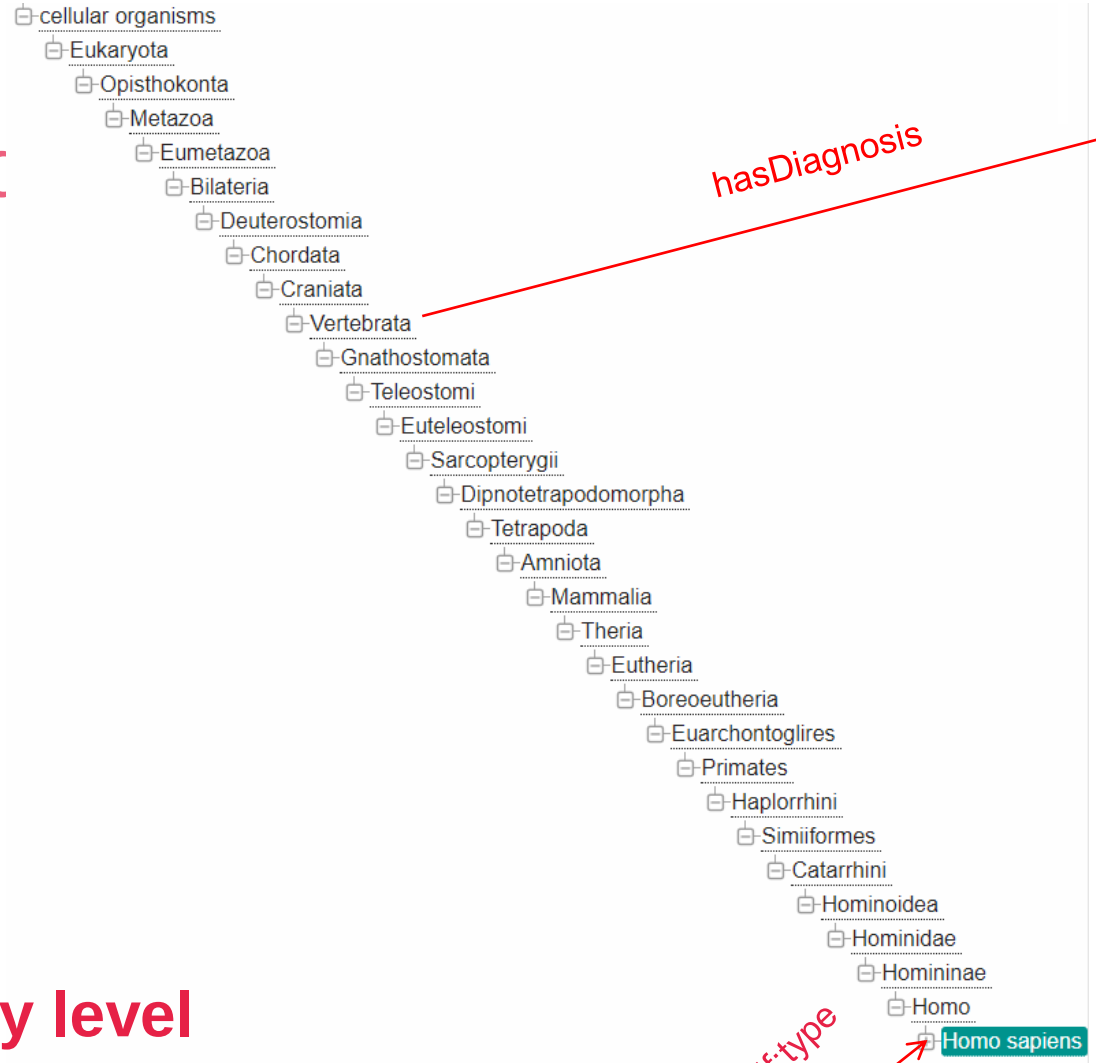
Data and **ontologies**

data level

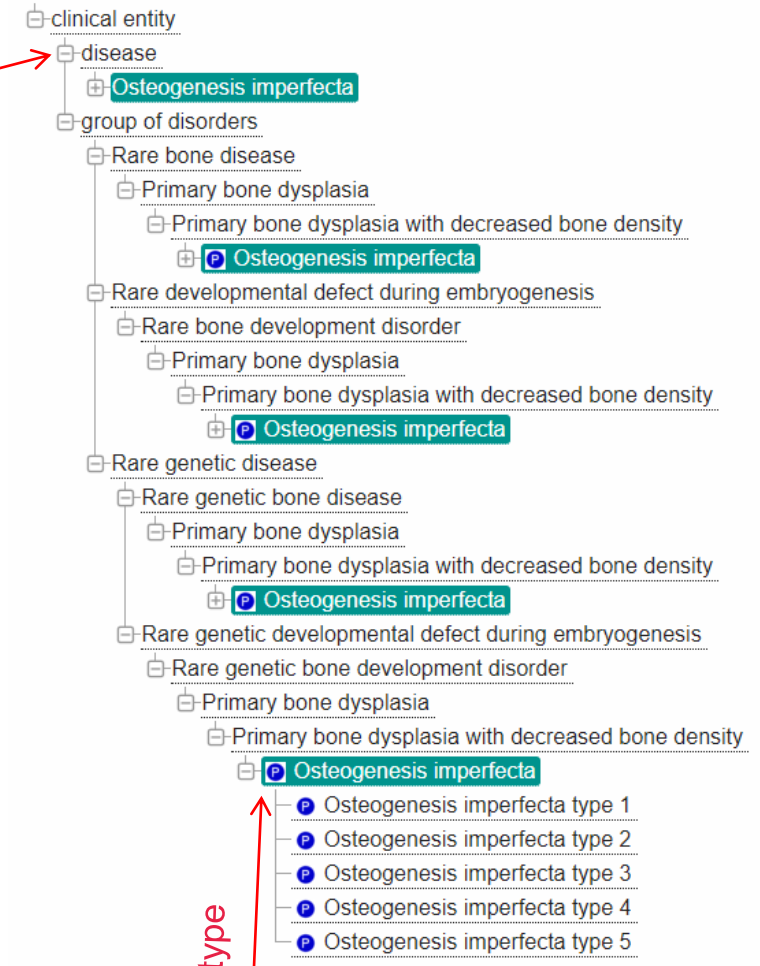


Instances

Data anc



hasDiagnosis



Classes

ontology level

rdf:type

rdf:type

data level

hasDiagnosis

John Doe

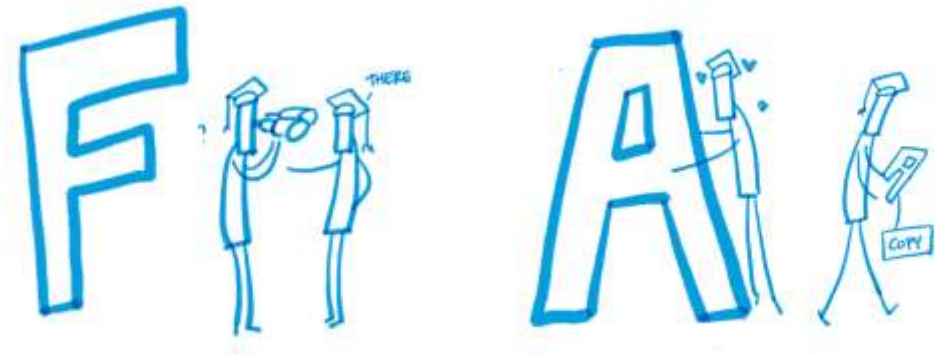
Osteogenesis Imperfecta

Instances

European Rare Diseases Research Alliance

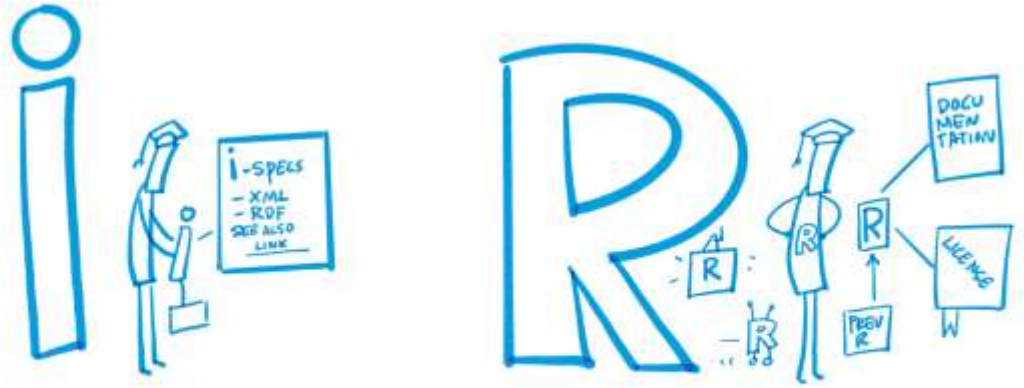
Coding for data AND metadata: Make data FAIR

Metadata and data should be easy to find for both humans and computers.



The user needs to know how data can be accessed, possibly including authentication and authorisation.

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.



To optimise the reuse of data, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

Image source:
How to be FAIR with your data. A teaching and training handbook for higher education institutions

Summary

1. Data standardization – Representing data in a standardized way
   Using schemas to relevant data elements to be structured
2. Data codification – Using codes for standardization
   Adding standard representation to structure
   Using web-based codes makes (meta)data linkable, adding synonyms, descriptions and mapping information
3. Disease identification – Using codes for standardized representation of diseases
   Leverages hierarchical structure for storage and retrieval

   Makes data FAIR

Contact info

Prof. Ronald Cornet, PhD
Full Professor ; Principal Investigator ; Principal Educator ;
Education director Medical informatics; Program Leader APH Digital Health
Department of Medical Informatics - Reusable Health Data
Amsterdam Public Health Research Institute
Amsterdam UMC
Amsterdam Public Health Research Institute
Location AMC | Meibergdreef 15, 1105 AZ Amsterdam
+31 20-566 52 05 | r.cornet@amsterdamumc.nl
kik.amsterdamumc.org/home/rcornet | www.amsterdamumc.nl

10.5281/zenodo.14927480