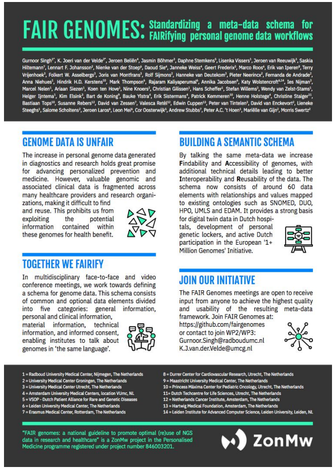
FAIR Genomes F2F @ Utrecht Feb. 6th WP2 & WP3 (Gurnoor, Joeri, et al.)

WP2/WP3 have been busy

- → FAIR genomes WP2/WP3 zoom video meetings:
  - ◆ 2 oct 2019: “Personal and Clinical”
  - ◆ 8 nov 2019: “Materials”
  - ◆ 22 jan 2020: “Technical”
- → V1.1 of meta model, mostly ontologized
- → 2 posters @ HealthRI 2020
- → A few open issues to be discussed
- → GitHub page updated



FAIR GENOMES: Standardizing a meta-data schema for FAIRifying personal genome data workflows

Gurnoor Singh[1*], K. Joeri van der Velde[2*], Jeroen Beliën[4], Jasmin Böhmer[3], Daphne Stemkens[5], Lisenka Vissers[1], Jeroen van Reeuwijk[1], Saskia Hiltemann[7], Lennart F. Johansson[2], Nienke van der Stoep[6], Daoud Sie[4], Janneke Weiss[4], Geert Frederix[3], Marco Roos[6], Erik van Iperen[8], Terry Vrijenhoek[3], Folkert W. Asselbergs[3], Joris van Montfrans[3], Rolf Sijmons[2], Hanneke van Deutekom[3], Pieter Neerincx[2], Fernanda de Andrade[2], Anna Niehues[1], Hindrik H.D. Kerstens[10], Mark Thompson[6], Rajaram Kaliyaperumal[6], Annika Jacobsen[6], Katy Wolstencroft[6,14], Ies Nijman[3], Marcel Nelen[1], Ariaan Siezen[1], Koen ten Hove[1], Nine Knoers[2], Christian Gilissen[1], Hans Scheffer[1], Stefan Willems[3], Wendy van Zelst-Stams[1], Helger Ijntema[1], Kim Elsink[3], Bart de Koning[9], Bauke Ylstra[4], Erik Sistermans[4], Patrick Kemmeren[10], Henne Holstege[4], Christine Staiger[11], Bastiaan Tops[10], Susanne Rebers[12], David van Zessen[7], Valesca Retèl[12], Edwin Cuppen[13], Peter van Tintelen[3], David van Enckevort[2], Lieneke Steeghs[1], Salome Scholtens[2], Jeroen Laros[6], Leon Mei[6], Cor Oosterwijk[5], Andrew Stubbs[7], Peter A.C. 't Hoen[1], Mariëlle van Gijn[2], Morris Swertz[2]

GENOME DATA IS UNFAIR

The increase in personal genome data generated in diagnostics and research holds great promise for advancing personalized prevention and medicine. However, valuable genomic and associated clinical data is fragmented across many healthcare providers and research organizations, making it difficult to find and reuse. This prohibits us from exploiting the potential information contained within these genomes for health benefit.

TOGETHER WE FAIRIFY

In multidisciplinary face-to-face and video conference meetings, we work towards defining a schema for genome data. This schema consists of common and optional data elements divided into five categories: general information, personal and clinical information, material information, technical information, and informed consent, enabling institutes to talk about genomes in ‘the same language’.

BUILDING A SEMANTIC SCHEMA

By talking the same meta-data we increase **F**indability and **A**ccessibility of genomes, with additional technical details leading to better **I**nteroperability and **R**eusability of the data. The schema now consists of around 60 data elements with relationships and values mapped to existing ontologies such as SNOMED, DUO, HPO, UMLS and EDAM. It provides a strong basis for digital twin data in Dutch hospitals, development of personal genetic lockers, and active Dutch participation in the European '1+ Million Genomes' Initiative.

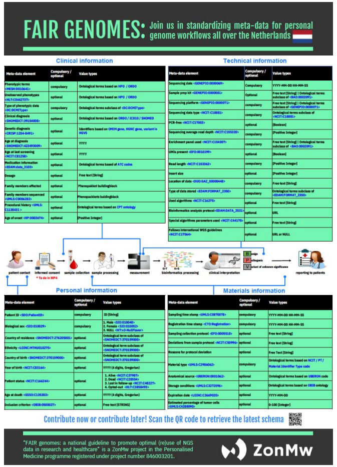
JOIN OUR INITIATIVE

The FAIR Genomes meetings are open to receive input from anyone to achieve the highest quality and usability of the resulting meta-data framework. Join FAIR Genomes at:
https://github.com/fairgenomes
or contact to join WP2/WP3:
Gurnoor.Singh@radboudumc.nl
K.J.van.der.Velde@umcg.nl

1 = Radboud University Medical Center, Nijmegen, The Netherlands
2 = University Medical Center Groningen, The Netherlands
3 = University Medical Center Utrecht, The Netherlands
4 = Amsterdam University Medical Centers, location VUmc, NL
5 = VSOP - Dutch Patient Alliance for Rare and Genetic Diseases
6 = Leiden University Medical Center, The Netherlands
7 = Erasmus Medical Center, Rotterdam, The Netherlands
8 = Durrer Center for Cardiovascular Research, Utrecht, The Netherlands
9 = Maastricht University Medical Center, The Netherlands
10 = Princess Máxima Center for Pediatric Oncology, Utrecht, The Netherlands
11= Dutch Techcentre for Life Sciences, Utrecht, The Netherlands
12 = Netherlands Cancer Institute, Amsterdam, The Netherlands
13 = Hartwig Medical Foundation, Amsterdam, The Netherlands
14 = Leiden Institute for Advanced Computer Science, Leiden University, Leiden, NL

“FAIR genomes: a national guideline to promote optimal (re)use of NGS data in research and healthcare” is a ZonMw project in the Personalised Medicine programme registered under project number 846003201.

ZonMw



FAIR GENOMES: Join us in standardizing meta-data for personal genome workflows all over the Netherlands

Clinical information

Meta-data element	Compulsory / optional	Value types
Phenotypic terms <MESH:D010641>	compulsory	Ontological terms based on HPO / ORDO
Unobserved phenotypes <HL7:C0442737>	optional	Ontological terms based on HPO / ORDO
Type of phenotypic data <DC:DCMIType>	compulsory	Ontological terms subclass of <DC:DCMIType>
Clinical diagnosis <SNOMEDCT:39154008>	optional	Ontological terms based on ORDO / ICD10 / SNOMED
Genetic diagnosis <CRISP:1254-8491>	optional	Identifiers based on OMIM gene, HGNC gene, variant in HGVS
Age at diagnosis <SNOMEDCT:423493009>	optional	YYYY
Age at last screening <NCIT:C81258>	optional	YYYY
Medication information <EDAM:data_3103>	optional	Ontological terms based of ATC codes
Dosage	optional	Free text [String]
Family members affected	optional	Phenopakket buildingblock
Family members sequenced <UMLS:C0086282>	optional	Phenopackkets buildingblock
Procedural history <UMLS:C1138431 >	optional	Ontological terms based on CPT ontology
Age of onset <HP:0003674>	optional	[Positive Integer]

Technical information

Meta-data element	Compulsory / optional	Value types
Sequencing date <GENEPIO:0000069>	Compulsory	YYYY-MM-DD HH-MM-SS
Sample prep kit <GENEPIO:0000081>	Optional	Free text [String] | Ontological terms subclass of <BAO:0002091>
Sequencing platform <GENEPIO:0000071>	compulsory	Free text [String] | Ontological terms subclass of <GENEPIO:0000071>
Sequencing data type <NCIT:C18881>	compulsory	Ontological terms subclass of <NCIT:C18881>
PCR-free <NCIT:C17003>	optional	[Boolean]
Sequencing average read depth <NCIT:C155320>	compulsory	[Positive Integer]
Enrichment panel used <NCIT:C154307>	compulsory	Free text [String] | Ontological terms subclass of <BAO:0002091>
UMIs present <EFO:0010199>	optional	[Boolean]
Read length <NCIT:C153362>	compulsory	[Positive Integer]
insert size	optional	[Positive Integer]
Location of data <DUO:GAZ_0000448>	compulsory	Free text [String]
Type of data stored <EDAM:FORMAT_2350>	compulsory	Ontological terms subclass of <EDAM:FORMAT_2350>
Used algorithms <NCIT:C16275>	optional	Free text [String]
Bioinformatics analysis protocol<EDAM:DATA_2531>	optional	URL
Special algorithms parameters used <NCIT:C44175>	optional	Free text [String]
Follows international WGS guidelines <NCIT:C17564>	optional	URL or NULL

Personal information

Meta-data element	Compulsory / optional	Value types
Patient ID <SDO:PatientID>	compulsory	ID [String]
Biological sex <SIO:010029>	compulsory	1. Male <SIO:010048> 2. Female <SIO:010052> 3. NULL <hl7:v3-NullFlavor>
Country of residence <SNOMEDCT:276205001>	optional	Ontological term subclass of <SNOMEDCT:370159000>
Ethnicity <LOINC:MTHU010275>	optional	Ontological term subclass of <SNOMEDCT:370159000>
Country of birth <SNOMEDCT:370159000>	optional	Ontological term subclass of <SNOMEDCT:370159000>
Year of birth <NCIT:C83164>	optional	YYYY (4 digits, Gregorian)
Patient status <NCIT:C166244>	optional	1. Alive <NCIT:C37987> 2. Dead <NCIT:C28554> 3. Lost in follow-up <NCIT:C48227> 4. Opted-out <HL7:C3858493>
Age at death <GSSO:C135383>	optional	YYYY (4 digits, Gregorian)
Inclusion criterion <OBIB:0500027>	optional	Free text [STRING]

Materials information

Meta-data element	Compulsory / optional	Value types
Sampling time stamp <UMLS:C0870078>	compulsory	YYYY-MM-DD HH-MM-SS
Registration time stamp <CTO:Registration>	compulsory	YYYY-MM-DD HH-MM-SS
Sampling collection protocol <EFO:0005518>	optional	Free text [String]
Deviations from sample protocol <NCIT:C50996>	optional	Free text [String]
Reasons for protocol deviation	optional	Free Text [String]
Material type <UMLS:C2986062>	compulsory	Ontological terms based on NCIT / PT / Material Identifier Type code
Anatomical source <UBERON:0001062>	optional	Ontological terms based on UBERON code
Storage conditions <UMLS:C3272596>	optional	Ontological terms based on OBIB ontology
Expiration date <LOINC:C3669020>	optional	YYYY-MM-DD
Estimated percentage of tumor cells <UMLS:C4288090>	optional	0-100 [Integer]

Contribute now or contribute later! Scan the QR code to retrieve the latest schema

"FAIR genomes: a national guideline to promote optimal (re)use of NGS data in research and healthcare" is a ZonMw project in the Personalised Medicine programme registered under project number 846003201.

ZonMw

Link to working document

FAIR Genomes working document:

https://docs.google.com/spreadsheets/d/1rnLsmE62t15jCwJfx4mCL5USYSeiXNctCA0XPcgprds/edit#gid=914344861

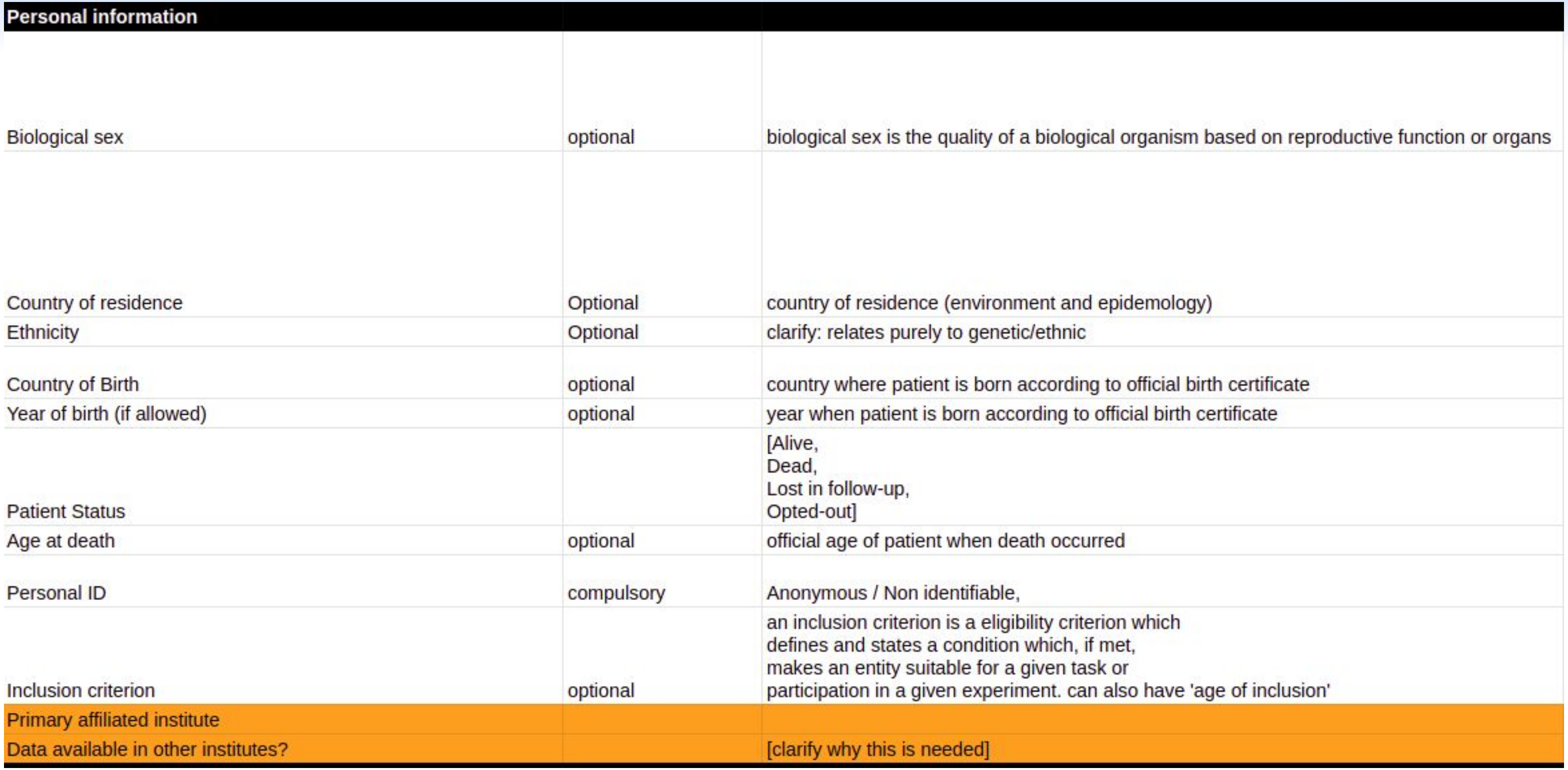
→ **Metadata_version 1.1**

“Meta-data attribute name”

Slide format

Description	*Description or example of this attribute*
Compulsory / Optional	*“Compulsory” or “Optional” ...*
Ontology Term (Meta-data)	*Link to ontology and term*
Ontology Term (Data)	*Value type, could be ontology (sub)term or data type like INTEGER, DATE, STRING etc*
Defined in other projects	*Is this attribute present in other projects and if so, how is it represented?*

Personal information



Personal information		
Biological sex	optional	biological sex is the quality of a biological organism based on reproductive function or organs
Country of residence	Optional	country of residence (environment and epidemology)
Ethnicity	Optional	clarify: relates purely to genetic/ethnic
Country of Birth	optional	country where patient is born according to official birth certificate
Year of birth (if allowed)	optional	year when patient is born according to official birth certificate
Patient Status		[Alive, Dead, Lost in follow-up, Opted-out]
Age at death	optional	official age of patient when death occurred
Personal ID	compulsory	Anonymous / Non identifiable,
Inclusion criterion	optional	an inclusion criterion is a eligibility criterion which defines and states a condition which, if met, makes an entity suitable for a given task or participation in a given experiment. can also have 'age of inclusion'
Primary affiliated institute		
Data available in other institutes?		[clarify why this is needed]

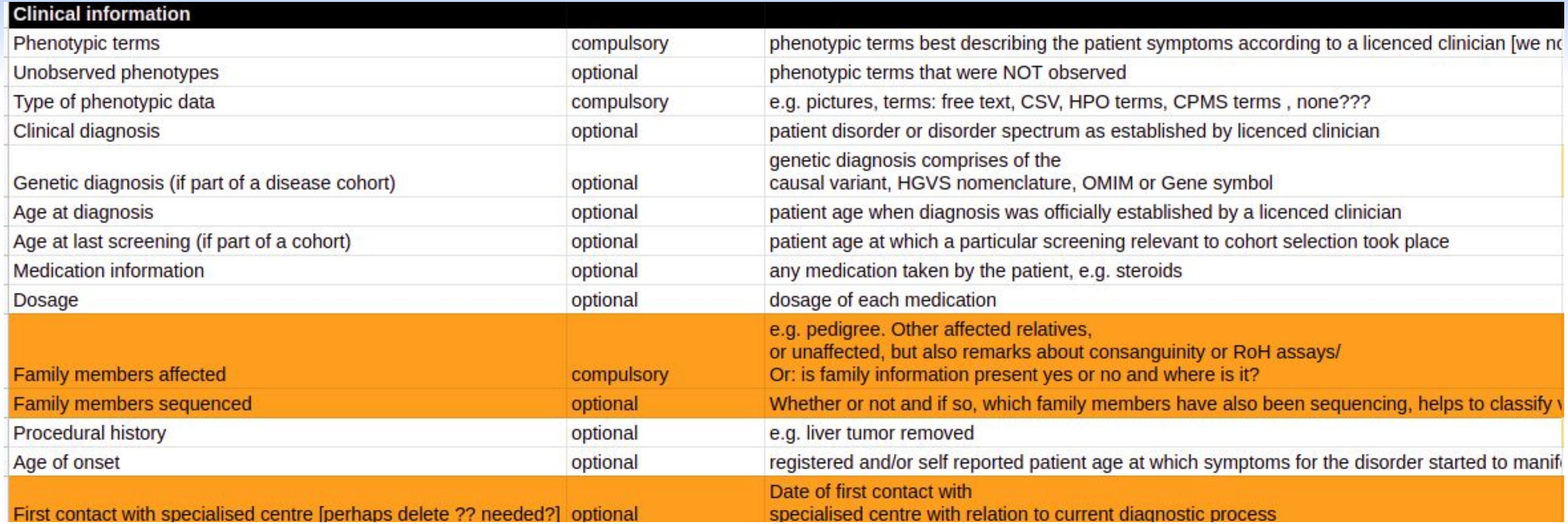
Personal information

Feedback @ HealthRI 2020

- Biological sex types: Currently we just have male, female and null flavor. There should be more like, “raised as male/female”, undetermined, unknown. Not sure if null-flavor can cover both undetermined, unknown
- Biological sex could be renamed as Gender
- Patient can be renamed as Individual
- Ethnicity can be renamed as Population

No issues (?)

Clinical information

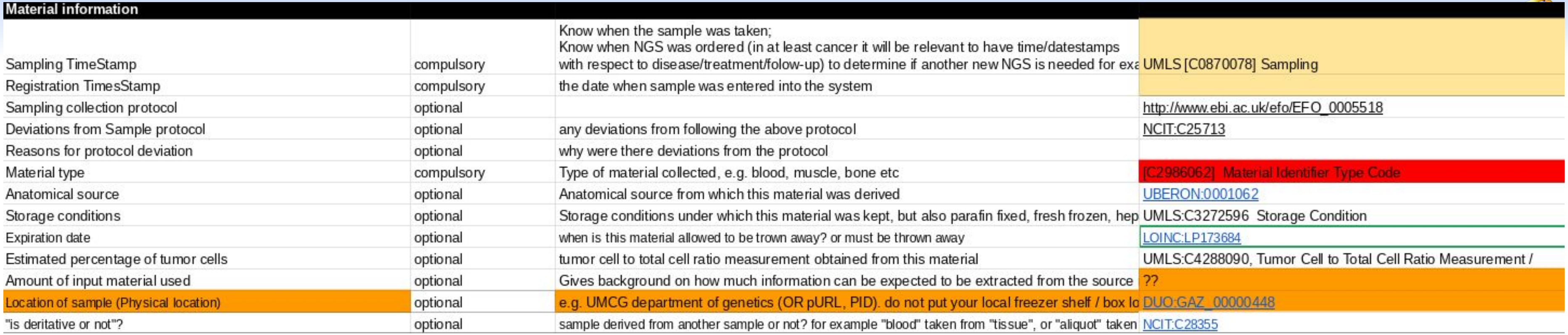


Clinical information		
Phenotypic terms	compulsory	phenotypic terms best describing the patient symptoms according to a licenced clinician [we no
Unobserved phenotypes	optional	phenotypic terms that were NOT observed
Type of phenotypic data	compulsory	e.g. pictures, terms: free text, CSV, HPO terms, CPMS terms , none???
Clinical diagnosis	optional	patient disorder or disorder spectrum as established by licenced clinician
Genetic diagnosis (if part of a disease cohort)	optional	genetic diagnosis comprises of the causal variant, HGVS nomenclature, OMIM or Gene symbol
Age at diagnosis	optional	patient age when diagnosis was officially established by a licenced clinician
Age at last screening (if part of a cohort)	optional	patient age at which a particular screening relevant to cohort selection took place
Medication information	optional	any medication taken by the patient, e.g. steroids
Dosage	optional	dosage of each medication
Family members affected	compulsory	e.g. pedigree. Other affected relatives, or unaffected, but also remarks about consanguinity or RoH assays/ Or: is family information present yes or no and where is it?
Family members sequenced	optional	Whether or not and if so, which family members have also been sequencing, helps to classify v
Procedural history	optional	e.g. liver tumor removed
Age of onset	optional	registered and/or self reported patient age at which symptoms for the disorder started to manif
First contact with specialised centre [perhaps delete ?? needed?]	optional	Date of first contact with specialised centre with relation to current diagnostic process

Issues:

- *Family information*
- *First contact*

Material information

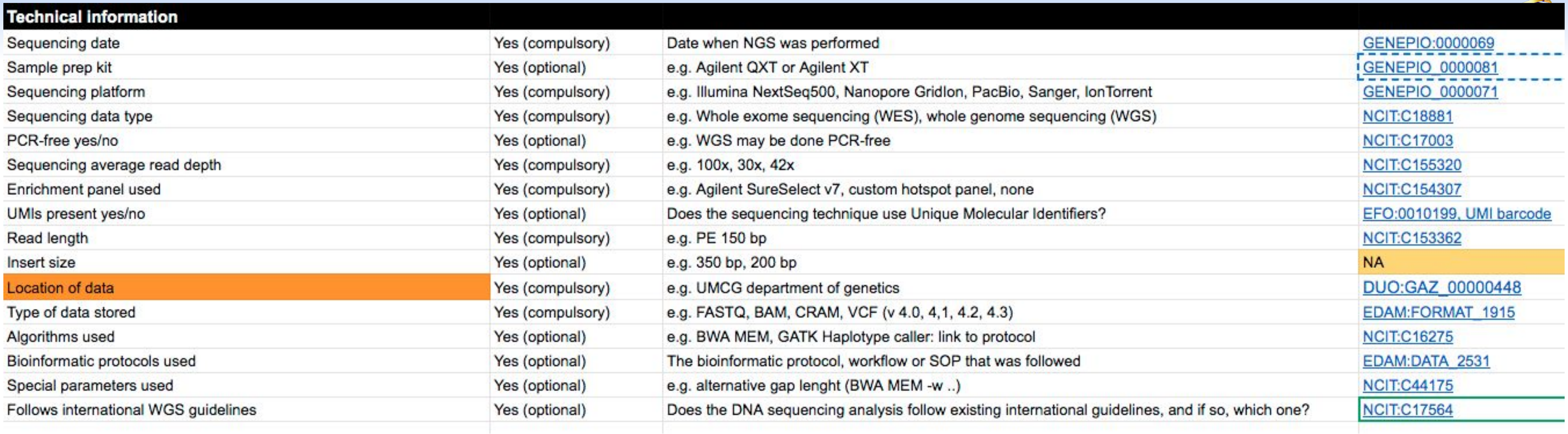


Material information			
Sampling TimeStamp	compulsory	Know when the sample was taken; Know when NGS was ordered (in at least cancer it will be relevant to have time/datestamps with respect to disease/treatment/folow-up) to determine if another new NGS is needed for exa	UMLS [C0870078] Sampling
Registration TimesStamp	compulsory	the date when sample was entered into the system	
Sampling collection protocol	optional		http://www.ebi.ac.uk/efo/EFO_0005518
Deviations from Sample protocol	optional	any deviations from following the above protocol	NCIT:C25713
Reasons for protocol deviation	optional	why were there deviations from the protocol	
Material type	compulsory	Type of material collected, e.g. blood, muscle, bone etc	[C2986062] Material Identifier Type Code
Anatomical source	optional	Anatomical source from which this material was derived	UBERON:0001062
Storage conditions	optional	Storage conditions under which this material was kept, but also parafin fixed, fresh frozen, hep	UMLS:C3272596 Storage Condition
Expiration date	optional	when is this material allowed to be trown away? or must be thrown away	LOINC:LP173684
Estimated percentage of tumor cells	optional	tumor cell to total cell ratio measurement obtained from this material	UMLS:C4288090, Tumor Cell to Total Cell Ratio Measurement /
Amount of input material used	optional	Gives background on how much information can be expected to be extracted from the source	??
Location of sample (Physical location)	optional	e.g. UMCG department of genetics (OR pURL, PID). do not put your local freezer shelf / box lo	DUO:GAZ_00000448
"is deritative or not"?	optional	sample derived from another sample or not? for example "blood" taken from "tissue", or "aliquot" taken	NCIT:C28355

Issues:

- ***Sample location** vs **data location** in technical*

Technical information



Technical information			
Sequencing date	Yes (compulsory)	Date when NGS was performed	GENEPIO:0000069
Sample prep kit	Yes (optional)	e.g. Agilent QXT or Agilent XT	GENEPIO_0000081
Sequencing platform	Yes (compulsory)	e.g. Illumina NextSeq500, Nanopore Gridlon, PacBio, Sanger, IonTorrent	GENEPIO_0000071
Sequencing data type	Yes (compulsory)	e.g. Whole exome sequencing (WES), whole genome sequencing (WGS)	NCIT:C18881
PCR-free yes/no	Yes (optional)	e.g. WGS may be done PCR-free	NCIT:C17003
Sequencing average read depth	Yes (compulsory)	e.g. 100x, 30x, 42x	NCIT:C155320
Enrichment panel used	Yes (compulsory)	e.g. Agilent SureSelect v7, custom hotspot panel, none	NCIT:C154307
UMIs present yes/no	Yes (optional)	Does the sequencing technique use Unique Molecular Identifiers?	EFO:0010199, UMI barcode
Read length	Yes (compulsory)	e.g. PE 150 bp	NCIT:C153362
Insert size	Yes (optional)	e.g. 350 bp, 200 bp	NA
Location of data	Yes (compulsory)	e.g. UMCG department of genetics	DUO:GAZ_00000448
Type of data stored	Yes (compulsory)	e.g. FASTQ, BAM, CRAM, VCF (v 4.0, 4,1, 4.2, 4.3)	EDAM:FORMAT_1915
Algorithms used	Yes (optional)	e.g. BWA MEM, GATK Haplotype caller: link to protocol	NCIT:C16275
Bioinformatic protocols used	Yes (optional)	The bioinformatic protocol, workflow or SOP that was followed	EDAM:DATA_2531
Special parameters used	Yes (optional)	e.g. alternative gap lenght (BWA MEM -w ..)	NCIT:C44175
Follows international WGS guidelines	Yes (optional)	Does the DNA sequencing analysis follow existing international guidelines, and if so, which one?	NCIT:C17564

Issues:

- *Insert size ontology /* *Insert size should be* ***Observed insert size***
- ***Sample location*** *vs* ***data location*** *in technical*

Special mention: nullflavors!

For **any empty** value, please elaborate

https://www.hl7.org/fhir/v3/NullFlavor/cs.html

some examples:

NAV	temporarily unavailable	Information is not available at this time but it is expected that it will be available later.
ASKU	asked but unknown	Information was sought but not found (e.g., patient was asked but didn't know)
NASK	not asked	This information has not been sought (e.g., patient was not asked)
NA	not applicable	Known to have no proper value (e.g., last menstrual period for a male).
MSK	masked	There is information on this item available but it has not been provided by the sender due to security, privacy or other reasons. There may be an alternate mechanism for gaining access to this information.
TRC	trace	The content is greater than zero, but too small to be quantified.

In progress: connecting to…

→ 1+ MG
→ SolveRD
→ EJP-RD
→ X-omics
→ GA4GH
→ CINECA
→ Phenopackets
→ Biosamples

→ RIVM WG?
→ Illumina WG?
→ BioSchemas (!)
→ WP5: ‘FAIR variants’ ?

see

https://raw.githubusercontent.com/LUMC-BioSemantics/ERN-common-data-elements/master/images/complete_data_model_v2.0.png

https://github.com/Xomics/GenomicsMeta-data

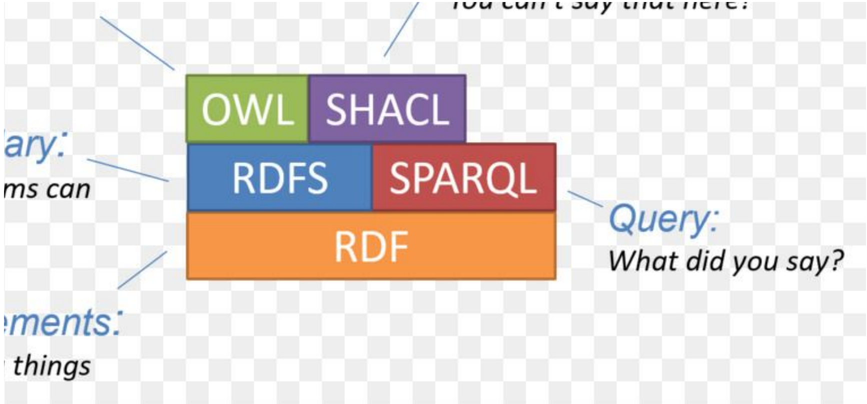
https://phenopackets-schema.readthedocs.io/en/latest/building-blocks.html

https://submission.ebi.ac.uk/api/docs/guide_getting_started.html

How to publish guidelines?

How to publish FAIR genomes guidelines for standardizing meta-data?

- Meta-data release
  - Github [https://github.com/fairgenomes/information] and Github wiki
- Data-model
  - Reviewed by stakeholders
    - (clinicians, diagnostics, pathologist)
  - Technical review
  - Community review



Future work

- Discuss further in FAIR genomes & with other experts
  - Finalize attributes, descriptions, values, link to other projects / standards
  - Integrate with WP4

- Demonstrator projects
  - Try out the FAIRgenomes schema on real genomics datasets

- From meta-data to a ‘real’ data model ?
  - Semantic data model (RDF or JSON-LD)
  - Rules based validator (RDF Shacl or JSON-LD markups)

- Apply to daily practice, use in care & research
  - (VKGL, VKGN, KMBP……)

- Publish to share our insights & FAIRify the rest of the world

All done!

Thanks