Deliverable

D4.7 All foundational standards selected and implemented across the project	
Version | Status	V1 | final
Work package	WP4
Lead beneficiary	EMBL-EBI (Thomas Keane)
Due date	30.06.2019 (M18)
Date of preparation	16.08.2019
Target Dissemination Level	Public
Author(s)	Dylan Spalding (EMBL-EBI), Thomas Keane (EMBL_EBI)
Reviewed by	Anthony Brookes (ULEIC), Morris Swertz (UMCG)
Approved by	Anthony Brookes (ULEIC)

The Solve-RD project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 779257.

Explanation according to GA Annex I:

Provide all foundational standards selected and implemented across the project.

Abstract:

Underpinning all activities at Solve-RD, from data submission, quality control, data dissemination to appropriate resources, discovery and finally distribution, a set of defined standards are required to ensure smooth and efficient data flow and interoperability between resources within Solve-RD and external resources. This deliverable focuses on defining and implementing the set of standards required to facilitate these processes, and here we describe how these standards have been established and implemented across Solve-RD.

Introduction:

Solve-RD will include many diverse data types including biosamples, patients, experimental methodology, consent, NGS measurement outcomes, and associated files. A standardised way of storing these data (e.g. file formats), describing these via metadata, representing the associated information, analysing the data, recording sample information, detailing the applicable attributes of the subject, and discovering or querying these data is of paramount importance for Solve-RD as a distributed and federated project (Table 1). Standards will also enable the data to be consistently and reproducibly analysed across the project, and provide the basis for improved data sharing and interoperability with external resources by ensuring Solve-RD data systems are FAIR (Findable, Accessible, Interoperable, Reuseable).

Table 1: Examples of the type of data and standardisation required for Solve-RD.

Data Type	Example Attributes
Sample data	Sample source, location, tissue type, data or collection
Subject Phenotype data	Disease and phenotype, including images
Subject Pedigree data	Familial relationships, consanguinity
Experimental metadata	Platform, library, assay type and conditions
NGS Data	File types (FASTQ, bam, and cram files)
Variation data	File types (VCF / BCF, gVCF), representation (HGVS)
Analysis data	Analysis pipelines (Alignment, QC, variant calling software)
Consent data	Data sharing conditions, incidental findings
Organisational metadata	Contacts, resources, data location
Identity information	Authorisation and Authentication

A key goal for WP4 of Solve-RD is to use existing community standards, and adapt where necessary to ensure applicability for Solve-RD use-cases. We surveyed the standards already in use by a diverse range of resources, such as the RD-Connect Sample Catalogue[1], the RD-Connect Genome-phenome Analysis Platform[2] (GPAP), the European Genome-phenome Archive[3] (EGA), BBMRI-ERIC[4], ELIXIR[5], the International Rare Disease in Research Consortium[6] (IRDiRC), and the Global Alliance for Genomics and Health[7] (GA4GH).

2 Data Standards for Solve-RD

Standards are continually evolving, as new technologies and use-cases emerge. The main goal of this deliverable is to document and collate the current standards that apply to Solve-RD. Standards are not singular and universally agreed. Instead, multiple standards often exist for the same or similar aspects of human data sharing. Choices therefore have to be made within Solve-RD over which standards to use, adapt or create for use by the project.

2.1 Intra Solve-RD Standards

To ensure standards initially chosen by the Solve-RD project would enable the greatest interoperability between Solve-RD partners and with other projects (including both rare disease and other health related projects), we initially compiled existing metadata standards, and mapped links between these standards (see Appendix 1 & 2). We started by mapping standards employed by existing resources within Solve-RD (e.g. EGA, RD-Connect Sample catalogue and GPAP, and MOLGENIS[8]) which provides the basis of the project's analytical sandbox. This process ensures that the data flow between the infrastructures (e.g. genome/phenome analysis platform - RD-Connect, EGA), through to the 'sandbox' used to analyse the data, can be semantically harmonised to enable joint research and clinical analysis applications. We worked with WP1 to ensure that the data collected from the European Reference Networks, (ERNs) (Figure 1), conforms to the standards we have defined and that these data are FAIR (Findable, Accessible, Interoperable, Reuseable) compliant. These standards were then used to help develop a Rare Disease Data Database (RD3) model (Figure 2) to be employed to track analysis within the Sandbox (using MOLGENIS), to enable a Federated Discovery Environment based on the Rare Disease Network for EXploring the UnSeen[9] software (RD-NEXUS) and a project metadata catalog (see D4.5 Metadata catalog operational, with initial content).

[1] https://samples.rd-connect.eu/
[2] https://platform.rd-connect.eu/
[3] https://ega-archive.org/
[4] http://www.bbmri-eric.eu/
[5] https://elixir-europe.org/
[6] http://www.irdirc.org/
[7] https://www.ga4gh.org
[8] https://www.molgenis.org/
[9] https://rd.discovery-nexus.org/

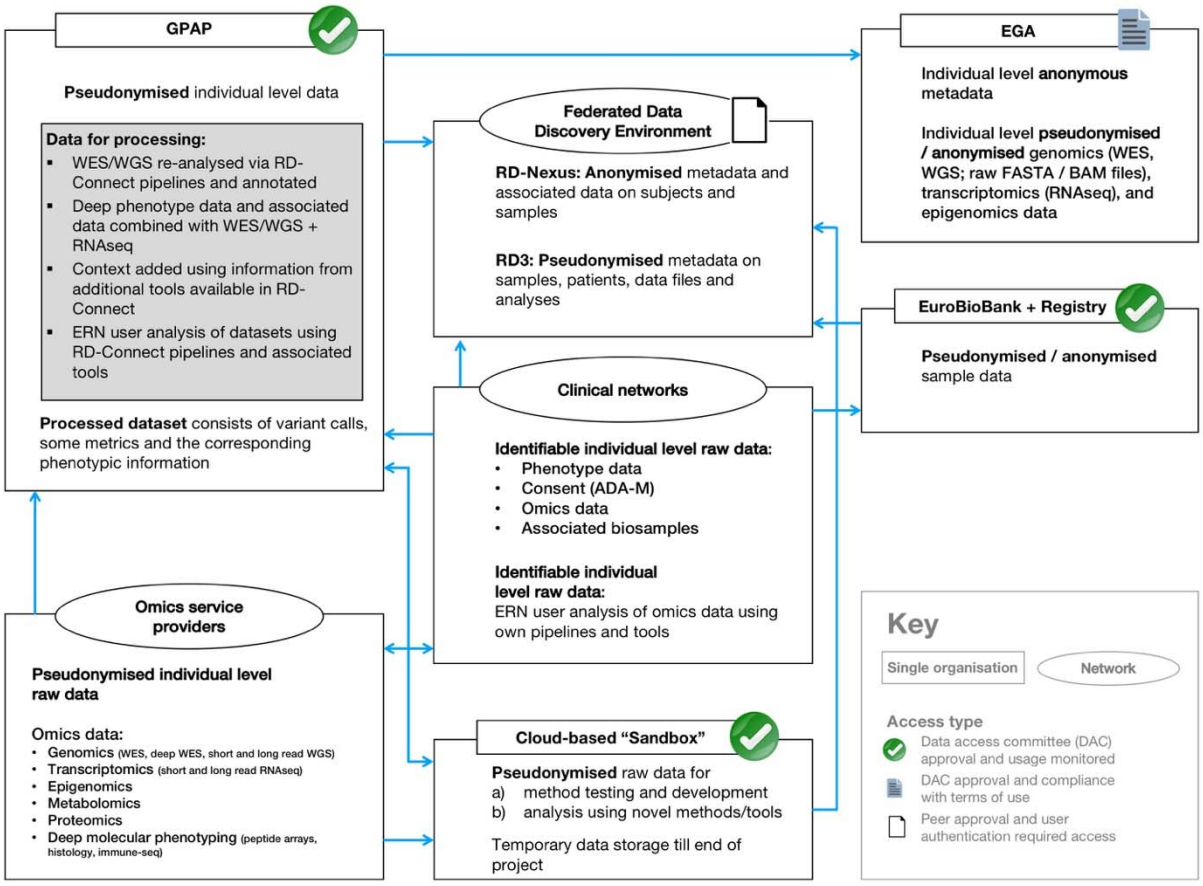

Figure 1: Data flow from the ERNclinical networks into Solve-RD.

2.2 Mapping to external initiatives

The next stage was to map the standards employed by the different Solve-RD resources to key external standards so that we maximise the interoperability of Solve-RD to other resources from the start. To do this we also mapped the metadata standards from Bioschemas[10], BioCADDIE DAT[11], and the evolving standards in the GA4GH, such as schemablocks[12] (Appendix 3). One key aim for Solve-RD is to have interoperability with the European Joint Programme on Rare Disease[13] (EJP-RD). As EJP-RD is a driver project for GA4GH, the activity of monitoring and mapping the GA4GH standards with respect to the Solve-RD standards not only allows interoperability with GA4GH compliant resources, but also ensure Solve-RD is compatible with EJP-RD workflows, maximising the opportunity for sharing Solve-RD data within the wider rare disease community and *vice versa*. We have collaborated with EJP-RD to apply the work done on mapping standards for this deliverable (D4.7) to help define the metadata standards being deployed by EJP-RD (https://github.com/ejp-rd-vp/resource-metadata-schema) with the aim of ensuring that the standards employed by both projects are consistent and interoperable.

[10] https://www.bioschemas.org/

[11] https://github.com/biocaddie/WG3-MetadataSpecifications/

[12] https://schemablocks.org/

[13] http://www.ejprarediseases.org/

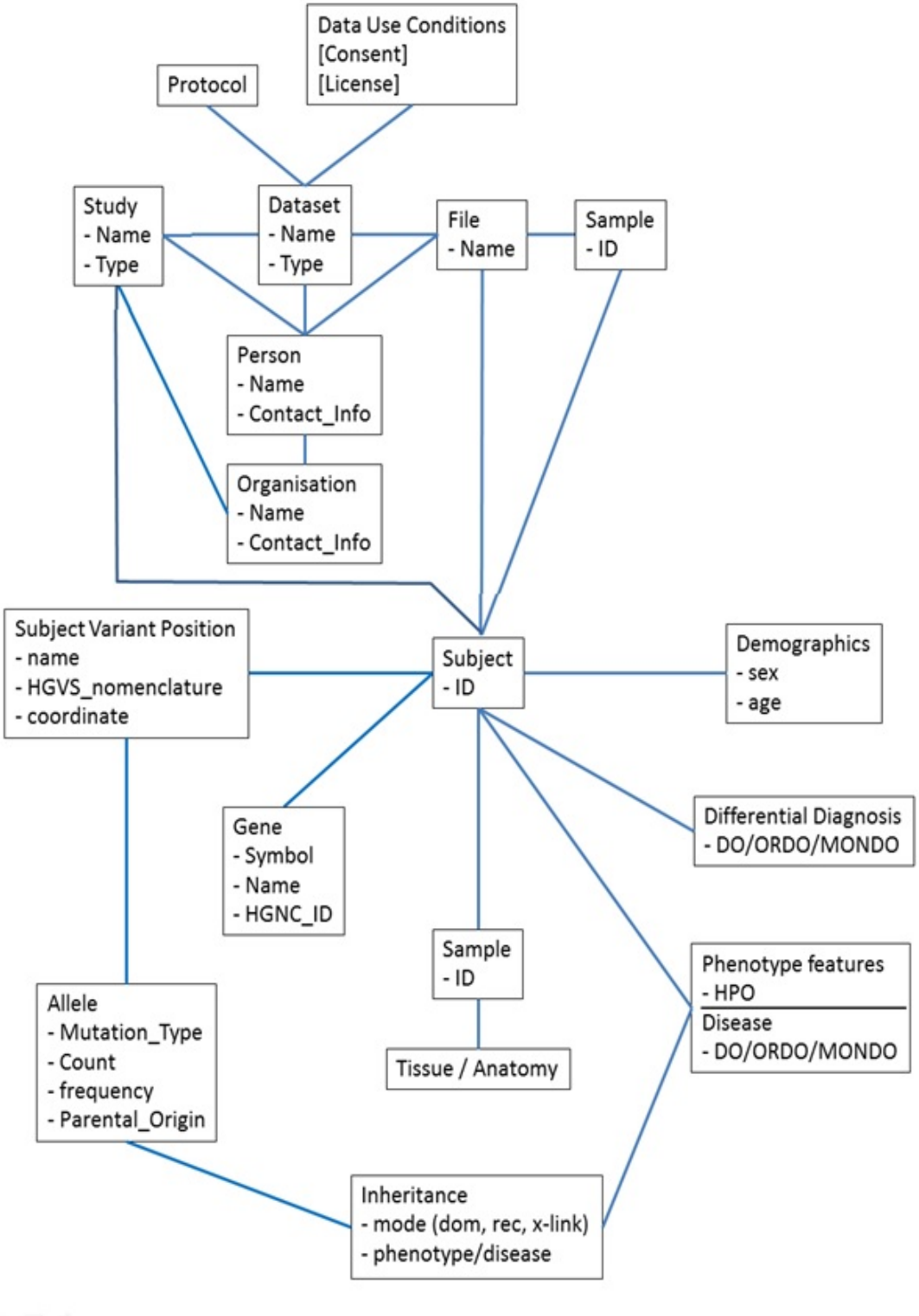

Figure 2: Version 1 of the RD3 data model. This design was informed by the standards requirements of the ERNs and participant resources in Solve-RD, and the mapping of these standards between each other.

For the discovery layer in Solve-RD, we defined the required standards to make the data discoverable and queryable by defining the 'findable facets' and associated AP respectively. https://github.com/ejp-rd-vp/resource-metadata-schema

The systematic use of public ontologies is a key enabler of cross resource harmonisation and interoperability. To describe rare diseases, we selected the Orphanet Rare Disease Ontolo-

gy[14] (ORDO) as it allows mapping via the Monarch Disease Ontology[15] (MONDO) to the widely used Disease Ontology[16] (DO) if required. For phenotypes of interest we support the use of the Human Phenotype Ontology[17] (HPO) and Online Mendelian Inheritance in Man[18] (OMIM), and for genes identifiers from the HUGO Gene Nomenclature Committee[19] (HGNC). For high level data use requirements we support the Data Use Ontology[20] (DUO), but for more extensive consent and data use requirements we use the Automatable Discovery and Access Matrix[21] (ADA-M) which has been implemented by or is undergoing evaluation by BBMRI, the Melbourne Genomics network, the Canadian Care4Rare-SOLVE[22] and the UK Tissue Directory and Coordination Center[23]. There are tools that can be used for mapping different ontologies, such as the Ontology Xref Service[24] (OxO), which helps support interoperability while supporting diversity where that is required, ensuring that standards development can be accommodated within Solve-RD.

The majority of human genetic data is subject to controlled access in accordance with the participant consent agreements. In Europe, the ELIXIR Authentication and Authorisation Infrastructure (AAI) is an established service that connects hundreds of research organisations, human data resources, EC H2020 projects across Europe. It is compliant with the GA4GH IT security implementation recommendations[25]. Therefore Solve-RD has decided to align with the ELIXIR AAI to enable access control interoperability. This has been chosen as it is based on industry standard OAuth2.0/OpenID Connect technology, and it will be compatible with the evolving GA4GH Researcher Identity standard which will allow additional claims to be attached to an identity (such as requested data use), and will be also used by the EJP-RD project. This ensures maximum interoperability and standardisation of the authentication and authorization process within Solve-RD. EGA, the RD-Connect GPAP and MOLGENIS are technically interoperable with the ELIXIR AAI.

The project's Genotype and Phenotype Analysis Platform (GPAP), which extends the RD-Connect database, is to execute a set of standard analysis pipelines to all incoming data to make the make it possible to do comparative analysis. These analyses include standardised alignment, variant calling, and annotation. Additionally a standard validation pipeline for ensuring the cloud sandboxes are operational and concordant with other sandboxes is defined here: https://molgenis.gitbooks.io/ngs_dna/ngs-protocols.html, which means the results of the same analysis on the same dataset performed on different sandboxes (on possibly different infrastructures) are identical.

Conclusion

This report (D4.7) combined with previous Solve-RD deliverables (D1.5 - Guidelines for Collection of Experimental Data; D1.4 - Deployment of PhenoTips custom forms according to the ERNs specification) from WP1, and WP4 deliverables (D4.3 - Central RD-Connect database serving Solve-RD, including user authentication and authorization; D4.5 - Metadata catalog operational, with initial content; and D4.1 - Principle Cloud services operational) document the foundational standards required for Solve-RD that are in place and operational,

[14] https://www.ebi.ac.uk/ols/ontologies/ordo
[15] https://www.ebi.ac.uk/ols/ontologies/mondo
[16] https://disease-ontology.org/
[17] https://hpo.jax.org/app/
[18] https://www.omim.org/
[19] https://www.genenames.org/
[20] https://github.com/EBISPOT/DUO
[21] https://www.nature.com/articles/s41525-018-0057-4
[22] https://care4rare.ca/solve/
[23] https://biobankinguk.org/
[24] https://www.ebi.ac.uk/spot/oxo/
[25] https://www.ga4gh.org/wp-content/uploads/2016May10_REV_SecInfrastructure.pdf

allowing data to be submitted to and exploited by Solve-RD in a consistent, efficient and unified way.

The standards relating to rare disease, genetics, and health continue to be develop and evolve. For example the GA4GH has eight new draft standards under review at the present time, and a subset of these are likely to become adopted standards across Europe. Therefore a key task for WP4 is to continue to monitor and contribute to the standards development landscape. As a driver projects for the GA4GH, both the EGA and EJP-RD with whom we are connected, can help drive the continued development of these standards, ensuring they are applicable to the Solve-RD and rare disease use cases in general.

Appendix 1: Mapping from the RD-Connect Genome-Phenome Analysis Platform (GPAP) to EGA

RD-Connect		EGA	
Object	Attribute	Object	Attribute
Participant	Phenotips_ID	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="PhenoTips_ID"
Participant	Submitter_ID	SAMPLE	SAMPLE_SET->SAMPLE->alias
Participant	MME	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="MME"
Participant	Registry_ID	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="Registry_ID"
Participant	Patient_Registry	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="Patient_Registry"
Participant	sex	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="gender"
Participant	phenotype	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="phenotype"
Participant	inheritance	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="mode_of_inheritance"
Participant	consanguinity	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="consanguinity"
Participant	parent	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="this patient is the parent of patient with id"
Participant	child	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="this patient is the child of patient with id"
Participant	sibling	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="this patient is the sibling of patient with id"
Participant	twin	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="this_patient_is_twin_of_patient_with_id"
Participant	cousin	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="cousin"
Participant	aunt_uncle	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="this patient is the aunt/uncle of patient with id"
Participant	niece_nephew	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="niece_nephew_of_patient"
Participant	grandparent	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="PhenoTips_ID"
Participant	grandchild	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="this patient is the grand-child of patient with id"
Participant	solved	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="case_solved"
Participant	gene_id	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="gene id"
Participant	disorder	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="disorder"
Participant	ORDO	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="ORDO"
Experiment	RD_Connect_ID_Experiment	EXPERIMENT	EXPERIMENT_SET->EXPERIMENT@alias
Experiment	Phenotips_ID	EXPERIMENT	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="PhenoTips_ID"
Experiment	Submitter_ID	EXPERIMENT	EXPERIMENT_SET->EXPERIMENT->EXPERIMENT_ATTRIBUTES->EXPERIMENT_ATTRIBUTE->TAG="Submitter Experiment ID"
Experiment	EGA_ID	EXPERIMENT	
Experiment	BioBank	EXPERIMENT	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="BioBank"
Experiment	Sample_ID	EXPERIMENT	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="BioBank Sample ID"
Experiment	library_source	EXPERIMENT	EXPERIMENT_SET->EXPERIMENT->DESIGN->LIBRARY_DESCRIPTOR->LIBRARY_SOURCE
Experiment	library_selection	EXPERIMENT	EXPERIMENT_SET->EXPERIMENT->DESIGN->LIBRARY_DESCRIPTOR->LIBRARY_SELECTION
Experiment	library_strategy	EXPERIMENT	EXPERIMENT_SET->EXPERIMENT->DESIGN->LIBRARY_DESCRIPTOR->LIBRARY_STRATEGY
Experiment	library_contruction_protocol	EXPERIMENT	EXPERIMENT_SET->EXPERIMENT->DESIGN->LIBRARY_DESCRIPTOR->LIBRARY_CONSTRUCTION_PROTOCOL
Experiment	design_description	EXPERIMENT	EXPERIMENT_SET->EXPERIMENT->DESIGN->DESIGN_DESCRIPTION
Experiment	read_insert_size	EXPERIMENT	EXPERIMENT_SET->EXPERIMENT->EXPERIMENT_ATTRIBUTES->EXPERIMENT_ATTRIBUTE->TAG="Read Insert Size"
Experiment	kit	EXPERIMENT	EXPERIMENT_SET->EXPERIMENT->EXPERIMENT_ATTRIBUTES->EXPERIMENT_ATTRIBUTE->TAG="Kit"
Experiment	project	EXPERIMENT	EXPERIMENT_SET->EXPERIMENT->EXPERIMENT_ATTRIBUTES->EXPERIMENT_ATTRIBUTE->TAG="Project"
Experiment	LOADDATE	EXPERIMENT	EXPERIMENT_SET->EXPERIMENT->EXPERIMENT_ATTRIBUTES->EXPERIMENT_ATTRIBUTE->TAG="Load Date"
Experiment	months_until_embargo	EXPERIMENT	EXPERIMENT_SET->EXPERIMENT->EXPERIMENT_ATTRIBUTES->EXPERIMENT_ATTRIBUTE->TAG="Lmonths_until_embargo"
Experiment	POSTEMBARGO-DATE	EXPERIMENT	EXPERIMENT_SET->EXPERIMENT->EXPERIMENT_ATTRIBUTES->EXPERIMENT_ATTRIBUTE->TAG="POSTEMBARGODATE"
Experiment	tissue	EXPERIMENT	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="Source Tissue"
Run / File	RD_Connect_ID_Experiment	RUN	EXPERIMENT@alias
Run / File	library_name	RUN	EXPERIMENT_SET->EXPERIMENT->EXPERIMENT_ATTRIBUTES->EXPERIMENT_ATTRIBUTE->TAG="Submitter Library name"
Run / File	file_name	RUN	RUN_SET->RUN->DATA_BLOCK->FILES->FILE@filename
Run / File	file_type	RUN	RUN_SET->RUN->DATA_BLOCK->FILES->FILE@filetype
Run / File	library_layout	RUN	EXPERIMENT_SET->EXPERIMENT->DESIGN->LIBRARY_DESCRIPTOR->LIBRARY_LAYOUT->SINGLE or

			PAIRED
Run / File	read_length	RUN	EXPERIMENT_SET->EXPERIMENT->DESIGN->LIBRARY_DESCRIPTOR->LIBRARY_LAYOUT->PAIRED@NOMINAL_LENGTH
Run / File	adapter	RUN	EXPERIMENT_SET->EXPERIMENT->EXPERIMENT_ATTRIBUTES->EXPERIMENT_ATTRIBUTE->TAG="Adapter"
Run / File	trimed	RUN	EXPERIMENT_SET->EXPERIMENT->EXPERIMENT_ATTRIBUTES->EXPERIMENT_ATTRIBUTE->TAG="trimmed"
Run / File	bqsr	RUN	EXPERIMENT_SET->EXPERIMENT->EXPERIMENT_ATTRIBUTES->EXPERIMENT_ATTRIBUTE->TAG="bqsr"
Run / File	instrument_model	RUN	RUN_SET->RUN->PLATFORM->ILLUMINA->INSTRUMENT_MODEL
Run / File	RD_Connect_ID_Experiment	RUN	EXPERIMENT@alias
Run / File	library_name	RUN	EXPERIMENT_SET->EXPERIMENT->EXPERIMENT_ATTRIBUTES->EXPERIMENT_ATTRIBUTE->TAG="Submitter Library name"
Run / File	file_name	RUN	RUN_SET->RUN->DATA_BLOCK->FILES->FILE@filename
Run / File	file_type	RUN	RUN_SET->RUN->DATA_BLOCK->FILES->FILE@filetype
Run / File	library_layout	RUN	EXPERIMENT_SET->EXPERIMENT->DESIGN->LIBRARY_DESCRIPTOR->LIBRARY_LAYOUT->SINGLE or PAIRED
Run / File	read_length	RUN	EXPERIMENT_SET->EXPERIMENT->DESIGN->LIBRARY_DESCRIPTOR->LIBRARY_LAYOUT->PAIRED@NOMINAL_LENGTH
Run / File	adapter	RUN	EXPERIMENT_SET->EXPERIMENT->EXPERIMENT_ATTRIBUTES->EXPERIMENT_ATTRIBUTE->TAG="Adapter"
Run / File	trimed	RUN	EXPERIMENT_SET->EXPERIMENT->EXPERIMENT_ATTRIBUTES->EXPERIMENT_ATTRIBUTE->TAG="trimmed"
Run / File	bqsr	RUN	EXPERIMENT_SET->EXPERIMENT->EXPERIMENT_ATTRIBUTES->EXPERIMENT_ATTRIBUTE->TAG="bqsr"
Run / File	instrument_model	RUN	RUN_SET->RUN->PLATFORM->ILLUMINA->INSTRUMENT_MODEL

Appendix 2: Mapping from the EGA to MOLGENIS (the basis of the sandbox)

MOLGENIS		EGA	
Object	**Attribute**	**Object**	**Attribute**
File	FileID	FILE	EGA File accession
File	FileName	OBJECT	OBJECT->FILE->filename
File	FileType	OBJECT	OBJECT->FILE->filetpye
File	ServerName	OBJECT	OBJECT_SET->OBJECT->OBJECT_ATTRIBUTES->OBJECT_ATTRIBUTE->TAG=ServerName
File	Md5CheckSum	OBJECT	OBJECT->unencrypted_checksum
File	FileEntryDate	OBJECT	OBJECT_SET->OBJECT->OBJECT_ATTRIBUTES->OBJECT_ATTRIBUTE->TAG=FileEntryDate
File	FileLastModifyDate	OBJECT	OBJECT_SET->OBJECT->OBJECT_ATTRIBUTES->OBJECT_ATTRIBUTE->TAG=FileLastModifyDate
File	FileNotes	OBJECT	OBJECT_SET->OBJECT->OBJECT_ATTRIBUTES->OBJECT_ATTRIBUTE->TAG=FileNotes
File	N/A	FILE	OBJECT->FILE->checksum_type=md5
Sample	SampleID	SAMPLE	SampleID
Sample	PersonID	SAMPLE	PersonID
Sample	MaterialType	SAMPLE	MaterialType
Sample	TissueType	SAMPLE	TissueType
Sample	SampleDate	SAMPLE	SampleDate
Sample	TimePoint	SAMPLE	TimePoint
Sample	SampleEntryDate	SAMPLE	SampleEntryDate
Sample	SampleLastModify-Date	SAMPLE	SampleLastModifyDate
Sample	SampleNotes	SAMPLE	SampleNotes
Person	PersonID	SAMPLE	PersonID
Person	MotherID	SAMPLE	MotherID
Person	FatherID	SAMPLE	FatherID
Person	PseudoID	SAMPLE	PseudoID
Person	Sex	SAMPLE	Sex
Person	Age	SAMPLE	Age
Person	FamilyID	SAMPLE	FamilyID
Person	PersonEntryDate	SAMPLE	PersonEntryDate
Person	PersonLastModifyDa-te	SAMPLE	PersonLastModifyDate
Person	PersonNotes	SAMPLE	PersonNotes
Person-Consent	PersonConsentNotes	SAMPLE / DATASET	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE-ATTRIBUTE->TAG=PersonConsentNotes, POLICY_SET->POLICY->DATA_USES->DATA_USE->@ontology/@code/@version
Assessment	PersonID	SAMPLE	Sample Alias
Assessment	AssignStrategy	SAMPLE	AssignStrategy
Assessment	AssessmentDate	SAMPLE	AssessmentDate
Assessment	AssessmentEntryDa-te	SAMPLE	AssessmentEntryDate
Assessment	AssessmentLastMo-difyDate	SAMPLE	AssessmentLastModifyDate
Assessment	AssessmentNotes	SAMPLE	AssessmentNotes
Hpo	HpoCode	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_LINKS->SAMPLE-LINK->XREF
Hpo	HpoName	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE-ATTRIBUTE->TAG=HpoName
Hpo	VersionNumber	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE-ATTRIBUTE->TAG=VersionNumber
Hpo	HpoEntryDate	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE-ATTRIBUTE->TAG=HpoEntryDate
Hpo	HpoLastModifyDate	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE-ATTRIBUTE->TAG=HpoLastModifyDate
Hpo	HpoNotes	SAMPLE	SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE-ATTRIBUTE->TAG=HpoNotes
Collection	CollectionID	DATASET	DATASET alias
Collection	ConsentID	DATASET	POLICY_SET->POLICY->DATA_USES->DATA_USE->@ontology/@code/@version
Collection	CollectionName	DATASET	DATASET title
Collection	CollectionDescription	DATASET	DATASET description
Collection	CollectionEntryDate	DATASET	DATASET_SET->DATASET->DATASET_ATTRIBUTES->DATASET-ATTRIBUTE->TAG=CollectionEntryDate
Collection	CollectionOwner	DATASET	DATASET_SET->DATASET->DATASET_ATTRIBUTES->DATASET-ATTRIBUTE->TAG=CollectionOwner
Collection	CollectionLastMo-difyDate	DATASET	DATASET_SET->DATASET->DATASET_ATTRIBUTES->DATASET-ATTRIBUTE->TAG=CollectionLastModifyDate
Collection	CollectionNotes	DATASET	DATASET_SET->DATASET->DATASET_ATTRIBUTES->DATASET-ATTRIBUTE->TAG=CollectionNotes

Appendix 3: Initial mapping of bioCADDIE DATS, RD3, EGA, RD-Connect GPAP and BioSchemas

BioCADDIE DATS		RD3	EGA	RD-Connect (GPAP)	BioSchema
Entity	Attribute				
Dataset	title	CollectionName	Dataset title	RD-Connect-Experiment-ID	https://schema.org/name
Dataset	types			Experiment->experiment_type	https://schema.org/additionalType
Dataset	->information	ProtocolNotes	EXPERIMENT_SET->EXPERIMENT->DESIGN->DESIGN_DESCRIPTION	Experiment->design_description	
Dataset	->method	ProtocolName	EXPERIMENT_SET->EXPERIMENT->DESIGN->LIBRARY_DESCRIPTOR->LIBRARY_STRATEGY	Experiment->library_strategy	
Dataset	->platform		RUN_SET->RUN->PLATFORM->ILLUMINA->INSTRUMENT_MODEL	Experiment->kit	
Dataset	->instrument			Experiment->file->instrument_model	
Dataset	->extraProperties			library_source,library_selection,library_construction,library_layout,read_length	
Dataset	creators	User	DATASET->submission_account		https://schema.org/creator
Dataset	->Person->identifier	UserID	submision_account_id	CAS->username	
Dataset	->Person->fullName	concat(FirstName LastName)		CAS->full name	
Dataset	->Person->firstName	FirstName		CAS->First Given Name	
Dataset	->Person->lastName	LastName		CAS->Surname	
Dataset	->Person->email	Email		CAS->email address	
Dataset	->Person->affiliations			CAS->group	
Dataset	->Person->roles	Role			
Dataset	->Person->extraProperties	Notes:UserNotes, CanGivePermission:boolean			
Dataset	identifier (identifiersInformation)				http://schema.org/identifier
Dataset	alternateIdentifiers (AlternateIdentifiersInformation)				http://schema.org/identifier
Dataset	relatedIdentifers (RelatedIdentifiersInformation)				http://schema.org/identifier
Dataset	version				https://schema.org/version
Dataset	date				http://schema.org/license
Dataset	->Date->date	CollectionEntryDate,CollectionLastModifyDate		Experiment->load date (?not sure which date is referring to, sampling,uploading,..)	https://schema.org/Date
Dataset	->Date->type	CollectionEntryDate,CollectionLastModifyDate			https://schema.org/category
Dataset	description	CollectionDescription	DATASET description		https://schema.org/description
Dataset	keywords				https://schema.org/keywords
Dataset	isAbout	distinct(HpoCode) = Disease	distinct(SAMPLE_SET->SAMPLE->SAMPLE_ATTRIBUTES->SAMPLE_ATTRIBUTE->TAG="phenotype") = disease	distinct(phenotype) = disease	https://schema.org/about
DatasetDistribution	identifier (identifiersInformation)				http://schema.org/identifier

DatasetDistribution	alternateIdentifiers (AlternateIdentifiersInfor-mation)				http://schema.org/identifier
DatasetDistribution	relatedIdentifers (Rela-tedIdentifiersInformation)				http://schema.org/identifier
DatasetDistribution	version				https://schema.org/version
DatasetDistribution	title				https://schema.org/name
DatasetDistribution	description				https://schema.org/description
DatasetDistribution	date				http://schema.org/license
DatasetDistribution	->Date->date	CollectionEntryDa-te,CollectionLastModifyDate			https://schema.org/Date
DatasetDistribution	->Date->type	CollectionEntryDa-te,CollectionLastModifyDate			https://schema.org/category
DatasetDistribution	storedIn		EGA	RD-Connect	https://schema.org/includedInDataCatalog
DatasetDistribution	access				https://schema.org/accessibilityAPI
DatasetDistribution	-access->identifier	ConsentID	Policy Accession		
DatasetDistribution	licenses				Thing > Property > license
DatasetDistribution	formats	FileType	file_type		Thing > Property > fileFormat
DatasetDistribution	size		sum(file_size)		Thing > Property > contentSize
DatasetDistribution	unit				Thing > Property > unitText