RD Connect

RD-Connect Sample Catalogue

Guide for *Biobank data managers*: preparing and uploading sample data

In this guide

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 305444 (RD-Connect).

1. Sharing sample data via the RD-Connect Sample Catalogue

Biological materials from rare disease patients are precious resources which play important role in driving research on understanding disease pathogenesis, validation of diagnoses and development of treatments. RD biobanks exist for the purpose of collecting and sharing of biological samples and clinical information from individuals with rare diseases, but it is often difficult for people to find out what each resource contains and how to access it. To tackle this bottleneck, RD biobanks are encouraged sign up and share their biobank information via the RD-Connect Registry & Biobank Finder, and sample collections via the RD-Connect Sample Catalogue. Information on how RD biobanks can express interest to participate and the process of participation can be found on the RD-Connect project website (www.rd-connect.eu).

Participating RD biobanks can make biological samples available to a wider scientific community via our searchable and dynamic RD-Connect Sample Catalogue, a tool which provides essential information on the biological samples and collections can be shared among researchers . Other advantages include:

- Become a member of EuroBioBank.
- Increase sample and data discovery and distribution.
- Establish new collaborations and enhance networking activities in the field of RD research.
- state-of-art tool connecting bioinformatics data generated from samples to clinical phenotypes.

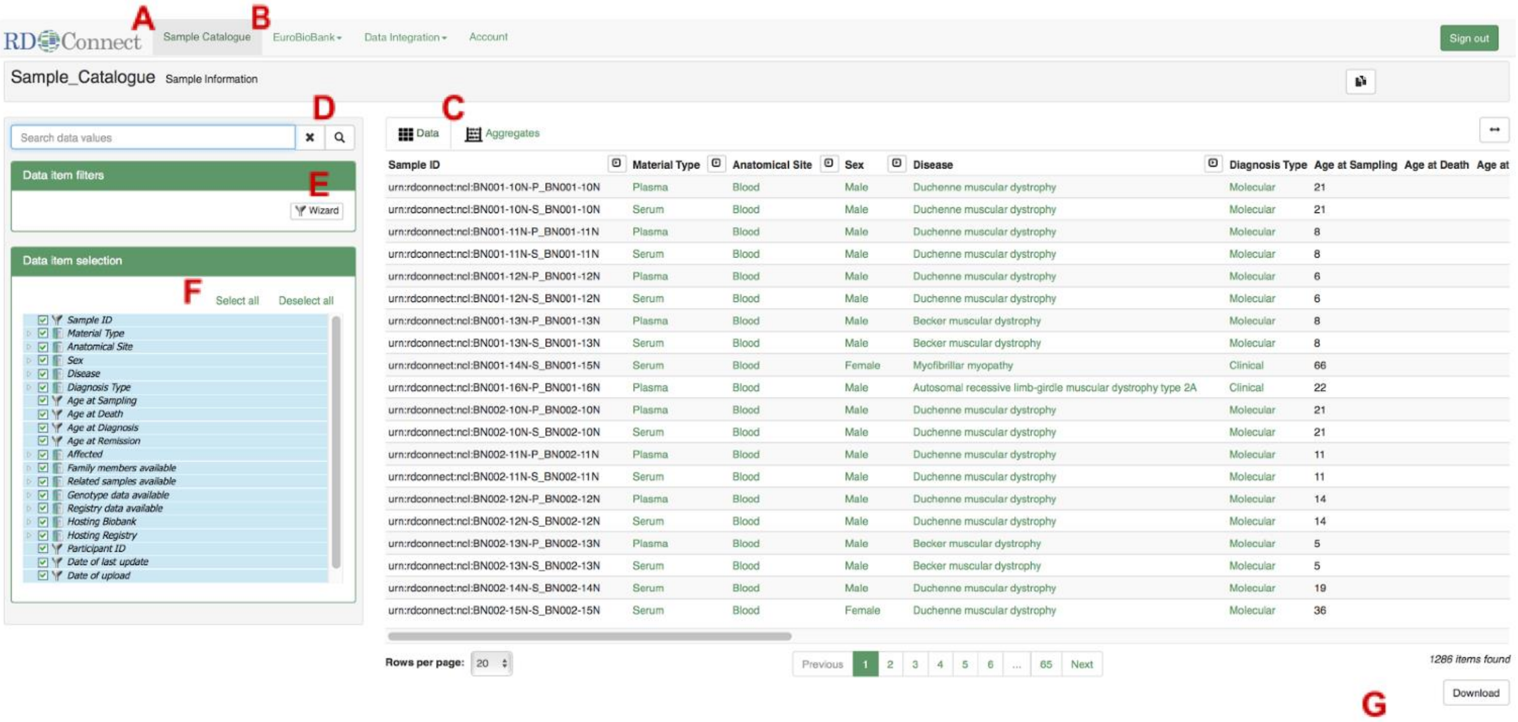
This guide is aimed for Biobank Data Managers to help them prepare the minimum sample dataset from their biobanks for sharing to the RD-Connect Sample Catalogue.

https://samples.rd-connect.eu/

2. Access and Registration

Once the RD Biobank has been admitted to participate in the platform, access requests to the Sample Catalogue is handled through the RD-Connect Project Coordination office at Newcastle.

3. Using the Sample Catalogue



Sample_Catalogue Sample Information

Search data values

Data item filters

Data item selection

Select all Deselect all

- ☑ Sample ID
- ☑ Material Type
- ☑ Anatomical Site
- ☑ Sex
- ☑ Disease
- ☑ Diagnosis Type
- ☑ Age at Sampling
- ☑ Age at Death
- ☑ Age at Diagnosis
- ☑ Age at Remission
- ☑ Affected
- ☑ Family members available
- ☑ Related samples available
- ☑ Genotype data available
- ☑ Registry data available
- ☑ Hosting Biobank
- ☑ Hosting Registry
- ☑ Participant ID
- ☑ Date of last update
- ☑ Date of upload

Sample ID	Material Type	Anatomical Site	Sex	Disease	Diagnosis Type	Age at Sampling	Age at Death	Age at
urn:rdconnect:ncl:BN001-10N-P_BN001-10N	Plasma	Blood	Male	Duchenne muscular dystrophy	Molecular	21		
urn:rdconnect:ncl:BN001-10N-S_BN001-10N	Serum	Blood	Male	Duchenne muscular dystrophy	Molecular	21		
urn:rdconnect:ncl:BN001-11N-P_BN001-11N	Plasma	Blood	Male	Duchenne muscular dystrophy	Molecular	8		
urn:rdconnect:ncl:BN001-11N-S_BN001-11N	Serum	Blood	Male	Duchenne muscular dystrophy	Molecular	8		
urn:rdconnect:ncl:BN001-12N-P_BN001-12N	Plasma	Blood	Male	Duchenne muscular dystrophy	Molecular	6		
urn:rdconnect:ncl:BN001-12N-S_BN001-12N	Serum	Blood	Male	Duchenne muscular dystrophy	Molecular	6		
urn:rdconnect:ncl:BN001-13N-P_BN001-13N	Plasma	Blood	Male	Becker muscular dystrophy	Molecular	8		
urn:rdconnect:ncl:BN001-13N-S_BN001-13N	Serum	Blood	Male	Becker muscular dystrophy	Molecular	8		
urn:rdconnect:ncl:BN001-14N-S_BN001-15N	Serum	Blood	Female	Myofibrillar myopathy	Clinical	66		
urn:rdconnect:ncl:BN001-16N-P_BN001-16N	Plasma	Blood	Male	Autosomal recessive limb-girdle muscular dystrophy type 2A	Clinical	22		
urn:rdconnect:ncl:BN002-10N-P_BN002-10N	Plasma	Blood	Male	Duchenne muscular dystrophy	Molecular	21		
urn:rdconnect:ncl:BN002-10N-S_BN002-10N	Serum	Blood	Male	Duchenne muscular dystrophy	Molecular	21		
urn:rdconnect:ncl:BN002-11N-P_BN002-11N	Plasma	Blood	Male	Duchenne muscular dystrophy	Molecular	11		
urn:rdconnect:ncl:BN002-11N-S_BN002-11N	Serum	Blood	Male	Duchenne muscular dystrophy	Molecular	11		
urn:rdconnect:ncl:BN002-12N-P_BN002-12N	Plasma	Blood	Male	Duchenne muscular dystrophy	Molecular	14		
urn:rdconnect:ncl:BN002-12N-S_BN002-12N	Serum	Blood	Male	Duchenne muscular dystrophy	Molecular	14		
urn:rdconnect:ncl:BN002-13N-P_BN002-13N	Plasma	Blood	Male	Becker muscular dystrophy	Molecular	5		
urn:rdconnect:ncl:BN002-13N-S_BN002-13N	Serum	Blood	Male	Becker muscular dystrophy	Molecular	5		
urn:rdconnect:ncl:BN002-14N-S_BN002-14N	Serum	Blood	Male	Duchenne muscular dystrophy	Molecular	19		
urn:rdconnect:ncl:BN002-15N-S_BN002-15N	Serum	Blood	Female	Duchenne muscular dystrophy	Molecular	36		

Rows per page: 20

Previous 1 2 3 4 5 6 ... 65 Next

1286 items found

A Navigate to the Sample Catalogue

B Navigate to the original sample data from the biobanks

C In the ‘Data’-tab you can view the Samples, in the ‘Aggregates’ you can view the counts over the sample data.

D Search through the data.

E Filters can be set using the ‘Wizard’. Active filters are shown in the ‘Data item filters’ window.

F In the ‘Data item selection’ window you can select or deselect columns of interest. Here you can also set filters for specific columns by clicking the Y or ▤ icon.

G Download the data in csv or excel format. Only selected columns will be downloaded.

4. Uploading data to the Sample Catalogue

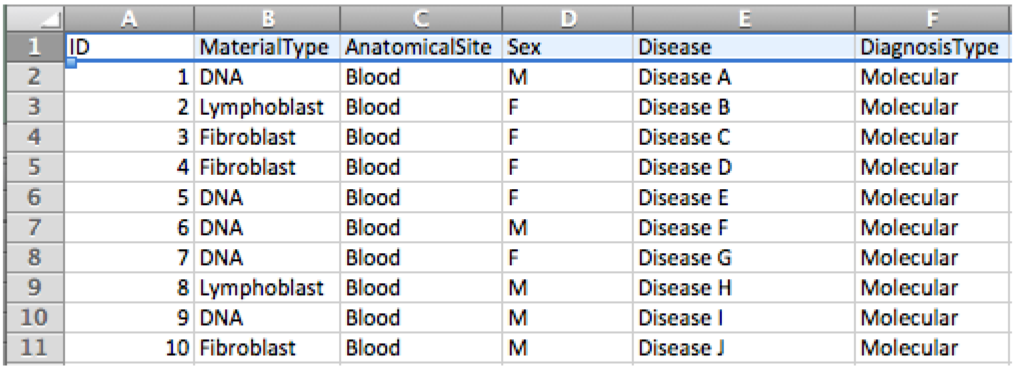
This Sample Catalogue is intended to facilitate the discovery of samples and samples data from Rare Diseases biobanks. It also provides information about sample collections and studies done on the registered samples.

Sample data will be loaded to the catalogue and will exist in two formats: 1) the original format in which data is uploaded, and 2) the format of the RD-Connect Sample Catalogue. Initially, the data is uploaded to the sample catalogue in the exact format it was received from the biobank. This will only be visible for data managers from the concerned biobank. After this first step, the data will be formatted to the sample catalogue model. This will be done by 'mapping' the data, which means JavaScript queries are used to convert the columns from the original format to the catalogue format. To create the catalogue format, the data will be mapped after uploading.
This manual only describes how to upload your data.

To upload your (raw) sample data login to https://samples.rd-connect.eu/. Click the 'upload' button and select the file you want to import. For more information about how to import your data, please also check: https://molgenis.gitbooks.io/molgenis/content/user_documentation/guide-explore.html.

4.1.Modelling data with EMX format

Sample data can be uploaded in an EMX format (full documentation here), in the data model format of the RD-Connect Sample Catalogue. This is a spreadsheet format using excel, zip or .tsv/csv files. The model of the data is defined in a sheet called 'attributes'. Each attribute describes the characteristics of a column. The easiest way to create an attributes sheet is to open your data in excel and copy all the headers from your data (*Figure 1*).



	A	B	C	D	E	F
1	ID	MaterialType	AnatomicalSite	Sex	Disease	DiagnosisType
2	1	DNA	Blood	M	Disease A	Molecular
3	2	Lymphoblast	Blood	F	Disease B	Molecular
4	3	Fibroblast	Blood	F	Disease C	Molecular
5	4	Fibroblast	Blood	F	Disease D	Molecular
6	5	DNA	Blood	F	Disease E	Molecular
7	6	DNA	Blood	M	Disease F	Molecular
8	7	DNA	Blood	F	Disease G	Molecular
9	8	Lymphoblast	Blood	M	Disease H	Molecular
10	9	DNA	Blood	M	Disease I	Molecular
11	10	Fibroblast	Blood	M	Disease J	Molecular

***Figure 1:** Example dummy data.*

Step 1: Open your dataset in excel, copy all the headers from the dataset (*Figure 1*).
Step 2: Create a second sheet with the name 'attributes' (*Figure 2*).
Step 3: Next, paste the headers transposed (*Figure 3a, b)* in the new sheet, in the first column, this column will be called 'name' (*Figure 4*).

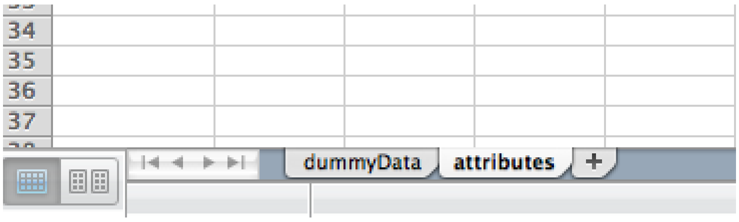


***Figure 2:** Select attributes sheet to enter data description.*

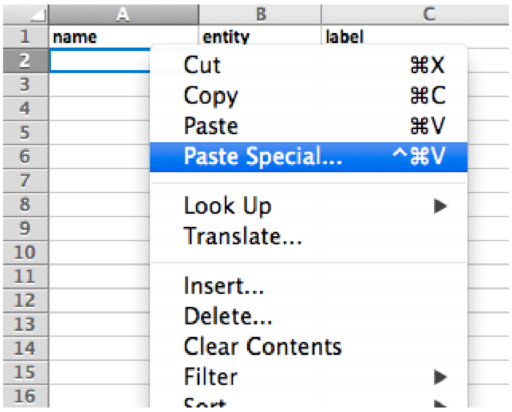
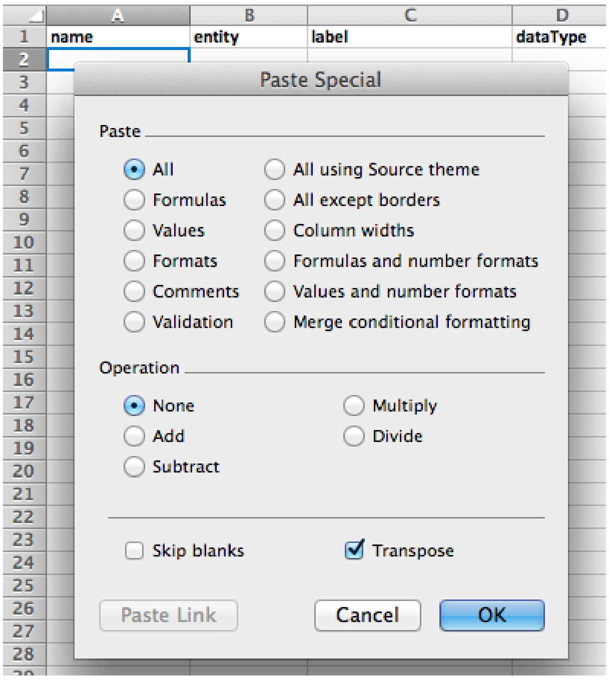


***Figure 3a:** Paste the headers transposed.*



Paste Special

Paste

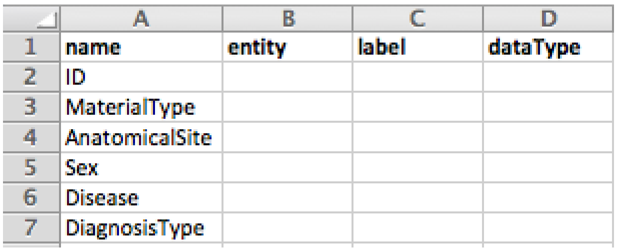
● All ○ All using Source theme
○ Formulas ○ All except borders
○ Values ○ Column widths
○ Formats ○ Formulas and number formats
○ Comments ○ Values and number formats
○ Validation ○ Merge conditional formatting

Operation

● None ○ Multiply
○ Add ○ Divide
○ Subtract

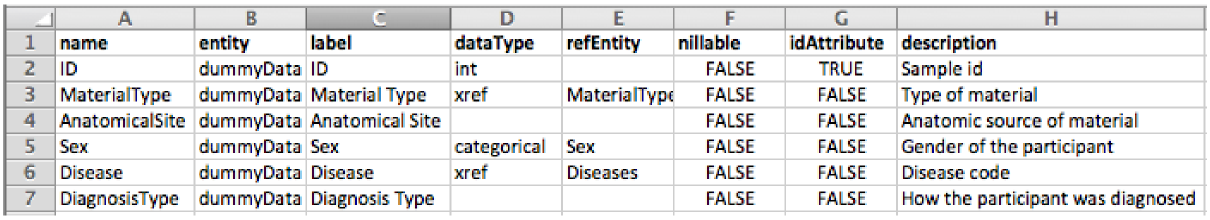
□ Skip blanks ☑ Transpose

***Figure 3b:** Paste the headers transposed.*



	A	B	C	D
1	name	entity	label	dataType
2	ID			
3	MaterialType			
4	AnatomicalSite			
5	Sex			
6	Disease			
7	DiagnosisType			

***Figure 4:** Entering your data attributes.*



	A	B	C	D	E	F	G	H
1	name	entity	label	dataType	refEntity	nillable	idAttribute	description
2	ID	dummyData	ID	int		FALSE	TRUE	Sample id
3	MaterialType	dummyData	Material Type	xref	MaterialTyp	FALSE	FALSE	Type of material
4	AnatomicalSite	dummyData	Anatomical Site			FALSE	FALSE	Anatomic source of material
5	Sex	dummyData	Sex	categorical	Sex	FALSE	FALSE	Gender of the participant
6	Disease	dummyData	Disease	xref	Diseases	FALSE	FALSE	Disease code
7	DiagnosisType	dummyData	Diagnosis Type			FALSE	FALSE	How the participant was diagnosed

***Figure 5:** Adding details for each attribute.*

Step 4: Now fill in the other columns in the attributes sheet. In the example of *Figure 5* the following characteristics of each attribute are defined:

- **name:** the name of the attribute. Make sure the names contain no special characters, only letters, numbers, '_' and '#' are allowed. Spaces are not allowed.
- **entity:** name of the sheet the data is located
- **label:** how the name of the attribute that is shown in the data explorer. Here special characters and spaces are allowed.
- **dataType:** the type of data. 'string' is the default data type. 'int' are natural numbers. 'xref' refer to another entity, where the possible values for the attribute are defined. In this way the options for the attribute are delimited and makes the data easily searchable. The same applies for 'categorical', but is usually used when there are only a few options possible, for example to define gender, when only 'male' or 'female' are possible options (*Figures 6-8*).
- **refEntity:** the entity that is referred to.
- **nillable:** defines if the attribute can be null or not. idAttributes cannot be null.
- **idAttribute:** defines the id attribute. Every entity needs an unique id attribute. If there is no id attribute available, the id can be automatically generated by setting the idAttribute on 'AUTO'.
- **description:** description of the attribute.

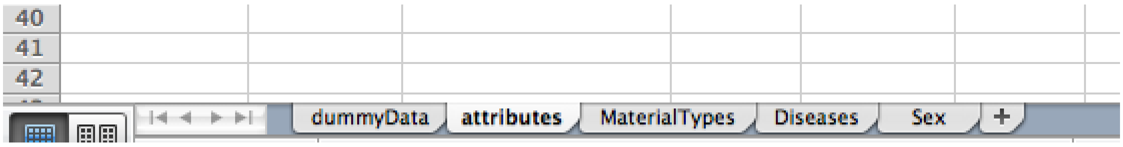
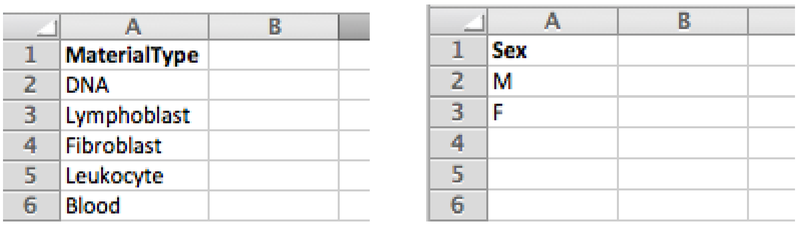


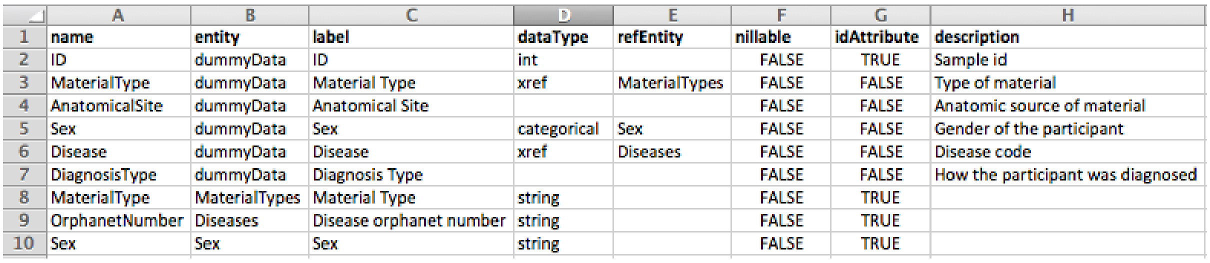
Figure 6: *Add the reference entities*



	A	B
1	**MaterialType**	
2	DNA	
3	Lymphoblast	
4	Fibroblast	
5	Leukocyte	
6	Blood	

	A	B
1	**Sex**	
2	M	
3	F	
4		
5		
6		

Figure 7: *Examples of reference entities.*



	A	B	C	D	E	F	G	H
1	**name**	**entity**	**label**	**dataType**	**refEntity**	**nillable**	**idAttribute**	**description**
2	ID	dummyData	ID	int		FALSE	TRUE	Sample id
3	MaterialType	dummyData	Material Type	xref	MaterialTypes	FALSE	FALSE	Type of material
4	AnatomicalSite	dummyData	Anatomical Site			FALSE	FALSE	Anatomic source of material
5	Sex	dummyData	Sex	categorical	Sex	FALSE	FALSE	Gender of the participant
6	Disease	dummyData	Disease	xref	Diseases	FALSE	FALSE	Disease code
7	DiagnosisType	dummyData	Diagnosis Type			FALSE	FALSE	How the participant was diagnosed
8	MaterialType	MaterialTypes	Material Type	string		FALSE	TRUE	
9	OrphanetNumber	Diseases	Disease orphanet number	string		FALSE	TRUE	
10	Sex	Sex	Sex	string		FALSE	TRUE	

Figure 8: *Make sure the reference entities are also defined in the attributes sheet.*

An excel file is not required. Different file extensions are possible. For example, another option for data upload is a .zip file with .csv files. Using this example, the zip file would contain an 'attributes.csv' file and a file called 'dummyData.csv'. Important: the file with the sample data should have the name of the entity.
In most cases an attributes sheet is enough to define the model. But if you want to structure the data even more, you can add an 'entities' sheet to describe the different entities, our add the entities to a package by defining a "packages" sheet.
More information about entities, attributes, data types and EMX formats in general can be found at https://molgenis.gitbooks.io/molgenis/content/user_documentation/ref-emx.html.

4.2. Delivering raw sample data

Data can be delivered in excel, csv or tsv format. The data will be uploaded in EMX format. Preferably the EMX is already created by the biobank.

The model is based on the MIABIS format, this represents the minimum information required for biobank data sharing.
To conform with the catalogue data model, a number of fields are required:

Attribute	Possible values	Description
Sample ID (pseudonymised)		
Material type	Leukocyte, Plasma, Serum, Fibroblast, RNA, Myoblast, DNA, Lymphoblast	
Anatomical Site	Muscle Tissue, Blood	Anatomic source of the material
Disease	ORPHA codes	Disease code from the Orphanet ontology
Hosting Biobank	Identifier from ID-Card	
Patient ID (pseudonymised)		
Date of last update		

The sample ID's and patient ID's should be pseudonymised. However, it is the biobank's responsibility to do the pseudonymisation in accordance with the RD-Connect Code of Practice.

The following attributes are not fields to be provided, but will always have a value. The bold values will be assigned when no value is provided by the biobank:

Attribute	Possible values	Description
Diagnosis type	Autoptic, Biochemical, Clinical, Cytogenetics, Echographic, Electrophysiological, Enzymatic, Histological, Molecular, Neuroradiological, **Not Specified**, Radiological	Classification of how the participant was diagnosed
Sex	Male, Female, Ambiguous, **Unknown**	
Affected	Yes, No, Unknown, **Not available**, not asked	Anatomic source of the material
Genotype	Yes, No, Unknown, **Not available**, not asked	Disease code from the Orphanet ontology
Family	Yes, No, Unknown, **Not available**, not asked	
Related Samples	Yes, No, Unknown, **Not available**, not asked	

Not required information:

Attribute	Possible values	Description
Registry	Identifier from ID-Card	Only if multiple registries are uploaded
Age at Sampling		
Age at Death		
Age at Diagnosis		
Age at Remission		

5. Data model of the RD-Connect Sample Catalogue

Attribute	Required	Possible values	Description
Sample ID (anonymous!)	Yes		
Material Type	Yes	DNA, Lymphoblast, Fibroblast, Leukocyte, Blood	
Anatomical Site	Yes	Muscle Tissue, Blood	Anatomic source of the material
Diagnosis type	Yes		Classification of how the participant was diagnosed
Disease (will be updated in the model)	Yes	Orphanet, OMIM, ICD10	Disease code from an established ontology
Sex	Yes	Male, Female, Unknown	Biological sex of the participant
Age at Sampling			
Age at Death			
Age at Diagnosis			
Age at Remission			
Affected	Yes	Yes, No, Unknown, Not available, not asked	Person showing a disease phenotype?
Genotype	Yes	Yes, No, Unknown, Not available, Not asked	Genotype data available?
Family	Yes	Yes, No, Unknown, Not available, Not asked	Is information from relatives available?
Related Samples	Yes	Yes, No, Unknown, Not available, Not asked	Are related samples available?
Registry	Only if multiple registries are uploaded		
Hosting Biobank	Yes		Id from ID-Card
Patient ID (anonymous!)	Yes		
Date of last update	Yes		

6. Useful resources

RD-Connect Sample Catalogue	https://samples.rd-connect.eu/
RD-Connect Registry & Biobank Finder	http://catalogue.rd-connect.eu/
RD-Connect Genome-Phenome Analysis Platform	https://platform.rd-connect.eu/
RD-Connect EuroBioBank	http://www.eurobiobank.org/

Bioportal	http://bioportal.bioontology.org/
UBERON ontology (anatomy)	http://purl.bioontology.org/ontology/UBERON/
Orphanet (rare disease)	http://purl.bioontology.org/ontology/ORDO
ICD-10	http://purl.bioontology.org/ontology/ICD10
OMIM	http://purl.bioontology.org/ontology/OMIM
MIABIS 2.0	https://github.com/MIABIS/miabis/wiki

7. Contact the team

If you require extra support on how to structure your sample data please write to:
Mariska Slofstra m.k.slofstra@umcg.nl

If you are interested to hear about how RD biobanks can participate in RD-Connect please write to:
Mary Wang mwang@telethon.it

If you have enquiries about user accounts and access please write to:
John Dawson platform@rd-connect.eu

Guide authors: Marije van der Geest (v1, 2016), Chiuhui Mary Wang (v2, 4/2017; v2.1 11/2017)