EOSC-LIFE: BUILDING A DIGITAL SPACE FOR THE LIFE SCIENCES

EOSC-LIFE WP4 TOOLBOX:
Pilot study protocol for the evaluation of the categorisation system

WP4 – **Policies, specifications and tools for secure management of sensitive data for research purposes**
Lead Beneficiary: **ECRIN, BBMRI**
Date of publication: **7 December 2020**

Grant agreement no. 824087
Horizon 2020
H2020-INFRAEOSC-2018-2
Type of action: RIA

Authors: **C. Ohmann (ECRIN, pilot study coordinator), S. Canham (ECRIN), J.-W. Boiten (EATRIS/Lygature), M. Cano Abadia (BBMRI), G. Chassang (INSERM), M.L. Chiusano (EMBRC), R. David (ERINHA), M.Th. Mayrhofer (BBMRI), L. Pireddu (CRS4-BBMRI)**

Table of Contents

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

Pilot study protocol for the evaluation of the categorisation system

Authors: C. Ohmann (ECRIN, pilot study coordinator), S. Canham (ECRIN), J.-W. Boiten (EATRIS/Lygature), M. Cano Abadia (BBMRI), G. Chassang (INSERM), M.L. Chiusano (EMBRC), R. David (ERINHA), M.Th. Mayrhofer (BBMRI), L. Pireddu (CRS4-BBMRI)

Status: Version 1, final

Date: 7 December 2020

ABSTRACT

The Horizon 2020 project EOSC-Life brings together the 13 Life Science 'ESFRI' research infrastructures to create an open, digital and collaborative space for biological and medical research. Sharing sensitive data is a specific challenge within EOSC-Life. For that purpose, a toolbox is developed, providing information to researchers who wish to share or use sensitive data. An essential component of this toolbox will be a knowledge base of categorised (tagged) resources linked to sensitive data sharing in the research areas of EOSC-Life. The categories are designed to support consistent labelling and categorisation of the stored resources, in terms most relevant to data sharing tasks and activities, so that they are available to users (e.g., as on screen filters) when searching the system. Objective of the pilot study is to evaluate the applicability and appropriateness of the categorisation system, as well as the coverage of the different research areas involved in EOSC-Life. Six research infrastructures will participate in the pilot study. Three infrastructures (BBMRI, EATRIS, ECRIN) will each nominate two experts who will independently assess 25 self-selected resources. The same procedure will be applied by the other infrastructures (EMBRC, ERINHA and Euro-BioImaging) for 10 self-selected resources. The assessments will be documented per infrastructure and a summary statistics of agreements and disagreements between experts will be provided. The results will be used as a basis for an improvement and update of the categorisation system.

Keywords:

Sensitive data, tags, categorisation system, life sciences, toolbox, data sharing, EOSC-Life, pilot study protocol, research infrastructure

1. BACKGROUND

The EOSC-Life Toolbox is intended to provide guidance to:

- Data providers: Researchers or public/private institutions that make their sensitive data available for future reuse (enabling future sharing of data);
- Data users: Researchers and public/data institutions and stakeholders that wish to make use of sensitive data made available by the data provider (enabling actual sharing of data).

The toolbox will consist of references to recommendations, procedures, best practices, and links to software (tools) to support these data sharing and reuse operations. The toolbox will not be designed *de novo*; instead, it will help scientists to navigate to previously collected high-quality content available throughout our collective infrastructure landscape (1).

An essential component of this toolbox will be a knowledge base of categorised (tagged) resources linked to data sharing in the research areas of EOSC-Life (2). The categories are designed to support consistent labelling and categorisation of the stored resources, in terms most relevant to data sharing tasks and activities, so that they are available to users (e.g. as on screen filters) when searching the system. They are designed to be used in conjunction with traditional text-based searching methods, e.g. of the resources' titles.

2. OBJECTIVES

Aim of the pilot study is to evaluate the categorisation system with respect to:

- Practical applicability of the categorisation system
- Coverage of the different research areas involved in EOSC-Life (e.g. health research, zoology, microbiology)
- Applicability and appropriateness of the structure and answer categories of the categorisation system
- Consistency check of the assessment results by independent assessors (interobserver variability)
- Flexibility and accuracy of the categorisation system

The results of the pilot study will be used to test the meaningfulness and to improve and update the categorisation system.

3. PROCEDURE

The pilot study will be performed within EOSC-Life WP4, supported by the partners of this WP and coordinated by ECRIN. This will be done according to the following procedure:

- Each involved infrastructures/institution will nominate 2 experts, willing to perform the assessment. The infrastructures/institutions are free to select the experts. The experts should

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

have adequate expertise to perform the categorisation.

Infrastructure/ institution	Number of resources to be assessed	Number of experts involved
BBMRI	25	2
EATRIS	25	2
ECRIN	25	2
EMBRC	10	2
ERINHA	10	2
Euro-BioImaging	10	2

Table 1: Number of resources and experts involved in the assessment

- The experts from the involved infrastructures/institutions will select the given number of resources by themselves, spanning the range of resource types (e.g. legislation & regulations, position papers, policies and principles, background & explanatory material, recommended practice, systems, tools & services, repositories or other resources). It is proposed that the infrastructures/institutions with more person-months (BBMRI, EATRIS, ECRIN) select 25 resources and the others (EMBRC, ERINYA, UNITO, MU) 10 resources. The selection of resources will be coordinated by the pilot study coordinator.
- The list of resources selected by the experts will be sent to the pilot study coordinator, who will produce a separate Zotero private library of the resources for each participating infrastructure / institution. The library will contain a metadata description of the resource and a link to the pdf, if available, or a URL. The Zotero private libraries are prepared in a way that the categories of the categorisation system are included in the list of possible tags of Zotero, which later can be selected for assessment of resources. Links to the private libraries are sent to the experts.
- The experts of the involved infrastructures / institutions assess the resources independently of each other. This is done via the tags of Zotero. It is important that all dimensions of the categorisation system are covered for a resource. Multiple tags per dimension are possible. Details are explained in the document, describing the categorisation system (2). If ready, the private Zotero libraries are exported and sent to the pilot study coordinator together with comments, if necessary.

Before the start of the pilot study a telephone conference is performed, where the details of the pilot study are explained and an introduction into the use of Zotero is given.

It is planned to pre-publish the protocol of the pilot study in OSF in order to be able to support a possible publication of the results in a journal with adequate impact.

4. ANALYSIS

The data from the private Zotero libraries will be transferred to spreadsheets by the pilot study coordinator. For each participating infrastructure/institution a separate sheet will be provided (see table 2).

Re-source	**Expert 1**	**Expert 2**	**Consen-sus**	**……**	**……..**	**Expert 1**	**Expert 2**	**Consen-sus**
	Cat. 1	Cat. 1				Cat. 8	Cat. 8	
Res. 1								
- KW 11	best practice	guidance	no			data steward	data steward	yes
- KW 12						data provider	-----	no
…..						------	data consumer	no
Res. 2								
……								
Res. 25								

Table 2: Documentation of individual results per infrastructure
Res. = Resource, KW = keyword, Cat. = Category
Cat. 1: Resource type, Cat. 8: Targeted group

A summary statistics of agreements/disagreements between experts will be provided and the results will be documented in Table 3.

Category	**Infrastructure**	**Agreement between experts of an infrastructure**	**Comment**

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

	(no. of resources assessed)	**Yes N (%)**	**Partial N (%)**	**No N (%)**	
Resource type	A (25)				
	B (25)				
	C (25)				
	D (10)				
	E (10)				
	Total (95)				
Resource field	A (25)				
	B (25)				
	C (25)				
	D (10)				
	E (10)				
	Total (95)				
Resource design	A (25)				
	B (25)				
	C (25)				
	D (10)				
	E (10)				

	Total (95)				
Data type	A (25)				
	B (25)				
	C (25)				
	D (10)				
	E (10)				
	Total (95)				
Stage in DS life cycle	A (25)				
	B (25)				
	C (25)				
	D (10)				
	E (10)				
	Total (95)				
Geographical scope	A (25)				
	B (25)				

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

	C (25)				
	D (10)				
	E (10)				
	Total (95)				
Specific topics	A (25)				
	B (25)				
	C (25)				
	D (10)				
	E (10)				
	Total (95)				
Targeted group	A (25)				
	B (25)				
	C (25)				
	D (10)				
	E (10)				
	Total (95)				

Table 3: **Consistency in the assessment of the resources by experts**
(Percentage of agreements per infrastructure and per category)

In addition, the number of terms allocated per category by each expert will be determined and the results will be summarised in Table 4. This analysis should give an indication for which categories usually a single answer is adequate and sufficient and where multiple allocations have to be expected.

Ex-pert	**Category**							
	Re-source type **Median (range)**	**Re-source field** **Median (range)**	**Re-source design** **Median (range)**	**Data type** **Median (range)**	**Stage in DS life cycle** **Median (range)**	**Geo-graph. scope** **Median (range)**	**Specific topics** **Median (range)**	**Target-ed group** **Median (range)**
A1								
A2								
B1								
B2								
C1								
C2								
D1								
D2								
E1								
E2								

Table 4: Number of terms allocated per category by each expert (median(range))

The results are reported back to the experts with the target to get consensus for disagreements (if possible). The result of the consensus exercise and the issues encountered are reported back to the coordinator of the pilot study.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

A report of the pilot study results, including the consensus exercise, is prepared and will be used as a basis for an improvement and update of the categorisation system. This version will be discussed in the categorisation subgroup and finalised within EOSC-Life WP 4.

5. CONCLUSIONS

The pilot study evaluates version 1 of the categorisation system. After this pilot, it is needed to envision a test for keeping the toolbox accurate and up to date. Therefore, the value of the categorisation system should not be restricted to the pilot study but it should also be explored for its sustainability. The issue about sustainability should be addressed already in the design phase of the tool. Somebody has to be responsible for potential updates or adaptations of the tagging system and the effects on already assessed resources as well as to ensure the validity of the assessment of resources through regular inspections and update (e.g. document becomes obsolete and has to be removed). Adequate resources and reliable agreements are necessary to ensure the involvement of experts on a regular basis.

6. REFERENCES

1. Boiten JW (EATRIS), Ohmann, C (ECRIN), Ludwig R (EATRIS) Pla AS (EATRIS), Fratelli M (EATRIS), Matei M (ECRIN), Panagiotopoulou M (ECRIN), Mayrhofer M (BBMRI), Holub P (BBMRI), Schlünder I (BBMRI), Adeniran A (BBMRI), Abadia MC (BBMRI), Chassang G (BBMRI), Merchant A (ELIXIR), David R (ERINHA), Chiusano ML (EMBRC), Tsamis G (EMBRC), Gribbon P (WP1), Pireddu L (WP2), David R (WP6), Wagener (WP7), Ludwig R (WP9): Toolbox for data sharing of sensitive data - a concept. https://drive.google.com/drive/folders/12oogU-UC5r0G-aK8QIyu9sAUXLsH-P8F

2. Ohmann, Christian, Canham, Steve, Boiten, Jan-Willem, Cano Abadía, Mónica, Chassang, Gautier, Chiusano, Marie Luisa, … Pireddu, Luca. (2020, December 8). EOSC-Life WP4 Toolbox: Categorisation system for resources to be referenced in the toolbox for sharing of sensitive data (Version 1). Zenodo. http://doi.org/10.5281/zenodo.4311094

7. **APPENDIX:** CATEGORISATION SYSTEM (SHORT FORM, VERSION 1)

EOSC-LIFE WP4 toolbox: Categorisation system (short form, version 1, final, 30 November 2020)							
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Resource type	**Research field**	**Research design**	**Data type**	**Stage in DS life cycle**	**Geographical scope**	**Specific topics**	**Perspective**
• Legislation/ regulation • Position papers/ policies/ principles • Best practice • Guidance/ recommend-dations • Systems/tool/ services • Repositories/ other infra-structures • Other resource type • Not applicable	• Health research • Pre- clinical research • General life sciences • Plant sciences • Zoology • Marine/ water biology • Microbiology • Ecology • Other • Not specified/ not clear	• Experimental/ interventional • Observational research • Secondary research • Modelling research • Other research designs • Not specified/ not clear • Not applicable	**A. Data from/about living humans** • Real world/ routine health data • Clinical research data • Biobank/registry data • Human population-level health/ socio-economic data • Data including images of humans • Genetic and molecular biology data **B. No data from/about identifiable human beings** • Omics data generated by basic research • Pre.clinical research data	• Preparation for DS • Planning for DS • Data preparation at the end of the study • Transfer of data to repository • Managing data access • Access to data for re-use • Publication of re-use • Monitoring DS • Discovering the data • Any • Other • Not applicable	• Local • Global • Continental • Region in the world • National • Sub-national • Not applicable -	• Legal aspects • GDPR • Data transfer agreement • Data use agreement • Data storage agreement • Broad consent • Informed consent • Alternatives to consent • Ethics of data sharing • Planning for data re-use • Data governance • Data access committee • Metadata for data sharing • Attribution/credit for DS • Anonymisation • Pseudonymisation	• Resource funder • Policy maker • Coordination forum • Standardisation body • Research communities • Data service providers • Data steward • Data provider • Data consumer • Other group • Cannot be specified • Not applicable

			• Organism or species specific data • Eological/ environmental data • Other biological data • Non-personal sensitive data *(additional tag)* **C. Other data** • Other type of data • Not specified/not clear • Not applicable			• De-identification • Repository quality • Managing data access • Technical & organisational control measures • Other topics	