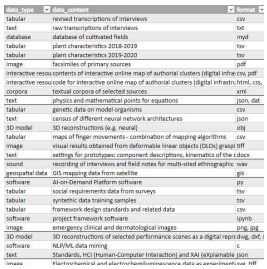
Dec 11, 2024 Version 2

Mapping Research Data at the University of Bologna: Protocol V.2



DOI
dx.doi.org/10.17504/protocols.io.n2bvj87jpgk5/v2

Sara Coppini[1], Bianca Gualandi[1], Giulia Caldoni[1], Mario Marino[1], Silvio Peroni[1], francesca.masini[1]

[1]University of Bologna

Censimento dati DMP

Sara Coppini
University of Bologna

OPEN ACCESS

DOI: dx.doi.org/10.17504/protocols.io.n2bvj87jpgk5/v2

Protocol Citation: Sara Coppini, Bianca Gualandi, Giulia Caldoni, Mario Marino, Silvio Peroni, francesca.masini 2024. Mapping Research Data at the University of Bologna: Protocol. **protocols.io https://dx.doi.org/10.17504/protocols.io.n2bvj87jpgk5/v2**Version created by **Sara Coppini**

Protocol status: Working
We use this protocol and it's working

Created: July 18, 2023

Last Modified: December 11, 2024

Protocol Integer ID: 85147

Keywords: research data management, data stewardship, data management plan, research data, data types

Abstract

Led by data stewards at the University of Bologna, this protocol was developed within an analysis of research data generated and managed within the institution with respect to the differences and commonalities between disciplines and potential challenges for institutional data support services and infrastructures. We are primarily mapping the type (e.g., image), content (e.g., scan of a manuscript) and format (e.g., .tiff) of managed data, thus sustaining the value of FAIR data as granular resources.

The analysis is based on data management plans (DMPs) produced by grantees of Horizon Europe and Horizon 2020 funding who are affiliated to the University of Bologna and are either project coordinators or partners in charge of the DMP. We are including in the study only the DMPs shared with us between May 2022 (when the team was created) and October 2023.

In short, we have selected variables of interest to be headers of a table that is progressively filled with information garnered through a close reading of the DMPs. Computational analysis (R version 4.2.2) on the collected data will produce graphs showing composition, relationship (bar graphs, pie charts and alluvial/sankey charts) and incidences (waterfall graph) of the different variables. The data and the software used will be published openly.

Guidelines

In this research project, we will analyse the data management plans (**DMPs**) produced by researchers affiliated to the University of Bologna (**UniBo**) who are taking part to European competitive projects (i.e. within **Horizon 2020 or Horizon Europe programmes**). These funding programmes require researchers to submit a **DMP within 6 months** from the beginning of the project (M6), but may also either require or suggest the update the DMP throughout the life of the project. Most of the documents we analyse are initial DMPs (produced within M6), but occasionally they may be updated to reflect a more advanced stage of the research project.

Here, we consider only those competitive projects in which **UniBO is either the coordinator or the partner responsible for the DMP as a deliverable**. Indeed, in both these scenarios, researchers can take advantage of our support as Data Stewards in managing their research data and drafting the DMP. This study analyses the DMPs that have seen the involvement of our Data Steward group, from its creation in May 2022 to October 31, 2023.

Our focus is on **digital data**, but within the complementary research we also take into account **non-digital outputs** in order to draw some considerations on these as well, but not so that they are analysed with the same categories as we have chosen for digital data. In the future, if the work is extended by integrating data from DMPs of non-Horizon Europe projects, we may consider changing the terminology used to one that is **more inclusive and more general, such as "digital research object"** instead of "data."

We consider **both newly generated data within the project and reused data** (which may also be mentioned in the DMP). However, in data analysis the focus is on newly generated data.

While analysis and application of this protocol will be on only DMPs, to develop this protocol and define a methodology for this research, we have **analysed a limited number of DMPs and of Grant Agreements** (documents regulating the administrative and financial aspects of EU-funded projects and describing in detail the planned research activities) in order to identify the type of information we want to collect, i.e. our variables of interest.

Please note: Two different categories of data can exist within the same dataset, e.g., a dataset collecting data about an interview may contain both a README file documenting the data (which we do not consider in this work), the *audio file* of the recorded interview and the *text file* of the transcript. The latter are two different components of the dataset and thus must be described separately.

We have defined **new** taxonomies when necessary, i.e., when we have not found any that adhere to our type of investigation in terms of purpose and method (e.g., "reasons of inaccessibility"). **Existing** taxonomies, either generalist (e.g., DataCite, **MIUR settori scientifico disciplinari**) or institutional (e.g., UniBO taxonomy for the 5 subject areas of academic research) have been reused when appropriate. We will expand these initial taxonomies or (occasionally) make changes to them if new typologies of data or other aspects not previously considered will emerge during the analysis.

For the field **"data type"** we reused the taxonomy proposed by DataCite, specifically we reused some of the controlled values for the element 10.a resourceTypeGeneral (**http://purl.org/dc/terms/**). We reused those in line with the definition of data chosen in this work, so we selected: Audiovisual, ComputationalNotebook, Image, InteractiveResource, Report, Software, Sound, Standard, Text, Workflow, Other, Model, Tabular. The latter has been renamed thus by us, whereas in the

original scheme it would be 'dataset'. We made this renaming choice because we found the term "dataset" confusing, since we already use it in the sense of "set or collection of data" (as it is understood in the DMPs that are the subject of our analysis) and since the datacite definition for "dataset" corresponds to the concept of "tabular data" or "structured data" (cf. "Data encoded in a defined structure", https://datacite-metadata-schema.readthedocs.io/en/4.5_draft/appendices/appendix_1/resourceTypeGeneral.html#dataset).

For some fields, the possible values are those of the UniBO taxonomies.

- **"creator's unit"** and **"project unit":** we used the departments in UniBO ("nd" when data is new but creator's name is yet to be defined, and "ext" for "external" when data is reused and thus created by a person external to unibo, since we are not interested in tracking that information if UniBO is not involved in the data reuse or generation)
- "**subject area**" for which we considered the 5 areas of research as defined by UniBo: Arts, Humanities, and Cultural Heritage (shortened: Humanities); Science; Economics and Management (shortened: Economics); Engineering; Medicine.

Departments of UniBO are: DISTAL (Agricultural and Food Sciences); DA (Architecture); BiGeA (Biological, Geological, and Environmental Sciences); DIBINEM (Biomedical and Neuromotor Sciences); CHIM (Chemistry "Giacomo Ciamician"); DICAM (Civil, Chemical, Environmental, and Materials Engineering); FICLIT (Classical Philology and Italian Studies); DISI (Computer Science and Engineering); DBC (Cultural Heritage); DSE (Economics), EDU (Education Studies "Giovanni Maria Bertin"); DEI (Electrical, Electronic, and Information Engineering "Guglielmo Marconi"); QUVI (Life Quality Studies); DiSCi (History and Cultures); CHIMIND (Industrial Chemistry "Toso Montanari"); DIN (Industrial Engineering); DIT (Interpreting and Translation); DSG (Legal Studies); DiSA (Management); MAT (Mathematics); DIMEC (Medical and Surgical Sciences); LILEC (Modern Languages, Literatures, and Cultures); FaBiT (Pharmacy and Biotechnology); FILCOM (Philosophy and Communication Studies); DIFA (Physics and Astronomy “Augusto Righi”); SPS (Political and Social Sciences); PSI (Psychology "Renzo Canestrari"); SDE (Sociology and Business Law); STAT (Statistical Sciences "Paolo Fortunati"); DAR (The Arts); DIMEVET (Veterinary Medical Sciences).

This protocol is related to:
1) the **dataset cited and descripted, deposited in Zenodo** (https://doi.org/10.5281/zenodo.14234555)
2) the **code used to run the computational analysis, also deposited in Zenodo** (https://doi.org/10.5281/zenodo.14234555)
3) the **DMP that documents the output management and the research worflow** (https://doi.org/10.5281/zenodo.14385803)

Before start

Before reusing this methodology, choose:

1. **What do you mean by 'data',** i.e. the object of analysis of this research as described in the data management plans on which it is based. We have chosen to consider "data" **all research outputs that are digital** (thus excluding physical and intangible research outputs) distinct from publications. This choice comes from the source materials on which the research is elaborated: DMPs of EU competitive projects.
2. **Which taxonomies to use to define the possible values of the fields/variables of the analysis**. We tried to reuse generalist and existing taxonomies whenever possible, but for three fields (creator unit, associated project unit, subject area) we chose to consider taxonomies defined for UniBO (list of departments and disciplinary areas of research).
3. **A computational analysis tool**. We chose R in the **4.2.2 version.**

For more information on the choices we made, please see section "guidelines".

Data collection

1 Using the DMPs and GAs of European projects as input, we structured the table in which to collect data information with the following **variables or fields** and their meaning or accepted values:

- **Project identifier** (*project_id*): alphanumeric string to identify the project to which the described data belong. Three-digit sequential numbering, independent of the other two identification fields
- **Dataset identifier** (*dataset_id*): alphanumeric string to identify the dataset to which the described data belong. Three-digit sequential numbering, independent of the other two identification fields
- **Entry identifier** (*entry_id*): alphanumeric string to identify the data category (i.e., file) described in the current row. Three-digit sequential numbering, independent of the other two identification fields
- **Creator's unit** (*creator_unit*): research unit (department, centre, etc.) of the person who created or reused or contributed to the dataset (values also accepted: "nd" when data is new but creator's name is yet to be defined, and "ext" for "external" when data is reused and thus created by a person external to unibo) - multiple values are accepted, when there are multiple creators from different known or unknown research units
- **Creator's SSD** (*creators_ssd*): disciplinary scientific sector ("settore scientifico disciplinare") of the person who created or reused or contributed to the dataset (values also accepted: "nd" when data is new but creator's name is yet to be defined, and "ext" for "external" when data is reused and thus created by a person external to unibo) - multiple values are accepted, when there are multiple creators from different known or unknown research units
- **Principal Investigator's SSD** (*pi_ssd*): disciplinary scientific sector ("settore scientifico disciplinare") of the principal investigator of the project
- **Project unit** (*project_unit*): research unit (department, centre, etc.) of the principal investigator of the project
- **Project programme** (*project_programme*): HE (Horizon Europe); H2020 (Horizon 2020)
- **Project type** (*project_type*): individual; consortium
- **Subject area** (*subject_area*): disciplinary or thematic area to which the project belongs
- **Month DMP is delivered** (*month_dmp*): e.g., M6 (sixth month), M12 (twelfth month), etc.
- **Public DMP** (*public_dmp*): 1 (True), 0 (False) - (ND is also accepted)
- **Data type** (*data_type*): typology of the data on a formal level, e.g. image - (ND is also accepted)
- **Data content** (*data_content*): (categorization of the data at the content level, and not on a content level, e.g., scanned image of a medieval manuscript) values are free-text descriptions
- **Format** (*format*)**:** refers to the format and specifically the extension (if there is more than one per data, they can all be entered separated by commas, without putting the dot before the extension name) - (ND is also accepted)

- **New data** (*new_data*)**:** 1 (True), 0 (False) - (ND is also accepted)
- **Contains personal data** (*personal_data*): 1 (True), 0 (False) - (ND is also accepted)
- **Personal data management strategy** (*p_d_strategy*): anonymization, pseudonymization, no strategy, consent to publish - (ND is also accepted)
- **Level of access** (*access*):

1. open (CC0, CC BY, CC BY-SA or equivalent - according to the **open definition** - equivalent for software are: MIT, Apache, MPL),
2. accessible under conditions (CC BY-NC, CC BY-NC-SA, CC BY-ND, CC BY-NC-ND or equivalent, e.g., for software: GPL, AGPL, JRL, AFPL),
3. controlled (authors or committee control access to the dataset),
4. embargoed,
5. unlicensed (when it's freely available but not "open" according to the open definition, so the data are not assigned a license),
6. unfiled (not deposited),
7. unknown (there is no information concerning accessibility and licesing of data)

- **Reason of inaccessibility** (*reason_inaccess*)**:** excessive size (therefore technical motivation), ethical issues, privacy, IPR (Intellectual Property Rights issues) - (ND is also accepted)
- **Size** (*size*): orders of magnitude for digital data (Bytes, KB, MB, GB, TB, PB, EB, ZB, YB) - (ND is also accepted)
- **Standard** (*standard*): name of standards used to organise and structure data (e.g., vocabularies, ontologies, taxonomies etc.) - (ND is also accepted)
- **Deposited** (*deposited*): 1 (True), 0 (False) - (ND is also accepted)
- **Chosen repository** (*chosen_repo*): alphanumeric string for the name of repository chosen by researchers to deposit data, as it is official stated in re3data.org - can be more than one, separated by commas - (ND is also accepted)
- **PID** (*pid*): 1 (True), 0 (False)
- **Associated publication** (*associated_pub*): 1 (True), 0 (False) - (ND is also accepted)
- **Notes** (*notes*): general notes concerning other unclassified issues

2 The variables identified form the header of a table, which is then filled in with the information from the DMPs, then formalised in a CSV file.
Here is a sample of 13 rows of a fictional table developed to test the first version of the code (the data it contains is fictitious):

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
	project_id	**dataset_id**	**entry_id**	**creator_unit**	**creators_ssd**	**pi_ssd**	**project_unit**	**project_prog**	**project_type**	**subject_area**	**month_DMP**	**public_DMP**	**data_type**	**data_content**	**format**	**new_data**	**personal_data**	**p_d_strategy**	**access**	**reason_inacc**	**size**	**standard**	**deposited**	**chosen_repo**	**pid**	associated				

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
								framme									**ata**	**gy**		**cc ess**							_pub				
	p_001	p_001_001	p_001_001_001	DISTATAL	AGR/01;AGR/01;AGR/01	AGR/02	DISTATAL	H2020	consortium	SocialSciences	M12	1	tabular	revisedtranscriptionsofinterviews	csv	1	0	nd	open	nd	KB	nd	0	nd	0	0				
	p_001	p_001_001	p_001_001_002	DISTATAL,nd	AGR/01;nd	AGR/02	DISTATAL	H2020	consortium	SocialSciences	M12	1	text	rawtranscriptionsofinterviews	txt	1	0	nd	open	nd	KB	nd	0	nd	0	0				
	p_001	p_001_002	p_001_002_0	STAT	MAT/01	AGR/02	DISTATAL	H2020	consortium	SocialSciences	M12	1	tabular	databaseofculti	myd	1	0	nd	embargoed	IPR	GB	nd	0	nd	0	0				

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
			01											vated fields																
	p_001	p_001_003	p_001_003_001	BIGeA	CHIM/02	AGR/02	DISTAL	H2020	consortium	SocialSciences	M12	1	tabular	plantcharacteristics2018-2019	tsv	1	0	nd	open	nd	MB	ISO345	1	Zenodo	1	1				
	p_001	p_001_003	p_001_003_002	DISTAL	AGR/02	AGR/02	DISTAL	H2020	consortium	SocialSciences	M12	1	tabular	plantcharacteristics2019-20200	tsv	1	0	nd	open	nd	MB	nd	1	Zenodo	1	1				
	p_002	p_002_001	p_002_001_	nd	nd	L-ART/02	FICLIT	HE	individual	Humanities	M6	0	image	facsimilesofpr	pdf	0	0	nd	controlled	IPR	GB	nd	0	nd	0	0				

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
			001											primary sources																
	p_002	p_002_002	p_002_002_001	FICLIT	M-DEA/01	L-ART/02	FICLIT	HE	individual	Humanities	M6	0	interactive resource	contents of interactive online map of authorial clusters (digital infrastructure)	csv, pdf	1	0	nd	open	nd	KB	nd	1	AMSActa	1	0				

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
	p_002	p_002_002	p_002_002_002	FICLIT	M-DEA/01	L-ART/02	FICLIT	HE	individual	Humanities	M6	0	interactive resource	code for interactive online map of authorial clusters (digital infrastructure)	html, css, js	1	0	nd	open	nd	MB	nd	1	AMSActa	1	0				
	p_002	p_002_003	p_002_003_001	FILCOM	M-GGR/01	L-ART/02	FICLIT	HE	individual	Humanities	M6	0	text	textual corpora of selecte	xml	1	0	nd	open	nd	MB	nd	0	ILC-CNR for CLARIN	0	0				

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
														dsources										-IT						
	p_003	p_003_001	p_003_001_001	DIFA	FIS/06	FIS/06	DIFA	HE	individual	Science	M6	1	text	physicsandmathematicalpointsforequations	json,dat	1	0	nd	open	nd	KB	nd	0	nd	0	0				
	p_004	p_004_001	p_004_001_001	nd	nd	FIS/05	DIMEC	H2020	consortium	Medicine	M6	0	tabular	geneticdataonmodelorganisms	csv	1	0	nd	open	nd	MB	nd	0	Openneuro,Neuromorpho	0	1				

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
	p_004	p_004_002	p_004_002_001	nd	nd	FIS/05	DIMEC	H2020	consortium	Medicine	M6	0	text	censusofdifferentneuralnetworkarchitectures	json	1	0	nd	open	nd	MB	nd	0	nd	0	0				

Data analysis

3 With the tabular data structured within the data collection phase as input, the data analysis will be descriptive statistics to investigate various research questions.

Set up instructions:

Remember to set your work directory before start, for instance:

```
setwd("C:/Users/john.doe/Desktop/cens/sim")
```

Remember to install all library/packages before run them

```
library(dplyr)
library(tidyr)
library(plotly)
library(ggplot2)
```

Import data, in Italy the default separator is ";" so we use "read.csv2". For instance:

```
data <-
read.csv2("C:/Users/john.doe/Desktop/cens/sim/20240710_cens.csv")
```

Define custom colors for the bars, for instance:

```
custom_colors <- c("lightblue", "lightgreen", "lightcoral",
                   "lightsalmon")
```

4 RESEARCH QUESTION **1: Types of data, re-used data *vs* new and data formats: an overview**

Related analysis are:

1.1 What types of data are managed by researchers at the University of Bologna?

1.1.1 What are the most popular types of data?
Group the data by data type and count occurrences across datasets
Create a bar plot of appearances of each data format and display histogram with ggplot() and geom_bar()
You can sort the data by Appearances column in descending order

1.1.2 How often do we find different data types in the same dataset?
Find number of data type unique for each dataset
Plotting with ggplot() and geom_bar()

1.1.3 How often do we find different data types in the same project?

Find number of data type unique for each dataset
Plotting with ggplot() and geom_bar()

1.1.4 How are data types distributed across single-beneficiary and collaborative projects?
Filter the data by project_type == "individual" or "consortium"
Aggregate the filtered data
Rename columns for better interpretation
Create and display the plot with ggplot() and geom_bar()

1.1.5 How are data types distributed across subject areas?
Calculate the count of each data type within each subject area (aggregate)
Rename columns for better interpretation
Create and display the plot using ggplot()

1.2: How many data entries include re-used data in the DMP and what is the ratio of new to re-used data?
Compute the ratio as ratio <- new_count / reused_count
Create the stacked bar chart with ggplot() and geom_bar()
Add the ratio as a text annotation on the plot with annotate()
Display the plot

1.3: How many projects have already made decisions about formats and have standard and open formats been chosen?
1.3.1 Are the formats precisely defined at the month 6 DMP?
Filter the data for DMP at month 6 with subset()
Calculate the percentage of formats that have been precisely defined with sum() and nrow()
Create a data frame to attempt a waterfall chart with data.frame()
Calculate cumulative values
Create and display the chart with ggplot() and geom_bar()

1.3.2 Are they standard and open formats?
Use library(wordcloud) and library(wordcloud2)
Split the 'format' column into multiple columns using the comma as a separator
Gather all the formats into a single column
Count the occurrences of each format
Create and display the word cloud

5 **RESEARCH QUESTION 2: what could be trends of problems and patterns useful to improve the Data Stewardship service?**

Related analysis are:

2.1.1 *How many projects involve treatment of* ***personal data****?*

Count with table()

Create and display plot with barplot()

2.1.2 *How many projects choose to* ***anonymise*** *data and publish them?*

Create table of anonymization counts with table() and sort()

Create and display bar plot with barplot()

2.1.3 *Which personal data management* ***strategies*** *are preferred? (no graph, only percentages are calculated)*

2.2.1 *How many datasets are kept* ***closed*** *and what are the main* ***reasons****?*

Use library(networkD3)

Filter out "unknown" entries from reason_inaccess

Count occurrences of access and reason_inaccess pairs

Create nodes and links for Sankey diagram with data.frame()

Create and display Sankey plot with sankeyNetwork()

2.3.1 *Is data* ***size*** *a recurrent issue in choosing data repository?*

Count with table()

Create bar plot with barplot()

2.4.1 *Which* ***repositories*** *are the most popular among researchers?*

Count with table()

Create bar plot with barplot()

2.5.1 *How many researchers make their* ***DMP public****?*

make it for DMP with (filter project_id)

Aggregate by public DMP and project ID with table()

Create and display plot with barplot()

2.6.1 *What is the rate of* ***projects*** *using at least one* ***standard****? (no graph, just counting since number of cases is low)*

6 **RESEARCH QUESTION 3: is there interdisciplinarity in data production at UniBO?**
We consider only the data produced by UniBO, hence the rows of the table where the value of "new" is 1.

Related analysis are:
How often does the ***principal investigator's SSD*** *coincide with the* ***creator's SSD****?*
How often ***creators with different SSDs*** *collaborated in generating a dataset?*
Is there interdisciplinarity between the ***types of data*** *produced by the various units, or is the type of data produced strictly related to the* ***subject area****?*
Is there more interdisciplinarity in ***single-beneficiary projects or in collaborative projects?***

For this research question, it is good to focus on two theoretical premises:
1) difficulty in finding an unambiguous and established definition of interdisciplinarity
2) in the sample collected, there are few cases of interdisciplinarity

For these reasons, there are no graphic visualisations for this research question, as it is primarily a qualitative analysis.

Data publication

7 The results will be organized in a CSV file and the code developed for analysis will then be deposited in an appropriate data repository and will be accompanied by accurate documentation, e.g., a README file specifying meaning of fields and values.
We also expect to be able to publish an article on this subject in a suitable journal.

Protocol references

Borghi, J., & Gulick, A. V. (2022). Promoting Open Science Through Research Data Management. *Harvard Data Science Review*, *4*(3). https://doi.org/10.1162/99608f92.9497f68e

Coyle, K. (2022). Works, Expressions, Manifestations, Items: An Ontology. *The Code4Lib Journal*, *53*. https://journal.code4lib.org/articles/16491

Fecher, B., Friesike, S., & Hebing, M. (2015). What Drives Academic Data Sharing? *PLOS ONE*, *10*(2), e0118053. https://doi.org/10.1371/journal.pone.0118053

Godin, B., & Gingras, Y. (2000). The place of universities in the system of knowledge production. *Research Policy*, *29*(2), 273–278. https://doi.org/10.1016/S0048-7333(99)00065-7

Habermann, T. (2022, June 24). Universities@DataCite. *Metadata Game Changer*. https://metadatagamechangers.com/blog/2022/6/23/universitiesdatacite

Parland-von Essen, J., Fält, K., Maalick, Z., Alonen, M., & Gonzalez, E. (2018). Supporting FAIR data: Categorization of research data as a tool in data management. *Informaatiotutkimus*, *37*(4). https://doi.org/10.23978/inf.77419