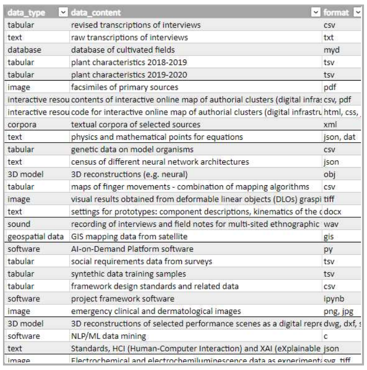
Methodology for DMPs analysis

Sara Coppini[1], bianca.gualandi[1], Giulia Caldoni[1], Mario Marino[1], Silvio Peroni[1], francesca.masini[1]

[1]University of Bologna

Sara Coppini
University of Bologna



data_type	data_content	format
tabular	revised transcriptions of interviews	csv
text	raw transcriptions of interviews	txt
database	database of cultivated fields	myd
tabular	plant characteristics 2018-2019	tsv
tabular	plant characteristics 2019-2020	tsv
image	facsimiles of primary sources	pdf
interactive resou	contents of interactive online map of authorial clusters (digital infras	csv, pdf
interactive resou	code for interactive online map of authorial clusters (digital infrastru	html, css,
corpora	textual corpora of selected sources	xml
text	physics and mathematical points for equations	json, dat
tabular	genetic data on model organisms	csv
text	census of different neural network architectures	json
3D model	3D reconstructions (e.g. neural)	obj
tabular	maps of finger movements - combination of mapping algorithms	csv
image	visual results obtained from deformable linear objects (DLOs) graspi	tiff
text	settings for prototypes: component descriptions, kinematics of the c	docx
sound	recording of interviews and field notes for multi-sited ethnographic	wav
geospatial data	GIS mapping data from satellite	gis
software	AI-on-Demand Platform software	py
tabular	social requirements data from surveys	tsv
tabular	syntethic data training samples	tsv
tabular	framework design standards and related data	csv
software	project framework software	ipynb
image	emergency clinical and dermatological images	png, jpg
3D model	3D reconstructions of selected performance scenes as a digital repre	dwg, dxf, s
software	NLP/ML data mining	c
text	Standards, HCI (Human-Computer Interaction) and XAI (eXplainable	json
image	Electrochemical and electrochemiluminescence data as experimenta	svg, tiff

JUL 14, 2023

DOI:
dx.doi.org/10.17504/protocols.io.n2bvj87jpgk5/v1

Protocol Citation: Sara Coppini, bianca.gualandi, Giulia Caldoni, Mario Marino, Silvio Peroni, francesca.masini 2023. Methodology for DMPs analysis . **protocols.io** https://dx.doi.org/10.17504/protocols.io.n2bvj87jpgk5/v1

Protocol status: In development
We are still developing and optimizing this protocol

Created: May 08, 2023

Last Modified: Jul 14, 2023

PROTOCOL integer ID: 81563

ABSTRACT

All eyes on data: Unleashing the untapped potential of research at the University of Bologna

Led by data stewards at the University of Bologna, this *preliminary 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 (.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 23 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 4.2.3) 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.

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

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 the present day (***October 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 developed **taxonomies** to be used as possible values for each variable. Again, they have been created through the analysis of the DMPs and Grant Agreements. 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) 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. Plus one added by us: "Tabular".

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

- **"creator's unit"** and **"project unit":** we used the departments in UniBO (TBD and NA are also accepted - TBD when data is new but creator's name is yet to be defined, and ND when data is reused since we are not interest in tracking that information if data is not generated by UniBO)
- "**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)

BEFORE START INSTRUCTIONS

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 **4.2.3 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
- **Dataset identifier** (*dataset_id*): alphanumeric string to identify the dataset to which the described data belong
- **Entry identifier** (*entry_id*): alphanumeric string to identify the data category (i.e., file) described in the current row
- **Creator's unit** (*creator_unit*): research unit (department, centre, etc.) of the principal investigator who created or reused the dataset (TBD and NA are also accepted)
- **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)
- **Data type** (*data_type*): typology of the data on a formal level, e.g. image
- **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)

- **New data** (*new_data*)**:** 1 (True), 0 (False)
- **Contains personal data** (*personal_data*): 1 (True), 0 (False)
- **Personal data management strategy** (*p_d_strategy*): anonymization, pseudo-anonymization, no strategy
- **Level of access** (*access*): open (CC BY or equivalent), controlled (CC BY-SA, CC BY-NC or equivalent), embargoed, unfiled
- **Reason of inaccessibility** (*reason_inaccess*)**:** excessive size (therefore technical motivation), ethical issues, privacy, IPR (Intellectual Property Rights) issues
- **Size** (*size*): orders of magnitude for digital data (Bytes, KB, MB, GB, TB, PB, EB, ZB, YB)
- **Deposited** (*deposited*): 1 (True), 0 (False)
- **Chosen repository** (*chosen_repo*): alphanumeric string for the name of repository chosen by researchers to deposit data
- **PID** (*pid*): alphanumeric string - PID of the deposited entry
- **Associated publication** (*associated_pub*): alphanumeric string - PID of the publication associated
- **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
	project_id	dataset_id	entry_id	creator_unit	project_unit	project_programme	project_type	subject_area	month_DMP	public_DMP	data_type	data_content	format	new_data	personal_data	p_d_strategy	access	reason_inaccess	size	deposited	chosen_repo	pid	associated_pub	notes		
	100	110	111	DISTAL	DISTAL	H2020	consortium	Social Sciences	M12	1	tabular	revised transcriptions of interviews	csv	1	0	nd	open	nd	KB	0	nd	nd	nd			

	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
	100	110	112	DISTAL	DISTAL	H2020	consortium	Social Sciences	M12	1	text	raw transcriptions of interviews	txt	1	0	nd	open	nd	KB	0	nd	nd	nd			
	100	120	121	STAT	DISTAL	H2020	consortium	Social Sciences	M12	1	tabular	database of cultivated fields	myd	1	0	nd	embargoed	IPR	GB	0	nd	nd	nd			
	100	130	131	BIGeA	DISTAL	H2020	consortium	Social Sciences	M12	1	tabular	plant characteristics 2018 - 2019	tsv	1	0	nd	open	nd	MB	1	Zenodo	12345678	doi/mlkn123			
	100	130	132	DISTAL	DISTAL	H2020	consortium	Social Sciences	M12	1	tabular	plant characteristics 2019 - 2020	tsv	1	0	nd	open	nd	MB	1	Zenodo	12345679	doi/mlkn123			

	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
	200	210	211	nd	FICLIT	HE	individual	Humanities	M6	0	image	facsimiles of primary sources	pdf	0	0	nd	controlled	IPR	GB	0	nd	nd	nd			
	200	220	221	FICLIT	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	1	AMSActa	amsacta123	nd			

	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
	200	220	222	FICLIT	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	1	AMSActa	amsacta124	nd			
	200	230	231	FILCOM	FICLIT	HE	individual	Humanities	M6	0	text	textual corpora of selected sources	xml	1	0	nd	open	nd	MB	0	ILC-CNR for CLARIN-IT	nd	nd			

	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
	300	310	311	DIFA	DIFA	HE	individual	Science	M6	1	text	physics and mathematical points for equations	json,dat	1	0	nd	open	nd	KB	0	nd	nd	nd			
	400	410	411	tbd	DIMEC	H2020	consortium	Medicine	M6	0	tabular	genetic data on model organisms	csv	1	0	nd	open	nd	MB	0	Openneuro,Neuromorpho	nd	pubmed3456			
	400	420	421	tbd	DIMEC	H2020	consortium	Medicine	M6	0	text	census of different neural network architectures	json	1	0	nd	open	nd	MB	0	nd	nd	nd			

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.
As for the first research question - **what types of data are produced and managed by UniBO** - related analysis are:

1) How often do we find different types of data in the same dataset? Do researchers organise datasets with several files of different formats with similar content or do they prefer the same data type within a single dataset?
2) How much do data types vary within a single project across all datasets produced and reused? How do they vary with respect to the subject area, the project's framework programme or the type of project (monobeneficiary, collaborative)?
3) Are the formats precisely defined at the month 6 DMP? Are they standard and open formats?
4) How many projects include re-used data in the DMP? What is the ratio of new data to re-used data?

To answer these questions, computational analysis will produce graphs showing composition and relationship, thus: bar graphs (with stacks), pie charts and alluvial/sankey charts. On the other hand, incidences are also calculated and represented with a waterfall graph.

4 As for the second research question - **identifying trends of problems and patterns to improve the Data Stewardship service** - related analysis are:

1) how many projects involve treatment of personal data? how many projects choose to anonymise data and publish them? Which personal data management strategies are preferred?
2) how many datasets are kept closed and what are the main reasons?
3) is data size a recurrent issue in choosing data repository? May it require infrastructural adjustments?
4) how many researchers make their DMP public?
5) which kind of repository are the most chosen ones?

To answer these questions, computational analysis will produce graphs showing composition and relationship, thus: bar graphs (with stacks), pie charts and alluvial/sankey charts.

5 As for the third research question - **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:

1) How often does the project department coincide with the dataset creator's department?
2) is there interdisciplinarity between the types of data produced by the various departments, or is the type of data produced strictly related to the subject area?
3) Does the diversity between data types and departments of the creators or the project vary according to the project's framework programme?
4) is there more interdisciplinarity in single-beneficiary projects or in collaborative projects? i.e. how does the relationship between the variables data_type, project_unit and creator_unit change, depending on the type of project (collaborative or single-beneficiary)?

To answer these questions, computational analysis will produce graphs showing composition and relationship, thus: bar graphs (with stacks), pie charts and alluvial/sankey charts.

Data publication

6 The results will be organized in a CSV file and the graphs derived from the analyses saved as image files in non-proprietary open formats. Everything 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.