MOAP Horizon 2020 Database Documentation

As created for the study:

"Monitoring the Open Access Policy of Horizon 2020" (SPECIFIC CONTRACT No RTD/2019/SC/021 implementing Framework contract No 2018/RTD/A2/OP/PP-07001-2018)

PPMI, Athena Research Center, and UNU-MERIT

TABLE OF CONTENTS

LIST OF FIGURES

1. DATABASE DESCRIPTION

DATABASE DESCRIPTION	
Study details	"Monitoring the Open Access Policy of Horizon 2020" (SPECIFIC CONTRACT No RTD/2019/SC/021 implementing Framework contract No 2018/RTD/A2/OP/PP-07001-2018)
Database DOI	10.5281/zenodo.4899767

2. TABLE OF ABBREVIATIONS

TABLE OF ABBREVIATIONS	
EC	European Commission
ORDP	European Commission's open research data pilot
APC	article processing charge
ID	identifier
PID	persistent identifier
PMID	unique identifier used in PubMed[1]
DOI	digital object identifier
CC	creative commons[2]

[1] https://pubmed.ncbi.nlm.nih.gov/
[2] https://creativecommons.org/

3. SCHEMA AND DOCUMENTATION

3.1. Schemata and Table Descriptions

The MOAP Horizon 2020 database was created under the auspices of the Study "Monitoring the Open Access Policy of Horizon 2020" (SPECIFIC CONTRACT No RTD/2019/SC/021 implementing Framework contract No 2018/RTD/A2/OP/PP-07001-2018)

For the purposes of the study, we created a relational database comprised of three schemata:

1. *moap_org* which contains mainly the data from OpenAIRE,[3]
2. *moap_ec* which contains mainly the data coming from the European Commission, and
3. *moap_final* which tables of indicators that lead to the analysis by different facets of interest.

Below we describe the three schemata, the tables the contain and the fields for each table. These have all be deposited as separate CSV files.

3.1.1. moap_org schema

For the purposes of this study, we created a relational database containing the subset of the OpenAIRE research graph[4] that is relevant to the study. We did not copy all the research outcomes in OpenAIRE but only to the publications and datasets that were linked to Horizon 2020 projects The moap_org schema is a relational adaptation of the OpenAIRE Graph, a detailed description of which can be found further below.

The schema of the database is fully normalized and contains one table for each of the main entities of the graph (result, data source, project, organization) and a large number of satellite tables that are used to store the either the multivalued attributes of the main entities or the many-to-many relations between the main entities.

[3] https://www.openaire.eu/
[4] https://graph.openaire.eu/

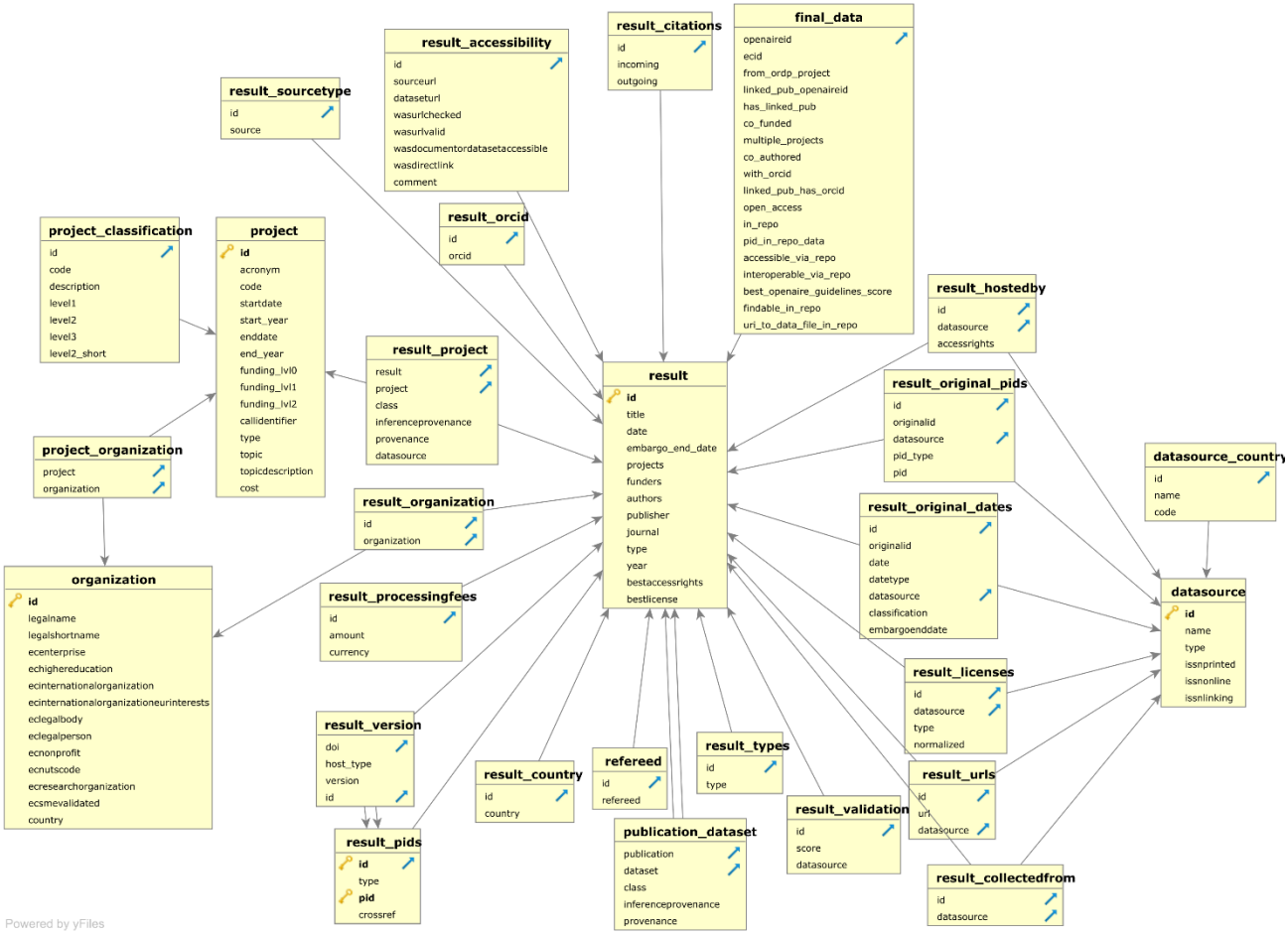

FIGURE 1: TABLES IN moap_org SCHEMA

- **result** Contains the publications and datasets relevant to this study and their main metadata.
- **datasource** Contains information about the data sources (e.g., repositories, journals, aggregators, etc) that either host the results or are the sources OpenAIRE collects the metadata of the results from.[5]
- **datasource_country** Contains the country of each data source.
- **organization** Contains information about the organizations that are: affiliated (through the authors) with a result, participate in a Horizon 2020 project, or manage a data source.
- **project** Contains information on Horizon 2020 projects.

[5] https://www.openaire.eu/category/content-providers

- **project_classification** Contains the Horizon 2020 classification of the projects by programme and the levels above.
- **project_organization** Links the projects to the participating organizations.
- **publication_dataset** Contains relations between publications and datasets.
- **refereed** Contains information on whether the result was peer reviewed or not.
- **result_collectedfrom** Links a result with the data sources it was harvested or collected from.
- **result_country** Contains the countries of the organizations that the result authors are affiliated with.
- **result_hostedby** Links a result with the data sources that host the result. In some cases, the hosting data source is the same that OpenAIRE harvested the metadata from. However, there are cases (e.g., when harvesting another aggregator) where the "hostedby" data source differs from the "collectedfrom" one.
- **result_licenses** Contains the licenses of the results (both the original values collected from the data sources and the cleaned values) and also a link to the data source that hosts the result with the particular license.
- **result_orcid** Contains the list of ORCID identifiers[6] of the authors of the results.
- **result_organization** Links a result to its affiliated organizations.
- **result_original_dates** Contains a list of relevant dates for each result, along with a link to the data source that hosts the result.
- **result_original_pids** Contains a list of all the PIDs for each result, along with a link to the data source that hosts the result.
- **result_pids** Contains a list of all the PIDs for each result, regardless of the data source that hosts the result.
- **result_processingfees** Contains the processing fees (when available) of the results as integrated in OpenAIRE from OpenAPC.[7]
- **result_project** Links a result to its funding project, along with a link to the data source where the link was harvested from.
- **result_sourcetype** Contains information on whether the result was collected from a repository or an open access journal.
- **result_types** Contains the types of the results (e.g., article, preprint, patent, etc.). Note that due to deduplication, a result may have more than one type, one for each merged instance of the result.

[6] https://orcid.org/
[7] https://openapc.net/

- **result_urls** Contains the URLs of the content of the publications and datasets, along with a link to the data source that hosts the result.
- **result_version** Contains information of the version of a document by datasource type and DOI.

Additionally, a number of tables contain the results of the processing performed on the metadata and content of the publications:

- **result_accessibility** Contains the results of the accessibility assessment.
- **result_citations** Contains the number of references and number of citations for each publication
- **result_fos** Contains the results of the classification of publications to scientific fields, according to the Frascati Manual and EuroSciVoc.[8]
- **result_validation** Contains the score of the validation of the metadata of the results against the OpenAIRE guidelines.[9]

3.1.2. moap_ec schema

This schema contains a number of tables containing the data coming from the European Commission as reported in the European Commission's System of Grant Management (SyGMA). Additionally, it contains the results of the accessibility assessment that was performed for both publications and the datasets.

[8] https://www.oecd.org/sti/inno/frascati-manual.htm, https://op.europa.eu/en/web/eu-vocabularies/euroscivoc
[9] https://www.openaire.eu/validator-registration-guide, https://guidelines.openaire.eu/en/latest/

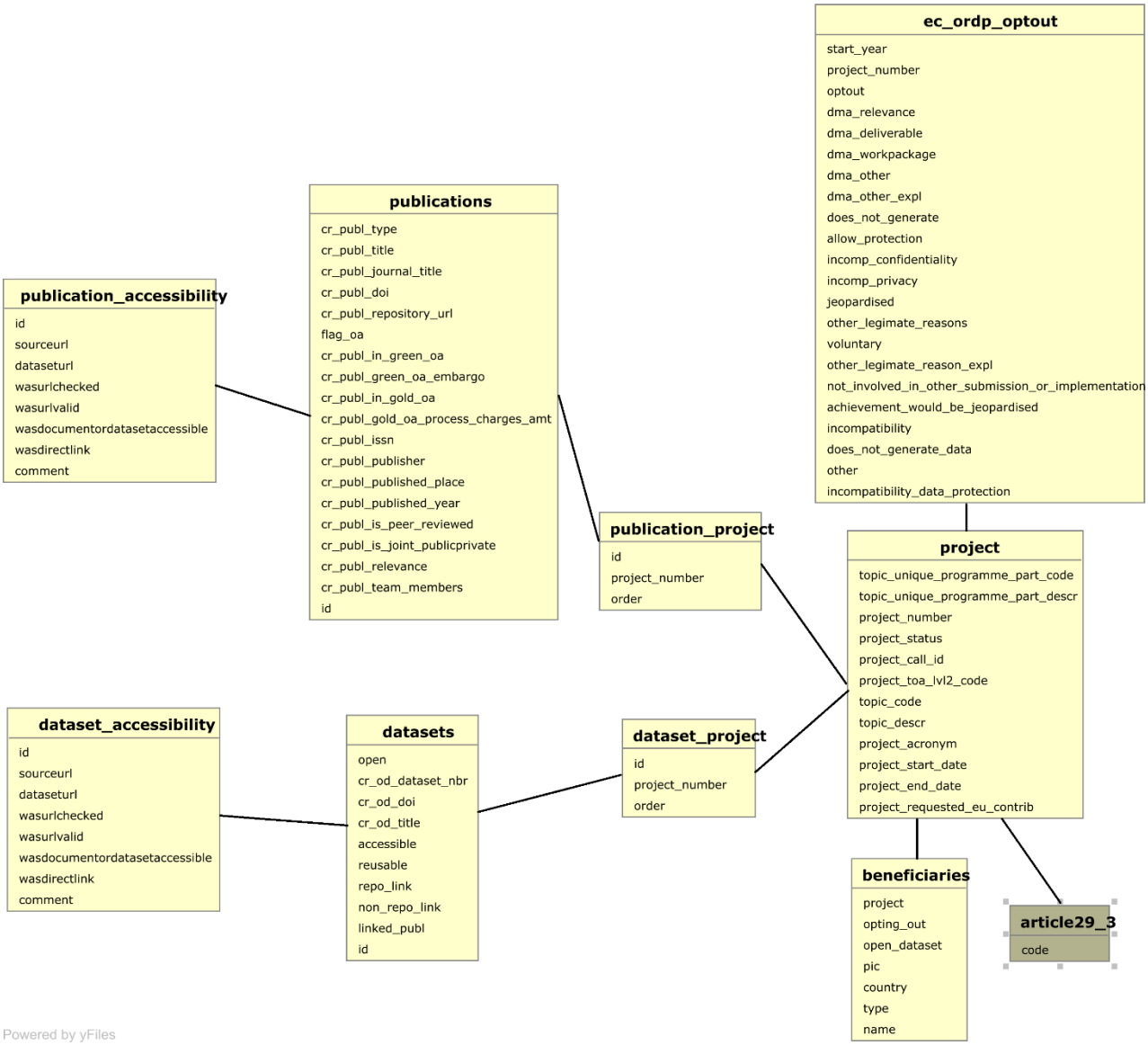

FIGURE 2: TABLES IN moap_ec SCHEMA

One issue we encountered during the creation of the schema was that it included duplicate publications and datasets, with a DOI not present for all of the publications. This fact prevented us from creating a unique identifier for each publication and dataset which in turn prevented us from creating primary and foreign keys in the tables of the schema. As a result, the lines in the previous figure do not represent proper foreign keys but rather an abstract link between the tables.

A brief description of each table in the schema follows:

- **article29_3** Declares whether a project participates in the European Commission's Open Research Data Pilot (ORDP).

- **beneficiaries** Contains a list of beneficiaries for projects.
- **dataset_accessibility** Contains the results of the accessibility assessment for datasets.
- **dataset_project** Links datasets to projects.
- **datasets** Contains the datasets that are linked to Horizon 2020 projects.
- **ec_ordp_optout** Contains information about participation in the ORDP and opt out reasons.
- **project** Contains information on Horizon 2020 projects.
- **publication_accessibility** Contains the results of the accessibility assessment for publications.
- **publication_project** Links publications to projects.
- **publications** Contains the list of publications that are funded by Horizon 2020 projects.

3.1.3. *moap_final* schema

This schema contains a small number of auxiliary tables that were built to support the creation of the final report of the study, most importantly the indicators analysis.

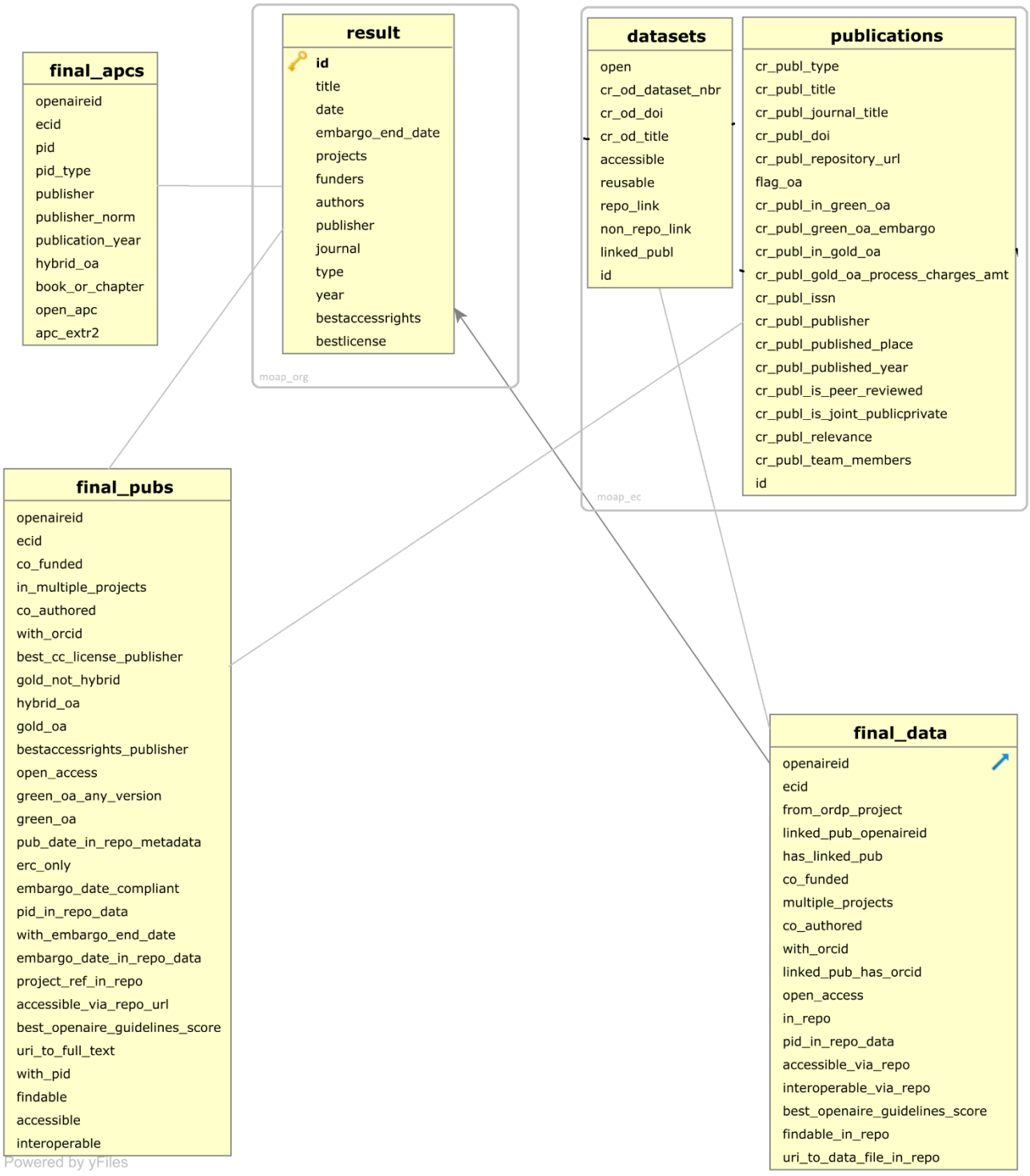

FIGURE 3: TABLES IN moap_final SCHEMA

As was the case for the moap_ec schema, it was not possible to create foreign keys to the moap_ec.publications and moap_ec.datasets tables. As a result, the links between the final_* tables and the European Commission publications and datasets do not represent actual foreign keys in the database.

- **final_apcs** Contains the extrapolated APCs for 'gold' open access publications.
- **final_data** Contains the indicators for Horizon 2020 datasets.
- **final_pubs** Contains the indicators for Horizon 2020 peer-reviewed publications.

3.2. moap_org: per table field descriptions

3.2.1. result

	description	data type	refers to
result			
Contains the publications and datasets relevant to this study and their main metadata			
id	the OpenAIRE generated identifier of the nonpeer reviewed publication or dataset. acts as the primary key of the result	text	
title	the title of the publication	text	
date	the date of publication	datetime	
embargo_end_date	the date that the embargo ends (where applicable)	datetime	
projects	the number of projects that funded the result	integer	
funders	the number of funders that funded the result	integer	
authors	the number of authors of the result	integer	
publisher	the publisher of the result	text	
journal	the journal the result was published in (where applicable)	text	
type	the type of the result (publication/dataset)	text	
year	the year of publication	integer	
bestaccessrights	the most open access rights found for this result	text	
bestlicense	the most permissive license for this result	text	

3.2.2. datasource

	description	data type	refers to
datasource			
contains information about the data sources where the results are either hosted at or OpenAIRE collects the metadata from			
id	the OpenAIRE id of the datasource	text	
name	the name of the datasource	text	
type	the type of the datasource (repository, journal, aggregator, etc.)	text	
issnprinted	if the datasource is a journal, the printed ISSN.	text	
issnonline	if the datasource is a journal, the online ISSN.	text	
issnlinking	if the datasource is a journal, the linking ISSN.	text	

3.2.3. datasource_country

	description	data type	refers to
datasource_country			
contains the countries of the data sources			
id	the OpenAIRE ID of the datasource	text	datasource.id
name	the country name	text	
code	the 2-letter code of the country	text	

3.2.4. organization

	description	data type	refers to
organization			
contains information about the organization that are affiliated with results, participate in projects funding results or manage the data sources hosting the results			
id	the OpenAIRE ID of the organization	text	
legalname	the legal name of the organization	text	
legalshortname	a short name for the organization	text	
country	the country of the organization	text	
ecenterprise (collected from EC)	whether the beneficiary is an enterprise	text	
echighereducation (collected from EC)	whether the beneficiary is a higher education institution	text	
ecinternationalorganization (collected from EC)	whether the beneficiary is an international organization	text	
ecinternationalorganizationeurinterests (collected from EC)	whether the beneficiary is an internal organization with EU interests	text	
eclegalbody (collected from EC)	whether the beneficiary is a legal body	text	
eclegalperson (collected from EC)	whether the beneficiary is a legal person	text	
ecnonprofit (collected from EC)	whether the beneficiary is a non-profit	text	
ecnutscode (collected from EC)	the NUTS code of the beneficiary[10]	text	

[10] https://ec.europa.eu/eurostat/web/nuts/background

ecresearchorganization (collected from EC)	whether the beneficiary is a research organization	text	
ecsmevalidated (collected from EC)	whether the beneficiary is a validated SME	text	

3.2.5. project

	description	data type	refers to
project			
contains information about the projects that funded the results			
id	the OpenAIRE ID of the project	text	
acronym	the acronym of the project	text	
code	the 6-digit grant id of the project	text	
startdate	the starting date of the project	datetime	
start_year	the starting year of the project	integer	
enddate	the ending date of the project	datetime	
end_year	the ending year of the project	integer	
funding_lvl0	the top level of the funding stream (H2020)	text	
funding_lvl1	the second level of the funding stream (ERC, RIA, etc.)	text	
funding_lvl2	the third level of the funding stream (ERC-ADG etc.)	text	
callidentifier	the call identifier for the project	text	
type	the type of the project (innovation action, synergy grant, etc)	text	
topic	the topic of the project	text	
topicdescription	the description of the topic	text	
cost	the total cost of the project	numeric	

3.2.6. project_classification

	description	data type	refers to
project_classification			
contains information about the EC project classification			
id	the OpenAIRE ID of the project	text	
code	the programme code of the project	text	
description	the programme description of the project	text	
level1	the EC pillar (e.g., excellent science)	text	
level2	the EC programme (e.g., ERC)	text	
level3	the EC subprogramme (e.g., supply of non-energy and non-agricultural raw materials) where available	text	
level2_short	a less verbose (shorter) version of level2	text	

3.2.7. project_organization

	description	data type	refers to
project_organization			
contains the relations between projects and participating organizations			
project	the OpenAIRE ID of the project	text	
organization	the OpenAIRE ID of the organization	text	

3.2.8. publication_dataset

	description	data type	foreign key
publication_dataset			
contains the relations between publications and datasets			

publication	the OpenAIRE ID of the publication	text	result.id
dataset	the OpenAIRE ID of the dataset	text	result.id
class	the type of relation between the results	text	
provenance	the provenance of the relation (harvested, inferred, etc)	text	
inferenceprovenance	if the relation was inferred, the process used for the inference	text	

3.2.9. refereed

	description	data type	refers to
refereed			
contains information about peer review status of the result			
id	the OpenAIRE ID of the result	text	result.id
refereed	the peer review status (peerreviewed, nonpeerreviewed, unknown)	text	

3.2.10. result_collectedfrom

	description	data type	refers to
result_collectedfrom			
contains the datasource from which OpenAIRE acquired the metadata of the results			
id	the OpenAIRE ID of the result	text	result.id
datasource	the OpenAIRE ID of the datasource	text	datasource.id

3.2.11. result_country

	description	data type	refers to
result_country			

contains the countries that are affiliated with the results			
id	the OpenAIRE ID of the result	text	result.id
country	the 2-letter code of the country	text	

3.2.12. result_hostedby

	description	data type	refers to
result_hostedby			
contains the data sources in which the results are hosted			
id	the OpenAIRE ID of the result	text	result.id
datasource	the OpenAIRE ID of the datasource that hosts the result	text	datasource.id
accessrights	the accessrights of the result, as reported in the metadata of the specific instance of the result	text	

3.2.13. result_licenses

	description	data type	refers to
result_licenses			
contains the licenses of the results			
id	the OpenAIRE ID of the result	text	result.id
datasource	the OpenAIRE ID of the datasource that hosts the result	text	datasource.id
type	the license of the result, as reported in the metadata	text	
normalized	a cleaned-up version of the licenses, especially for the cc and most common publisher ones	text	

3.2.14. result_orcid

	description	data type	refers to
result_orcid			
contains the ORCID identifiers of the authors/creators of the results			
id	the OpenAIRE ID of the result	text	result.id
orcid	the ORCID ID of the author	text	

3.2.15. result_organization

	description	data type	refers to
result_organization			
contains the organizations that are affiliated with the authors of the results			
id	the OpenAIRE ID of the result	text	result.id
organization	the OpenAIRE ID of the organization	text	organization.id

3.2.16. result_original_dates

	description	data type	refers to
result_original_dates			
contains all the relevant dates of the results			
id	the OpenAIRE ID of the result	text	result.id
originalid	the OpenAIRE ID of the result that was merged through deduplication with the current record	text	
date	the date	datetime	
datetype	the type of the date (publication, issued, etc)	datetime	

datasource	the OpenAIRE ID of the datasource that hosts the result	text	datasource.id
classification	the subtype of the result (article, book, book chapter, dataset, etc)	text	
embargoenddate	the date that the embargo ends, according to the metadata of the datasource hosting the result with the original ID	datetime	

3.2.17. result_original_pids

	description	data type	refers to
result_original_pids			
contains the persistent identifiers of the results, along with the datasource that hosts the results			
id	the OpenAIRE ID of the result	text	result.id
originalid	the OpenAIRE ID of the result that was merged through deduplication with the current record	text	
datasource	the OpenAIRE ID of the datasource that hosts the result	text	datasource.id
pid_type	the type of the PID (DOI, PMID, etc)	text	
pid	the PID of the result	text	

3.2.18. result_pids

	description	data type	refers to
result_pids			
contains the persistent identifiers of the results			
id	the OpenAIRE ID of the result	text	result.id
pid_type	the type of the PID (DOI, PMID, etc)	text	
pid	the PID of the result	text	
crossref	if the PID is a DOI, whether it was issued by Crossref[11]	Boolean	

[11] https://www.crossref.org/

3.2.19. result_processingfees

	description	data type	refers to
result_processingfees			
contains the processing fees of the results			
id	the OpenAIRE ID of the result	text	result.id
amount	the processing fees	numeric	
currency	the currency of the fees	text	

3.2.20. result_project

	description	data type	refers to
result_project			
contains the relations between the results and their funding projects			
id	the OpenAIRE ID of the result	text	result.id
project	the OpenAIRE ID of the project	text	project.id
inferenceprovenance	if the relation was inferred, the process that was used	text	
provenance	whether this relation was harvested, inferred, etc	text	
datasource	if the relation was harvested, the datasource where the result is hosted	text	datasource.id

3.2.21. result_sourcetype

	description	data type	refers to
result_sourcetype			

contains the type of the datasource that the result was acquired from			
id	the OpenAIRE ID of the result	text	result.id
source	the type of the datasource (journal in DOAJ, repository, etc)	text	

3.2.22. result_types

	description	data type	refers to
result_types			
contains the subtypes of the results			
id	the OpenAIRE ID of the result	text	result.id
type	the subtype of the result (article, book, chapter, etc)	text	

3.2.23. result_urls

	description	data type	refers to
result_urls			
contains the URLs of the results			
id	the OpenAIRE ID of the result	text	result.id
url	the URL the result	text	
datasource	the OpenAIRE ID of the datasource that hosts the result	text	datasource.id

3.2.24. result_version

	description	data type	refers to
result_version			

contains the locations of the open access versions of the results from Unpaywall			
id	the OpenAIRE ID of the result	text	result.id
doi	the DOI of the result	text	
host_type	the type of the host hosting the result	text	
version		text	

3.2.25. result_accessibility

	description	data type	refers to
result_accessibility			
contains the results of the accessibility assessment			
id	the OpenAIRE ID of the result	text	result.id
sourceurl	the URL of the result, as specified in the metadata	text	
dataseturl	the actual URL where the result payload was located	text	
wasurlchecked	whether the URL was processed or not	Boolean	
wasurlvalid	whether the URL was a valid URL	Boolean	
wasdocumentordatasetaccessible	whether the URL was accessible	Boolean	
wasdirectlink	whether the URL in the metadata was a direct link to the payload	Boolean	
comment	in case the assessment was not complete, a reason for the failure	text	

3.2.26. result_citations

	description	data type	refers to
result_citations			
contains the number of incoming and outgoing citations of the results			

id	the OpenAIRE ID of the result	text	result.id
incoming	the number of incoming citations	integer	
outgoing	the number of outgoing citations	integer	

3.2.27. result_fos

	description	data type	refers to
result_fos			
contains the results of the classification of publications to scientific fields, according to the Frascati Manual and EuroSciVoc			
id	the OpenAIRE ID of the result	text	result.id
fos1	the first level of the classification	text	
fos2	the second level of the classsification	text	
fos3	the third level of the classification	text	

3.2.28. result_validation

	description	data type	refers to
result_urls			
contains the URLs of the results			
id	the OpenAIRE ID of the result	text	result.id
score	the OpenAIRE Validator score	double precision	
datasource	the OpenAIRE ID of the datasource that hosts the result	text	datasource.id

3.3. moap_ec: per table field descriptions

3.3.1. article29_3

	description	data type	refers to
article29_3			
the projects that participated and did not opt out of the ORDP (data shared by the EC)			
code	the project grant ID	text	moap_org.project.code, moap_ec.project.project_number

3.3.2. beneficiaries

	description	data type	refers to
beneficiaries			
beneficiary metadata for select projects (data shared by the EC)			
project	the project grant ID	text	moap_org.project.code, moap_ec.project.project_number
opting_out	whether the project opted out of the ORDP	Boolean	
pic	the PIC number[12] of the beneficiary	text	
country	the two-letter code of country of the beneficiary	text	
type	the type of activity of the beneficiary	text	
name	the name of the beneficiary	text	

3.3.3. dataset_accessibility

	description	data type	refers to
dataset_accessibility			
information on the datasets' URLs reported to the EC via SyGMA			

[12] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/grants/applying-for-funding/register-an-organisation/registration-of-organisation_en.htm

id	dataset ID	text	moap_ec.datasets.id
sourceurl	the URL of the dataset	text	moap_ec.datasets.repo_link, moap_ec.datasets.non_repo_link
dataseturl	the direct URL to the datafile	text	
wasurlchecked	whether the sourceurl was checked	text ('true', 'false')	
wasurlvalid	whether the sourceurl was valid	text ('true', 'false')	
wasdocumentordatasetaccessible	whether the data file was accessible via the sourceurl	text ('true', 'false')	
wasdirectlink	whether the sourceurl linked directly to the datafile	text ('true', 'false')	
comment	in case the assessment was not complete, a reason for the failure	text	

3.3.4. dataset_project

	description	data type	refers to
dataset_project			
table linking datasets and projects in moap.ec			
id	dataset ID	text	moap_ec.datasets.id
project_number	project ID	text	moap_ec.project.project_number
order	the project dataset order number cr_od_dataset_nbr as shared by the EC	text	

3.3.5. datasets

	description	data type	refers to
datasets			
datasets as reported to the EC on SyGMA (shared by the EC, id created in MOAP study_			
open	whether the dataset is open	Boolean	
cr_od_dataset_nbr	the project dataset number	text	
cr_od_doi	the cleaned out doi of the dataset	text	moap_org.result_pids.pid
cr_od_title	the dataset title	text	
accessible	whether the dataset is reported as accessible	Boolean	
reusable	whether the dataset is reported as reusable	Boolean	
repo_link	the repository URL link of the dataset (via OpenAIRE)	text	dataset_accessibility.sourceurl
non_repo_link	the non-repository URL link of the dataset	text	dataset_accessibility.sourceurl
linked_publ	the DOI of the publication linked to the dataset	text	
id	the ID of the dataset (as created in the MOAP study)	text	moap_org.result.id

3.3.6. ec_ordp_optout

	description	data type	refers to
ec_ordp_optout			
projects ORDP opting out information for select projects (as shared by the EC)			
start_year	project start year	text	
project_number	project grant id	text	moap_org.project.code, moap_ec.project.project_number
optout	whether the project opted out the ORDP	text	
dma_relevance	Are data management activities relevant to	text	

	your proposed project?		
dma_deliverable	Will a data management plan be delivered?	text	
dma_workpackage	Is data management part of a work package?	text	
dma_other	Will the data management plan be integrated in other activities?	text	
dma_other_expl	If so, which other activities?	text	
The columns below refer to reasons for opting out the ORDP.			
does_not_generate	The project does not generate any data	text	
allow_protection	To allow the protection of results (e.g., patenting)	text	
incomp_confidentiality	Incompatibility with the need for confidentiality linked to security	text	
incomp_privacy	Incompatibility with privacy/data protection	text	
jeopardised	Achievement of the project's main aim would be jeopardised	text	
other_legitimate_reasons	Other legitimate reasons	text	
voluntary		text	
other_legitimate_reasons_expl	Other legitimate	text	

	reasons explained		
not_involved_in_other_submission_or_implementation		text	
achievement would be jeopardised	Achievement of the project's main aim would be jeopardised	text	
incompatibility	Incompatibility with the need for confidentiality linked to security	text	
does_not_generate_data	The project does not generate any data	text	
other	Other legitimate reasons	text	
incompatibility_data_protection	Incompatibility with privacy/data protection	text	

3.3.7. project

	description	data type	refers to
project			
project information as shared by the EC			
topic_unique_programme_part_code	the programme/subprogramme code	text	
topic_unique_programme_part_descr	the description of the programme/subprogramme	text	
project_number	the project grant ID	text	moap_org.project.code
project_status	whether the project is closed	text	
project_call_id	the call the project responded to	text	
project_toa_lvl2_code	EC TOA level 2 code	text	

topic_code	EC topic code	text	
topic_descr	EC topic description	text	
project_acronym	project acronym	text	
project_start_date	the start date of the project	text	
project_end_date	the end date of the project	text	
project_requested_eu_contrib	the amount (in EUR) of the requested EU contribution of the project	text	

3.3.8. publication_accessibility

	description	data type	refers to
publication_accessibility			
information on the publications' URLs reported to the EC via SyGMA			
id	publication ID	text	moap_ec.publications.id
sourceurl	the URL of the publication	text	moap_ec.publications.cr_publ_repository_url
dataseturl	the direct URL to the text file	text	
wasurlchecked	whether the sourceurl was checked	text ('true', 'false')	
wasurlvalid	whether the sourceurl was valid	text ('true', 'false')	
wasdocumentordatasetaccessible	whether the text file was accessible via the sourceurl	text ('true', 'false')	
wasdirectlink	whether the sourceurl linked directly to the text file	text ('true', 'false')	
comment	in case the assessment was not complete, a reason for the failure	text	

3.3.9. publication_project

	description	data type	refers to
publication_project			
table linking publications and projects in moap_ec			
id	publication ID	text	moap_ec.publications.id
project_number	project ID	text	moap_ec.project.project_number
order	the project publication order number as shared by the EC	text	

3.3.10. publications

	description	data type	refers to
publications			
publications as reported to the EC on SyGMa (shared by the EC, ID created in MOAP study			
cr_publ_type	the type of the publication	text	
cr_publ_title	the title of the publication	text	
cr_publ_journal_title	the venue title of the publication	text	
cr_publ_doi	the DOI of the publication (cleaned out in the moap study)	text	publication_accessibility.sourceurl
flag_oa	whether the publication is reported as open access	text	
cr_publ_green_oa	whether the publication is reported as green open access	text	
cr_publ_green_oa_embargo	whether the publication is ported as green oa with an embargo	text	

cr_publ_in_gold_oa	whether the publication is reported as being published in gold open access	text	
cr_publ_gold_oa_process_charges_amt	the reported processing charge of the gold open access publication	text	
cr_publ_issn	the ISSN of the venue of publication	text	
cr_publ_publisher	the publisher of the venue	text	
cr_publ_published_place	the country the publication was published	text	
cr_publ_published _year	the year of publication	text	
cr_publ_is_peer_reviewed	whether the publication is reported as peer-reviewed	text	
cr_publ_is_join_publicprivate	whether the publication was the result of a public private partnership	text	
cr_publ_relevance	publication relevance as shared by the EC	text	
cr_publ_team_members	team members	text	
id	ID of publication as created in the moap study	text	

3.4. moap_final: per table field descriptions

3.4.1. final_apcs

	description	data type	refers to
final_apcs			
contains the actual and extrapolated apcs for Horizon 2020 peer-reviewed publications.			
openaireid	the OpenAIRE ID of the result	text	moap_org.result.id
ecid	the EC ID of the result	text	moap_ec.publications.id

pid	the PID ot the result	text	moap_org.result_pids.pid, moap_ec.publications.cr_publ_doi
pid_type	the type of the PID	text	
publisher	the publisher of the result	text	
publisher_norm	the cleaned out and grouped publisher of the result (e.g. nature and springer, grouped to springer - nature)	text	
publication_year	the year of publication of the result	text	
hybrid_oa	whether the publication is hybrid open access or not	numeric (0,1)	
book_or_chapter	whether the publication is a book, part of a book or chapter of a book	numeric (0,1)	
open_apc	the processing fees of the publication (in EUR) as it is listed in the openapc database	numeric	
apc_extr2	the extrapolated APC for the publication (in EUR)	numeric	

3.4.2. final_data

	description	data type	refers to
final_data			
contains the indicators for Horizon 2020 datasets			
openaireid	the OpenAIRE ID of the result	text	moap_org.result.id
ecid	the EC ID of the result	text	moap_ec.datasets.id
from_ordp_project	whether the dataset was produced in a project that participated in the open research data pilot	numeric (0,1)	
linked_pub_openaireid	the OpenAIRE ID of the publication linked to the dataset	text	moap_org.result.id
has_linked_pub	whether the dataset is linked to a publications	numeric (0,1)	
co_funded	whether the result was funded by more than one funder	numeric (0,1)	
multiple_projects	whether the result is linked to more than one project	numeric (0,1)	

co_authored	whether the result has more than one authors	numeric (0,1)	
with_orcid	whether the result has in its metadata the ORCID ID of at least one of its authors	numeric (0,1)	
linked_pub_has_orcid	whether the publication linked to the dataset has in its metadata the ORCID ID of at least one of its authors	numeric (0,1)	
open_access	whether at least one instance of the result is open access	numeric (0,1)	
in_repo	whether the result has been deposited in a repository harvested by OpenAIRE	numeric (0,1)	
pid_in_repo	whether the result has a PID in its repository metadata (null if result not in a repository)	numeric (0,1)	
accessible_via_repo	whether the data file of the result can be access via a URL in its repository metadata (null if result not in a repository)	numeric (0,1)	
interoperable_via_repo	whether the data file of the result accessed via a URL in its metadata is in a machine-readable format	numeric (0,1)	
best_openaire_guidelines_score	the best validator score[13] of available metadata records for the result, using the OpenAIRE guide archives[14]	numeric	
findable_in_repo	whether the result has a PID and a valid URL and its repository metadata (null if result not in a repository)		
uri_to_data_file_in_repo	whether the result has a valid URL in its repository metadata (null if result not in a repository)		

3.4.3. final_pubs

	description	data type	refers to
final_pubs			

[13] https://www.openaire.eu/validator-registration-guide
[14] https://guidelines.openaire.eu/en/latest/data/index.html

contains the indicators for h2020 publications			
openaireid	the OpenAIRE ID of the result	text	moap_org.result.id
ecid	the EC ID of the result	text	moap_ec.publications.id
co_funded	whether the result was funded by more than one funder	numeric (0,1)	
in_multiple_projects	whether the result is linked to more than one project	numeric (0,1)	
co_authored	whether the result has more than one authors	numeric (0,1)	
best_cc_license_publisher	the most open CC licence that has been found in the journal or publisher metadata[15]	text	
gold_not_hybrid	a publication published in an open access journal	numeric (0,1)	
hybrid_oa	an open access publication that is not in an open access journal has a CC license in the journal or publisher metadata	numeric (0,1)	
gold_oa	a publication that can be found as open access at the journal of publisher	numeric (0,1)	
bestaccessrights_publisher	the best available access rights at the journal or publisher metadata[16]	text	
open_access	whether the result is open access (access rights of the publisher are given priority over the access rights in a repository, i.e. if closed access at publisher and open access at the repository, open_access=0)	numeric (0,1)	
green_oa_any_version	whether the result can be found open access in a repository	numeric (0,1)	
green_oa	whether the version of record (VOR) or author-accepted manuscript (AAM) of the result can be found open access in a repository (null if no info on version deposited)	numeric (0,1)	
pub_date_in_repo_metadata	whether the publication date of the result can be found in its repository metadata (null if result not found in a repository)	numeric (0,1)	
erc_only	whether the result is linked only to ERC projects	numeric (0,1)	
embargo_date_compliant	whether open access to the result is given within the time limits of Article 29.2 of the Model Grant Agreement[17] (null if not embargo end date available)	numeric (0,1)	
pid_in_repo_data	whether the result has a PID in its repository metadata (null if result not found in a repository)	numeric (0,1)	

[15] Ranking: https://commons.wikimedia.org/wiki/file:creative_commons_license_spectrum.svg
[16] Ranking: open access, embargo, restricted, closed access
[17] https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf

with_embargo_end_date	whether the result has an embargo end date in its metadata	numeric (0,1)	
embargo_end_date_in_repo_data	whether the result has an embargo end date in its repository metadata (null if result not found in a repository)	numeric (0,1)	
project_ref_in_repo	whether the result has the linked project id in its repository metadata (null if result not found in a repository)	numeric (0,1)	
accessible_via_repo_url	whether the full text file of the result can be access via a url in its repository metadata (null if result not in a repository)	numeric (0,1)	
best_openaire_guidelines_score	the best validator score[18] of available metadata records for the result, using the OpenAIRE guidelines for literature repositories[19]	numeric	
uri_to_full_text	whether the result has a valid URL in its metadata	numeric (0,1)	
with_pid	whether the result has a PID in its metadata	numeric (0,1)	
findable	whether the result has a PID and a valid URL in its metadata	numeric (0,1)	
accessible	whether the full text file of the result can be access via a URL in its metadata	numeric (0,1)	
interoperable	whether the full text of the result accessed via a URL in its metadata is in a machine-readable format	numeric (0,1)	

3.5. OpenAIRE Research Graph

The OpenAIRE Research Graph includes metadata and links between scientific products (e.g., literature, datasets, software, and "other research products"), organizations, funders, funding streams, projects, communities, and (provenance) data sources - the details of the graph data model can be found in Zenodo.org.

The Graph is available and obtained as an aggregation of the metadata and links collected from ~1500 trusted sources, further enriched with metadata and links provided by:

- OpenAIRE end-users, e.g., researchers, project administrators, data curators providing links from scientific products to projects, funders, communities, or other products;
- OpenAIRE Full-text-mining algorithms over around ~10Mi open access Article full-texts;
- Research infrastructure scholarly services, bridged to the graph via OpenAIRE, exposing metadata of products such as research workflows, experiments, research objects, software, etc.

[18] https://www.openaire.eu/validator-registration-guide
[19] https://guidelines.openaire.eu/en/latest/literature/index.html

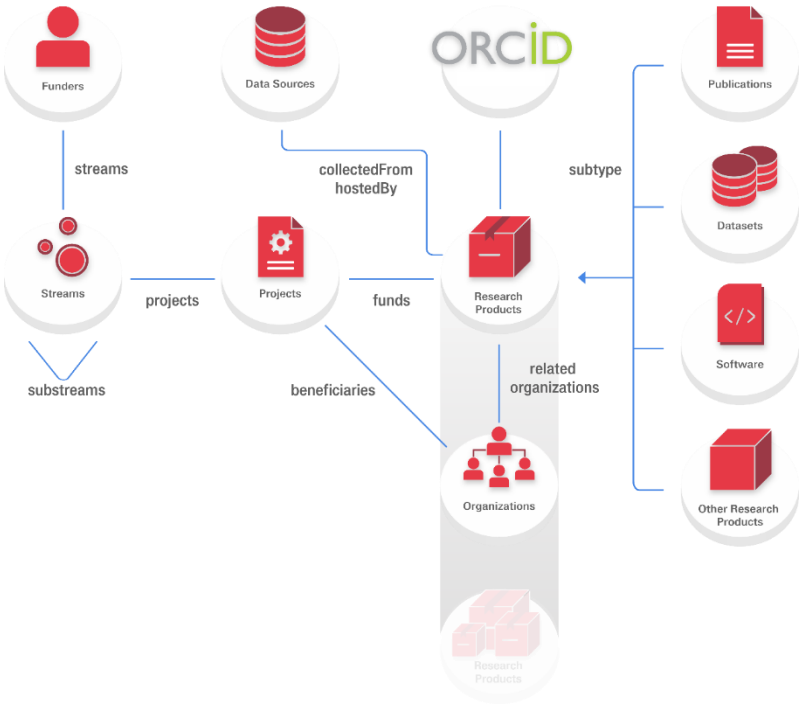

FIGURE 4: THE OPENAIRE RESEARCH GRAPH

3.5.1. OpenAIRE entities, relationships and types

The entities in the data model belong to the following categories:

Core entities: the entities whose information is continuously and incrementally fed to the ORG (OpenAIRE Research Graph) and is of interest to OpenAIRE end-users; namely **Result** (Literature, Dataset, Software, Other products), **Organization**, **Data Source**, **Projects**, **Funder, Funding Stream**.

Linking entities: entities used to model relationships, used to connect in a semantic-agnostic way two or more main entities; namely, those denoted by an Entity1_Entity2 notation.

Types: types are used to define structured values for entity properties. Structured values differ from objects in the sense they do not have an identity, i.e., cannot be referred to by relationships from other objects.

3.5.2. Core Entities

In this section the core entities of the data model are introduced by describing the concept or real-world entity they represent and introducing their descriptive properties and relationships with other entities.

Results are intended as digital objects, described by metadata, resulting from a scientific process. Its sub-entity types are Literature, Dataset, Software, and Other Research Product, which inherit all Result properties and relationships with other entities and add their specific ones:

- **Literature** includes all digital research artefacts whose intended use is narrative storytelling of a research activity and its results. Examples are scientific Articles, reports, slides, data papers, patents etc. Although there are exceptions, as each scientist has a large degree of freedom in publishing and interlinking his artefacts, it can be generally assumed that literature artefacts are published with a narrative intent. For those specific cases where literature is intended for different use, we in general do not expect scientists to publish such artefacts as literature artefacts. For example, when an Article is a carrier of readable datasets (e.g., Articles with tables) the Article is often deposited a second time in a data repository, assigned a new DOI, and marked as a dataset of type "textual"; in the case Articles full-texts are used for natural language processing (NLP), scientists will likely create a dataset of type "collection of Articles".

- **Datasets** include digital research artefacts encoding experimental or real-world observations/measures (e.g., primary data), secondary data derived from programmatic processing of other datasets, or more generally digital representations of facts to be interpreted by a program. The definition is cross-discipline, hence spans across multiple interpretations of datasets, where typologies and granularity obey different scientific facets. Examples include, but are not limited to: databases (e.g., Worms), records of databases (e.g., proteins in the UniProt database), table files, queries over databases (time-series slices, geospatial maps, SQL queries), media (e.g., images, videos) or collections of media.

- **Software** entities represent research software, i.e., software that is an output of a research activity. Examples include, but are not limited to: code scripts, source code of web services and/or web applications.

- **Other research products** include any research output that is not literature, data, or software. Examples include, but are not limited to: algorithms, scientific workflows/pipelines, protocols, standard operating procedure (SOP), simulations, mathematical and statistical models, but also research packages. Research packages can group a set of research artefacts, but can also include the encoding of a composition logic that binds them together. For example, an instance of a workflow is a package that describes the combination of specific artefacts to implement a scientific process, execute an experiment, etc.

- **Communities** i.e., are intended as groups of people with a common research intent and can be of two types: *research initiatives* or *research communities*. The former is intended to capture a view of the information space that is "research impact"-oriented, i.e., all products generated due to my research initiative, the latter "research activity" oriented, i.e., all products that may be of interest or related to my research initiative. For example, the organizations supporting a research infrastructure fall in the first category, while the researchers involved in a discipline fall in the second.

- **Organizations** include companies, research centres or institutions involved as project partners or as responsible for operating data sources. Information about organizations is collected from funder databases like CORDA, registries of data sources like OpenDOAR and re3Data, and current research information systems (CRIS[20]), as being related to projects or data sources.

- **Funders, funding streams and projects**. Of crucial interest to OpenAIRE is also the identification of the funders (e.g., European Commission, WellcomeTrust, FCT Portugal, NWO The Netherlands) that co-funded the projects that have led to a given result. Funders can be associated with a list of funding streams (e.g., FP7, Horizon 2020 for the European Commission), which identify the strands of funding. Funding streams can be nested to form a tree of sub-

[20] https://www.eurocris.org/why-does-one-need-cris

funding streams. Projects are typically associated with the funding stream "leaves" of such trees.

- **Data sources**. OpenAIRE entity instances are created out of data collected from various data sources of different kinds, such as publication repositories, dataset archives, CRIS systems, funder databases, etc. Data sources export information packages (e.g., XML records, HTTP responses, RDF data, JSON) that may contain information on one or more of such entities and possibly relationships between them. For example, a metadata record about a project carries information for the creation of a Project entity and its participants (as Organization entities). It is important, once each piece of information is extracted from such packages and inserted into the OpenAIRE information space as an entity, for such pieces to keep provenance information relative to the originating data source. This is to give visibility to the data source, but also to enable the reconstruction of the very same piece of information if problems arise.