RISIS

RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

DOCUMENTATION OF RISIS DATASETS

EUPRO_FP

A. Wolfmayr,
M. Barber, B. Heller-Schuh, Th. Scherngell, G. Zahradnik (AIT)

June 2022

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 824091

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Outline

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

1 Basic Characteristics

Name and short description of the infrastructure

The development of the EUPRO database has been started by AIT back in 2005 comprising information on R&D projects and all participating organizations funded by the European Framework Programmes (FP). Within RISIS, EUPRO has been extended significantly by including data of other European, and recently also national funding initiatives (https://zenodo.org/record/3934327#.Y3zSz32ZMZU).

Currently EUPRO consists of the following four modules:

- data on FP projects (1984-2021), including H2020 as the main update to the prae-RISIS version of EUPRO,
- data on R&D projects funded by EUREKA (1985-2019), an intergovernmental network supporting market-oriented R&D and innovation projects across all technological sectors,
- data on R&D projects funded through the Joint Technology Initiatives ARTEMIS, ENIAC and ECSEL (JTI) (2008-2014), long-term public-private partnerships, which support large-scale multinational research activities in areas of major interest to European industrial competitiveness and issues of high societal relevance,
- data on R&D actions funded by the European Cooperation in Science and Technology (COST) (1971-2014), a pan-European networking initiative supporting transnational cooperation among researchers, engineers and scholars across Europe, and
- data on nationally funded projects in the so-called NATPRO module of EUPRO (https://zenodo.org/record/4610742#.Y3zTC32ZMZU)

Aim of the database

The EUPRO database is a significant asset of the Center for Innovation Systems & Policy of AIT used for basic oriented research projects and contract research for national and international customers, such as the European Commission. It facilitates the analysis of participation patterns of organisations in and across different European funding initiatives and the investigation of collaborative network structures, including their evolution over time and the development of the European Research Area (ERA).

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Legal name of operating organization

The database is operated and maintained by AIT Austrian Institute of Technology GmbH, located at Giefinggasse 4, 1210 Vienna, Austria.

2 Database content

The documentation for the complete EUPRO database (comprising all four European programme modules), last updated in 2021, can be found here:

> Heller-Schuh, B., Barber, M., Bilalli Shkodra, Xh., Scherngell, Th. & Zahradnik, G. (2020). Documentation of RISIS datasets: EUPRO (1.0). Zenodo. https://doi.org/10.5281/zenodo.4428394.

This documentation focuses on the current EUPRO_FP module (EUPRO_FP_2022) comprising all project data for FP7 and H2020, publicly available in May 2022. Moreover, note that NATPRO as new module of EUPRO will have its own documentation (currently under development: https://zenodo.org/record/4610742#.Y3zTC32ZMZU)

2.1 Definition and description of observations

Units and definition of observations

Basically, EUPRO_FP covers information on **projects** (such as project objectives and achievements, project costs, total funding, start and end date, contract type, information on the programme call), **participants** (name of the participating organisation, organisation type, geographical location, and funding) and **project output** (such as title, author/applicant name and DOI or access URLs of project reports, deliverables, publications and patents).

Number of observations

EUPRO_FP_2022 comprises information on ***61,163 projects*** and ***313,084 participations*** *for FP7 and H2020*. Note that information on former programmes (FP1-FP6) is still available upon request (https://zenodo.org/record/5828177#.Y3zSeX2ZMZU).Table 1 disaggregates the units of observation by different FPs and other European funding initiatives.

Table 1: EUPRO database - number of projects and participations

Programme	Period	Projects	Participations
FP7	2007 - 2013	25,785	140,048
H2020	2014 - 2020	35,378	173,084
Total	**2007 - 2020**	**61,163**	**313,132**

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

2.2 Data acquisition and processing (e.g. data cleaning)

Where the data are retrieved from

The data on FP projects for FP7 and H2020 are publicly available and can be downloaded as excel files from the CORDIS projects database (http://cordis.europa.eu/). Basic raw data on FP7 and H2020 projects, participants, and project outcomes of the current version (EUPRO_FP_2022) was downloaded in May 2022 in xlsx-format.

How the data are processed in terms of data cleaning

The quality of the raw data extracted from the different programmes websites is not generally sufficient for policy-relevant analyses. AIT has undertaken substantial efforts to improve quality of the data and to retrieve and add missing data. Data quality of data was improved by cleaning the netto EC Contributions for H2020 data, by cleaning the country code and by extending, cleaning and harmonizing the type of participating organizations. Furthermore, organisation names have been linked to organisations listed in OrgReg, the Register of European Public Research and Higher Education Actors, and FirmReg, the list of businesses included in the firms' datasets which are part of RISIS[1].

Data cleaning and harmonisation includes four major steps:

- harmonization of the participants' activity type and the country code provided by CORDIS
- cleaning of the netto EC Contributions provided by CORDIS
- linking of organisation names to OrgReg and FirmReg entities, and
- regionalisation (i.e., assignment to European NUTS3 regions).

The participants' activity type was harmonized across organisations with the same organisation name within the CORDIS raw data. For this the organisation name was cleaned, i.e., punctuation, blanks and accents were removed where necessary. The cleaned names were then grouped and organisations with different or missing activity types were identified and harmonised. Moreover, in some cases the activity types were different for the same organisation in data from FP7 and H2020. In this case, the activity type was harmonised to the ones provided in the data from H2020. Furthermore, activity types written in the following way

[1] for more information on OrgReg and FormReg, see Lepori, B. (2022). OrgReg Methodological manual. Zenodo. https://doi.org/10.5281/zenodo.6396703 and Camerani, R. and Guerini, M. (2022), FirmReg Handbook. Zenodo, https://doi.org/10.5281/zenodo.6424802.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

e.g., HES; HES were harmonised to HES to facilitate grouping within an activity type. Similarly, country codes provided in CORDIS in the form of e.g., AT; AT were harmonised to AT.

For the H2020 projects the net EC Contributions and EC Contributions were cleaned such that the sum of all net EC Contributions for the participant and its third parties is equal to the participant's EC Contribution. If the sum did not match, the data was manually compared to the data provided on the CORDIS website (last retrieved June 2022) and harmonized accordingly. In 48 cases the sum did not match due to missing third parties in the CORDIS raw data, but these are named on the CORDIS website. They were manually added. The EUPRO_FP_2022 database thus covers 173,084 participants instead of the 173,036 provided in the CORDIS raw data (downloaded in May 2022). As the CORDIS website gets updated regularly, it was not possible to clean everything. In fact, the data on the CORDIS website did not match the downloaded data for net EC and EC contributions for two projects with the project IDs 825575 and 945541. Thus, for these two projects cleaning was not possible.

The linking between EUPRO organisation names and the entities in OrgReg and FirmReg was conducted by applying specific matching algorithms developed by AIT. These algorithms are based on statistical properties such as the frequency of adjacent characters in the organisation names and are used to identify similar organisation names that can be attributed to the same organisation. All algorithmically identified name matches are manually checked for accuracy.

For the data regionalisation, we used a two-stage process. First, we used a mapping from postal codes to NUTS3 regions. Where that was unsuccessful, we made use of RISIS-developed geolocalisation tools, specifying their geographical locations by giving their latitude and longitude coordination. This facilitates all kind of spatial analyses of project-based R&D networks, e.g., the investigation of the network at the level of functional urban areas.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

2.3 Information on all variables/indicators

Programmes table

Table 2: Description of variables providing information about programmes

Variable	Description
prgId	internal unique identifier for FP subprogrammes[2]
prgType	code for the names of the specific framework programme types (FP7 and H2020) in which the subprogrammes were funded[2]
prgName	full name of subprogramme areas in each of the framework programmes (e.g., Specific Programme "Cooperation": Health)[3]
prgAcr	subprogramme acronym (e.g., FP7-HEALTH)[3]

Projects tables

Table 3: Description of variables providing information about projects

Variable	Description
projectId	unique identifier for each project in the database, identical with unique identifier of all projects in the CORDIS projects database[3]
title	full title of the project[3]
projAcr	(non-unique) project acronym or abbreviation of the project title[3]
startDate, endDate	day, month and year of project start and end[3]
totalProjectCosts	official project costs as indicated in the project proposal[3]
projectECFunding	factual financing contribution of the European Union; since not all projects are financed completely, figures in "projectECFunding" are equal to or smaller than figures in "totalProjectCosts".[3]
fundingSchemeId	funding scheme identifier (corresponding Funding Schemes see Table 4)[2]
masterCall	call identifier for H2020[4,3]
subCall	subcall identifier for FP7 and H2020[3]
objective	conceptual orientation of the project[3]
projectUrl	official website of the project sourced from CORDA[3]
rcn	unique identifier (record control number) for each project in the database[3]

[2] introduced and/or processed by AIT

[3] provided by source data set, unchanged

[4] no MasterCall data for FP7 available

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Table 4: Description of variables providing information about funding_schemes

Variable	Description
fundingSchemeId	internal unique funding scheme identifier[2]
fundingSchemeCode	abbreviation of funding schemes[3]
fundingSchemeName	different types of contracts which regulate size, financing and funding of the research projects (e.g., RIA - research and innovation action, CP – collaborative project)[3]

Participations tables

Table 5: Description of variables providing information about participations

Variable	Description
participId	internal unique identifier of each participation[2]
projectId	link to projects table (Table 3)[3]
role	role of participant in the project; differentiates between "coordinator", "participant", "partner" (in MSC-Actions), "thirdParty" (commissioned by participant) and "internationalPartner (outside Europe)[3]
orgId	link to organisations table (Table 6)[2]
endOfParticipation	indicates that organisations left the projects, but also changes of organisations' names or addresses during project run-time[3]
geoId	link to geography table (Table 8)[2]
address, postcode, city	street level address information[3]
ecContribution	amount of overall EC funding on participant level[3]
netEcContribution	amount of net EC funding for each participant and third party in H2020; for FP7 programmes netEcContributions are the same as ECcontributions as netEcContributions are not available for FP7 projects.
orgUrl	link to the organisation's website[3]
cordisOrganisationId	9-digit Participant Identification Code used for all FP participants in EU programmes[1]
vatNumber	Value Added Tax Registration Number; unique number that identifies a taxable person (business) or non-taxable legal entity that is registered for VAT[3]
nutsAssignmentId	Link to table defining strategy used to identify NUTS3 region for the participation (Table 9)[2]

Table 6: Description of variables providing information about organisations

Variable	Description
orgId	internal unique identifier for each organisation; corresponds with the entries in the field orgId in the participations table (Table 5)[2]
orgName	CORDIS organisation name[3]
orgTypeId	link to organisation type (Table 7)[2]

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

orgRegId	link to RISIS-OrgReg, the Register of European Public Research and Higher Education Actors[2]
firmRegId	link to RISIS-FirmReg, the Register of businesses included in RISIS firms' datasets[2]
ctry	country

Table 7: Description of variables providing information about the organisation type

Variable	Description
orgTypeId	internal unique identifier for organisation types; corresponds with the entries in the field orgTypeId in the organisations table[2]
orgType	CORDIS activity type[3]
orgTypeDescription	explanation of organisation types[3]

Geography tables

Table 8: Description of variables providing information about the geography of participations

Variable	Description
geoId	internal unique identifier for each geographical location of the participants[2]
ctryId	link to ctry table (Table 10)[2]
NUTS3	Regional EUROSTAT classification using the 2021 NUTS definitions[2]

Table 9: Description of variables providing information about the method used to assign NUTS3 regions to participations

Variable	Description
nutsAssignmentId	internal unique identifier for strategy used to assign a NUTS3 region to a participation[2]
description	Explanation of strategy used to assign a NUTS3 region[2]

Table 10: Description of variables providing information about the country of origin

Variable	Description
ctryId	internal unique identifier for each participating country[2]
ctryCode_eu	standardised country codes of the participating organisational units; used by the European Commission[5]; in the case of multinational organisations the participating national branches are listed[2]
ctryCode_iso	standardised country codes of the participating organisational units; abbreviations are given ISO 3166-1 Alpha-2 country codes[6]; in the case of multinational organisations the participating national branches are listed[2]

[5] https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Country_codes

[6] https://www.iso.org/obp/ui/#search/code/

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

ctryName	official name of countries[2]
ctryTypeId	link to ctryTypes table[2]

Table 11: Description of variables providing information about the country type

Variable	Description
ctryTypeId	internal unique identifier for each participating country[2]
ctryType	differentiation between EU member states, associated countries and third countries (including subcategories) – abbreviations[2]
ctryTypeDesc	differentiation between EU member states, associated countries and third countries (including subcategories) – full text[2]

Project outputs tables

Table 12: Description of variables providing information about deliverables[7]

Variable	Description
cordisDelivId	Identifier for each project deliverable[3]
projectId	link to projects table (Table 3)[3]
title	deliverable title[3]
description	short description of the content[3]
deliverableType	Documents, reports; open data; Websites, patent fillings, videos etc.[3]
url	direct link to download the document (only for FP7)[3]
lastUpdateDate	date of last update[3]

Table 13: Description of variables providing information about publications

Variable	Description
cordisPublId	Identifier for each project publication[3]
projectId	link to projects table (Table 3)[3]
title	publication title[3]
authors, journalTitle, journalNumber, publishedYear, publishedPages, issn, doi	bibliographic information in the case of open access publications[3]
url	direct link to download the document[1]
isPublished	type of publication (Peer reviewed articles, Conference proceedings, Thesis dissertations, etc.)[3]
lastUpdateDate	date of last update[3]

[7] data only for H2020 available

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Table 14: Description of variables providing information about patents

Variable	Description
applicationId	Unique identifier for each project patent application[3]
projectId	link to projects table (Table 3)[3]

Table 15: Description of variables providing information about reports

Variable	Description
cordisReportId	Unique identifier for each project report[3]
projectId	link to projects table (Table 3)[3]
title	report title[3]
relatedFile	link to related illustrations, images, announcements, etc.[3]
lastUpdateDate	date of last update[3]

Thematic classifications tables

Table 16: Description of variables describing the thematic orientation (FoS) of projects (fosTerms)

Variable	Description
fosCode	unique identifier of each Fields of Science (FoS) Term[3]
fosTerm	verbal representation of fosCode[3]
hierarchyCode	placement of fosCode in the FoS taxonomy[8,3]

Table 17: Description of variables providing information about the hierarchy of FoS (fosCodeClosure)

Variable	Description
ancestorFosCode	fosCode of the superordinate Fos term[3]
fosCode	link to fosTerms table[3]
depth	number of hierarchy levels between ancestorFosCode and fosCode[3]

Table 18: Description of variables providing information about programme topics

Variable	Description
topicId	internal unique identifier for each topic[2]
prgId	link to programme table (Table 2)
topicCode	topic identifier within subprogrammes[3]
topicName	name of topic within subprogrammes[3]

[8] for a complete representation of the EuroSciVoc Taxonomy see **Fehler! Verweisquelle konnte nicht gefunden werden.** in the Appendix

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

2.4 Sectorial, temporal and geographical coverage

Information on the sectorial classifications used

Table 19: Organisation type

OrgTypes	Description
HES	Higher or Secondary Education Establishments
REC	Research Organisations
PRC	Private for-profit entities (excluding Higher or Secondary Education Establishments)
PUB	Public bodies (excluding Research Organisations and Secondary or Higher Education Establishments)
OTH	Other
NA	Not available

Table 20: Fields of Science (EuroSciVoc) taxonomy[9]

Level 0	Level 1	Level 2
agricultural sciences	agricultural biotechnology	agricultural genetics
		biomass
		marker assisted selection
	agriculture, forestry, and fisheries	agriculture
		fisheries
		forestry
	animal and dairy science	apiculture
		dairy
		pets
	other agricultural sciences	
	veterinary science	
engineering and technology	chemical engineering	biochemical engineering
		chemical engineering software
		chemical process engineering
		separation technologies
	civil engineering	architecture engineering
		construction engineering
		structural engineering
		transportation engineering
		water engineering
		urban engineering
	electrical engineering, electronic engineering, information engineering	electrical engineering
		electronic engineering
		information engineering

[9] for the complete EuroSciVoc 6-level table see https://op.europa.eu/en/web/eu-vocabularies/concept-scheme/-/resource?uri=http://data.europa.eu/8mn/euroscivoc/40c0f173-baa3-48a3-9fe6-d6e8fb366a00

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

	environmental biotechnology	bioremediation
		biosensing
	environmental engineering	energy and fuels
		geological engineering
		geotechnics
		mining and mineral processing
		natural resource management
		remote sensing
		waste management
		air pollution engineering
		carbon capture engineering
		water treatment processes
	industrial biotechnology	biomaterials
		bioprocessing technologies
		metabolic engineering
	materials engineering	ceramics
		coating and films
		colors
		composites
		crystals
		fibers
		liquid crystal
		metallurgy
		nanocomposites
		paper and wood
		synthetic dyes
		textiles
	mechanical engineering	applied mechanics
		manufacturing engineering
		mechatronics
		thermodynamic engineering
		tribology
		vehicle engineering
	medical engineering	medical laboratory technology
		diagnostic imaging
		wearable medical technology
	nanotechnology	nano-materials
		nano-processes
		nanoelectromechanical systems
		nanoelectronics
		nanophotonics
	other engineering and technologies	food and beverages
		microtechnology
		nuclear engineering
humanities	arts	architectural design
		art history
		musicology

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

		modern and contemporary art
		performing arts
		visual arts
	history and archaeology	archaeology
		history
	languages and literature	languages - general
		linguistics
		literature - general
	other humanities	library science
	philosophy, ethics and religion	ethics
		philosophy
		religion
medical and health sciences	basic medicine	anatomy and morphology
		immunology
		medical genetics
		medicinal chemistry
		neurology
		pathology
		pharmacology and pharmacy
		physiology
		toxicology
	clinical medicine	allergology
		anaesthesiology
		andrology
		angiology
		cardiology
		clinical microbiology
		clinical neurology
		critical care medicine
		dentistry
		dermatology
		embryology
		emergency medicine
		endocrinology
		gastroenterology
		general medicine
		gerontology
		hematology
		hepatology
		integrative & complementary med.
		internal medicine
		nephrology
		gynaecology
		obstetrics
		odontology
		oncology
		ophthalmology

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

		orthopaedics
		otorhinolaryngology
		paediatrics
		peripheral vascular disease
		physiotherapy
		pneumology
		psychiatry
		radiology
		rheumatology
		surgery
		transplantation
		urology
		venereal diseases
	health sciences	dietetics
		epidemiology
		health care sciences
		health care services
		infectious diseases
		inflammatory diseases
		medical ethics
		nursing
		nutrition
		parasitology
		personalized medicine
		public and environmental health
		social biomedical sciences
		speech-language pathology
		sport and fitness sciences
		substance abuse
		tropical medicine
	medical biotechnology	cells technologies
		genetic engineering
		implants
		medical bioproducts
		nanomedicine
		prosthetics
		tissue engineering
	other medical sciences	forensic science
		history of medicine
natural sciences	biological sciences	biochemistry
		biodiversity conservation
		behavioural sciences biology
		biological morphology
		biology
		biophysics
		botany
		cell biology

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

		developmental biology
		ecology
		evolutionary biology
		freshwater biology
		genetics and heredity
		histology
		marine biology
		microbiology
		molecular biology
		morphology
		neurobiology
		reproductive biology
		synthetic biology
		zoology
		other biological topics
	chemical sciences	analytical chemistry
		catalysis
		electrochemistry
		inorganic chemistry
		nuclear chemistry
		organic chemistry
		physical chemistry
		polymer science
	computer and information sciences	artificial intelligence
		computational science
		computer security
		data science
		databases
		internet
		knowledge engineering
		software
	earth and related environmental sciences	atmospheric sciences
		environmental sciences
		geochemistry
		geology
		geophysics
		hydrology
		oceanography
		palaeontology
		physical geography
		soil science
	mathematics	applied mathematics
		pure mathematics
	other natural sciences	
	physical sciences	acoustics
		astronomy
		atomic physics

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

		classical mechanics
		condensed matter physics
		electromagnetism and electronics
		molecular and chemical physics
		nuclear physics
		optics
		plasma physics
		quantum field theory
		quantum mechanics
		relativistic mechanics
		theoretical physics
		thermodynamics
social sciences	economics and business	business and management
		economics
	educational sciences	didactics
		inclusive education
		pedagogy
		special education
	law	admiralty law
		constitutional law
		criminology
		human rights
		international law
		law enforcement agencies
		penology
	media and communications	graphic design
		information science
		journalism
	other social sciences	social sciences interdisciplinary
	political science	government systems
		political communication
		political policies
		political transitions
		public administration
	psychology	behavioural psychology
		cognitive psychology
		developmental psychology
		ergonomics
		psycholinguistics
		psychotherapy
		social psychology
	social and economic geography	cultural and economic geogr.
		transport
		urban studies
	sociology	anthropology
		demography
		family studies

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

		gender studies
		globalization
		governance
		ideologies
		industrial relations
		social problems
		social work

Information on the temporal coverage used

EUPRO module	First call year	Last call year
FP 2022	1984	2022

Information on the geographical coverage and classifications used

Since we have information on the geographical location of the project participants in the EUPRO database, we can analyse their geographical distribution across Europe at the country-level as well as on the regional level by assigning organisation to European NUTS regions[10] using NUTS classification revision 2016[11].

EUPRO covers participations from the following countries[12]:

- *EU 27 Member States*
- *European Free Trade Association (EFTA)*
- *United Kingdom*
- *EU candidate countries*
- *Potential candidates (all other Western Balkan countries)*
- *European Neighbourhood Policy (ENP)-East countries[13]*
- *European Neighbourhood Policy (ENP)-South countries[13]*
- *Other European countries*
- *Non-European countries*
- *Overseas Countries and Territories linked to the Member States (EU 27 OCT)*
- *Overseas Countries and Territories linked to UK (UK OCT)*

[10] including the analogous territorial descriptions for Switzerland and Norway

[11] History of NUTS, http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/history_nuts (accessed: 24/04/2014)

[12] The classification is based on the typology of Eurostat: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Country_codes and extended by the categories EU 27 OCT and UK OCT. Countries not included in Eurostat's categorisation are added manually to the fitting categories..

[13] https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:European_Neighbourhood_Policy_(ENP)

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

2.5 Quality and accuracy of data

Information on the number of missing values[14]

Table 21: Number and ratio of missing values of FP project data

Variable	Missing values	
	Count	Ratio
rcn	0	0.0%
projectReference	0	0.0%
title	0	0.0%
projAcr	0	0.0%
startDate	130	0.2%
endDate	130	0.2%
totalProjectCosts	938[15]	1.5%
projectECFunding	0	0.0%
fundingSchemeId	0	0.0%
masterCall	44,836[16]	73.3%
subCall	7	0.0%
objective	2	0.0%
projectUrl	35,395	57.9%

Table 22: Number and ratio of missing values of FP participation data

Variable	Missing values	
	Count	Ratio
participId	0	0.0%
rcn	0	0.0%
role	0	0.0%
orgId	0	0.0%
endofParticipation	2,433	0.8%
geoId	5	0.0%
address	821	0.3%
postcode	2,619	0.8%
city	95	0.0%
ecContribution	15,969	5.1%
netEcContribution	66,512	21.1%

[14] This section covers information on the variables of the two main tables in each of the data sets, projects and participations. Information on variables of the remaining tables can be provided upon request.

[15] Those which had total ProjectCost=0 (shown in CORDIS as no data), except for one (projectId= 304096) which had totalProjectCost=0 and projectECfunding=0 (shown in CORDIS as 0€)

[16] No MasterCall data available for FP7, FP7 data only includes subCall data

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

orgUrl	76,735	24.5%
cordisOrganisationId	472	0.2%
vatNumber	54,332	17.4%

Estimation of data quality issues with respect to data acquisition, reliability of retrieving system

In general, the CORDIS database constitutes a reliable resource on all R&D projects and participations. However, in some few cases information is incomplete or inconsistent, as for instance the address data, project costs, project funding and funding on the participants' level.

3 Technical Specifications

3.1 Information on the data base system

Current data base system used

The EUPRO database is realised as Microsoft Access 2016 database.

Planned future technical changes concerning data base system

As part of the RCF Platform, EUPRO is one of the datasets to be incorporated. For that purpose, an API – Application Programming Interface will be developed for the transfer of the data to the Platform. While having a MS Access version of the EUPRO database has its benefits, for server-side web programming, an appropriate relational database management system (RDBMS) and a database server needs to be used. Therefore, a MySQL version for the four modules of the EUPRO database will be provided as well.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

3.2 Technical variable definition

Labelling and data type of all variables[17]

Table 23: Data type of variables providing information about projects

Variable	Data type
projectId	Number
rcn	Number
title	Text
projAcr	Text
startDate	Date
endDate	Date
totalProjectCosts	Number
projectECFunding	Number
fundingSchemeId	Number
MasterCall	Text
subCall	Text
objective	Long Text
projectUrl	Text

Table 24: Data type of variables providing information about participations

Variable	Data type
participId	Number
projectId	Number
role	Text
orgId	Number
endofParticipation	Text
geoId	Number
address	Text
postcode	Text
city	Text
ecContribution	Number
netEcContribution	Number
orgUrl	Text
cordisOrganisationId	Text
vatNumber	Text
nutsAssignmentId	Number

[17] This section covers information on the variables of the two main tables in each of the data sets, projects and participations. Information on variables of the remaining tables can be provided upon request.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

3.3 Description of the Entity Relationship Model

FP 2022 currently consists of 24 tables (Figure 1). The relation between *projects*, *programmes*, *participations*, and project outputs (*reports*, *deliverables*, *publications, patents*) and thematic classifications (*fosTerms*, *topics*) is realised via projectId, which is a unique identifier for each project in the database and identical to the unique identifier for the project in the CORDIS projects database. The thematic classification of projects is provided in the *fosTerms* table, linked by fosCode to the *projects* table. Further information on FP *topics* are linked by topicId to the *projects* table.

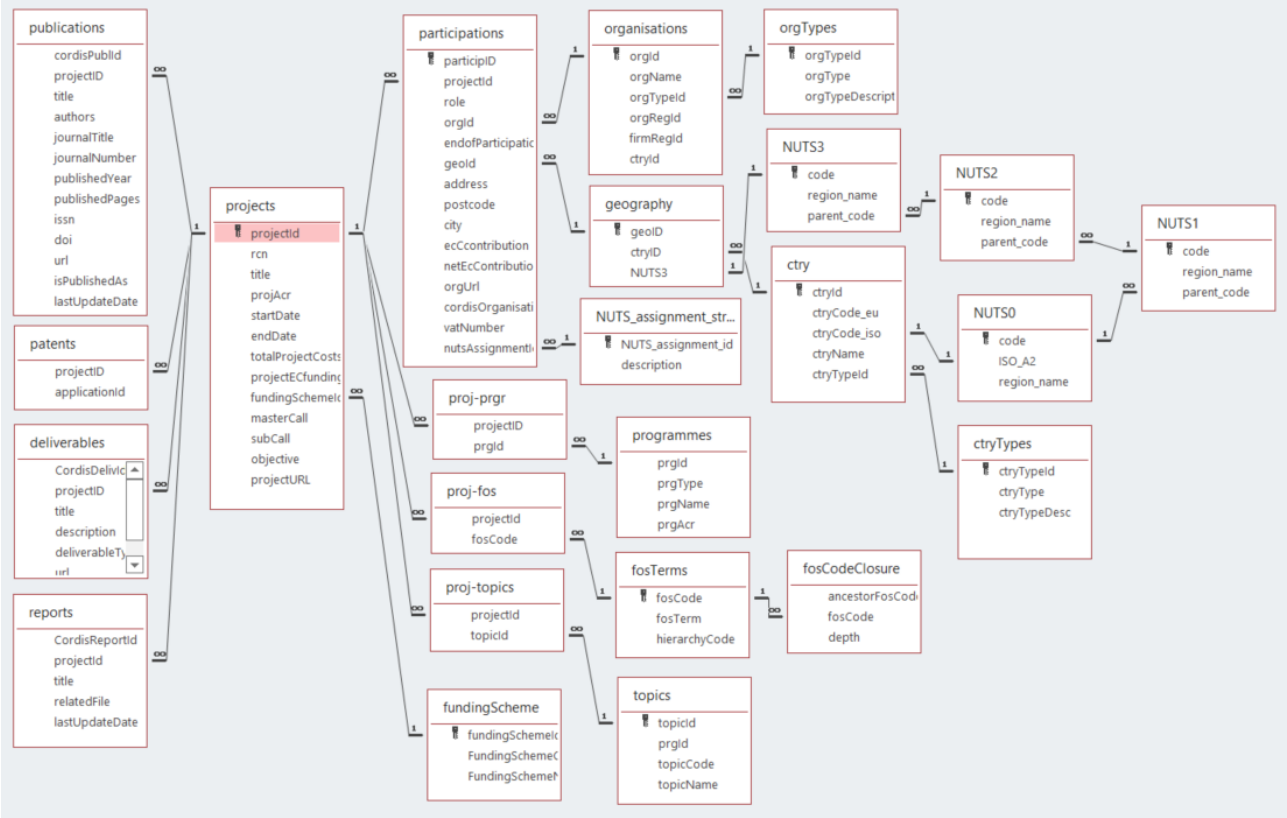
The *participations* table is linked by geoId to the *geography* table, which holds the geocoded data of the participants on city level and is linked (1) to the *regions* table, to provide the respective NUTS 3 code in which the city is located, and (2) to the *ctry* table, which holds the names of the country of origin. By the table *ctryTypes*, countries of origin are assigned to different country categories (EU27, associated countries and third countries).

Every organisation is linked by orgId to the *organisations* table, which provides the link to RISIS-OrgReg, the Register of European Public Research and Higher Education Actors, to RISIS-FirmReg, the Register of businesses included in RISIS firms' datasets, and to the *orgTypes* table, which differentiates the organisations in higher and secondary education establishments, research organisations, public bodies, etc.

Data of the scientific output of each project is comprised in *reports*, *deliverables*, *publications, and patents*.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Figure 1: FP Entity Relationship Model



3.4 Interfaces for access and to other infrastructures

Technical information on interfaces with other infrastructures

Integration of EUPRO within RISIS has been core in order to increase the scientific value of EUPRO for cross-dataset empirical analyses, on the one hand, and to be able to gain from RISIS developed facilities, such as geolocalisation tools, for the further advancement of EUPRO, on the other hand. Inter-operability with other datasets is considered as a key element for the further establishment and sustainable attractivity of EUPRO for new research endeavours, in particular those relating to the investigation of impacts of publicly funded R&D networks.

The following integration activities has been conducted during RISIS II:

- The link to **RISIS-OrgReg** enables the direct inter-linking with other RISIS core datasets featuring public research organisation, in particular LeidenRank and RISIS-ETER. At the moment, organisations included in EUPRO_FP_2022 are linked to RISIS-OrgReg by orgRegId in the *organisations* table. Accordingly, companies included in the RISIS firms' datasets are linked to **RISIS-FirmReg** in the *organisations* table by firmregId.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

- With the application of the new geolocalisation tools developed in RISIS to EUPRO, we have been able to analyse the spatio-temporal development of FP networks in much more flexible manner and integrate it in geographical terms with other RISIS datasets. One promising example in this context has been a combination of EUPRO with the **Nano S&T dataset**, investing Nanoscience and -technology networks in publications and patents (coming from Nano S&T dataset) with Nanoscience and –technology FP networks (coming from EUPRO) (see Villard et al. 2017).
- EUPRO has been inter-linked directly at the organisational level with **RISIS-ETER** in order to analyse the relationships between several characteristics of European Higher Education Institutions (HEIs) coming from RISIS-ETER, and their participation patterns in the EU FP coming from EUPRO. First research endeavors in this direction have been conducted within in RISIS (see Lepori et al. 2015), further research projects in this direction are currently under way.
- **JoREP** has – in contrast to EUPRO – a programme perspective on joint transnational research programmes, i.e. it gathers systematic information on a series of transnational programmes from 2000 to 2014. In EUPRO we have established a direct link (programme identifier) of programmes available in EUPRO and also in JoREP. These are at the current stage EUREKA and JTIs, and enables the relational investigation of networks of performers in these programmes (coming from EUPRO), and programme characteristics (coming from JoREP).
- **RISIS-KNOWMAK** draws upon project data from the EUPRO FP module to compute and visual indicators of knowledge creation in the European Research Area (see Lepori et al. 2017). Projects are annotated[18] with topical classes relevant for Key Enabling Technologies (KET) and Societal Grand Challenges (SGC), include detailed subclasses. The link to RISIS-KNOWMAK has been further extended to include annotation of all projects based on their descriptions, including FP projects outside the scope of RISIS-KNOWMAK as well as EUREKA projects. With the full annotation, alternative indicator definitions may be used, and specific projects may be associated with KETs and SGCs.

Integration with RCF

The EUPRO version that is made available for access to researchers in RISIS is foreseen to be fully incorporated in RCF, under the condition of controlled access

[18] https://gate.ac.uk/projects/knowmak

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

and that security of usage is given (i.e. access for selected users with a concrete research project to the parts of the dataset needed for the research). Note that underlying cleaning and standardisation data (e.g. name variants) will not be made available via RCF. Linking to other datasets in the RCF will be realized via the RISIS registers (providing the respective identifiers to the registers in EUPRO). Most probably, as mentioned above in the planned future technical changes section, there will be an API (Application Programming Interface) developed for the transfer of the EUPRO data to the RCF Platform. Anyhow, the technical issues for incorporation of EUPRO into RCF (e.g. database system, how can a user access which parts of the dataset, etc.) are to be defined in more details in close cooperation with WP4 beginning with autumn.

4 Scientific use cases and main references

Summarizing the research activities of external researchers accessing EUPRO via RISIS, we can distinguish the following main research directions for which EUPRO has been mobilized:

- Observing and characterizing **structure and dynamics of knowledge creation** and networks, disaggregated across different topics and/or geographical spaces (e.g., for climate change, biodiversity, Nanoscience; done by researchers from University of Coimbra, Université Paris-Est Marne-la-Vallée)
- Observing **FP participation patterns and networking of firms** in specific industries (e.g., pharmaceutical and chemical industries, done by researcher from Université Paris-Est Marne-la-Vallée based on the inter-linking of EUPRO with CIB)
- Using EUPRO to quantify and model **impacts of publicly funded R&D networks** on knowledge creation and technological diversification (e.g., done by researchers from Utrecht University, Université Paris-Est Marne-la-Vallée)
- Analysis of **country-specific participation patterns in the FP**, with a special focus on topical orientation and main partners (e.g., done for Israel by researchers from the Samuel Neaman Institute), increasingly as compared to nationally funded projects recorded in the new NATPRO module of EUPRO collecting respective information at country level (see https://zenodo.org/record/6549168#.Y3zQ_H2ZMZU)
- Observing **R&D hot spots** and activities funded by FP projects, at an organizational level in specific topics and geographical locations across Europe (e.g., on marine biotechnology, done by researchers from the EMBRIC project)
- Using real-world network data to test **novel statistical models for dynamic network analysis**, and to test these models in concrete empirical applications (done by researchers from University of Lugano)

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

- Tracing and investigating **characteristics of universities in terms of FP funding**, e.g., relation of FP funding to other university characteristics, disciplinary background of funding, etc. (done by researchers from Vrije University Brussels)

Selected recent references to publications using EUPRO

Barber, M., Guffarth, D. and Scherngell, T. (2015): Determinants of spatial distribution of European aerospace: contrasting R&D and supply. 55th Congress of the European Regional Science Association, 26-29 August, Lisbon, Portugal, 2015

Ciffolilli, A., & Muscio, A. (2018). Industry 4.0: national and regional comparative advantages in key enabling technologies. European Planning Studies, 26(12), 2323-2343.

Freitas, F. and Carrozza, C. (2017): Tracing research and development impacts using geographic data and a FP7 dataset. 2017 Annual Conference of the EU-SPRI Forum, June 7-9, Vienna Austria

Heller-Schuh, B., Reale, E. Scherngell, T., Spinello, A., Varinetti, E., Zahradnik, G. and Zinilli, A. (2022): Insights into project-based R&D funding from RISIS datasets: Some evidence from EFIL and NATPRO, RISIS Policy Brief, Issue 10 / zenodo, DOI: 10.5281/zenodo.6549168

Heller-Schuh, B., Lepori, B. and Neuländtner, M. (2020): Mergers and acquisitions in the public research sector. Toward a comprehensive typology. ***Research Evaluation***, https://doi.org/10.1093/reseval/rvaa015

Heringa, P. W., Hessels, L. K., & van der Zouwen, M. (2016). The influence of proximity dimensions on international research collaboration: an analysis of European water projects. Industry and Innovation, 23(8), 753-772.

Lata, R., Scherngell, T. and Brenner, T. (2015): Integration Processes in European R&D: A comparative spatial interaction approach using project based R&D networks, co-patent networks and co-publication networks, ***Geographical Analysis*** 47, 349-375

Lepori, B, Veglio, V., Heller-Schuh, B., Scherngell, T. and Barber, M. (2015): Participations to European Framework Programs of Higher Education Institutions and their association with organizational characteristics, ***Scientometrics*** 3, 2149-2178

Lepori, B. and Guerini, M. (2017): KNOWMAK Manual. KNOWMAK project deliverable D5.2. https://project.knowmak.eu/wp-content/uploads/2018/02/KNOWMAK-D5.2-KNOWMAK-Manual.pdf

Muscio, A., & Ciffolilli, A. (2018). Technological diversity in Europe: Empirical evidence from agri-food research projects. Regional Studies, 52(3), 374-387.

Neuländtner, M. and Scherngell, T. (2022) R&D networks and their effects on knowledge exploration versus knowledge exploitation: Evidence from a spatial econometric perspective. ***Industry and Innovation***, [forthcoming]

Neuländtner M. and Scherngell T. (2020) Geographical or Relational: What drives technology-specific R&D collaboration networks? ***Annals of Regional Science***, https://doi.org/10.1007/s00168-020-01002-5

Robinson, D., Schoen, A., Laurens, P. and Laredo, P. (2017): Developing global and local STI indicators for profiling the territorial embedding of marine biotechnology research centres. STI 2017, Paris [Fr], 6-8 September

Uhlbach, W.-H., Balland, P.-A. and Scherngell, T. (2017): The effects of the European Union Framework Programmes on the technological diversification of European regions. 2017 Annual Conference of the EU-SPRI Forum, June 7-9, Vienna Austria

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Uhlbach, W.-H., Balland, P.-A. and Scherngell, T. (2017): R&D Policy and Technological Trajectories of Regions: Evidence from the EU Framework Programmes. STI 2017, Paris [Fr], 6-8 September

Uhlbach, W. H., Balland, P. A., and Scherngell, T. (2022) Public R&D funding and new regional specialisations: The contingent role of technological relatedness. ***Industry and Innovation***, [forthcoming]

Wanzenböck, I., Scherngell, T. and Brenner, T. (2014): Embeddedness of regions in European knowledge networks. A comparative analysis of inter-regional R&D collaborations, co-patents and co- publications, ***The Annals of Regional Science*** 53, 337-368

Wanzenböck, I., Scherngell, T. and Lata, R. (2015): Embeddedness of European regions in EU funded R&D networks: A spatial econometric perspective, ***Regional Studies*** 49, 1685-1705

Wanzenböck, I. and Piribauer, P. (2017): R&D networks and regional knowledge production in Europe. Evidence from a space-time model. ***Papers in Regional Science*** [doi 10.1111/pirs.12236]

Wanzenböck, I., Scherngell, T. and Dünser, M. (2017): Impacts of EU funded R&D networks on the generation of Key Enabling Technologies: Empirical evidence from a regional perspective, EMAEE Conference 2017, Strasbourg [FR], 31 May 2017

Wanzenböck, I., Scherngell, T. and Dünser, M. (2017): RTI policy for co-creation activities and its effects on Key Enabling Technologies: Regional evidence from the 7th EU Framework Programme. 2017 Annual Conference of the EU-SPRI Forum, June 7-9, Vienna Austria

Wanzenböck, I., Neuländtner, M. and Scherngell, T. (2020): Impacts of EU funded R&D networks on the generation of Key Enabling Technologies: Empirical evidence from a regional perspective. ***Papers in Regional Science***, https://doi.org/10.1111/pirs.12473

Villard, L., Perruchas, F., Scherngell, T., Barber M. and Laredo, P. (2017): The role of European Programmes in the European distribution of knowledge, the case of nanoscience and technology. STI 2017, Paris [Fr], 6-8 September

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Appendix

Table A 1: List of countries participating in FP7 and H2020

Country	Code (eu)	Type of Country
Switzerland	CH	EFTA
Iceland	IS	EFTA
Liechtenstein	LI	EFTA
Norway	NO	EFTA
Georgia	GE	ENP-East
Ukraine	UA	ENP-East
Armenia	AM	ENP-East
Azerbaijan	AZ	ENP-East
Belarus	BY	ENP-East
Moldova	MD	ENP-East
Tunisia	TN	ENP-South
Algeria	DZ	ENP-South
Egypt	EG	ENP-South
Israel	IL	ENP-South
Jordan	JO	ENP-South
Lebanon	LB	ENP-South
Libya	LY	ENP-South
Morocco	MA	ENP-South
Palestine	PS	ENP-South
Syria	SY	ENP-South
Poland	PL	EU
Malta	MT	EU
Netherlands	NL	EU
Estonia	EE	EU
Portugal	PT	EU
Romania	RO	EU
Sweden	SE	EU
Slovenia	SI	EU
Slovakia	SK	EU
Austria	AT	EU
Belgium	BE	EU
Bulgaria	BG	EU
Cyprus	CY	EU
Denmark	DK	EU
Germany	DE	EU
Czechia	CZ	EU
Spain	ES	EU
Finland	FI	EU
France	FR	EU
Ireland	IE	EU

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Croatia	HR	EU
Greece	EL	EU
Hungary	HU	EU
Latvia	LV	EU
Luxembourg	LU	EU
Lithuania	LT	EU
Italy	IT	EU
Wallis and Futuna	WF	EU 27 OCT
Aruba	AW	EU 27 OCT
Bonaire; Saint Eustatius and Saba	BQ	EU 27 OCT
Curaçao	CW	EU 27 OCT
Albania	AL	EU candidate
Turkey	TR	EU candidate
Serbia	RS	EU candidate
Montenegro	ME	EU candidate
North Macedonia	MK	EU candidate
Kosovo	XK	EU potential candidate
Bosnia and Herzegovina	BA	EU potential candidate
Qatar	QA	Non-European countries
Pakistan	PK	Non-European countries
San Marino	SM	Non-European countries
Saudi Arabia	SA	Non-European countries
Seychelles	SC	Non-European countries
Sudan	SD	Non-European countries
Singapore	SG	Non-European countries
Rwanda	RW	Non-European countries
Senegal	SN	Non-European countries
Somalia	SO	Non-European countries
El Salvador	SV	Non-European countries
Tajikistan	TJ	Non-European countries
Eswatini	SZ	Non-European countries
Togo	TG	Non-European countries
Thailand	TH	Non-European countries
Philippines	PH	Non-European countries
Malaysia	MY	Non-European countries
Suriname	SR	Non-European countries
Niger	NE	Non-European countries
Taiwan	TW	Non-European countries
Trinidad and Tobago	TT	Non-European countries
Mauritius	MU	Non-European countries
Maldives	MV	Non-European countries
Malawi	MW	Non-European countries
Mexico	MX	Non-European countries
Mozambique	MZ	Non-European countries
Namibia	NA	Non-European countries
New Caledonia	NC	Non-European countries

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Papua New Guinea	PG	Non-European countries
Nigeria	NG	Non-European countries
Nicaragua	NI	Non-European countries
Nepal	NP	Non-European countries
New Zealand	NZ	Non-European countries
Oman	OM	Non-European countries
Panama	PA	Non-European countries
Peru	PE	Non-European countries
French Polynesia	PF	Non-European countries
Mauritania	MR	Non-European countries
Dominican Republic	DO	Non-European countries
Democratic Republic of the Congo	CD	Non-European countries
Guatemala	GT	Non-European countries
Guinea	GN	Non-European countries
The Gambia	GM	Non-European countries
Greenland	GL	Non-European countries
Ghana	GH	Non-European countries
French Guiana	GF	Non-European countries
Gabon	GA	Non-European countries
Fiji	FJ	Non-European countries
Guyana	GY	Non-European countries
Ecuador	EC	Non-European countries
Hong Kong	HK	Non-European countries
Cape Verde	CV	Non-European countries
Cuba	CU	Non-European countries
Costa Rica	CR	Non-European countries
Colombia	CO	Non-European countries
China	CN	Non-European countries
Cameroon	CM	Non-European countries
Chile	CL	Non-European countries
Turkmenistan	TM	Non-European countries
Congo	CG	Non-European countries
Ethiopia	ET	Non-European countries
Cambodia	KH	Non-European countries
Macao	MO	Non-European countries
Myanmar/Burma	MM	Non-European countries
Mali	ML	Non-European countries
Marshall Islands	MH	Non-European countries
Madagascar	MG	Non-European countries
Lesotho	LS	Non-European countries
Sri Lanka	LK	Non-European countries
Laos	LA	Non-European countries
Kazakhstan	KZ	Non-European countries
Guinea-Bissau	GW	Non-European countries
South Korea	KR	Non-European countries
Central African Republic	CF	Non-European countries

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Kyrgyzstan	KG	Non-European countries
Kenya	KE	Non-European countries
Japan	JP	Non-European countries
Jamaica	JM	Non-European countries
Iran	IR	Non-European countries
India	IN	Non-European countries
Indonesia	ID	Non-European countries
Haiti	HT	Non-European countries
Honduras	HN	Non-European countries
Kuwait	KW	Non-European countries
South Africa	ZA	Non-European countries
Sierra Leone	SL	Non-European countries
Côte d'Ivoire	CI	Non-European countries
Paraguay	PY	Non-European countries
Mongolia	MN	Non-European countries
Liberia	LR	Non-European countries
Iraq	IQ	Non-European countries
Grenada	GD	Non-European countries
Djibouti	DJ	Non-European countries
Belize	BZ	Non-European countries
Canada	CA	Non-European countries
Zambia	ZM	Non-European countries
São Tomé and Príncipe	ST	Non-European countries
Yemen	YE	Non-European countries
Samoa	WS	Non-European countries
Vanuatu	VU	Non-European countries
Vietnam	VN	Non-European countries
Venezuela	VE	Non-European countries
Uzbekistan	UZ	Non-European countries
Uruguay	UY	Non-European countries
United States	US	Non-European countries
Uganda	UG	Non-European countries
Tanzania	TZ	Non-European countries
Zimbabwe	ZW	Non-European countries
Benin	BJ	Non-European countries
Botswana	BW	Non-European countries
Bhutan	BT	Non-European countries
Brazil	BR	Non-European countries
Solomon Islands	SB	Non-European countries
Brunei	BN	Non-European countries
Chad	TD	Non-European countries
Burundi	BI	Non-European countries
Bahrain	BH	Non-European countries
Burkina Faso	BF	Non-European countries
Afghanistan	AF	Non-European countries
Barbados	BB	Non-European countries

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Australia	AU	Non-European countries
Argentina	AR	Non-European countries
Angola	AO	Non-European countries
United States Minor Outlying Islands	UM	Non-European countries
United Arab Emirates	AE	Non-European countries
Bolivia	BO	Non-European countries
Bangladesh	BD	Non-European countries
Holy See/Vatican City	VA	Other European countries
Faeroe Islands	FO	Other European countries
Monaco	MC	Other European countries
Russia	RU	Other European countries
United Kingdom	UK	UK
Gibraltar	GI	UK OCT
British Virgin Islands	VG	UK OCT
Anguilla	AI	UK OCT
Jersey	JE	UK OCT