Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Consiglio Nazionale delle Ricerche

Exploring research infrastructure data through advanced statistical applications

FOSSR Online Training

26-27 September 2024

FOSSR

Fostering Open Science in Social Science Research

Innovative tools and services to investigate economic and societal change

Aim of the course

The main aim of this course is to provide an **overview of the opportunities for exploiting shared data within Research Infrastructures (RI)**. Specifically, the course aims to raise awareness of these opportunities and encourage the implementation of **data-driven approaches** through the application of advanced statistical techniques.

The focus is primarily on the **RISIS European Infrastructure**, on topics related to Science, Technology, and Innovation (STI), one of the three infrastructures that initiated the **FOSSR project**, and on advanced statistical techniques that FOSSR is treating in its research work packages. Drawing inspiration from the data available within RISIS, the course will explore aspects of **data access, interoperability and processing** in depth.

The course will guide participants through the process of **engaging with a RI and making reasoned use of its resources for research purposes**. Furthermore, it will present the application of **advanced statistical techniques on infrastructure data, particularly *Network Models* and *Bayesian Modeling***.

Schedule - Day 1 – Thursday 26 September 2024

10:00 – 12:30

Module 1. Research Infrastructures and Open Science

The module will outline the contours of the new paradigm for scientific knowledge production, emphasizing openness, the value of collaboration, and data availability. It will explore the features of data-intensive science, which underpin the creation and strengthening of Research Infrastructures (RI). The characteristics of these sociotechnical platforms will be discussed by analyzing both the Italian and European contexts, with reference to FOSSR's efforts in developing the Italian Open Cloud for Social Sciences.

--- *Lunch break* ---

14:00 – 16:00

Module 2. Accessing and querying interoperable RI data

The module will present examples of accessing and querying data from research infrastructures, highlighting the importance of data interoperability. Specific cases will illustrate the use of the RISIS infrastructure for science, technology and innovation studies. We will present cases in which research questions are addressed differently depending on the nature of the data, analysing the ways of managing datasets and their enrichment with data external to the infrastructure. The content presented will highlight the value of shared database access.

Schedule - Day 2 – Friday 27 September 2024

10:00-12:30

Module 3. Network models applied to RI data

The module aims to illustrate the basic concepts and statistical measures of network science and provide an overview of the main statistical network models. The module will conclude with two applications where networks are analysed using data from research infrastructures. The two applications that will be covered in this module are:

- *Application 1: Complex networks and academic project funding*
- *Application 2: Research collaborations and research productivity*

--- Lunch break ---

14:00 – 16:30

Module 4. Causal Bayesian networks and applications to RI data

The module focuses on Bayesian networks as a tool for modelling complex causal relationships. A comparison between causal Bayesian networks and potential outcomes is carried out to highlight how the two approaches can be implemented synergistically. The module will include two applications that employ causal Bayesian networks on research infrastructure data. The two applications covered in this module are:

Application 1: Research collaborations and research productivity
Application 2: Remote working and firm revenues during Covid

Presentation of the team

Teachers of the course are five researchers from CNR-IRCrES:

- Dr. Andrea Orazio Spinello (*course curator*)
- Dr. Emanuela Varinetti
- Dr. Lucio Morettini
- Dr. Antonio Zinilli
- Dr. Lorenzo Giammei

In addition to the course curator, the Organizing Committee is composed of CNR-IRCrES personnel:

- Dr. Serena Fabrizio
- Dr. Alessia Fava
- Dr. Rita Giuffredi
- Dr. Alessandra Maria Stilo (FOSSR communication WP leader).

This course is organized within the framework of the FOSSR project, WP8 Capacity building and Training, Task 8.1 – Online Training Courses.
WP8 leader: Dr. Andrea Orazio Spinello (CNR-IRCrES); WP8 Task 8.1 led by Dr. Valentina Tudisca (CNR-IRPPS).
Training evaluation committee: Dr. Adriana Valente (CNR-IRPPS), Dr. Emanuela Reale and the WP8 leader (CNR-IRCrES).

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Consiglio Nazionale delle Ricerche

Module I: Research Infastructures and Open Science

Part I: Value and potential of Research Infrastructures in the Social Sciences

FOSSR
Fostering Open Science in Social Science Research
Innovative tools and services to investigate economic and societal change

Andrea Orazio Spinello
Researcher at CNR-IRCrES

«Exploring research infrastructure data through advanced statistical applications»

FOSSR Online Training

26-27 September 2024

A new frontier for knowledge production

The nature of scientific research processes is increasingly transitioning toward a **peculiar mode of knowledge production** (Gibbons et al., 1994) characterized by:

- Emergence of diverse **sites for knowledge production and circulation**, fostering interactions among researchers previously hindered by physical and technical constraints;
- Emphasis on **transdisciplinarity**;
- Innovative **research designs** leveraging technological advancements in the scientific field.

Furthermore, due to the technical evolution of the Internet, the **relationship between information and communication technologies and the scientific world** has taken on new distinctive features.

The **exchange of data, information, and research results** is now commonplace and occurs through personal interactions or by leveraging collaboration platforms.

A path for science under a... «data deluge»

A new paradigm is emerging based on the **sharing of information wealth** and the intensity of data (the so-called “**data deluge**”) with a greater potential for the resolution of scientific issues (*The Fourth Paradigm*, Hey et al., 2009).

The traditional static view of the scientific process, with rigid roles and disciplines, is being replaced.

Instead, we see the rise of **distributed and collaborative knowledge networks**, where researchers’ skills interconnect (Nielsen, 2011).

The progressive institutionalization of **research data as a common and shared good** has given relevance to the **concept of openness** in terms of data accessibility oriented towards ever wider audiences of researchers and society (Boulton et al., 2012).

Science as an open enterprise

Knowledge production is **no longer confined solely to academic institutions**. Universities now **collaborate** more closely with civil society, sharing knowledge, skills, and research outcomes to maximize the benefits of scientific research.

➢ Royal Society's report *Science as an open enterprise* (Boulton et al., 2012) is very close to a Manifesto for the new path for science:

- There's a call for **greater openness among scientists**, both in terms of collaboration and **engagement with society**;
- The **collection, analysis, and communication** of scientific data are increasingly recognized as valuable;
- **Common standards** are encouraged for sharing information to make it widely accessible;
- Data should be published in **reusable formats;**
- Researchers need expertise in managing **digital data;**
- Appropriate tools are essential for **analyzing large datasets**.

Data-intensive science and networked science

The term **‘data-intensive’** characterizes a research approach focused on the value of data. In this approach, hypotheses are not only tested through data collection and analysis, but increasingly, research outputs result from combining and **extracting existing and accessible data sets**.

Collaborations and networking, enabled by new technologies, highlight the evolution of scientific practice as a **large collaboration** in which scientists share and build a vast common good of information. Scientific problem-solving processes are increasingly based on a new approach characterized by collaborative methods that leverage cognitive diversity and the wealth of information and skills available globally through the Internet.

The effective implementation of this collaboration model is founded on **‘modularization’** of expertise**.** The collaboration thus outlined becomes self-stimulating and is characterized by dynamics of **‘planned serendipity’** allowing for faster resolution of many previously unsolvable problems by directing open questions to more qualified individuals (Nielsen, 2011).

The mobilization towards Open Science

Building upon these foundations (data deluge and expanded networking), an increasingly ‘open’ scientific process takes shape, appropriately named **Open Science -** an ‘umbrella term’ that:

- Captures the trend toward making scientific research and the dissemination of data and results **accessible at all levels**, from citizens and amateur scientists to industry professionals.
- Encompasses practices such as **openly publishing research**, advocating for open access, encouraging scientists to engage in **open notebook science** (where raw data collected during research phases are shared), and facilitating the publication and dissemination of scientific information.
- Extends beyond researchers’ willingness to share; **it also involves research organizations, funding entities, and the general public in the domain of science**, influencing their approach toward the knowledge production.
- Is not an entirely novel concept in the scientific landscape but rather the **result of a long-term evolutionary process in scientific conduct**. Recent years have emphasized this shift due to enhanced infrastructural capabilities linked to information technologies.

Open science in SSH between knowledge sharing and social impact

The concept of **'open science'** encompassing **knowledge sharing** within scientific communities and its interaction between science and society, holds a crucial role in the contemporary science, especially in the fields of **Social Sciences and Humanities (SSH)**.

Within scientific endeavors, open sharing of knowledge, including **data and resources**, has the potential to enable the exploration of **new research questions**, drive **interdisciplinary collaboration** and foster the exploration of **innovative analytical tools**, thereby positively influencing research processes in general.

From a societal standpoint, it becomes important to ensure **research findings are accessible to policymakers**, thus amplifying the **impact of scientific work on decisions that affect citizens**.

E.g, Open data availability can be exploited for socio-economic policies through collaborative utilization by researchers and the interest of policymakers.

Resistances and critical aspects related to open data

Especially in the field of SSH, scientists are still reluctant to share their data. The process of opening data still requires **promotion through mechanisms such as the obligation** imposed by national or international funding agencies (Schmidt et al., 2016; Chawinga & Zinn, 2019).

In the context of open data, there are **key dimensions to take into account**: users, legal aspects, dissemination methods, access standards, management and conservation, sharing scale, material references, and related benefits (Pasquetto et al., 2015).

Commonly used **solutions for access and reuse** include adopting an *Open Data Commons license*, which mandates source citation and considerations for privacy and ethical restrictions. These approaches aim to strike a balance between openness and responsible data handling.

Data handling, opening and GDPR rules

GDPR-compliant data refers to data that adheres to the requirements outlined in the **General Data Protection Regulation (GDPR).**

The GDPR applies to **personal data of individuals** within the European Union (EU) and the European Economic Area (EEA). It covers both *data controllers* (organizations that collect and process data) and *data processors* (entities that process data on behalf of controllers).

GDPR grants individuals several **rights regarding their personal data**, including the right to access, rectify, and erase their data. It also includes the "right to be forgotten," allowing individuals to request the deletion of their data from an organization's records.

Data Protection Principles: Lawfulness, Fairness, and Transparency; Purpose Limitation; Data Minimization; Accuracy; Storage Limitation; Integrity and Confidentiality.

Research Infrastructures as a key component for the development of open science

A significant boost to the development of 'open science' comes from the **creation and enhancement of Research Infrastructures (RIs)**. The practice of 'open science,' facilitated by RIs, is realized through activities enabling researchers or interested parties to access data, tools and services.

What are RIs? A simple definition could be:

"RIs are sociotechnical platforms (composed of data, services, tools and people) crucial for conducting high-quality research, often based on data of considerable size and complexity sourced from diverse origins".

The objective of RIs extends **beyond mere scientific data sharing**, encompassing the provision of **innovative tools and services** for research, fostering initiatives to **build communities**, and engaging **social and political actors** (public engagement).

Research Infrastructures a new (but old) concept

Ris has become increasingly popular in research policy literature since the start of the 21st century. Additionally, it has been a topic of discussions regarding research funding (Hallonsten and Cramer, 2020; Franssen, 2020). However, they actually represent, in various forms, a **continuous thread in the way knowledge has been produced** for more than two thousand years.

The first infrastructure built to serve the advancement of knowledge is generally identified as the Mouseion erected in Alexandria, Egypt, in 307 BC by Ptolemy I. Following the example of the **Alexandrian library**, since the beginning of modern science, certain unique ‘research supports’ have been created and perceived by scientists as enablers of the knowledge production process, such as physical libraries and data collections.

It is important to note from the outset that there is **no single accepted and shared definition of RI** in the literature (Renschler et al., 2013).

Research Infrastructures as a political construct

Definitions of RI tend to be **operational** and **influenced by actors seeking funding** for their specific infrastructure projects. Some scholars contend that the RI concept is primarily a **political construct**, and the labeling of RIs serves as a tool for governing and funding research by public authorities (Franssen, 2020).

In its Regulation (EU) 2021/695 establishing Horizon Europe,, RIs are defined as
"**facilities** that provide **resources** and **services** for the **research communities** to conduct research and foster innovation in their fields, including the associated human resources, major equipment or sets of instruments; knowledge-related facilities such as collections, archives or scientific data infrastructures; computing systems, communication networks and any other infrastructure of a unique nature and **open to external users**, essential to achieve excellence in R&I; they may, where relevant, be used beyond research, for example for education or public services and they may be **single sited, virtual or distributed**".

Calibration of RI references in the field of SSH

In the domain of Social Sciences and Humanities (SSH), research infrastructures encompass various categories, including **academic material collections**, **data repositories**, **archives**, **platforms**, and **structured/interlinked research databases**. Examples include national and international statistical facilities, research data service facilities, and survey-led studies.

These platforms serve as long-term institutions, accessible to scientific communities and other stakeholders, supporting collaboration and data sharing. They can include, beyond data, services, tools, instruments for collaboration, related activities of training and engagement.

Further classification may be relevant when differentiating between ***project-driven RIs (as sets of databases)***, derived from specific projects limited in time and scope, and **RIs of European interest or of international standing.**

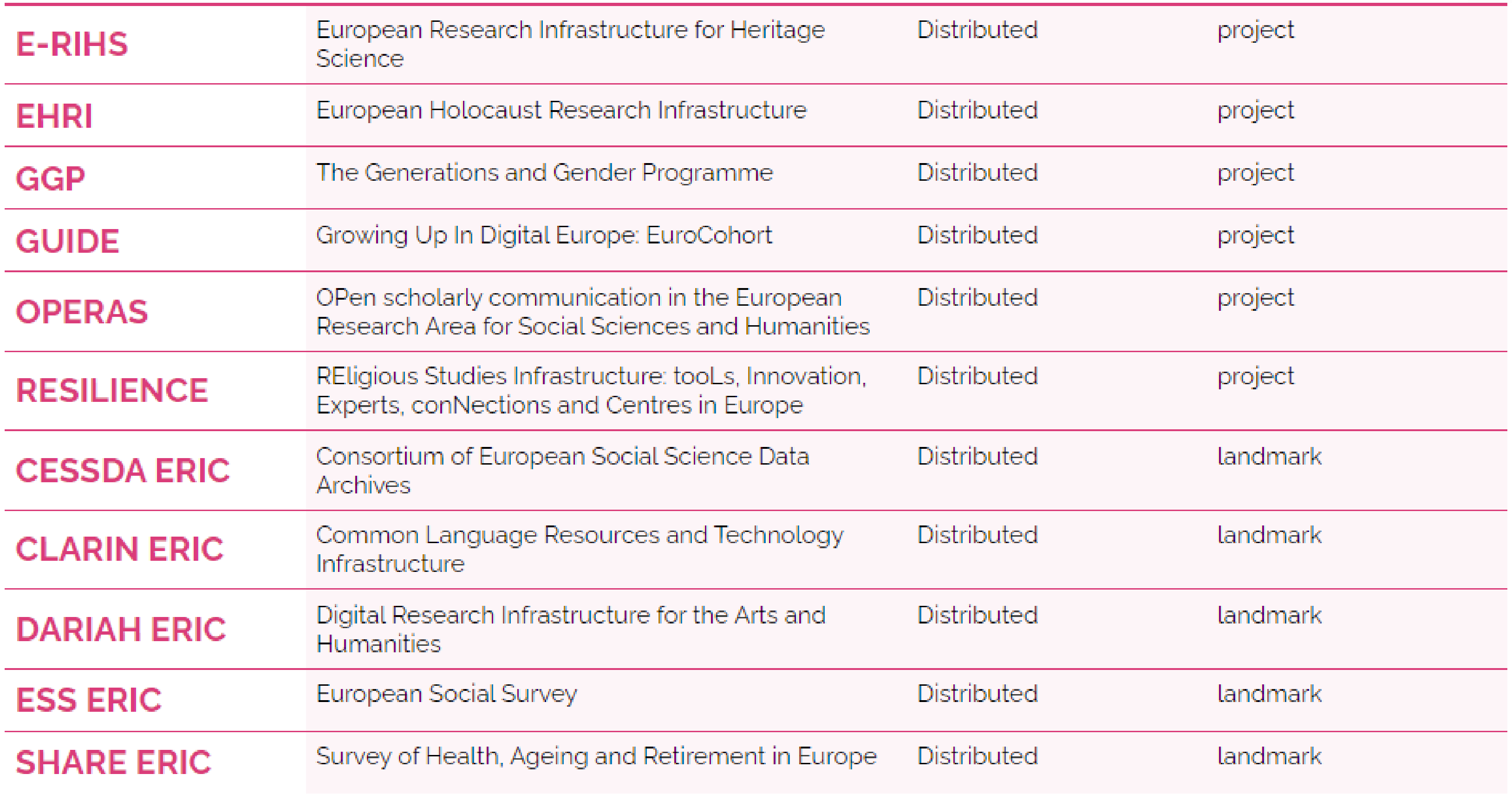
ESFRI projects and landmarks

At the European level, there is a growing presence of institutionalized RIs characterized by **openness to the research community, regular updates, and interoperability features**.

ESFRI (*European Strategy Forum on Research Infrastructures*) identifies research infrastructures of pan-European interest across all scientific areas, including **projects and landmarks**.

- **Projects** are selected based on scientific excellence and maturity, and they are included in the ESFRI roadmap to emphasize their strategic importance for the European research infrastructure system and support timely implementation.
- **Landmarks** are already developed infrastructures that play a crucial role in competitiveness within the European research space. They require continuous support for completion, operation, and updates in line with optimal management and maximum return on investment.

ESFRI projects and landmarks in the area «Social and Cultural Innovation»



E-RIHS	European Research Infrastructure for Heritage Science	Distributed	project
EHRI	European Holocaust Research Infrastructure	Distributed	project
GGP	The Generations and Gender Programme	Distributed	project
GUIDE	Growing Up In Digital Europe: EuroCohort	Distributed	project
OPERAS	OPen scholarly communication in the European Research Area for Social Sciences and Humanities	Distributed	project
RESILIENCE	REligious Studies Infrastructure: tooLs, Innovation, Experts, conNections and Centres in Europe	Distributed	project
CESSDA ERIC	Consortium of European Social Science Data Archives	Distributed	landmark
CLARIN ERIC	Common Language Resources and Technology Infrastructure	Distributed	landmark
DARIAH ERIC	Digital Research Infrastructure for the Arts and Humanities	Distributed	landmark
ESS ERIC	European Social Survey	Distributed	landmark
SHARE ERIC	Survey of Health, Ageing and Retirement in Europe	Distributed	landmark

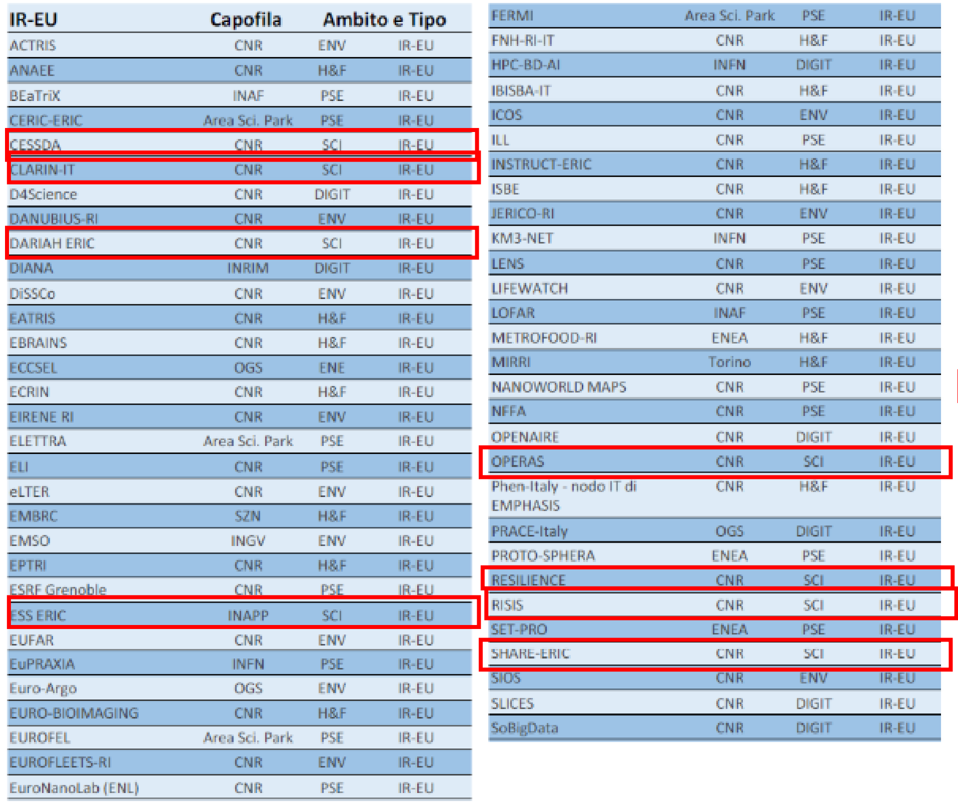
PNIR – The Italian National Plan for Research Infrastructures

In Italy, the **PNIR (National Plan for Research Infrastructures)** is a strategic document that provides an **overview of Italy's research infrastructure landscape (national, European or global RIs).** These infrastructures are affiliated with public research organizations and universities.
The PNIR offers strategic guidance for research infrastructure policies, aiming to enhance the overall quality of Italian research and its international competitiveness

Eight criteria that have been considered to identify RIs of **high/medium/low relevance** in Italy:

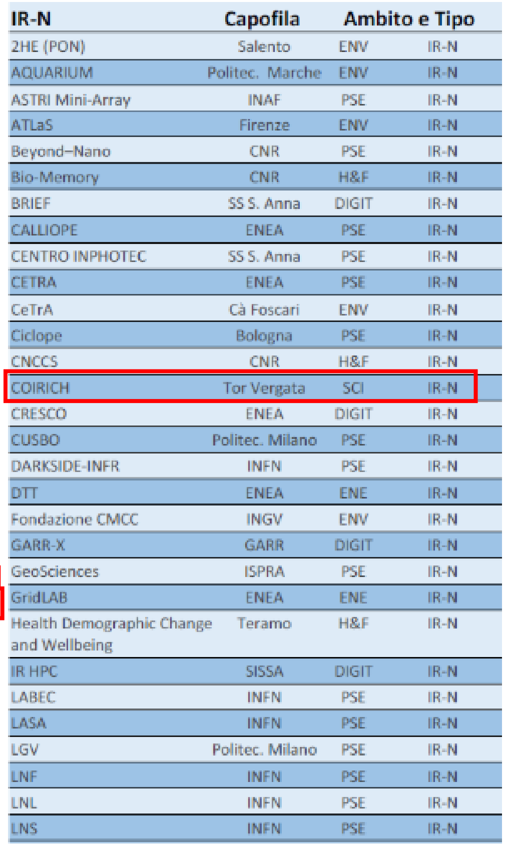
- Scientific excellence
- Socio-economic impact
- Critical analysis of history and prospects
- Completeness of access policies
- International relations and pan-European relevance
- Political commitment and financial support from participating countries
- Governance, management, and human resource management
- Financial aspects

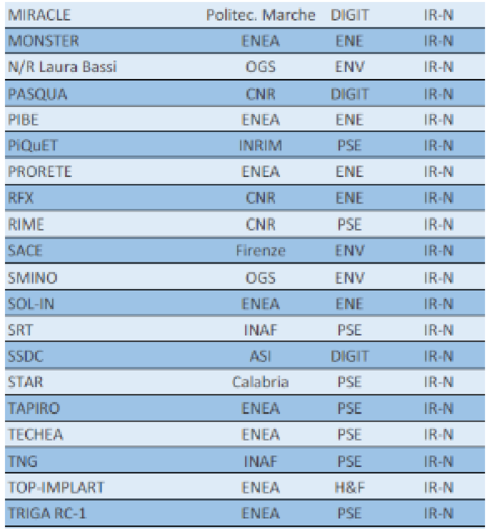
Research Infrastructures recognized in the PNIR 2021-2027



IR-EU	Capofila	Ambito e Tipo	
ACTRIS	CNR	ENV	IR-EU
ANAEE	CNR	H&F	IR-EU
BEaTriX	INAF	PSE	IR-EU
CERIC-ERIC	Area Sci. Park	PSE	IR-EU
CESSDA	CNR	SCI	IR-EU
CLARIN-IT	CNR	SCI	IR-EU
D4Science	CNR	DIGIT	IR-EU
DANUBIUS-RI	CNR	ENV	IR-EU
DARIAH ERIC	CNR	SCI	IR-EU
DIANA	INRIM	DIGIT	IR-EU
DiSSCo	CNR	ENV	IR-EU
EATRIS	CNR	H&F	IR-EU
EBRAINS	CNR	H&F	IR-EU
ECCSEL	OGS	ENE	IR-EU
ECRIN	CNR	H&F	IR-EU
EIRENE RI	CNR	ENV	IR-EU
ELETTRA	Area Sci. Park	PSE	IR-EU
ELI	CNR	PSE	IR-EU
eLTER	CNR	ENV	IR-EU
EMBRC	SZN	H&F	IR-EU
EMSO	INGV	ENV	IR-EU
EPTRI	CNR	H&F	IR-EU
ESRF Grenoble	CNR	PSE	IR-EU
ESS ERIC	INAPP	SCI	IR-EU
EUFAR	CNR	ENV	IR-EU
EuPRAXIA	INFN	PSE	IR-EU
Euro-Argo	OGS	ENV	IR-EU
EURO-BIOIMAGING	CNR	H&F	IR-EU
EUROFEL	Area Sci. Park	PSE	IR-EU
EUROFLEETS-RI	CNR	ENV	IR-EU
EuroNanoLab (ENL)	CNR	PSE	IR-EU
FERMI	Area Sci. Park	PSE	IR-EU
FNH-RI-IT	CNR	H&F	IR-EU
HPC-BD-AI	INFN	DIGIT	IR-EU
IBISBA-IT	CNR	H&F	IR-EU
ICOS	CNR	ENV	IR-EU
ILL	CNR	PSE	IR-EU
INSTRUCT-ERIC	CNR	H&F	IR-EU
ISBE	CNR	H&F	IR-EU
JERICO-RI	CNR	ENV	IR-EU
KM3-NET	INFN	PSE	IR-EU
LENS	CNR	PSE	IR-EU
LIFEWATCH	CNR	ENV	IR-EU
LOFAR	INAF	PSE	IR-EU
METROFOOD-RI	ENEA	H&F	IR-EU
MIRRI	Torino	H&F	IR-EU
NANOWORLD MAPS	CNR	PSE	IR-EU
NFFA	CNR	PSE	IR-EU
OPENAIRE	CNR	DIGIT	IR-EU
OPERAS	CNR	SCI	IR-EU
Phen-Italy - nodo IT di EMPHASIS	CNR	H&F	IR-EU
PRACE-Italy	OGS	DIGIT	IR-EU
PROTO-SPHERA	ENEA	PSE	IR-EU
RESILIENCE	CNR	SCI	IR-EU
RISIS	CNR	SCI	IR-EU
SET-PRO	ENEA	PSE	IR-EU
SHARE-ERIC	CNR	SCI	IR-EU
SIOS	CNR	ENV	IR-EU
SLICES	CNR	DIGIT	IR-EU
SoBigData	CNR	DIGIT	IR-EU

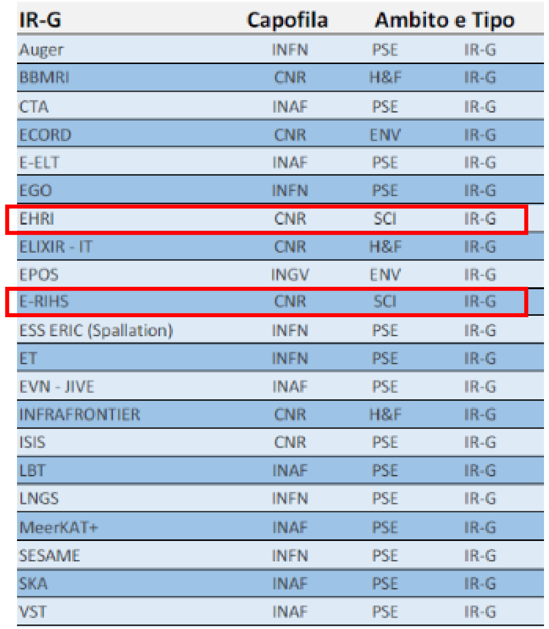
Tabella 4: Elenco delle IR di categoria "europea"





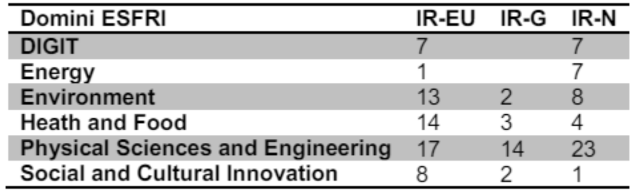
IR-N	Capofila	Ambito e Tipo	
2HE (PON)	Salento	ENV	IR-N
AQUARIUM	Politec. Marche	ENV	IR-N
ASTRI Mini-Array	INAF	PSE	IR-N
ATLaS	Firenze	ENV	IR-N
Beyond–Nano	CNR	PSE	IR-N
Bio-Memory	CNR	H&F	IR-N
BRIEF	SS S. Anna	DIGIT	IR-N
CALLIOPE	ENEA	PSE	IR-N
CENTRO INPHOTEC	SS S. Anna	PSE	IR-N
CETRA	ENEA	PSE	IR-N
CeTrA	Cà Foscari	ENV	IR-N
Ciclope	Bologna	PSE	IR-N
CNCCS	CNR	H&F	IR-N
COIRICH	Tor Vergata	SCI	IR-N
CRESCO	ENEA	DIGIT	IR-N
CUSBO	Politec. Milano	PSE	IR-N
DARKSIDE-INFR	INFN	PSE	IR-N
DTT	ENEA	ENE	IR-N
Fondazione CMCC	INGV	ENV	IR-N
GARR-X	GARR	DIGIT	IR-N
GeoSciences	ISPRA	PSE	IR-N
GridLAB	ENEA	ENE	IR-N
Health Demographic Change and Wellbeing	Teramo	H&F	IR-N
IR HPC	SISSA	DIGIT	IR-N
LABEC	INFN	PSE	IR-N
LASA	INFN	PSE	IR-N
LGV	Politec. Milano	PSE	IR-N
LNF	INFN	PSE	IR-N
LNL	INFN	PSE	IR-N
LNS	INFN	PSE	IR-N
MIRACLE	Politec. Marche	DIGIT	IR-N
MONSTER	ENEA	ENE	IR-N
N/R Laura Bassi	OGS	ENV	IR-N
PASQUA	CNR	DIGIT	IR-N
PIBE	ENEA	ENE	IR-N
PiQuET	INRIM	PSE	IR-N
PRORETE	ENEA	ENE	IR-N
RFX	CNR	ENE	IR-N
RIME	CNR	PSE	IR-N
SACE	Firenze	ENV	IR-N
SMINO	OGS	ENV	IR-N
SOL-IN	ENEA	ENE	IR-N
SRT	INAF	PSE	IR-N
SSDC	ASI	DIGIT	IR-N
STAR	Calabria	PSE	IR-N
TAPIRO	ENEA	PSE	IR-N
TECHEA	ENEA	PSE	IR-N
TNG	INAF	PSE	IR-N
TOP-IMPLART	ENEA	H&F	IR-N
TRIGA RC-1	ENEA	PSE	IR-N

Tabella 5: Elenco delle IR di categoria "Nazionale"[37]



IR-G	Capofila	Ambito e Tipo	
Auger	INFN	PSE	IR-G
BBMRI	CNR	H&F	IR-G
CTA	INAF	PSE	IR-G
ECORD	CNR	ENV	IR-G
E-ELT	INAF	PSE	IR-G
EGO	INFN	PSE	IR-G
EHRI	CNR	SCI	IR-G
ELIXIR - IT	CNR	H&F	IR-G
EPOS	INGV	ENV	IR-G
E-RIHS	CNR	SCI	IR-G
ESS ERIC (Spallation)	INFN	PSE	IR-G
ET	INFN	PSE	IR-G
EVN - JIVE	INAF	PSE	IR-G
INFRAFRONTIER	CNR	H&F	IR-G
ISIS	CNR	PSE	IR-G
LBT	INAF	PSE	IR-G
LNGS	INFN	PSE	IR-G
MeerKAT+	INAF	PSE	IR-G
SESAME	INFN	PSE	IR-G
SKA	INAF	PSE	IR-G
VST	INAF	PSE	IR-G

Tabella 6: Elenco delle IR di categoria "Globale"[38]



Domini ESFRI	IR-EU	IR-G	IR-N
DIGIT	7		7
Energy	1		7
Environment	13	2	8
Heath and Food	14	3	4
Physical Sciences and Engineering	17	14	23
Social and Cultural Innovation	8	2	1

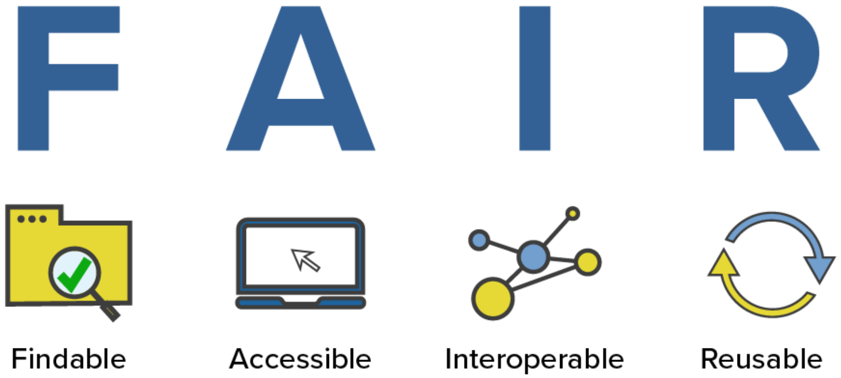
Tabella 3: distribuzione delle IR censite in base al dominio ESFRI

The importance of RIs for high-quality research

Research infrastructures play a crucial role in advancing high-quality research.

- **Allowing access to high-quality resources**: they provide researchers with access to state-of-the-art facilities, cutting-edge equipment, and specialized services. These resources are often expensive and require specific expertise to manage effectively.
- **Fostering Innovation**: By offering researchers the tools they need, RIs promote innovation promoting methods and tools and enable the development of groundbreaking technologies. This fosters scientific progress and drives solutions to global challenges.
- **Collaboration and Networking**: RIs facilitate collaboration among scientists, both nationally and internationally. Researchers from different backgrounds can work together, exchange ideas, and tackle complex problems more effectively.

RIs and FAIR data



- **FINDABLE**: Data and materials enriched with metadata assigned with a unique identifier.

- **ACCESSIBLE**: Data and metadata stored in a trusted repository with an open and free protocol accessible by machines and humans.

- **INTEROPERABLE**: Using vocabularies and public domain ontologies the metadata can be referenced and linked.

- **REUSABLE:** Additional documentation and protocols describing the acquisition of data, licensed with a detailed provenance.

The reutilization of research data

Data reutilization involves repurposing existing data for new investigations. It proves particularly valuable when addressing phenomena that require budget-intensive data collection or when developing studies using an **inductive approach.** It minimizes duplication of effort and promotes efficiency.

Cost-Efficiency: reusing existing data reduces the need for costly new experiments or data collection. Researchers can build upon prior work, saving time and resources.

Meta-Analysis: combining data from multiple studies enables meta-analyses, revealing broader trends and patterns. It enhances our understanding of complex phenomena**.**

Interdisciplinary Insights: data reutilization encourages collaboration across disciplines. Researchers can explore connections beyond their own field.

Accessing Research Infrastructures

Open access to resources (data, articles, standards, procedures, tools) and facilities is an essential condition.

Each RI, with its unique characteristics and combinations, essentially offers three different types of access:

- **Virtual Access**: Researchers can *access* data, tools, and digital products directly through communication networks.
- **Physical Access**: Researchers can visit the RI's laboratories and facilities in person, using on-site equipment, receiving training, and specialized support.
- **Remote Access**: Even when not physically present at the facility, researchers can utilize specific infrastructure services and equipment (e.g. high-performance computers) from a distance.

When competitive user selection is foreseen, experts evaluate requests based on:

- **Scientific Excellence**: Assessing the quality and socio-economic relevance of research at national and European levels (*excellence-driven*).
- **Technical Needs**: Ensuring high-quality analyses, precise measurements, and reliable data (*need-driven*).
- **Private Sector Demands**: Considering applications' relevance for potential innovation impacts and adapting access to specific user needs (*market-driven*).

Supplying training activities within Research Infrastructures

- Delivering a **set of training activities** is crucial for promoting the contents and services of a RI, with the goal of improving individual skills and nurturing the **development of communities**.
- In the context of RI, **training for researchers and stakeholders** should be designed to empower infrastructure users with the necessary knowledge, skills, and competencies, enabling them to effectively utilize scientific data and derive benefits from them.
- Knowledge on methods, techniques, and services for data collection, management, preservation, and analysis is essential for conducting cutting-edge research in social science, driving social innovation, and fortifying synergies between scholarly endeavors and societal stakeholders.
- A **non-scientific audience** (societal and political actors) has the potential to enhance the understanding of available data, consequently paving the way to introduce evidence-based knowledge into society.

Investigation on the use of RIs by social scientists in Italy / 1 (Spinello, 2019)

- The use of research infrastructures, either recognized by PNIR or project-driven, has had a **growing impact over time on the production of new scientific outputs** among researcher at early or mid-career level. Scholars with more experience reported a more limited impact.

- Nevertheless, social researchers are increasingly considering **pre-organized databases** during the design phase of their research, as an alternative to implementing original data collections, especially in times of inadequate funding.

- The infrastructure used has facilitated a more **inductive approach** to implementing studies compared to the past. Research questions that were not initially anticipated in the designs have been formulated post hoc through reasoning and reflections on existing data.

Investigation on the use of RIs by social scientists in Italy / 2 (Spinello, 2019)

- The availability of Ris in the field of social sciences, especially those of European interest, **does not appear to have significantly fostered the establishment of new aggregative dynamics**. In most instances, these infrastructures have served as vehicles to substantiate collaborative instances already induced *ex ante* through shared participation in projects, demonstrating their ability to strengthen existing networks by creating common outputs.

- However, researchers believe that the utilization of infrastructures has **enhanced their ability to conduct high-quality research**. Indeed, leveraging these infrastructures has led to significant advantages in representing phenomena due to a sample coverage that is rarely achievable through individual studies. Additionally, it has facilitated regional or international comparisons by ensuring methodological consistency in data collection.

- The study depicted a scientific community still grappling with the **initial stages of transitioning** toward a knowledge production model based on widely shared research infrastructures.

Challenges for the future of SSH RIs (Spinello et al., 2021)

- **Dissemination and awareness**. Effective dissemination ensures that individuals are aware of available data repositories and have the necessary skills to access and utilize them. Simply making data available doesn't guarantee its use. Researchers, policymakers, and practitioners need to understand the contents and possibilities of these resources.

- **Incentives for data sharing**. Some researchers are still concerned about someone "stealing" their data, rather than having a vision open to the concept of "sharing" from which it is possible to obtain mutual benefits in terms of knowledge growth. There is a need to include open science practices in research evaluation for career advancement.

- **Interoperation and searching tools.** The construction of common data structures for interoperability and reusable data flows are essential for implementing a system of Ris which is accessible and user-friendly. Common agreed rules on metadata can also facilitate the processes of searches.

References

- Boulton, G., Campbell, P., Collins, B., Elias, P., Hall, W., Laurie, G., O'Neill, O., Rawlins, M., Thornton, J., Vallance, P. & Walport, W. (2012). Science as an open enterprise. London: The Royal Society.
- Chawinga, W.D., & Zinn, S. (2019). Global perspectives of research data sharing: A systematic literature review. Library & Information Science Research, 41(2), pp. 109-122.
- Franssen, T. (2020). Research infrastructure funding as a tool for science governance in the humanities: A country case study of the Netherlands. In K. Cramer & O. Hallonsten (Eds.) Big Science and Research Infrastructures in Europe, Edward Elgar Publishing.
- Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., & Trow, M. (1994). The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies, London: SAGE.
- Hallonsten, O. & Cramer, K. (2020). Big science and research infrastructures in Europe: Conclusions and outlook. In K. Cramer & O. Hallonsten (Eds.), Big Science and Research Infrastructures in Europe, Edward Elgar Publishing.
- Hey, T., Tansley, S., & Tolle, K. (2009). Jim Gray on e-science: A transformed scientific method. In Hey, T., Tansley, S., & Tolle, K. (Eds.). The fourth paradigm. Data-intensive scientific discovery. Redmond, Washington: Microsoft Research, pp. xvii-xxxi.
- Nielsen, M. (2011), Reinventing discovery: The new era of networked science. New Jersey: Princeton University Press.
- Pasquetto, I.V., Sands, A.E., & Borgman C.L. (2015). Exploring openness in data and science: What is "open," to whom, when, and why? In Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community. Article No. 141.
- Renschler, B., Kleiner, I., Wernli, B., Farago, P., & Joye, D. (Eds.) (2013). Understanding Research Infrastructures in the Social Sciences, Zurich: Seismo Press.
- Schmidt, B., Gemeinholzer, B., & Treloar, A. (2016). Open data in global environmental research: The Belmont Forum's open data survey. PLoS ONE, 11 (1), Article e0146695.
- Spinello, A.O. (2019). Le infrastrutture di ricerca nelle scienze sociali: uno studio di caso su utilizzo ed effetti nella produzione scientifica. Welfare e Ergonomia, 1/2019, pp. 67-78.
- Spinello, A. O., Giglitto, D., Lockley, E. (2021). Management of open access research infrastructures in large EU projects: the "CultureLabs" case (CNR-IRCrES Working Paper 9/2021). Istituto di Ricerca sulla Crescita Economica Sostenibile.

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Consiglio Nazionale delle Ricerche

THANK YOU!

✉ fossr.dissemination@ircres.cnr.it

@fossrproject

fossr.eu

fossr-eu

@fossr

zenodo zenodo.org/communities/fossr

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Consiglio Nazionale delle Ricerche

Module I: Research Infastructures and Open Science

Part II: Example of a Novel Sociotechnical Platform - FOSSR

FOSSR

Fostering Open Science in Social Science Research

Innovative tools and services to investigate economic and societal change

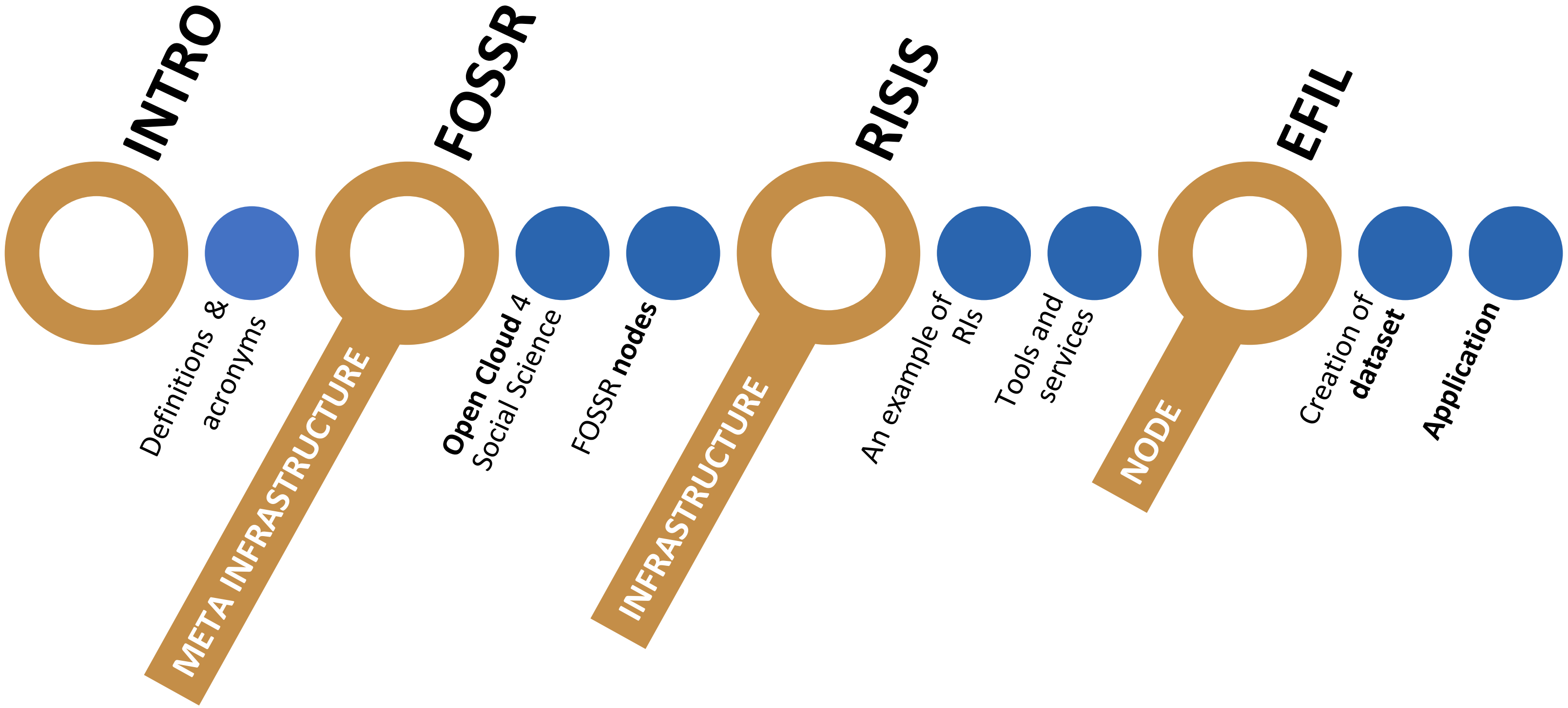
Emanuela Varinetti
Researcher at CNR-IRCrES

«Exploring research infrastructure data through advanced statistical applications»

FOSSR Online Training

26-27 September 2024

TABLE OF CONTENTS



INTRO

Definitions & acronyms

FOSSR

META INFRASTRUCTURE

Open Cloud 4 Social Science

FOSSR **nodes**

RISIS

INFRASTRUCTURE

An example of RIs

Tools and services

EFIL

NODE

Creation of **dataset**

Application

A definition from EU commission

Research infrastructures are facilities that provide resources and services for the research communities to conduct research and foster innovation in their fields.

These include:

- major equipment or sets of instruments
- knowledge-related facilities such as collections
- archives or scientific data infrastructures
- computing systems
- communication networks

https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/european-research-infrastructures_en

WHAT IS FOSSR

Fostering Open Science in Social Science Research

Research (meta)Infrastructures/sociotechnical platform with open resources of high quality FAIR data, information, services, indicators, key resources to improve the evidence-based decision-making capabilities in different policy domains.

Italian Open Science Cloud inspired by the European Open Science Cloud (EOSC).

Value of FOSSR is the creation of a network of the RIs in social sciences that :

a) develop new open services and resources, highly innovative not yet existing in Italy and essential for robust analysis and research investigations in social sciences

b) build an Open Cloud linked to a network of data centers mainly located in the South of Italy, with the aim to improve the computing facilities existing at local level.

Aims

The aim of FOSSR is to provide innovative tools and services to investigate issues related to the economic and societal change of contemporary societies:

- demographic analysis and the structure of economy (innovative firms and fast growing firms, innovation processes and outcome, new modes of knowledge production)
- society and social behaviors (ageing, wealth distribution, inequalities, education, migration, etc.)
- models for social simulation
- design, implementation, and assessment of public policies (e.g., R&I policies, health policies)

Finally, FOSSR aim is to connect three European infrastructures CESSDA, SHARE and RISIS, and then implement others, as well as panels and services currently under development and construction.

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Consiglio Nazionale delle Ricerche

CESSDA- ERIC

Consortium of European Social Science Data Archives

Member countries seek to increase the scientific excellence and efficacy of European research in the social sciences, as well as to expand easy access to data and metadata regardless of borders. They want to provide a research infrastructure for their researchers, and join forces among their (national) data service providers.

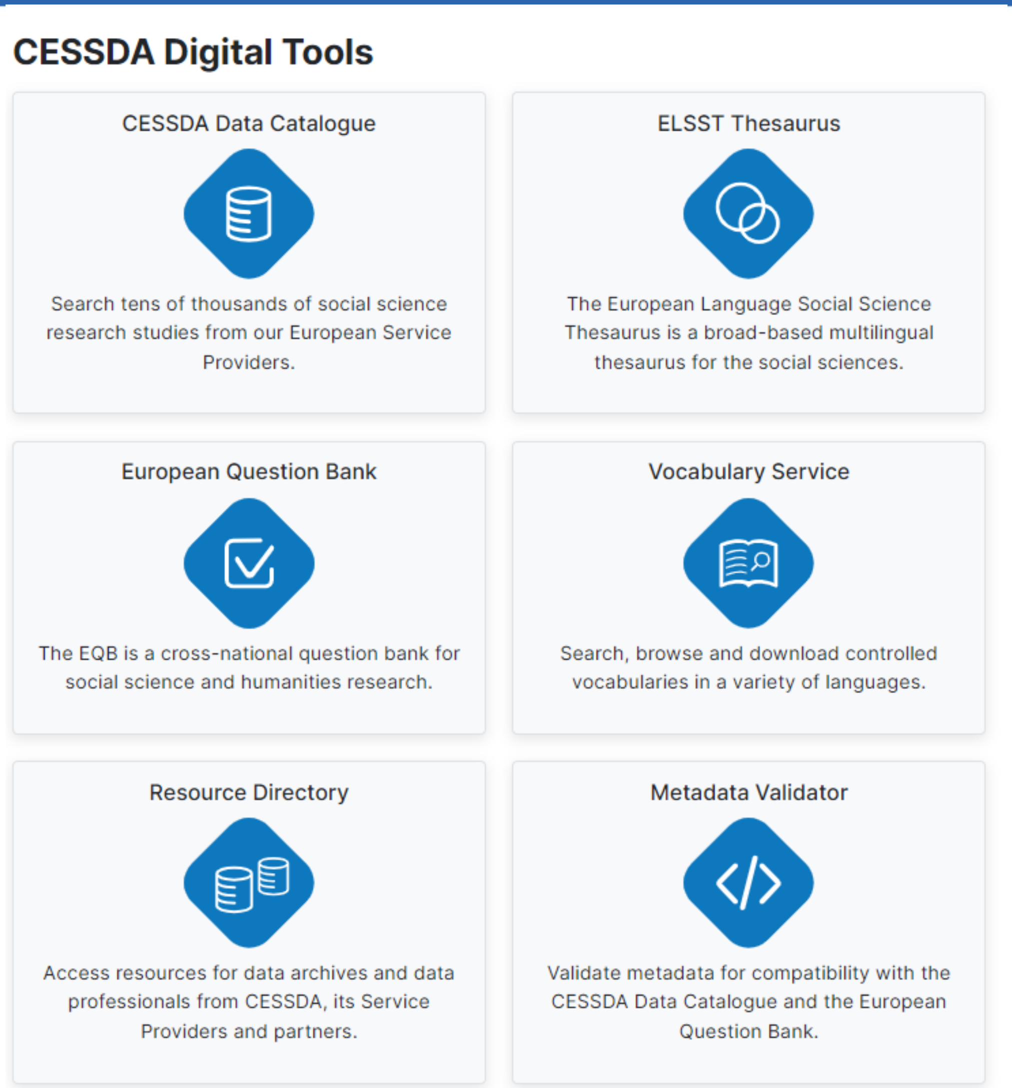
Sciences Italy - DASSI

DASSI (Data Archive Social Sciences Italy) is a Joint Research Unit composed of the Italian National Research Council (CNR) and the University of Milano Bicocca.

The **Interdepartmental Centre UniData** – Bicocca Data Archive is a joint project coming from eight departments of the University of Milano-Bicocca.

The project aims to create a centre for excellence in data sharing, enhance the secondary analysis and promote responsible data use in social, economic and environmental studies.

https://www.cessda.eu/



CESSDA Digital Tools

CESSDA Data Catalogue

Search tens of thousands of social science research studies from our European Service Providers.

ELSST Thesaurus

The European Language Social Science Thesaurus is a broad-based multilingual thesaurus for the social sciences.

European Question Bank

The EQB is a cross-national question bank for social science and humanities research.

Vocabulary Service

Search, browse and download controlled vocabularies in a variety of languages.

Resource Directory

Access resources for data archives and data professionals from CESSDA, its Service Providers and partners.

Metadata Validator

Validate metadata for compatibility with the CESSDA Data Catalogue and the European Question Bank.

SHARE- ERIC

Survey of Health, Ageing and Retirement in Europe

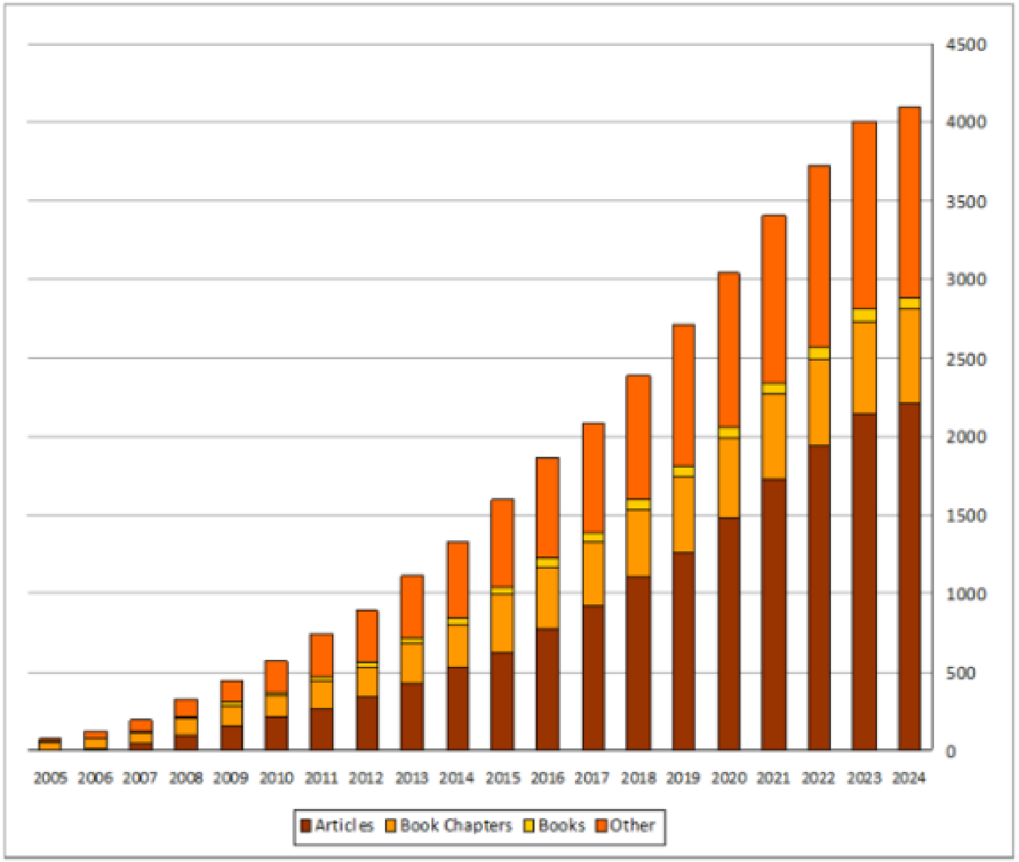
SHARE, the Survey of Health, Ageing and Retirement in Europe, is a research infrastructure for studying the effects of health, social, economic and environmental policies over the life-course of European citizens and beyond.

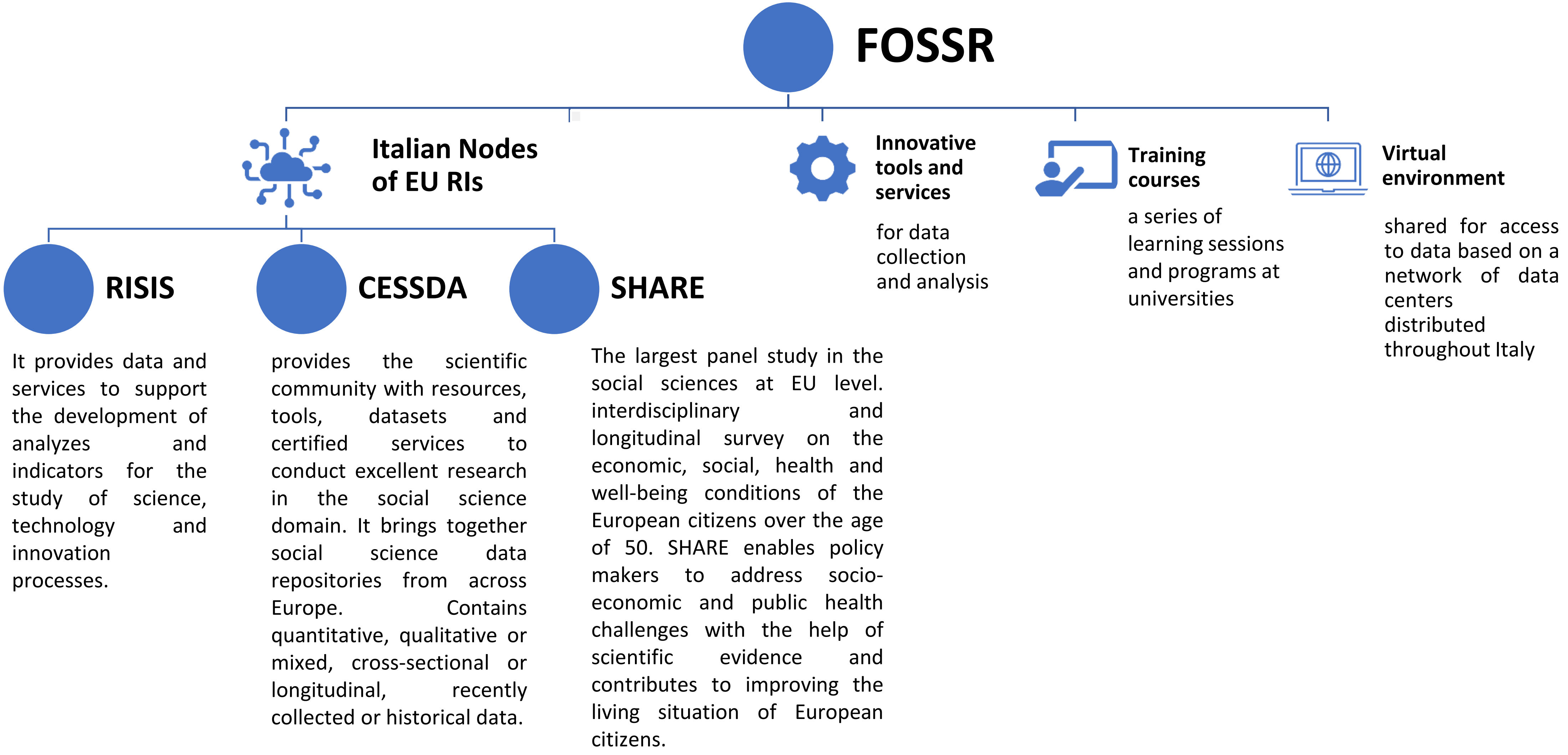
COUNTRY TEAM

- Prof. Guglielmo Weber, Ph.D. Nancy Zambon - Universita' degli Studi di Padova Dipartimento di Scienze Economiche
- Prof. Agar Brugiavini, Ph.D. - Università Ca' Foscari di Venezia Department of Economics

https://share-eric.eu/

Graphic 2: Publications by Year (Cumulative), June 2024





FOSSR

Italian Nodes of EU RIs

RISIS

It provides data and services to support the development of analyzes and indicators for the study of science, technology and innovation processes.

CESSDA

provides the scientific community with resources, tools, datasets and certified services to conduct excellent research in the social science domain. It brings together social science data repositories from across Europe. Contains quantitative, qualitative or mixed, cross-sectional or longitudinal, recently collected or historical data.

SHARE

The largest panel study in the social sciences at EU level. interdisciplinary and longitudinal survey on the economic, social, health and well-being conditions of the European citizens over the age of 50. SHARE enables policy makers to address socio-economic and public health challenges with the help of scientific evidence and contributes to improving the living situation of European citizens.

Innovative tools and services

for data collection and analysis

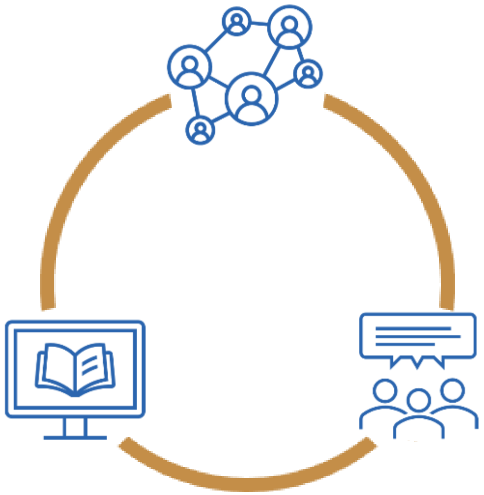
Training courses

a series of learning sessions and programs at universities

Virtual environment

shared for access to data based on a network of data centers distributed throughout Italy

Participatory sessions, seminars and webinars



Co-working spaces where the various actors collaborate and share knowledge and new ideas.

The Project offers:

Learning sessions and **action research sessions** are participatory training events to share reflections about the use of RIs based on the point of views of researcher and different stakeholders

Policymakers Sessions are online events that aim to inform policy makers of the potential uses of FOSSR Open Cloud data and services for policy design and evaluation.

The **Online Seminars** aim to promote the exchange of research ideas and activate a collective process of knowledge co-creation between researchers, students and other stakeholders within and outside the FOSSR community.

The **webinars** are intended to give visibility to the project and discuss the specific tools and services developed within FOSSR to investigate economic and social change.

Post-graduate university programmes

A key objective of FOSSR is to train a **new generation of young social science researchers** capable of exploiting the opportunities related to Open Science. To this end a series of agreements with some prestigious Italian universities are signed.

A Second Level **Master programme in Social Data Science**, has been activated at the University of Milan-Bicocca.

Positions in **PhD courses** related to the social sciences sector, on topics related to the contents of the FOSSR Project

Innovative tools and services

FOSSR incorporate tools (hardware and software) and methods useful to research practices and traceable to the paradigms of eScience, behavioural economics, and computational social sciences.

In operational terms, the functionalities are:

- Data collection and data integration
- Data analysis
- Data curation and data sharing

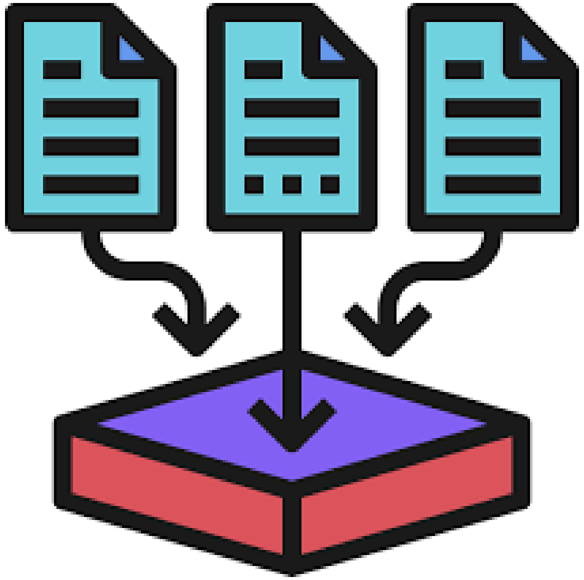
Data collection and data integration

Collection of complementary **data** (in terms of spatial and temporal resolution, typology, etc.) with **innovative methods** compared to those currently used.

Among methods: web scraping techniques and social media analysis.

Moreover, **new tools,** will be developed:

- ORP - Online Research Panel
- IOPP - Italian Online Probability Panel



ORP

Online Research Panel

Multidisciplinary approach that allows:

- To obtain a more complete and in-depth view of social and economic dynamics
- to understand processes in the long-term context

providing a solid basis for policies development and implementation of targeted interventions.

ORP is composed of selected Italian citizens, interviewed regularly through online surveys in order to collect FAIR data on:

- behaviour,
- preferences and attitudes of the population
- with respect to the most important issues on the Italian and international scene (politics, health, economy, society and environment)

IOPP

The Italian Online Probability Panel

The Italian Online Probability Panel (IOPP) will provide a multi-purpose tool enabling the execution of surveys on the Italian population, characterized by the highest scientific standards in social sciences fostering the creation in Italy of a strong infrastructural research hub for Italian panel studies in the social sciences.

Items:

- Family structure
- housing condition
- employment and economic situation
- inequalities
- social vulnerability
- gender equality

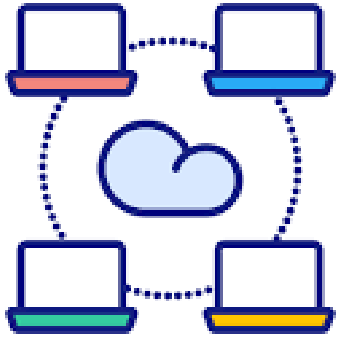
Data analysis

Analysis of data collected **through** the two main declinations (data-driven and model driven) of **computational social science** shall provide important new resources for science and innovation studies (RISIS) and to exploit data provided by CESSDA and the survey results of SHARE.

More in detail **methods** that will be open to the users are:

- data mining (machine learning, natural language processing)
- agent-based social simulation models
- social network analysis and spatial analysis
- complexity modeling
- semantic knowledge graph (SKG)

Data curation and data sharing

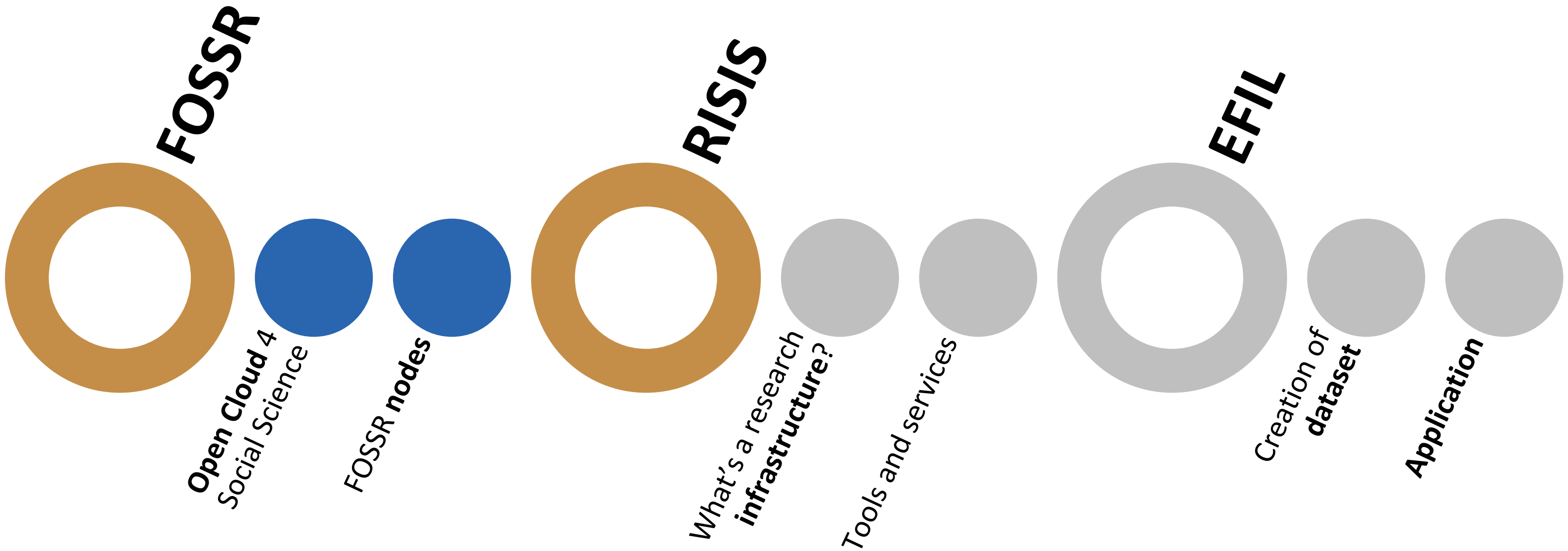


Data and analysis results are available for user through an open platform:

- in line with the model of open science and the enhancement of public information assets
- accessible to national and international users, thus improving the attraction of the RIs involved in the thematic network
- providing a unique integrated access to different data, tools, and services for research in several domains of social sciences.

FOSSR website:

https://www.fossr.eu/



FOSSR

Open Cloud 4 Social Science

FOSSR **nodes**

RISIS

What's a research **infrastructure?**

Tools and services

EFIL

Creation of **dataset**

Application

What is RISIS - *Research Infrastructure for Science and Innovation policy Studies*

- European research infrastructure focusing on services and data for science and innovation studies
- 18 partner - including universities and the most innovative research centres in Europe https://www.risis2.eu/partners/
- Initiated in 2014 and renewed in 2018 until 2023
- Transformation in AISBL (*Association internationale sans but lucratif*)

It provides European researchers:

Freely **online** accessible **services**

Project-based **access to** curated & enriched **datasets**

Tools for methodological advances

Registers on research organizations and firms

Ontologies, visualization maps

Training, research and awareness raising **events**

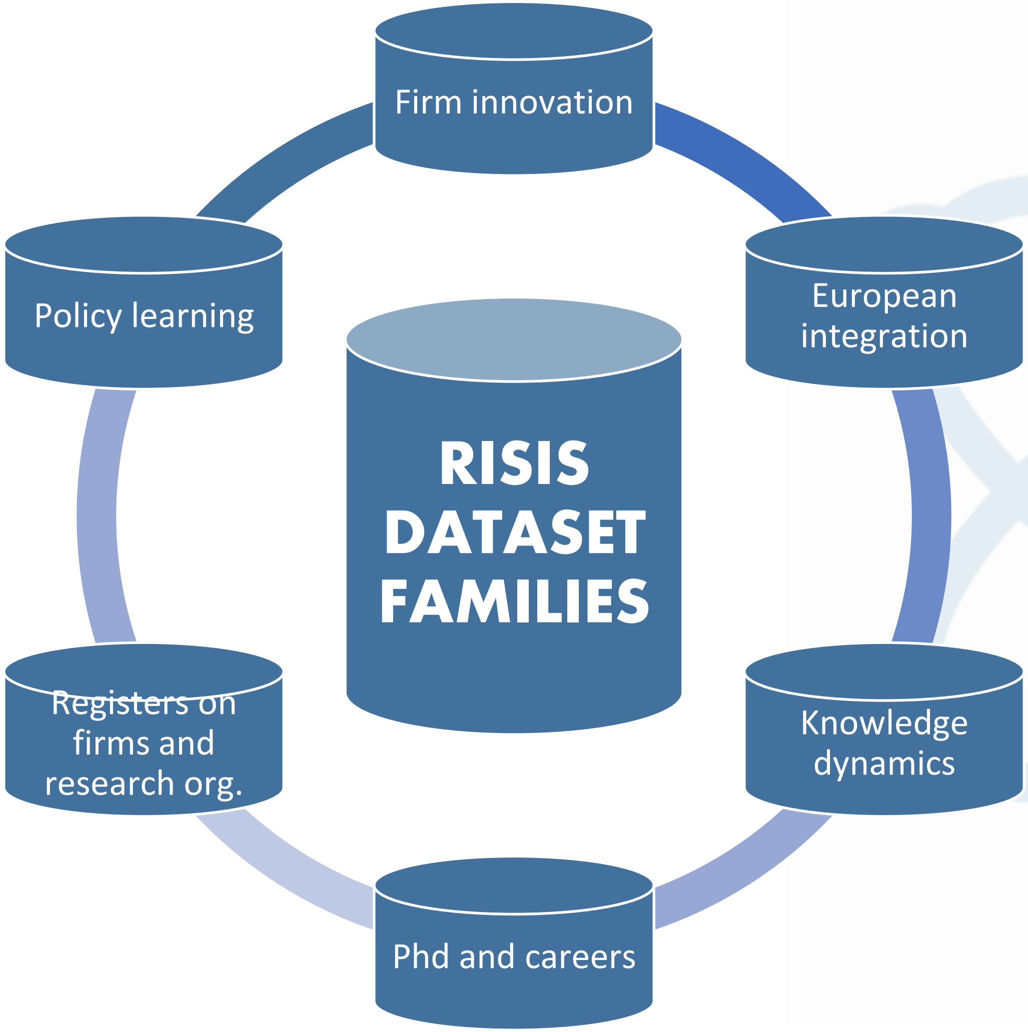
RISIS structure

Access to **15 RISIS datasets** for studying science and innovation

A unique data and services infrastructure

Supporting the development of a new generation of analyses and indicators

https://www.risis2.eu/risis-datasets/



RISIS DATASET FAMILIES

Firm innovation

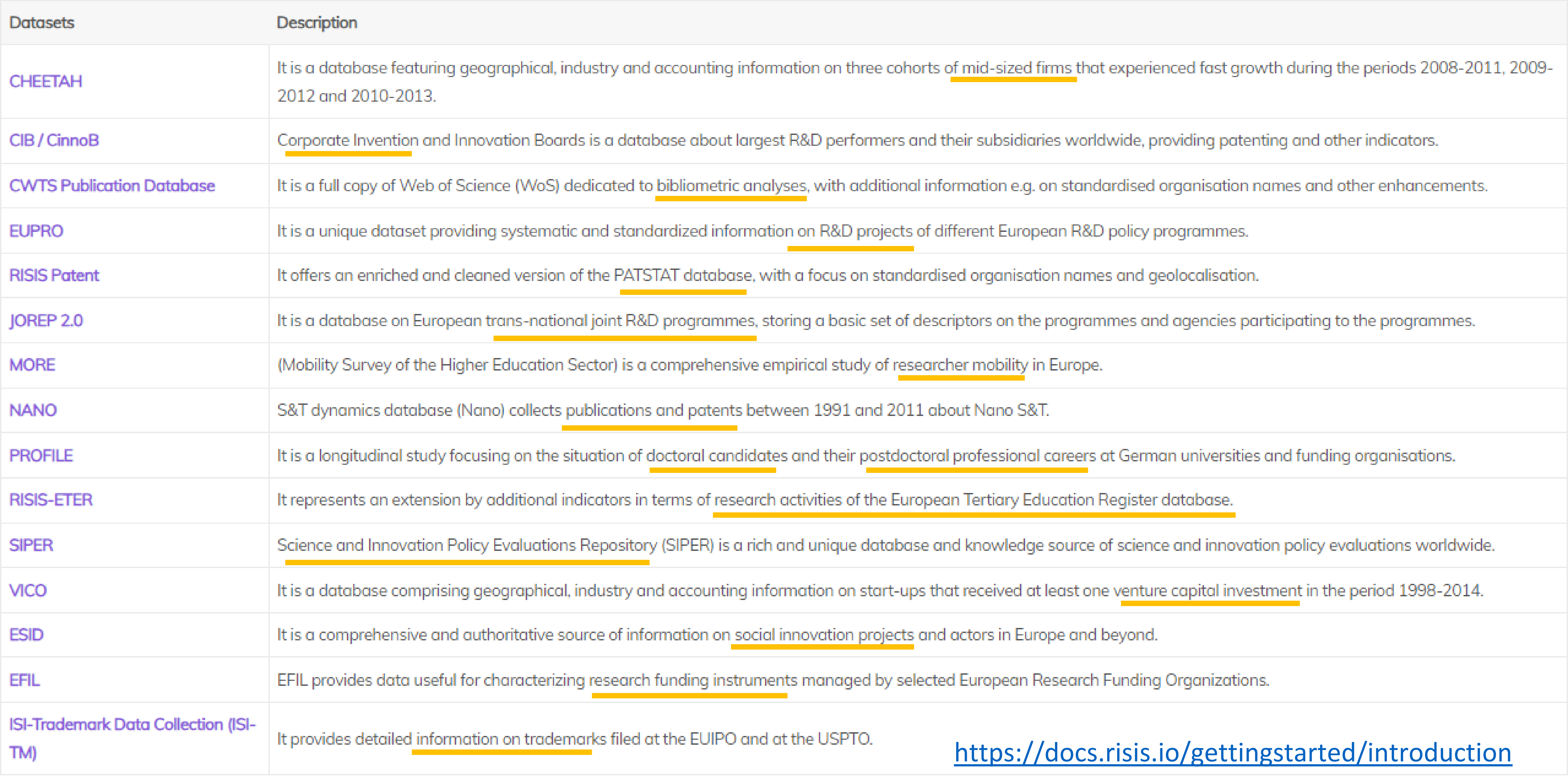
European integration

Knowledge dynamics

Phd and careers

Registers on firms and research org.

Policy learning



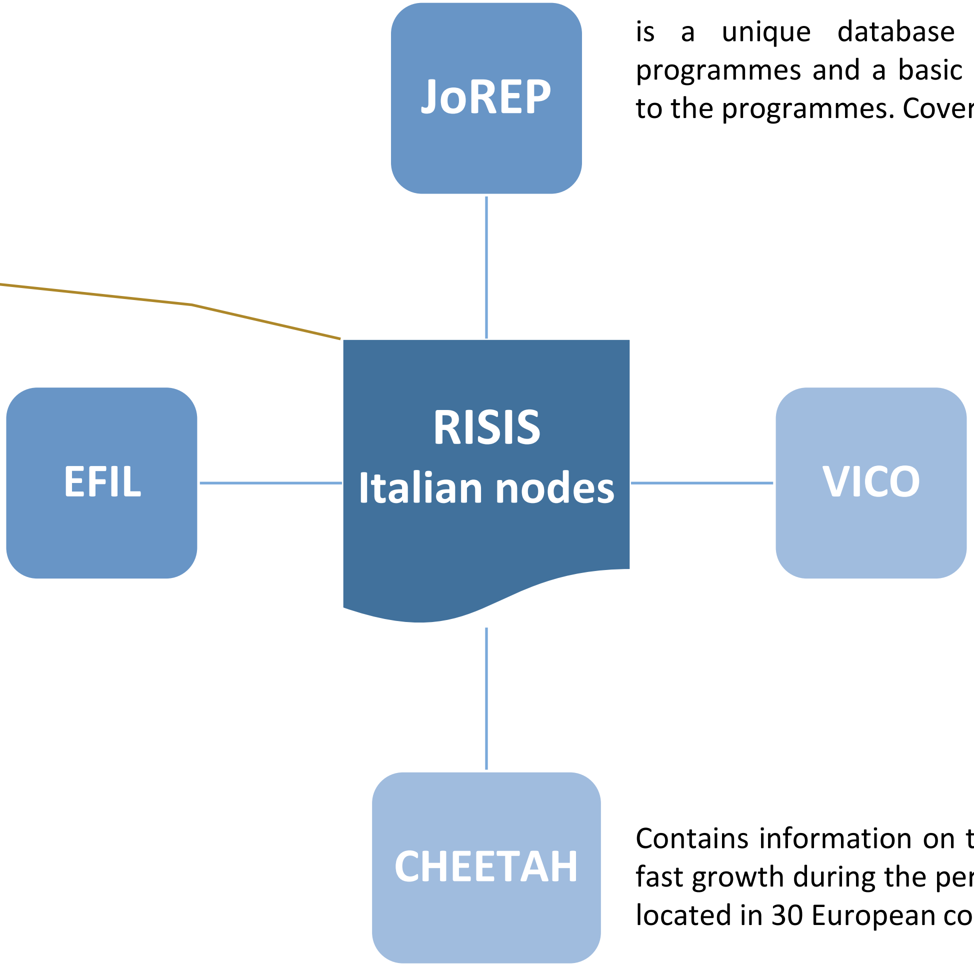
Datasets	Description
CHEETAH	It is a database featuring geographical, industry and accounting information on three cohorts of mid-sized firms that experienced fast growth during the periods 2008-2011, 2009-2012 and 2010-2013.
CIB / CinnoB	Corporate Invention and Innovation Boards is a database about largest R&D performers and their subsidiaries worldwide, providing patenting and other indicators.
CWTS Publication Database	It is a full copy of Web of Science (WoS) dedicated to bibliometric analyses, with additional information e.g. on standardised organisation names and other enhancements.
EUPRO	It is a unique dataset providing systematic and standardized information on R&D projects of different European R&D policy programmes.
RISIS Patent	It offers an enriched and cleaned version of the PATSTAT database, with a focus on standardised organisation names and geolocalisation.
JOREP 2.0	It is a database on European trans-national joint R&D programmes, storing a basic set of descriptors on the programmes and agencies participating to the programmes.
MORE	(Mobility Survey of the Higher Education Sector) is a comprehensive empirical study of researcher mobility in Europe.
NANO	S&T dynamics database (Nano) collects publications and patents between 1991 and 2011 about Nano S&T.
PROFILE	It is a longitudinal study focusing on the situation of doctoral candidates and their postdoctoral professional careers at German universities and funding organisations.
RISIS-ETER	It represents an extension by additional indicators in terms of research activities of the European Tertiary Education Register database.
SIPER	Science and Innovation Policy Evaluations Repository (SIPER) is a rich and unique database and knowledge source of science and innovation policy evaluations worldwide.
VICO	It is a database comprising geographical, industry and accounting information on start-ups that received at least one venture capital investment in the period 1998-2014.
ESID	It is a comprehensive and authoritative source of information on social innovation projects and actors in Europe and beyond.
EFIL	EFIL provides data useful for characterizing research funding instruments managed by selected European Research Funding Organizations.
ISI-Trademark Data Collection (ISI-TM)	It provides detailed information on trademarks filed at the EUIPO and at the USPTO.

https://docs.risis.io/gettingstarted/introduction

RISIS** is identified among the high and medium priority infrastructures, within the **PNIR 2021 2027

***RISIS** with **CESSDA, SHARE** are RIs in social sciences coordinated by CNR.*

*The integration of this pool of resources **will contribute to the realization of Open Science Cloud for scholars in social sciences** proposed by **FOSSR***



RISIS
Italian nodes

JoREP

is a unique database on European trans-national joint R&D programmes and a basic set of descriptors of agencies participating to the programmes. Covers data for the period 2000-2014 (IRCrES)

EFIL

VICO

contains information on start-ups that received at least one venture capital investment during 1998-2014 (POLIMI)

CHEETAH

Contains information on three cohorts of mid-sized firms that experienced fast growth during the period 2008/2013. It includes 42,369 mid-sized firms located in 30 European countries and Israel (POLIMI)

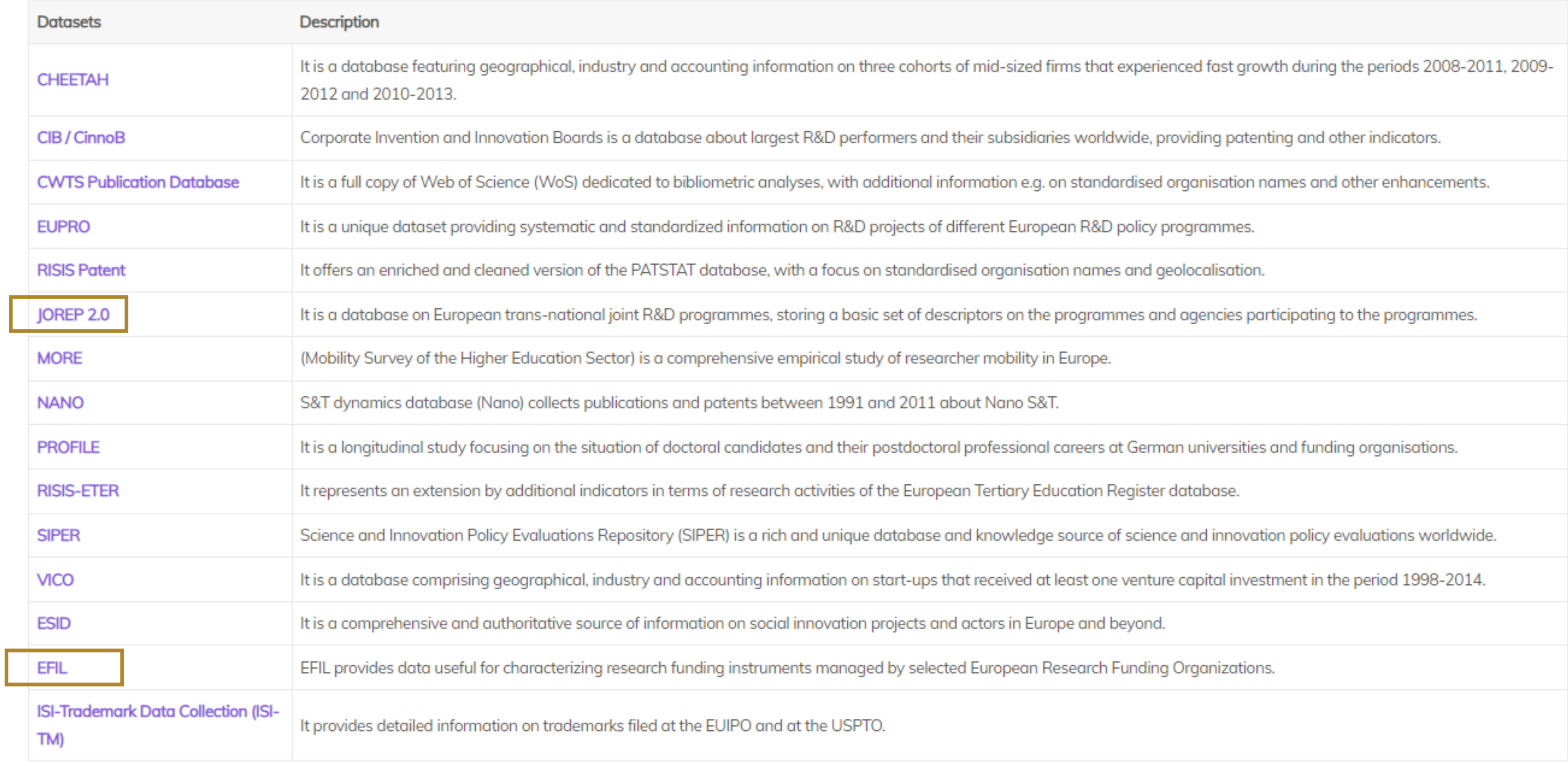
CNR IRCrES Istituto di Ricerca sulla Crescita Economica Sostenibile

IRCrES CNR is reference institute for the FOSSR **RISIS** Italian node, in which Politecnico di Milano also participates with VICO and CHEETAH.

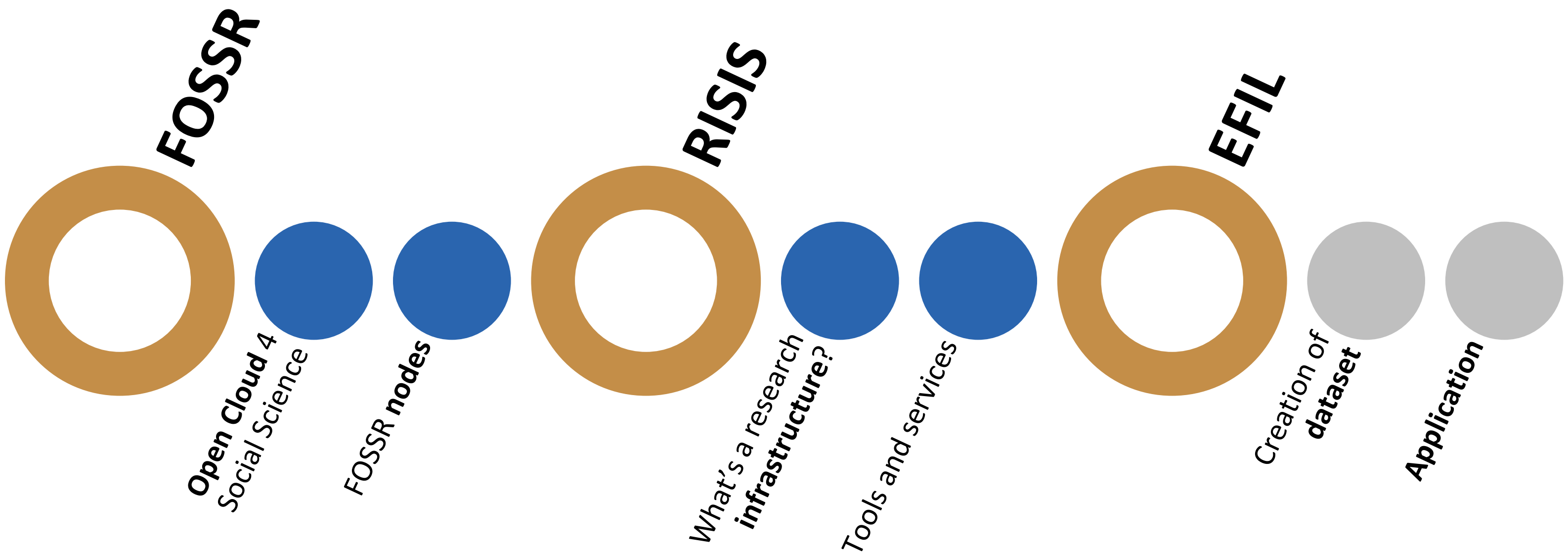
IRCrES participated in the development of the design of **RISIS** in the context of two contracts **financed by** the European Union with the 7th Framework Program of **Horizon 2020.**

Dataset manager and scientific coordinator of two Datasets:

JOREP and **EFIL**



Datasets	Description
CHEETAH	It is a database featuring geographical, industry and accounting information on three cohorts of mid-sized firms that experienced fast growth during the periods 2008-2011, 2009-2012 and 2010-2013.
CIB / CinnoB	Corporate Invention and Innovation Boards is a database about largest R&D performers and their subsidiaries worldwide, providing patenting and other indicators.
CWTS Publication Database	It is a full copy of Web of Science (WoS) dedicated to bibliometric analyses, with additional information e.g. on standardised organisation names and other enhancements.
EUPRO	It is a unique dataset providing systematic and standardized information on R&D projects of different European R&D policy programmes.
RISIS Patent	It offers an enriched and cleaned version of the PATSTAT database, with a focus on standardised organisation names and geolocalisation.
JOREP 2.0	It is a database on European trans-national joint R&D programmes, storing a basic set of descriptors on the programmes and agencies participating to the programmes.
MORE	(Mobility Survey of the Higher Education Sector) is a comprehensive empirical study of researcher mobility in Europe.
NANO	S&T dynamics database (Nano) collects publications and patents between 1991 and 2011 about Nano S&T.
PROFILE	It is a longitudinal study focusing on the situation of doctoral candidates and their postdoctoral professional careers at German universities and funding organisations.
RISIS-ETER	It represents an extension by additional indicators in terms of research activities of the European Tertiary Education Register database.
SIPER	Science and Innovation Policy Evaluations Repository (SIPER) is a rich and unique database and knowledge source of science and innovation policy evaluations worldwide.
VICO	It is a database comprising geographical, industry and accounting information on start-ups that received at least one venture capital investment in the period 1998-2014.
ESID	It is a comprehensive and authoritative source of information on social innovation projects and actors in Europe and beyond.
EFIL	EFIL provides data useful for characterizing research funding instruments managed by selected European Research Funding Organizations.
ISI-Trademark Data Collection (ISI-TM)	It provides detailed information on trademarks filed at the EUIPO and at the USPTO.



FOSSR

Open Cloud 4 Social Science

FOSSR **nodes**

RISIS

What's a research **infrastructure?**

Tools and services

EFIL

Creation of **dataset**

Application

What is EFIL?

- EFIL – **European dataset of public R&D funding instruments** aims at enabling users to investigate public R&D funding in Europe at the level of project funding instruments and Research Funding Organizations (RFO), addressing questions related to policy design and policy implementation.
- **Main objectives** of EFIL are: – re-composing and characterizing the portfolios of funding instruments of relevant RFOs from selected European countries (**Austria, Switzerland, Czech Republic, Denmark, Estonia, Germany, Italy, Norway, United Kingdom, France**); – producing evidence of the structural, procedural, and allocational aspects of funding instruments, as well as organizational profiles.
- The **temporal coverage : 2021 backwards to 2010**.
- EFIL is complemented by a **repository of official documents** hosted on a cloud – composed of instrument calls, guidelines for participants, descriptions on official webpages, etc. – that is accessible to the database user (useful to characterize the instruments through text analysis, key words, and vocabularies).

PORTFOLIO RECOMPOSITION PROCESS

RFO Portfolio

RFO portfolio from RFO official sources

Funding Route
Meso level of representation

UNITS

Large funding instruments; Aggregations of scheme

Funding Instrument
Micro-level of representation

UNITS

Funding instruments; Aggregations disentangled

Key definitions

RFOs are entities that distribute – through funding instruments – public project funds for research and development

“R&D funding instruments” are funding schemes for R&D, within the total public R&D allocations, having proper characteristics in terms how they are managed, the beneficiaries, and how they are allocated

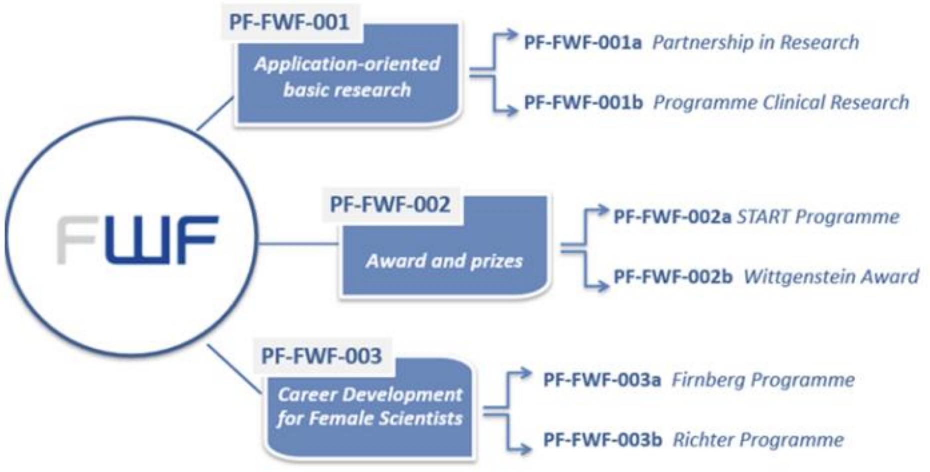
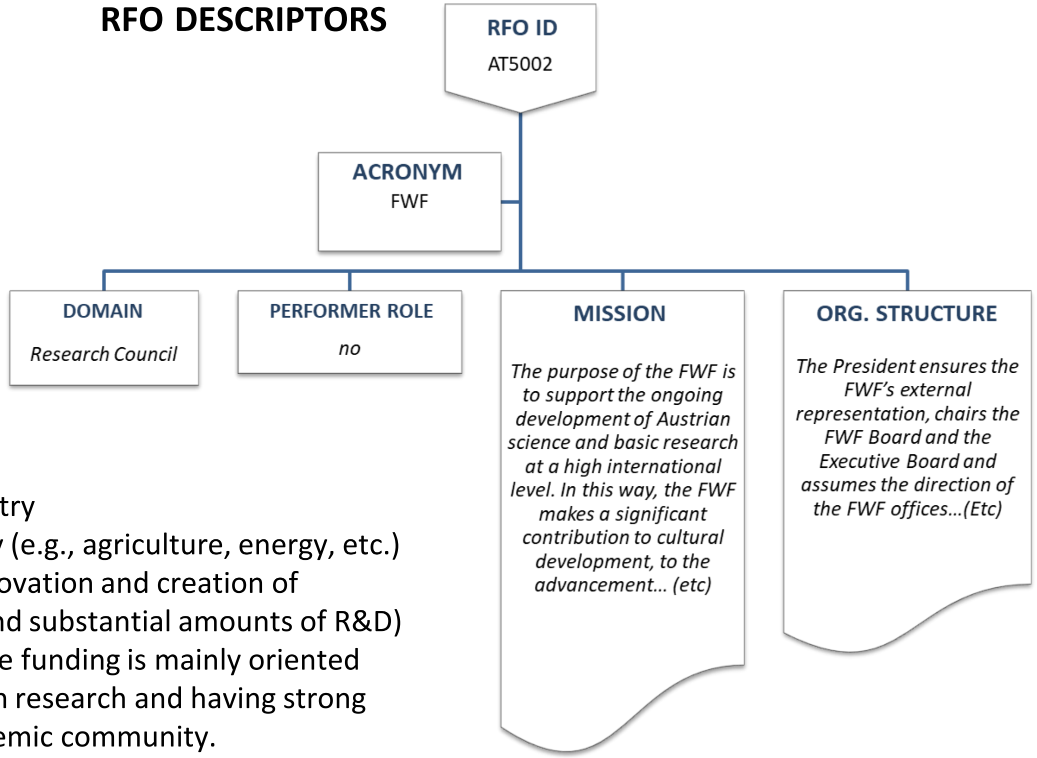

FWF
PF-FWF-001
Application-oriented basic research
PF-FWF-001a Partnership in Research
PF-FWF-001b Programme Clinical Research
PF-FWF-002
Award and prizes
PF-FWF-002a START Programme
PF-FWF-002b Wittgenstein Award
PF-FWF-003
Career Development for Female Scientists
PF-FWF-003a Firnberg Programme
PF-FWF-003b Richter Programme

Figure 5. Example of treatment of IDs of route/instruments in EFIL.

Descriptors at a glance

RFO DESCRIPTORS



RFO ID
AT5002

ACRONYM
FWF

DOMAIN
Research Council

PERFORMER ROLE
no

MISSION
The purpose of the FWF is to support the ongoing development of Austrian science and basic research at a high international level. In this way, the FWF makes a significant contribution to cultural development, to the advancement… (etc)

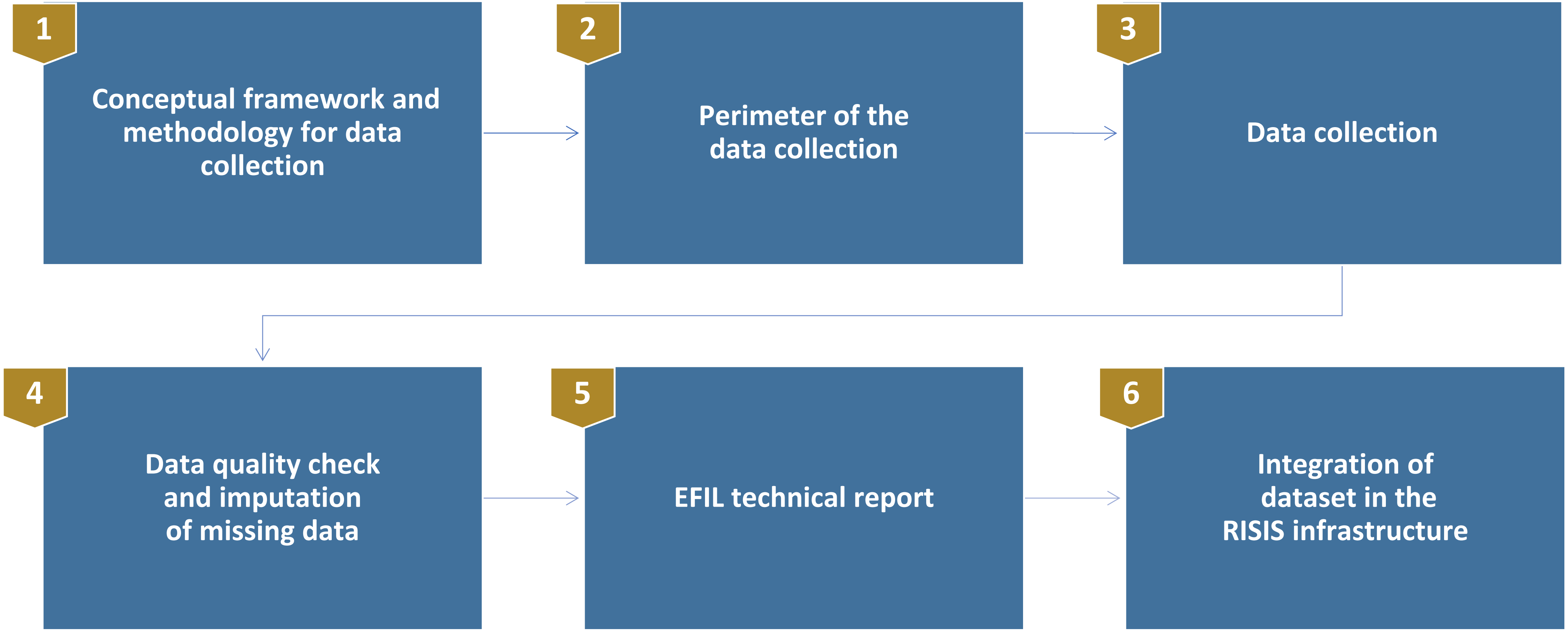
ORG. STRUCTURE
The President ensures the FWF's external representation, chairs the FWF Board and the Executive Board and assumes the direction of the FWF offices…(Etc)

- National research ministry
- National sector ministry (e.g., agriculture, energy, etc.)
- Innovation agency, (innovation and creation of economic value, but fund substantial amounts of R&D)
- Research Council, whose funding is mainly oriented towards curiosity-driven research and having strong connection to the academic community.
- Sectoral RFO – related to specific topic (energy, environment, etc.), e.g., sectoral regulatory agencies or sectoral funding agencies.

ROUTES/INSTRUMENTS DESCRIPTORS (main ones)

- Route/Instrument name (ENG + lang);
- Route/instruments goal;
- Route aim;
- Instrument KETs, SGCs, SDGs;
- Type of transfer;
- Academic-private cooperation;
- Composition of the decision-making body;
- Assessment criteria for funding allocation;
- Assessment methods for funding allocation;
- Level of openness;
- Eligible sectors;
- Type of funding

Dataset construction stages



1 Conceptual framework and methodology for data collection

2 Perimeter of the data collection

3 Data collection

4 Data quality check and imputation of missing data

5 EFIL technical report

6 Integration of dataset in the RISIS infrastructure

2 PERIMETER

The **EFIL perimeter** is based on the RFOs operating in European countries including Central and Eastern countries.

The **first phase focused** on RFOs with ‘important’ funding capabilities, already integrated in ORGREG.

The perimeter was enlarged in a second phase when the whole structure has been defined and tested.

As already mentioned, EFIL stores data on **55 relevant RFOs** from ten countries. EFIL provides information on **386 funding routes** and **700 funding instruments** for these RFOs.
Table 1 details the "EFIL perimeter" by country: number of RFOs selected as relevant, as well as the number of funding routes and single funding instruments for each of the countries.

Country	Number of RFOs	Funding routes	Funding instrument
Austria	4	44	100
Czech Republic	9	37	45
France	3	11	57
Denmark	6	87	88
Estonia	4	10	20
Germany	6	35	78
Italy	3	26	29
Norway	2	16	99
Switzerland	4	20	57
United Kingdom	14	100	127
TOTAL	**55**	**386**	**700**

Table 1. Number of observation for the composition of the EFIL perimeter.

new entry coming soon: **US- National Science Fundation**

3 DATA COLLECTION

Pilot data collection. The **feasibility** of the conceptual framework was explored, and the most problematic items were figured out. F**easibility** study to implement **text analysis** (through RISIS services) to extract from the calls policy implementation and mission orientation of R&D funding.

EFIL data collection is based on a **non-automated procedure developed through a web-exploration** of publicly available information dispersed into multiple resources. This collides with the **dispersion of the contents into multiple locations and sources** or, in the worst cases, with the **elimination of the contents** from the websites.

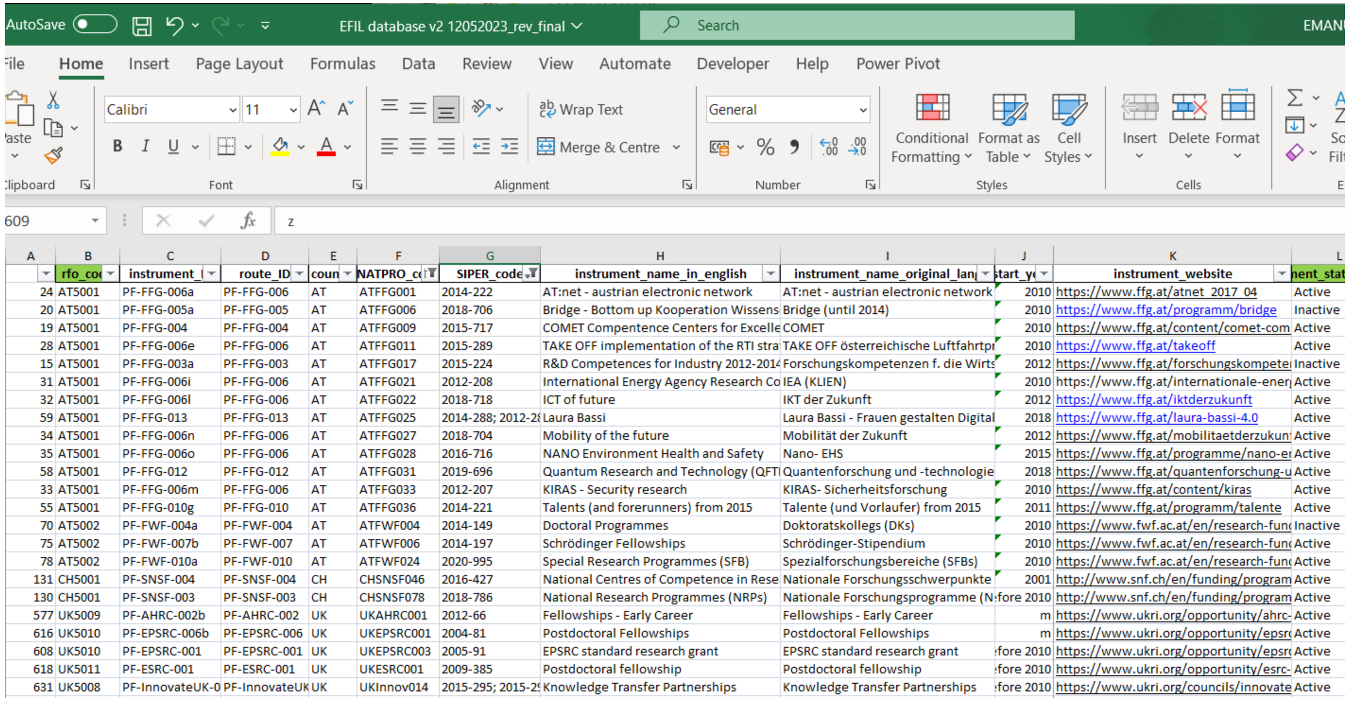
Two waves of collection have been performed:

• the **first** data collection (completed for the release 1.0 on March 2022) regarded data on instruments active in 2017-2018 and was developed to acquire budgetary data backwards to 2010;

• the **second** data collection (completed for the release 2.0 on May 2023) regarded data on new instruments active in 2019-2020-2021 and the update of data of the instruments collected during the first wave. In this collection are also included 15 instruments announced in 2021 but that will become active from 2022

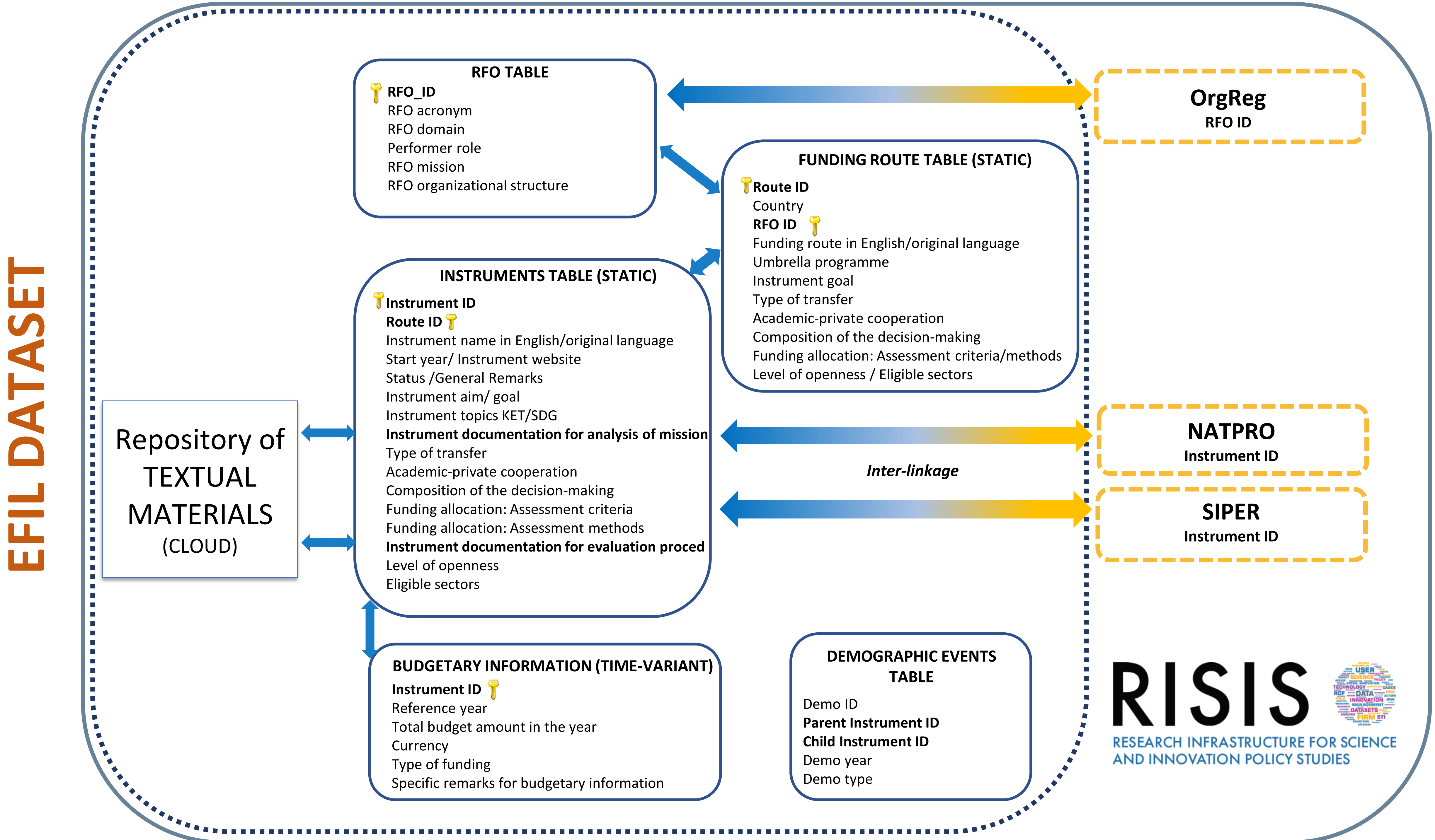
6 Integration of dataset in/with RISIS infrastructure

ORGREG through the RFOs managing the funding schemes (standardization of RFOs and classification)
NATPRO because of the complementarities between the two facilities (linking projects and funding instruments)
Other integrations **SIPER** repository –reference to funding scheme in the evaluation



	rfo_co	instrument_	route_ID	coun	NATPRO_c	SIPER_code	instrument_name_in_english	instrument_name_original_lan	start_y	instrument_website	nent_stat
24	AT5001	PF-FFG-006a	PF-FFG-006	AT	ATFFG001	2014-222	AT:net - austrian electronic network	AT:net - austrian electronic network	2010	https://www.ffg.at/atnet_2017_04	Active
20	AT5001	PF-FFG-005a	PF-FFG-005	AT	ATFFG006	2018-706	Bridge - Bottom up Kooperation Wissens	Bridge (until 2014)	2010	https://www.ffg.at/programm/bridge	Inactive
19	AT5001	PF-FFG-004	PF-FFG-004	AT	ATFFG009	2015-717	COMET Compentence Centers for Excelle	COMET	2010	https://www.ffg.at/content/comet-com	Active
28	AT5001	PF-FFG-006e	PF-FFG-006	AT	ATFFG011	2015-289	TAKE OFF implementation of the RTI stra	TAKE OFF österreichische Luftfahrtpr	2010	https://www.ffg.at/takeoff	Active
15	AT5001	PF-FFG-003a	PF-FFG-003	AT	ATFFG017	2015-224	R&D Competences for Industry 2012-2014	Forschungskompetenzen f. die Wirts	2012	https://www.ffg.at/forschungskompeten	Inactive
31	AT5001	PF-FFG-006i	PF-FFG-006	AT	ATFFG021	2012-208	International Energy Agency Research Co	IEA (KLIEN)	2010	https://www.ffg.at/internationale-ener	Active
32	AT5001	PF-FFG-006l	PF-FFG-006	AT	ATFFG022	2018-718	ICT of future	IKT der Zukunft	2012	https://www.ffg.at/iktderzukunft	Active
59	AT5001	PF-FFG-013	PF-FFG-013	AT	ATFFG025	2014-288; 2012-28	Laura Bassi	Laura Bassi - Frauen gestalten Digital	2018	https://www.ffg.at/laura-bassi-4.0	Active
34	AT5001	PF-FFG-006n	PF-FFG-006	AT	ATFFG027	2018-704	Mobility of the future	Mobilität der Zukunft	2012	https://www.ffg.at/mobilitaetderzukunf	Active
35	AT5001	PF-FFG-006o	PF-FFG-006	AT	ATFFG028	2016-716	NANO Environment Health and Safety	Nano- EHS	2015	https://www.ffg.at/programme/nano-en	Active
58	AT5001	PF-FFG-012	PF-FFG-012	AT	ATFFG031	2019-696	Quantum Research and Technology (QFTI	Quantenforschung und -technologie	2018	https://www.ffg.at/quantenforschung-u	Active
33	AT5001	PF-FFG-006m	PF-FFG-006	AT	ATFFG033	2012-207	KIRAS - Security research	KIRAS- Sicherheitsforschung	2010	https://www.ffg.at/content/kiras	Active
55	AT5001	PF-FFG-010g	PF-FFG-010	AT	ATFFG036	2014-221	Talents (and forerunners) from 2015	Talente (und Vorlaufer) from 2015	2011	https://www.ffg.at/programm/talente	Active
70	AT5002	PF-FWF-004a	PF-FWF-004	AT	ATFWF004	2014-149	Doctoral Programmes	Doktoratskollegs (DKs)	2010	https://www.fwf.ac.at/en/research-fund	Inactive
75	AT5002	PF-FWF-007b	PF-FWF-007	AT	ATFWF006	2014-197	Schrödinger Fellowships	Schrödinger-Stipendium	2010	https://www.fwf.ac.at/en/research-fund	Active
78	AT5002	PF-FWF-010a	PF-FWF-010	AT	ATFWF024	2020-995	Special Research Programmes (SFB)	Spezialforschungsbereiche (SFBs)	2010	https://www.fwf.ac.at/en/research-fund	Active
131	CH5001	PF-SNSF-004	PF-SNSF-004	CH	CHSNSF046	2016-427	National Centres of Competence in Rese	Nationale Forschungsschwerpunkte	2001	http://www.snf.ch/en/funding/program	Active
130	CH5001	PF-SNSF-003	PF-SNSF-003	CH	CHSNSF078	2018-786	National Research Programmes (NRPs)	Nationale Forschungsprogramme (N	fore 2010	http://www.snf.ch/en/funding/program	Active
577	UK5009	PF-AHRC-002b	PF-AHRC-002	UK	UKAHRC001	2012-66	Fellowships - Early Career	Fellowships - Early Career	m	https://www.ukri.org/opportunity/ahrc-	Active
616	UK5010	PF-EPSRC-006b	PF-EPSRC-006	UK	UKEPSRC001	2004-81	Postdoctoral Fellowships	Postdoctoral Fellowships	m	https://www.ukri.org/opportunity/epsrc	Active
608	UK5010	PF-EPSRC-001	PF-EPSRC-001	UK	UKEPSRC003	2005-91	EPSRC standard research grant	EPSRC standard research grant	fore 2010	https://www.ukri.org/opportunity/epsrc	Active
618	UK5011	PF-ESRC-001	PF-ESRC-001	UK	UKESRC001	2009-385	Postdoctoral fellowship	Postdoctoral fellowship	fore 2010	https://www.ukri.org/opportunity/esrc-	Active
631	UK5008	PF-InnovateUK-0	PF-InnovateUK	UK	UKInnov014	2015-295; 2015-2	Knowledge Transfer Partnerships	Knowledge Transfer Partnerships	fore 2010	https://www.ukri.org/councils/innovate	Active

- The combined use of EFIL and NATPRO datasets may help in shedding light on the characteristics of national R&D funding systems, as well as national R&D projects and participations.
- The link between programs and projects allows for interpreting on how government goals for science policy are translated into concrete research activities by project beneficiaries.
- The joint information from the two datasets may reveal the mismatch between the policy orientation and research practices of scholars' communities.



EFIL DATASET

RFO TABLE

RFO_ID
RFO acronym
RFO domain
Performer role
RFO mission
RFO organizational structure

FUNDING ROUTE TABLE (STATIC)

Route ID
Country
RFO ID
Funding route in English/original language
Umbrella programme
Instrument goal
Type of transfer
Academic-private cooperation
Composition of the decision-making
Funding allocation: Assessment criteria/methods
Level of openness / Eligible sectors

INSTRUMENTS TABLE (STATIC)

Instrument ID
Route ID
Instrument name in English/original language
Start year/ Instrument website
Status /General Remarks
Instrument aim/ goal
Instrument topics KET/SDG
Instrument documentation for analysis of mission
Type of transfer
Academic-private cooperation
Composition of the decision-making
Funding allocation: Assessment criteria
Funding allocation: Assessment methods
Instrument documentation for evaluation proced
Level of openness
Eligible sectors

Repository of TEXTUAL MATERIALS (CLOUD)

BUDGETARY INFORMATION (TIME-VARIANT)

Instrument ID
Reference year
Total budget amount in the year
Currency
Type of funding
Specific remarks for budgetary information

DEMOGRAPHIC EVENTS TABLE

Demo ID
Parent Instrument ID
Child Instrument ID
Demo year
Demo type

OrgReg
RFO ID

NATPRO
Instrument ID

Inter-linkage

SIPER
Instrument ID

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

Collection of textual materials

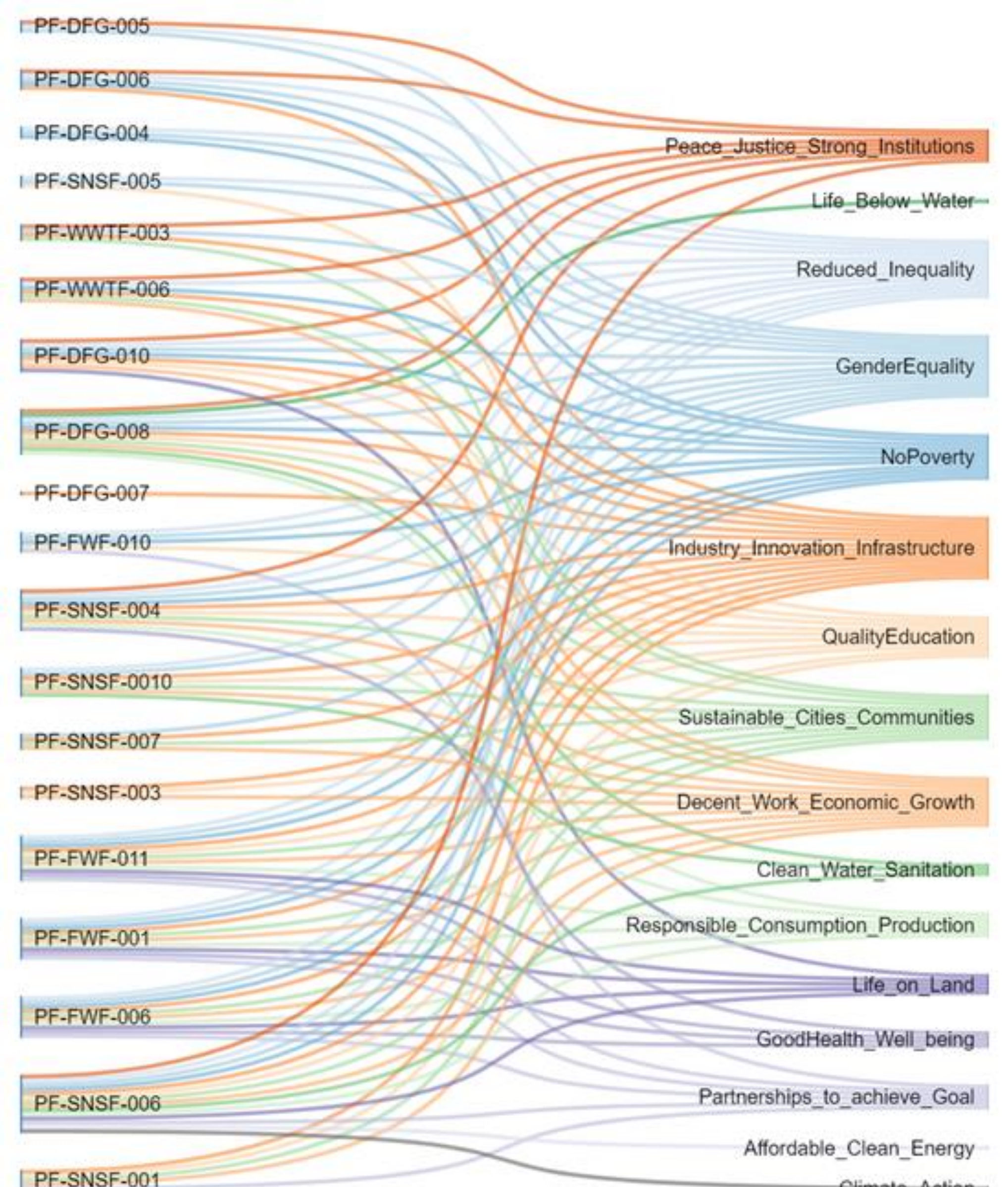
- EFIL hosts a “side-collection” in the form of a repository of **official textual documents regarding the funding instruments** that allow deepening factors related to policy implementation and R&D funding mission orientation.
- Possibility of characterizing the instruments through **text analysis, key words, vocabularies**.
- The documents – **instrument calls, guidelines for participants, descriptions on official webpages**, etc. – are intended to shed light on instruments mission and aim, as well as the elements related to proposal selections (e.g., evaluation criteria)
- The database user will have direct access to a repository hosted on a cloud or can follow a direct link to download the documents from the Internet.

EFIL as a source of textual data

EFIL allows performing text analysis procedures in order to analyze the instruments' call (or other official documents) to understand policy implementation and mission orientation of R&D funding. For each funding instrument we have three indicators on Key Enabling Technologies (**KET**s), Societal Grand Challenges (**SGC**s) and Sustainable Development Goals (**SDG**s).

Construction of the three indicators:

- the selection of the KET or SGC categories are based on the keywords found on selected official documents of the FIs (mainly the call of the RFOs) referring to the respective SGC or KET subclasses ontologies developed within the KNOWMAK project
- the selection of the SDGs categories are based on keywords developed by United Nations Sustainable Development Solutions Network (SDSN) and Monash University

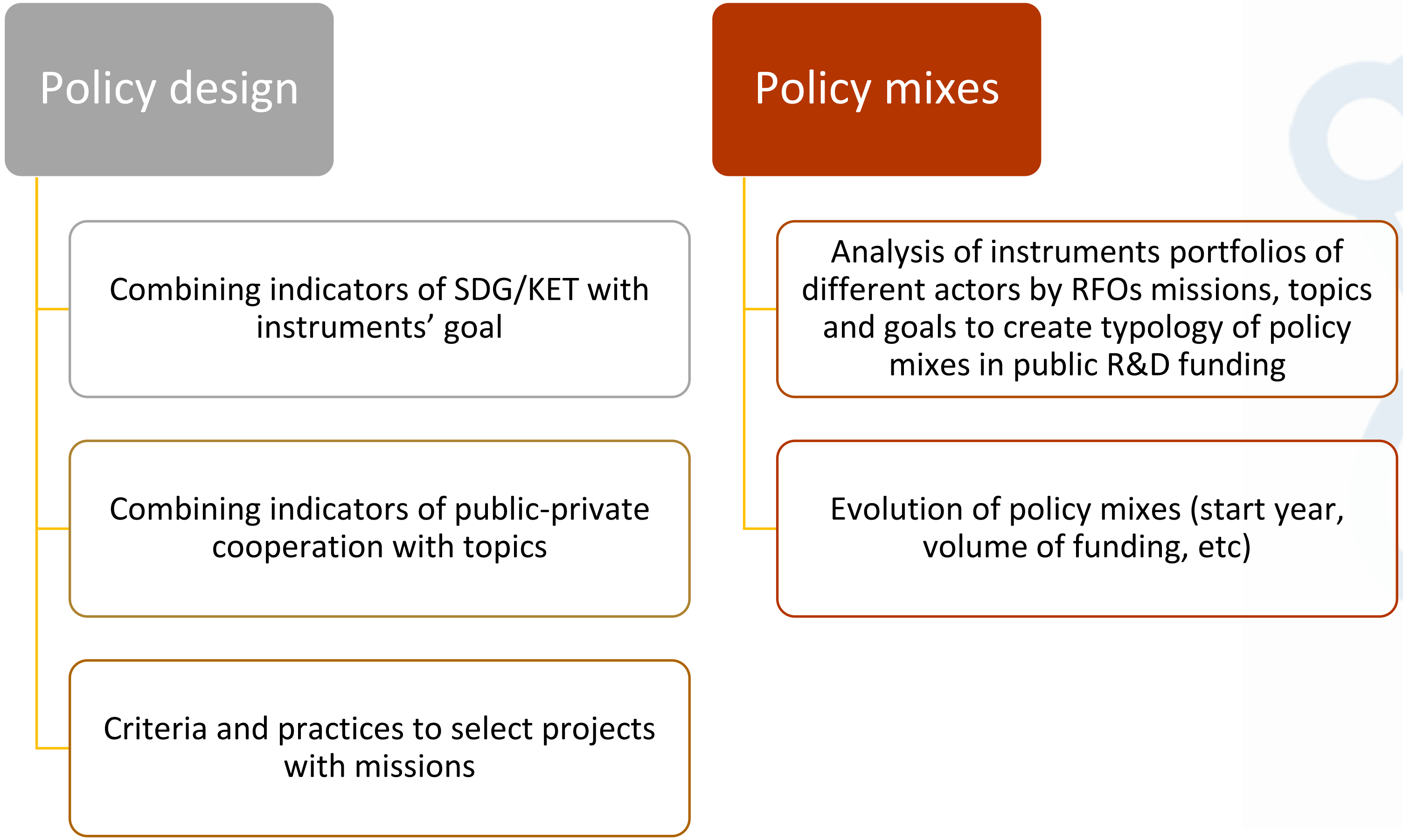


SDGs analysis

Semantic analysis of natural language text based on ontologies performed on selected official documents (calls for proposals, description of the instruments) to delve deeper into the thematic orientation.

Documents related to 25 instruments from 4 RFOs – FWF, WWTF, DFG, SNSF, analyzed for identifying themes from keywords associated with the Sustainable Development Goals (SDGs).

Example of possible usage cases



Policy design

- Combining indicators of SDG/KET with instruments' goal
- Combining indicators of public-private cooperation with topics
- Criteria and practices to select projects with missions

Policy mixes

- Analysis of instruments portfolios of different actors by RFOs missions, topics and goals to create typology of policy mixes in public R&D funding
- Evolution of policy mixes (start year, volume of funding, etc)

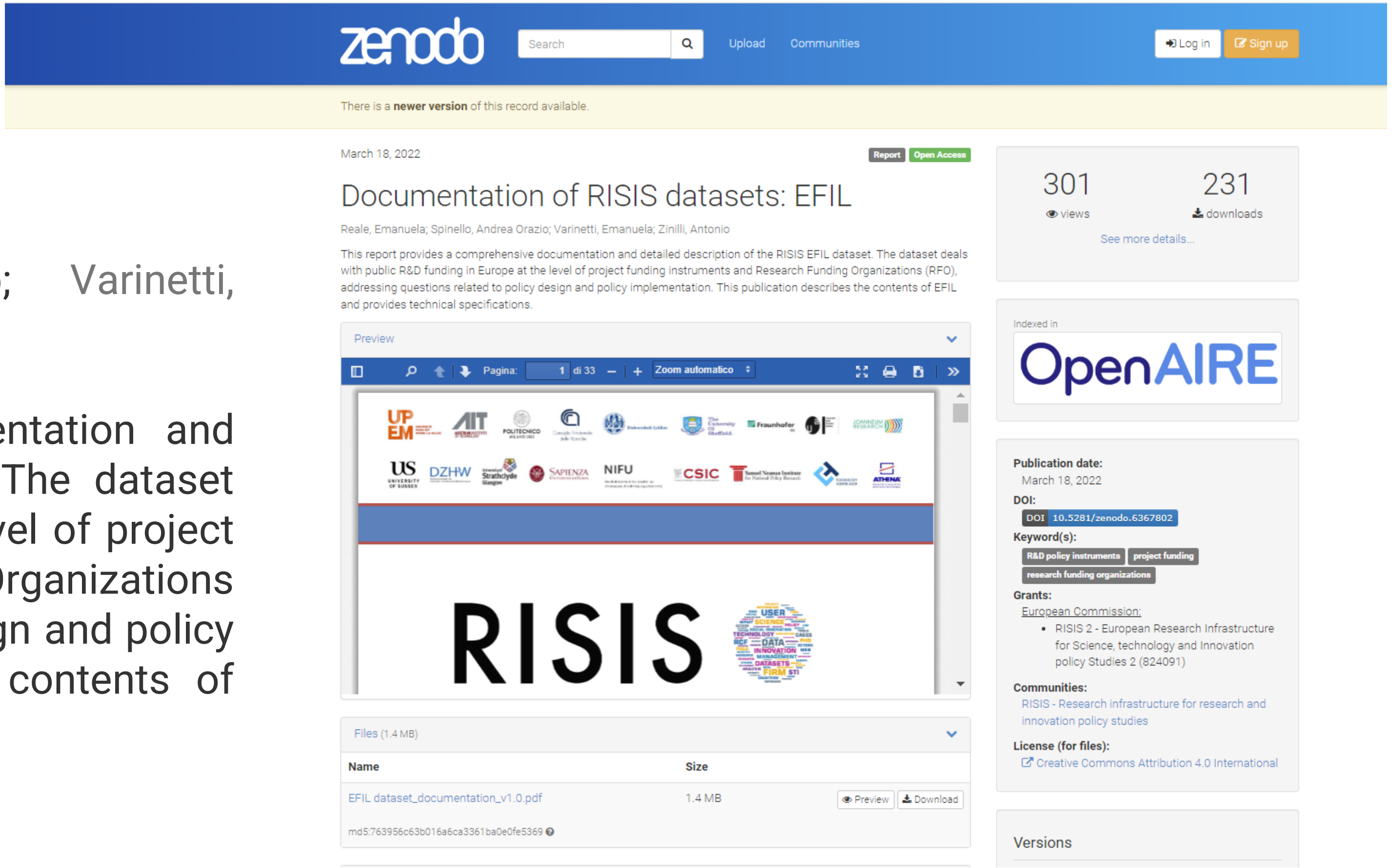
5 EFIL technical report

Documentation of RISIS datasets: EFIL

Reale, Emanuela; Spinello, Andrea Orazio; Varinetti, Emanuela; Zinilli, Antonio

This report provides a comprehensive documentation and detailed description of the RISIS EFIL dataset. The dataset deals with public R&D funding in Europe at the level of project funding instruments and Research Funding Organizations (RFO), addressing questions related to policy design and policy implementation. This publication describes the contents of EFIL and provides technical specifications.

https://zenodo.org/record/6367802



There is a **newer version** of this record available.

March 18, 2022

Documentation of RISIS datasets: EFIL

Reale, Emanuela; Spinello, Andrea Orazio; Varinetti, Emanuela; Zinilli, Antonio

This report provides a comprehensive documentation and detailed description of the RISIS EFIL dataset. The dataset deals with public R&D funding in Europe at the level of project funding instruments and Research Funding Organizations (RFO), addressing questions related to policy design and policy implementation. This publication describes the contents of EFIL and provides technical specifications.

301 views

231 downloads

See more details...

Files (1.4 MB)

Name	Size	
EFIL dataset_documentation_v1.0.pdf md5:763956c63b016a6ca3361ba0e0fe5369	1.4 MB	Preview Download

Publication date:
March 18, 2022

DOI:
DOI 10.5281/zenodo.6367802

Keyword(s):
R&D policy instruments; project funding; research funding organizations

Grants:
European Commission:
- RISIS 2 - European Research Infrastructure for Science, technology and Innovation policy Studies 2 (824091)

Communities:
RISIS - Research infrastructure for research and innovation policy studies

License (for files):
Creative Commons Attribution 4.0 International

How to access RISIS data

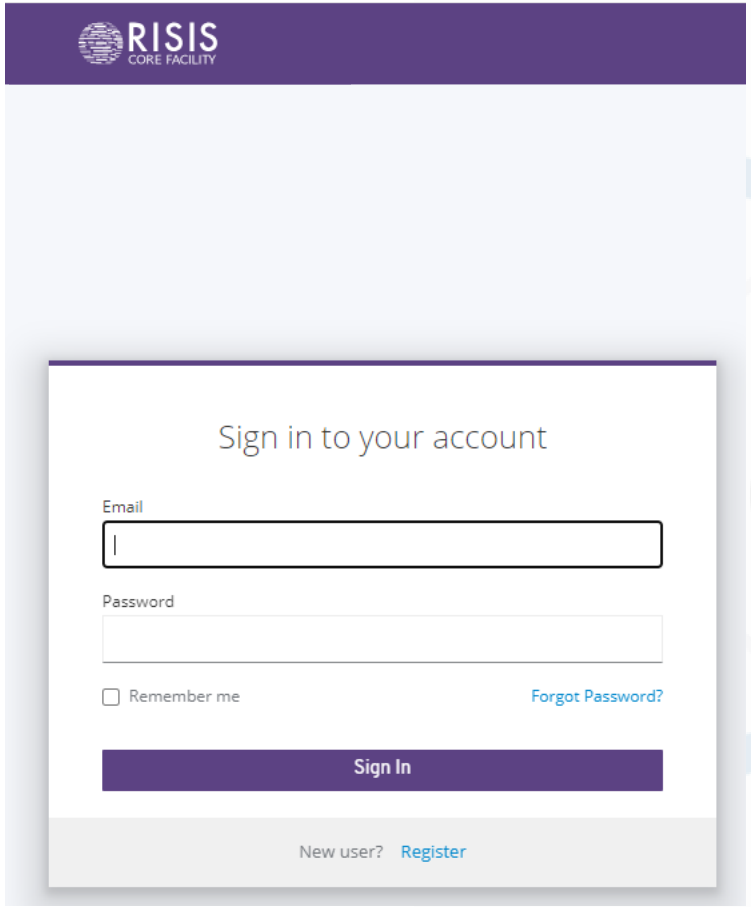
RISIS infrastructure must provide a **unique entry point** online

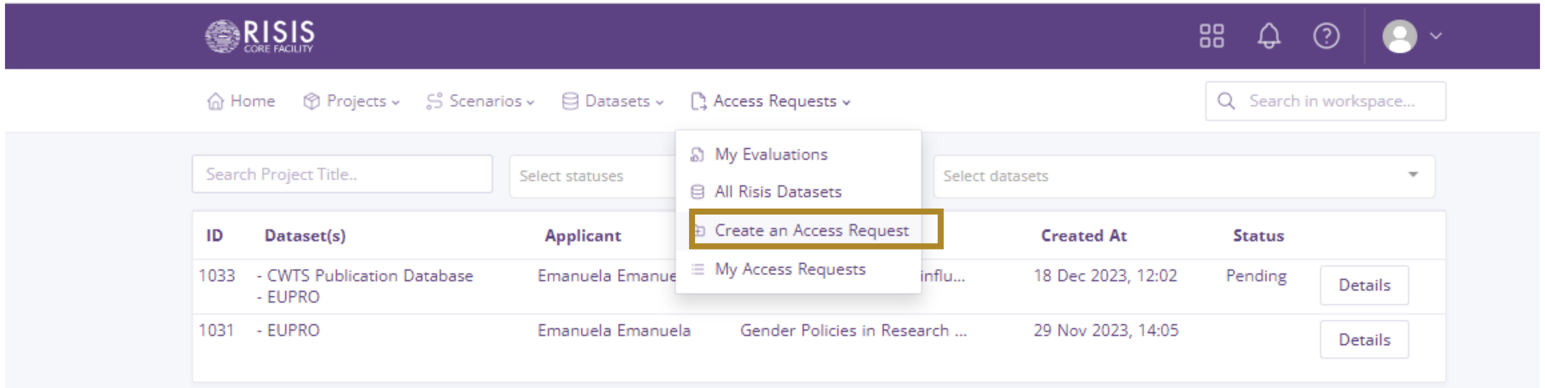
Users can access to a monitored and secured **workspace.**

This workspace is designed to provide services to users interested in jointly exploiting different RISIS datasets and various Linked Open Data resources with the goal to explore, retrieve and visualize results of data analysis for their research purposes

Login into **RISIS Core Facilities**

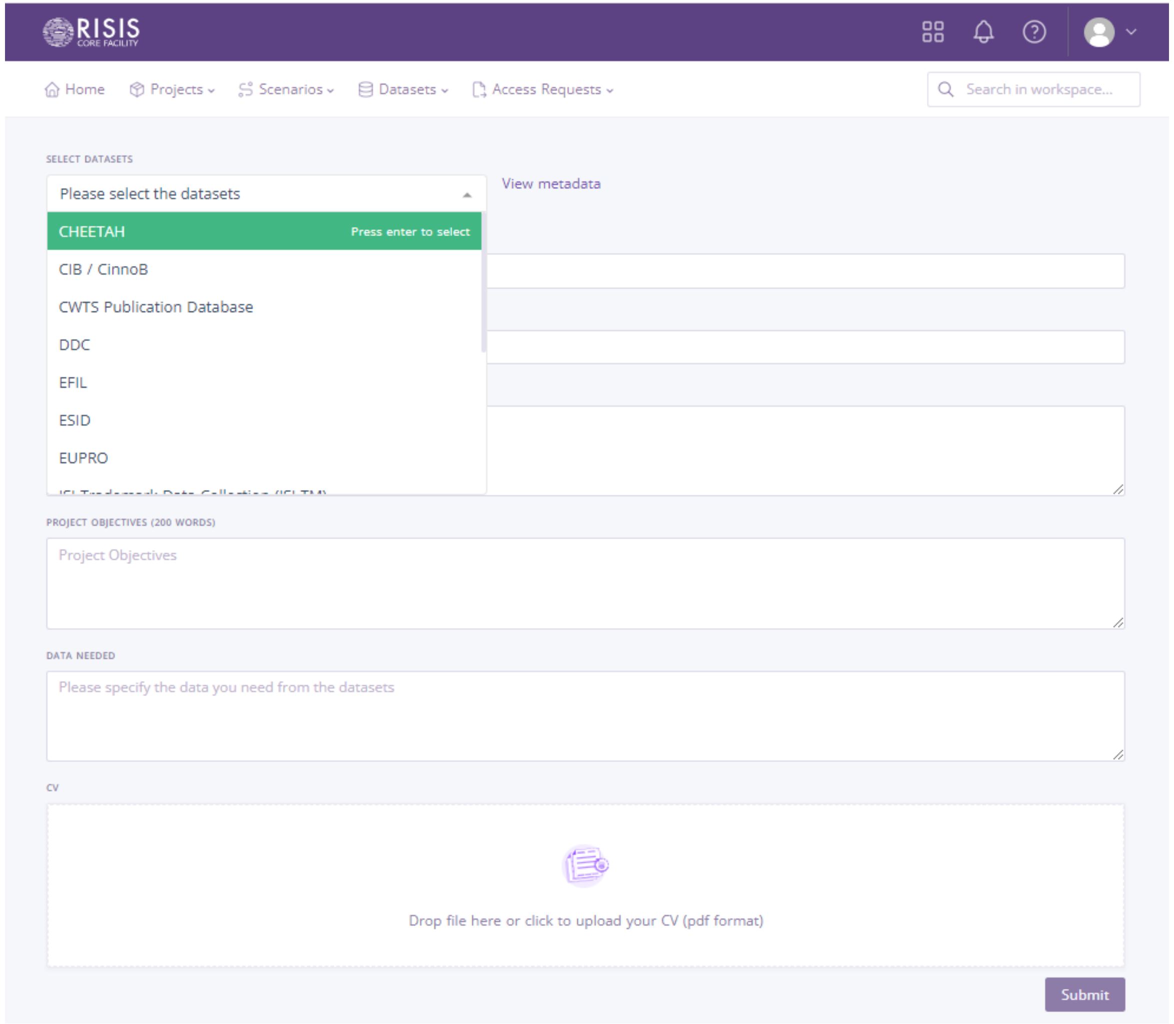
https://rcf.risis.io





ID	Dataset(s)	Applicant		Created At	Status	
1033	- CWTS Publication Database - EUPRO	Emanuela Emanue	influ...	18 Dec 2023, 12:02	Pending	Details
1031	- EUPRO	Emanuela Emanuela	Gender Policies in Research ...	29 Nov 2023, 14:05		Details

Click 'Access Request' item in navbar, then 'Create an Access Request'



SELECT DATASETS

Please select the datasets

View metadata

- CHEETAH — Press enter to select
- CIB / CinnoB
- CWTS Publication Database
- DDC
- EFIL
- ESID
- EUPRO
- ISI Trademark Data Collection (ISI-TM)

PROJECT OBJECTIVES (200 WORDS)

Project Objectives

DATA NEEDED

Please specify the data you need from the datasets

CV

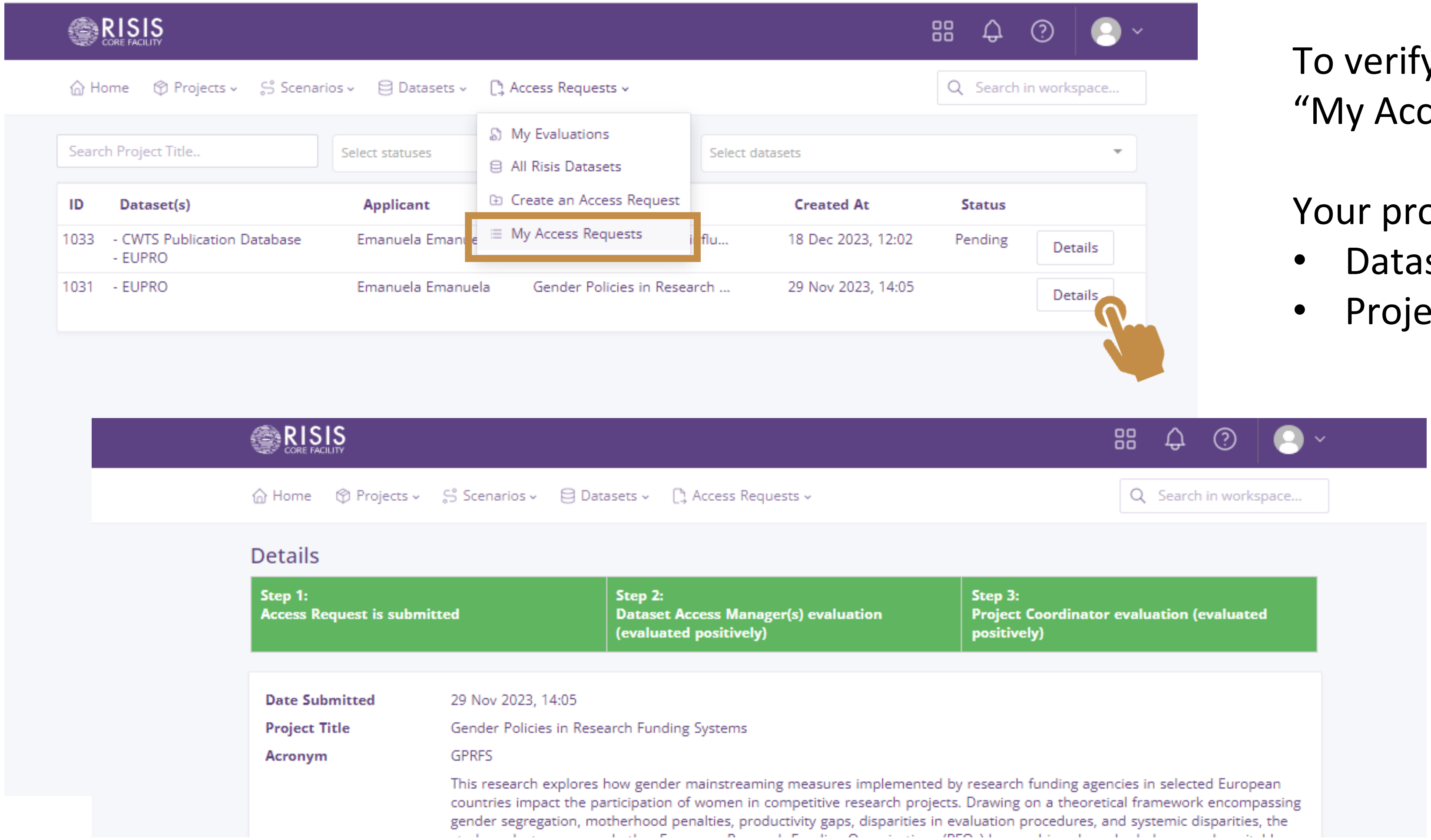
Drop file here or click to upload your CV (pdf format)

Submit

First, the project idea!

Explore all datasets descriptions and let be inspired by data

- Access Request form page
- Select at least one dataset
- Proceed to fill the form
- Brief research project
- Identify Data needed from dataset
- Upload the CV
- Submit



ID	Dataset(s)	Applicant		Created At	Status	
1033	- CWTS Publication Database - EUPRO	Emanuela Emanu	flu...	18 Dec 2023, 12:02	Pending	Details
1031	- EUPRO	Emanuela Emanuela	Gender Policies in Research ...	29 Nov 2023, 14:05		Details

Details

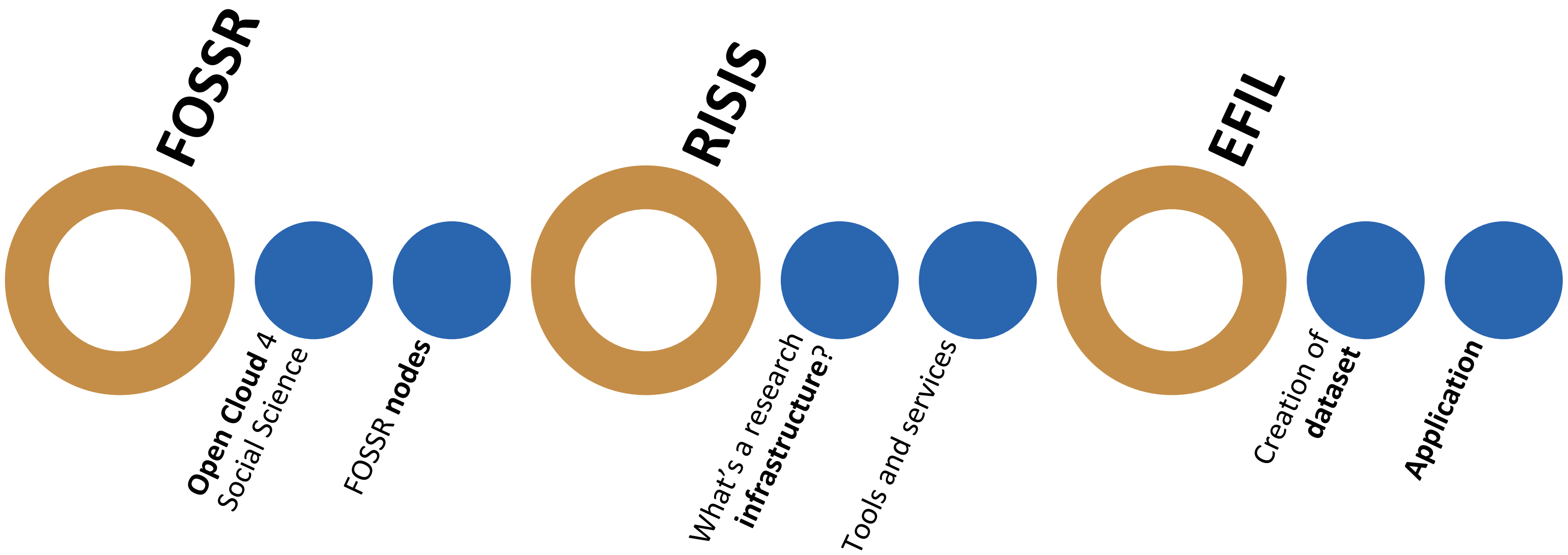
Step 1: Access Request is submitted	Step 2: Dataset Access Manager(s) evaluation (evaluated positively)	Step 3: Project Coordinator evaluation (evaluated positively)

Date Submitted	29 Nov 2023, 14:05
Project Title	Gender Policies in Research Funding Systems
Acronym	GPRFS
	This research explores how gender mainstreaming measures implemented by research funding agencies in selected European countries impact the participation of women in competitive research projects. Drawing on a theoretical framework encompassing gender segregation, motherhood penalties, productivity gaps, disparities in evaluation procedures, and systemic disparities, the

To verify the status of you request go to “My Access Requests”

Your project is evaluated by two figures:

- Dataset access manager
- Project coordinator



FOSSR

Open Cloud 4 Social Science

FOSSR **nodes**

RISIS

What's a research **infrastructure?**

Tools and services

EFIL

Creation of **dataset**

Application

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Consiglio Nazionale delle Ricerche

THANK YOU!

✉ fossr.dissemination@ircres.cnr.it

@fossrproject

fossr.eu

fossr-eu

@fossr

zenodo.org/communities/fossr

Untying acronyms

Initiatives, strategies and networks

European Research Infrastructure Consortium (ERIC)
A specific legal form that facilitates the establishment and operation of Research Infrastructures. More about ERIC

European Open Science Cloud (EOSC)
Cloud database for research in Europe. More about EOSC

Association of European-Level Research Infrastructures Facilities (ERF-AISBL)
A not-for-profit association promoting the development and visibility of European infrastructures providing access to external users. The ERF members are open at an international level and include national infrastructures as well as European networks and consortia of research infrastructures. Every year the ERF member organisations serve over 20,000 academic and industrial users from Europe and overseas. More about ERF-AISBL

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Consiglio Nazionale delle Ricerche

Module II: Accessing and querying interoperable RI data

Part I: EFIL DATA exploitation

Gender Policy and Practice in R&D Funding Instruments in Europe

FOSSR
Fostering Open Science in Social Science Research
Innovative tools and services to investigate economic and societal change

Emanuela Varinetti
Researcher at CNR-IRCrES

«Exploring research infrastructure data through advanced statistical applications»

FOSSR Online Training

26-27 September 2024

Goal

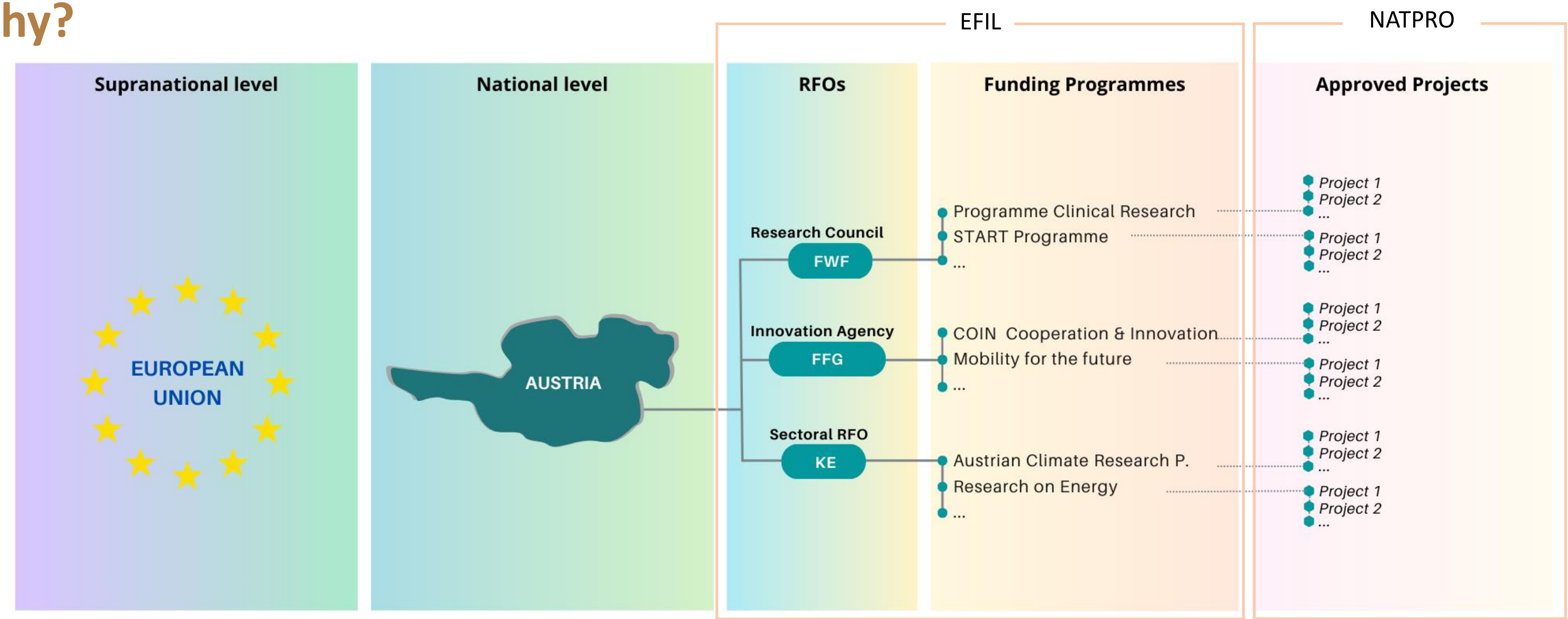
The study investigates how Research Funding Organizations (RFOs) have transposed the strategies promoted by the European Union to reach gender equality in R&D and achieve goal 5 “gender equality” of SDG 2030

Through an analysis of research funding programmes promoted by RFOs in ten European countries, we intend to identify:

- whether there is an orientation to equal opportunities both in the strategic action plans and in the promotion of the gender dimension as a cross-cutting element of the funding instruments

The study aims to identify the policy orientation of RFOs towards gender equality by examining competitive funding programmes

Why?



Need for more research on the effects of **EU measures and implementation activities in specific organizations** to identify how they contribute to a convergence **of gender equality policies and practices among R&I organizations and national R&I systems** (EIGE GEAR tool Analytical Paper, 2022) Resources distributed towards competitive funding programmes, addressing targeted research objectives, should improve the government's ability to control content of research activities developed by researchers, as well as producing effects on society (Braun, 2006) importance to underline the role of the funders in shaping scientific research toward "targeting" thematic orientation (Aagaard et al., 2021)

Research questions

- ***Q1*** *How have European RFOs at the national level transposed strategies promoted by supranational institutions (European Union and UN SDG 2030) regarding gender equality in R&D competitive funding programmes and in their strategic action plans?*

- **H1**: There is a growing orientation on the promotion of gender dimension as cross-cutting items both in strategic action plans and in **competitive funding programmes** promoted by RFOs. Through these actions, a clear political orientation of the RFO is outlined.

Data and Methods

Data sources:

- EFIL dataset exploration for analysis on research funding programmes

Temporal coverage: 2017/18 – 2020/21

Geographical coverage:

- Analysis of programmes : 10 European countries considered in EFIL perimeter, with focus on 4 selected RFOs from Austria (FWF), Germany (DFG), Italy (MUR), France (ANR)

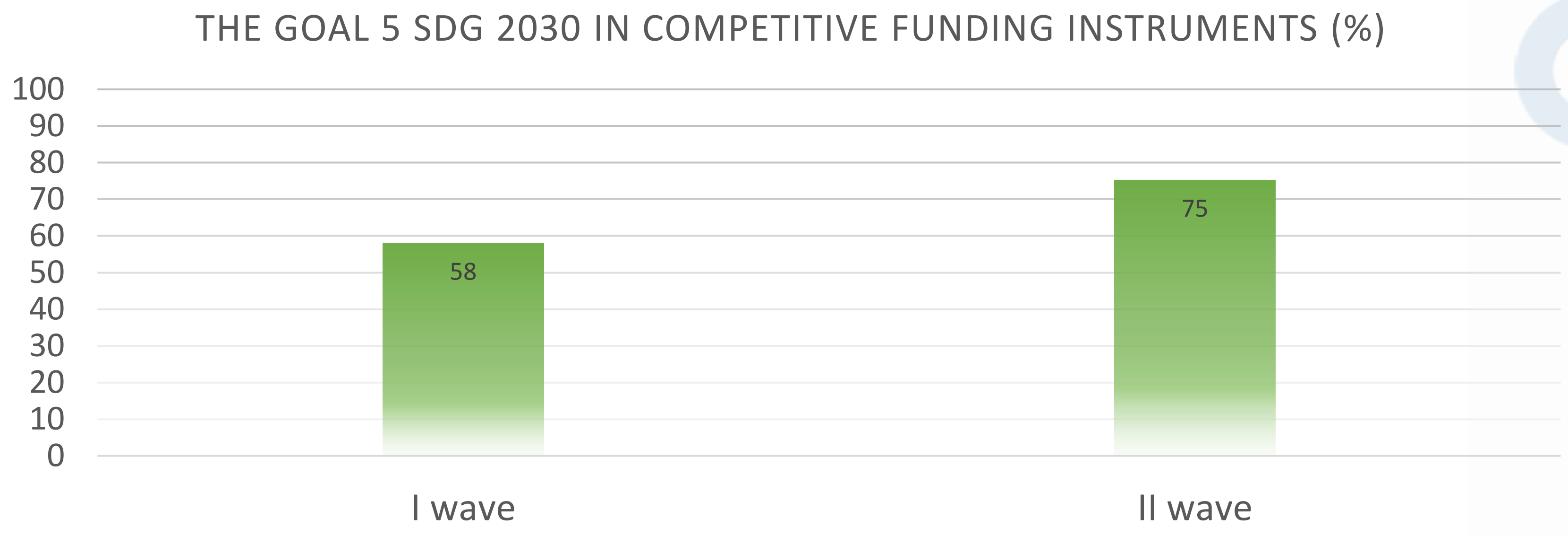
EFIL SDGs keywords extraction

EFIL store a repository of official documents pertaining to funding instruments retrieved from RFO websites (e.g., instruments calls, reports, and guidelines for applicants)
Documents allows a deeper understanding of factors relating to policy implementation and R&D funding orientation.
*Based on the official documents, an automated text analysis process was used to generate SGC, KET**, SDG** descriptors in database. The **selection of the SDGs is based on RISIS2 ontology***

Two waves of data and documents collection:
First wave : data on instruments active in 2017-2018
Second wave: new instruments active in 2020-2021 and the update of data of the instruments collected during the first wave

EFIL perimeter: 55 relevant RFOs from 10 countries: Austria (AT), Czech Republic (CZ), Denmark (DK), Estonia (EE), France (FR), Germany (DE), Italy (IT), Norway (NO), Switzerland (CH), United Kingdom (UK).

Gender equality keywords in funding programmes 1/3



Gender items in calls increased by 17% between 2017–18 and 2020–21

Perimeter of observation: all ten countries selected in EFIL

FUNDING PROGRAMMES LEVEL

Gender keywords in competitive funding programmes 2/3

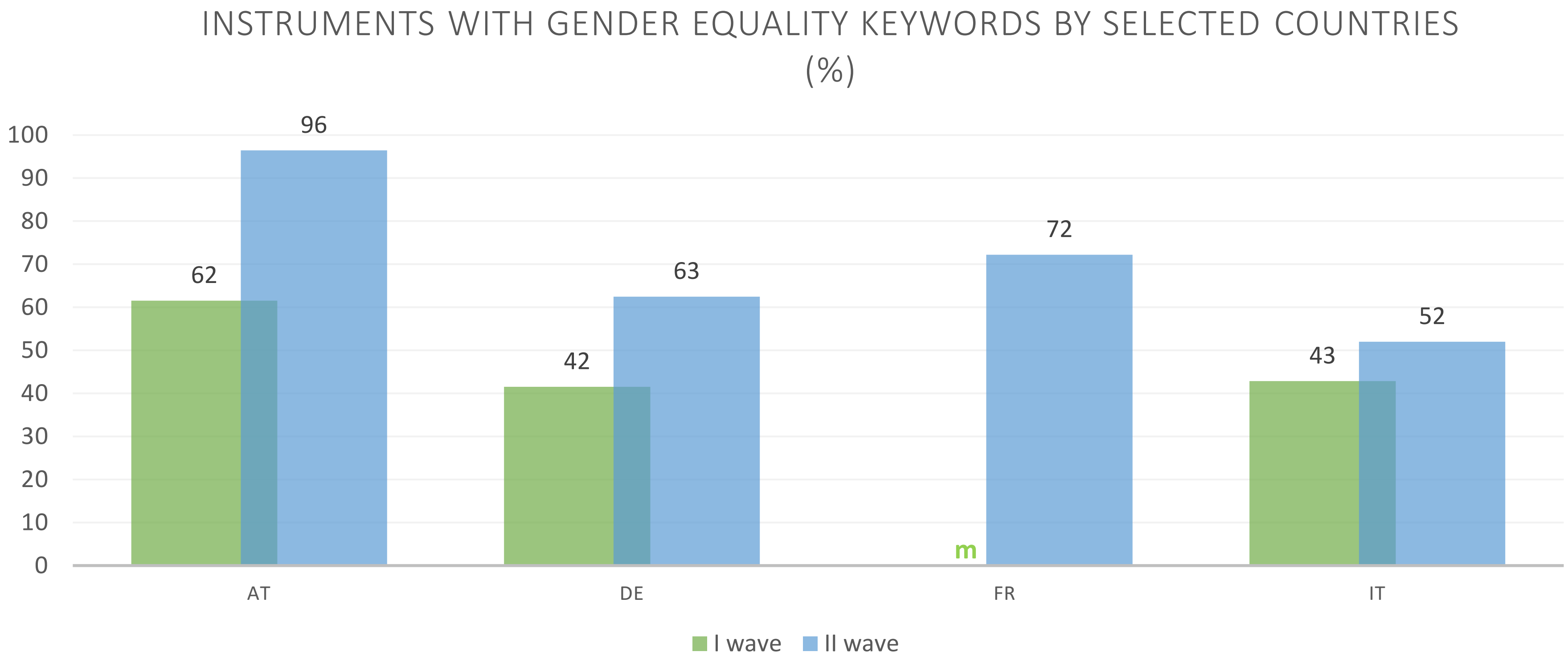
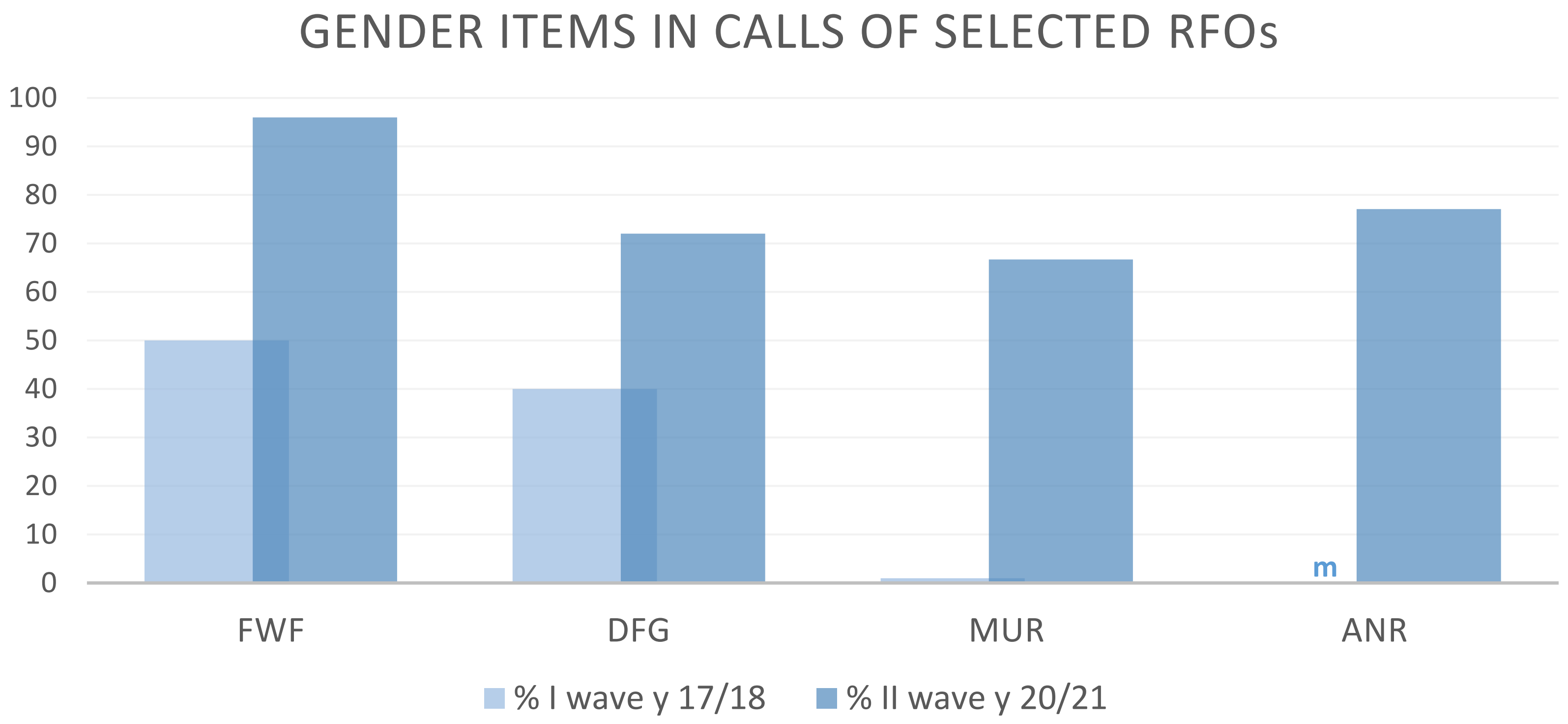


Figure shows the orientation towards SDG goal 5 of **all RFOs** of four selected countries

FUNDING PROGRAMMES LEVEL

Gender keywords in competitive funding programmes 3/3



PROGRAMMES/ RFOs LEVEL

Next step

NATPRO –text analysis of abstract of project financed by competitive funding programmes

Connecting projects (Natpro) with programmes (EFIL)

Data on funding instruments gender oriented will be combined with data on participation level of different types of performers.

To observe if this orientation is reflected in the research projects funded by the RFOs and understanding how research performers transpose and implement these measures in their research proposals

References

1. Spinello A. O., Varinetti E., Zinilli A., Reale E., (2023), Il finanziamento competitivo della Ricerca e Sviluppo in Italia per le sfide sociali e tecnologiche, in Consiglio Nazionale delle Ricerche, Relazione sulla ricerca e l'innovazione in Italia. In corso di stampa
2. Reale, E., Spinello, A.O., Varinetti, E., Zinilli, A. (2023). Documentation of RISIS datasets: EFIL. Version 2.0 RISIS 2 - European Research Infrastructure for Science, technology and Innovation policy Studies 2, grant agreement 824091. Available on Zenodo: https://zenodo.org/record/7938134#.ZGtopnZByUl
3. Zinilli, A., Reale, E., Spinello, A.O., Varinetti, E. (2022). "Diversification vs. specialization from the perspective of research programmes: a complexity approach". In N. Robinson-Garcia, D. Torres-Salinas, & W. Arroyo-Machado (Eds.), Proceedings of the 26th International Conference on Science and Technology Indicators, STI 2022(sti2256).
4. Spinello, A.O., Reale, E., Zinilli, A. (2021). Outlining the orientation toward socially relevant issues in competitive R&D funding instruments. Frontiers in Research Metrics and Analytics, 6:712839
5. Zinilli, A. (2021). Imputation methods for estimating public R&D funding: evidence from longitudinal data. Qual Quant 55, 707–729.

https://www.ircres.cnr.it/index.php/it/?option=com_content&view=article&id=422

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Consiglio Nazionale delle Ricerche

Module II: Accessing and querying interoperable RI data

Part II: Part II: The use for research of two RISIS datasets - Features of MORE and Cheetah

FOSSR
Fostering Open Science in Social Science Research
Innovative tools and services to investigate economic and societal change

Lucio Morettini
Researcher at CNR-IRCrES

«Exploring research infrastructure data through advanced statistical applications»

FOSSR Online Training

26-27 September 2024

Introduction

Presentation of the characteristics of two datasets present in RISIS - Research Infrastructure for Science and Innovation policy Studies:

- **MORE** – Mobility Survey of the Higher Education Sector.
- **Cheetah** – Medium-sized fast growing firms dataset.

Main elements of the presentations:

- Structure of the datasets;
- Strengths of the available data;
- Scientific interest.

Context

MORE (Mobility Survey of the Higher Education Sector) is an empirical study that extensively analyses the mobility of European researchers.

- The analysis of the factors that influence **researchers' career paths** has progressively required the possibility of identifying more detailed elements that characterize researchers' scientific and professional development.

- **MORE** responds to the need to identify at **what point in their careers** researchers have been able to enrich their experience through mobility, both towards other countries and in fields other than academia.

MORE 3 dataset, created between 2016 and 2018 and integrated into RISIS in 2019, is the evolution of previous datasets focused on the international mobility of researchers.

Structure of MORE 3

MORE 3 is an original data collection based on information collected through the administration of a 106-question questionnaire.

- The dataset contains information relating to **10,394 valid responses** from researchers in **31 European countries** (EU28 plus Switzerland, Iceland and Norway). An estimated population frame of 1.373 million researchers for EU28 and 1.439 million in total for the 31 countries.

- The questionnaire is the result of a **work of improvement of previous experiences** (MORE and MORE 2), which involved a larger audience of researchers and a more detailed search for information on the typology of international mobility experiences, placing them temporally and geographically.

Main elements of MORE 3

The main focus of the dataset is Academic Mobility and Career Paths, with questions aimed at collecting data on:

a. Mobility during the PhD;
b. Mobility during the career, with a sub-focus on Short-term Mobility (<3 months);
c. Intersectoral Mobility;
d. Comparisons with forms of virtual mobility (distance collaborations);
e. Reasons for non-mobility choices.

Refer to mobility experiences in the 15 years prior to the time of the interview, including motivations, barriers and effects of mobility.

Main elements of MORE 3

Detailed contextualization with respect to the elements of the research career. MORE 3 contains information related to working life:

- Current employment activities and working conditions;
- Timing of work experiences and mobility episodes;
- Definition of previous career stages through standardization of job positions developed by the European Commission:

  R1: First Stage Researcher (up to the point of PhD),
  R2: Recognized Researcher (PhD holders or equivalent who are not yet fully independent),
  R3: Established Researcher (researchers who have developed a level of independence),
  R4: Leading Researcher (researchers leading their research area or field).

So, what is MORE 3?

Although conceived as a data collection focused on researchers' mobility during their career, MORE has become a dataset that compares the details of the professional paths of European researchers, with the possibility of identifying:

- standardized classification of the transitions from one career stage to the next,
- timing between two stages,
- details on international mobility, extra-sectoral experiences and work characteristics of each career path.

Lack of information in MORE 3:

- Protection of respondents' privacy that does limits details on scientific productivity at individual and institutional level;
- Impossibility to interconnect the dataset with other RISIS datasets.

How to use MORE 3

Elements that favored the development of a research hypothesis:

- timing between career steps;
- characterization of professional activities in each phase;
- contextualization of the academic career.

Final research hypothesis: analyze gender differences in the timing of career progression.

Reference paper under review, related working paper: Morettini L. & Tani M.; Gender and Career Progression in Academia: European Evidence, IZA DP No. 16206, 2023

How to use MORE 3

Sample selection:

- coherence in career steps;
- limitation of the sample to researchers;
- elements that guarantee the contextualization of careers:

  inclusion in the sample of data external to MORE 3

  → share of women in total researchers (source: OECD)

  → expenditure in R&D (source: Eurostat)

How to use MORE 3

Main **characteristics of the sample**:

Career stage	Observation	Women	Men
R2	1,575	742	833
R3	3,792	1,635	2,157
R4	1,567	527	1,040

Analysis **strategy**:

estimation of the impact of the professional and personal characteristics of each researcher on the **duration** of each step **subject to the probability** of having a career progression.

Discussion

The results of our analysis show two main features:

- women are less likely to overcome the bottleneck of selections between steps, but those who pass do so faster than their male colleagues;
- the transition from the second to the third career stage is governed by different dynamics than the transition from the third to the fourth stage.

These results show how MORE allows to analyze the development of research careers from an internal perspective, going beyond the intentions of its authors.

Despite some limitations, MORE offers a broad perspective on the characteristics of research careers, allowing to overcome differences in academic paths of different countries.

The possibility of contextualizing the different elements that form research careers temporally and geographically, allows to enrich the dataset with data from other sources.

Context

Cheetah dataset presents geographical, industry, accounting and ownership information of medium-sized firms that experienced fast growth rates in the periods 2008 - 2020.

The characterizing elements of the companies contained in Cheetah:

Medium size → number of employees between 50 and 249 and either a turnover of not exceeding €50 million (Eurostat definition) or number of employees between 250 and 4999 and either a turnover of not exceeding € 1.5 billion (Entreprise de taille intermédiaire definition).

Fast-growth firms → firms with average annualized growth of greater than 20% over a three-year period. Growth is measured by either the number of employees or by turnover.

Context

The aim of Cheetah is to cover the long-term economic performance of this peculiar class of firms, as one of the main pillars of the European industrial and technological system, providing an entry point for RISIS in the analysis of **research and innovation policies for and towards enterprises**.

Why medium-sized firms? Firms that are already well-structured and are in a **transitional phase**, more inclined to **innovation**.

Why fast-growth firms? Firms that experience a **growth shock** and must manage the **sudden change** they are subjected to, for which different forms of innovation can represent the cause of the shock or the solution to the shock.

Structure of Cheetah

Source of data: Orbis - developed by Bureau van Dijk based on globally collected balance sheet data

Cheetah contains observational data based on three-year cohorts, starting from the 2008 - 2011 cohort up to the 2017 - 2020 cohort (latest version: Cheetah v. 3.0), for a total of **10 cohorts** and **13 years** of observation.

Total sample: 129,752 firms, in 30 European countries.

Main steps:
1. Identification of European medium-sized firms.
2. Selection of fast growing firms by applying the OECD definition.
3. Data collection and organization of accounting information.
4. Data collection and organization of ownership information.
5. Data collection and organization of information on M&A activity.
6. Geocoding.

Limitations of information in Cheetah

The careful selection work to create the dataset does not allow to avoid some biases that must be addressed:

- Data source: balance sheet data of firms;
- Data harmonization procedures at international level;
- Inconsistency or lack of data.

Scientific Interest in Cheetah

Peculiarities of the sample: Cheetah firms are not a minority but an elite.

Completeness of the sample: firms are not observed only in fast-growth spurt but all along their life.

Point-based analysis structuring: dataset structure structure allows for point-based analysis on the spatial or sectoral or procedural level.

Possibility of interconnecting Cheetah with other datasets present in RISIS: **RISIS Patent**.
RISIS Patent database derives from the EPO PATSTAT. The database is designed for the analysis of technological knowledge creation, using patent as a proxy.
The information is directly traceable to other RISIS datasets such as Cheetah, providing a framework for the innovative effort undertaken by fast-growing firms to relate it to the growth shock.

An early use of Cheetah

Reference paper: Morettini L., Potì B., Gabriele R.; Process and strategies of growth in medium-sized fast-growing firms. (2023) Journal of Industrial and Business Economics 50.

Approach to the dataset in its first version, data available from 2008 to 2013. Sample limited to only 1,666 Italian companies, for a total of 7,914 observations.

Despite the limitation in data availability, in the paper we have addressed the issue of the effects of fast growth on the internal organization of firms.

We structured the analysis to observe how the growth shock pushes firms to review their internal structure in terms of work organization, external financial resources, and reorganization of the corporate structure.

Discussion

The main result of our analysis was the identification of a shared reaction among Italian fast-growing firms, tending to compensate the effect of growth through reorganization strategies represented by an increase in spending on external components.

From a broader perspective, the test demonstrates that Cheetah allows an analysis of firm behavior in exceptional conditions, allowing behavioral indicators to be found among balance sheet data.

The enrichment of Cheetah with new and more detailed information and the development of interoperability between RISIS databases opens the way to studies on the role of innovation of a group of highly versatile companies such as Fast-Growing Firms.

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Consiglio Nazionale delle Ricerche

Module II: Accessing and querying interoperable RI data

Part III: Combining interlinked dataset in a RI: the case of RISIS-EFIL and RISIS-NATPRO

FOSSR
Fostering Open Science in Social Science Research
Innovative tools and services to investigate economic and societal change

Andrea Orazio Spinello
Researcher at CNR-IRCrES

«Exploring research infrastructure data through advanced statistical applications»

FOSSR Online Training

26-27 September 2024

Context

- **Project-based R&D activities supported by public research funding organisations (RFOs)** has become a core research issue in science and innovation policy studies

- RISIS infrasructure has put a strong emphasis in responding to this growing interest and increased demand from the research community but also policy makers by maintaining and developing datasets on **project-based R&D**

→ **EUPRO** is an established RISIS dataset, widely used by the RISIS research community and important impetus to RISIS indicator tools (KNOWMAK, OrgReg and FirmReg), but missing nationally funded R&D projects -> **Creation of the NATPRO module on national R&D project**

→ **EFIL** dataset created between 2019 and 2022 to collect more systematic and comprehensive information on **R&D funding programmes** and their manifold characteristics

A joint use case: Funding and project patterns of the German RFO DFG - German Research Foundation

- Presentation of some exemplary empirical illustrations of the two datasets, mobilizing their inter-linking in a rough analysis of the funding patterns observed for the German Research Foundation (DFG) as one main example of newly collected data in both datasets.

- **Data from EFIL** with present some peculiar characteristics of the DFG funding instrument portfolio in terms of aims, type of funding and budget mobilized.

- **Data from NATPRO** will shed light on the performers' side, the level of projects and participations in the DFG funding opportunities.

- Reference period for the data presented is 2016-2018.

All the following data and figures are from
Heller-Schuh, B., Reale, E., Scherngell, T., Spinello, A.O., Varinetti, E., Zahradnik, G., Zinilli, A. (2022). New insights into project-based R&D funding from RISIS datasets: Some evidence from EFIL and NATPRO. RISIS Policy Brief Series, Issue #10.

DFG funding portfolio at a glance

- DFG provides a wide range of funding instruments for individuals, research groups, and institutions.
  - ➢most of the funding schemes pursue a bottom-up approach;
  - ➢globally the portfolio is aimed at curiosity-driven research;
  - ➢low orientation to economic innovation or specific policy goals.

- 30% of agency investment on “Individual Research Grants”

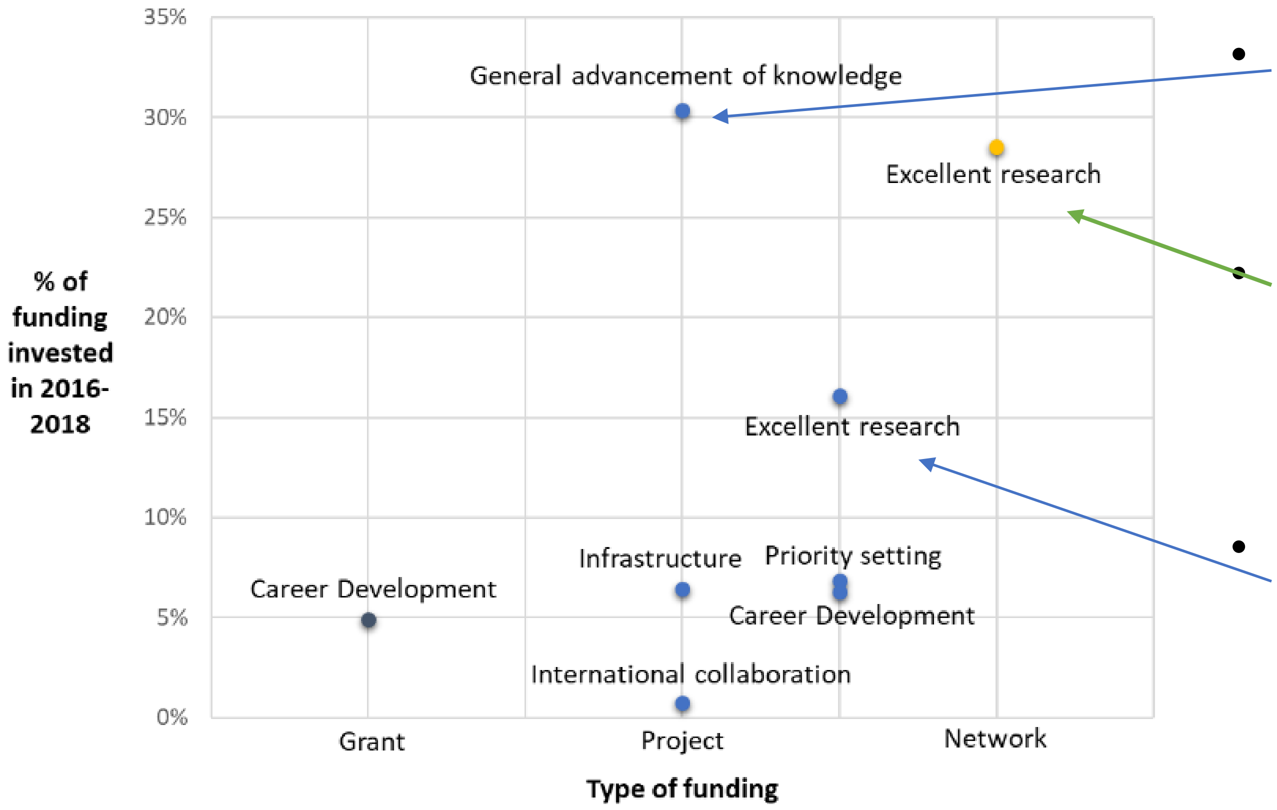
- Small set of programs targeted to individuals for career development

- Presence of funding schemes devoted to scientific excellence and the fulfillment of relevant thematic priorities (a top-down approach is pursued for “Priority programmes”)

- Coordinated programs that enable the formation of long-term networks of researchers (e.g., “Research Units”, “Clinical Research Units”) or institutional networks - e.g., “Collaborative Research Centers”; “DFG Research Centers”.

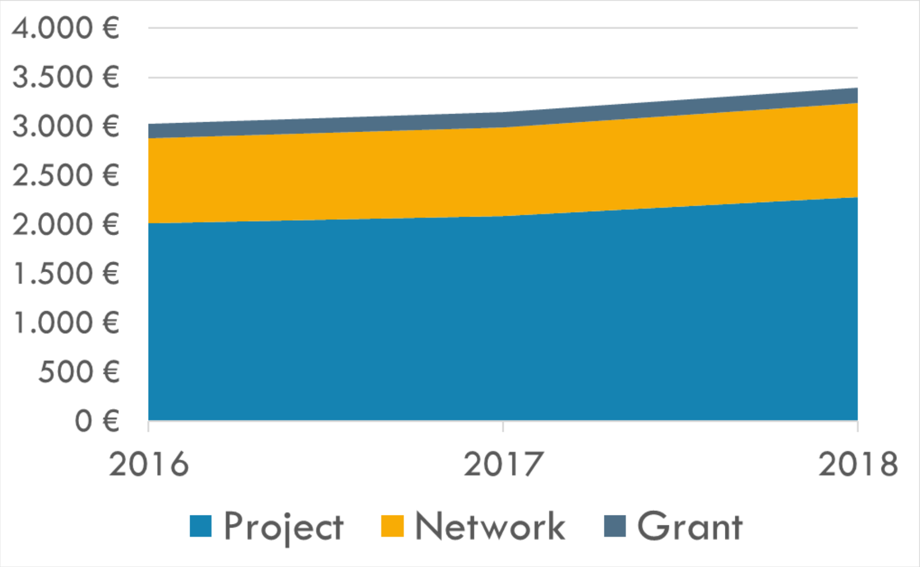
DFG funding by instrument aim and type of funding 2016-2018

- The agency invested significantly in basic research.
- Excellence and internationalization of research are at the core of DFG portfolio.



- Funding projects limited in time and scope pursue the general advancement of knowledge in all disciplinary fields.
- A consistent part of the DFG budget is targeted to cooperative excellence research between different organizations forming a network and for the establishment of long-term research units (28%)
- The Excellence Initiative (17% of the 2018 budget) promotes cutting-edge research contributing to German universities' international competitiveness.

Evolution of funding by type



Amounts expressed in thousand Euro

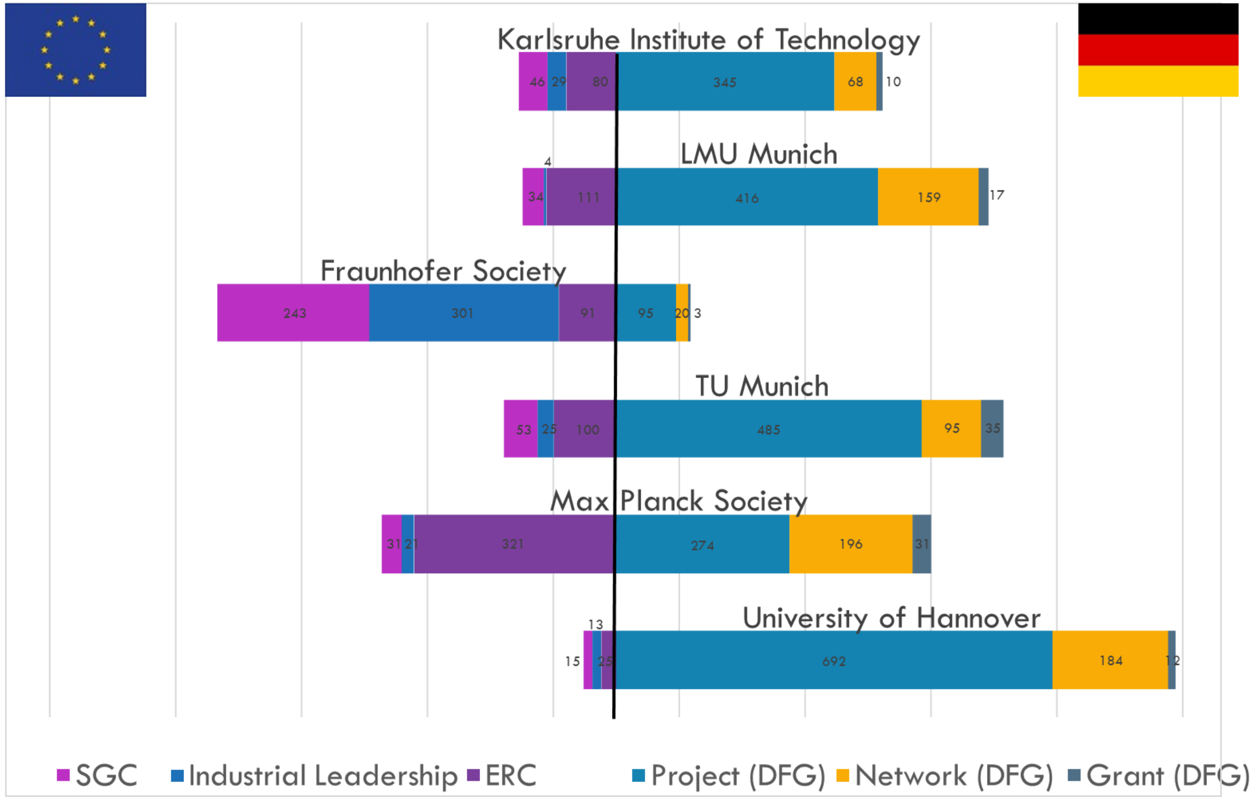
- ○ Project = typical project funding with a scope and duration
- ○ Grant = funding to people for career advancement plus awards and prizes
- ○ Network = establishment of long-term cooperative research

- From 2016 to 2018 there has been an increase of the total DFG funding (12%).

- The various types of funding (project, grant, network) have not shown significant proportional differences, with the smaller percentage of funding allocated to personal grants for career development

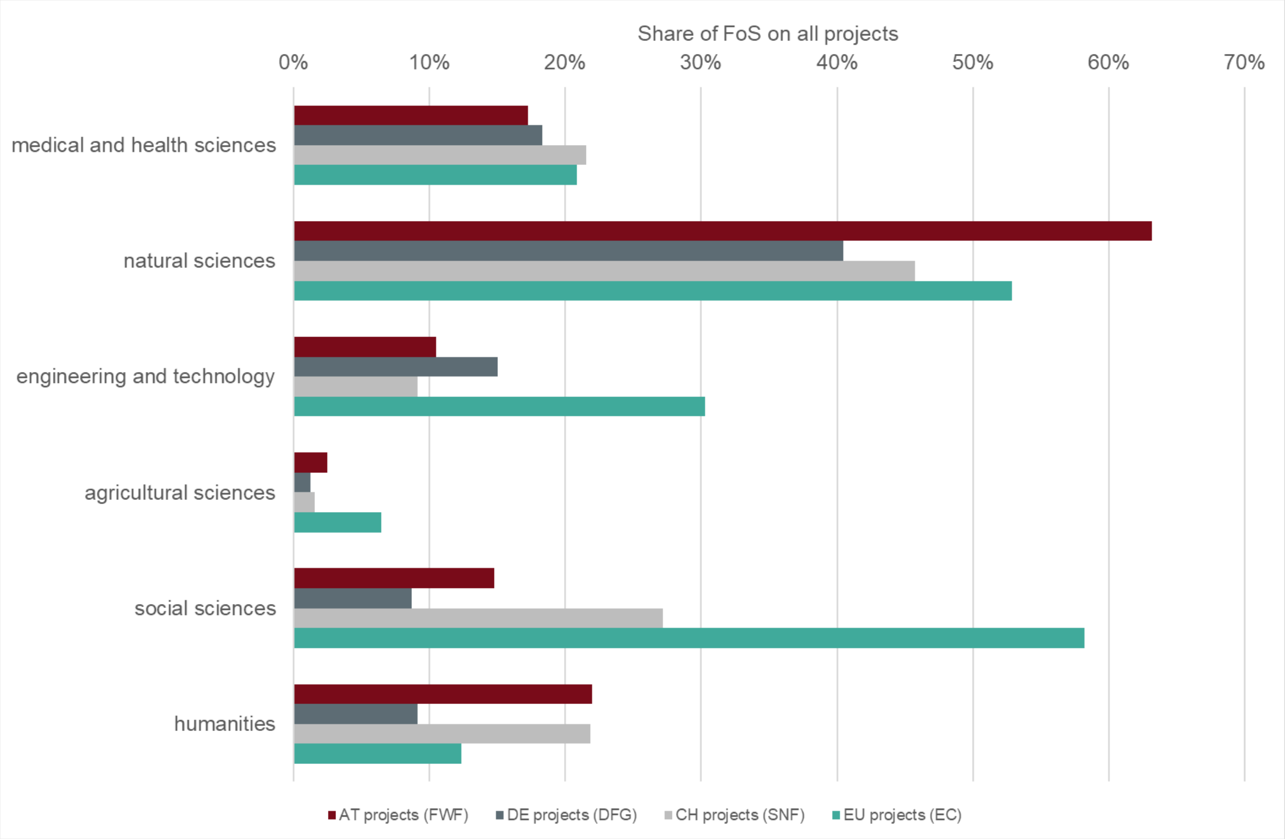
- Cooperative excellence research between different organizations forming a network and for the establishment of long-term research units keep on representing an asset for the agency

Performance side: Projects and Participations



- For all top 3 universities DFG funding (in terms of number of participations) is more important than EU funding.
- Complemented with EU funding (in particular ERC grants) for TU Munich and LMU Munich (but not University of Hannover)
- Public research organisations show higher share of EU funding
- Applied research organizations (Fraunhofer) participate in more innovation-oriented programs (Industrial Leadership, SGC); ERC grants more important for basis research organzisation (Max Planck Society)

Thematic orientation of funding agencies



- Thematic orientation of funders can be analyzed using Fields of Sciences (FoS)
- Projects are assigned to one or multiple FoS
- FoS systems enables very detailed analysis (up to 6 levels)
- Not all national funders provide FoS (in particular innovation agencies and national information systems) but national thematic classifications
- Analysis can be combined with actor level or regional level

Discussion

- As regard to the DFG example, **data on aim and orientation of the funding instruments were combined with data on participation level of different types of performers**. The results presented here can be used as a starting point for further research into the relationship between research policy design and the effects of research policies to the beneficiaries.

- EFIL and NATPRO integration allows for new possibilities for addressing research and policy questions, in particular opens for even more opportunities for relevant policy studies.

- The **combined use of the two datasets** may help in shedding light on the characteristics of national R&D funding systems, as well as national R&D projects and participations. The link between programs and projects allows for interpreting on how government goals for science policy are translated into concrete research activities by project beneficiaries.

- The joint information from the two datasets may reveal the mismatch between the policy orientation and research practices of scholars' communities.

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Consiglio Nazionale delle Ricerche

THANK YOU!

✉ fossr.dissemination@ircres.cnr.it

@fossrproject

fossr.eu

fossr-eu

@fossr

zenodo zenodo.org/communities/fossr

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Consiglio Nazionale delle Ricerche

Module III: Network models applied to RI data

Antonio Zinilli
Researcher at CNR-IRCrES

«Exploring research infrastructure data through advanced statistical applications»

FOSSR Online Training

26-27 September 2024

FOSSR

Fostering Open Science in Social Science Research
Innovative tools and services to investigate economic and societal change

Introducing Network Analysis: understanding Complex Systems

No a specific definition, but there is general consensus on the following observed properties:

1. large scale
2. evolving over time
3. power law degree distributions
4. small world properties

other properties depend on the kind of network being discussed

Properties of complex networks

1. Large scale: relative to order and size

- web graph: order > trillion
  - some sense infinite: number of strings entered into Google
- Facebook: > 1 billion nodes; Twitter: > 500 million nodes
  - much denser (higher average degree) than the web graph
- protein interaction networks: order in thousands

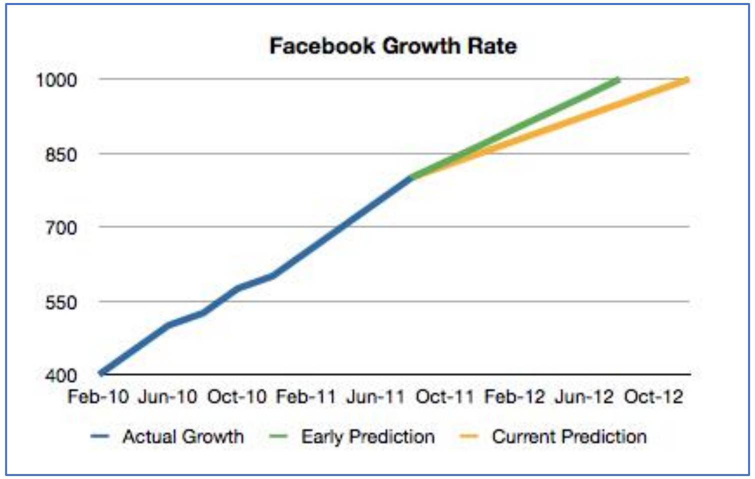
2. Evolving: networks change over time

- web graph: billions of nodes and links appear and disappear each day

- Facebook: grew to 1 billion users
  - denser than the web graph

- protein interaction networks:

  order in the thousands
  - evolves much more slowly



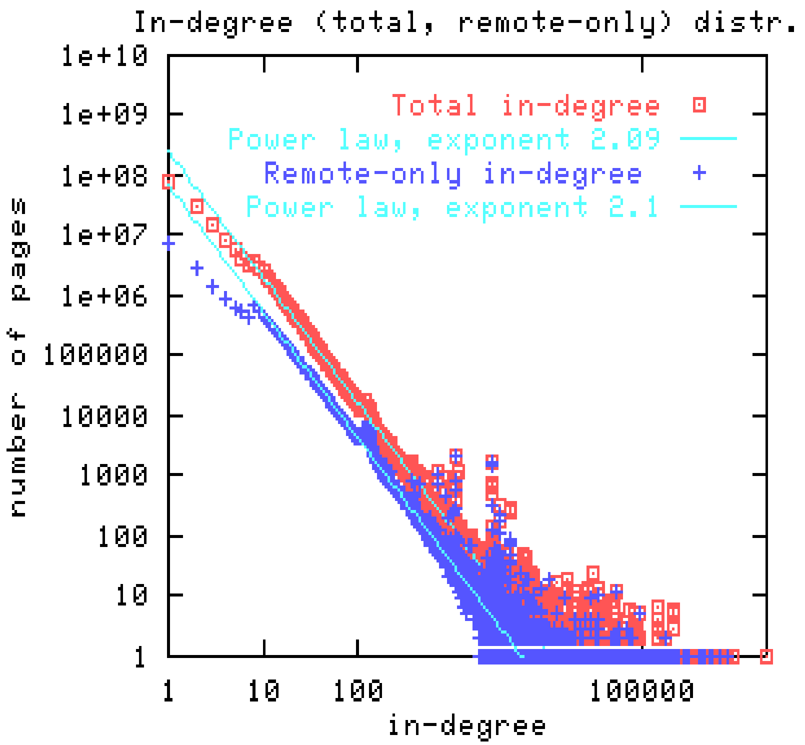
3. Power law degree distribution

- Many of networks in economic, physical, technological and social systems have
- been found to have a power-law degree distribution. That is, the number of
- vertices N(m) with m edges is given by

$$N(m) \approx m^{-\alpha}$$

α is called the exponent of the power law

Power law degree distribution in the web graph:



(Broder et al, 2001)

reported an exponent α = 2.1 for the in-degree distribution (in a 200 million nodes)

4. Small world property

Small World networks introduced by social scientists Watts & Strogatz in 1998

- low distances

diam(G) = O(log n) (Diameter is the shortest distance between the two most distant nodes)

- higher clustering coefficient than random graph with same expected degree

The Strength of Weak Ties

Mark S. Granovetter (The American Journal of Sociology, 1973) interviewed people and asked:

“How did you find your job?”

Kept getting the same answer: “through an acquaintance, not a friend”

Main Idea

Links can have a wide range of possible strengths, but for conceptual simplicity we’ll categorize all links in the social network as belonging to one of two types:

- **stronger links,** correspond to friend
- **weaker links,** correspond to acquaintances

The Strength of Weak Ties

- Many weak ties → more access to wider community's ideas, resources, etc.

- Our weak ties are with people whose ties are with those socially distant to us. Weak ties bring us knowledge of our community not available through friends

The Strength of Weak Ties

- Do leads for new jobs come through strong or weak contacts?
  - Strong: More motivation to help you, since they know you better
  - Weak: Likely less overlap with leads you can easily get elsewhere

- Study by author shows that weak wins
  - Most job referrals come through those who we see rarely: old school friends, former co-workers, etc.

The Strength of Weak Ties

- Removal of weak ties raises path lengths more than removal of strong ties
- Assume: probability of info passing successfully between two node is proportional to the number of paths connecting the two nodes
- Conclusion: removal of a weak edge damages the connectivity more than the removal of a strong edge

Dependencies are important

We know that for many contexts observations depend on other observations.

A lot of individual behaviour depend on other behaviour:

- Innovation (spillover);

- Smoking behavior;

- ...

- ...

Social Networks

- **Social**

Friendship, romantic relationships

- **Government**

Political alliances, government agencies

- **Markets**

Trade: flow of goods, supply chains

Labor markets: getting jobs

- **Organizations and teams**

Interlocking Employees

Email exchange

Other Networks

Internet

Transportation: Airline networks

Metabolic networks

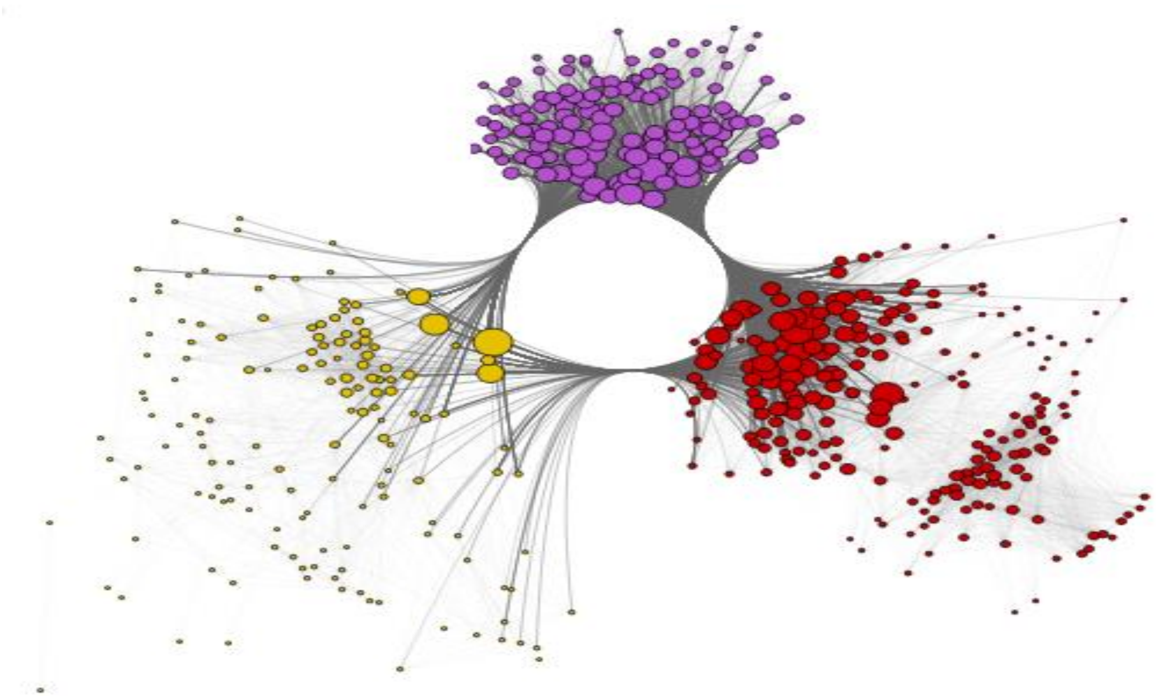
Neural networks

Protein interaction

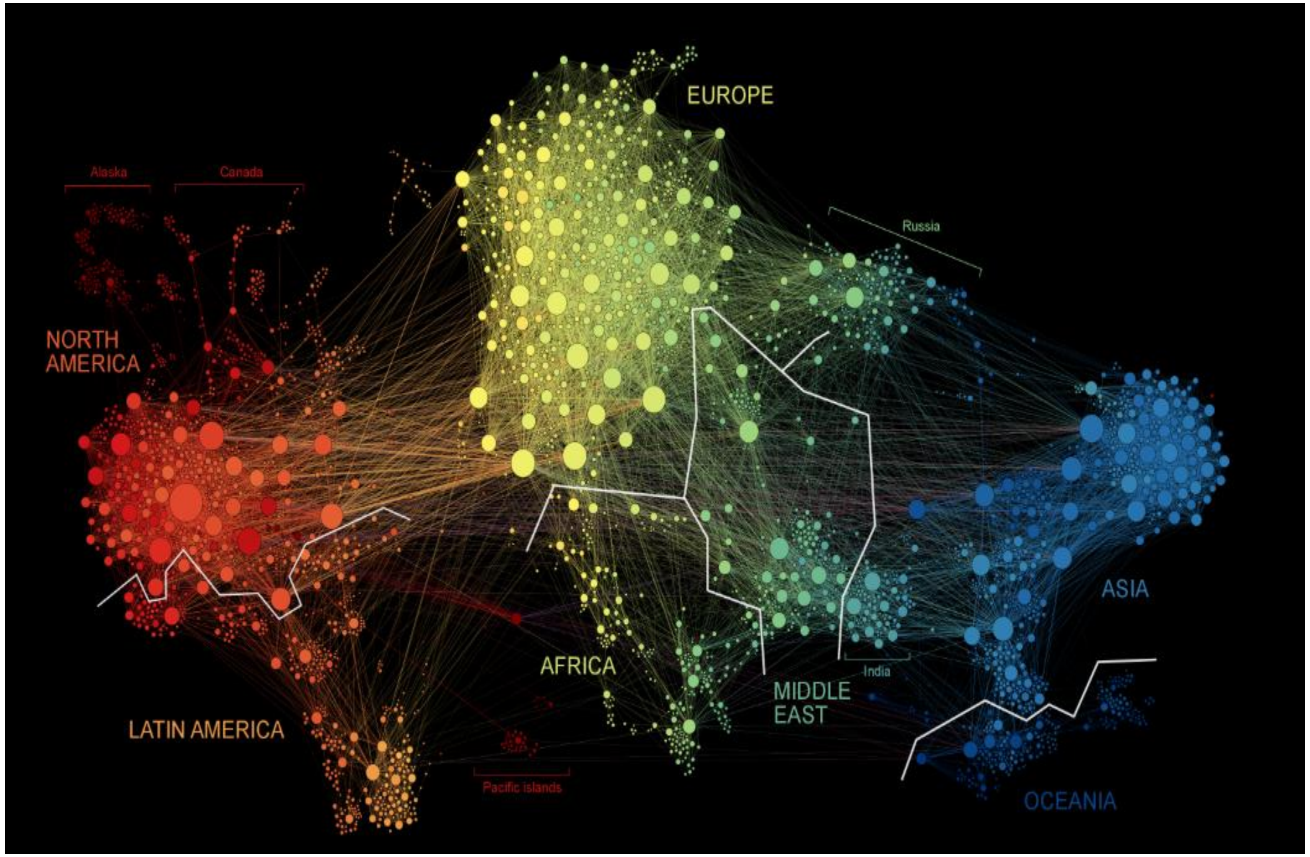
- ...

- ...

Communities in Facebook Friend Network



Air Transportation Cluster



What Are Networks?

Networks are patterns of relationships that connect individuals, institutions, or objects (or leave them disconnected).

EXAMPLES

- Individuals’ co-memberships in organizations
- Relationship between countries, regions, cities

When to Study Networks?

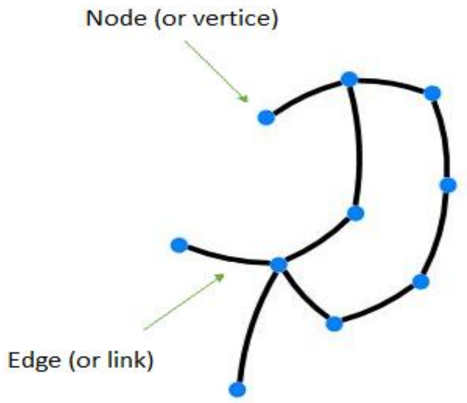
It is possible applied network analysis techniques in more contexts. But the question we want to ask is: when in the network aspect of phenomenon particularly pertinent to the social dynamics that matter to us?

Network analysis tends to place a strong emphasis on the *relationship* (or “the dyad”) as a unit of analysis.

Graphs

Social networks can be represented as graphs

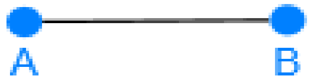
Graphs are made up of **nodes** (i.e., actors, cities, organizations, articles etc.) that are connected by **links** (i.e., relationships, membership, citations etc.).



Types of Links

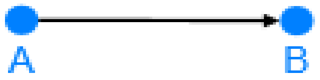
Undirected vs. directed links

- **Undirected links,** identified with a simple straight line**,** are used when there is a symmetric relationship:



E. g. Facebook has always used a symmetric model, if you add someone as a friend they have to add you as a friend as well.

- **Directed links**, identified with arrows, are used when there is an asymmetry in a relationship:



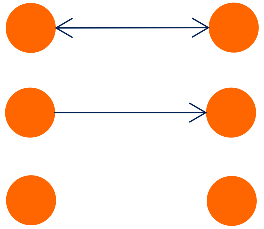
E. g. On Twitter you can “follow” someone else without them following you back.

Dyad for directed links

- A dyad is a pair of actors (i, j) in the network, plus the configuration of the tie variables (y_{ij}, y_{ji}) between them.

- Dyads can be of three types:

- *mutual*
- *asymmetric*
- *null*

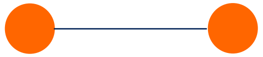


Dyad for undirected links

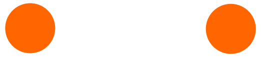
- A dyad is a pair of actors (*i, j*) in the network, plus the configuration of the tie variables (*yij, yji*) between them.

- Dyads can be of two types:

- *Mutual*



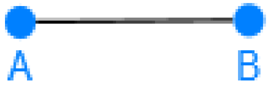
- *null*



Dichotomous vs. Valued Links

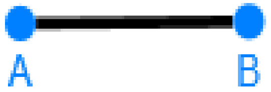
Dichotomous vs Valued Links

Dichotomous Links: either a link exists or it doesn’t
(e.g. if we are friends or we are not friends, there is a collaboration or not etc..)



Valued Links: the links vary on the based of a weight (strength)

(e.g. our friendship may be strong or weak, the number of times that each pair of countries cooperate)



Main parts of a graph

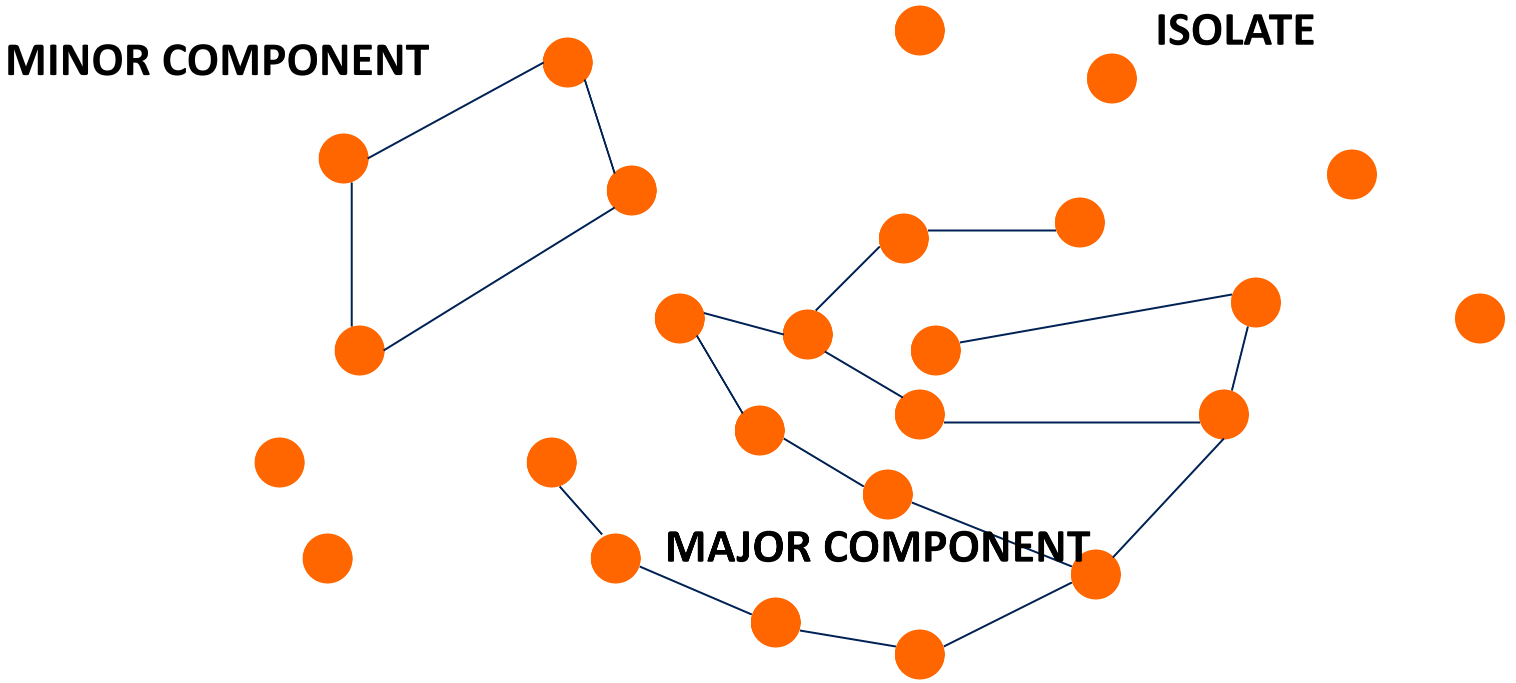
- **Component**: all nodes that assemble a connected subgraph within a network:

  main component is the largest component within a network;

  minor component is a component that is smaller than the main component. Usually there are more minor components.

- **Isolate:** a node that has no links to the other nodes within the network

Main parts of a graph



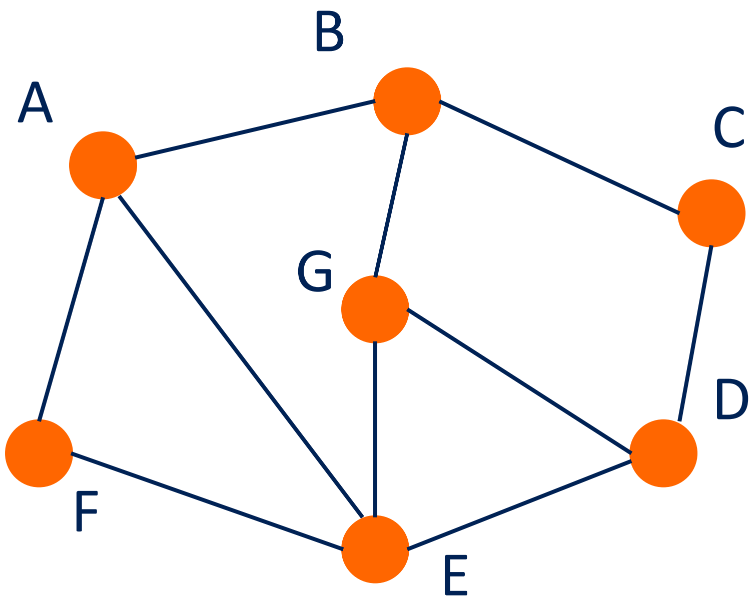
Main Graph Implementation Strategies

- Edge List

- Adjacency Matrix

Edge List

A	A	A	B	B	C	E	E	E	G
B	E	F	G	C	D	F	G	D	D



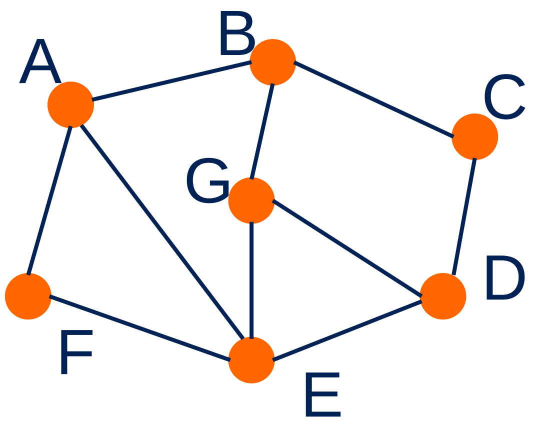
Matrices

The most basic matrix is an adjacency matrix: an *nxn* matrix where:

- the nondiagonal entry aij is the number of edges joining vertex *i* and vertex *j* (or the weight of the edge joining vertex *i* and vertex *j*). 1 indicates the presence of a link, while a 0 indicates the absence of a link

- the diagonal entry akk corresponds to the number of loops (self-connecting edges) at vertex *k*. Usually loops are not counted

$$
\begin{array}{cc}
 & \begin{array}{ccccccc} \mathrm{A} & \mathrm{B} & \mathrm{C} & \mathrm{D} & \mathrm{E} & \mathrm{F} & \mathrm{G} \end{array} \\
M = \begin{array}{c} A \\ B \\ C \\ D \\ E \\ F \\ G \end{array} & \left\{ \begin{array}{ccccccc}
0 & 1 & 0 & 0 & 1 & 1 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 1 & 1 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 1 & 1 & 0 & 0
\end{array} \right\}
\end{array}
$$



Matrices

	Austria	France	Germany	Italy
Austria	0	0	1	0
France	0	0	0	1
Germany	1	0	0	0
Italy	0	1	0	0

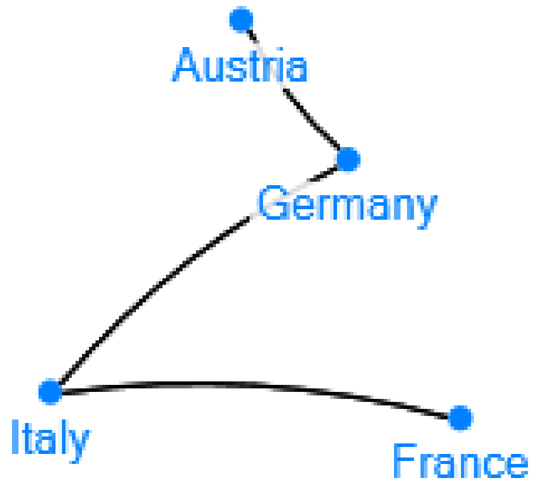
1 indicates the presence of a link, while 0 indicates the absence of a link

	Austria	France	Germany	Italy
Austria	0			
France	0	0		
Germany	1	0	0	
Italy	0	1	0	0

If matrices are symmetric, they may be represented by upper or lower triangle only

One-Mode Network

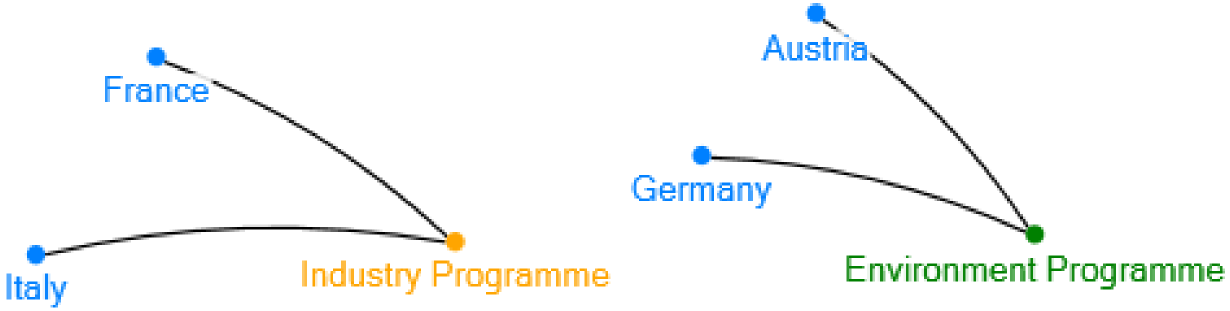
- Network analysis typically involves only one mode. A mode is a class of nodes in a network.



Two-Mode Networks

Mode 1	Mode 2
People	Events
Students	Universities
Countries	Projects or Programs

Example: Two-mode data have countries and research programmes. France and Italy collaborate at the same programme.



Advantages vs. Disadvantages

- **Advantages of Going from 2-mode to 1-mode**
- Reduce the dimension of the data
- Make it easier to visualize
- Focus on what really matters

- **Disadvantages of going from 2-mode to 1-mode**
- Lose information
- Confuse the reader
- Removing the important relationships

- **Depends entirely on your case**

Converting Data From One Mode to Two Modes

This two-mode network

	Math	Physics	Politics	English
Fabio	1	1	0	0
Antonio	0	0	1	1
Claudia	0	1	1	0
Valentina	1	1	0	1

Can be reduced to this one-mode matrix

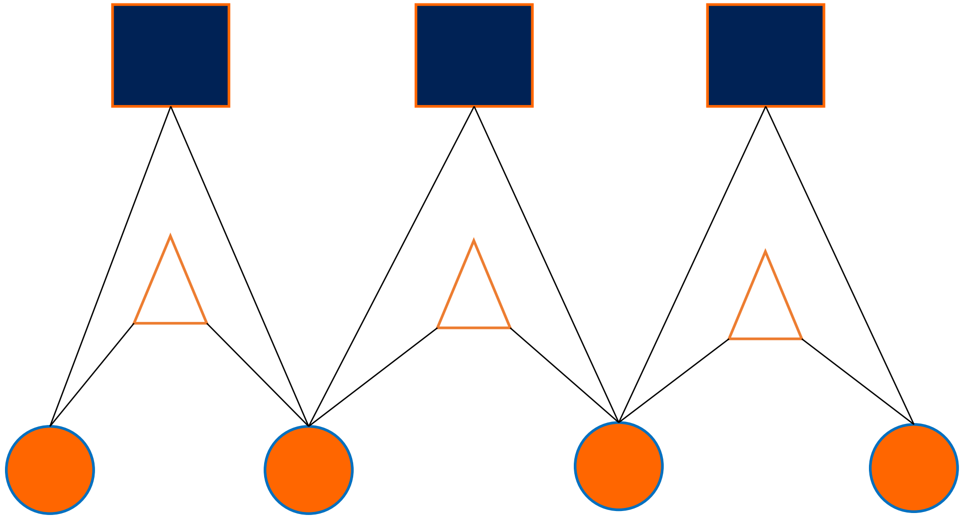
	Fabio	Antonio	Claudia	Valentina
Fabio	2			
Antonio	0	2		
Claudia	1	1	2	
Valentina	2	1	1	3

Or this one

	Math	Physics	Politics	English
Math	2			
Physics	2	3		
Politics	0	1	2	
English	1	1	1	2

More than Two Modes

It is possible for network data to have more than two modes.



Example

Mode 1: People

Mode 2: Organizations

Mode 3: Ideologies

The Limits of Multi-Modal Analysis

Almost all network analysis can be conducted using when one-mode data is on hand.

In many network software programs two-mode measures (e.g., centrality) can be easily generated. But progress in this area is still moving forward.

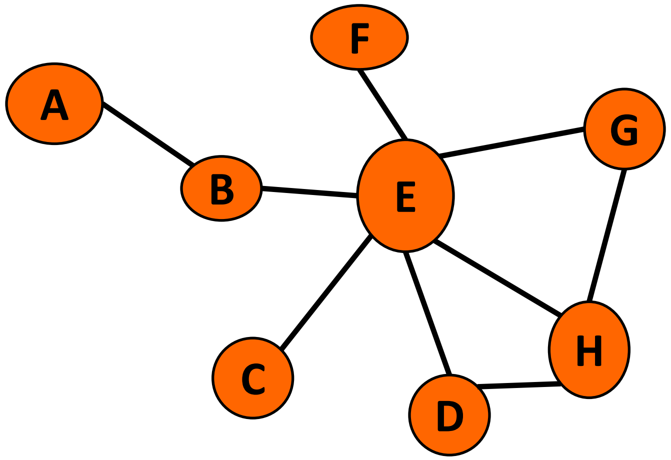
Extant models of three-mode data is generally are confined to lattices and other relatively complex mathematical forms.

Higher-order modes are conceivable, but work needs to be done to make their analysis practical for social scientists.

Basic Network Statistics

- Path
- Geodesic distance
- Density
- Degree centrality
- Closeness centrality
- Betweenness centrality
- Clustering Coefficient
- Eigenvector centrality

Path



- ABEDHG is a path from A to G. There are multiple paths from A to G.

- Path length is the number of steps in a path. The path length of ABEDHG is equal to 5.

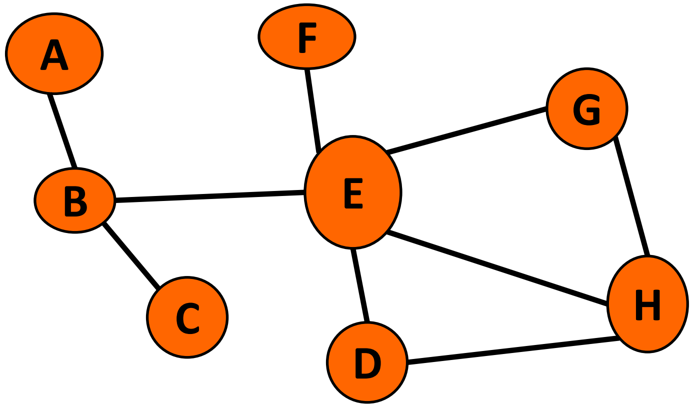
Geodesic distance

- **Geodesic distance** is the shortest path from one node to another node.

- The **shortest path** is the path that achieves that distance.

- The average network **diameter** is the average of shortest path lengths over all pairs of nodes in a network.

Geodesic distance: an example



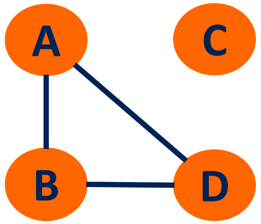
ABEG is the geodesic from A to G

Density

- Density is a property of a network.
- The proportion of links in a network relative to the total number possible.

Density: an example

In this example we have 3 links and 4 vertices



Network Density:

$$\frac{\textit{Actual connections}}{\textit{Potential connections}}$$

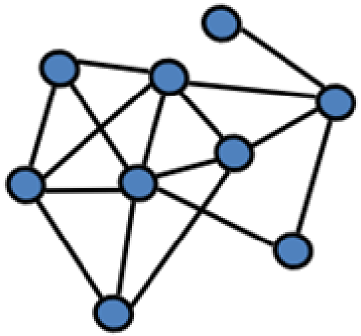
Potential connections:

$$PC = \frac{n(n-1)}{2}$$

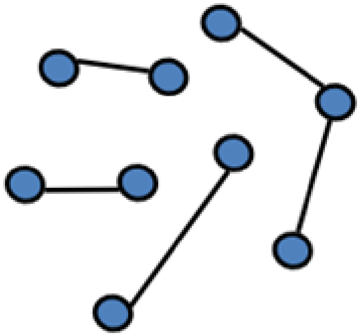
- Potential connections = [(4 (4-1))/2] = 12/2 = 6
- Density: 3/6=0,5
- This graph has one half of all possible links.

Density

- Proportion of ties in a graph



High density (44%)



Low density (14%)

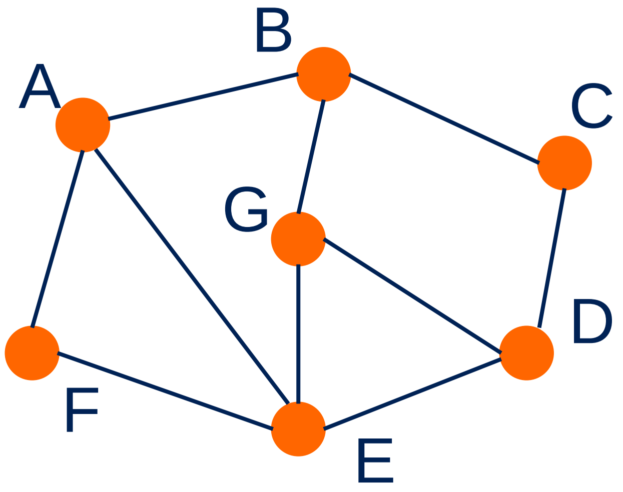
Centrality

Well connected actors are in a structurally advantageous position.

- *Getting jobs*
- *Better informed*
- *Higher status*

Centrality

Degree is a property of a node. The degree of a node is equal to the number of links that it has.



B has a degree of 3.

The average degree of a graph is given by

$$\bar{k} = \frac{1}{g}\sum_{i=1}^{g} k(n_i)$$

Degree distribution

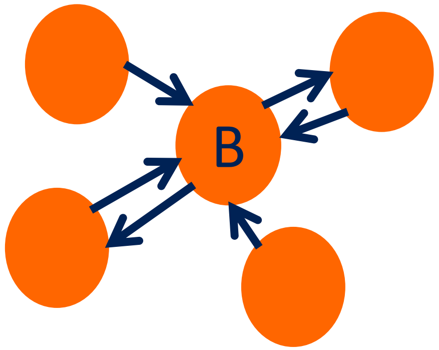
- A degree distribution a property of a network.
- A degree distribution is the number of nodes of a network that have each degree level.
- A degree distribution may be a good way of summarizing the activity of nodes in a network. This measure provides a first indication on the importance of a node; important nodes are those that have a greater influence to the flow of information in a network.
- May be a good way of comparing networks to one another.

Indegree and Outdegree

- **Directed networks only**

- **Indegree**: The number of links that a node **receives** from other nodes

- **Outdegree:** The number of links that a node **sends** to other nodes



- What is B’s indegree? Answer: 4

- What is B’s outdegree? Answer: 2

Closeness centrality

Closeness is based on the length of the average shortest path between a vertex and all vertices in the graph

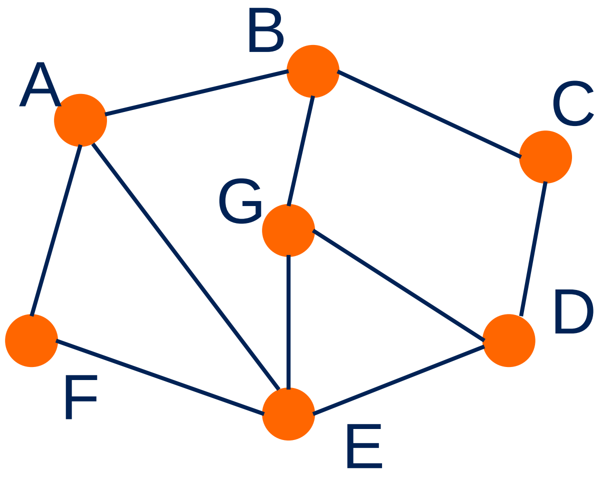
Closeness Centrality:

$$C_c(i) = \left[\sum_{j=1}^{N} d(i, j) \right]^{-1}$$

Normalized Closeness Centrality

$$C_C^{'}(i) = (C_C(i)) / (N - 1)$$

Closeness centrality: an example



$$C_c^{'}(A) = \left[\frac{\sum_{j=1}^{N} d(A, j)}{N-1} \right]^{-1} = \left[\frac{\mathrm{d(AB)+d(AC)+d(AD)+d(AE)+d(AF)+d(AG)}}{N-1} \right]^{-1} = \left[\frac{1+2+2+1+1+2}{6} \right]^{-1} = \left[\frac{9}{6} \right]^{-1} = 0.66$$

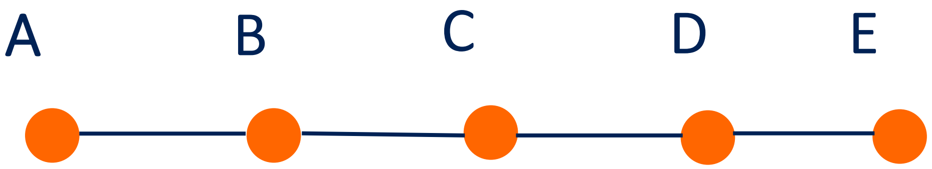
Betweenness centrality

Betweenness centrality indicates the extent to which a vertex lies on paths between other vertices.

$$C_B(i) = \sum_{j<k} \overbrace{g_{jk}(i)}^{\text{paths between j and k that pass through i}} / \underbrace{g_{jk}}_{\text{all paths between j and k}}$$

Betweenness centrality: an example

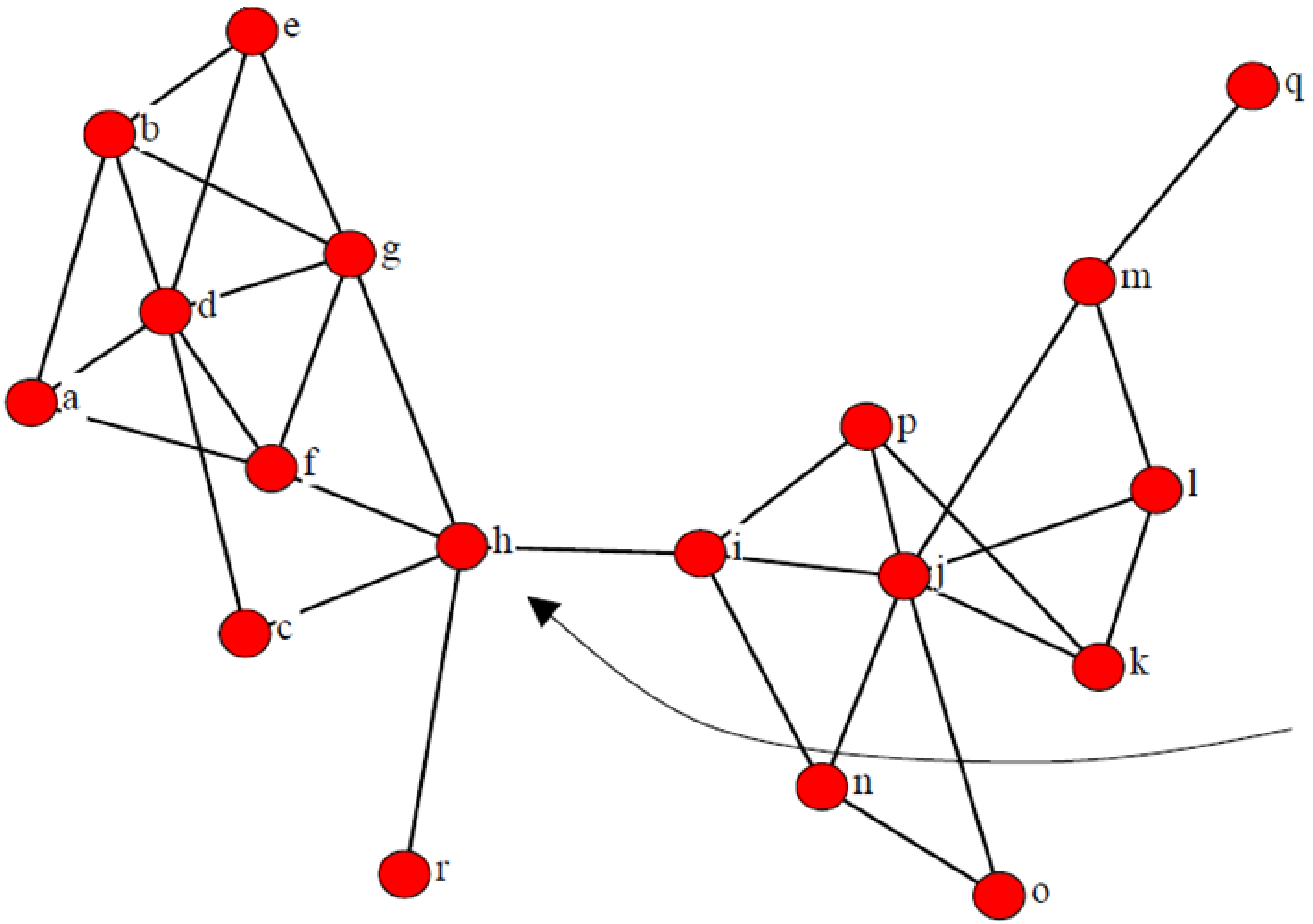
- In this simple example there are no alternate paths.



- A lies between no two other vertices
- B lies between A and 3 other vertices: C, D, and E - (AC), (AD), (AE)
- C lies between 4 pairs of vertices: A, B, D, E - (A,D),(A,E),(B,D),(B,E)
- D lies between E and 3 other vertices: A, B, and C - (AE), (BE), (CE)
- E lies between no two other vertices

- We can conclude that C gets full credit.

Betweenness centrality : an example



Highest Betweenness Centrality

Eigenvector centrality

Eigenvector: is a measure of centrality that takes into account the centrality of other nodes to which a node is connected.

is an adjacency matrix with n nodes

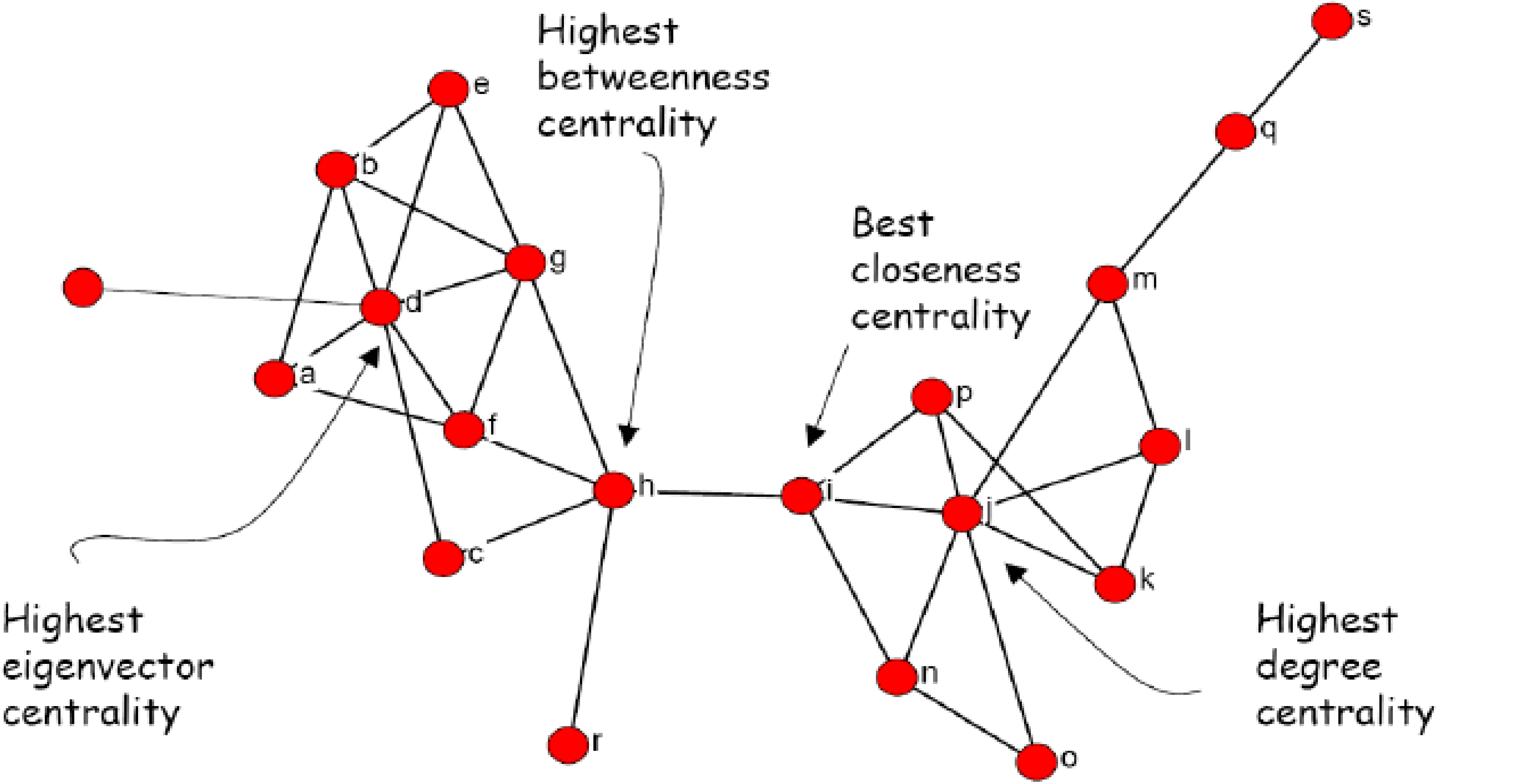
$$E_i = \frac{1}{\lambda}\sum_{j=1}^{N} a_{ij} x_j$$

i is a node that is distinct from j

is the realized value of link in the network

(Constant)-Eigenvector solved through an interactive algorithm

Main centrality measures


Highest betweenness centrality
Best closeness centrality
Highest eigenvector centrality
Highest degree centrality
s
q
m
e
b
g
d
a
f
h
i
p
l
j
k
c
n
r
o

Clustering coefficient

Clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together.

We can speak about:

- Local Clustering coefficient
- Global Clustering coefficient

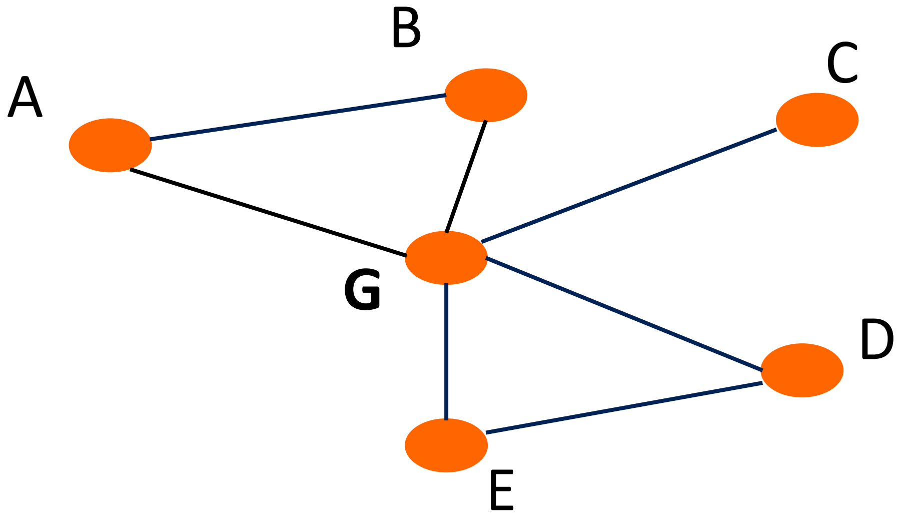
Local Clustering coefficient

The local clustering coefficient is defined as the ratio of the observed connections between all neighbours, and all possible connections between the neighbours

$$C_i = \frac{\dfrac{b_i}{n!}}{k!(n-k)!}$$

with bi indicates the number of observed links between all neighbours of a specific node; the denominator is the binomial coefficient and indicates all possible connections between the neighbours.

Example



$$C_i = \frac{2}{\frac{n!}{k!(n-k)!}} = \frac{2}{\frac{5!}{2!(5-2)!}} = \frac{2}{\frac{5!}{2!(5-2)!}} = \frac{2}{\frac{120}{12}} = 0,2$$

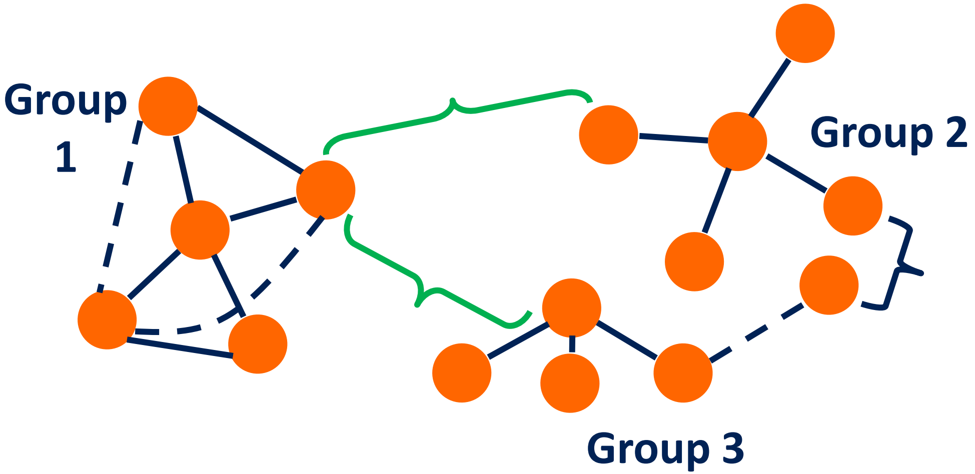
Global Clustering coefficient

The global clustering of a graph is given by the average of all local clustering coefficients

$$\bar{C} = \frac{1}{g} \sum_{i=1}^{g} C_i$$

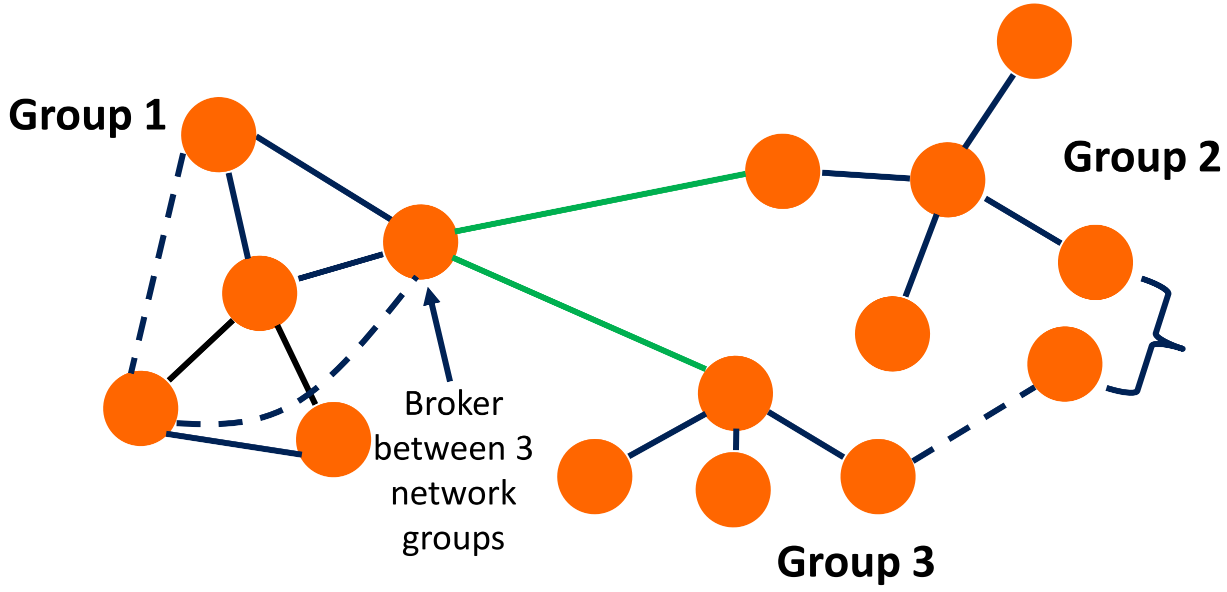
Structural Holes

- Structural holes describe the situation where there are gaps, or holes, between different clusters of network partners.
- People usually focus on activities inside their own network group, which creates "holes" in the information flow between groups, called, structural holes (Burt & Ronchi, 2005).
- Structural holes are important because they present areas where diverse information relative to the focal actor may reside.
- A lack of structural holes in one's network → redundant information flow, and the potential to miss important information relative to the industry, market, technology, etc.



Bridging Ties

Bridging ties describe the situation where an actor is tied to another actor who has no other links with that cluster.



Benefits of Bridging Ties

1. Actors holding bridging positions are *more likely to receive* novel information vs. the rest of the network
2. Bridging actors *more likely to receive* new information *earlier* than others in the network
3. This leads to more power and control benefits for the actors holding the bridge position (e.g., the brokering position)

Drawbacks of Bridging Ties

1. If you do not hold the bridging position, you may be in a weak position

Other Measures of centrality

There are a large number of other possible measures of centrality.

K-star, Transitivity etc.

Usually, these different measures are highly correlated

More information

Usually, measures of centrality are used as independent variable.

Usually, the network ties are used as dependent variable.

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Consiglio Nazionale delle Ricerche

CASE STUDY

Disentangling the relationship between collaboration and research productivity: direct and indirect effects

Aim of the work

This work aims to study the relationship between university's collaborations and university's research productivity.

- Focusing on two key forms of knowledge exchange networks: those stemming from EU-funded projects and scholarly publications

- Within this analytical framework, we refine our examination to include two specific domains within the European Research Council (ERC) sphere:

  Physical Sciences and Engineering (PE) and Life Sciences (LS)

Literature

The relationship between collaboration and research productivity has been a central focus of academic research in recent years (Landry et al., 1996; Lee and Bozeman, 2005; Ductor, 2015; Yadav et al., 2023)

However, the directionality and nature of this relationship remain subjects of ongoing debate (Abramo et al., 2017)

On the one hand collaboration on research projects boosts the visibility of participating institutions and increases their chances of publishing in prestigious, high-impact journals (Lee Bozeman, 2005; Barjak Robinson, 2008; Abramo et al., 2009; Abramo et al., 2017)

On the other hand highly productive researchers are more likely to attract collaborative opportunities, creating a cumulative advantage (Abramo et al., 2011; Zinilli, 2016)

Research questions

Q1: how do different types of network centrality (e.g., Degree Centrality UniversityUniversity and University-Firm, Closeness Centrality, Betweenness Centrality) influence research productivity?

Q2: is there evidence of a bidirectional relationship between collaboration and research productivity?

Q3: how do these relationships vary across different scientific domains, specifically between Life Sciences and Physics and Engineering?

Data

Data were collected from three databases hosted within the RISIS infrastructure and FOSSR over three years (2016-2018)

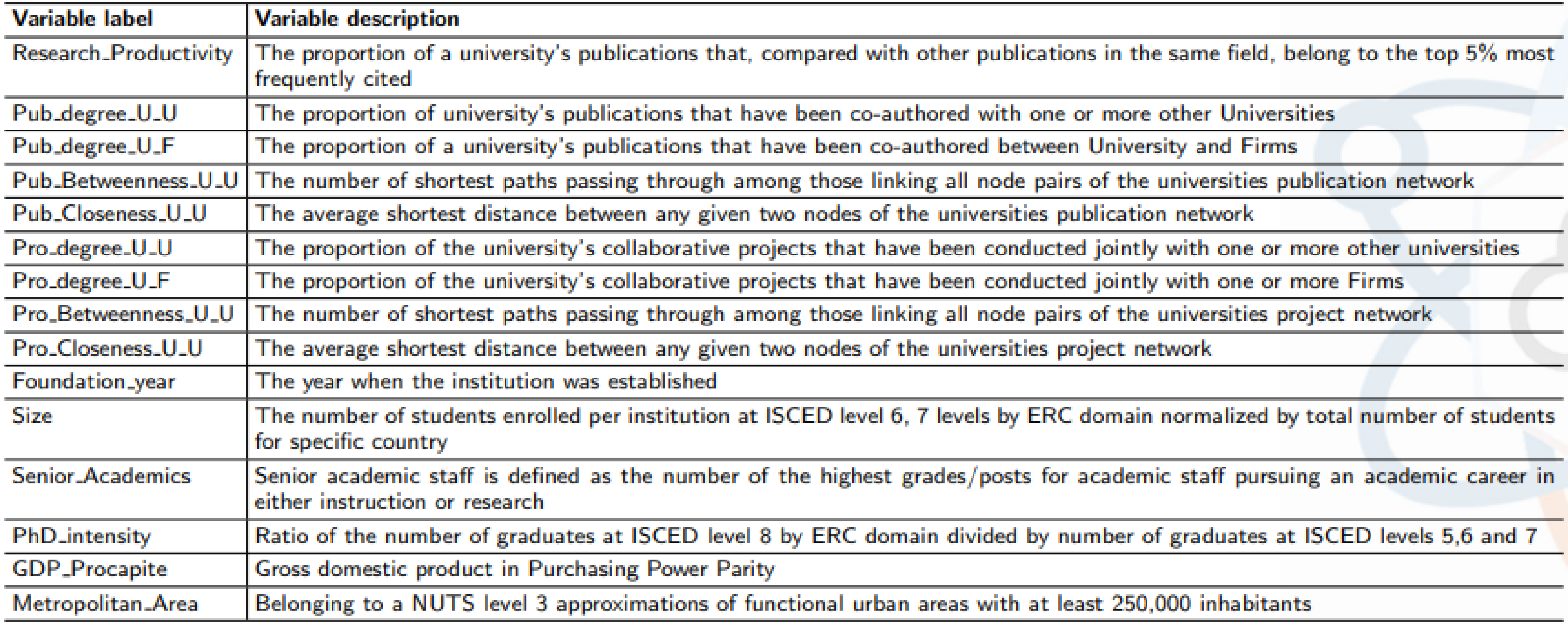
1. EUPRO containing information on European funded programs
2. CWTS containing information on publications
3. ETER is a database providing information on Higher Education Institutions (HEIs).

The collaboration variables are: university-university (U-U) and university-firm (U-F) degree, closeness and betweenness centralities for both project and publication networks

The research productivity variable is based on CWTS citation indicators and is composed of three levels:

- ✓ low productivity
- ✓ medium productivity
- ✓ high productivity

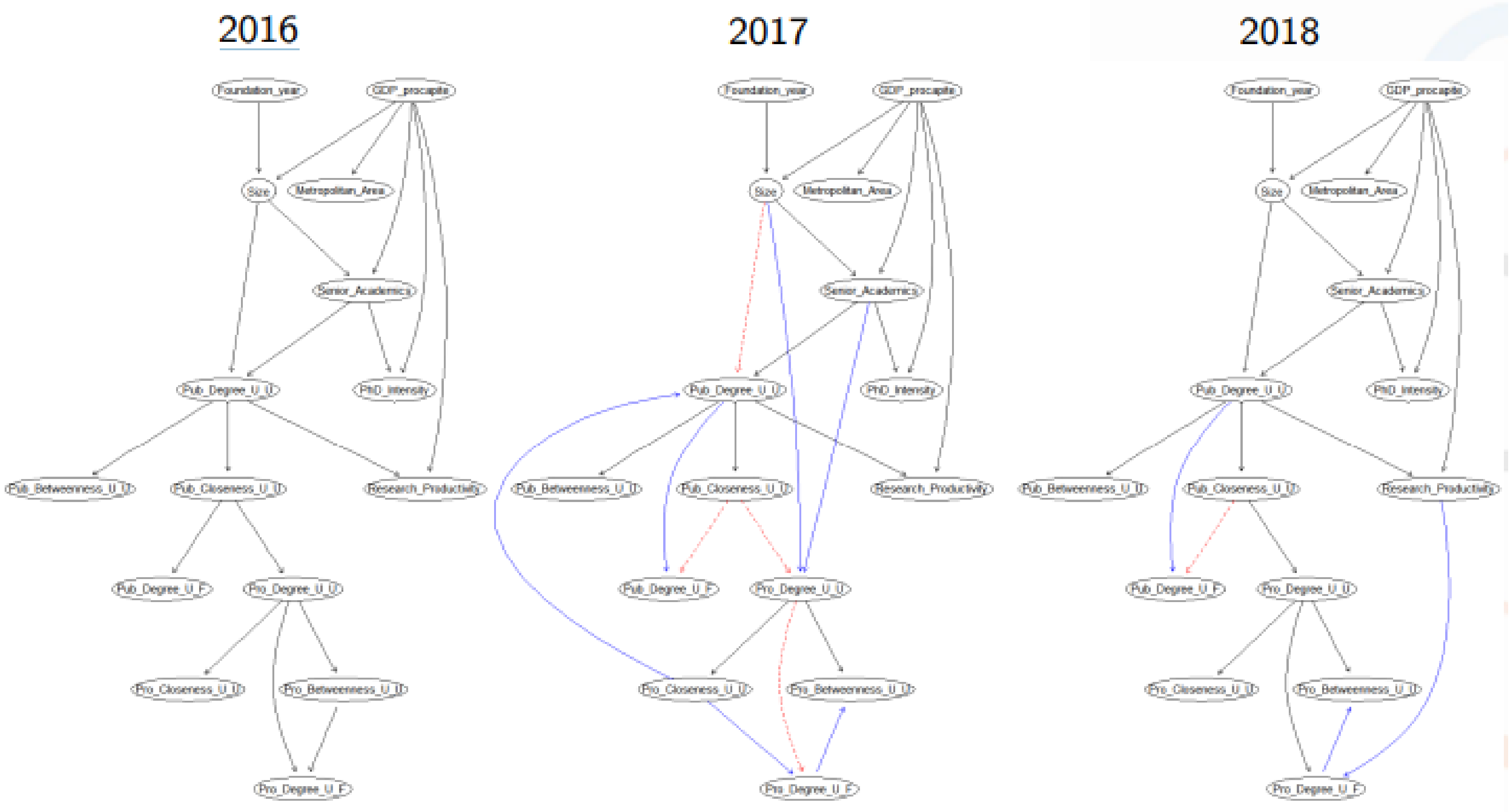
Variable Description



Variable label	Variable description
Research_Productivity	The proportion of a university's publications that, compared with other publications in the same field, belong to the top 5% most frequently cited
Pub_degree_U_U	The proportion of university's publications that have been co-authored with one or more other Universities
Pub_degree_U_F	The proportion of a university's publications that have been co-authored between University and Firms
Pub_Betweenness_U_U	The number of shortest paths passing through among those linking all node pairs of the universities publication network
Pub_Closeness_U_U	The average shortest distance between any given two nodes of the universities publication network
Pro_degree_U_U	The proportion of the university's collaborative projects that have been conducted jointly with one or more other universities
Pro_degree_U_F	The proportion of the university's collaborative projects that have been conducted jointly with one or more Firms
Pro_Betweenness_U_U	The number of shortest paths passing through among those linking all node pairs of the universities project network
Pro_Closeness_U_U	The average shortest distance between any given two nodes of the universities project network
Foundation_year	The year when the institution was established
Size	The number of students enrolled per institution at ISCED level 6, 7 levels by ERC domain normalized by total number of students for specific country
Senior_Academics	Senior academic staff is defined as the number of the highest grades/posts for academic staff pursuing an academic career in either instruction or research
PhD_intensity	Ratio of the number of graduates at ISCED level 8 by ERC domain divided by number of graduates at ISCED levels 5,6 and 7
GDP_Procapite	Gross domestic product in Purchasing Power Parity
Metropolitan_Area	Belonging to a NUTS level 3 approximations of functional urban areas with at least 250,000 inhabitants

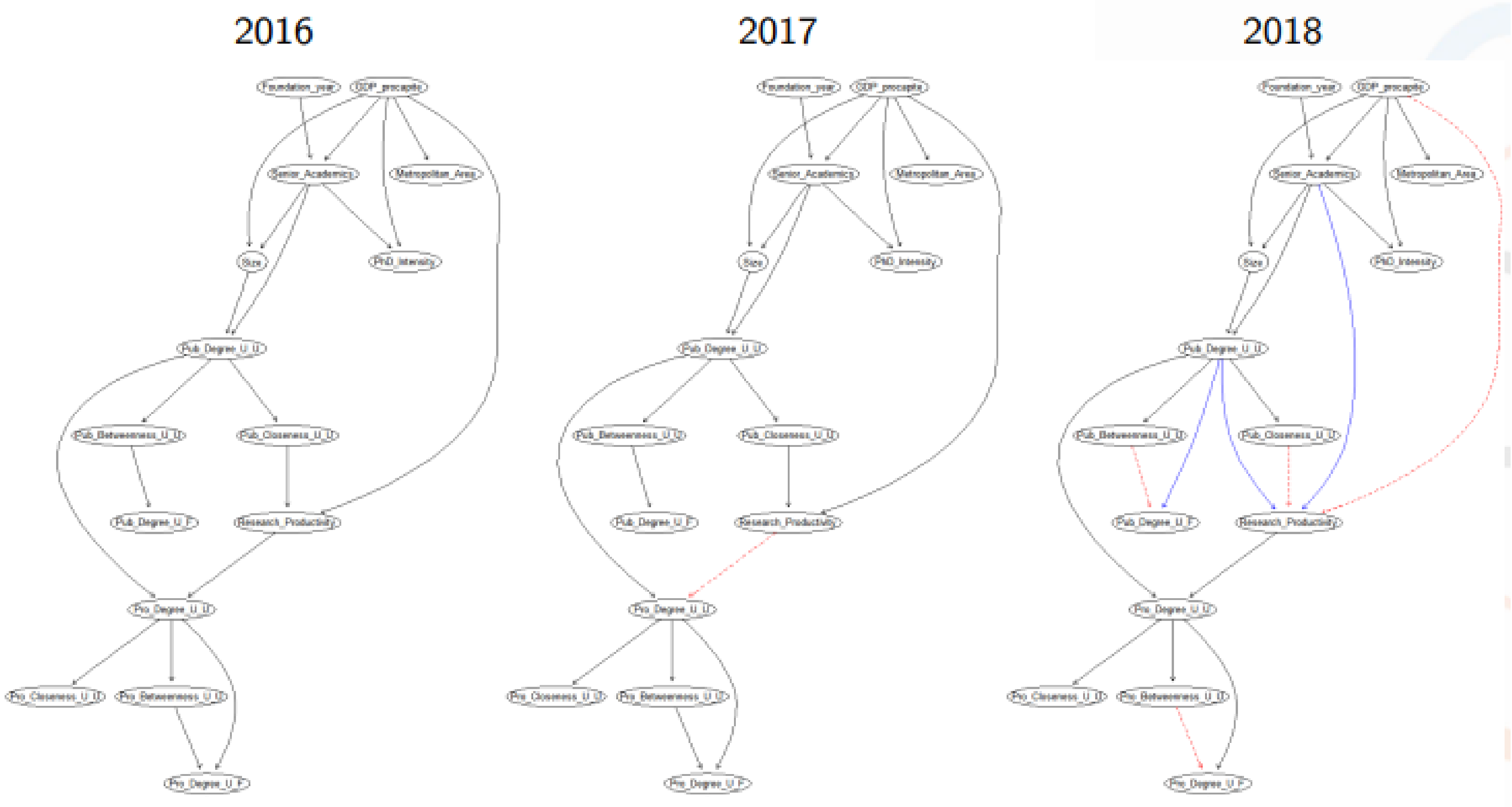
Obtained PE graphs

- Degree centrality is driving the other network indices
- In 2016 and 2017 publication collaborations drive project collaborations
- Publication degree centrality U-U affects research productivity
- In 2018 research productivity affects in turn firm project degree centrality



Obtained LS graphs

- As in PE, there is a hierarchy in the collaboration variable structure
- Research productivity is affected by publication closeness and degree between universities
- Research productivity affects project degree between universities



Discussion

Q1: How do different types of network centrality influence research productivity?
A1: Network centrality measures appear to be positively correlated with research productivity, emphasizing the significance of direct and indirect inter-institutional connections

Q2: Is there evidence of a bidirectional relationship between collaboration and research productivity?
A2: The models reveal that collaboration in publication affects research productivity which then affects collaboration in projects.

Q3: How do these relationships vary across different scientific domains?
A3: In both domains collaboration present similar hierarchical structures and positively affects research productivity. However research productivity in turn positively affects project collaboration with firms in PE and project collaboration with other universities in LS.

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Consiglio Nazionale delle Ricerche

THANK YOU!

✉ fossr.dissemination@ircres.cnr.it

@fossrproject

fossr.eu

fossr-eu

@fossr

zenodo zenodo.org/communities/fossr

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Consiglio Nazionale delle Ricerche

Module IV: Causal Bayesian networks and applications to RI data

Lorenzo Giammei
Researcher at CNR-IRCrES

≪Exploring research infrastructure data through advanced statistical applications≫

FOSSR Online Training

26-27 September 2024

FOSSR
Fostering Open Science in Social Science Research
Innovative tools and services to investigate economic and societal change

Outline

Causal inference and Potential Outcomes

Causal Bayesian networks: Fundamentals

Causal Bayesian networks: Structural learning

- Application I: Collaboration and research productivity

Causal Bayesian networks: Causal effect estimation

- Application II: Home based working and expected revenues

Outline

Causal inference and Potential Outcomes

Causal Bayesian networks: Fundamentals

Causal Bayesian networks: Structural learning

- Application I: Collaboration and research productivity

Causal Bayesian networks: Causal effect estimation

- Application II: Home based working and expected revenues

What is causal inference and why is it important?

- **Correlation does not imply causation**
- Investigating causality usually means assessing if and how a certain intervention, often called treatment, affects an outcome of interest
- Investigating relations of cause and effect motivates most of the research in social and biomedical sciences.
- Two main contexts:
  1. Randomized experiments
  2. Observational data
- The preeminent approaches to deal with causality are **Potential Outcomes** (PO) and **Causal Graphs** (CG).

Potential Outcomes

Rubin 1974; Imbens and Rubin, 2015

- Originates from the work of Neyman and Fisher on **randomized controlled trials**
- The name of the framework comes from its peculiar notation $Y_i(t)$ that denotes the **potential outcome** for unit i when receiving the treatment level $T = t$
- Potential Outcomes notation in case of a binary treatment: $Y_i(0)$ and $Y_i(1)$
- The causal effect of T on Y can therefore be computed by **comparing summary statistics** of the potential outcomes distribution
- The resulting causal estimate is usually called the **average treatment effect** (ATE) and can be expressed in different ways:
  1. $ATE = E[Y_i(1) - Y_i(0)]$
  2. $ATE = \frac{E[Y_i(1)]}{E[Y_i(0)]}$

PO: Fundamental problem of causal inference and RCT

- **Fundamental problem**: $Y_i(0)$ and $Y_i(1)$ cannot be observed for the same unit i
- Units receive only one level of treatment, creating a missing data problem
- **ATE can be estimated in randomized controlled trials (RCT)**
- In RCT treatment is assigned randomly to the units of the sample, thus rendering T independent of the potential outcomes: $T_i \perp\!\!\!\perp (Y_i(0), Y_i(1))$.
- If we add the assumption that there is no interference between units (SUTVA) an unbiased estimate of the ATE can be obtained by computing the difference

$$\bar{Y}_t - \bar{Y}_c, \quad \text{with} \quad \bar{Y}_t = \frac{1}{N_t} \sum_{i:T_i=1} Y_i \quad \text{and} \quad \bar{Y}_c = \frac{1}{N_c} \sum_{i:T_i=0} Y_i.$$

PO and observational data

- What can we do **if data are not experimental?**
- The PO framework also provides several solutions to deal with observational data
- PO methods that deal with observational data aim at **emulating an experimental context** under specific assumptions
- What generally prevents observational data from being treated as experimental data is the presence of **confounders**
- Confounders are variables that affect both the treatment and the outcome and can lead to biased causal estimates if not adequately accounted for
- The concern worsens when confounders are unobserved since, in this situation, treatment effects could be impossible to identify

PO: Unconfoundedness

- The assumptions that tackles directly the problem of confounders is called **unconfoundedness**:

$$T_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) | X_i$$

- Unconfoundedness states that the treatment T_i is independent of the potential outcomes, given a set of pre-treatment variables X_i
- The condition allows estimating the ATE as

$$ATE = E[Y_i(1) - Y_i(0)] = E[E[Y_i | T_i = 1, X_i] - E[Y_i | T_i = 0, X_i]]$$

- The formula is also called **adjusting for** X and as long as unconfoundedness holds, it ensures an unbiased estimation of the ATE in the presence of confounders
- Adjustment can be performed through various methods, including regression and matching

PO identification strategies and assumptions

- The unconfoundedness assumption **cannot be tested**
- This implies that justifying it becomes difficult if a priori knowledge is missing
- As the number of variables in the model increases, assessing the assumption validity turns out to be a **challenging task**
- Other common identification strategies include:
  1. Instrumental variable
  2. difference-in-differences
  3. regression discontinuity
  4. synthetic control
- These methods provide solutions to very specific causal problems and usually impose additional functional-forms restrictions on probability distributions, such as linearity, monotonicity or additivity

Outline

Causal inference and Potential Outcomes

Causal Bayesian networks: Fundamentals

Causal Bayesian networks: Structural learning

- Application I: Collaboration and research productivity

Causal Bayesian networks: Causal effect estimation

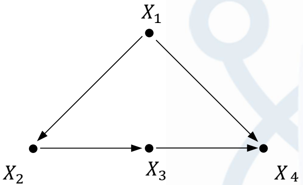
- Application II: Home based working and expected revenues

Causal Bayesian networks

Pearl, 1995

- **A causal Bayesian network** (BN) or causal graph is composed by:
  1. A directed acyclic graph (DAG)
  2. A joint probability distribution that can be factorized as

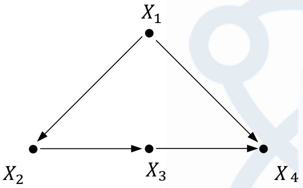
$$P(x_1, \dots, x_n) = \prod_i P(x_i | pa_i)$$



- A DAG is a collection of nodes and oriented edges that does not contain cycles
- In the context of causal graphs DAGs are employed to **represent causal structures**
- The vertices represent random variables, the edges describe the causal relations between them

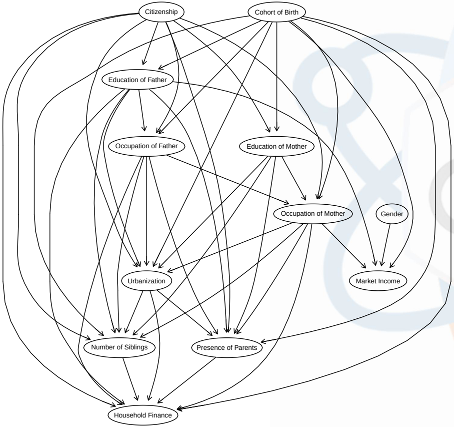
A simple DAG

- The path p along the ordered sequence of nodes (X_1, X_2, X_3, X_4) is a **directed path** since all the edges are oriented in the same direction along the path
- X_1 is called an **ancestor** of each node belonging to$\{X_2, X_3, X_4\}$ since it precedes them in p and the vertices in $\{X_2, X_3, X_4\}$ are **descendants** of X_1
- We can also say that X_1 is a direct cause of X_2 and X_4 or that X_1 is a **parent** of X_2 and X_4
- The same is true for every ordered pair of random variables (X_i, X_j) connected by a directed edge that goes from X_i to X_j in the DAG



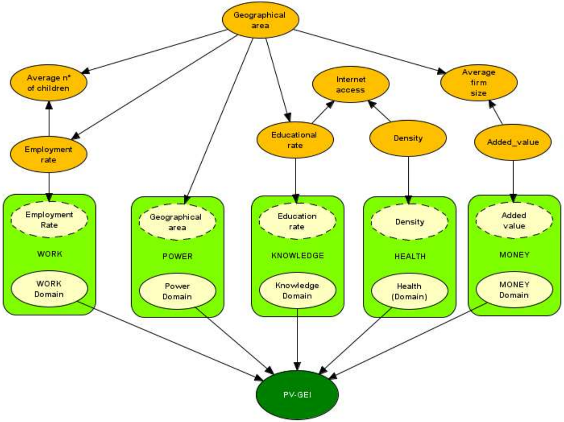
Example I: A Bayesian network to unveil inequality of opportunity

- The Bayesian network describe the relations between **socio-economic background and income**
- It possible to identify the **channels** that activate inequality of opportunity, and inform the policy about what actions can effectively reduce it



Example II: A Bayesian network to investigate gender equality

- The Bayesian Network replicates the structure of the **Gender Equality Index**
- The network allows the **computation** of the index and the investigation of **interactions** between domains



Outline

Causal inference and Potential Outcomes

Causal Bayesian networks: Fundamentals

Causal Bayesian networks: Structural learning

- Application I: Collaboration and research productivity

Causal Bayesian networks: Causal effect estimation

- Application II: Home based working and expected revenues

The origin of a DAG

- If **complete knowledge** of the subject matter is available
  1. the DAG structure can be outlined directly
  2. the joint probability distribution can be defined
- **If knowledge is partial** or totally missing
  1. **structural learning algorithms** can be employed to retrieve the structure of the graph
  2. the joint probability distribution can be obtained through the **EM algorithm**

Structural learning algorithms

- We want to investigate the **multivariate relations** between the variables belonging to a set $\mathbf{X}$ from a dataset $D(\mathbf{X})$
- We assume the existence of an **unknown underlying model** described by a DAG $G(V, E)$ and a joint probability distribution $P(V)$, from which $D(\mathbf{X})$ has been sampled
- Once the graph is learnt, a joint probability distribution over the nodes of the graph can be obtained through maximum likelihood estimation. This phase usually involves computing maximum likelihood estimates subject to the independence constraints encoded in the graph
- Structural learning algorithms can be divided in three families: *constraint-based* algorithms, *score-based* algorithms and *hybrid algorithms*.

Structural learning algorithms assumptions

- The assumptions usually focus on **the relation between the graph and the distribution** of the data employed to learn it
- A usually required assumption is **faithfulness**. A graph G faithfully represents a dataset D, if all the independence relations embedded in D are entailed by the structure of the graph G
- When dealing with causal graphs, another key assumption is **causal sufficiency**. The assumption states that a given set of variables $\mathbf{X}$ is causally sufficient for a population if and only if in the population every common cause of any two or more variables belonging to $\mathbf{X}$ is in $\mathbf{X}$
- Structural learning can be employed with **both discrete and continuous data**. The latter require the normal distribution assumption on each node. Also mixed data can be handled through the conditional gaussian distribution.

Constraint-based algorithms

- Constraint-based algorithms learn the graph's structure via **conditional independence statements emerging from data**
- They usually start with a complete graph, and then if two variables turn out to be **marginally or conditionally independent**, the edge connecting them is deleted.
- This procedure is repeated iteratively until a **stopping criterion** is satisfied.
- Constraint-based algorithms require making statistical decisions concerning how to assess conditional independence. Several tests can be employed to check if conditional independence holds, and violations of the assumptions required by the tests can generate unreliable independence statements
- Constraint-based algorithm include the PC algorithm, the IG algorithm, and the most recent Grow-Shrink algorithm

Score-based algorithms

- Score-based algorithms rely on a given **score function** that measures how well a certain DAG describes a dataset
- Common choices for the score function are the likelihood function or the Bayesian Information Criterion (BIC)
- These algorithms usually begin by computing **the score of an initial graph**.
- **The diagram is then modified** by introducing, deleting or reversing edges, and its score is computed again for each modification.
- The graph recording the best score at the end of the procedure is retained as the algorithm's output.
- Algorithms belonging to this family include the greedy search, the simulated annealing and genetic algorithms

Hybrid algorithms

- Hybrid algorithms aim to exploit the advantages of score-based and constraint-based algorithms by **merging** them in a single procedure
- Generally, they begin with a **restrict phase** where the parents of each node are selected through tests of conditional independence, similarly to what happens in constraint-based algorithms.
- The second phase is called **maximize** and consists in selecting a DAG in the restricted DAG family outlined by phase one by optimizing a given score function.
- The most used hybrid algorithms are the Sparse Candidate (SC) algorithm and the Max-Min Hill-Climbing algorithm (MMHC)

Background knowledge

- Structural learning algorithms are employed when **information concerning the causal graph is not available or partial**
- Partial knowledge can be introduced in structural learning procedures by **imposing constraints** on the structure of the obtained network:
  1. If is known that a variable X_i cannot cause a second variable X_j, the directed edge that goes from X_i to X_j is forced to be absent.
  2. If background knowledge suggests that X_i affects X_j, a directed edge from X_i to X_j can be imposed
- A consequence of including previous knowledge in the learning phase is that the graph is not entirely obtained through the information contained in the data.
- The constraints on the structure of the graph **restrict the search space** of the algorithms and often reduce both uncertainty and computational time.

Outline

Causal inference and Potential Outcomes

Causal Bayesian networks: Fundamentals

Causal Bayesian networks: Structural learning
- Application I: Collaboration and research productivity

Causal Bayesian networks: Causal effect estimation
- Application II: Home based working and expected revenues

Application I: Collaboration and research productivity

Aim of the work

- This work aims to study the relationship between **university's collaborations** and university's **research productivity**.
- Focusing on two key forms of knowledge exchange networks: those stemming from **EU-funded projects** and scholarly **publications**
- Within this analytical framework, we refine our examination to include **two specific domains** within the European Research Council (ERC) sphere:
  1. Physical Sciences and Engineering (PE)
  2. Life Sciences (LS)

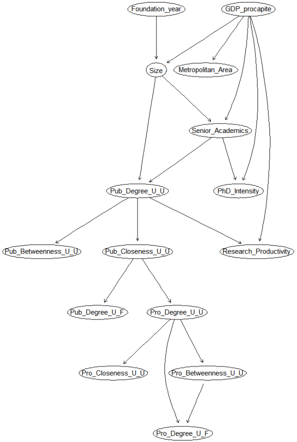
Data

- Data were collected from **three databases** hosted within the **RISIS** infrastructure and **FOSSR** over three years (2016-2018):
  1. EUPRO containing information on European funded programs
  2. CWTS containing information on publications
  3. ETER is a database providing information on Higher Education Institutions (HEIs).
- The **collaboration** variables are: university-university (U-U) and university-firm (U-F) degree, closeness and betweenness centralities for both project and publication networks
- The research productivity variable is based on **CWTS citation indicators**
- Yearly Bayesian networks are learnt through the **Tabu Search algorithm** with a *BIC score*
- **Background knowledge** concerning variable relationship is inserted through constraints

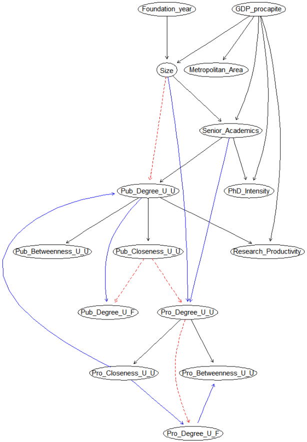
Obtained PE graphs

- **Degree centrality** is driving the other network indices
- In 2016 and 2017 **publication collaborations** drive project collaborations
- **Publication degree centrality U-U** affects research productivity
- In 2018 **research productivity** affects in turn firm project degree centrality

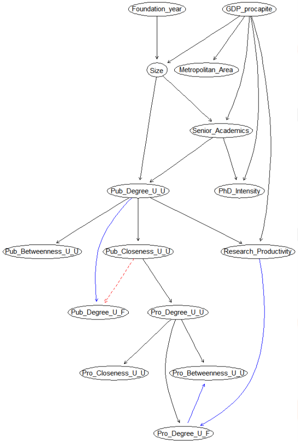
2016



2017

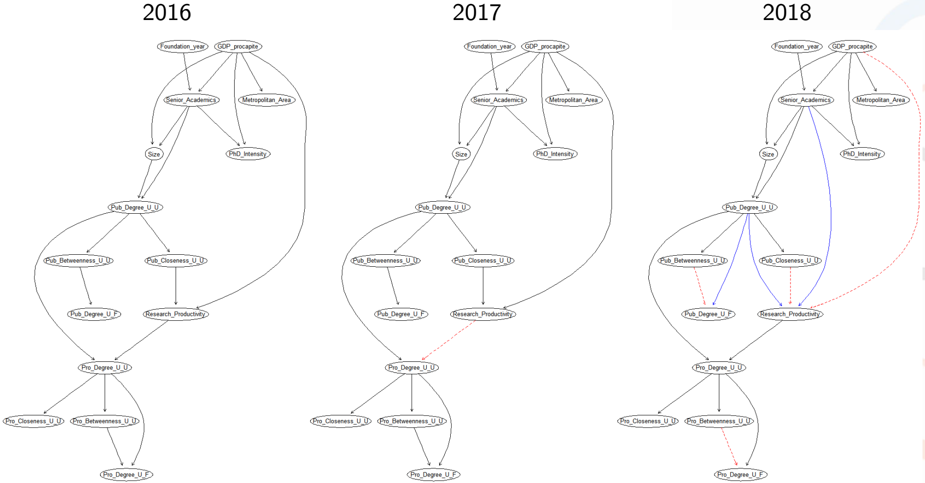


2018



Obtained LS graphs

- As in PE, there is a **hierarchy** in the collaboration variable structure
- **Research productivity is affected** by publication closeness and degree between universities
- **Research productivity affects** project degree between universities



Discussion

- **Q1**: How do different types of network centrality influence research productivity?
  A1: Network centrality measures appear to be positively correlated with research productivity, emphasizing the significance of direct and indirect inter-institutional connections
- **Q2**: Is there evidence of a bidirectional relationship between collaboration and research productivity?
  A2: The models reveal that collaboration in publication affects research productivity which then affects collaboration in projects.
- **Q3**: How do these relationships vary across different scientific domains?
  A3: In both domains collaboration present similar hierarchical structures and positively affects research productivity. However research productivity in turn positively affects project collaboration with firms in PE and project collaboration with other universities in LS.

Outline

Causal inference and Potential Outcomes

Causal Bayesian networks: Fundamentals

Causal Bayesian networks: Structural learning
- Application I: Collaboration and research productivity

Causal Bayesian networks: Causal effect estimation
- Application II: Home based working and expected revenues

Causal graph analysis at interventional level

- Pearl (2000) introduces the **do-operator** $do(X = x)$ to indicate that a variable X is forced by intervention to take value x
- The do-operator allows writing $P(Y|do(T = t))$ to denote the distribution of Y given an **intervention** that sets $T = t$
- This is different form $P(Y|T = t)$ that instead represents **the observational distribution** of Y given $T = t$
- The **causal effect** of T on Y can be obtained by comparing the quantity $P(Y|do(T = t))$ for different values of t
- Similar to what is done in the PO framework where instead $Y(t)$ was the quantity of interest.
- When dealing with non-experimental data, causal effects cannot be estimated directly from data since the interventional distribution of Y is not an observed quantity.

Back-door Criterion

- Causal graphs can be employed to **express interventional distributions in terms of observational quantities**
- This is a crucial result since conditional distributions such as $P(Y|T = t)$ can be directly available in a non-experimental context
- A graphical condition can be applied to causal graphs to test if a subset of its nodes is sufficient for identifying $P(Y|do(T = t))$ from observational data

Back-door Criterion

A set of variables $\mathbf{S} \subseteq \mathbf{X}$ satisfies the back-door criterion relative to a graph G with node set $\mathbf{X}$, a treatment variable $T \in \mathbf{X}$ and an outcome variable $Y \in \mathbf{X}$ if:

1. no node in $\mathbf{S}$ is a descendant of T; and
2. $\mathbf{S}$ blocks all the paths between T and Y that contain a directed edge pointing towards T.

BD Criterion and adjustment

- If the back-door criterion is satisfied by a set **S** (adjustment set) then interventional quantities can be expressed through observational ones:

$$P(y|do(T=t)) = \sum_{\mathbf{s}} P(y|t,\mathbf{s})P(\mathbf{s})$$

- Obtaining an adjustment set **S** through the back-door criterion also ensures that **S satisfies the unconfoundedness** condition for estimating the effect of T on Y
- The adjustment set can then be used to derive the interventional distribution through the adjustment formula, or directly estimate the ATE with a method of choice, such as regression, matching or inverse probability weighting.

Do-calculus

- Combined and iterative use of back-door and front-door criterion constitute the building block to identify causal effects on complex DAGs
- Pearl (2000) describes a set of rules based on the two criteria, also called **do-calculus**, that allows expressing interventional distributions in terms of observational distributions only, in an automated way
- The procedure has been proved to be **sound and complete** meaning that an algorithmic iteration of the rules of do-calculus always return a solution for the identification of causal effects, if such solution exists

Outline

Causal inference and Potential Outcomes

Causal Bayesian networks: Fundamentals

Causal Bayesian networks: Structural learning

Application I: Collaboration and research productivity

Causal Bayesian networks: Causal effect estimation

Application II: Home based working and expected revenues

Application II: Home based working and expected revenues

Context of the application

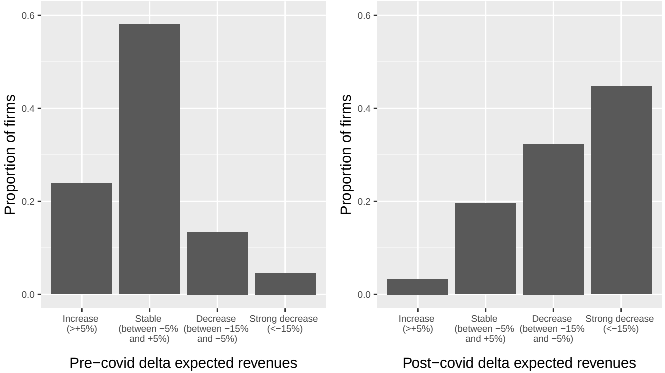
- The outbreak of **Covid-19** in March 2020 had **unprecedented consequences** on the Italian economy
- Firms tried to do everything possible to minimize losses
- **Home-based working** (HBW) has been one of the key firms' countermeasures
- The **implications of switching to HBW** have been thoroughly studied over the past years and its related literature has spiked in Covid-19 times
- However a **rigorous causal evaluation** of how home-based working affects firm performance seems to be missing

The ATE of Home Based Working

- The objective of the work is to study **the effect of home based working (HBW) on firm expected revenues** during the pandemic
- The analysis employs a **firm-level** dataset that covers:
  - Firms' characteristics, including financial and strategic components
  - The immediate effect of the covid shock on firms' organizations
  - The change in their future expectations

Selected variables (1/2)

- **The outcome variable Y** describes post-covid expectations towards future variation in revenues
- The same variable has been observed right before the pandemic outbreak and **the changes are pictured in the figure**



Selected variables (2/2)

- **The treatment** T is a binary variable that denotes if a firm has implemented home-based working for a portion of its employees
- The following **logical groups** contain the remaining variables:

Precovid demographics $\not\leftarrow$	Other precovid features $\not\leftarrow$	Precovid expectations $\not\leftarrow$	Postcovid features
Size (n. of employees)	Innovation, R&D	Pre-covid Δ expcted revenues	Confirmed Covid infections
Geographical area	Credit rationing		Essential business sector
Business sector	Export		
Manager education	Δ number of employees		
	Digital litteracy		
	Past Δ revenues		

- The average treatment effect (ATE) will be estimated through a mix of **Causal Bayesian Networks and Potential Outcomes**

Graph Learning

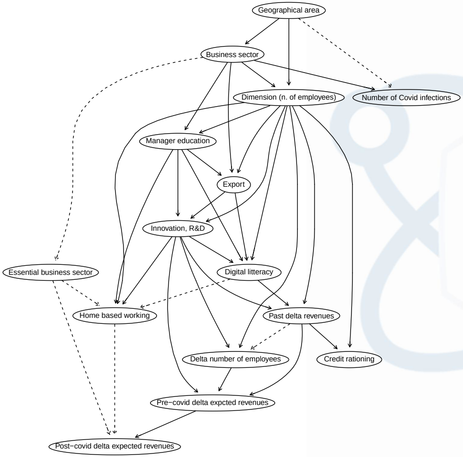
- **If the graph is not known it can be learnt from data** through a structural learning algorithm
- The **Tabu Search algorithm** with a *BIC score* is employed
- A set of constraints has been derived from the logical variable groups

Precovid demographics ↚	Other precovid features ↚	Precovid expectations ↚	Postcovid features
Size (n. of employees)	Innovation, R&D	Pre-covid delta expcted revenues	Confirmed Covid infections
Geographical area	Credit rationing		Essential business sector
Business sector	Export		
Manager education	Delta number of employees		
	Digital litteracy		
	Past delta revenues		

- **Additional constraints** based on known variable relations have been also added to to the model

Obtained Causal Graph

- The graph sheds light on the relational structure between variables
- Applying the back-door criterion to the graph outputs an **adjustment set** to estimate causal effects
- **Unconfoundedness is thus ensured** as long as structural learning recovers the correct graph structure


Geographical area
Business sector
Dimension (n. of employees)
Number of Covid infections
Manager education
Export
Innovation, R&D
Essential business sector
Digital litteracy
Home based working
Past delta revenues
Delta number of employees
Credit rationing
Pre−covid delta expcted revenues
Post−covid delta expected revenues

Causal effect estimation (1/2)

Adjustment set selection for matching

- The **adjustment set for the effect of T on Y** is

$$Z = \{\textit{Essential business sector}; \textit{Pre-covid delta expected revenues}\}$$

- **Full matching** is employed for causal effect estimation
- The weights deriving from full matching are employed in a **weighted regression model** for Y on T and S_{adj}
- The model is used to predict the interventional distributions of the outcome with **g-computation**

Causal effect estimation (2/2)

- The **ATE is then computed** as

$$ATE = \frac{E[Y_1]}{E[Y_0]}$$

Table: ATE estimates of HBW implementation on post-covid Δ expected revenues

Post-covid delta expected revenues	Point Estimate	95% C.I.
Increase (>+5%)	1.67	(1.45,1.89)
Stable (between -5% and +5%)	1.30	(1.22,1.37)
Decrease (between -15% and -5%)	1.22	(1.16,1.29)
Strong decrease (<-15%)	0.71	(0.67,0.75)

- HBW partly **counterbalances** the impact of the covid shock
- Coherent with the idea that it **increases productivity and flexibility**

THANK YOU!

✉ fossr.dissemination@ircres.cnr.it

@fossrproject

fossr.eu

fossr-eu

@fossr

zenodo zenodo.org/communities/fossr

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Consiglio Nazionale delle Ricerche

Exploring research infrastructure data through advanced statistical applications

September 26-27, 2024
online training course

FOSSR
Fostering Open Science in Social Science Research
Innovative tools and services to investigate economic and societal change

Image credits: S T on Unsplash; visual design: Rita Giuffredi.

Agenda

Set. 26

10:00 **Module 1.** Research Infrastructures and Open Science

Andrea Orazio Spinello, Emanuela Varinetti (CNR-IRCrES)

The module will outline the contours of the new paradigm for scientific knowledge production, emphasizing openness, the value of collaboration, and data availability. It will explore the features of data-intensive science, which underpin the creation and strengthening of Research Infrastructures (RI). Thse characteristics of these sociotechnical platforms will be discussed by analyzing both the Italian and European contexts, with reference to FOSSR's efforts in developing the Italian Open Cloud for Social Sciences.

12:30 *lunch break*

14:00 **Module 2.** Accessing and querying interoperable RI data

Lucio Morettini, Andrea Orazio Spinello, Emanuela Varinetti (CNR-IRCrES)

The module will present examples of accessing and querying data from research infrastructures, highlighting the importance of data interoperability. Specific cases will illustrate the use of the RISIS infrastructure for science, technology and innovation studies. We will present cases in which research questions are addressed differently depending on the nature of the data, analysing the ways of managing datasets and their enrichment with data external to the infrastructure. The content presented will highlight the value of shared database access.

16:00 *end of day 1*

Set. 27

10:00 **Module 3.** Network models applied to RI data

Antonio Zinilli (CNR-IRCrES)

The module aims to illustrate the basic concepts and statistical measures of network science and provide an overview of the main statistical network models. The module will conclude with two applications where networks are analysed using data from research infrastructures. The two applications that will be covered in this module are:

- **Application 1**: Complex networks and academic project funding
- **Application 2**: Research collaborations and research productivity

12:30 *lunch break*

14:00 **Module 4.** Causal Bayesian networks and applications to RI data

Lorenzo Giammei (CNR-IRCrES)

The module aims to illustrate the basic concepts and statistical measures of network science and provide an overview of the main statistical network models. The module will conclude with two applications where networks are analysed using data from research infrastructures. The two applications that will be covered in this module are:

- **Application 1**: Complex networks and academic project funding
- **Application 2**: Research collaborations and research productivity

16:30 *end of the training*

www.fossr.eu | @fossrproject | fossr.eu | fossr-eu

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
Consiglio Nazionale delle Ricerche

Course description

The concept of '**open science**', which encompasses knowledge sharing within scientific communities and the interaction between science and society, plays a crucial role in contemporary research, particularly in the field of Social Sciences and Humanities (SSH). Open sharing of knowledge, including data and services for research, enables the exploration of new research questions, drives interdisciplinary collaboration, and fosters the development of innovative analytical tools. From a societal perspective, **it is essential to ensure that research findings are accessible not only to the scientific community but also to social and political stakeholders, thereby amplifying the impact of scientific work on decisions that affect citizens**. A significant boost to the development of 'open science' comes from the creation and enhancement of **Research Infrastructures** (RIs).

RIs facilitate 'open science' by **enabling researchers and other stakeholders to access high-quality data, tools, and services**. In Italy, the SSH community is still in the early stages of transitioning toward a knowledge production model based on widely shared research infrastructures. **The FOSSR project aims to sustain and encourage this process by enhancing the Italian nodes of three existing European RIs and implementing new services, tools, and resources to create a conducive environment for the development of open science in SSH in Italy**. Effective dissemination regarding the opportunities linked to RIs is essential to ensure that individuals are aware of available data and resources. Additionally, users must have the skills and expertise to use infrastructure data appropriately.

The course will guide participants through the process of **engaging with a RI and making reasoned use of its resources for research purposes**. Furthermore, it will present the application of advanced statistical techniques on infrastructure data, particularly **Network Models and Bayesian Modeling**.

Training objectives

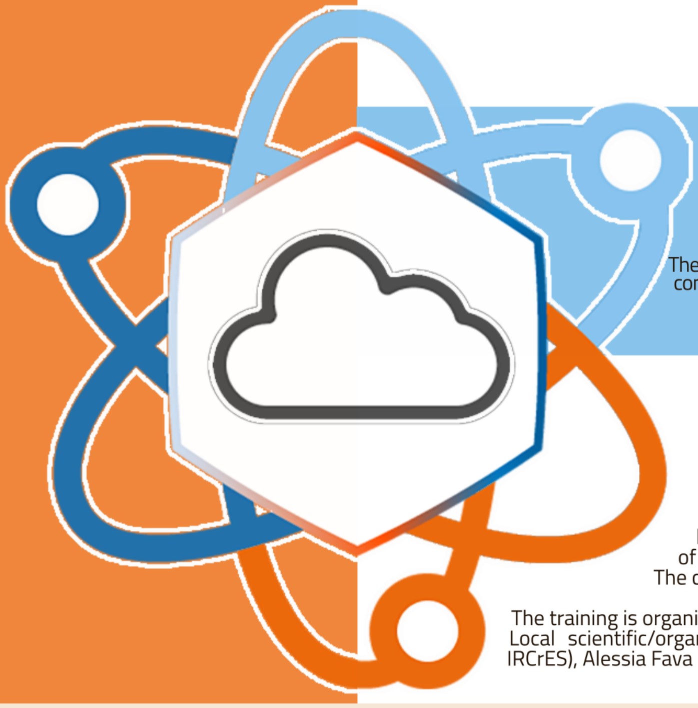
The main aim of this methodological course is to provide an overview of the opportunities for exploiting shared data within RIs.

Specifically, the course aims to raise awareness of these opportunities and **encourage the implementation of data-driven approaches through the application of advanced statistical techniques**. The focus is primarily on the **RISIS European Infrastructure, on topics related to Science, Technology, and Innovation (STI)**, one of the three infrastructures that initiated the FOSSR project, and on advanced statistical techniques that FOSSR is treating in its research work packages. Drawing inspiration from the data available within RISIS, the course will explore aspects of **data access, interoperability and processing in depth**. From data exploration to data processing, the process will be substantiated by the exemplary application of advanced statistical techniques, particularly Network Science and Bayesian Modelling, that will be treated in theory and practise.

After completing this course, learners will be aware of the opportunities linked to RIs in SSH, particularly the RISIS infrastructure, and will gain in-depth knowledge of techniques developed within the FOSSR research Work Packages.

Audience

The course is designed for individuals who are or wish to be involved in **creating, capturing, analysing, or generally managing research data within the social science disciplines**. The target audience includes, but is not limited to, early-career researchers, researchers aspiring to advance their careers, technicians, data stewards, and data managers.



How to apply

Applications should be submitted **within September 19th, 2024** through the form available at: ***https://l.cnr.it/fossr-training-set24***

More information: ***www.fossr.eu/eventi/fossr-training-sept24/***
Contact person: **andrea.spinello@ircres.cnr.it**

The online training is open to everyone, with no specific requirements on previous knowledge or competences. However, due to the format of the training course, **the number of participants is limited to a maximum of 35**. A selection will be made on the basis of the closeness of the scientific field covered by the submitted curriculum vitae to the subject matter of the course. Notification of acceptance: September 20th.

Organisational details

The course will be structured in four on-line training modules distributed in two days, each one based on frontal lessons and several interactions. The course will be held online; the link will be provided to participants in due time before the beginning of the course.
The course is conducted in Italian, while the materials will be published in English.

The training is organised by CNR-IRCrES in the frame of the FOSSR project.
Local scientific/organising committee: Andrea Orazio Spinello (CNR-IRCrES), Serena Fabrizio (CNR-IRCrES), Alessia Fava (CNR-IRCrES), Rita Giuffredi (CNR-IRCrES), Alessandra Maria Stilo (CNR-IRCrES).

This event was supported by FOSSR (Fostering Open Science in Social Science Research), funded by the European Union - NextGenerationEU under NRRP Grant agreement n. MUR IR0000008.