SOBIGDATA
RESEARCH INFRASTRUCTURE

TERRITORI APERTI

Open Data, European Initiatives and SoBigData RI,

Conferenza Nazionale di Presentazione delle Attività di Ricerca

Roberto Trasarti - ISTI CNR, Pisa
Coordinator SoBigData RI

SoBigData.eu receives funding from the European Union's grant agreements No. 654024, 871042 and 101079043.
Is also funded by the National Recovery and Resilient Programme for the Central Hub: "SoBigData.it"

Open Data/Science:

Freely available to everyone to use, reuse, and redistribute without any restrictions...

...At least this is what ideally should be in research.

Why Open Data/Science is fundamental

Data access is fundamental for researchers to study **social dynamics and detect systemic risks** in the digital sphere, with the goal to build a **future proof society** in which technology benefits all.

Multi-disciplinary approach requires sharing data and knowledge.

The principles

F **Findable**

Data and metadata must be easy to find by both humans and machines, also in an automated way using machine-readable metadata

A **Accessible**

Once found, it must be possible to access to the data/ metadata in an easy way, with clearly defined conditions

I **Interoperable**

The data/metadata must be in such a format that it is possible to combine it with other data, also using automated means

R **Reusable**

It must be possible to reuse the data/metadata for future research and to further process it also using automated means

F **Fair**

The results must be impartial and non-discriminatory, and efforts must be made to avoid unfair results even if they are computationally correct

A **Accurate**

The results must be correct and up to date, so misleading conclusions are avoided

C **Confidential**

Privacy, trade secrets and confidentiality are respected throughout all the process of data analysis

T **Transparent**

The results can be explained and understood in a transparent manner, so they can be trustworthy

The data and metadata used should respect the **FAIR** guiding principles which foster the reusability of information both by humans and machines.

Data science activities should be carried out in an ethical and secure manner so they can benefit society without violating fundamental values and individual' rights. To reach this goal, we should foster **FACT** data science

To be FAIR… it is not an easy task

F1. (Meta)data are assigned a globally **unique and persistent identifier**

F2. Data are described with **rich metadata**

F3. (Meta)data are **registered or indexed in a searchable resource**

A1. (Meta)data are retrievable by their identifier using a **standardised communications protocol**

A1.1 The **protocol is open, free, and universally implementable**

A1.2 The protocol allows for an **authentication and authorisation procedure**, where necessary

A2. **Metadata are accessible, even when the data are not available**

I1. **(Meta)data use a formal,** accessible, shared, and broadly applicable language for knowledge representation.

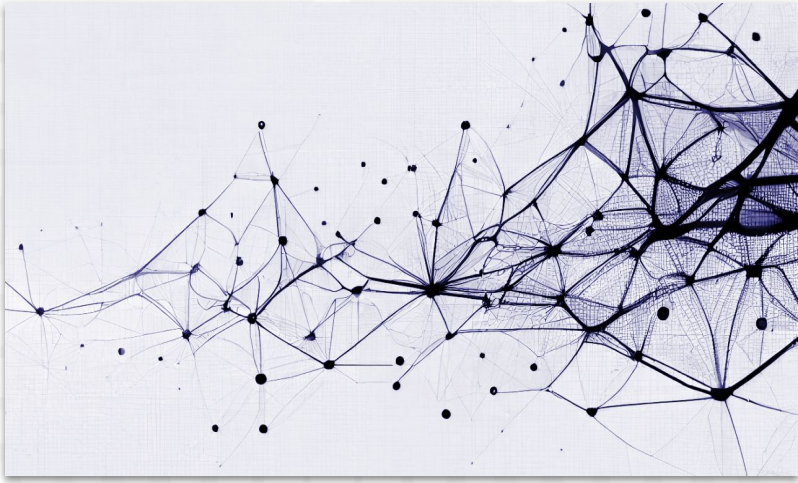
I2. (Meta)data include qualified **references to other** (meta)data (linked data)

R1 (Meta)data are released with a clear and **accessible data usage license**

R2 (Meta)data are associated with detailed **provenance**

Not only for Humans...

We need to be **FAIR for ML and AI technologies** which require data that can be meaningfully and accurately interpreted by machines. Having a AI-aware and AI-ready data management is the key to **unlock potentialities** of Retrieval Augment Generation (RAG), Knowledge Graphs, and other LLM components.



Barriers for FAIR Data

1. Data fragmentation and interoperability issues: Data is often scattered across various platforms, databases, and file formats, making it challenging to locate and access. A lack of standardized data models, ontologies, and controlled vocabularies.

2. Limited accessibility: Data access may be restricted due to proprietary or privacy concerns, leading to limited availability: Legal and ethical issues, data security, confidentiality, and intellectual property

4. Data quality and documentation: Inadequate documentation, incomplete metadata, and inconsistent data formats can affect data quality and reliability.

5. Infrastructure and resources: Labs may lack the necessary infrastructure, resources, and technical expertise to implement effective data management practices.

6. Cultural and incentive barriers: The scientific community traditionally emphasizes publishing research outcomes rather than sharing raw data.

European Strategy for Data

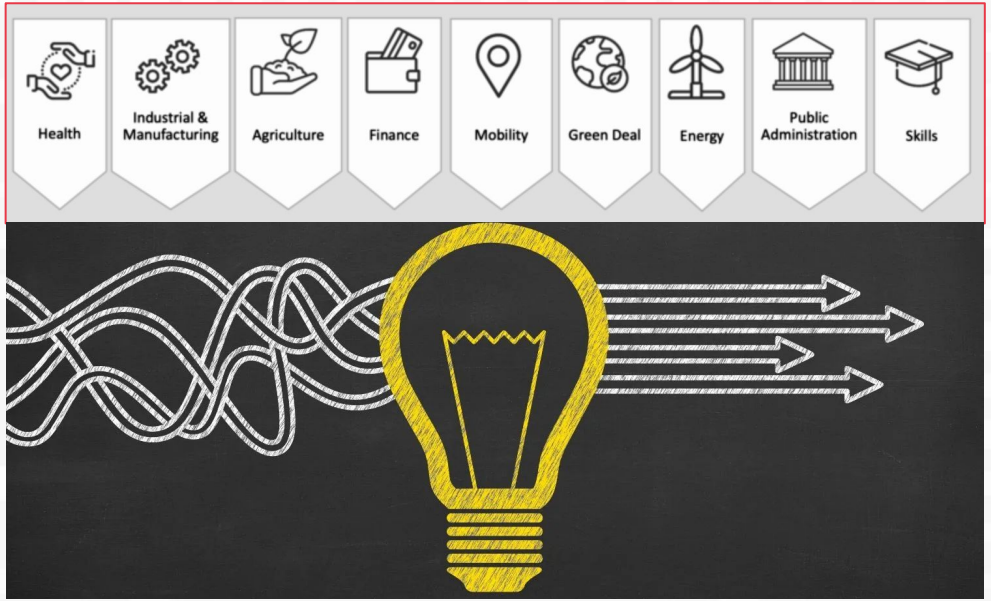
Proposed by the **European Commission** in 2020 aims at creating a single market for data that will ensure Europe's global competitiveness and data sovereignty. Its legislative pillars are two Regulations: the **Data Governance Act** (23rd of June 2023) and the **Data Act** (27th of November 2023).

In order to lay down a list of specific **high-value datasets and the arrangements for their publication and reuse**, the EU Commission adopted the **Implementing Act on High-value Datasets Under the Open Data Directive** (Regulation (EU) 2023/138).

Common European Data spaces

the European strategy for data of February 2020 set out the path to the creation of common European data spaces in a number of strategic fields: health, agriculture, manufacturing, energy, mobility, financial, public administration.

Working on Project Proposal Calls and synergies between them to create commons



EOSC: The European Open Science Cloud

It arises from the European Commission's stated intention to increase the circulation and exploitation of knowledge by promoting open access to the data resulting from publicly funded research under Horizon 2020.

Work on Collaborative working groups and associations

ESFRI: European Strategy Forum on Research Infrastructures

It is a strategic instrument to develop the scientific integration of Europe and to strengthen its international outreach. ESFRI has established a European Roadmap for **research infrastructures** for the next 10-20 years.

Work on roadmaps for Research infrastructures

The Digital Infrastructure for Social Mining: **SoBigData RI**

A **distributed**, **Pan-European**, **multi-disciplinary** research infrastructure aimed at using **social mining and big data** to understand the **complexity** of our contemporary, globally interconnected **society**

SoBigData RI objective

Use big data to **understand** the **complexity** of society and **offer no-profit** services to researchers, industry, public bodies, and citizens through the creation of a **multidisciplinary scientific community** according to the European vision of an **ethical**, **legal**, and **open data and science platform.**

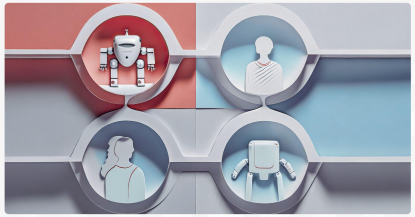
Research Spaces

Vertical contexts fostering **Data**, Methods and Technologies towards grand societal challenges

Societal debate
and misinformation

Demography, Economy
and Finance 2.0

Sustainable Cities for Citizens



Social Impact of AI and
explainable machine learning

Health Studies

Societal and Industrial Impact
of Next-Gen. Internet
& beyond 5G Networks

Pervasive Intelligence
in Cyber-Physical Systems
for Future Society

Disaster response
and recovery

SoBigData RI users

- Creating a **multi-disciplinary** scientific community
- **Boosting FAIR data sharing** and **research quality** among the nodes
- Embracing the **European strategy on ethical and legal** principles
- Connect to **EOSC initiative**

- Promoting the **technological transfer and data sharing**
- Facilitating the creation of new **business opportunities**

- Giving **value to public data** to understand the territory
- Creating **indicators** for policy makers
- Designing new **services for citizens**

Distributed over 13 Countries

UK

Sweden

Poland

Netherlands

Italy

Greece

France

Finland

Spain

Estonia

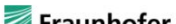
Germany

Belgium

Austria

SoBigData

Main Research Institutes



Main Companies

SoBigData Research Infrastructure

SoBigData.it RRNP

Funds: **20 Million**
Institutions: 11
Central Node,
National Project

SoBigData PPP HE

Funds: **3.2 Million**
Institutions: 28
Countries: 14

SoBigData H2020

Funds: **6 Million**
Institutions: 11
Countries: 6

SoBigData ++ H2020

Funds: **10 Million**
Institutions: 23
Countries: 12

PLATFORM USERS

12'000+ users

2015 — 2020 — 2022 — TODAY — 2025

SBD RI established in 2015 as an Horizon 2020 project to integrate existing EU platforms on various social mining topics

SBD RI selected to be part of the ESFRI roadmap as relevant digital infrastructure in the European Research Area (ERA)

SBD RI included in the top relevant infrastructures in Italy and found with the National Recovery and Resilience Plan (NRRP) to strengthen its capabilities

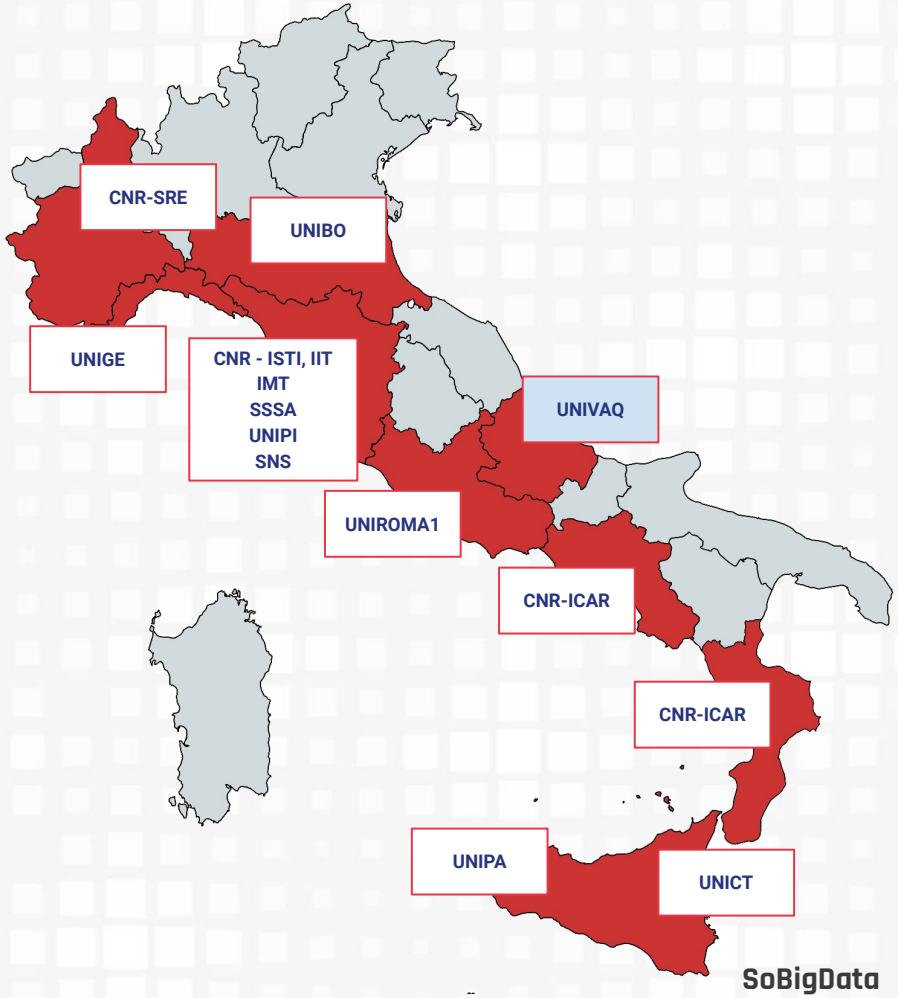
E R I C

SoBigData.it:
Strengthening the Italian RI for Social Mining and Big Data Analytics

(NRRP - Central Hub)

The Objective is the scientific and technological empowerment of the Italian node of the infrastructure creating a strong italian community with a multidisciplinary approach to social challenges.

In particular the **synergies** with UNIVAQ created a new data center (about 2,500,000€) and training initiatives (about 500,000€) making them one most important site in the SoBigData Italian Node.



SOBIGDATA
RESEARCH INFRASTRUCTURE

250+	researchers and PhDs	220+	big datasets
900+	scientific publications	200+	methods and apps
12,500+	RI users	2 mln+	peak access to apps
5,200+	young researchers trained	175+	pilot projects

Updated on 31 January 2024

info@sobigdata.eu

www.sobigdata.eu

SOBIGDATA
RESEARCH INFRASTRUCTURE

Coordinator: *Roberto Trasarti*
Management team: *Valerio Grossi and Michela Natilli*
Communication team: *Daniele Fadda, Katia Genovali, Beatrice Rapisarda*
Institute of Information Science and Technologies (ISTI),
National Research Council (CNR), Pisa, Italy

SoBigData.eu receives funding from the European Union's grant agreements No. 654024, 871042 and 101079043.
Is also funded by the National Recovery and Resilient Programme for the Central Hub: "SoBigData.it"

SoBigData