Introducing the opportunities and challenges of sensitive data in the social sciences

Bonnie Wolff-Boenisch
Executive Director of CESSDA

Assises Nationales des données ouvertes de la recherche Paris, December 2025

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'ESPACE
université PARIS-SACLAY
FRANCE

CESSDA

Consortium of European Social Science (SoS) Data Archives

- Is a virtual, distributed RI with seat in Bergen, Norway
- 23 Member States and their data archiving infrastructures (Service Providers) - for France: Progedo
- Supports 11 partner countries across Europe including the Ukraine Data Archive

MISSION

to provide a full-scale sustainable RI that enables the research community to conduct high-quality research in the SoS contributing to solutions to the major challenges society is facing today.

HOW?

- CESSDA provides data services to social scientists (and beyond) across the whole Research Data Lifecycle with focus on data collection, curation and long term data preservation.

Is a member of

- EOSC (European Open Science Cloud),
- RDA (Research Data Alliance)
- DDI (Data Documentation Initiative)
- MoU with OPERAS aisbl
- MoU with CODATA (the Committee on Data) of ISC

What is the point?

While there are **Open Data** requirements from funders and journals, there is data that is **never shared** and made available for **secondary analyses or replication** because they are too sensitive.

This represents a significant loss of scientific potential.

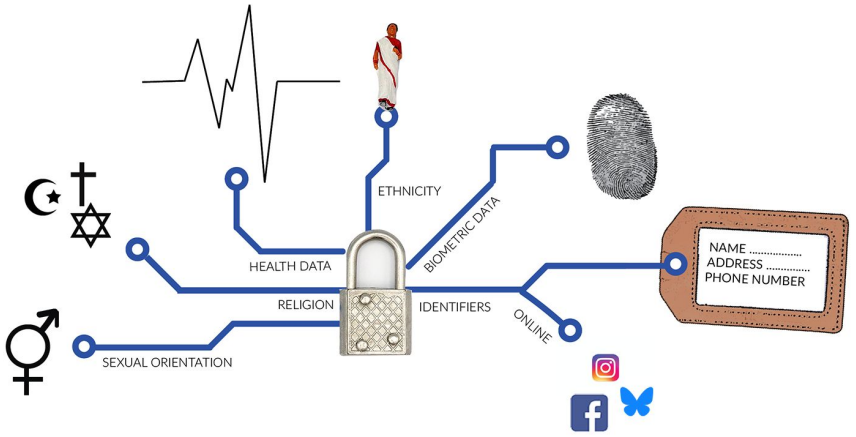
Working definition of sensitive data

“Sensitive data” is used here as

- a shorthand for a broad category of data that cannot be shared because of data protection
- or other legislation, including data that are personal and identifying, or copyright protected

Examples of sensitive data in the context of Social Sciences Research

- **Personal data:**
  - Identifiers such as names or identification numbers; physical, physiological, genetic, mental, economic, cultural or social characteristics
  - Location data from GPS or mobile phones
- **Confidential data:** trade secrets (business processes, algorithms, or financial information that could confer competitive advantage, investigations, data protected by IP
- **Protected Information**: national safety, military information
- Data sensitive that are not legally based, for example:
  - **Group harm**: Even anonymized data might stigmatize (a)
  - or even harm communities (b)
    - a) e.g., linking ethnicity or location to crime or disease rates
    - totalitarian regimes, interviews



Combination of different variables in different datasets that can be combined into sensitive or personal data

The challenges of making sensitive data available

- Sensitive data is usually made available through **Trusted Research Environments** (TREs)
  1. **Policy and legal constraints** to be checked:
     - Are there legal restrictions or restrictions imposed by the data owners or both?
     - Do you have the rights to onward share the data?
     - Do the survey consents allow sharing for research use?
     - If using an external infrastructure, are you allow to transfer the data?

  2. **Resource** constraints for running a TRE
     - Limited funding
     - Low pool of trained professionals

  3. **Technical** constraints:
     - Many decisions needed - e.g.,
       - Build in-house vs. working with external solutions
       - What do you need now vs. what you might need in the future

Example of combining cross-disciplinary sensitive data

A study (Johnny Downs (2019) - https://bmjopen.bmj.com/content/9/1/e024355 that linked clinical mental health service data of ~35,500 children and young people to education/social-care data from the National Pupil Database in England.

- Linkage between health and education records offers a powerful tool for evaluating the impact of mental health on school function.
- Collaborative research with data providers is needed to develop linkage methods that minimise potential biases in analyses of linked data.

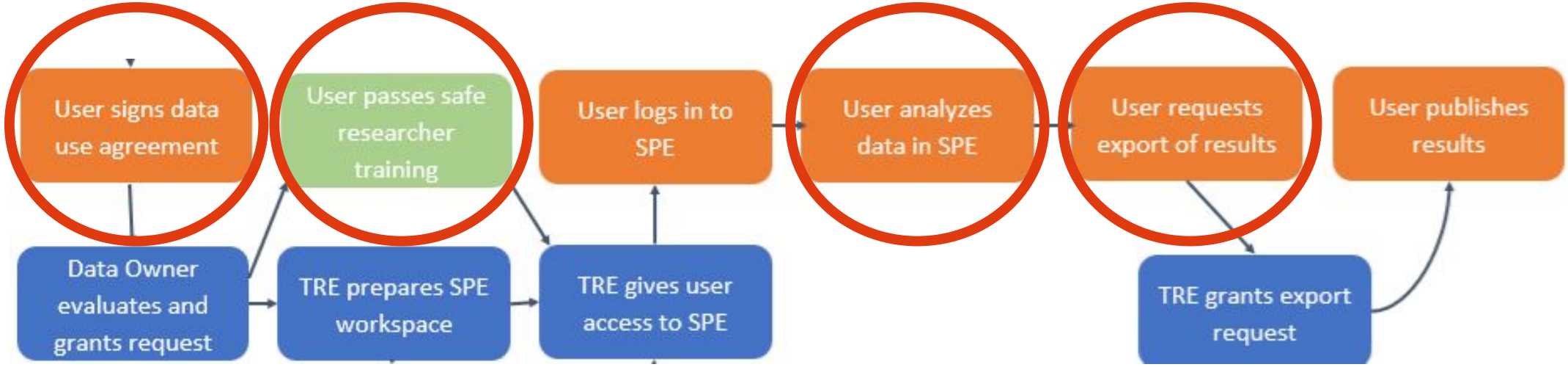
- This type of research is only possible through a secure data infrastructure, which enabled lawful, privacy-protected linkage of identifiable records across sectors.

Current landscape in Europe

- There are national archives such as CESSDA's SPs that can handle data that can be sufficiently anonymised
- They often lack the infrastructure for preserving and disseminating sensitive data
  - Such an infrastructure could include
    - safe rooms or remote access
    - with highly restrictive conditions
    - with data stored on secure servers

Good Practice in Action - GESIS Secure Data Center

- Physical access to sensitive data via a Safe Room at GESIS, Cologne
  - Secure computing platforms where data never leaves the environment



*Secure Protected Environment (SPE)

A few current initiatives to redress the infrastructure problem

- CESSDA sensitive data working group (SDWG)
- A few networks: International Secure Data Facility Professionals Network (ISDFPN), International Data Access Network (IDAN)
- EOSC-ENTRUST project

EOSC-ENTRUST - towards an European Trusted Research Environment

- **Breaking national data silos**: researchers to access and compare sensitive socio-economic data across European countries through standardized secure environments
- **Proving universal applicability**: Uses social sciences' unique heterogeneity (administrative, survey, mixed methods data) to validate that one framework can serve all research disciplines
  https://eosc-entrust.eu/

International access to sensitive data

Internationally, researchers often face significant hurdles in terms of:

- time and financial burden having to travel to a Safe Room, and
- being associated with a Higher Education institution of the other country

Synthetic Data

- Synthetic data is **artificial data** that is generated from original data and a model that is trained to reproduce the characteristics and structure of the original data.

  - complementary to TREs - not a replacement
  - mimics statistical properties of real data without containing actual individual records
  - training / educational purposes

Where to go from there?

- More research is needed to establish procedures for defining the balance between privacy protection and analytical value of synthesized dataset - when do we really need access to original data
- Support activities for investments in infrastructure for sensitive data at service providers for federated access
- Promote FAIR sensitive data for the social sciences
- Provide guidance in compliance with European and international standards
- Improve the exchange of data

Thank you for your attention!

cessda.eu

@CESSDA_Data

@CESSDA ERIC

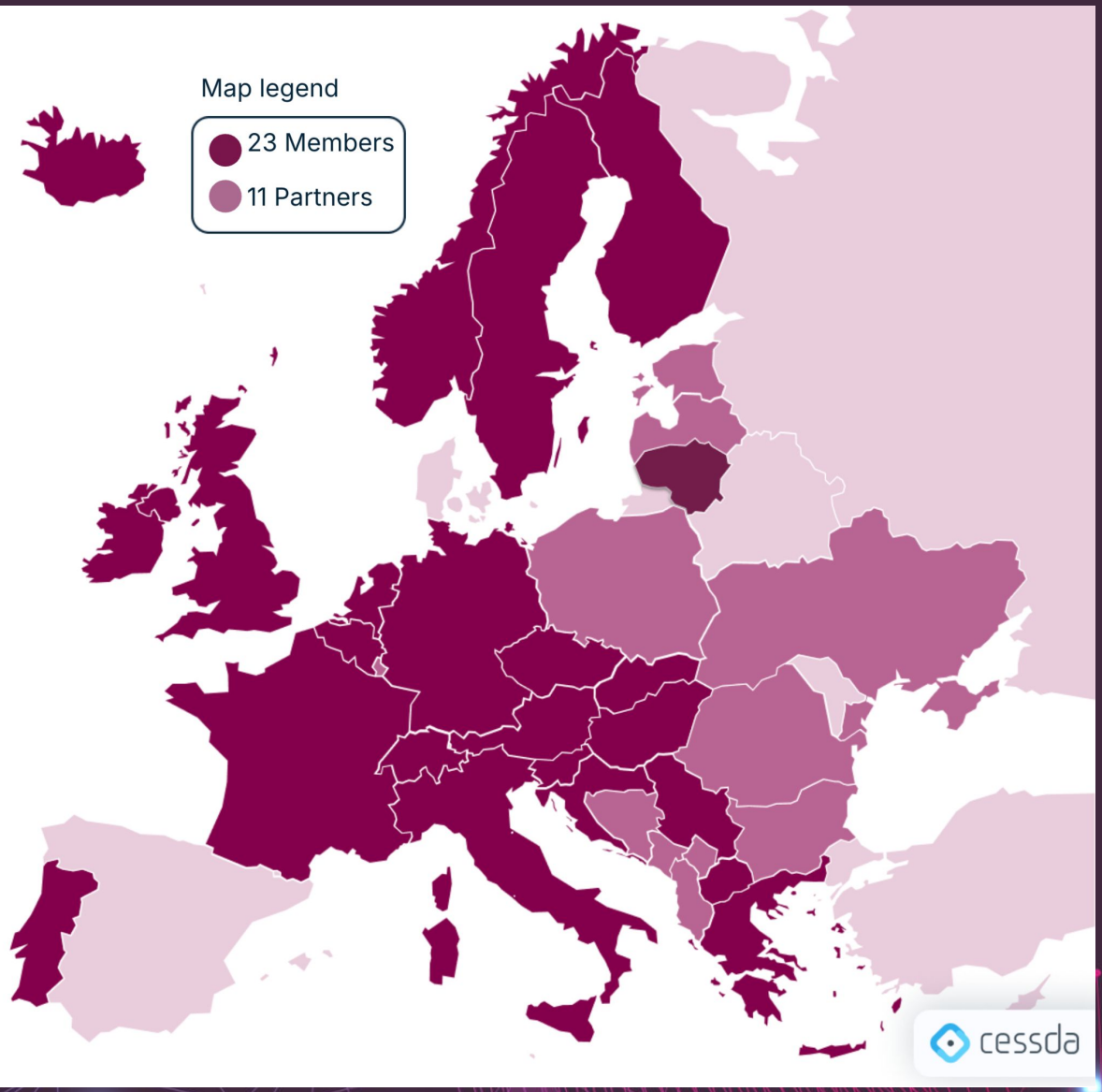
CESSDA’s Membership

23 European Member States representatives in the GA

Austria
Belgium
Croatia
Czech Republic
Denmark
France
Finland
Germany
Greece
Hungary
Iceland
Ireland
Italy
Lithuania
Netherlands
North Macedonia
Norway
Portugal
Serbia
Slovakia
Slovenia
Switzerland
Sweden
UK

11 Partners

Albania
Bosnia and Herzegovina
Bulgaria
Estonia
Kosovo
Latvia
Luxembourg
Montenegro
Poland
Romania
Ukraine



What are the opportunities investing into infrastructures for sensitive data?

- Greater compliance with funder and journal requirements on data sharing
  - More data shared, and greater re-use, replication, and transparency.
- Availability of higher-quality and richer data that are less aggregated due to the de-identification processes needed to make data anonymous.
- Acceleration of scientific discovery and innovation.
- With more potential for reuse, less additional data collections are required, and the risk of vulnerable populations is reduced.