Data Science Competence Center (DSCC)

Access to Relevant Data

Prof. Dr. Bertrand Loison

Vice-Director General at the Swiss Federal Statistical Office

Head of Data Science, AI & Statistical Methods Division

7th International Conference on Big Data and Data Science for Official Statistics, 05.11.2022, Yogyakarta, Indonesia

How does DSCC work in partnership with private sector and academia?

Why did FSO decide to establish the DSCC?

What data does FSO traditionally get access to?

What are the objectives of DSCC?

How is DSCC organized?

Does FSO have experience with access to mobile phone data?

What are the deliverables of DSCC?

What new data sources does DSCC get access to?

What is the importance for DSCC to work with privacy-enhancing technologies?

Agenda

1. Introduction
2. From an FSO's Working Group to a Federal Strategy on Data Science
3. National Data Management
4. Conclusion

United Nations E/CN.3/2016/6*

Economic and Social Council

Distr.: General
17 December 2015

Original: English

Statistical Commission
Forty-seventh session
8-11 March 2016

Big data methodology and estimation

21. Another conclusion reached in the panel discussions was that big data needs official statistics as much as official statistics need big data. This is not only because the production of official statistics is anchored in internationally agreed quality frameworks and methodologies and based on principles of professional independence and trust; it is the official statistics using traditional source data that allow methods and techniques for generating statistics from big data sources to be calibrated, "trained" and, ultimately, validated. Other findings were that statistical methodology can turn big data into small data, for example, through sampling, and that the transfer of data is not always necessary, as the method or algorithm can be applied at the location of the data source.

A Changing Society

« Ensuring statistics accurately reflect a changing economy is one of the hardest challenges NSIs face.

The economy's complexity and structure are becoming increasingly difficult to capture within the basic conceptual framework that underpins the national accounts. When the statistical framework was first devised, the economy was one in which most businesses were engaged in the production of reasonably homogenous goods in a single country.

The reality today is rather different, with many businesses operating across national borders and producing a range of heterogeneous goods and services that may be tailored to the tastes of individual consumers. »

Independent Review of UK Economic Statistics

Professor Sir Charles Bean

Source: Charles Bean, Independent Review of UK Economic Statistics, (2016). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/507081/2904936_Bean_Review_Web_Accessible.pdf, pp. 116.

Challenges of the Data Revolution

Source: The Economist, May 2017

Four technical trends

- Data volumes are increasing
- The nature of data is changing
- Processing power is improving
- Cloud computing

Three analytics trends

- Real-time analytics will be in demand
- Analytics at the edge will be the new normal
- Complex and agent-based models will be needed

Agenda

1. Introduction
2. From an FSO's Working Group to a Federal Strategy on Data Science
3. National Data Management
4. Conclusion

… and the Response of the Federal Administration

2015

Internal FSO working group on big data

2017

Data Innovation Strategy
Pilot projects

2020

Data Science Competence Center (DSCC)

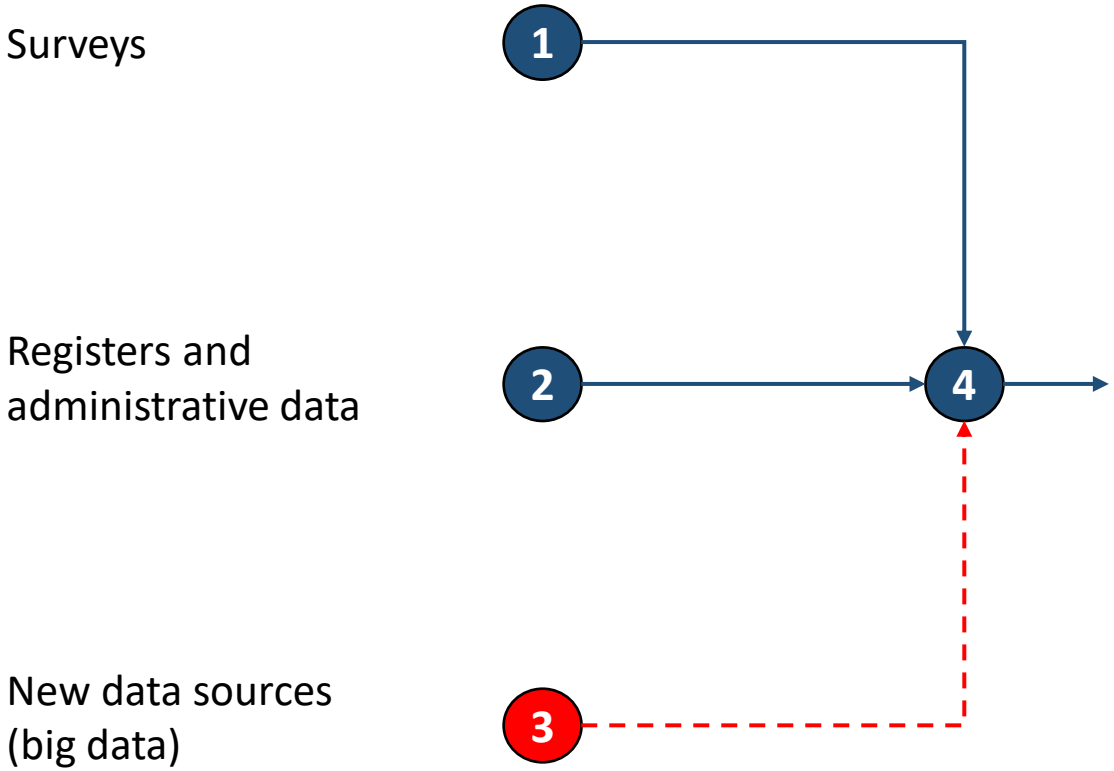
2021

Competence Network for Artificial Intelligence (CNAI)

2023 - …

Federal Data Science Strategy (DSStB)

Are Big Data and Data Science Really Useful?



	Survey data	Administrative data	Big data
Specification	Statistical products specified ex-ante	Statistical products specified ex-post	Statistical products specified ex-post
Purpose	Designed for statistical purposes	Designed for other purposes	Organic (not designed) or designed for other purposes
Byproducts	Lower potential for by-products	Higher potential for by-products	Higher potential for by-products
Methods	Classical statistical methods available	Classical statistical methods available, usually depending on the specific data	Classical statistical methods not always available
Structure	Structured	A certain level of data structure, depending on the objective of data collection	A certain level of data structure, depending on the source of information
Comparability	Weaker comparability between countries	Weaker comparability between countries	Potentially greater comparability between countries
Representativeness	Representativeness and coverage known by design	Representativeness and coverage often known	Representativeness and coverage difficult to assess
Bias	Not biased	Possibly biased	Unknown and possibly biased
Error	Typical types of errors (sampling and non-sampling errors)	Typical types of errors (non-sampling errors, e.g., missing data, reporting errors and outliers)	Bother typical errors (e.g., missing data, reporting errors and outliers) although possibily less frequently occurring, and new types of errors
Persistence	Persistent	Possibly less persistent	Less persistent
Volume	Manageable volume	Manageable volume	Huge volume
Timeliness	Slower	Potentially faster	Potentially must faster
Cost	Expensive	Inexpensive	Potentially inexpensive
Burden	High burden	No incremental burden	No incremental burden

Source: Rob Kitchin, The opportunities, challenges and risks of big data for official statistics, National University of Ireland, Maynooth, (2015). https://www.researchgate.net/publication/282421109_The_opportunities_challenges_and_risks_of_big_data_for_official_statistics, pp. 9.

… and the Response of the Federal Administration

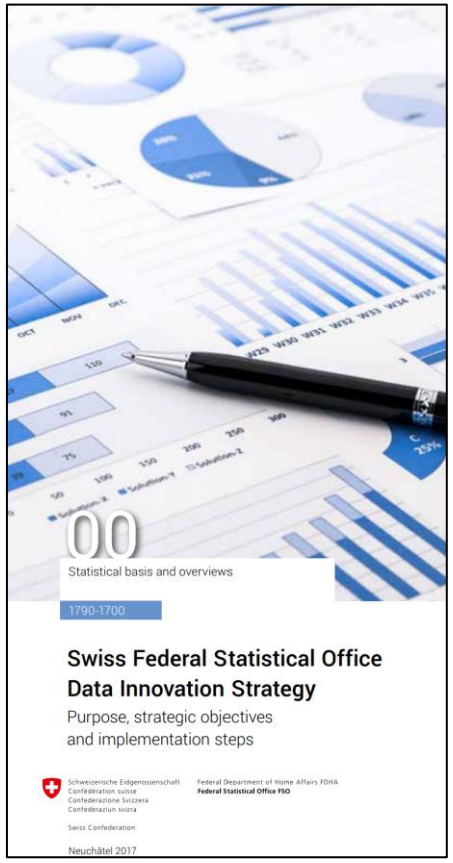
2015

Internal FSO working group on big data

Data Innovation Strategy
Pilot projects

2017

The FSO's Data Innovation Strategy



Content of the Data Innovation Strategy (2017 – 2020)

1. Definining FSO's data typology (primary, secondary)
2. Prioritizing pilot projects on already known primary and secondary data
3. Pilot projects were prioritized with the clear focus to generate new statistical insights and publishing them as experimental statistics and then putting them into production
4. No need to make the «buzz» with non-traditional (big data) sources

The main goal of the Data Innovation Strategy was to augment and/or complement FSO's official statistics and to increase FSO's expertise with non-traditional statistical methods (e.g. from data science, machine learning and/or AI) with already well known primary and identifiable secondary data.

The Pilot Projects in 2017

Area statistics deep learning
Use AI to automate (even partially) visual interpretation of aerial images in order to detect and classify changes.

Data validation with machine learning
Extend and speed up administrative data validation, and improve data quality.

Automation of NACE coding
Automation of coding of economic activity of enterprises using ML methods on already available FSO data (surveys, commercial register, etc.).

EXPERIMENTAL STATISTICS

Machine learning social security system
Clustering of typical prospective trajectories patterns concerning the receipt of benefits in the Swiss social security system.

EXPERIMENTAL STATISTICS

Small area estimation methods for the job statistics
Produce reliable estimates of the total number of jobs and FTEs for Swiss cantons, major towns and NACE (NOGA) levels that were not anticipated in the sampling plan.

… and the Response of the Federal Administration

2015

Internal FSO working group on big data

2017

Data Innovation Strategy
Pilot projects

2020

Data Science Competence Center (DSCC)

The New FSO since January 1st 2021

FSO's Director General

4 Divisions of Statistical Production

1

Statistics

Interoperability and Registers Division

2

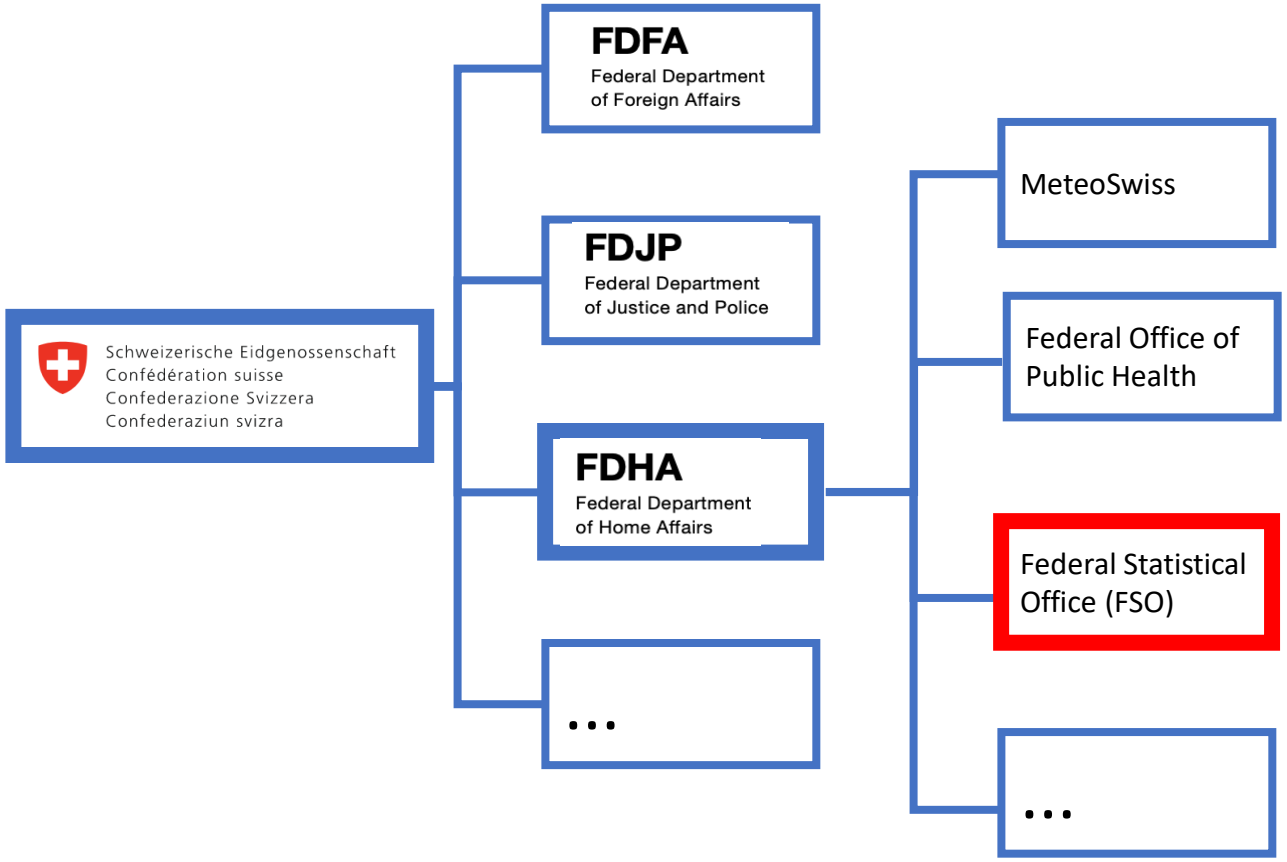
Data Manage-ment

Data Science, AI and Statistical Methods Division

3

Data Science

A Competence Center for The Entire Swiss Public Sector



Structure of the Federal Administration

- 7 government departments, close to 90 federal agencies
- 38,000 employees

Topics covered by the Federal Administration

- Public transport
- Energy
- Migration
- Finance
- Health
- Social security
- Justice
- Telecommunication
- …

Coordination with regional administrations

- 26 cantons
- 2200 communes

Data Science Under the Rule of Law

DSSC's Code of Conduct

Privacy

- Data Governance
- Data Privacy
- Data Security

Transparency

- Explainability
- Open Source
- Reproducibility

Ethics

- Non-discrimination
- Objectivity
- Representativeness

Rule of Law & Public Trust

Trust, Ethics, and Transparency

21.4645 POSTULAT

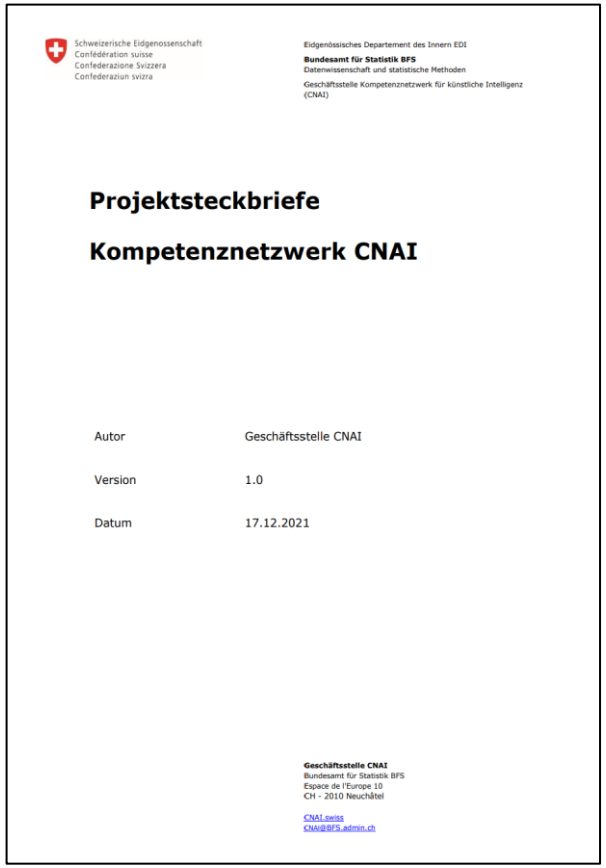
Für mehr Neutralität der Algorithmen sorgen

21.3239 INTERPELLATION

Braucht es eine unabhängige Kontrollinstanz für Algorithmen (Anwendungen künstlicher Intelligenz)?

Project database

- The DSCC maintains a list of Data Science and AI-relevant projects in the Federal Administration to create an overview of possible topics and methods and also to facilitate the exchange of experience.
- The project overview also creates transparency about Data Science and AI projects available in the Federal Administration.



Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement des Innern EDI
Bundesamt für Statistik BFS
Datenwissenschaft und statistische Methoden
Geschäftsstelle Kompetenznetzwerk für künstliche Intelligenz (CNAI)

Projektsteckbriefe

Kompetenznetzwerk CNAI

Autor	Geschäftsstelle CNAI
Version	1.0
Datum	17.12.2021

Geschäftsstelle CNAI
Bundesamt für Statistik BFS
Espace de l'Europe 10
CH - 2010 Neuchâtel

CNAI.swiss
CNAI@BFS.admin.ch

Provider of DSaaS for the Swiss Public Sector

Data Science as a Service (DSaaS)

Consulting	Methodological support	Project execution	Training
Advisory and appraisal services on strategic, tactical and operational application of innovative data science methods and procedures.	Methodological support (coaching and on the job training) for data science projects within public administrations.	Complete realization; from problem formulation to Minimal Viable Product of data science projects.	Application-oriented training (off the job training) on data science methods, techniques, practices, technologies and tools.

Provider of DSaaS for the Swiss Public Sector

Data science is a "concept to **unify statistics, data analysis, informatics**, and their related methods" in order to "**understand and analyze** actual *phenomena*" with **data**.

For more details see the DSCC website and the CNAI terminology.

Data Engineer

- High performance computing
- Data security
- Data bases
- Big data
- Cluster computing
- User Interface
- GPU / TPU

Data Scientist

- Statistics
- Signal processing
- Optimization
- Machine learning & AI
- Data Privacy
- Robotics / Automation
- Data visualizations
- Mathematics
- Computer vision

Domain Expertise

Evidence-based Policy Making

Data, combined with expertise, can be leveraged to improve and expediate the policy-making cycle.

Risk: relevant data and expertise are not used for policy-making, leading to ineffective / harmful policies being either adopted or allowed to stay in place.

Goal: support policy makers by providing policy evaluation and assessment of the impact of alternative policy scenarios.

The DSCC offers expertise in:

- Causal inference
- Risk Assessment
- Leveraging big data from non-traditional sources

The policy-making cycle

Privacy Preserving Data Science

Swiss citizens trust the Confederation with sensitive personal data.

Risk: sensitive data, such as health records or income status, are leaked to third parties, e.g., insurance agencies or banks.

Goal: maintain citizens' trust by ensuring data privacy.

The DSCC offers expertise in:

- machine learning algorithms with encrypted data,
- decentralized data analysis,
- protection measures against re-identification attacks.

Applications of Data Science and AI in the public sector

DSCC's Code of Conduct

Evidence-based policy

- Estimating and assessing the impact of policies
- Improved weather forecasts
- Prediction of the evolution of trends
- Higher resolution statistics
- Real time statistics and monitoring

AI/ML assisted tools

- Anomaly and fraud detection
- Chatbots
- Data matching
- Data preparation
- Data plausibility checks
- Satellite image classification

Safe and scalable computing and storage infrastructures

Roads Office (ASTRA – OFROU)

Automated highway traffic monitoring

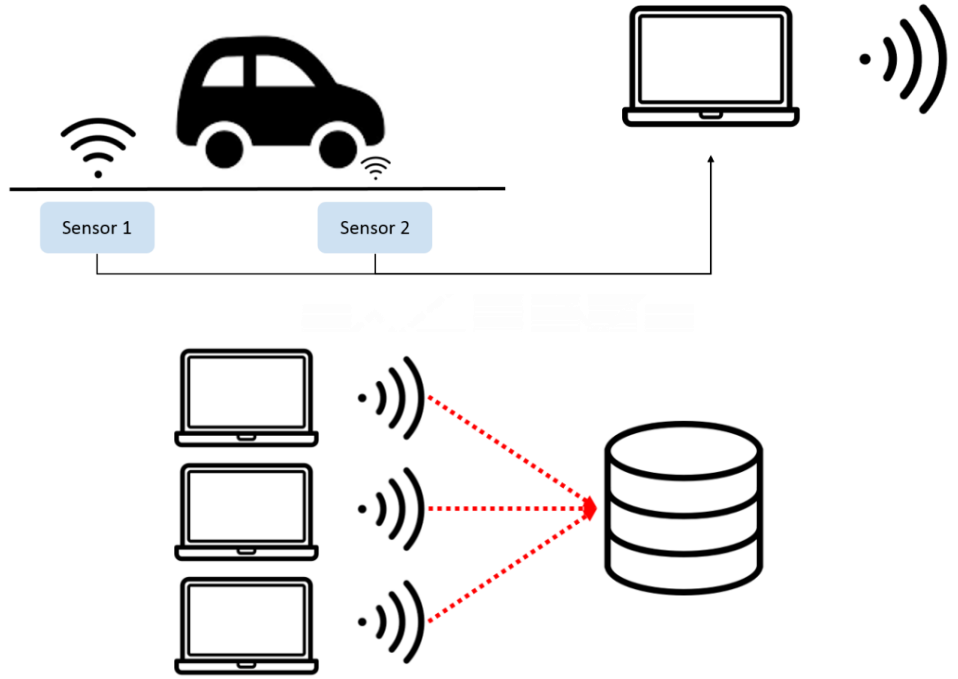
About 500 sensors on Swiss highways monitor the volume and class of vehicles in real time.

Goal: automatic detection of measurement errors and missing data reconstruction.

Current practice: data sanitation performed by human experts.

New solution: development of a toolbox in R for automatic:

- anomaly detection,
- data reconstruction,
- in depth statistics and visualizations.



Poverty Index Estimation (BFS – OFS)

Goal: Estimate poverty indicators for all cantons

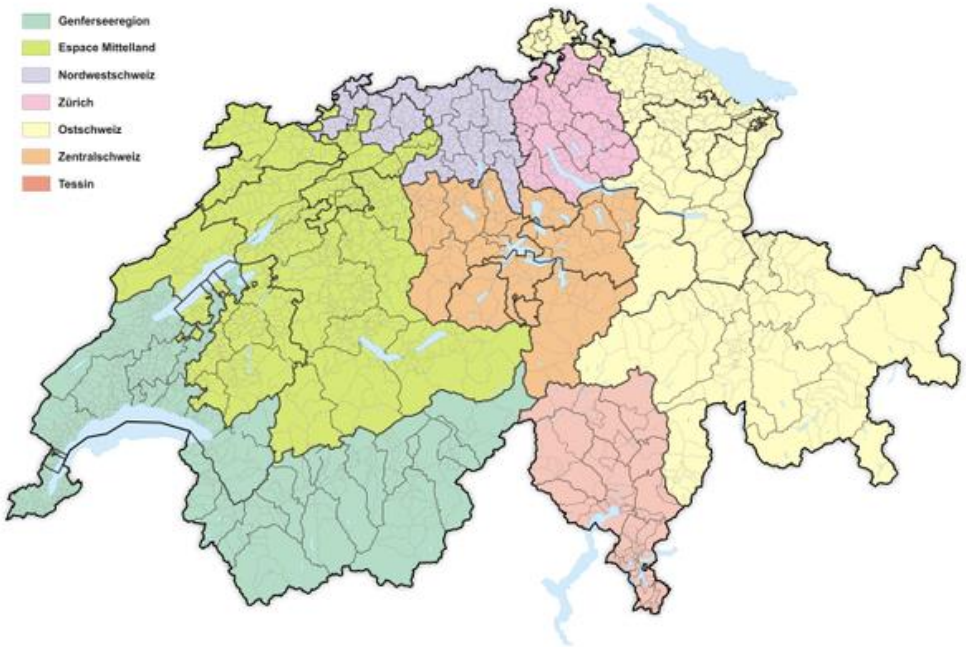
Current practice: Phone questionnaire with “only” 17’000 people.

- Robust only for ‘Large Regions’.
- Already quite expensive.

➡ Federal Council requests cantonal level.

New solution: Machine Learning algorithms to improve estimation

- Random Forest, Grading Boosting, Neural Network, Ensemble



Community “Data Science for Public Good” - Federal Level

Data Science for Public Good seminars

The Data Science Competence Center (DSCC) at FSO teams up with the Swiss Data Science Center (SDSC) to organize a series of seminars for the Swiss public administration and its academic partners. The Data Science for public good seminar series aims to focus on topics related to data and computer science. We invite Swiss and international speakers to present their latest innovative research with respect to the values of the Rule of Law.

- Data governance
- Data protection
- Ethical data analysis
- Explainability of algorithms
- Reproducibility
- Respect for the principles of non-discrimination
- Security of information
- Transparency

Community of Practice (CoP) – Federal Level

Special interest groups

An interest group enables researchers to gather around shared areas of interest in data science and AI with the aim to share ideas and knowledge and spark new ideas for collaborations and innovative projects. The Data Science Competence Center (DSCC) at FSO organizes communities of practice around a shared area of interest in data science and AI. The areas of interest can include but are not limited to the following methodological approaches:

- Causal analysis (causal inference)
- Evidence-based policy making
- Machine Learning
- Algorithmic design
- Privacy-preserving Data Science
- Statistical design, analysis, and modeling
- Computer Vision and Geodata science
- Infrastructure

Clusters and Topics of Collaborations with Universities

Strategy

Capacity Building

Community

Mission-Driven Research

Knowledge Infrastructure (FAIR)

… and the Response of the Federal Administration

2015

Internal FSO working group on big data

2017

Data Innovation Strategy
Pilot projects

2020

Data Science Competence Center (DSCC)

2021

Competence Network for Artificial Intelligence (CNAI)

… and the Response of the Federal Administration

2015

Internal FSO working group on big data

2017

Data Innovation Strategy
Pilot projects

2020

Data Science Competence Center (DSCC)

2021

Competence Network for Artificial Intelligence (CNAI)

2023 - …

Federal Data Science Strategy (DSStB)

Data Science Along The Process of Policy Making

Source: The five stages of the policymaking process or cycle (from Howlett & Giest, 2015)

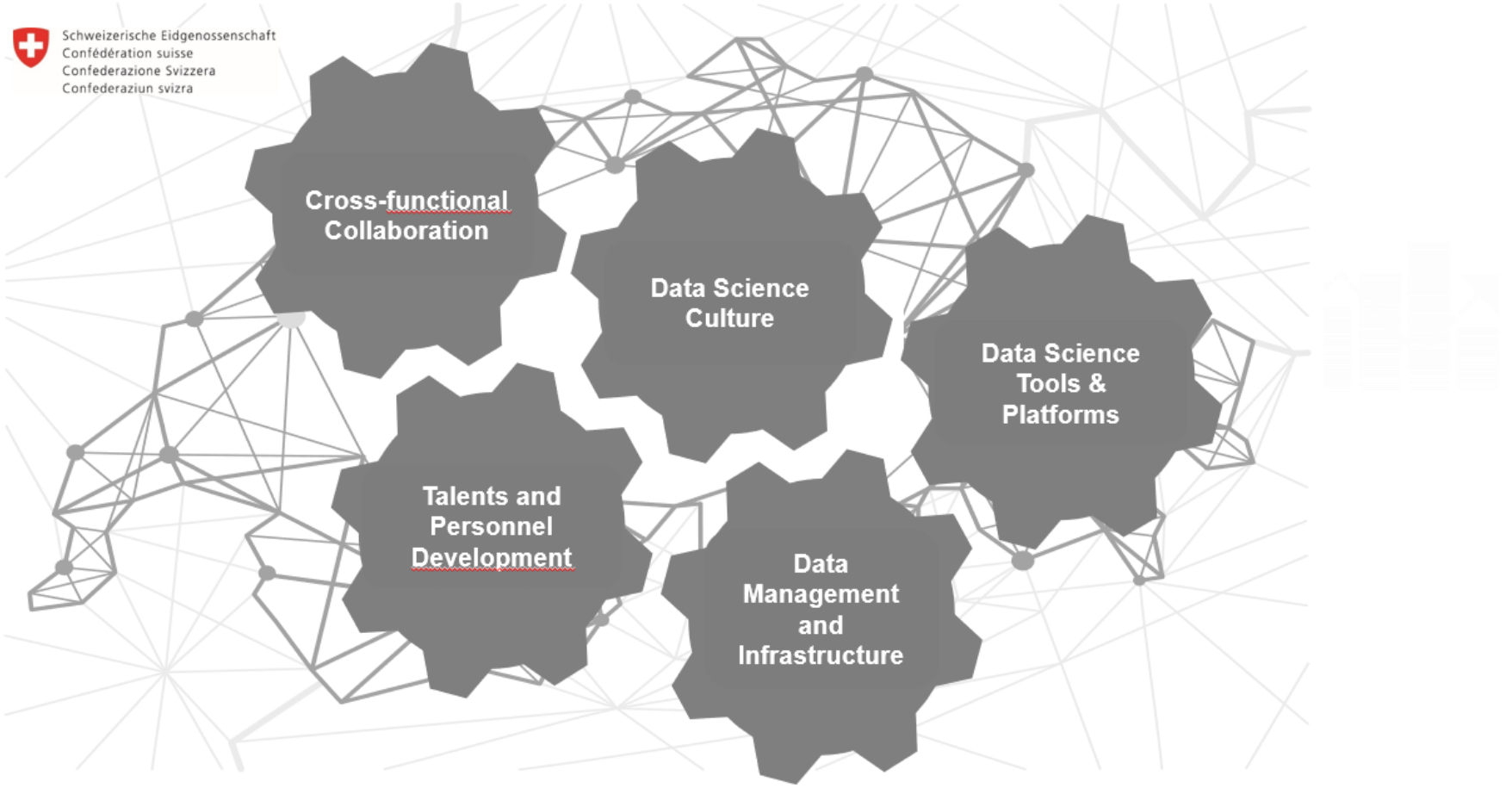
Public Policies

- Public transport
- Energy
- Migration
- Finance
- Education
- Health
- Social security
- Justice
- Telecommunication
- Customs
- …

The Needs to Scale Up at The Federal Level…

- **Data Science and AI Literacy**
- **Trust, Ethics, and Transparency**
- **Training, Competencies, and Capacity Development**
- **Community**
- **Data**
- **Algorithms**
- **Technology / Infrastructure / Security**
- **Governance**

… And to create a Data Science Ecosystem at The Federal Level



Agenda

1. Introduction
2. From an FSO's Working Group to a Federal Strategy on Data Science
3. National Data Management
4. Conclusion

National Data Management

ongoing

Political Support

Once Only – Principle
New strategic initiatives
within Swiss administration

ongoing

Data Harmonization

Domain concepts
Standardization

established

Collaboration Model

Data Ownership
Data Stewardship
Different government levels

2019 — 2020 — 2021 — 2022

established

“Bundesrats-beschluss”

New federal program:
“National Data Management (“NaDB”)
Orders and funding

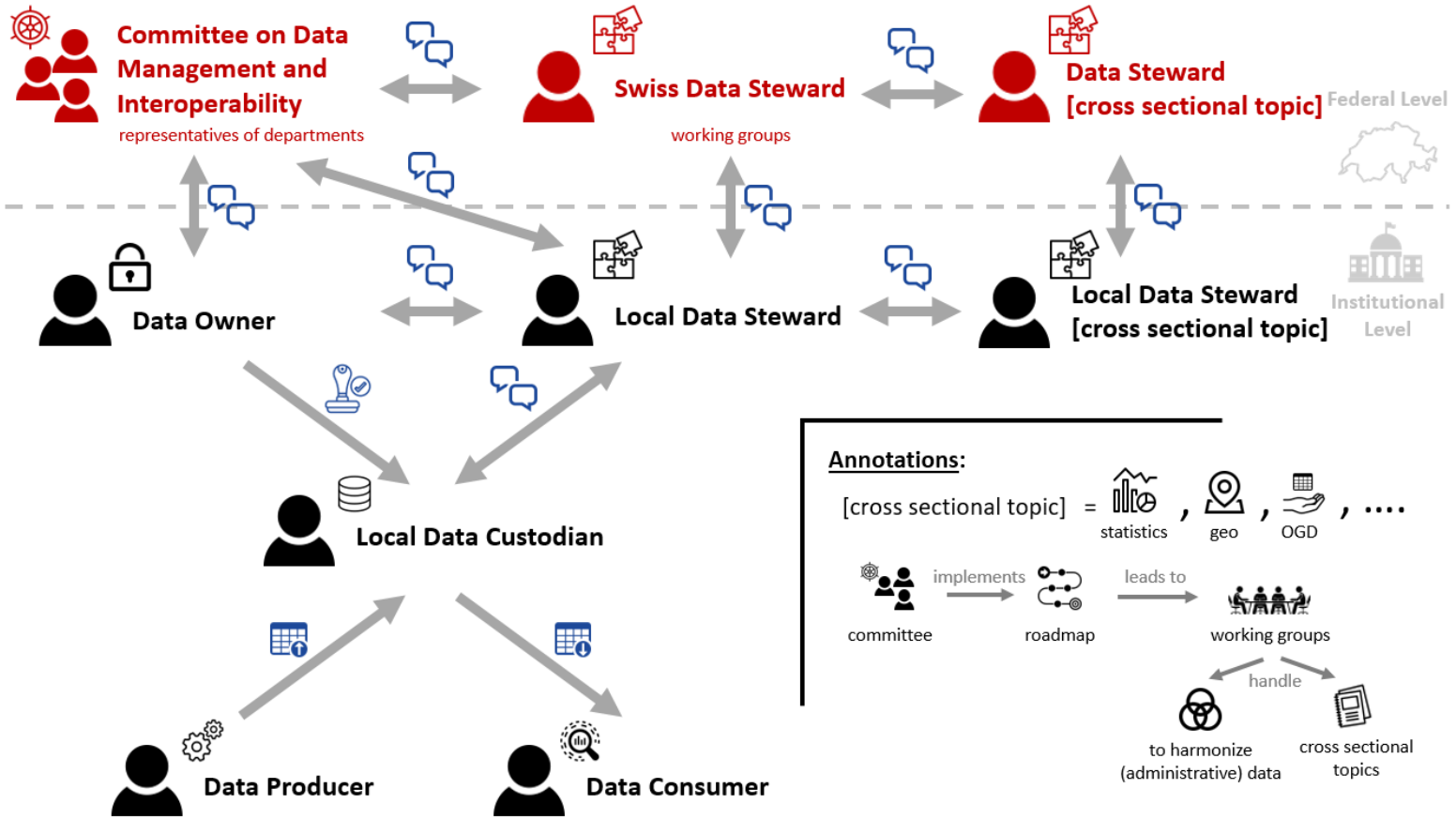
ongoing

Implementation

Interoperability
Metadata
API directory

Collaboration Model

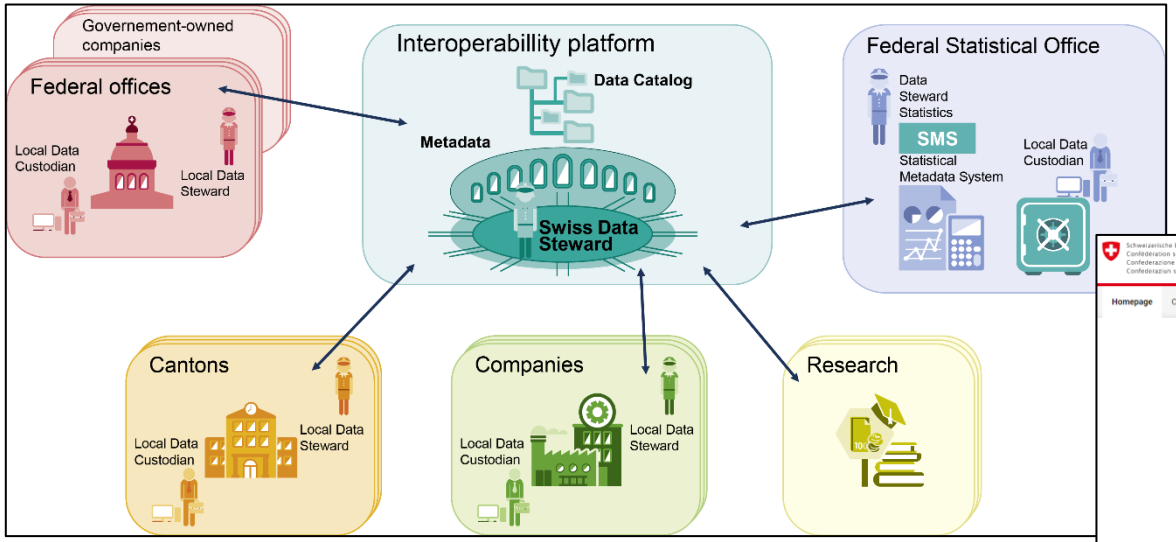
Data Stewardship Model (Governance)

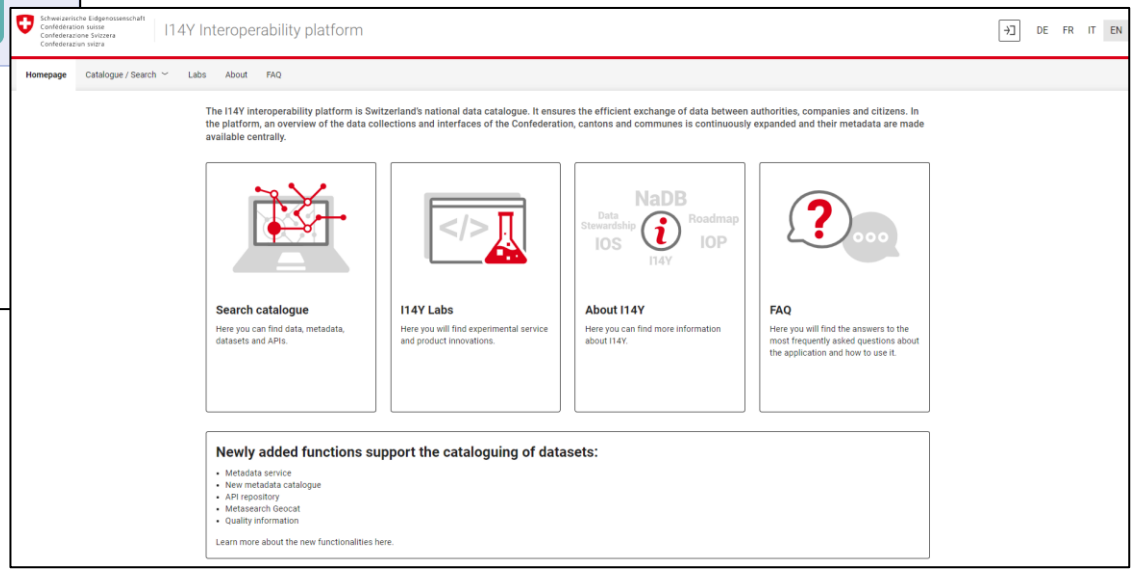


The FSO's Director General is the Swiss Data Steward.

Implementation

Interoperability Platform (Metadata Catalogue)





https://www.i14y.admin.ch

Collaboration Model

Implementation

Data Harmonization

Purposes, Scope, Access, Harmonization of Data Spaces

Ongoing

Finan… …culture Data

Energy Data

Health Data

…

Transport Data

Current Access to Data

Official Statisticians at FSO based on the federal statistics law…

… can launch surveys to obtain information that is not already available in public administrations. To do so, the FSO must update an ordinance that describes the number and purpose of the surveys it conducts.

… can access all administrative data available within the federal administration. This is not always the case for the administrative data of the cantons e.g. tax data.

… can access data held by private companies for statistical purposes only if they agree (there is no legal basis that would allow the FSO to force a company to make its data available).

… can express the wish to modify its legal bases to carry out Web Scraping.

The FSO, based on the federal statistics law…

… is the only federal agency authorized to match data for statistical purposes.

Current Access to Data - #2

Data Science Competence Center as Service Provider …

… can work with all types of data legally (legal basis) in possession of the principal (client) e.g. tax data, health data, transport data, energy data, mobile phone data, weather data, geospatial data, etc.

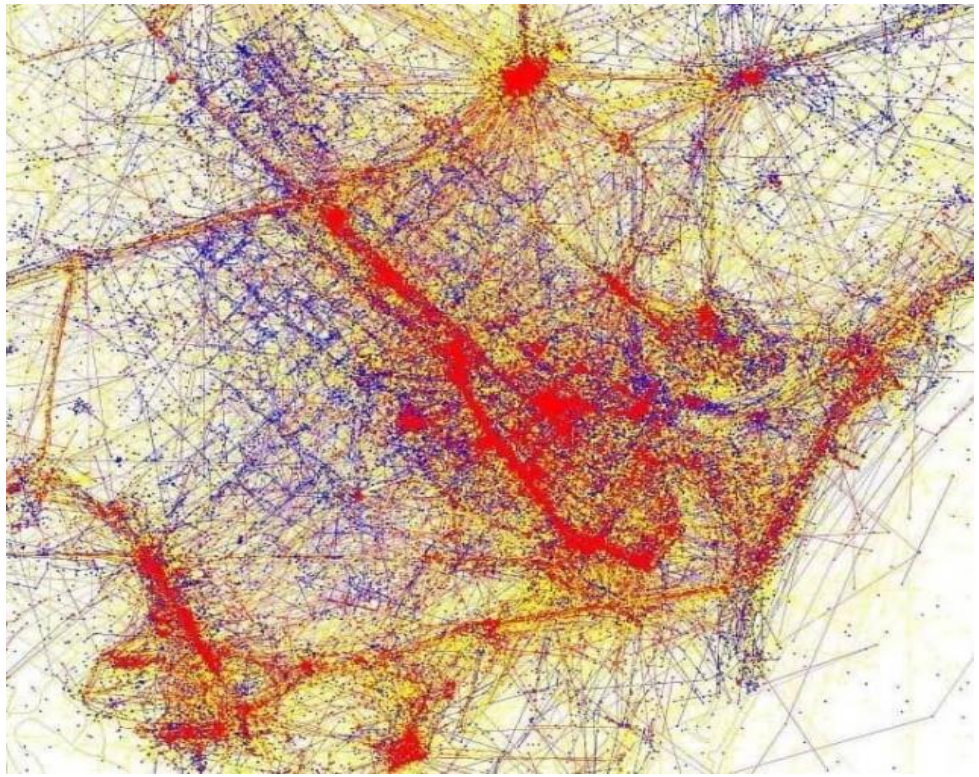
... may match all types of data for statistical or administrative purposes, provided that the principal has the appropriate legal basis e.g. fraud detection.

… never publishes the results of its work but always performs work as an agent for a principal.

Agenda

1. Introduction
2. From an FSO's Working Group to a Federal Strategy on Data Science
3. National Data Management
4. Conclusion

Public Trust in the Algorithmic Age



- Fit for the algorithmic age: understanding new issues and learning from more and more data/working out reliable information together.