DATAWEEK 24

JOIN.LEARN.SHARE.GET VALUE

Elevating Data Quality: A Paradigm Shift for Data Spaces and AI Needs

12/03/2024 12:00–13:00 (CET)

Daniel Alonso, Håkan Burden, Susanne Stenberg, Mark Dietrich, Kuldar Aas, Tobias Guggenberger

Funded by the European Union

The Data Spaces Support Centre receives funding from the European Union Digital Europe Programme under grant agreement n° 101083412

DSBA ■ BDV BIG DATA VALUE ASSOCIATION FIWARE FOUNDATION gaia-x INTERNATIONAL DATA SPACES ASSOCIATION

DATA SPACES SUPPORT CENTRE

Data Spaces Symposium

Unite. Innovate. Adopt.

Darmstadtium | Frankfurt region

Objectives of the session

- To approach data quality from a new perspective, addressing specific requirements coming from new applications and in view of AI Act and other regulations
- To consider data quality as a dynamic feature, different metrics to respond to final application and use cases needs (fit for purpose)
- Data spaces to support data quality, as a controlled and trusted environment that ensures governance, provenance, traceability, description, business needs, …

Data Spaces Symposium

Unite. Innovate. Adopt.

DATAWEEK24
JOIN.LEARN.SHARE.GET VALUE

Elevating Data Quality: A Paradigm Shift for Data Spaces and AI Needs

12 March 2024 | 12:00 - 13:00

Hakan Burden
RISE

Susanne Stenberg
RISE

Mark Dietrich
EGI

Kuldar Aas
Ministry of Economic Affairs and Communications for Estonia

Tobias Guggenberger
Fraunhofer

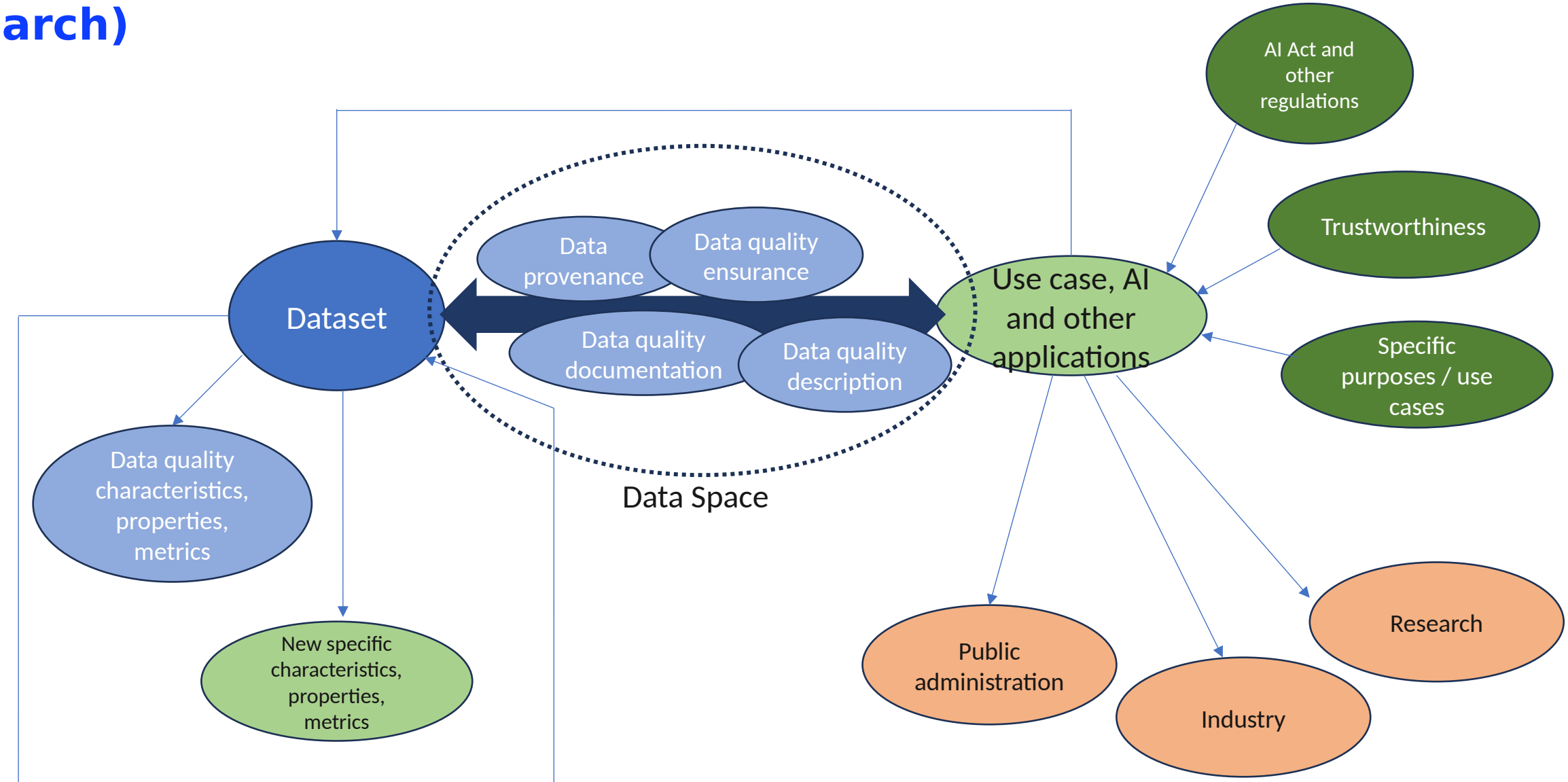
Patrick van der Smagt
Volkswagen AG / etami

Daniel Alonso
BDVA

Agenda of the session

Speaker	
Daniel Alonso	Intro and setting-up the scene
Hakan Burden / Susanne Stenberg	Data quality from a policy perspective – from law to standard. Concrete examples from the AI Act, EHDS and emerging standards. Examples of policies related to stakeholders involved in the assessment of data quality. Implications for policy development .
Mark Dietrich	Quality from data product paradigm. Importance of semantics. View from research, EOSC, GREAT.
Kuldar Ass	Estonian data infrastructure. New layers for secure data sharing. Fit for purposes, use cases
Tobias Guggenberger	Data quality from a business perspective. Impact of data quality on costs, resources and KPIs
All	Q&A

BDVA discussion paper (to be published by the end of March)



DATAWEEK 24

JOIN.LEARN.SHARE.GET VALUE

Hakan Burden / Susanne Stenberg

RISE

Funded by the European Union

The Data Spaces Support Centre receives funding from the European Union Digital Europe Programme under grant agreement n° 101083412

DSBA ■ BDV BIG DATA VALUE ASSOCIATION FIWARE FOUNDATION gaia-x INTERNATIONAL DATA SPACES ASSOCIATION

DATA SPACES SUPPORT CENTRE

Data Spaces Symposium

Unite. Innovate. Adopt.

A policy perspective on data quality

Håkan Burden & Susanne Stenberg

What signifies data quality?

And who should care?

RI.SE

Fit for purpose: Data quality implies human judgement

European Health Data Space:

”‘data quality’ means the degree to which the elements of electronic health data are assessed and considered **suitable for their intended** primary and secondary **use**”

AI Act:

”Training, validation and testing data sets [for high risk AI systems] shall be relevant, sufficiently representative, and to the best extent possible, free of errors and complete **in view of the intended purpose**”

Fit for purpose?

Standards help us take the right decisions

RI. SE

Data quality: ISO 8000-150:2022

Syntactic quality
Well-formed, right type, ...

Semantic quality
Interpretation of labels and values

Pragmatic quality
Is the data fit for purpose?

job_application = applicant:”Jane Doe”, ”123456”, job:”course assistant”, cv:””}

job_application = {applicant:”Jane Doe”, date:010203, job:”teacher”, cv:”cv.pdf”}

job_application = {applicant:”Jane Doe”, date:010203, job:”chef”, cv:”cv.pdf”}

RI.
SE

Data quality: ISO 5259-2 for ML

AI Act: data used for training, testing and validation should be ”relevant [...] and complete”

Completeness:

Ratio of data items of no presence of null data values in a dataset

Ratio of unlabelled or incompletely labelled samples in a dataset

job_application = {applicant:”John Doe”, date:010203, job:”chef”, cv:”cv.pdf”}

 Syntactic quality

Relevance:

Ratio of features or records in the dataset that are relevant to the given context

→Pragmatic quality

→Is ”chef” fit for purpose?
Requires the applicant’s intention, which might not be inferred by provenance nor the metamodel

RI.
SE

By Göteborgs konstmuseum (Gothenburg Museum of Art). Photo: Hossein Sehatlou, CC BY 4.0, https://commons.wikimedia.org/w/index.php?curid=112974394

Data quality according to whom?

Points of view according to the AI Act:

As a provider in your market assessment

As an AI competent authority in the regulatory sandbox

As a notified body in the compliance procedure

As a deployer fulfilling user obligations

As an AI competent authority conducting market surveillance

RI. SE

Our conclusions

Data quality is assessed on a case-by-case basis and implies human judgement

There are currently gaps in when data quality is a legal requirement, but also how it should be applied across domains and applications

There are opportunities to define the implementation of data quality (i.e. relevant threshold levels – quality metrics) and applications (what is good enough for my market?)

This opens for differing governance and
data quality models across and within

Thanks!

hakan.burden@ri.se
susanne.stenberg@ri.se

All pictures taken from https://commons.wikimedia.org/wiki/Category:Media_contributed_by_Göteborgs_konstmuseum

DATAWEEK 24

JOIN.LEARN.SHARE.GET VALUE

Mark Dietrich

EGI

Funded by the European Union

The Data Spaces Support Centre receives funding from the European Union Digital Europe Programme under grant agreement n° 101083412

DSBA ■ BDV BIG DATA VALUE ASSOCIATION FIWARE FOUNDATION gaia-x INTERNATIONAL DATA SPACES ASSOCIATION

DATA SPACES SUPPORT CENTRE

Data Spaces Symposium

Unite. Innovate. Adopt.

Elevating Data Quality:
A Paradigm Shift for Data Spaces and AI Needs

DataWeek 2024

Darmstadt, Germany

Mark Dietrich, Senior Advisor, EGI.eu

Elevating Data Quality

Is Quality In the Eye of the Beholder?

EGI.eu 2024

Breaking Quality Down

Some Measures of Quality

Quality of Single Data Points

- Declared Precision, Accuracy
- Declared Data Collection/Processing Standards

Dataset Statistics

- Missing, outdated data
- Coding errors
- Inter-/Intra Dataset Consistency
- Apparent Bias
- Representativeness
- Balanced

“Subjective” (wrt Use Case)

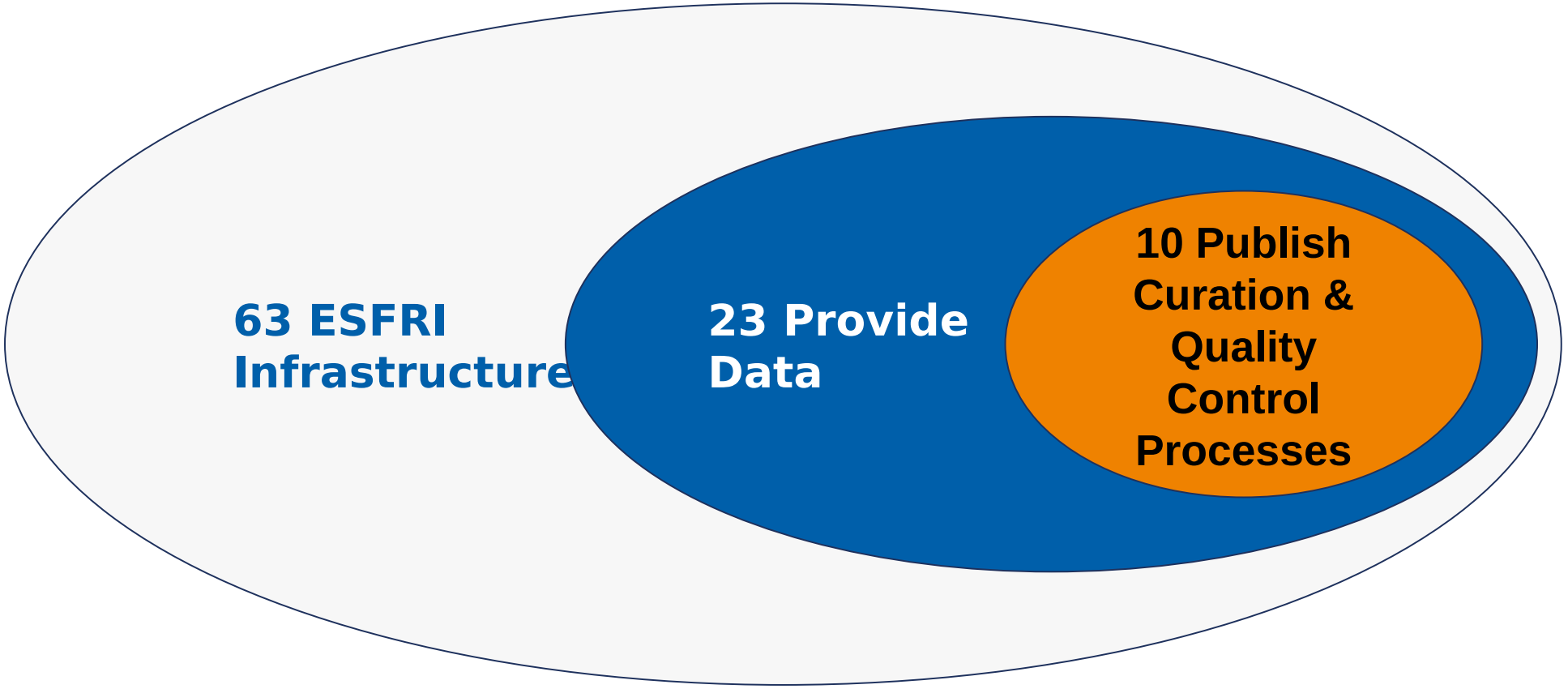
Metadata

- Incomplete, erroneous
- Vocabulary errors
- FAIR metrics

In EU, only ~30% of data is FAIR

Data Curation/QC Processes

Sample: European Strategy Forum on Research Infrastructures (ESFRI), 2021 Roadmap



What Does “Fit for Purpose” Mean?

Research, Public Sector

Purpose:

- *Advancement of Knowledge*
- *Scientific Method*
- *Reproducibility*

Quality required depends on specific study

- *Mismatches caught in peer review*

Sensitive Data

Purpose: must be declared every time you want to access sensitive data.

Fit for Purpose

- *Ethics Review*
- *Original reason for data collection must be compatible with proposed use*
- *Even if compatible, quality may not be sufficient*

Business, Industry

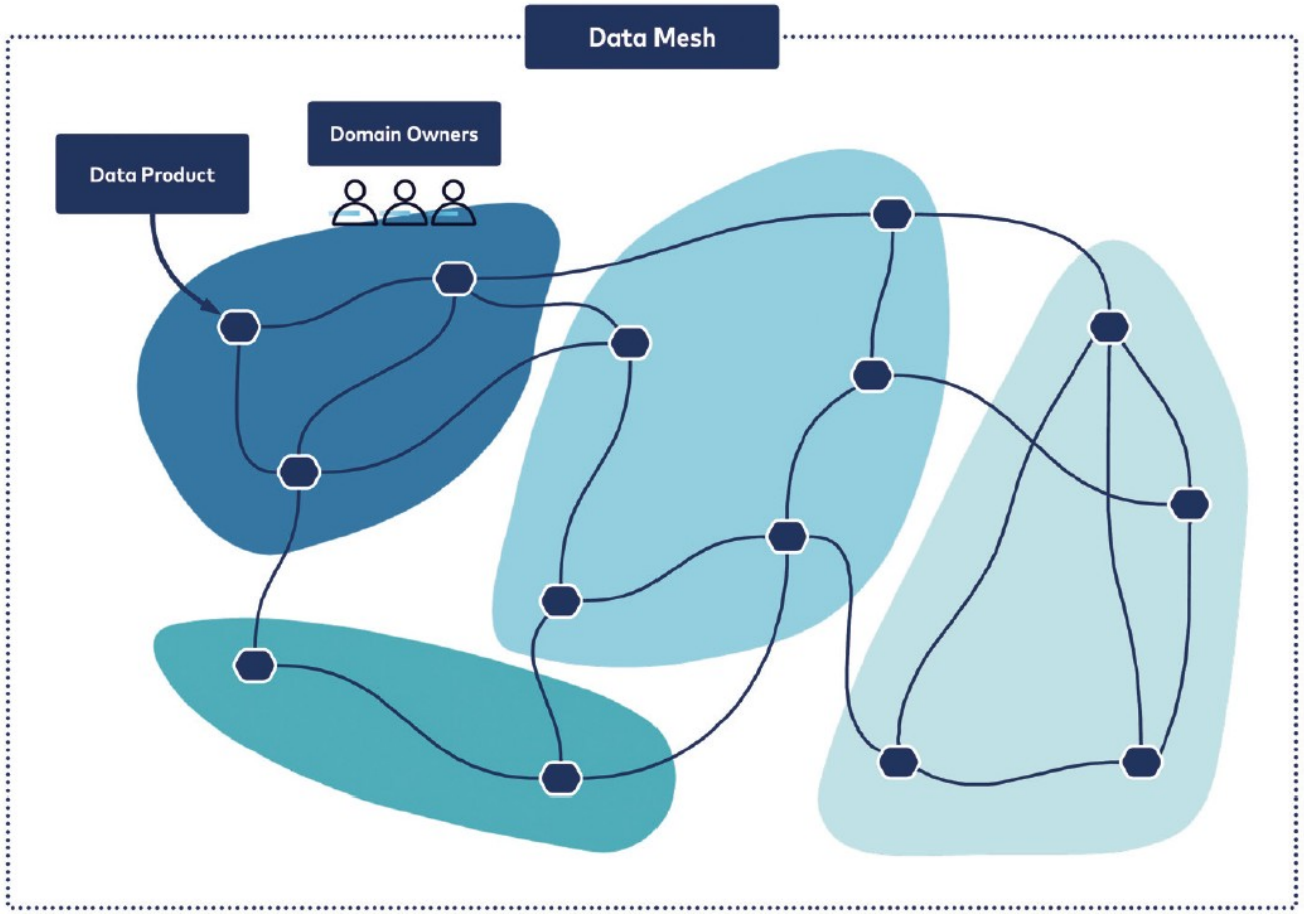
Purpose motivates the first use case

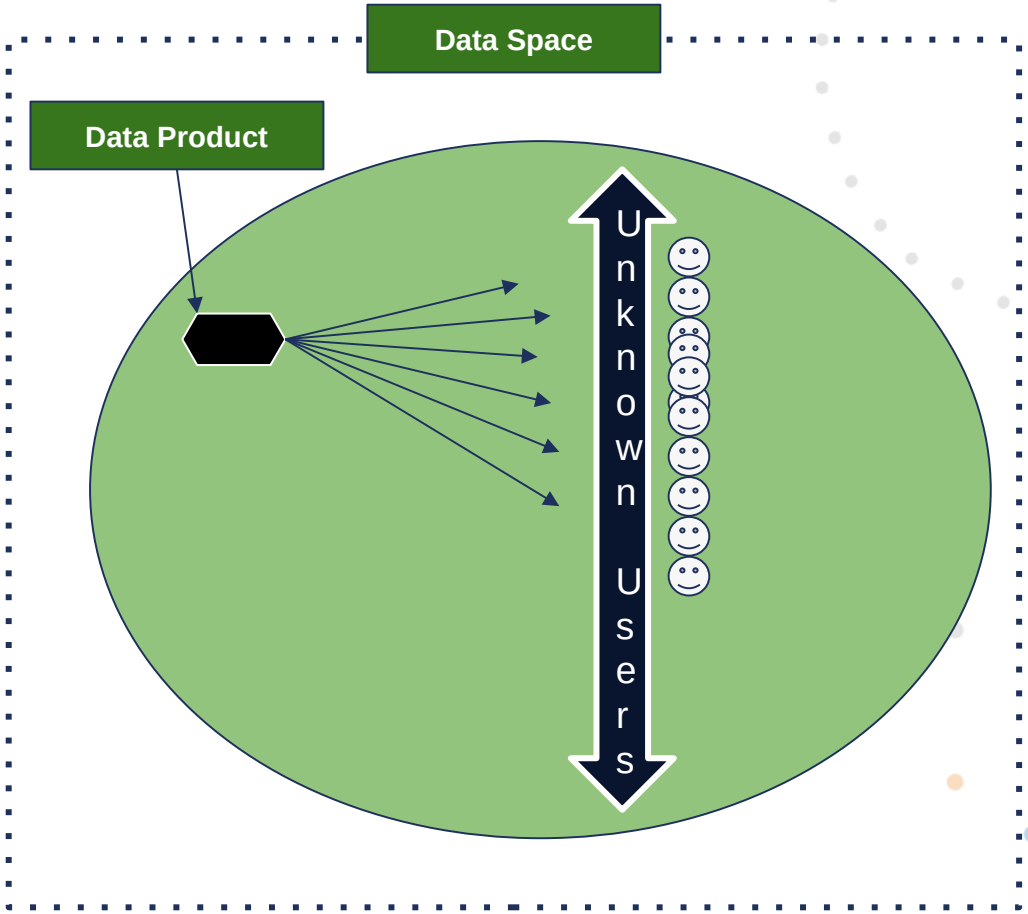
- *Data sovereignty enables later use cases*

Fit for Purpose?:

- *No review: ends justify the means*
- *Competitive advantage*
- *Bias, Opacity*
- *Disinformation, collusion?*

How Can We Operate at Scale?





Source: Adam Bellemare, “Practical Data Mesh: *Building Decentralized Data Architectures with Event Streams”*, 2022 Confluent Inc.

Conclusion

Quality CAN be defined and measured along multiple dimensions

- **For individual data points**
- **Across and between datasets**
- **For the metadata that describes the dataset**

But these dimensions are independent of each other (not necessarily correlated)

So Quality LABELS (high, medium, low) CANNOT be universally defined.

- **But labelling schemes might be set up for particular use cases**

Thank you!

Contact: mark.dietrich@egi.eu

About Mark

- Data Spaces Support Centre (DSSC)
- EOSC-Future, EOSC-Focus, EOSC EU Node Implementation
- GREAT (preparatory action for the Green Deal Data Space)
- HealthyCloud.eu (Five Safes)
- EGI-ACE
- HCloud.eu (Horizon Cloud)
  - First analysed *The European Data Strategy,* June 2020
- EUCloudEdgeIoT.eu
- Gaia-X
- Supercomputing, HPC (green computing, economic models)

Icons created by Dinosoftlabs, Eucalyp, Freepik, Icon Mania, Inkubators, Parzival' 1997: https://www.flaticon.com/free-icons/code

Funding provided by:

GREAT: Green Deal Data Space: EC grant agreement 101083927

EGI-ACE: EC grant agreement 101017567

EOSC Focus: EC grant agreement 101058432

EOSC Future: EC grant agreement 101017536

HealthyCloud: EC grant agreement 965345

EGI Support for Data Spaces

Technological Services

- EGI provides several services and technology solutions for the development and implementation of Data Spaces. These solutions, for example, enable the sharing and collaboration of data in a secure and trusted environment.

Policy Development Support

- EGI's policy development support activities help to ensure that data spaces are established and managed in a way that promotes trust, transparency, and accountability.

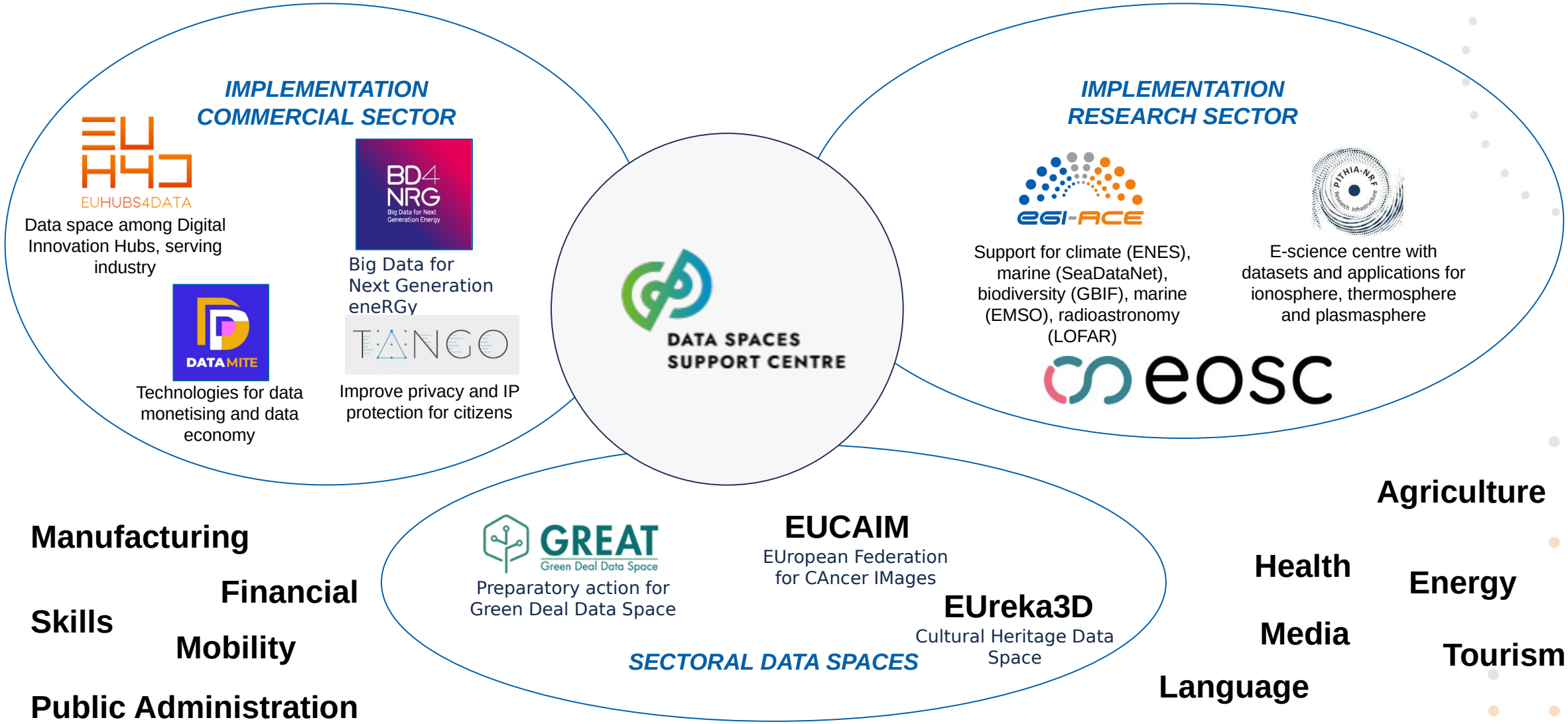
Data Space Landscape Harmonisation

- EGI helps to create synergies between different projects and initiatives, identifying common goals, aligning activities and standards, avoiding duplication of effort and maximizing the impact of funding and resources, thus harmonising the landscape of different Data Space related initiatives.

Contributions to Projects

- EGI leads or contributes to several projects aimed at developing thematic data spaces or contributing to the development of overarching, multidisciplinary data spaces.

Contributions to Data Space Projects



DATAWEEK 24

JOIN.LEARN.SHARE.GET VALUE

Kuldar Ass

Ministry of Economic Affairs and Communications (Estonia)

Funded by the European Union

The Data Spaces Support Centre receives funding from the European Union Digital Europe Programme under grant agreement n° 101083412

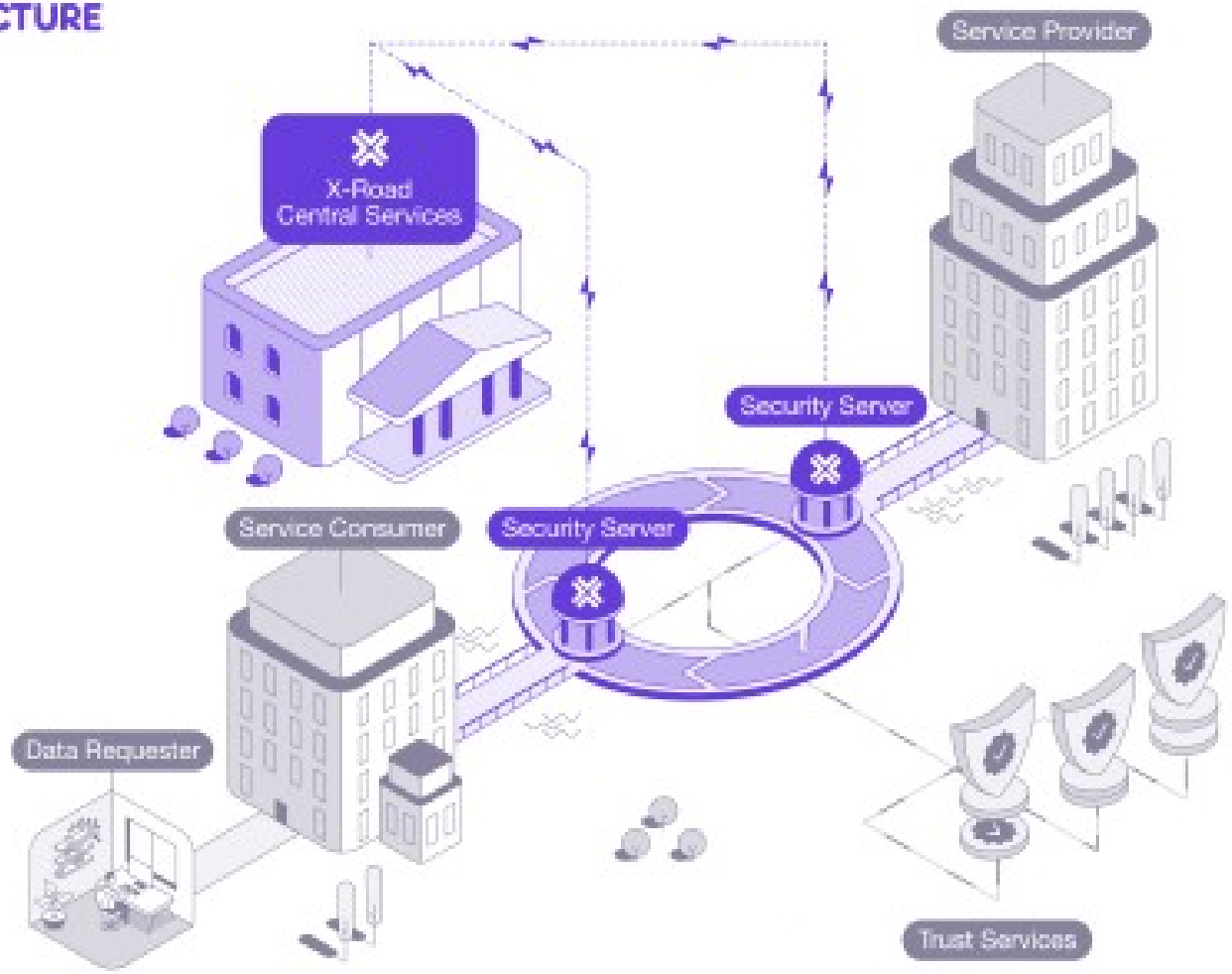
DSBA ■ BDV BIG DATA VALUE ASSOCIATION FIWARE FOUNDATION gaia-x INTERNATIONAL DATA SPACES ASSOCIATION

DATA SPACES SUPPORT CENTRE

Data Spaces Symposium

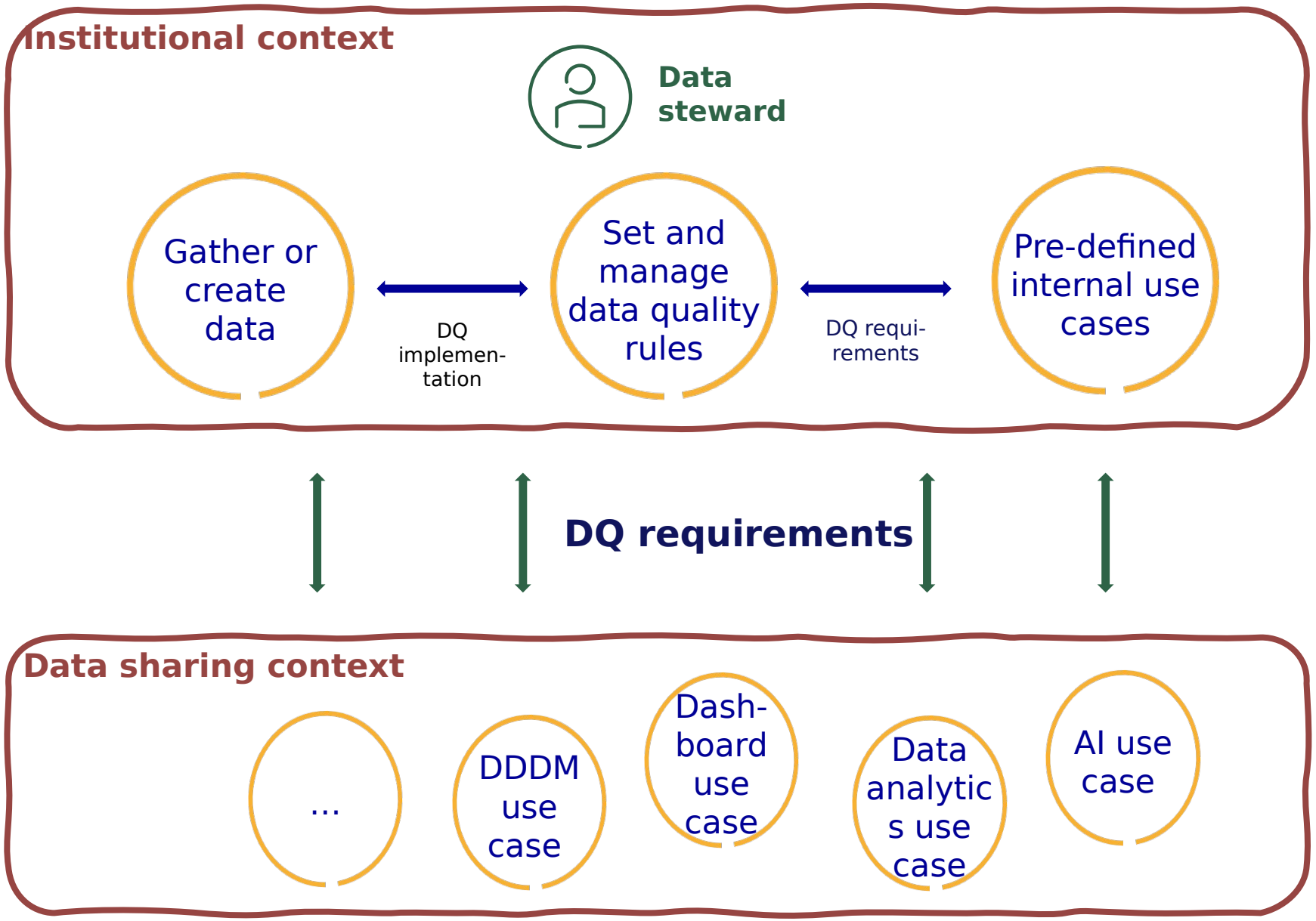
Unite. Innovate. Adopt.

X-ROAD ARCHITECTURE



Practical issues in data reuse

- Low quality of data descriptions → Can I find relevant data?
- Data quality unknown → Is the data actually useful for me?
- Data service quality undefined → Will I get the data when I need it?
- Limited responsiveness to issues → Will my problems get solved in a timely manner?
- Limited knowledge on DP → Will I go to prison for breaching GDPR?


Institutional context
Data steward
Gather or create data
DQ implemen-tation
Set and manage data quality rules
DQ requi-rements
Pre-defined internal use cases
DQ requirements
Data sharing context
...
DDDM use case
Dash-board use case
Data analytic s use case
AI use case

Way forward

- **Making it easier to assess data quality**
  - „public sector agencies to establish formal data quality procedures“ (01.2024)
  - „public sector agencies to publish their data quality procedures“ (2025?)
- **Making it easier to achieve data quality**
  - Providing data quality guidelines and training (continuous)
  - Establishing a support centre on data quality (2022)
  - Setting minimal requirements for assessing the value and ensuring the quality of data assets (2025?)
  - Supporting agencies in acquiring / developing data quality and profiling tools (and publishing according data quality reports) (2025)
- **Last but not least**
  - Let’s join the DSSC SSF and learn from others! (December 2023)

DATAWEEK 24

JOIN.LEARN.SHARE.GET VALUE

Tobias Guggenberger

Fraunhofer ISST

Funded by the European Union

The Data Spaces Support Centre receives funding from the European Union Digital Europe Programme under grant agreement n° 101083412

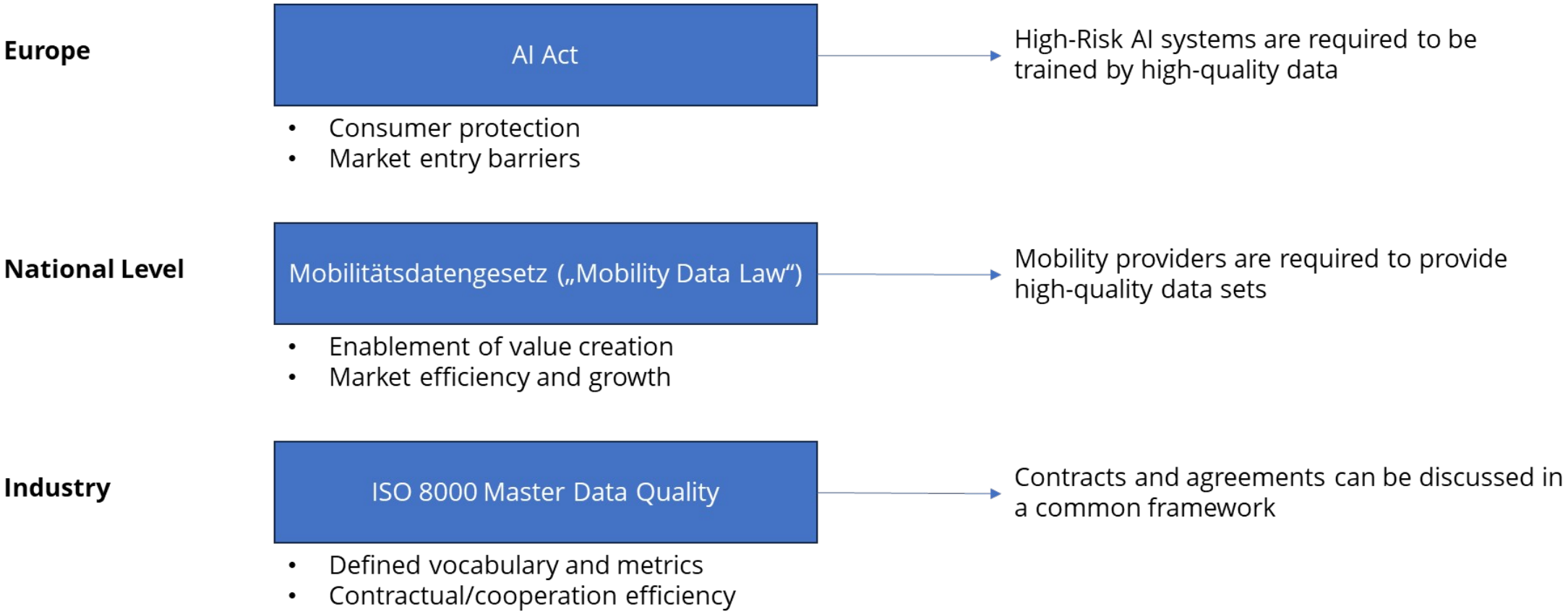
DSBA ■ BDV BIG DATA VALUE ASSOCIATION FIWARE FOUNDATION gaia-x INTERNATIONAL DATA SPACES ASSOCIATION

DATA SPACES SUPPORT CENTRE

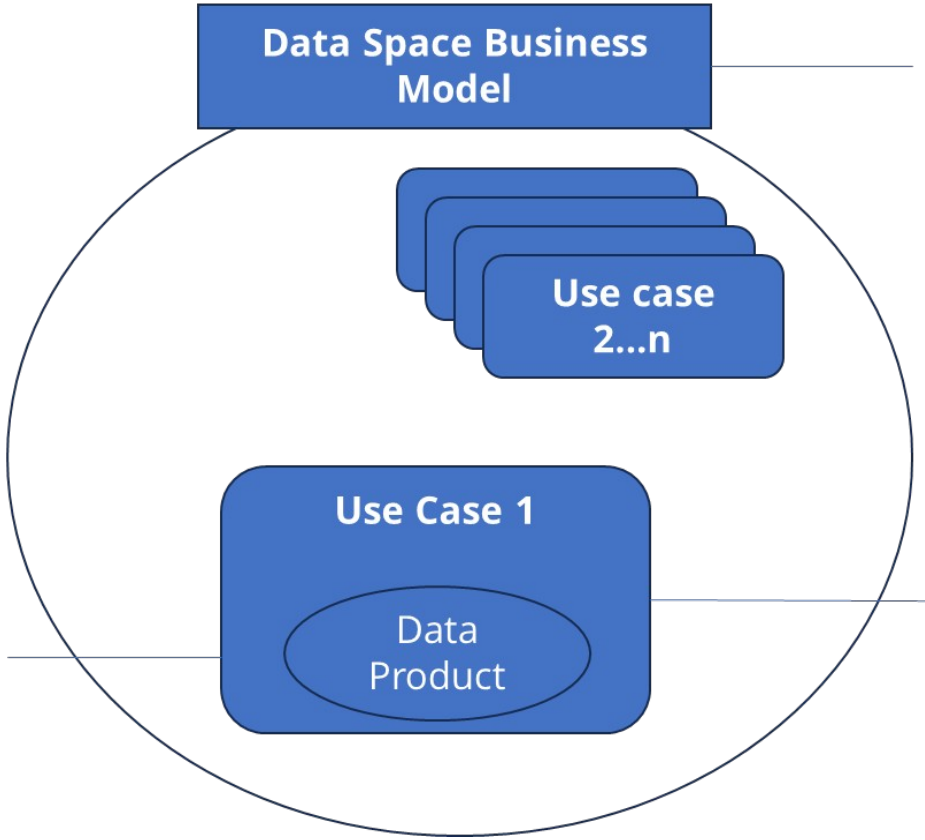
Data Spaces Symposium

Unite. Innovate. Adopt.

Legislation and Business



Levels of Data Quality



Requirement:
Definition of quality metrics, maybe minimum data quality
Why?
Information asymmetry in transactions lead to lower average quality and market efficiency. Trust in the DS

Requirement:
Definition of (use case independend) quality metrics
Why?
Quality is the „fitness for use“ and, thus, is measured against the use case

Requirement:
Definition of quality metrics, minimum data quality, and monitoring, assessment of output data
Why?
Matching the „quality level agreement“ of the outcome. Also regaulatory compliance if AI is involved

DATAWEEK 24

JOIN.LEARN.SHARE.GET VALUE

Thank you!

Funded by the European Union

The Data Spaces Support Centre receives funding from the European Union Digital Europe Programme under grant agreement n° 101083412

DSBA ■ BDV BIG DATA VALUE ASSOCIATION FIWARE FOUNDATION gaia-x INTERNATIONAL DATA SPACES ASSOCIATION

DATA SPACES SUPPORT CENTRE

Data Spaces Symposium

Unite. Innovate. Adopt.

Darmstadtium | Frankfurt region

Data Spaces Symposium

DATAWEEK24
JOIN.LEARN.SHARE.GET VALUE

13:00

Lunch

Feel free to grab your meals and relax.