
AI Ecosystem & Market Analysis
Quick scan of data sharing market to validate blueprint of the NL AIC
Working Group Data Sharing, December 2020
NL AI Coalition

Contents

Contents

NL AIC data sharing workgroup seeks validation that the blueprint on data sharing for AI that relies heavily on IDS is future proof

Situation

- AI applications are developing fast and data sharing ecosystem for AI specifically is currently still in development
- Data sharing blueprint that stipulates how data will be shared in the AI ecosystem is now being developed by the data sharing group of NL AIC
  - Blueprint provides an initial data sharing architecture that is (partially) based on IDS* and (partially) on expertise by TNO and market
  - Blueprint will serve as starting document for the development of a Trust Framework
  - Initial data sharing architecture as described in the blueprint will be validated and extended in collaboration with the market in 2021
- Data sharing group of NL AIC plans to start the development of a Trust Framework on data sharing for AI in an adaptive collaborative approach in 2021 to create a flourishing AI ecosystem by 2024

Uncertainty

- Due to the rapidly developing AI market in terms of both technologies and applications, the NL AIC data sharing workgroup is unsure whether the blueprint based on IDS and own expertise is the right foundation for a future proof architecture for data sharing for AI

Question

- Is the data sharing blueprint the right foundation for a future proof architecture for data sharing in a flourishing AI ecosystem?

* In the entire document, IDS refers to the IDS Reference Architecture Model (IDS-RAM)

NL AI Coalition

Quick scan shows AI data sharing market is still in flux, and blueprint based on IDS and market expertise provides good base for future

The **AI data sharing market is still in flux** and given the current available architectures, a **blueprint based on IDS and market expertise is a good foundation** for data sharing for AI. However, to be fully future proof the architecture should be continuously **updated on new market developments and needs.**

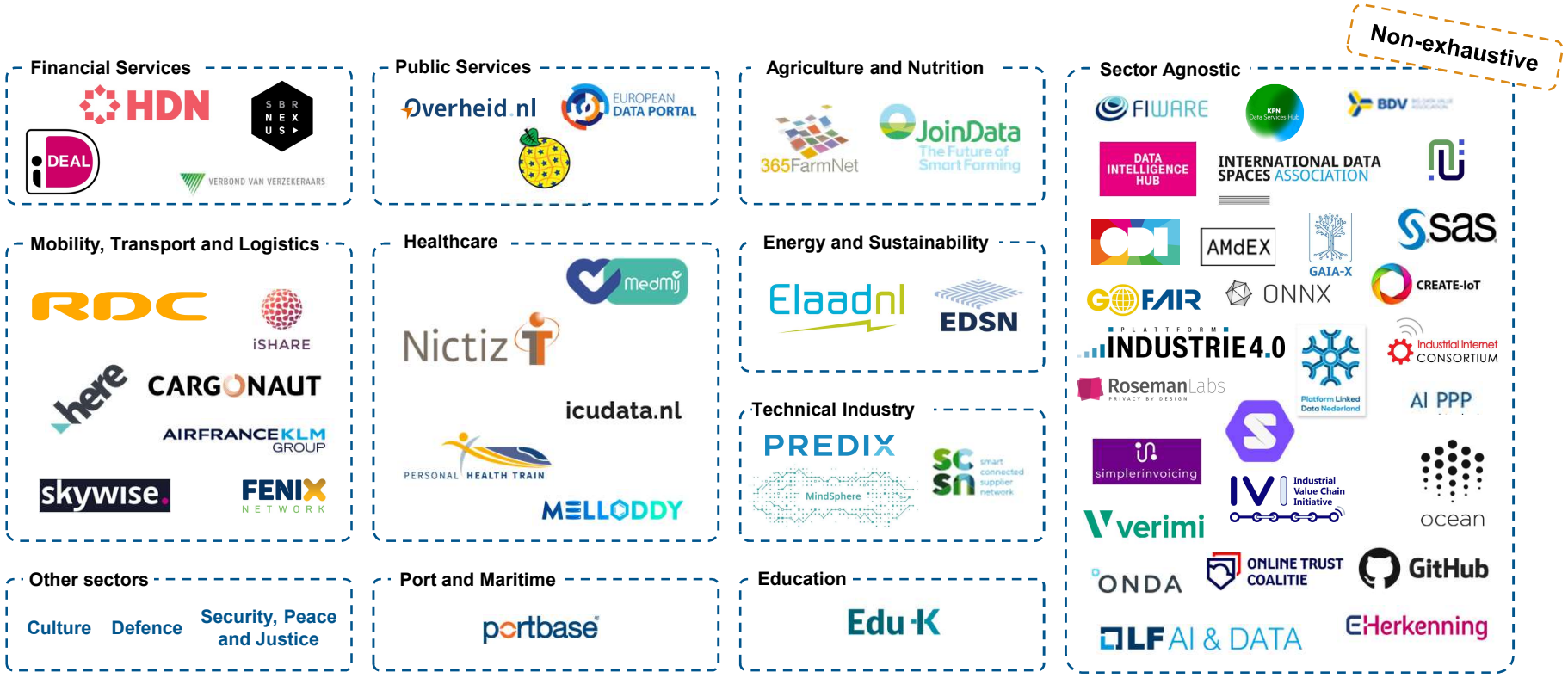
Quick scan of AI data sharing market shows wide variety of data sharing initiatives while market is still very much in flux. Initiatives suitable for AI data sharing are still in development

The market currently has no dominant architecture that is suitable for data sharing for AI and the few suitable architectures are still developing

In order to ensure that the architecture is future proof, an adaptive approach is needed in which market developments and future market needs are monitored and translated in the architecture

Contents

Quick scan shows wide variety of data sharing initiatives in the market each with its own specific domain focus

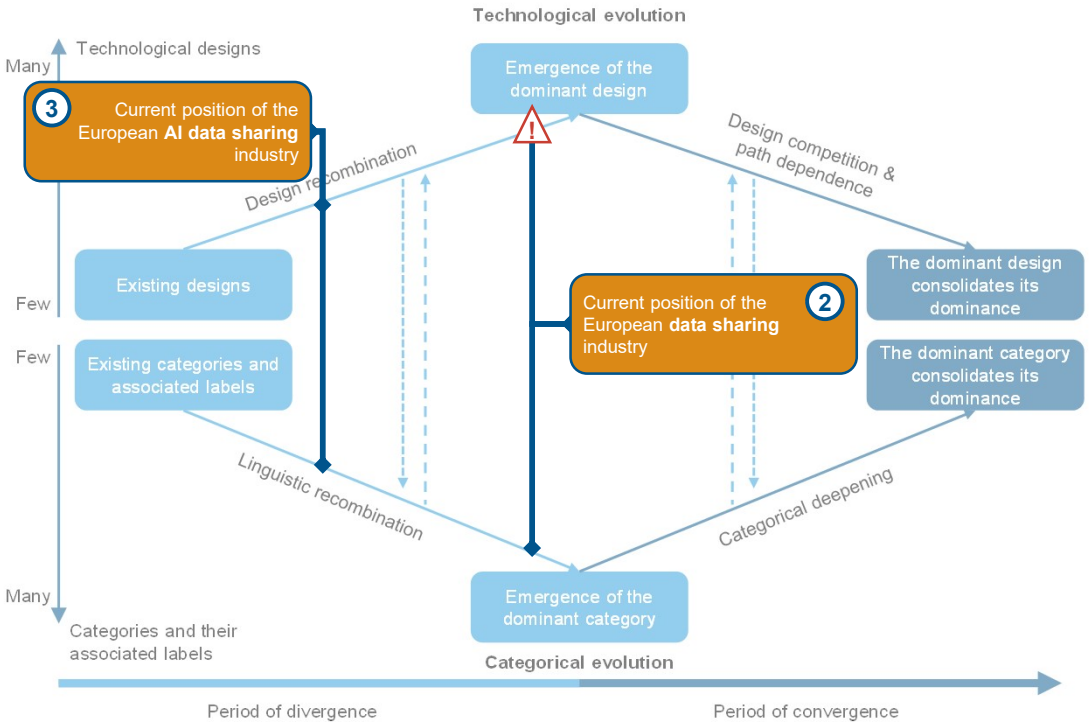


Note: this figure contains 56 initiatives because RAMI 4.0 is part of plattform industrie 4.0 and Centrum bestrijding verzekeringscriminaliteit is part of the verbond van Verzekeraars

NL AI Coalition

Quick scan shows that data sharing market is still in a divergent phase, which is especially the case for data sharing for AI

① The Coevolution of Technologies and Categories During Industry Emergence



Source: Grodal et al. 2015, INNOPAY & TNO analysis

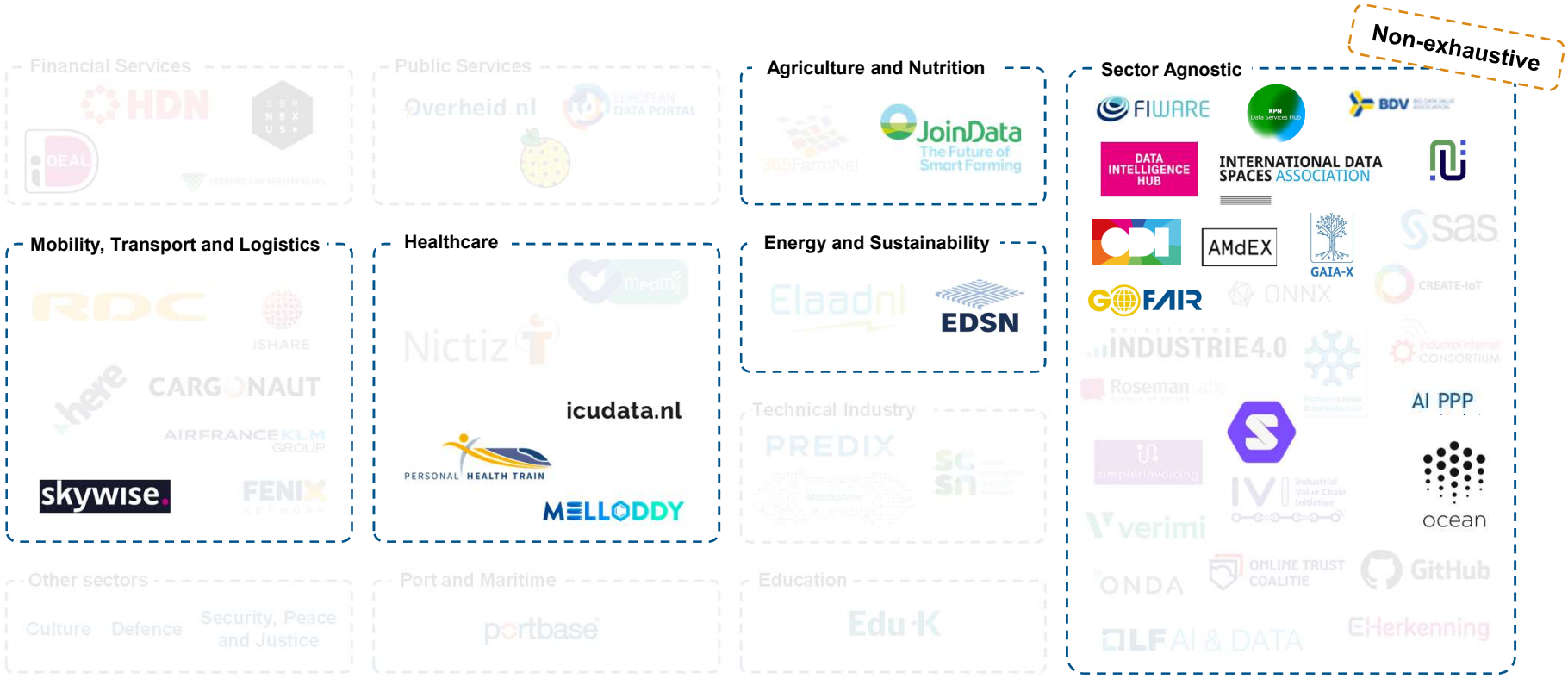
Relation to data sharing market and its initiatives

1. Most industries, including data sharing, evolve through a period of technological (identifying the ideal product/service) and categorical (identifying the ideal name/category) divergence followed by a period of convergence
2. Data sharing is still in a period of divergence where individuals and organisations continue to recombine existing technological designs and categories to identify a dominant design/category
3. 'Data sharing for AI' is still at an earlier stage of divergence, as new technologies and methods to share data for AI have only recently been discovered

⚠ *While data sharing is still in a divergent phase, several elements (e.g. data sovereignty, federated and distributed data access through data brokerage, coupled to open and inclusive interoperability) appear to emerge in multiple designs, indicating that the period of convergence is near.*

NL AI Coalition

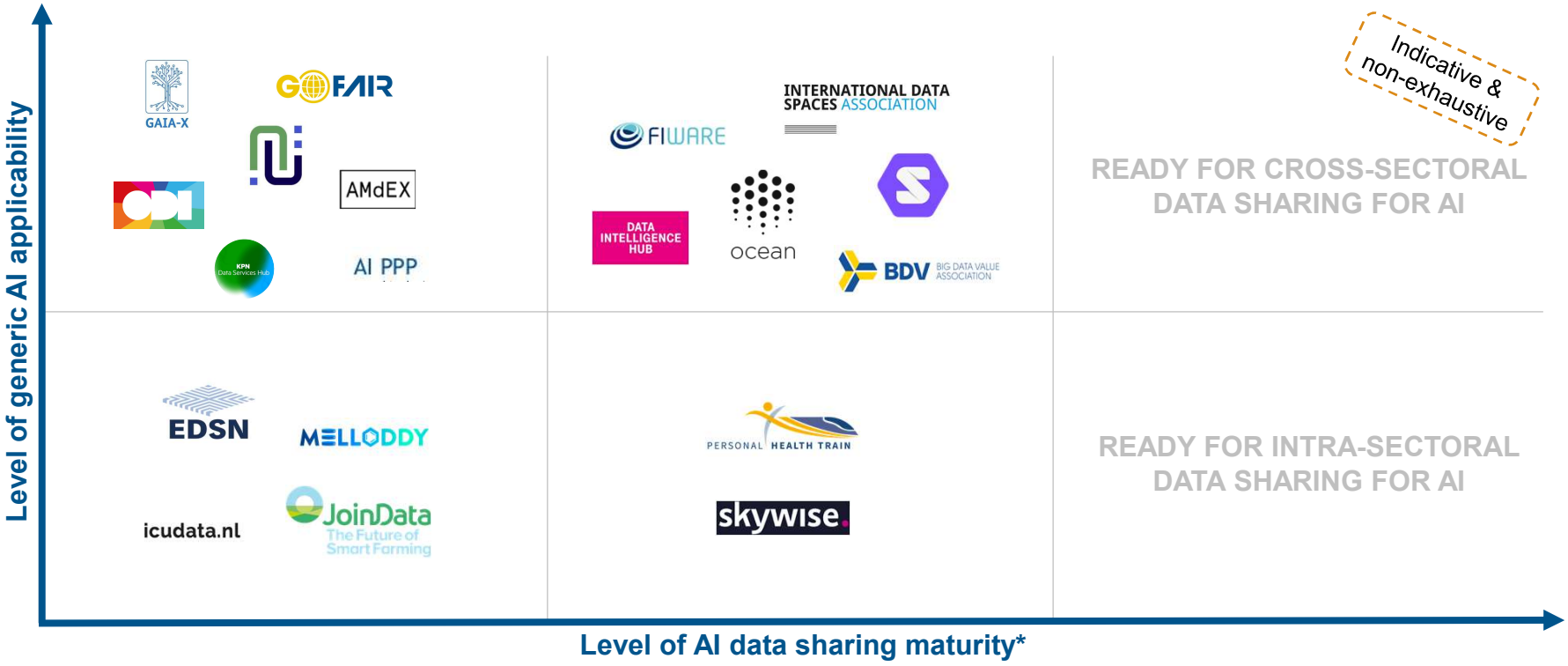
Only a few of the data sharing initiatives are, at the present time, suitable for data sharing for AI



NL AI Coalition

Contents

Although a range of data sharing initiatives show potential for AI, they currently lack the required maturity to be fully AI ready



Source: Quick scan of the data sharing market (see Appendix).

* = maturity is assessed on a low medium high scale based on level of adoption by market players

NL AI Coalition

Even the most mature initiatives and generic initiatives are still under development and none tackles all challenges for data sharing for AI

Indicative

	Challenge for AI data sharing	BDVA i-Spaces	FIWARE	IDS	Ocean Protocol	SOLID
Trust	Sovereignty, ownership of data					
	Protection of sensitive data, privacy					
Data quality	Reliability of data and quality of AI					
	Data life-cycle management					
	Verification and provenance					
Data quantity	Open data					
	Decentralised processing, access to data					
Collaboration	Data sharing frameworks, European and International coordination					
	Interoperability					
	Governance structure and profit mechanism					

Source: INNOPAY & TNO analysis, NL AIC 2020

Legend: Issue fully addressed | Issue partially adressed | Issue not addressed

Given the current available architectures, a blueprint based on IDS and market expertise is a good foundation for data sharing for AI

Quick scan shows that the market currently knows no dominant, fully developed architecture suitable for data sharing for AI

Given the challenges for data sharing for AI, the reviewed architectures FIWARE and IDS are both logical starting points for a blueprint for data sharing for AI

Considering that no architecture can be classified as future proof in this volatile market, international coordination – as done by IDS – improves the chances of successful future adoption

Contents

An adaptive approach of continuous market monitoring, operationalising and validating leads to a future proof architecture

Quick scan of data sharing initiatives only provides a snapshot of the current situation and a snapshot cannot guarantee a future proof architecture because the market is in flux and the situation can quickly change

How to realise a future proof architecture?

Monitor market

1. Snapshot of market should be continuously updated by monitoring market developments and needs

1

2

Operationalise findings

2. Findings on market on developments should be operationalised within NL AIC and validated in market

Contents

Appendix contains underlying information on data sharing market, (challenges for) data sharing for AI and reference architectures

1 **Data sharing market**

2 **Data sharing for AI**

3 **Reference architectures**

NL AI Coalition

Quick scan of data sharing market provides overview of the data sharing initiatives in the Netherlands and Europe

Approach to quick scan of data sharing market

The analysis focused on data sharing initiatives, defined as all types of initiatives that contribute to exchange of data, within the EU (with some exceptions)

58 Data sharing initiatives were identified through desk research and a reach out to data sharing experts

Interviews were conducted with internal and external experts on specific data sharing initiatives to gain deeper understanding of the initiatives

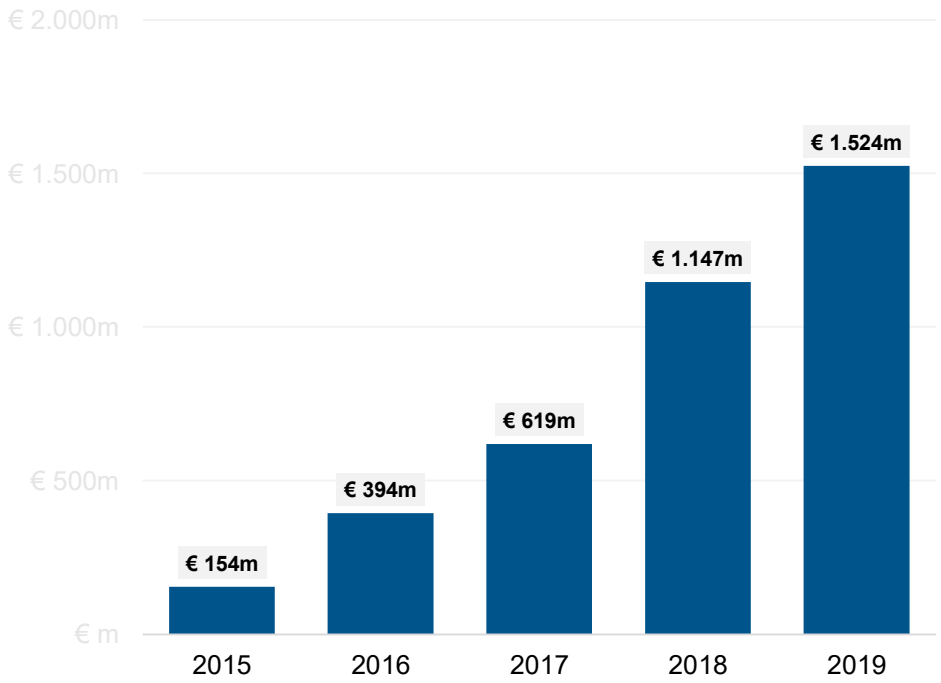
Findings were synthesized and conclusions were prepared based on the interviews, desk research and academic literature

Source: INNOPAY & TNO analysis

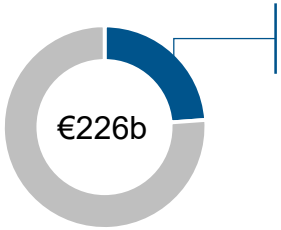
NL AI Coalition

Total investments in infrastructures for data sharing increases each year, reflecting growth of data sharing and data sharing market

Venture capital raised by data infrastructure* (DI) start-ups from 2015-2019



DI is more important for larger companies as well



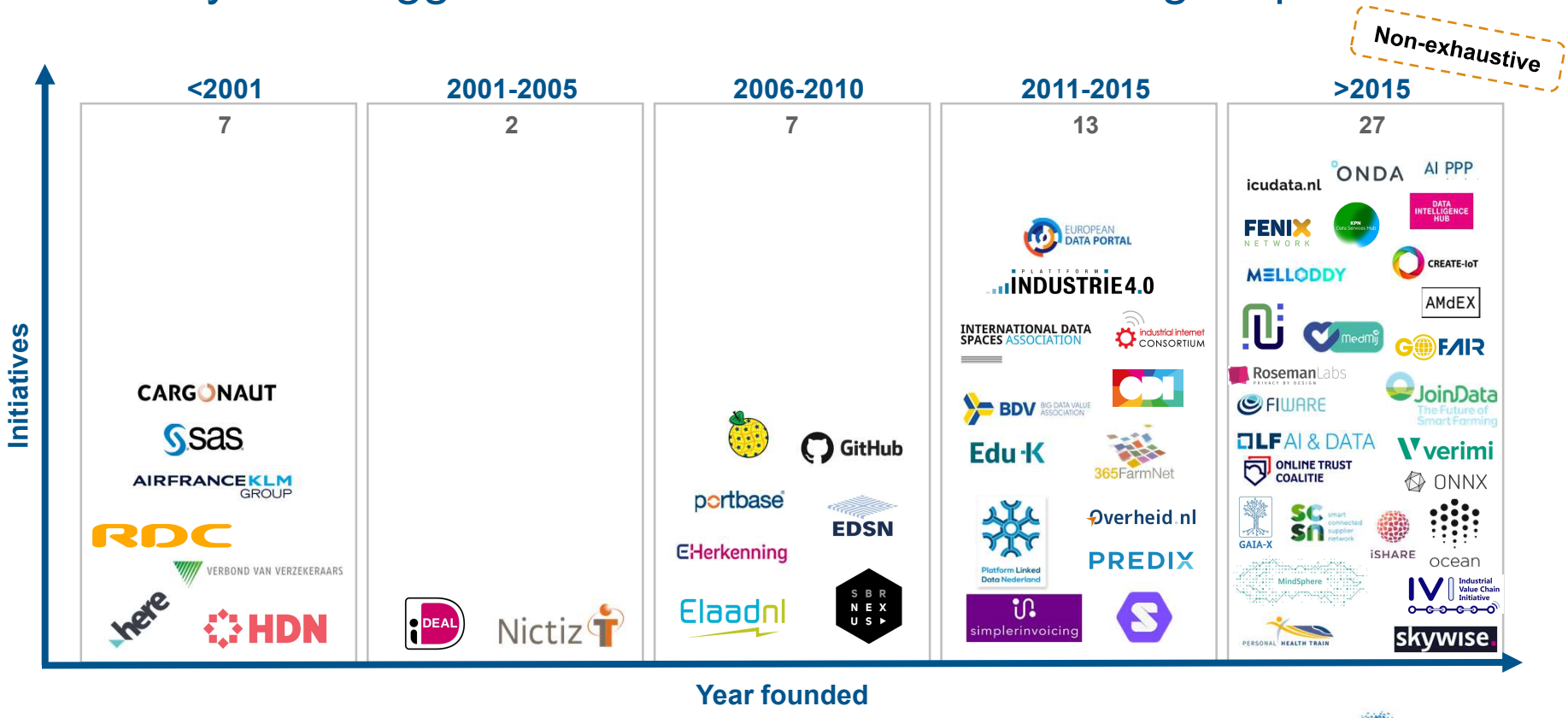
Data infrastructure spending hit a record high €54 billion in 2019

Global infrastructure software spending (2019)

Number of Fortune 1000 employing a Chief Data Officer

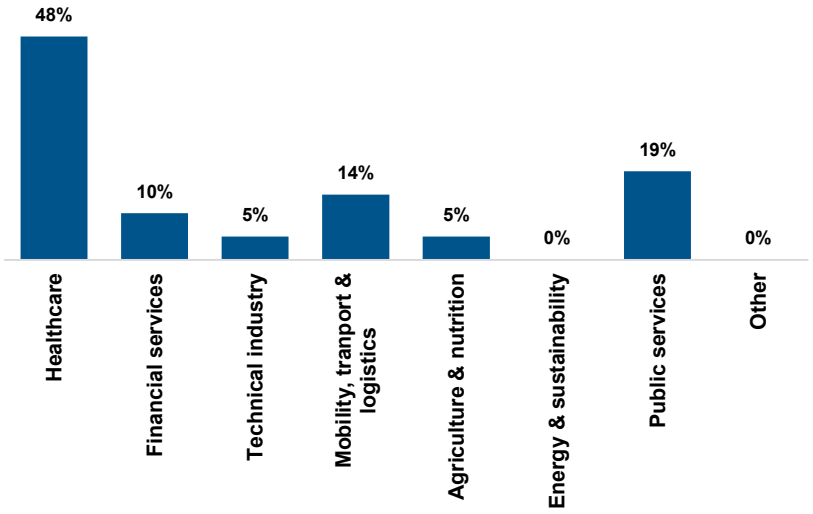
Source: Andreessen Horowitz 2020, Pitchbook 2020, Gartner 2020, NewVantage Partners 2020. *Data infrastructures are defined as analytic systems and operational systems for data sharing.

Increase of data sharing initiatives, implementing different designs, in recent years suggests that market is still in a divergent phase

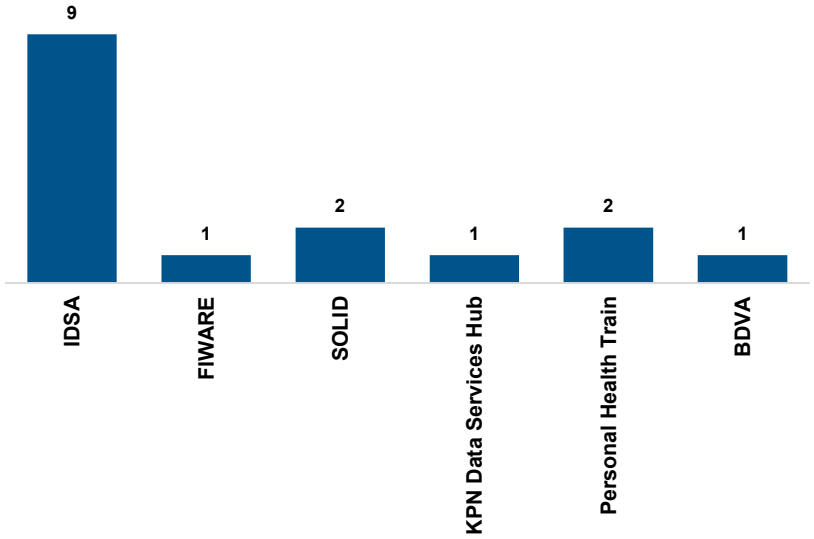


Participants of the NL AIC Data Sharing working group have different views on how the data sharing market will develop

Which sector do you consider to be best positioned to scale up the adoption of data sharing solutions for AI?



If you would be able to invest money in one data sharing initiative today, which one would it be?



Source: Mentimeter questionnaire results, Data Sharing working group session 15 December 2020

Longlist of 58 initiatives included in the analysis (non-exhaustive)

Longlist of (58) data sharing initiatives			
365 Farmnet	FENIX	KPN Data Services Hub	Predix
AI PPP	FIWARE	Linked Data	RAMI 4.0
Amdex	GAIA-X	MedMij	RDC
BDVA i-Spaces	Github	Melloddy	Roseman Labs
Cargonaut	GoFair	Mindsphere	SAS
Centrum bestrijding verzekeringscriminaliteit	HDN	Nictiz	SBR Nexus
CREATE IoT RA	HERE	Ocean Protocol	SCSN
Data Intelligence Hub	iDeal	ODI	Simplerinvoicing
Data Sharing Coalition	IDS	ONDA / DIAS	Skywise
Dataregister van de Nederlandse Overheid	IIRA	Online Trust Coalition (from ECP)	Solid
E-Herkenning	INSPIRE	ONNX (Open Neural Network Exchange)	The Dutch ICU Data Warehouse
Edu-K	iShare	Open source LF AI	Verbond van Verzekeraars
eLaad	IVRA	Personal Health Train	Verimi
Energie Data Services Nederland (EDSN)	JoinData	Plattform Industrie 4.0	
European Data Portal	KLM AirFrance	Portbase	

AI readiness map of data sharing initiatives

AI readiness map of longlist of (58) data sharing initiatives			
365 Farmnet	FENIX	KPN Data Services Hub	Predix
AI PPP	FIWARE	Linked Data	RAMI 4.0
Amdex	GAIA-X	MedMij	RDC
BDVA i-Spaces	Github	Melloddy	Roseman Labs
Cargonaut	GoFair	Mindsphere	SAS
Centrum bestrijding verzekeringscriminaliteit	HDN	Nictiz	SBR Nexus
CREATE IoT RA	HERE	Ocean Protocol	SCSN
Data Intelligence Hub	iDeal	ODI	Simplerinvoicing
Data Sharing Coalition	IDS	ONDA / DIAS	Skywise
Dataregister van de Nederlandse Overheid	IIRA	Online Trust Coalition (from ECP)	Solid
E-Herkenning	INSPIRE	ONNX (Open Neural Network Exchange)	The Dutch ICU Data Warehouse
Edu-K	iShare	Open source LF AI	Verbond van Verzekeraars
eLaad	IVRA	Personal Health Train	Verimi
Energie Data Services Nederland (EDSN)	JoinData	Plattform Industrie 4.0	
European Data Portal	KLM AirFrance	Portbase	

Source: INNOPAY & TNO Analysis

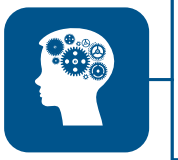
Legend: High AI DS potential | Medium AI DS potential | Low AI DS potential

NL AI Coalition

Data sharing initiatives closest to AI readiness differ in focus, region and solutions for the challenges of data sharing

Initiative	Region	Focus	Description
BDVA		**Business**	Trusted Data Incubators (DIH's) targeted to accelerate take up of data driven innovation in commercial sectors. These platforms host Closed as well as Open Data from Business and Public sources
FIWARE		**Business**	Open source initiative defining a universal set of standards for context data management. FIWARE mission, that is: “to build an open sustainable ecosystem around public, royalty-free and implementation-driven software platform standards that will ease the development of new Smart Applications in multiple sectors”
International Data Spaces		**Business**	Open, vendor-independent architecture for a peer-to-peer network which provides usage control of data from all domains
Ocean Protocol		**Business & Governments**	Ocean Protocol is a tokenized service layer that exposes data, storage, compute and algorithms for consumption with a set of deterministic proofs on availability and integrity that serve as verifiable service agreements
SOLID		**Consumer**	SOLID "aims to radically change the way Web applications work today, resulting in true data ownership as well as improved privacy" by developing a platform for linked-data applications that are completely decentralized and fully under users' control rather than controlled by other entities

Challenges for data sharing for AI* relate to trust, quality, quantity and collaboration



Challenges for AI

Trust: *making sure that sensitive data is protected from unwanted parties and prevent misuse*	**Sovereignty, ownership of data sets**	Both companies and individuals “provide” data. Given the more significant opportunities that AI provides to extract value, the importance of sovereignty under AI only increases. Secure access control is of great importance, especially with decentralised processing. Besides, AI may process data from multiple sources.
	Protection of sensitive data, privacy	Because AI can extract more information from data (including implicit sensitive information), dealing with sensitivity and privacy is even more important.
Quality: *the AI needs higher quality data to become better at predicting*	**Reliability of data and quality of AI**	Because AI is applied “ever deeper” in decision making, reliability is becoming ever more critical. Validation of (the origin of) data is essential. This also includes bias, especially in training data sets.
	Data life-cycle management	Data may not have been generated for sharing. That is why it is extra important to pay attention to the maturity of data services (cleaning, aggregation). The distinction between “training” and “production” is of great importance for data life-cycle management. It is of great importance for the quality of AI systems that they are not trained on outdated data.
	Verification and provenance	The reliability of the data itself and knowing its origin is essential for reliable results. Moreover, it also has a liability component.
Quantity: *the AI needs more data to become better at predicting*	**Open data**	Open data provides opportunities for AI because many AI algorithms have to perform tasks that require much data (think of ImageNet for training image recognition). However, it is also a potential threat because it makes it possible to re-identify anonymised data sets.
	Decentralised processing, access to data	Architectures to learn AI on distributed data, where access to data is even more critical because of the “data hunger” of many AI algorithms. Privacy, bandwidth and size of data play a role in consideration of decentralised processing.
Collaboration: *for Data sharing multi-stakeholder collaboration is key*	**Data sharing frameworks, European and International coordination**	Agreement frameworks that pursue the righteous goals collectively are crucial for AI. Especially now that there are discussions about a ‘European’ version of AI that differs from ‘capitalist’ AI (US) or ‘totalitarian’ AI (China).
	Interoperability	Because AI is not only data-driven but also knowledge-driven, the semantics of data is becoming more critical; interoperability is the basis for proper semantics.
	Governance structure and profit mechanism	The right governance structure and profit mechanism are required to ensure collaborative backing of the initiative.

Source: INNOPAY & TNO analysis, NL AIC 2020. *9 challenges have been identified in report, where 10th challenge (governance structure and profit mechanism has been added based on market expertise

NL AI Coalition

General findings on how challenges for data sharing for AI are addressed in broad data sharing market

Indicative & non-exhaustive

	Challenge for AI data sharing	Findings
Trust	**Sovereignty, ownership of data**	Data sovereignty is built-in for most generic infrastructures following European Commission Data Strategy. Algorithm ownership is not specifically addressed when the data is owned by multiple parties.
	Protection of sensitive data, privacy	Good examples are almost all medical data sharing focus on the sensitivity and privacy of data, such as the Personal Health Train. Other examples: a focus on compliance with GDPR privacy regulations, as enabled by secure multi-party computation, for example.
Data quality	**Reliability of data and quality of AI**	No data sharing initiative effectively tackles the challenges related to bias in data sets.
	Data life-cycle management	Most data sharing initiatives don't actively tackle this challenge, even though it will likely become more important in the future for AI data sharing. Therefore, this is a key point of attention for the NL AI Coalition.
	Verification and provenance	Verification and provenance are recurring problems that are addressed in supply chain data sharing initiatives.
Data quantity	**Open data**	Open data challenges are addressed in governmental public data and less in the private domain due to the competitive advantage of keeping data private. Indeed, the notion of "open data" may no longer be well aligned with the dominant data sharing design which offers strong protection against unwarranted data access.
	Decentralised processing, access to data	More data sharing initiatives are founded recently addressing the challenges related to decentralised processing. Next to the Health PoC of the NL AIC, MELLODDY is an initiative of some of the largest European pharmaceuticals to use federated learning for drug discovery.
Collaboration	**Data sharing frameworks, European and International coordination**	International coordination in a data sharing framework requires participation and collaboration with parties across Europe, for example in the standardization communities such as NEN, CEN, ETSI and ISO.
	Interoperability	Interoperability is a key element for current data sharing initiatives and it will only become more important in the future because of the increasing number of initiatives. Being interoperable across European data sharing initiatives will enable scaling-up and truly widespread cross-sectoral data sharing. As an architecture, IDS well positioned for interoperability because it easily connects to many data sharing initiatives (e.g. DSC, iSHARE, GAIA-X, FIWARE).
	Governance structure and profit mechanism	Having the right governance structure leads to a more collaborative approach. The collaborative approach is key as it increases the trust participants have in each other. For example, HDN, DSC and iSHARE all generate trust through collaboration, which is embedded in the governance structure. The profit mechanism (business model) needs to align objectives and incentives across all relevant stakeholders but this is difficult to achieve.

NL AI Coalition

The basic reference architectures for data sharing initiatives are summarised as a set of recommendations

1

ENSURE DATA SOVEREIGNTY FOR ALL PARTIES

- This goes further than the basic GDPR and keeps the data owner in control of which parties may access their data, what for, when, and what they gain in return
- A new challenge is to maintain ease of use for advanced sovereignty functionality, to enable access to these services for all parts of society

2

APPLY FEDERATED AND DISTRIBUTED DATA ACCESS USING DATA BROKERAGE

- This prevents a "winner takes all" platform from dominating the market through massive data ownership
- This represents a significant divergence from the current design of US-dominated data sharing platforms

3

WORK WITH OPEN AND INCLUSIVE INTEROPERABILITY, INCLUDING IDENTIFICATION, AUTHENTICATION AND AUTHORISATION

- By adhering to European standards for interoperability, including underlying semantic models, cross-industry data sharing will be enabled which will lead to important innovations
- This is a necessary condition to reduce risks and attract capital investment for data sharing

4

QUICKLY AGREE ON STANDARDS AND THEIR UNDERLYING SEMANTIC MODELS, INCLUDING FAIR

- Only with such open standards can data sharing initiatives truly scale up
- Supply side organisations, including telecom operators and system integrators, have an imporant role in ensuring speedy international alignment

5

ACQUIRE KNOWLEDGE AND EXPERTISE TO BE ABLE TO JOIN AND BENEFIT FROM DATA SHARING INITIATIVES

- Through pilot projects and interorganisational collaboration, relevant knowledge can be integrated into innovation projects
- Government investment is important to accelerate widespread adoption of data sharing initiatives

Source: INNOPAY & TNO analysis

NL AI Coalition

To build a future proof data sharing for AI architecture, existing data sharing architectures can provide relevant input

Data sharing architectures

#	Name of DS architecture
1	Reference Architectural Model Industrie 4.0
2	Industrial Internet Reference Architecture
3	Industrial Value Chain Reference Architecture
4	FIWARE Open Source Reference Architecture
5	CREATE-IoT RA
6	BDVA Reference Architecture
7	AI PPP Reference Architecture in SRIDA
8	IDS Reference Architecture Model
9	SOLID
10	Ocean Protocol

Commercial AI/ML architectures

#	Name of commercial AI/ML architecture
11	DAWN (Stanford University)
12	Michelangelo (Uber)
13	Mlflow (DataBricks)
14	FBLearner Flow (Facebook)
15	TFX (Google)
16	AI and ML Blueprint (combination)

For a complete overview of data sharing architectures and AI/ML architectures, see Report 'DS and AI reference architectures' by NL AIC Data Sharing group.

NL AI Coalition


NL AI Coalition