Data Market Design

Insights form Literature and Industry

06-07-2028

Stefan Driessen, Geert Monsieur, Willem-Jan van den Heuvel

The founders of JADS

JADS Jheronimus Academy of Data Science

TU/e Technische Universiteit Eindhoven University of Technology

Tilburg University

's-Hertogenbosch

Provincie Noord-Brabant

Table of Contents

The promises of Big Data and Data Science

- Big Data and Data Science will **change the way you do business**!
- Enterprises and organisations will become **data-driven**, enabling value creation
- **Share data across your organisation**: from production to BI analysts and Data Scientists

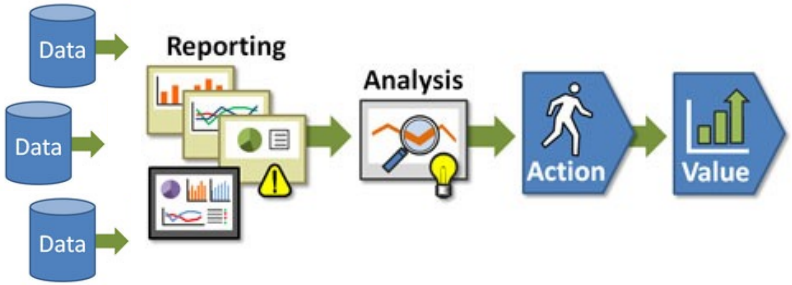
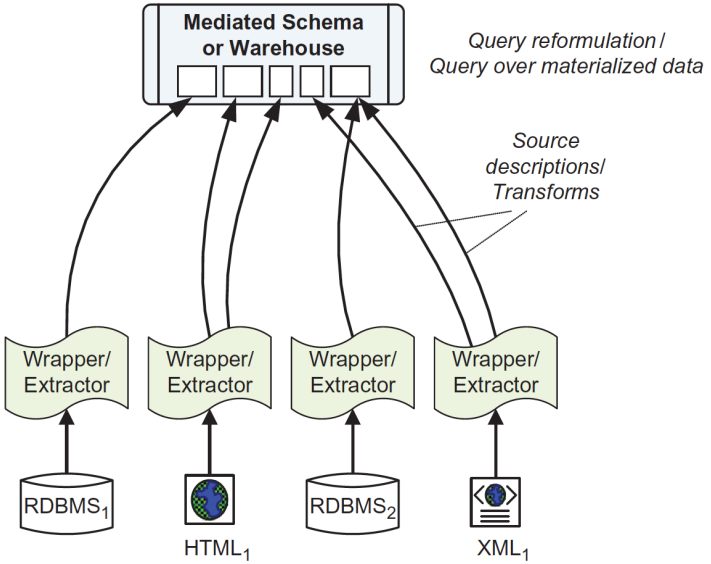


Figure based on https://www.oreilly.com/library/view/creating-a-data-driven/9781491916902/ch01.html

[3]

The Data Warehouses

- Provide a single interface to query over many sources.
- Mediated schema allows for interoperability.
- *But* tight coupling hinders scaling for big data.

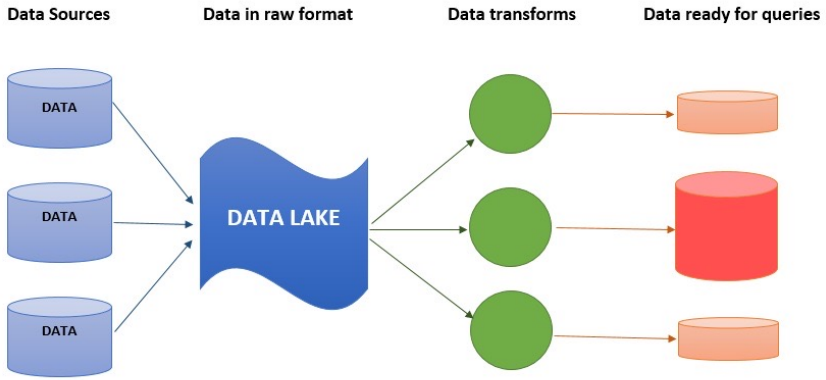


A. Doan, A. Halevy, and Z. Ives, Principles of Data Integration, 1st ed. Elsevier, 2012

[4]

The Data Lake

- Driven by 5 V's of Big Data
- On-board many data sources easily.
- Store structured, semi-structured and unstructured data as-is (raw format)
- Central "data office" enables the construction of pipelines for ingestion and consumption.



source: https://databricks.com/glossary/data-lake

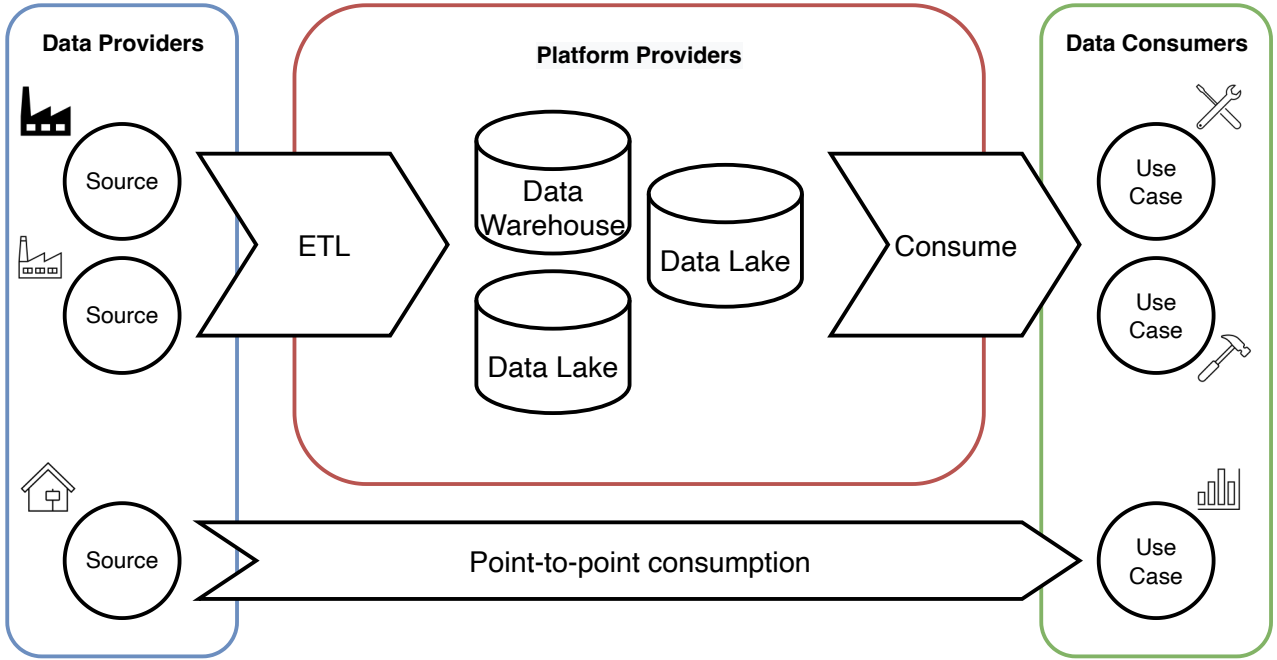
[5]

Is the data lake making us data-driven?

- Most companies and organisations have data warehouses and lakes, but few are fully data-driven.
- 87% of data science projects never make it into production (VentureBeat AI)
- 77% of businesses report that "business adoption" of big data and AI initiatives continues to represent a big challenge for business. (NewVantage)
- 80% of analytics insights will **not deliver business outcomes** and 80% of AI projects will ***“remain alchemy, run by wizards”*** (Gartner)

[6]

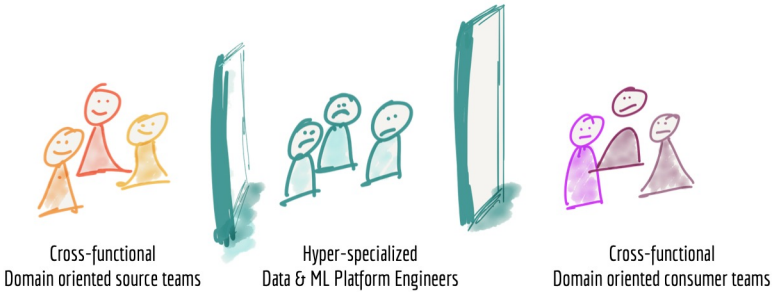
Insights from Industry



[7]

Monolithic Data Platforms do not scale well

- Monolithic platforms cannot support and harmonise **heterogeneous data** coming from **different domains**.
- Monolithic platforms cannot support **heterogeneous use cases** for data.
- **Data provider** expertise is separated from **data consumer** expertise.



Source: *https://martinfowler.com/articles/data-monolith-to-mesh.html*

[8]

Example 1: Monte-Carlo simulation for part tolerance with PD

- Using advanced techniques to investigate the tolerance of “hang-on” parts (headlights, taillights, glass roofs, etc.)
- Data was required from different plants and divisions that was known to be in the legacy systems.
- Request the same data from different teams, but get completely different data!

- Misunderstandings on requirements
- Different data sources to begin with
- No end-to-end overview

Data could not be combined

[9]

Example 2: Noise Vehicle Harshness with ITD/C

- Investigating the sound inside a driving car: NVH
- Underlying data is audio files with different quality, formats, taken in different scenarios.
- Existing data platform does not support experimentation with the data.

- Expertise of Data Providers is lost.
- Adding new functionality to existing platform is hard.
- Comparing based on metadata alone is insufficient.

Experimentation is tedious.

[10]

Decentralisation trend

- Application software architectures are shifting away from centralized monoliths and towards distributed microservices (a service mesh).

- Data architectures are following the same trend towards decentralization
  - Enterprise Data Markets — @SummerSOC
  - Data Spaces — GAIA-X (EU)
  - Data Mesh — Martin Fowler & More

From Data Assets to Data Products

Data Product = Data Asset that has been optimised for consumption

Data Asset = Data that has the potential to be valuable for the company / organisation

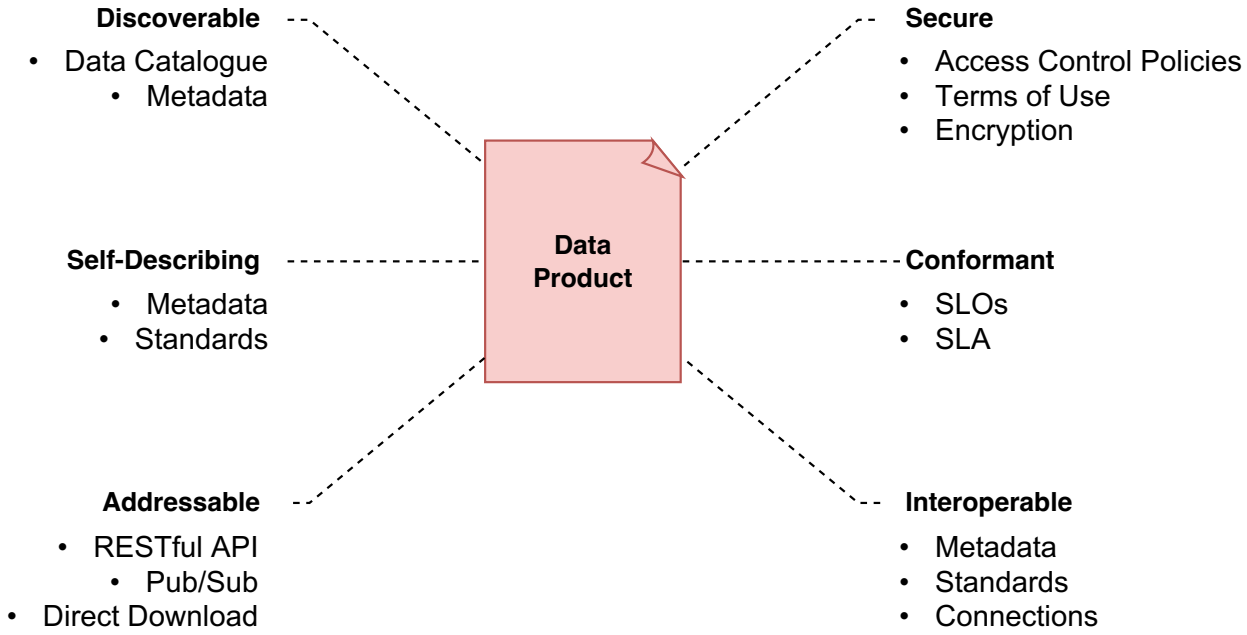
[12]

Data Products: a Tough Pill to Swallow?

A product is more than its content:

- Packaging
- Price
- Available in a store / market
- Brand
- Instructions
- Prescription
- Etc.

Data Products: Optimised for Consumption



[14]

Data Market Design: SLR

- Context & Domain of Data Markets in Literature
- Problems
- Solutions (State of the Art)
- Archetypical Data Markets

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 15, 2022, accepted March 12, 2022, date of publication March 22, 2022, date of current version March 30, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3161478

Data Market Design: A Systematic Literature Review

STEFAN W. DRIESSEN, (Graduate Student Member, IEEE), GEERT MONSIEUR, AND WILLEM-JAN VAN DEN HEUVEL

Jheronimus Academy of Data Science, Tilburg University, 5211 DA 's-Hertogenbosch, The Netherlands

Corresponding author: Stefan W. Driessen (s.w.driessen@jads.nl)

This work was supported in part by the European Union Horizon 2020 Project under Grant 825480.

ABSTRACT Data markets are platforms that provide the necessary infrastructure and services to facilitate the exchange of data products between data providers and data consumers from different environments. Over the last decade, many data markets have sprung up, capitalising on the increased appreciation of the value of data and catering to different domains. In this work, we analyse the existing body of scientific literature on data markets to provide the first comprehensive overview of research into the design of data markets, regardless of scientific background or application domain. In doing so, we contribute to the field in several ways: 1) We present an overview of the state of the art in academic research on data markets and compare this with existing market trends to identify potential gaps. 2) We identify important application domains and contexts where data markets are being put into practice. 3) Finally, we provide taxonomies of both design problems for data markets and the solutions that are being investigated to address them. We conclude our work by identifying common types of data markets and corresponding best practices for designing them. The outcome of this work is intended to serve as a starting point for software architects and engineers looking to design data markets.

INDEX TERMS Data market, data marketplace, data product, literature review.

I. INTRODUCTION

Nowadays, data is no longer viewed as an inept byproduct of (business) processes, but rather a valuable resource [1], [2]. A famous analogy proclaims data as the new oil,[1] and, like oil, it can be traded, processed and used in different contexts and applications. Indeed, the last decade has seen an incredible increase in both the amount of data being collected [3], [4], as well as the development of infrastructure necessary to process and share the vast amounts of collected data in new contexts [5], [6].

In the wake of these trends, many data markets have sprung up, facilitating data exchange between data providers and data consumers. These data markets capitalise on the increased appreciation of the value of data, catering to different domains (e.g., IoT [7], medical data [8] manufacturing data [9]) and contexts (e.g., national data [10], [11]). Therefore, it is not surprising that the scientific community has taken an interest

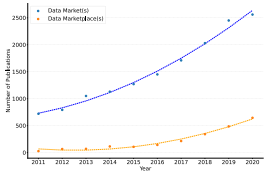


FIGURE 1. Research Trends for Data Markets, an exponential growth is observed. Source: Number of results for each query in google scholar.

in the phenomenon of data markets as well: as fig. 1 shows, there is a definite trend in scientific articles being published that have a term related to data market(places) in their title or keywords. In this work, we analyse the existing body of scientific literature on data markets to provide the first

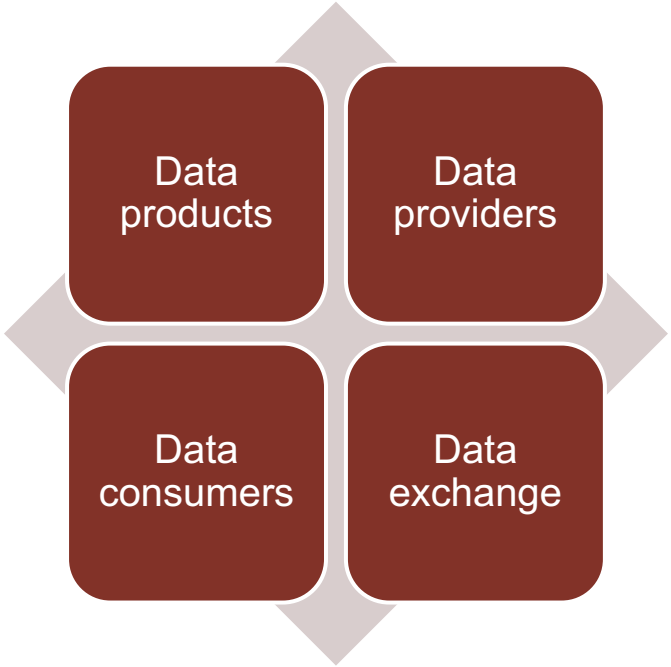
The associate editor coordinating the review of this manuscript and approving it for publication was Kostas Kolomvatsos.

[1]The Economist, "The world's most valuable resource is no longer oil, but data," may 2017

[21]

Four concepts in many formal and informal definitions

A data market is a platform that provides the necessary infrastructure and services to facilitate the exchange of data products between data providers and data consumers from different environments.



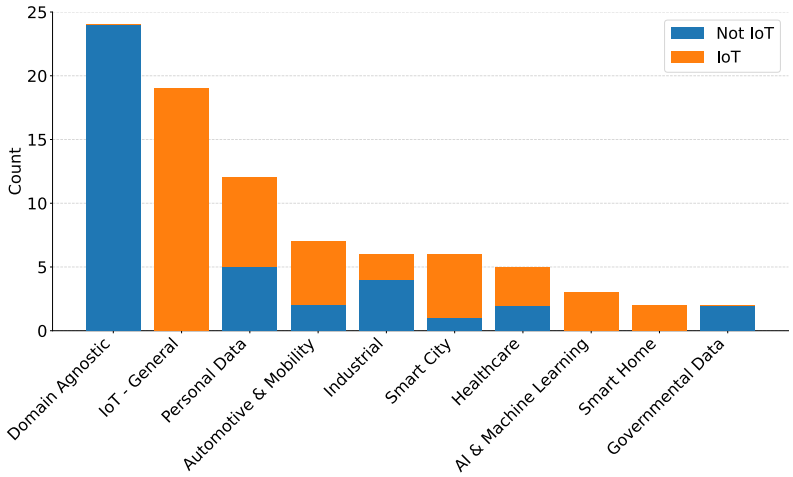
[22]

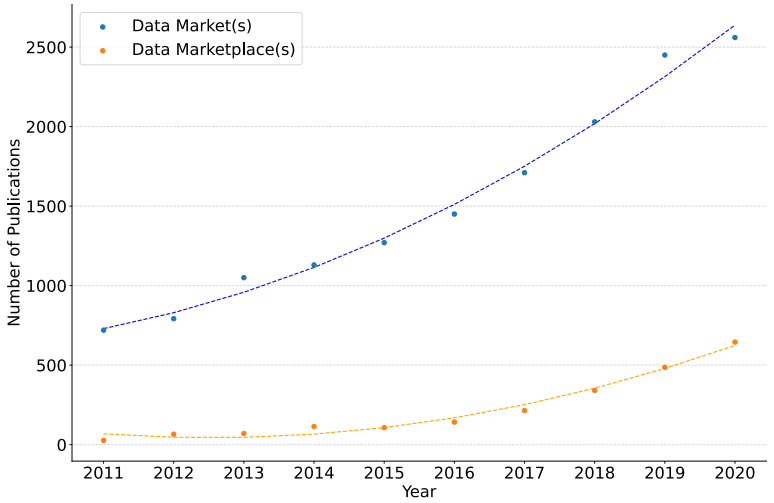
Examples

- Social Media Platforms consume your data in exchange for services
- European Initiative GAIA-X and Nokia Data Marketplace facilitate B2B data exchange.
- Decentralised Data Marketplaces allow individuals to exchange data.
- Internal Data Marketplaces facilitate data exchange inside organisations.

[23]

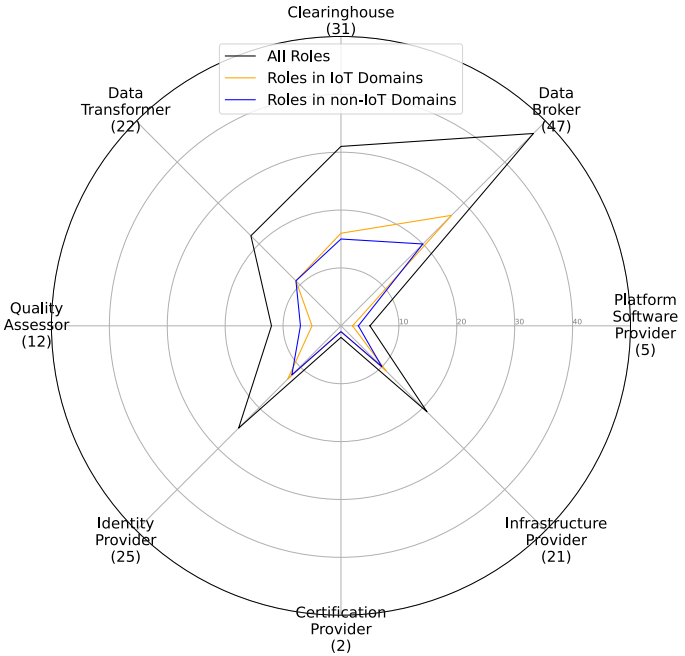
Data Markets are everywhere





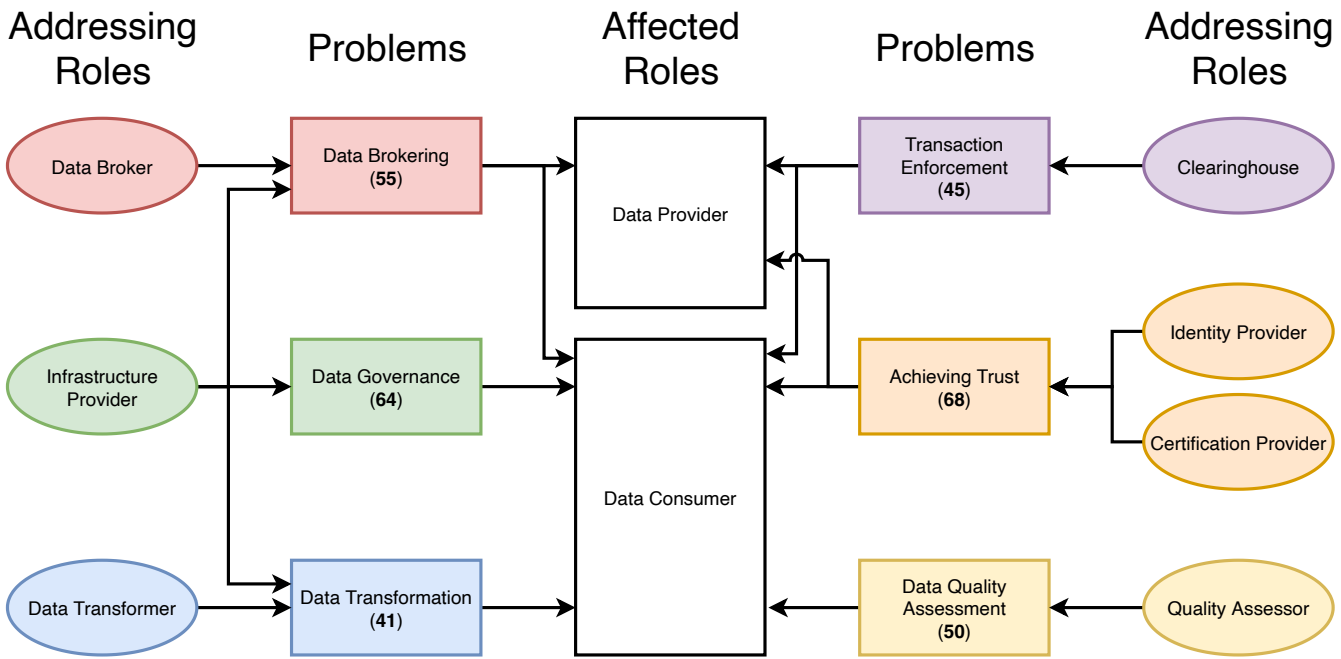
[25]

More roles / actors found in the literature



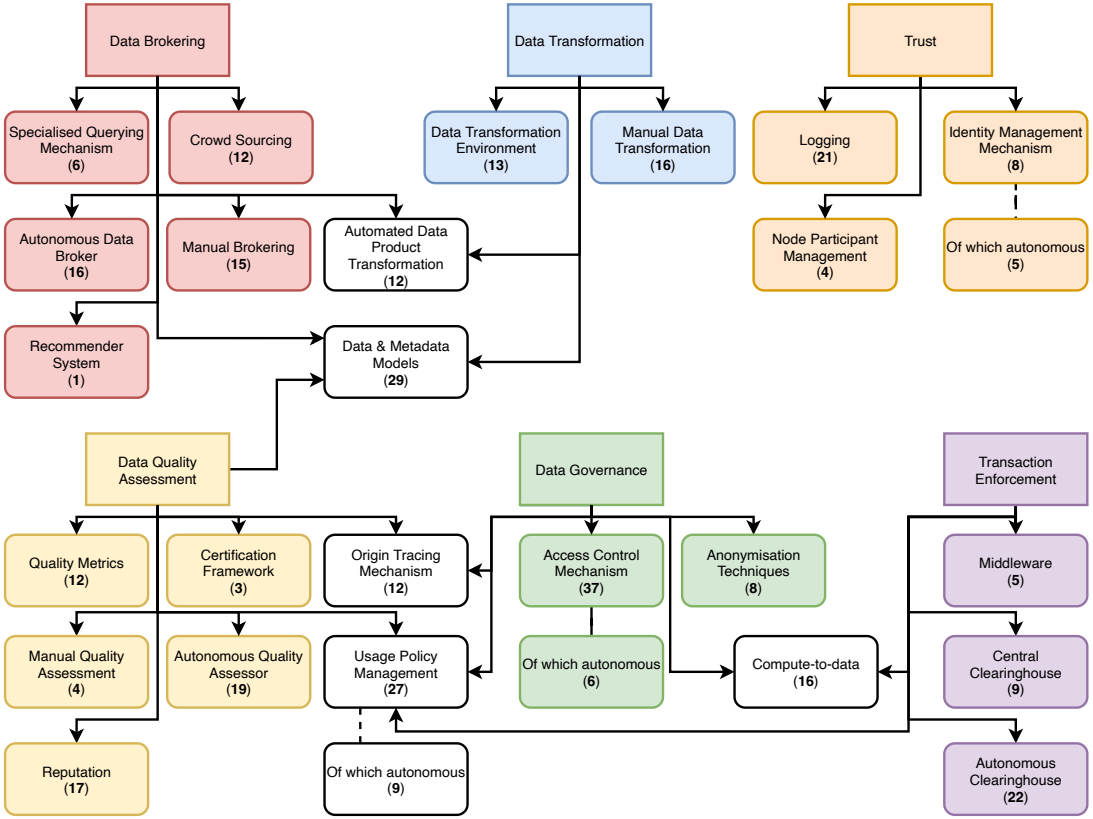
[26]

Problems in data market design



[27]

Solutions found in the literature



[28]

Five types of data markets

- Best practices
- Most commonly proposed roles
- Problems to focus on, and solutions that address these problems.
- Not mutually exclusive (e.g. a data market that is both a *specialist* and an *aggregator)*.

Generalist data market

Defining Characteristics	Heterogeneous data (as in a data mesh) Domain-agnostic Many-to-many matching
Central roles	Data Broker Clearing house
Critical problems	Data Brokering Transaction Enforcement
Typical solutions	Central Clearing house Specialised Querying Mechanism Manual Actors
Example works	Hayashi & Ohsawa [131], Spiekermann [16], Nguyen & Won [154]

[30]

Specialist data market

Defining Characteristics	Homogeneous data Single domain
Central roles	Domain Dependent Data Transformer (making it useful for many data consumers)
Critical problems	Domain Dependent Data Transformation
Typical solutions	Quality Metrics Automated data transformation Compute-to-data (works well with well-known data structures)
Example works	Ahmed & Shabani [9], Sakr [66], Sajan et al. [51], Alsharif & Nabil [91]

[31]

Industry data exchange data market

Defining Characteristics	Providers & consumers are companies/organisations Data from one domain, but heterogeneous structure Decentral architecture & many-to-many matching Consortium-owned Specialised software
Central roles	Infrastructure Provider Identity Provider Certificate Provider
Critical problems	Data Governance
Typical solutions	Identity Management Node Participation Management Certification Framework Usage Policies
Example works	Llewelyn et al. [49], Munoz-Arcentales et al. [111], Pillman et al. [83], Radhakrishnan & Das [97]

[32]

Enabler data market

Defining Characteristics	Many-to-many matching Small data products
Central roles	Clearing house Infrastructure Provider
Critical problems	Data Transformation Transaction Enforcement
Typical solutions	Middleware Central/automated clearing house Manual transformation Transformation Environment
Example works	Cao et al. [12], Jeong et al. [88], Figueredo et al. [89], Perera et al. [21]

[33]

Aggregator data market

Defining Characteristics	Many-to-one + one-to-many matching Extensive control of all processes Monopoly
Central roles	Data transformer
Critical problems	Data Transformation Data Governance
Typical solutions	Anonymisation Techniques Data Usage Policies
Example works	Eng et al. [61], Niu et al. [37], Thomas & Leiponen [13], Liang et al. [155]

[34]

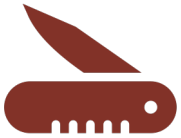
Recap

- Data Lakes and Data Warehouses often do not scale well enough to enabl big data-driven organisations because:
  - Monolithic platforms cannot support and harmonise **heterogeneous data** coming from **different domains**.
  - Monolithic platforms cannot support **heterogeneous use cases** for data.
  - **Data provider** expertise is separated from **data consumer** expertise.
- Decentral data exhanges such as enterprise data markets, data mesh and data spaces are promising alternatives but relatively untested.
- We can learn from data markets, which are better understood.

Future research

Incentivise Data Providers for internal Data Markets

Propose Architecture, Patterns and Solutions



Develop Tool-Suite for Data Product Management

[36]

JADS

Jheronimus
Academy
of Data Science

The founders of JADS

TU/e Technische Universiteit Eindhoven University of Technology

Tilburg University

's-Hertogenbosch

Provincie Noord-Brabant

Implications for Metadata Management: A balancing act

Central

- Single point of access for discoverability and metadata
- Global standards and Policies
- Link Metadata across domains

Decentral

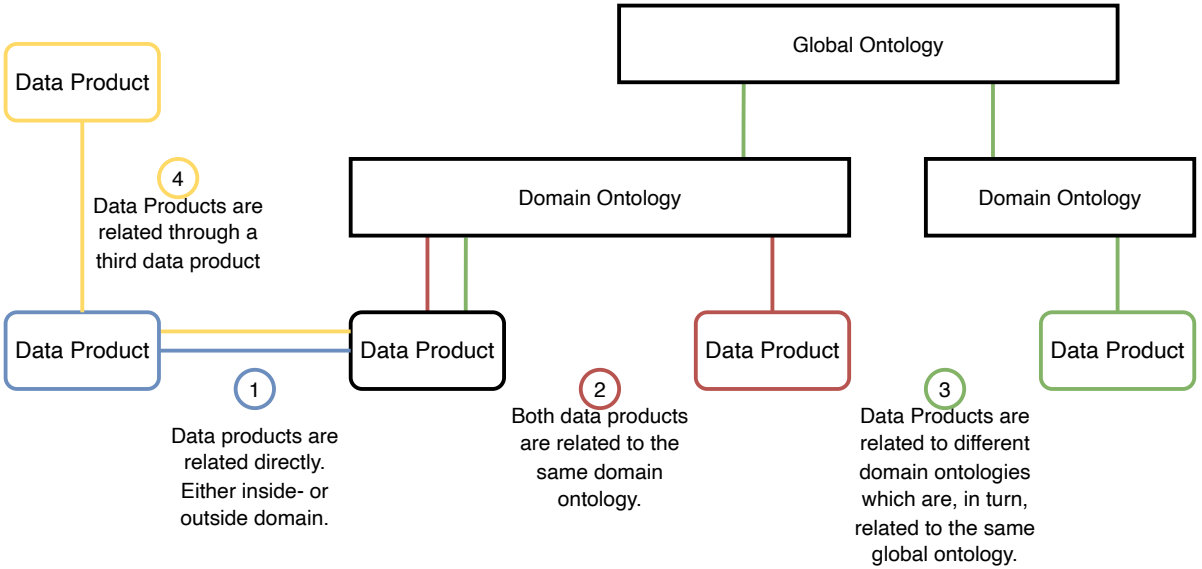
- Empower Data Providers to express their domain knowledge.
- Deviate from existing metadata standards.

Req.	Description	Addresses
R1.	The metadata management tools should allow data providers to capture domain expertise (i.e. semantic knowledge) as well as technical expertise (i.e. data schemas and statistics) in their models.	**G3**, **G4**, **G5**, **P1**, **P3**, **P4**, **P5**
R2.	The resulting models should allow data products to be connected on a data level, even when crossing domain or organisational boundaries.	**G5**, **G8**, **P5**, **P8**
R3.	The resulting models should relate data products semantically, even when crossing domain or organisation boundaries.	**G3**, **G4**, **G5**, **P1**, **P3**, **P4**
R4.	Metadata should be created autonomously by data providers and this should be as easy as possible.	**G2**, **G7**, **P2**, **P7**
R5.	The metadata management tools should allow data consumers to express data product requirements.	**G1**, **G6** **P1**, **P6**

Table 4. Five requirements were identified to help the actors reach their goals and overcome their respective problems. For each requirement, the goals and problems addressed by that requirement are shown in the final column.

[39]

Semantic-Web Technology as a Solution



[40]