The Data Marketplace Survey Revisited

Davide De Rosa

davide.derosa@studio.unibo.it

LM in Informatica

A.A. 2024/25

Introduction

Nowadays, **Data** has become a *tradable commodity*.

Recognizing this, **data providers** have established platforms for **buying**, **selling**, or **exchanging data**.

Introduction

This process is driven by several factors:

- High-quality information enables accurate decision-making, directly benefiting company revenues.
- Such reliable data can be difficult to source.
- This presents a business opportunity with significant potential.

Data Marketplaces aim to address this challenge.

Data Marketplaces History's

In the early 1990s, shortly after the advent of the Web, a new category of professionals emerged as **Information Intermediaries**.

These intermediaries were tasked with conducting searches on behalf of others for a fee. They would scour the Web for relevant information and provide the results to their clients.

The term *data marketplace* was likely first introduced in 1998 by **Armstrong** and **Durfee**. They developed a model for the exchange of information between digital libraries.

Data Marketplaces History's

Advancements in technology and the abundance of data have led to the emergence of numerous **modern data marketplaces**. These platforms act as intermediaries for buying and selling data, often processing, aggregating, and reselling data found on the Web.

They provide significant value by addressing two key challenges:

- **Aggregating** scattered data into cohesive, refined datasets, making it more accessible to customers.
- **Standardizing** access mechanisms and formats across diverse datasets, saving customers time and money.

Purpose of the three surveys

Surveying and **comparing** different *marketplaces* and *vendors* is an essential starting point for analyzing and gaining a deeper understanding of this evolving field.

By examining the current state of the market, researchers can begin to grasp its underlying dynamics, including **how it operates** and **adapts to change**.

This process also lays the groundwork for identifying **trends** that may shape the market in the future, enabling a more comprehensive understanding of its trajectory.

Classification Framework for D.M.

A **Classification Framework for Data Marketplaces** is based on later research conducted by the authors of the surveys, which build upon earlier findings and provide a deeper exploration of the topic.

Understanding the **rise** and **development** of data marketplaces is crucial for comprehending how this market is **evolving** and how it **functions**.

By systematically gathering and **evaluating** the characteristics of these platforms, researchers can gain a **clearer picture** of their structure, challenges, and potential.

Classification Framework for D.M.

Neo-Classical economics views marketplaces as both **physical** or **virtual spaces** where market activities take place.

According to this perspective, we can give two definitions.

Markets: *concrete place* where the **interactions** of buyers and sellers determine the **price** and the **quantity** of a good or a service. The focus is on a single product.

Marketplaces: for a given good is the **explicit place** of encounter in terms of time and location where market participants prepare and **execute transactions**. It provides the **infrastructure** for trading.

Classification Framework for D.M.

The market serves three different functions:

1. **Institution**: a market, as an institution, consists of a set of rules that govern the behavior of participants, assigning roles such as sellers, intermediaries, and buyers. These rules set expectations and protocols for how agents should behave, while also providing a medium for trade, helping participants meet their exchange goals.
2. **Transaction**
3. **Pricing mechanism**

Classification Framework for D.M.

The market serves three different functions:

1. **Institution**
2. **Transaction**: a market is defined by the total of all transactions that take place within it. These transactions are typically broken down into four phases:
    i. *Information phase*: agents gather information on products and express intentions to trade through bids and offers.
    ii. *Negotiation phase*: the product, contract terms, and price are negotiated and finalized into a contract.
    iii. *Transaction phase*: the contract is fulfilled, and the commodity is exchanged.
    iv. *After-sales phase*: customer service plays a key role in enhancing satisfaction and ensuring long-term commitment.
3. **Pricing mechanism**

Classification Framework for D.M.

The market serves three different functions:

1. **Institution**
2. **Transaction**
3. **Pricing mechanism**: markets serve as mechanisms where buyers and sellers interact to set prices, which act as the balancing force that coordinates their actions. Prices also signal the conditions of exchange to market participants.

Classification Framework for D.M.

The rapid development of information and communication technology (ICT) has enabled the creation of **virtual marketplaces** where products, services, and information can be traded, much like traditional goods.

Notably, **data**, in various forms, has become a new category of **tradable goods**. ICT has also led to more flexible pricing and faster transactions, with automation of information processing significantly increasing.

However, this shift presents challenges in **defining electronic markets**, as no universally accepted definition exists. The variety of interpretations complicates research, but it's suggested that the relationship between *electronic markets* and *marketplaces* mirrors the distinction between their *physical counterparts*.

Classification Framework for D.M.

Electronic Markets: submarkets qualified by the electronic infrastructure they are based upon.

Electronic markets **differ** from traditional markets in two key ways:

- The *Institutional function* is more complex because the widespread nature of electronic markets makes it challenging to establish rules and common language.
- *Pricing mechanisms* also differ. While price still serves as the main indicator of a good's value and conditions, the composition of pricing, especially regarding transaction costs and the cost structure of virtual goods, can vary.

Classification Framework for D.M.

An **Electronic Marketplace**, based on the earlier distinction between markets and marketplaces, refers to the specific platform or infrastructure that enables participants to meet and conduct market transactions in an electronic form.

However, the term is often used broadly to describe various concepts of e-commerce and market organization, or even as a synonym for electronic markets. **Wang** and **Archer** provide a summary of common definitions, grouping them into two main categories: electronic marketplaces as *governance structures* and electronic marketplaces as *business models*:

- The **business model** perspective defines an electronic marketplace as a concrete virtual platform where supply and demand meet.
- The **governance structure** perspective, on the other hand, refers to electronic markets in a more abstract sense, without directly referencing the concept of electronic marketplaces.

Classification Framework for D.M.

The evaluation of data marketplaces can be enriched by examining the relationship dynamics and the distinction between *market-based* and *hierarchical* systems.

In **market-based** systems, competitive forces freely determine the quantity and price of goods.

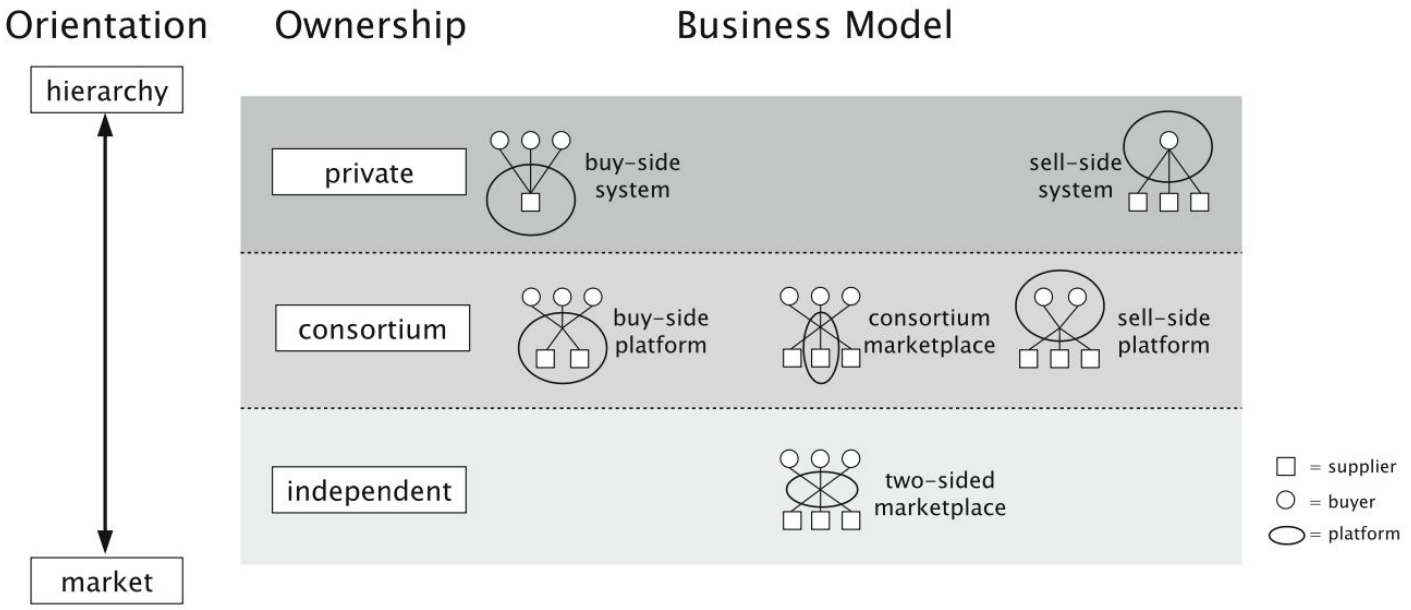
Conversely, **hierarchical** systems, which operate within organizational boundaries, give infrastructure operators (either suppliers or demanders) an inherent advantage, with predetermined constraints on prices and participants.

Transactions between suppliers and buyers fall into either market-based or hierarchical categories. The choice between these systems depends on transaction costs and the structure of the goods involved.

Classification Framework for D.M.

The model places providers on a spectrum from **hierarchy to market** and categorizes marketplaces by ownership type:

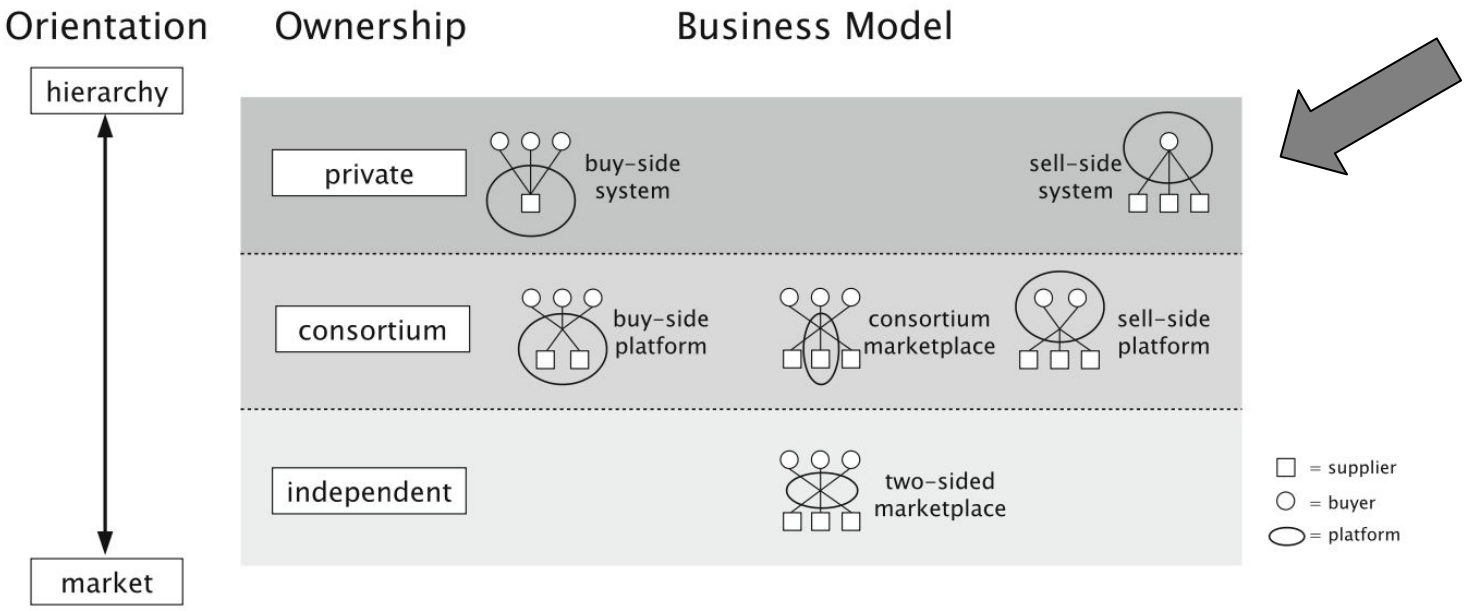
- **Private ownership**: operated by a single company (seller or buyer).
- **Consortium-based ownership**: shared by a small group of companies within the same industry.
- **Independent platforms**: neutral marketplaces without ties to sellers or buyers.



Classification Framework for D.M.

This ownership structure leads to **six** business models:

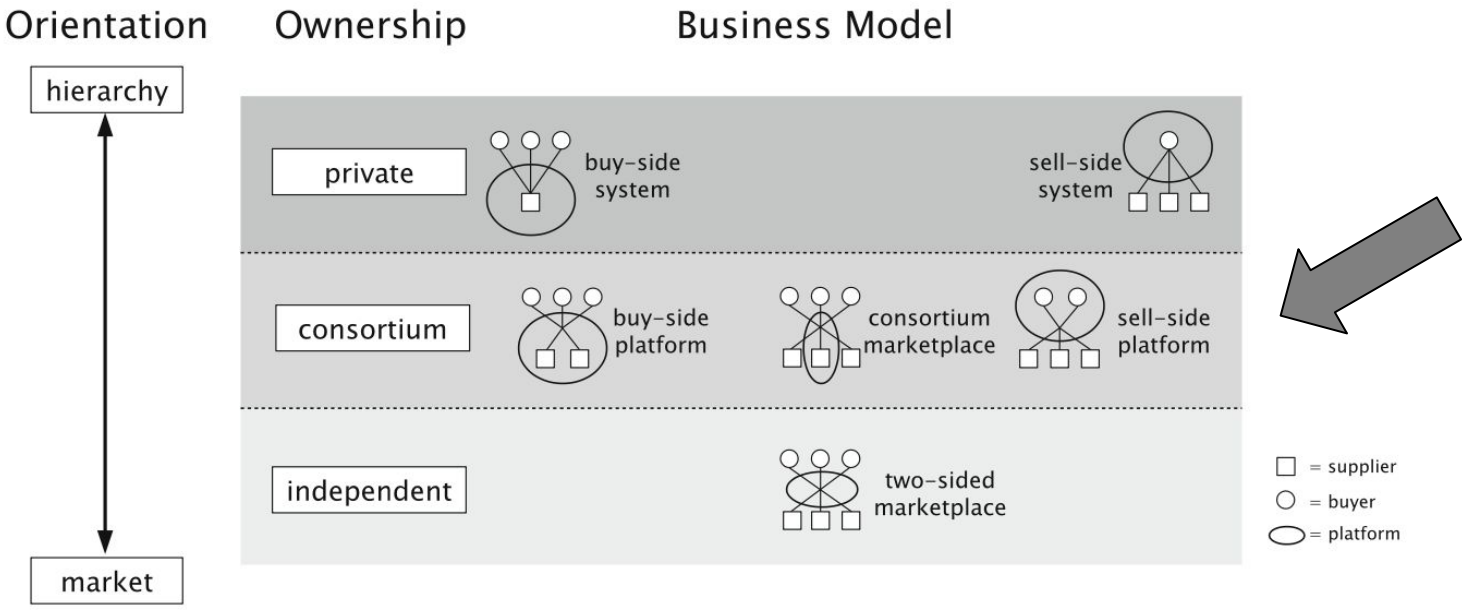
- **Hierarchy level**: privately owned platforms function in closed systems, facilitating procurement or sales for a single company, with *one-to-many* or *many-to-one* relationships.
- **Intermediate level**
- **Market level**



Classification Framework for D.M.

This ownership structure leads to **six** business models:

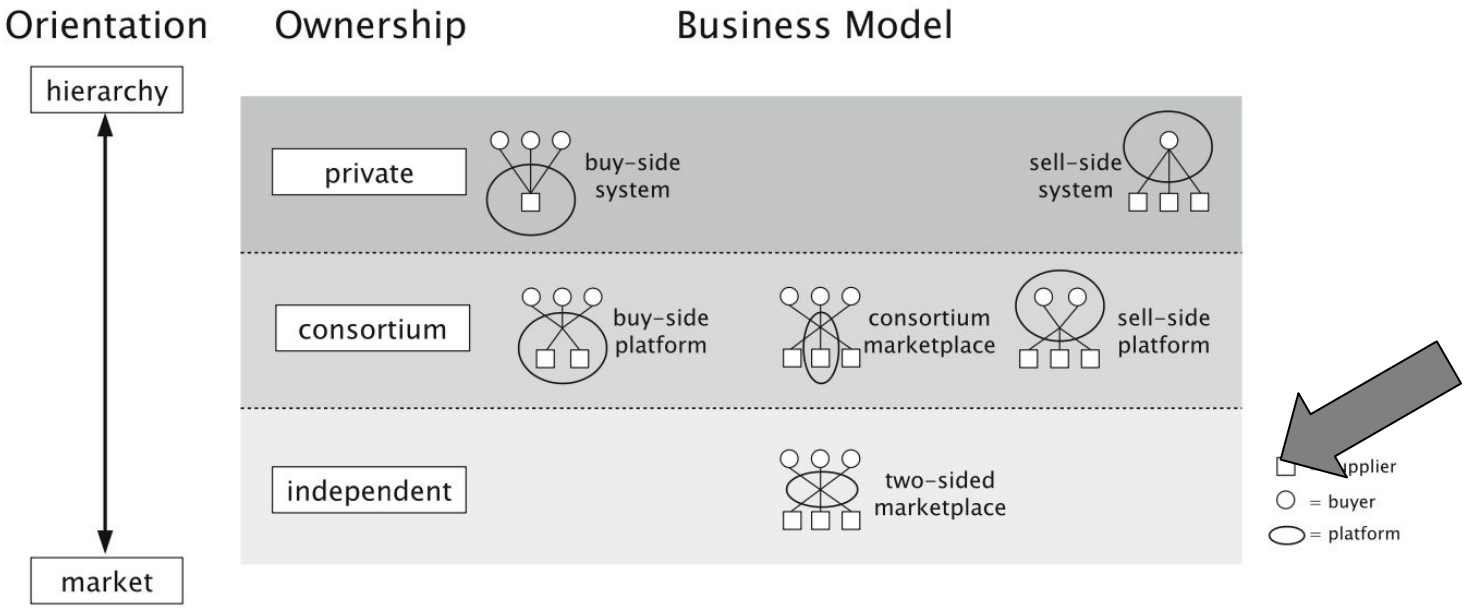
- **Hierarchy level**
- **Intermediate level**: consortium-based platforms feature *many-to-few* or *vice versa* relationships. Entry is theoretically possible but typically restricted to members of the consortium.
- **Market level**



Classification Framework for D.M.

This ownership structure leads to **six** business models:

- **Hierarchy level**
- **Intermediate level**
- **Market level**: *many-to-many* platforms are independent intermediaries with minimal entry restrictions.



Classification Framework for D.M.

This model **bridges the gap** between *overly complex theoretical frameworks* and *overly simplistic models*. It simplifies analysis while maintaining explanatory depth, providing a practical tool for empirical studies.

Also, to classify data marketplaces within the neo-classical framework for markets, some criteria are identified to distinguish them as electronic marketplaces:

- **Primary Business Model**: the provider must primarily **focus on offering data** and/or related services to qualify as a data marketplace.
- **Infrastructure Requirements**: data marketplaces must provide a **platform** that enables users to upload, browse, download, buy, and sell machine-readable data. The platform must **host** the data and **clearly identify** whether it originates from the community or the operator.

The Surveys

As already said before, there are **three different surveys**, all done consecutively. We are going to look at the first two initially, adding the results of the third later on.

The focus is on companies offering either a platform that allows users to **buy and/or sell data**, providing **raw data** in **any form**, or on **companies** offering **data enrichment tools**. The platform, or service, has to be **online**.

The vendors are divided in **12** categories (**14** in the 2nd survey).

The categories are **not mutually exclusive**.

Set of Dimensions

	Dimension	Categories	Question to be answered
objective	Type	Web Crawler, Customizable Crawler, Search Engine, Pure Data Vendor, Complex Data Vendor, Matching Vendor, Enrichment –Tagging, Enrichment –Sentiment, Enrichment Analysis, Data Market Place	What is the type of the core offering?
	Time Frame	Static/Factual, Up To Date	Is the data static or real-time?
	Domain	All, Finance/Economy, Bio Medicine, Social Media, Geo Data, Address Data	What is the data about?
	Data Origin	Internet, Self-Generated, User, Community, Government, Authority	Where does the data come from? Who is the author?
	Pricing Model	Free, Freemium, Pay-Per-Use, Flat Rate	Is the offer free, pay-per-use or usable with a flat rate?
	Data Access	API, Download, Specialized Software, Web Interface	What technical means are offered to access the data?
	Data Output	XML, CSV/XLS, JSON, RDF, Report	In what way is the data formatted for the user?
	Language	English, German, More	What is the language of the Web site? Does it differ from the language of the data?
	Target Audience	Business, Customer	Towards whom is the product geared?
	Pre-Purchase Testability	None, Restricted Access, Complete Access.	Can buyers test if the offer matches their needs?
subjective	Trustworthiness	Low, Medium, High	How trustworthy is the vendor? Can the original data source be tracked or verified?
	Size of Vendor	Startup, Medium, Big, Global Player	How big is the vendor?
	Maturity	Research Project, Beta, Medium, High	Is the product still in beta or already established?
	Pre-Purchase Information	Barely Any, Sparse Medial Information, Comprehensive Medial Information	To what degree take vendors measures to reduce information uncertainty of buyers?

Dimensions are divided in **objective** and **subjective**.

From the first survey, there have been added **2 dimensions**:

- *Pre-purchase testability*
- *Pre-purchase information*

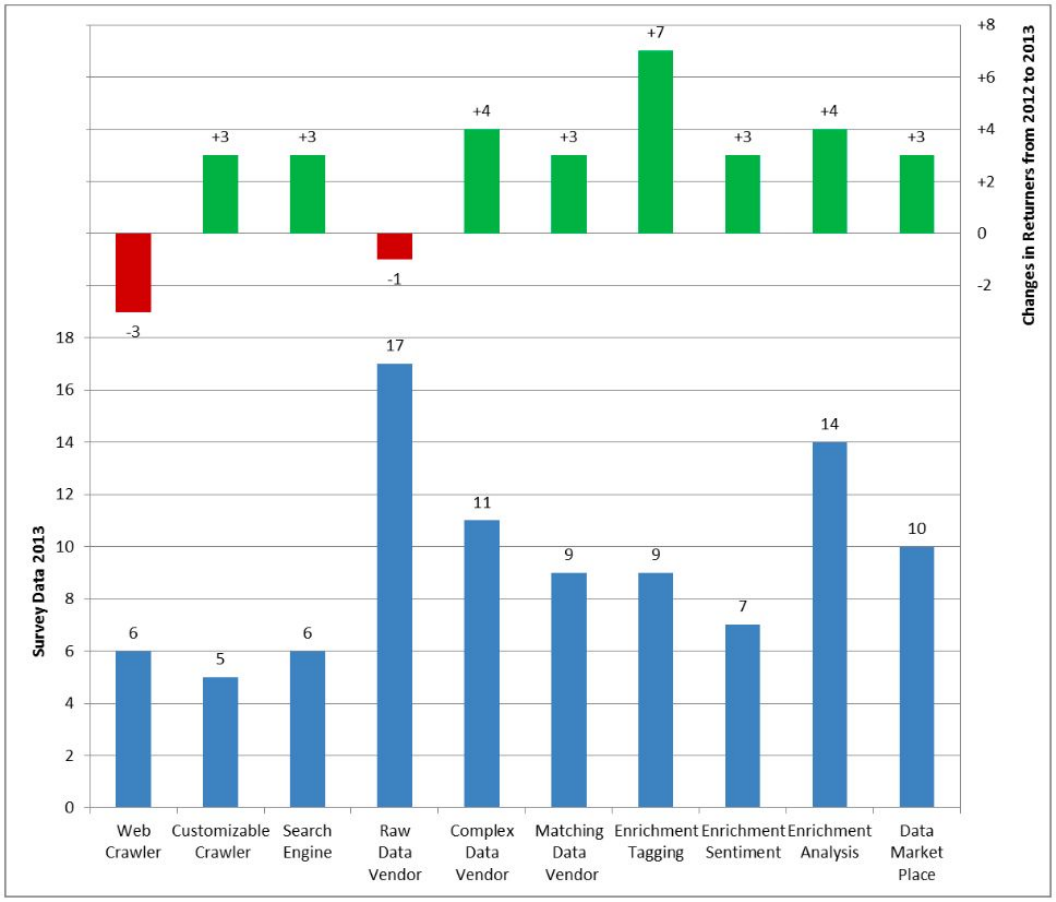
Also, some companies have since **ceased operations** or **shifted** their business **focus**, while new ones have been **added** to the latest survey.

Findings - Type

The Type dimension categorizes vendors **based** on their **core products**.

Vendors offering *Enhanced Data* are becoming more prevalent, while services focused on *Raw or Unprocessed Data* have slightly declined.

This trend likely reflects the **growing demand** for **high-quality**, **processed data** in an expanding market, where stagnation equates to falling behind.

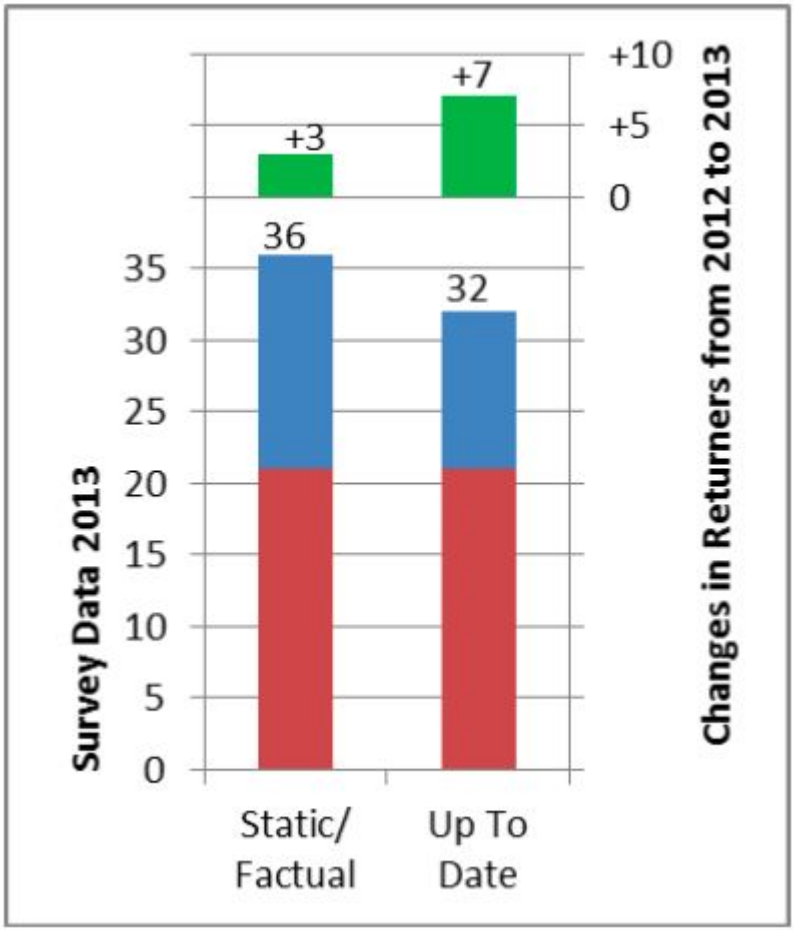


Findings - Time Frame

The Time Frame dimension categorizes data as either **long-term static/factual** or **short-term up-to-date**.

Notably, the percentage of vendors offering both types of data grew significantly, from **under 20%** in **2012** to about **45%** (21 vendors) in **2013**.

The gap between vendors offering only one type narrowed from **9 to 4**, with a stronger increase in up-to-date data among returning vendors.

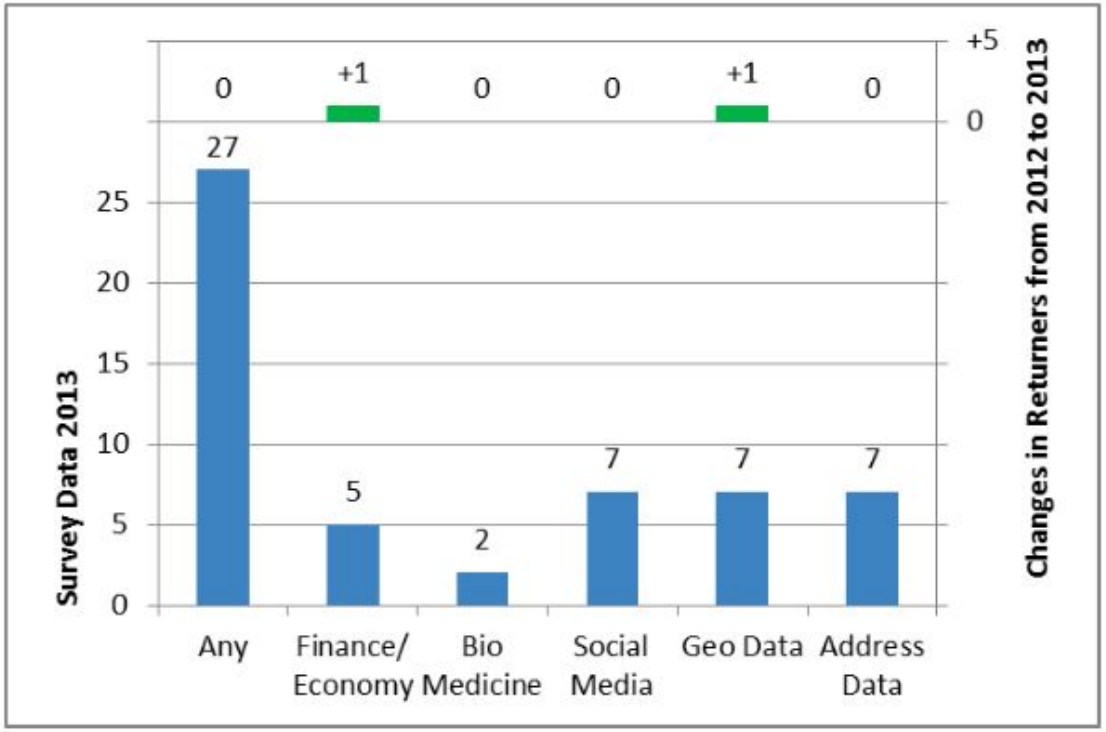


Findings - Domain

The Domain dimension identifies the **original application area** of the data.

Vendors in the **any** category, such as data marketplaces, are excluded from explicit domain counts.

Other domains are not mutually exclusive, allowing vendors to serve multiple areas.
The results are consistent with the first study, with **no notable changes** or emerging trends observed.

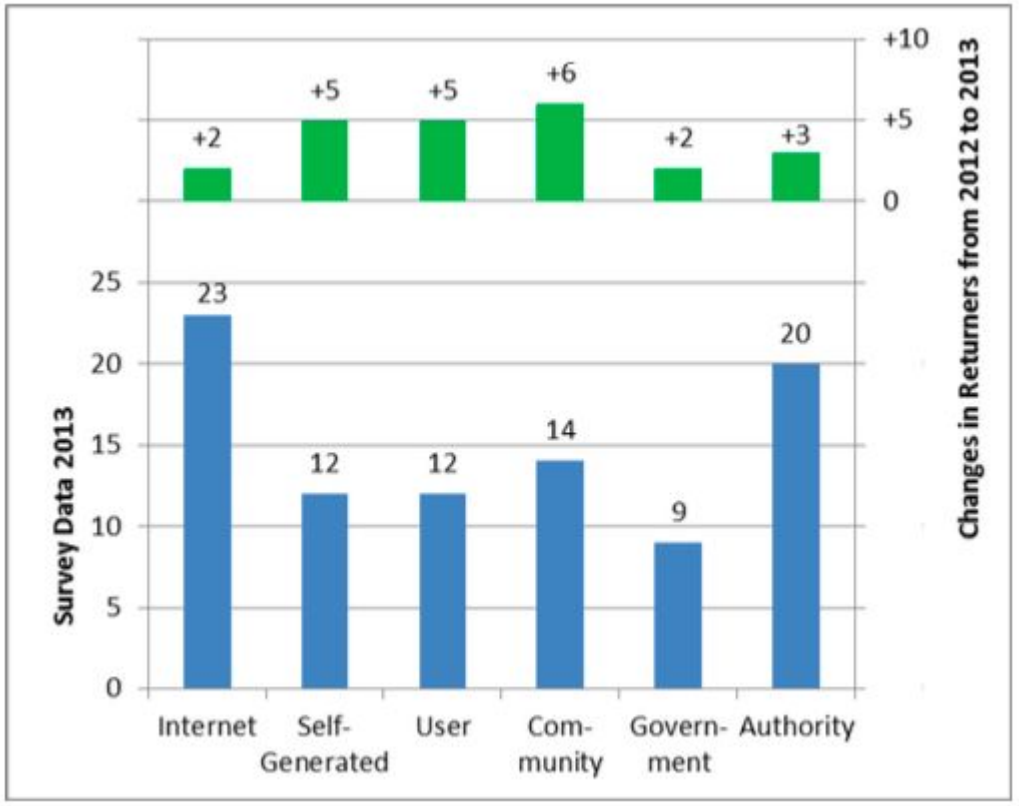


Findings - Data Origin

The Data Origin dimension classifies data by its **source**.

In the second survey, the most common sources remained **Internet** and **Authority**, valued for their **high accuracy**, **completeness**, and **credibility**.

However, there was an **80% increase** among returners in data sourced from self-generated, user, and community origins.
The rise in user data, often linked to enhancement services, indicates a growing demand for adding value to existing data.

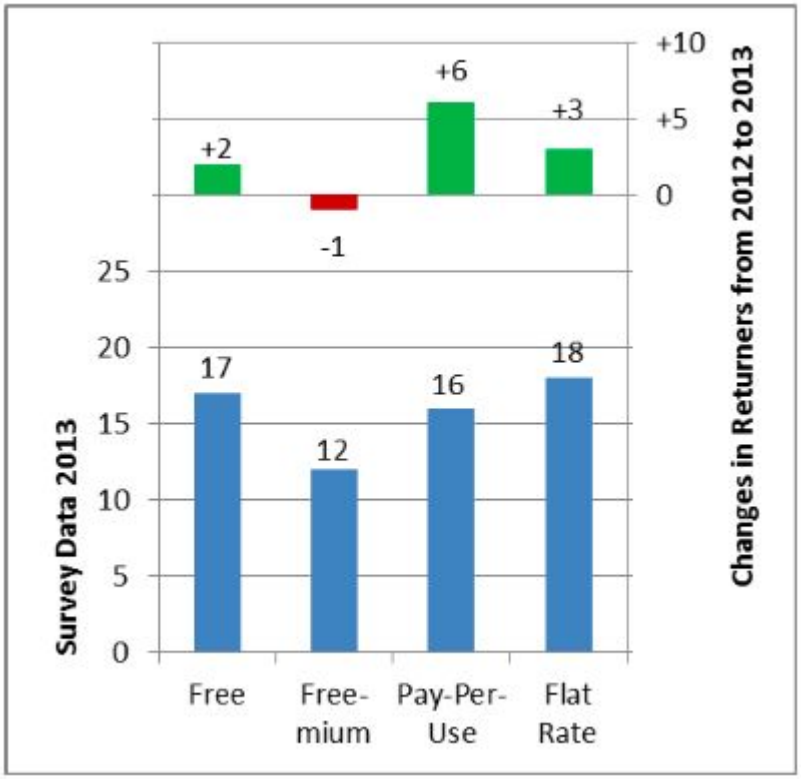


Findings - Pricing Model

The Pricing Model dimension classifies data by its **pricing model**.

Most models have remained stable, except for **freemium**, which has **declined** in popularity, and **pay-per-use**, which has seen significant **growth**.

This shift, observed both among returning vendors and overall, suggests **increased customer** trust in the quality of purchased data and a greater willingness to pay.



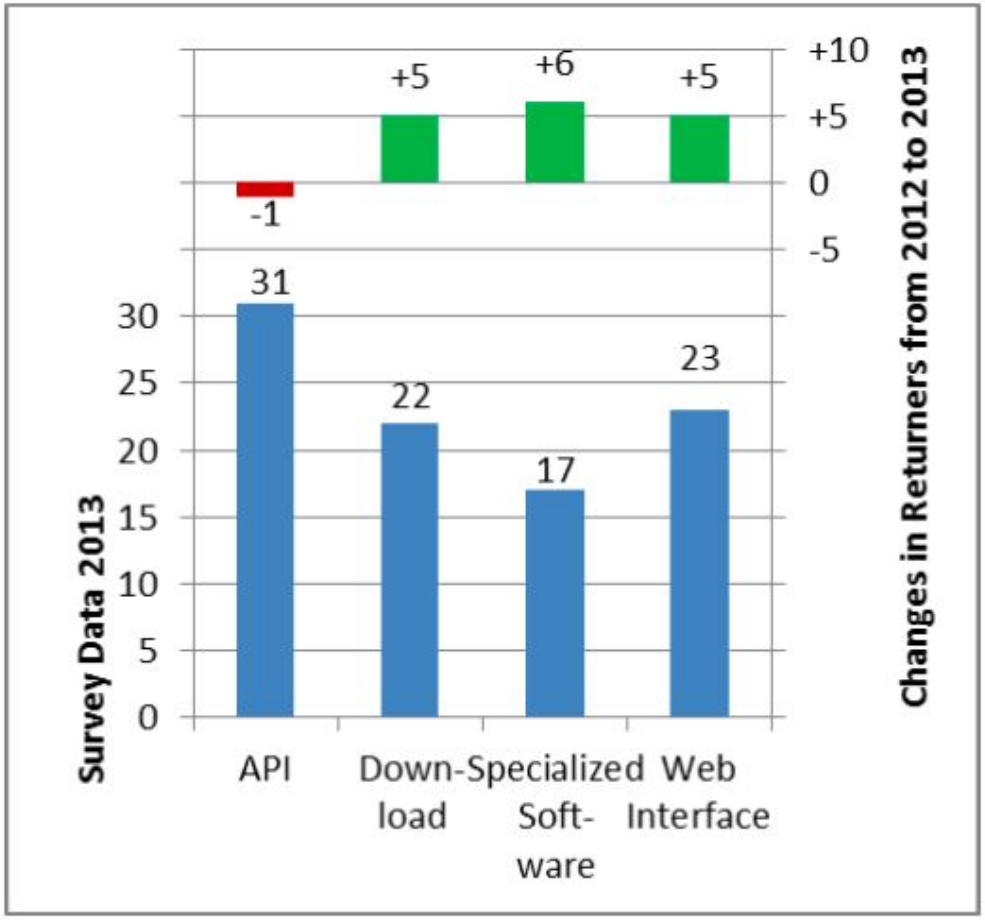
Findings - Data Access

The Data Access dimension describes how end-users **obtain data from vendors**.

APIs remain the **most common** access method but have declined in popularity compared to the first survey.

Proprietary access via specialized software saw the largest increase, though it remains the least common.

Additionally, 11% of vendors now offer **all access types**, providing customers with maximum flexibility.

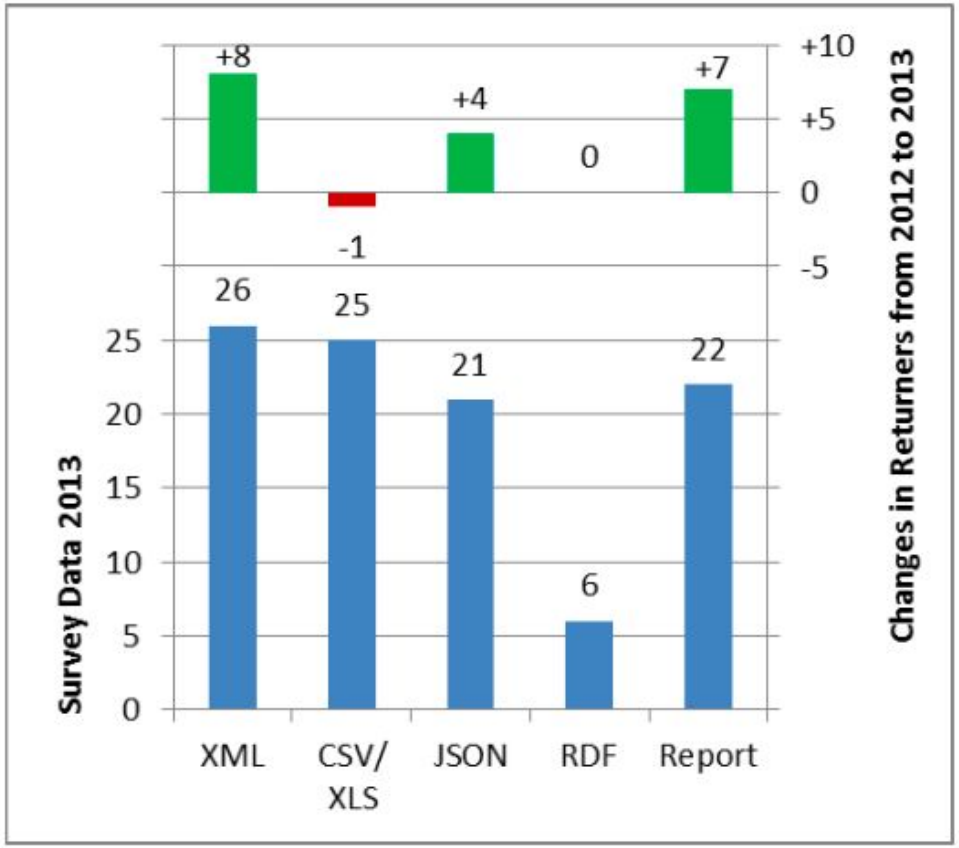


Findings - Data Output

The Data Output dimension describes how end-users **receive data from vendors**.

XML has surpassed CSV/XLS as the most popular data format, and with the rise of **JSON**, it appears that web standards are replacing traditional formats.

Two vendors even offer all data output formats. The growth in pre-formatted reports suggests that vendors are aiming to differentiate themselves and simplify data access for non-technical users, such as managers.

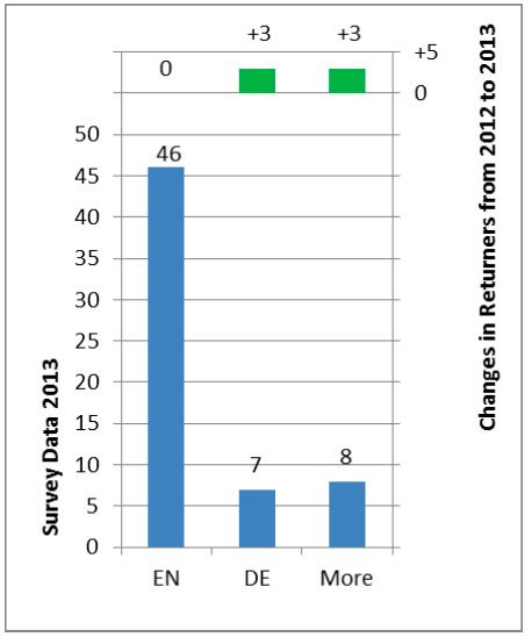


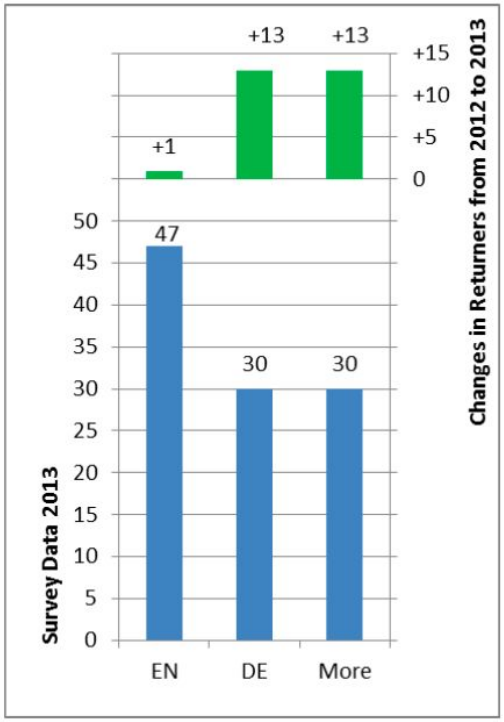
Findings - Language

The Language analysis differentiates between the **language of websites** and the **language of the data**.

As in the previous study, the figure shows that **English** remains the **dominant website language**, with only minor increases in German and other languages among returners.

However, the language of the data itself shows slight growth in English, while German and other languages have seen significant increases, indicating a **rising demand** for national, **non-English data**.





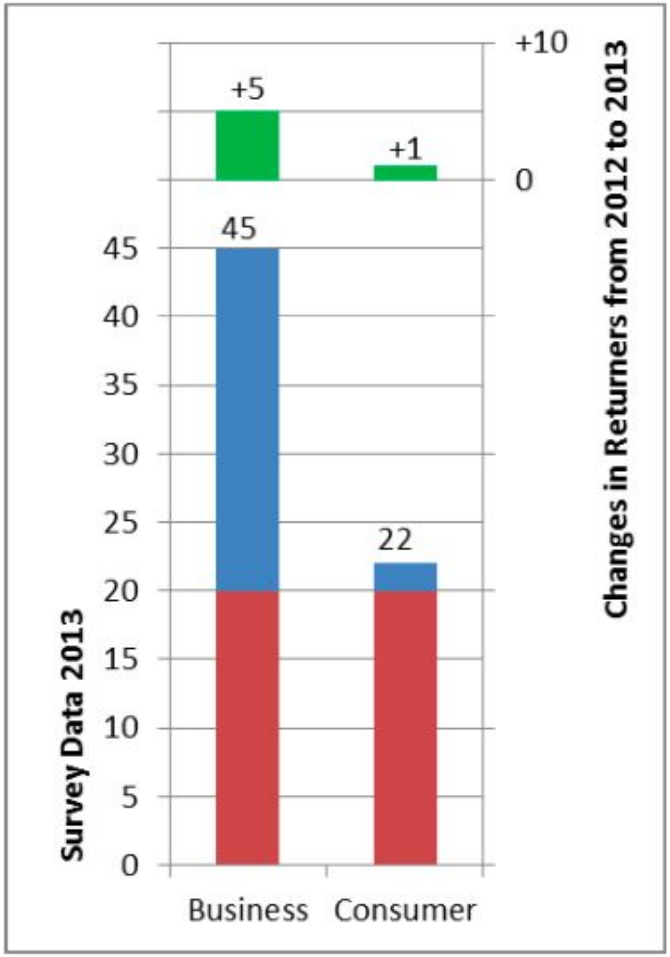
Findings - Target Audience

The Target Audience dimension distinguishes between **offerings** aimed at **business customers** (**B2B**) and **consumers** (**B2C**).

The figure shows that the percentage of vendors serving both categories increased from **28% to 43%**.

However, more than twice as many offerings **focus on B2B** customers than on consumers.

Based on the changes observed between the first and the second survey, it is reasonable to conclude that data services are, and will likely remain, a **B2B-centric market**.

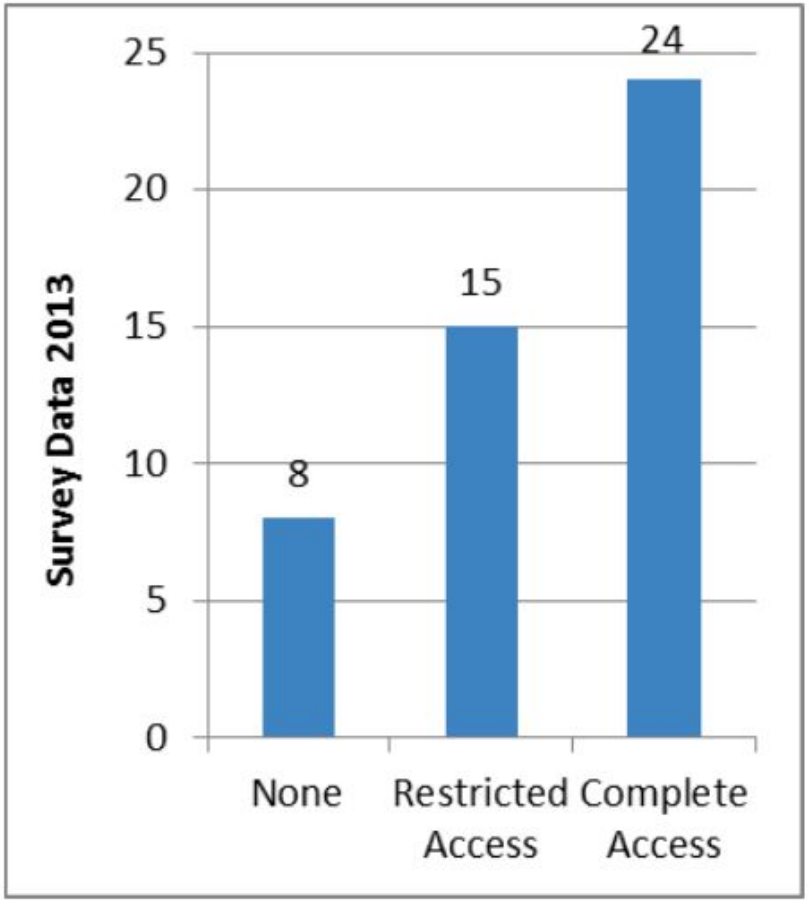


Findings - Pre Purchase Testability

The Pre-Purchase Testability dimension evaluates the extent to which data offerings can be **tested before purchase**.

Given that most buyers prefer as much information as possible before making a purchase, it's not surprising that over **80%** (39 vendors) offer at least **restricted access**.

However, **17%** (8 vendors) provide **no access before purchase**, relying on customers trusting their promises, which is surprisingly high.

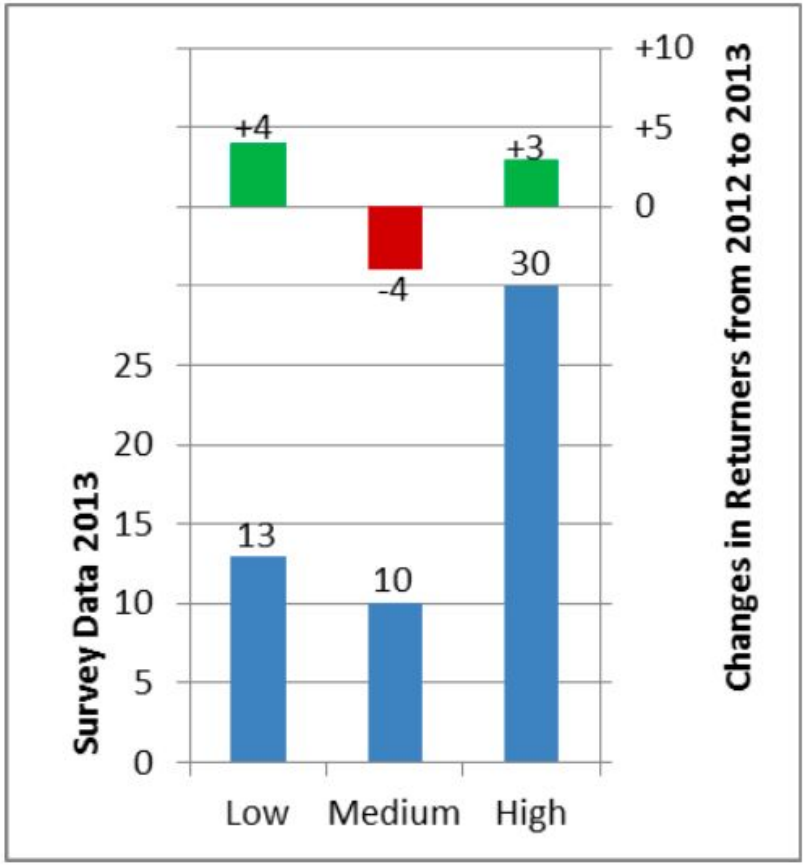


Findings - Trustworthiness

The Trustworthiness dimension evaluates the **trustworthiness** of vendors based on the **origin and processing of their data**.

Since the results are subjective and not quantifiable, they allow for multiple entries per vendor, as they may offer various services.

The figure shows no clear trend, although there is an **increase** in both **low and high trustworthiness ratings**.

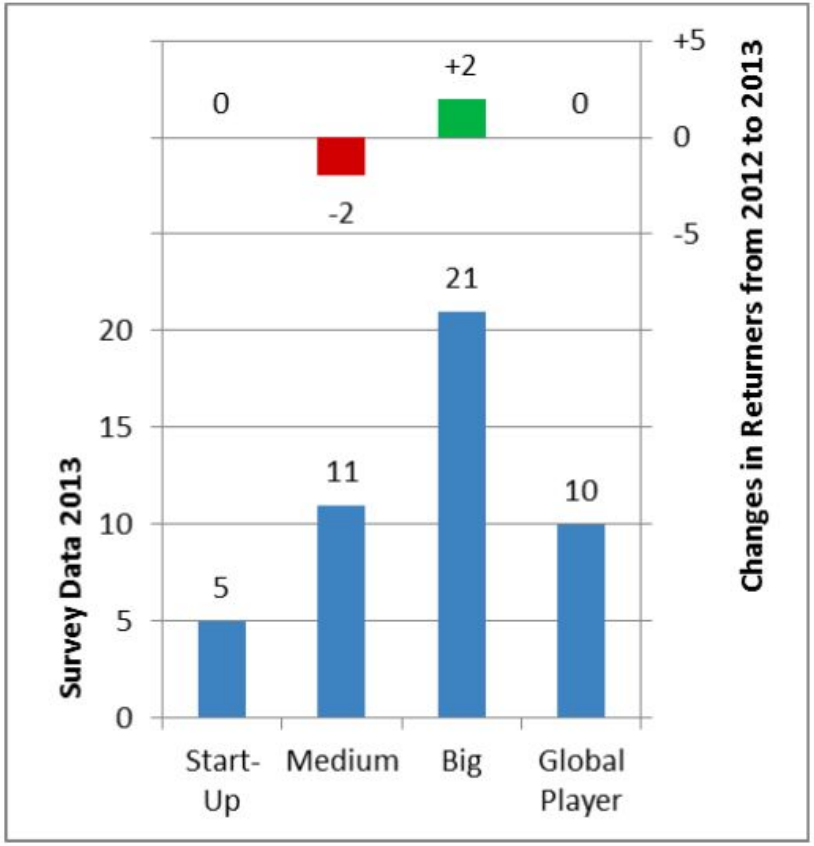


Findings - Size of Vendors

The Size of Vendors dimension classifies vendors by **size** based on their **web presentation**.

In the second survey, the overall results show a **stronger presence of large companies** compared to medium-sized ones, while the number of startups and global players remains stable.

This suggests that the **market is growing** and companies are evolving.

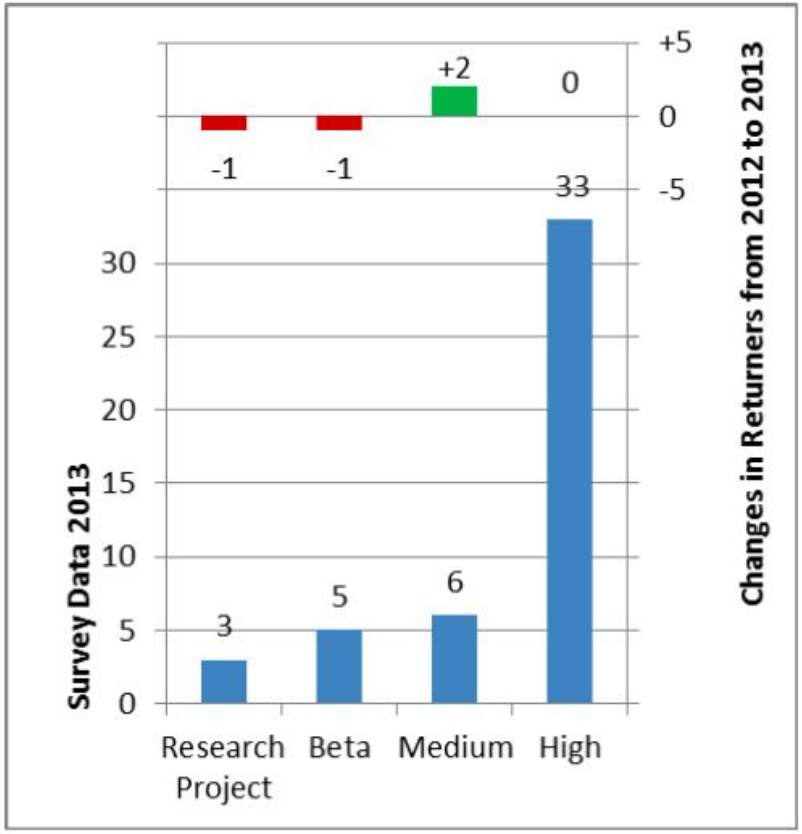


Findings - Maturity

The Maturity dimension classifies vendors by their **maturity level**.

There is a **slight increase** in **medium and high** maturity levels in the overall set, with returners also showing similar trends.

The figure illustrates this, reinforcing the idea that the **market and companies** are **growing** and **maturing**, though at a relatively **slow pace**.

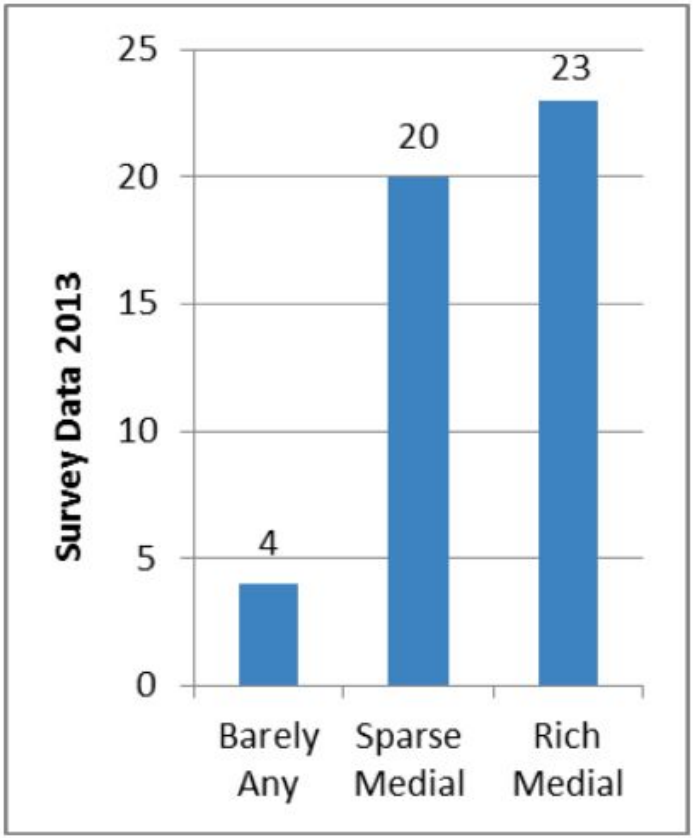


Findings - Pre Purchase Information

The Pre-Purchase Information dimension evaluates how well vendors **provide information before a purchase**, focusing on the **extent** rather than the **quality** of the information.

Since more information helps customers better assess a service, it is not surprising that **only 3 vendors** provide **minimal information**.

In contrast, nearly half (23 vendors) offer **comprehensive media content** to reduce uncertainty and assist in the purchase decision.



The Third Survey

The third survey has some **new refined dimensions**:

- The *Website Language* dimensions is **out**, considered **irrelevant**.
- The *Data Language* now refers to the **metadata**.
- A new dimensions (*Ownership*) has been added, thanks to the **Classification Framework** to evaluate biases.

Also, the sample got **bigger**. From **47 to 72** vendors.

This new survey offers new findings, such as **Trends** and a new **Statistical Analysis Method**.

Statistical Analysis Method

The survey utilizes categorical variables with binary responses (positive or negative). Since some dimensions allow multiple responses, methods for multiple response categorical variables are used for analysis.

The study combines dimensions to derive **insights based on their relevance**, such as understanding provider behavior.

Not all combinations yield meaningful results, with some being inherently correlated or unrelated.

Some dimensions combinations are shown next.

Type / Origin (1)

Investigates whether certain business models **obtain data from specific sources**.

	Origin											
	Internet		Self-Generated		User		Community		Government		Authority	
Type	#	%	#	%	#	%	#	%	#	%	#	%
Web Crawler	4	5.56	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
Custom. Crawler	4	5.56	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
Search Engine	4	5.56	2	2.78	1	1.39	2	2.78	3	4.17	3	4.17
Raw Data	7	9.72	16	22.22	1	1.39	6	8.33	6	8.33	11	15.28
Complex Data	1	1.39	6	8.33	2	2.78	0	0.00	2	2.78	5	6.94
Matching Data	0	0.00	6	8.33	7	9.72	0	0.00	0	0.00	5	6.94
Enr. – Tagging	4	5.56	0	0.00	2	2.78	2	2.78	0	0.00	0	0.00
Enr. – Sentiment	8	11.11	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
Enr. – Analysis	6	8.33	2	2.78	2	2.78	0	0.00	1	1.39	2	2.78
Marketplace	0	0.00	1	1.39	0	0.00	14	19.44	3	4.17	5	6.94

Type / Origin (2)

Enrichment services and **crawlers** rely primarily on the **Internet** for data collection.

Marketplaces focus on **community-curated** data.

While it initially seems that most Raw Data Vendors generate their data in-house, only six providers rely solely on self-generated data.

The majority aggregates data from online, federal, and institutional sources, reflecting a demand for aggregated and cleaned data rather than purely proprietary datasets.

Pricing / Domain

Examines the **prices for a given domain**.

Specialized domain data is rarely distributed for free, except for **Scientific Data**, where **80% is free**.

Social Media and Economic Data are often priced with flat rates, aligning with their need for regular updates.

No clear trend emerges for the *any* category, which shows an even distribution across pricing models.

	Domain											
	Any		Economic		Scientific		Social Media		Geo Data		Address Data	
Pricing	#	%	#	%	#	%	#	%	#	%	#	%
Free	10	13.89	1	1.39	4	5.56	0	0.00	1	1.39	0	0.00
Freemium	6	8.33	7	9.72	1	1.39	1	1.39	5	6.94	6	8.33
Pay-per-Use	4	5.56	7	9.72	0	0.00	2	2.78	7	9.72	9	12.50
Flat Rate	9	12.50	12	16.67	1	1.39	10	13.89	6	8.33	8	11.11

Origin / Domain

Examines where **domain draw data from**.

Some domains rely on diverse sources, while others are confined to specific types.

Address Data is most commonly paired with self-generated sources, reflecting limited transparency in the sourcing process.

Any data is predominantly obtained from communities, indicating minimal barriers to participation.

	Domain of Data											
	Any		Economic		Scientific		Social Media		Geo		Address	
Origin	#	%	#	%	#	%	#	%	#	%	#	%
Internet	8	11.11	4	5.56	0	0.00	11	15.28	1	1.39	3	4.17
Self-Generated	5	6.94	13	18.06	1	1.39	0	0.00	5	6.94	12	16.67
User	1	1.39	7	9.72	0	0.00	0	0.00	4	5.56	6	8.33
Community	11	15.28	3	4.17	1	1.39	0	0.00	4	5.56	4	5.56
Government	7	9.72	2	2.78	2	2.78	0	0.00	0	0.00	1	1.39
Authority	8	11.11	9	12.50	3	4.17	0	0.00	4	5.56	6	8.33

Trends

When looking at the results of the surveys over the course of the last three years, **five global trends** can be identified:

- Providers focus on only **one category** and limit themselves to only **one domain** and **one data source**.
- Providers who specialize in a single domain tend to **charge for their data** rather than offering it for free, especially in the **B2B market**.
- Providers prefer **flat rates** for their steady revenue stream, often combined with freemium models to minimize uncertainty and capitalize on **lock-in effects**.
- The data market is still largely dominated by **hierarchical** ("vertical") **relationships**, where providers control specific data offerings.
- Changes in data access types, with a shift towards web exchange formats like **JSON** and **CSV**, indicating a market orientation towards **non-technical users**.

Conclusions

The evolution of the data market over time presents a notable shift in the **size and composition of its providers**. Initially, the market was dominated by large, established companies, often from the software and hardware industries.

However, as the market matured, these entry barriers began to lower, allowing new players, particularly startups, to enter and establish themselves.

The growing presence of startups doesn’t contradict the overall trend of a maturing market. In fact, their emergence suggests that the data market has become more stable and established, with intermediaries now playing a central role in data trading.

Conclusions

In addition to this shift, the concept of **data commoditization** comes into play, especially when considering the two groups of data consumers: those seeking **highly specific**, **individualized data**, and those requiring **data of consistent quality**.

For the former group, commoditization may not be desirable, as highly specialized data is unlikely to follow a standardization path.

On the other hand, for the latter group, data that is more standardized and consistent would be more conducive to commoditization, likely increasing its exchange on marketplaces.

This could lead to a more competitive market for standardized data, further driving down implementation costs and increasing accessibility.

References

Schomm, F., Stahl, F., & Vossen, G. (2013). Marketplaces for data: an initial survey. ACM SIGMOD Record, 42(1), 15-26.

Stahl, F., Schomm, F., & Vossen, G. (2014). The data marketplace survey revisited (No. 18). ERCIS Working Paper.

Stahl, F., Schomm, F., Vossen, G., & Vomfell, L. (2016). A classification framework for data marketplaces. Vietnam Journal of Computer Science, 3, 137-143.

Stahl, F., Schomm, F., Vomfell, L., & Vossen, G. (2015). Marketplaces for digital data: Quo vadis? (No. 24). ERCIS Working Paper.