Working Papers

ERCIS – European Research Center for Information Systems
Editors: J. Becker, K. Backhaus, M. Dugas, B. Hellingrath,
T. Hoeren, S. Klein, H. Kuchen, U. Müller-Funk, H. Trautmann, G. Vossen

Working Paper No. 24

Marketplaces for Digital Data: Quo Vadis?

Florian Stahl, Fabian Schomm, Lara Vomfell, Gottfried Vossen

ISSN 1614-7448

cite as: Florian Stahl, Fabian Schomm, Lara Vomfell, Gottfried Vossen: Marketplaces for Digital Data: Quo Vadis?. In: Working Paper No. 24, European Research Center for Information Systems, Eds.: Becker, J. et al. Münster 2015.

Contents

List of Figures

Type

Research Report

Title

Marketplaces for Digital Data: Quo Vadis?.

Authors

Florian Stahl, Fabian Schomm, Lara Vomfell, Gottfried Vossen
contact via: {florian.stahl, fabian.schomm, lara.vomfell, vossen}@uni-muenster.de

Abstract

The newly emerging market for data is insufficiently researched up to now. The survey presented in this work - which is the third iteration of a a series of studies that started in 2012 - intends to provide a deeper understanding of this emerging type of market. Research questions concerning the provider manifestations and the commoditization of data are identified. The findings indicate that data providers focus on limited business models and that data remains individualized and differentiated. Nevertheless, a trend towards commoditization for certain types of data can be foreseen, which even allows an outlook to further developments in this area.

Keywords

Cloud Computing, Data as a Service, Data Marketplace, Data Marketplace Survey, Data Marketplace Development

1 Introduction

The Internet now enables almost ubiquitous transactions and exchanges of information, which has led to the emergence of new markets which were not conceivable before; additionally, it has initiated major transformations of existing markets. Due to the unprecedented supply of data, the Internet has not only altered how people relate to information, but has also allowed an increasing proliferation of data. Prior to the Web 2.0, the data market could be characterized as a private large-scale information exchange between major companies [15]. Increasingly, data is both supplied and demanded publicly on the Internet, which has led to the emergence of free databases like Wikipedia or Wolfram|Alpha, but also to the emergence of data marketplaces, i.e. virtual spaces of exchange between many actors on the supply and demand side [16], or simply electronic marketplaces where the commodity data is traded. This paper reports on the third and final study of data marketplaces and tries to answer the following questions:

1. What manifestations do data providers choose to operate on data markets?
2. Is there a progression of commoditization of data and if this is the case, how far has it advanced?
3. What is to be expected within the next three to five years?

Every new market is characterized by numerous participants entering and exiting while developing solutions and strategies for the number of challenges that every new market entails. The relatively high number of providers eventually leaving the field in the past few years illustrates that data markets appear to be particularly challenging. Interviews with founders of the visualization tool Swivel, closed in 2010, yielded that, aside from the “usual” management issues, the main obstacle to their business was that the number of users willing to pay for their services was ”in the single-digit area” [10]. The Internet, the very medium that has led to the transformation of data markets in the first place, is also one of the major threats to their economy: Users are accustomed to have constant access to information for free which results in a rather low willingness to pay for data. Companies with a focus on data provisioning need to find suitable strategies to make revenue from their offering.

Despite a lot of discussion in the blogosphere, systematic research on the landscape of data marketplaces is still scarce. Some evaluations on a small scale have been performed with notable examples being [7, 8, 15, 14, 18]; however, several of the offerings discussed there are already out of business. Until recently, a deficiency in the investigation of data marketplaces was the lack of a theoretical groundwork as well as the lack of clear definition. In order to mitigate those issues, we have developed both a theoretical and clear definition in [27] to transparently communicate the foundation of our surveys. In order to come to a clearer understanding of the market, it is crucial to analyze the solutions providers employ and the various business models, which we try to deliver in this paper.

Trading of data as an information good is often impeded because of varying utility value attributions on the consumers’ side, information asymmetries and particular cost structures that makes data pricing very complex. Providers willing to make the selling of data their primary business model need to find a pricing strategy that exploits the customers’ willingness to pay while considering that the marginal cost of data is zero [27, 13].

The first question stated above serves to identify whether certain forms of data provisioning are more reasonable than others and how providers deal with the issue of generating revenue from data. This relates to the topic of the data they sell, how they reduce buyers’ uncertainty, and which means of differentiation they adopt. The second question is concerned with the good *data* itself. Data is a rather abstract, digitized good the value of which is difficult to estimate. In turn, this factors into the issue of pricing. A process of product standardization, called commoditization,

has the potential to facilitate the issue of value attribution. The state of commoditization of data indicates whether the traded data is highly differentiated and unique, in which case one provider has only few direct competitors or whether data converges towards commodities, which entails a more perfect market. The last question is based on the first two and intends to provide an outlook of the direction(s) in which the market will move in the near future. This question will be answered with the results of all three surveys in mind.

Section 2 introduces the approach used here, its conduction, and its limitations. The various findings of the third survey are displayed in Section 3. Firstly, the distribution on every dimension is presented visually in histograms in Section 3.1 and several trends along the sample size are pointed out. Secondly, marginal tables which serve to shed light on the two questions of provider manifestation and data commoditization are presented as well as the results from the independence tests in Section 3.2. Section 4 looks at local as well as global trends; to this end, we first recapitulate findings from the previous survey in Section4.1; the global trends across all surveys are presented in Section 4.2. Finally, a conclusion is drawn in Section 5.

2 Methodology

The methodology of the survey reported in this paper is almost identical to the methodologies that were employed in the previous iterations in 2012 and 2013: Services that fulfill the provider definition are included in the sample and then categorized along dimensions based on an analysis of their respective web sites. Our approach is substantiated both in the need for comparability as well as in resource limitations. Two modifications to the previous methodology have been made for this iteration and are explained in Section 2.1: the refined provider definition developed in [27] is employed and the sample size has been expanded to allow for statistical analyses between the dimensions.

2.1 Provider Definition

For completeness' sake we repeat the provider definition of [27], according to which a provider must fit in one of the following categories:

1. The providers' primary business model needs to be providing data.

2. The providers offer an infrastructure to upload, browse or download machine-readable (e.g., RDF or XML) data to buy and sell. The data has to be hosted by the providers and it needs to be clear whether the specific data comes from the community or the operator. This type is the electronic marketplace in the narrow sense.

3. The providers offer or sell proprietary data they host themselves. Whether they have created or collected the data themselves or acquired it from other providers is not a criterion; there must, however, be transparency and traceability on the data sources. Providers of analyzed data must disclose the sources and methods of calculation.

4. The data analysis tools must be online tools and provide storable data as their main offering. They need to use proprietary data in their calculations; services like algorithm-based analyses of customers' data or the provision of crawling code do not classify for this type.

Additionally, several disqualifiers are necessary when trying to draw conclusions on the data marketplace as a distribution medium. Also, only legal, publicly accessible data providers are sur-

veyed.[1] One criterion is the machine-readability of the data. It can be argued that only machine-readable data successfully indicates the commoditization of data on marketplaces. Otherwise, the information is simply shared because users personally deem them useful without being concerned about allocability and effective distribution. This rule applies, for example, to Wikipedia: Its marketplace-like infrastructure allows users to upload or access information free of charge, which is not machine-readable though.

Data vendors only linking to data locations without hosting them (like KDnuggets.com's list of data sets) also do not fulfill the criteria. Providers that do not make their sources and methods transparent are excluded because no serious conclusions on their trustworthiness, on the data origin and sometimes even what type of data is offered can be drawn. Consumer credit agencies like `Schufa.de` or marketing agencies that rely on their data sources on relevant consumer groups as a unique selling point like `AlliantData.com` are common types of offerings excluded for that reason. Online tools that process the user's data or re-use them in software products without use of proprietary data like `OpenCalais.com` are other examples for providers not fully matching our criteria. Software-as-a-service (SaaS) suppliers are also disqualified based on the non-storability of their data.

Government agencies or non-government organizations providing free data are generally not considered as a data vendor, as they publish data as a side effect of their purpose in general and are not set on commoditizing data or even finding an appropriate business model. Nevertheless, data providers offering data collected originally by governments are still included in this survey.

A large number of cities, provinces, and countries – the Global Open Data Index counts 79 countries – participate in the Open Government movement [9]. This movement aims at publishing government data to allow for more transparent and citizen-orientated participation and innovation [19]. Transnational organizations like the United Nations or the World Bank and NGOs like interaction.org promote their objectives by sharing their findings. The evaluation of a small sample size yielded that all (non-)government agencies manifest in the same way with freely accessible data from a variety of sources within the institution in different formats. Due to the fact that the number of governmental and non-governmental organizations is quite large and they hardly differ in their services, they are generally disqualified. The research on this emerging field is still developing, two notable works are by [6] and [21]. Institutional data providers hold special relevance with regard to the number of participants as they are one of the few types that work in a consortium-like fashion.

Finally, financial institutions like stock exchanges are excluded from the survey as well, due to their sheer number: The World Federation of Exchanges alone counted 64 official and 15 affiliate members as of October 2014, not including the countless futures and online exchanges like `CMEgroup.com` [28]. They are disregarded entirely because it is impossible to reduce the sample set to a reasonable size.

2.2 Provider Acquisition

The provider lists of [22] and [26] formed the basis for the provider sample of this survey. As the statistical methods applied require a sufficiently large number of cases the sample was expanded from 47 to 72. To this end, ten keywords were identified by induction from their selection: "complex data analysis", "data crawler", "data market", "data marketplace", "data platform", "data provider", "data tagging", "data vendor", "data search engine", "sentiment analysis". They were taken as a basis for a keyword-based Web search and an analysis of the first 50 Google results. The results of the search and the lists of the previous surveys were matched against the criteria and a final list of 72 providers was assembled.

[1] Researchers claim an increasing proliferation of illegal data [20].

2.3 Limitations

The provider selection just described has several restrictions. Including a certain offering means including all of its competitors and similar offerings, leading to a sample size of possibly several thousand vendors which would go beyond the scope of this study. Consequently, a compromise between finding as many online offerings that provide data as possible while remaining within a reasonable and manageable scope is necessary which we achieve by a clear definition in conjunction with the presented disqualifiers.

The providers are evaluated solely by analyzing their online presences. Their self-portrayal on the respective websites does not necessarily reflect an objective assessment so the results of the survey can be biased. As a personal testing of the offerings cannot be covered due to resource restraints, an evaluation based on the Web presentation appears an appropriate solution. Four of the 15 categories are subjective so the results in these categories could be biased. Therefore, they are not included in the analyses and only serve to give an impression of the market and the providers.

Not all dimensions could be assessed for every provider so the data is not complete. Missing values are treated as N/A and disregarded in the analysis. Only 1.2% of all fields are counted as N/A, most of them in the Pricing, Data Access, and Data Output dimensions.

Even though the evaluation is intended as a continuation of previous surveys [22, 26], two changes lead to differences. First, the sample is significantly expanded while some of the previously surveyed providers are disregarded: Some, like Uberblic, are no longer in business, some no longer fit the provider definition, such as the governmental providers. Secondly, modifications to the dimensions of the preceding surveys have been undertaken. Dimension Website Language is removed entirely, while the related dimension Data Language is re-interpreted to strictly refer to the meta data available. It can be argued that the language reflects the specificity or nationalization of the data as in [22] but the national background of a provider does not necessarily imply a national focus of the data. Additionally, the dimension of Ownership has been added to allow for an analysis of the inherent bias of the providers. Depending on whether the operator allows other providers to participate on his platform, the business can be biased towards the operator [27].

2.4 Statistical Analysis Methods

The survey consists of categorical variables which only allow for positive (1) and negative (0) responses. As not all dimensions are exclusive, i.e., some dimensions are ”tick all that apply” questions, they are with methods for multiple response categorical variables (MRCVs) [3]. In a first step, two dimensions at a time are merged to show the combined responses to each category of the dimensions. The combinations of dimensions are picked based on three considerations: First, which dimension combinations return meaningful results at all, second which combinations can provide answers to the provider manifestation question and thirdly, which ones can give indicators as to the commoditization process of data. Traditionally, the commoditization of data can be inferred from knowledge about the competition situation and the standardization of data quality. Due to the fact that neither can be construed from web site evaluations, the associations between data domains serve to at least gauge information on the commoditization.

Based on the first consideration, subjective dimensions are left out because they are too imprecise to base analyses on. Additionally, not every combination of the 15 dimensions provides potential for meaningful interpretation. This is partly due to unrelated dimension combinations (like Ownership and Data Access) and partly due to the simultaneity of the data so for example the combination Data Output / Data Access returns a high number of providers offering CSV data via an API which is not a reasonable result.

The manifestations of providers as well as some indications on the competition situation can best be illustrated by the following combinations: Type / Domain shows whether certain business models are more likely to distribute a certain type of data. Whether certain business models obtain data from a specific source can be shown in Type / Origin. What audiences providers target can be inferred from Type / Audience. As an additional reference, Audience / Pricing is analyzed to determine which pricing models are more likely to be employed for different customer groups. Type / Pricing shows whether certain pricing strategies make more sense for some business models. Regarding the standardization of data, the evaluation is more difficult. Firstly, the source of certain data domains is regarded in Origin / Domain which could provide explanations on the specificity of the data. Secondly, the complexity of the surveyed data is considered in Time / Domain. The last combination, Pricing / Domain, looks at whether certain characteristics of data make more sense for certain pricing models. The results from that combination could shed light on the different purposes of data, i. e., whether it is procured regularly or only through spot-purchases. Other combinations (e. g., Domain / Access) have been tested but yield no meaningful results and are therefore not replicated here.

The tables in Section 3.2 are positive response tables computed with the MRCV package in R. R is a software environment for statistical analysis where codes typed into the interface are immediately executed. Through additional packages it is highly extensible. The MRCV package is focused on Multiple Response Categorical Variables and allows the computation of response tables with the following command: `marginal.table(data, I, J)`, where data refers to the target data set and I and J are respectively the variables to be counted [12]. The combination of the two variables is called manifestation combination and counts the number of observed cases for each positive manifestation combination. Additionally, it counts the according percentages based on the sample size. Because the providers in the sample size may manifest several times per dimension, both the absolute and the percentage margins do not sum up to the population size or 100%, as opposed to contingency tables.

Specifically, statistical independence among the individual variables is of interest. This is akin to asking the question "whether the probability of a positive response to each item changes depending on the responses to other questions" [4]. If two variables are statistically independent, knowing the value of one variable does not help to predict the value of another variable [23]. In the case of two MRCVs, independence means that each manifestation of a MRCV is independent of each manifestation of the other variable and that this holds for every response combination [3]. This hypothesis is called *simultaneous pairwise marginal independence* (SPMI) [3]. The independence is marginal because each manifestation is counted without regard to the other responses of the specific individual to the categories [4]. When testing for independence, neither the presumed association direction nor the "roles" of the variables are relevant [23]. This means that it is not important to know beforehand which variable is the explanatory one and which is the outcome variable.

The test for SPMI is computed in R with the MRCV package and the `MI.test(data, I, J, type, B=1999, summary.data=TRUE)` command. The test structure is briefly explained here, for more detailed explanations see [4] from where the following notation is derived.

Consider the case of two MRCVs W with I items and Y with J items. For item $i \in I$ and $j \in J$ the variables are referred to as W_i and Y_j respectively. Define the joint probability for $i = 1, \ldots, |I|$ and $j = 1, \ldots, |J|$ as $\gamma_{ij} = P(W_i = 1, Y_j = 1)$. Additionally, $\gamma_{i+} = P(W_i = 1)$ and $\gamma_{+j} = P(Y_j = 1)$ denote the marginal probabilities for the positive responses for the items. Let then the hypotheses for SPMI be:

H_0: $\gamma_{ij} = \gamma_{i+}\gamma_{+j}$ for $i = 1, \ldots, |I|$ and $j = 1, \ldots, |J|$
H_1: for at least one (i, j) pair the equality does not hold

where $\gamma_{ij} = \gamma_{i+}\gamma_{+j}$ specifies marginal independence. The hypotheses can be tested by a variety of different testing methods, including Rao-Scott Second-Order adjustments, Bonferroni adjustments, and bootstrap, a resampling algorithm under the assumption of independence [3]. For a large number of binary categories Rao-Scott adjustments are not realizable in R [12]. Bonferroni adjustments sometimes return more conservative critical values for small sample sizes while bootstrap p-values appear to have the highest power [3]. In order to show the results for both approaches, the Bonferroni p-values as well as p-values obtained under bootstrap are used.

3 Findings

This chapter presents the findings of the survey in two parts. Section 3.1 describes the 15 dimensions along which the providers have been classified and presents the findings in histogram form. Several trends are already pointed out and briefly compared to results of [22] and [26]. Section 3.2 presents the findings from the marginal tables and the tests for SPMI.

We mention that the two earlier surveys [22] [26] have established the initial framework which is used and developed further here. As will be seen, this framework allows meaningful comparisons of the results between the data providers over the years and to make predictions about future trends in this area.

3.1 Dimension Results in Histograms

The providers we considered are classified by 15 dimensions each of which consists of several categories. These dimensions originate from the [22] and [26] surveys and are split up into objective and subjective dimensions. The quantifiable, objective dimensions structure the surveyed data offerings into different types, while the subjective dimensions aim at capturing an impression of the respective company. All values are strictly Boolean, as an offering either fits a category or not. Most categories are not mutually exclusive and a single offering can cover several categories. When categories are exclusive it is pointed out in the respective description. The writing in the next two chapters is deliberate: When referring to a dimension, the name of that dimension is capitalized (e.g., Size). Categories are capitalized and italic (e.g., *Economic Data*). The setup of the histograms is identical for all figures: The abscissa maps the categories of the respective dimension and the ordinate maps the absolute number of cases. For the non-mutually exclusive categories the case numbers do not sum up to the 72 surveyed providers.

Type This dimension specifies the business model(s) of a data vendor. The categories are not mutually exclusive as one business model may cover several categories or one company may offer several services. *Crawlers* and *Customizable Crawlers* are offerings that search (crawl) a specific webpage or a set of webpages along links and extract data matching certain keywords into a given format. While *Crawlers* are bound to one domain of data, *Customizable Crawlers* can be set up to crawl for any content by the customer. A *Search Engine* returns lists of relevant content to the user's input of keywords. *Raw Data Vendors* offer data in a cleaned formatted way, usually in tables, but without further analysis. *Complex Data Vendors* in contrast process the data available in some way, for example by integrating various data sources or using statistical analysis. *Matching Data Services* sell the verification of customer input which they match against their own data, for example as address or business risk verification.

When data is merged, matched, or compared to other data, it is enriched and its value increases. Enrichment services differ from *Complex Data Vendors* in that they enrich the data by the customer's specification. *Enrichment – Tagging* provides meta data to mostly textual data by tagging additional information like geo coordinates to addresses or topics to Twitter posts. *Enrichment –*

Sentiment services capture sentiments and opinions towards a certain product or topics, usually based on social media mentions. *Enrichment – Analysis* are services that provide more additional information, using statistics or comparisons with historical data to enrich the data. *Data Marketplaces* as a category does not refer to the infrastructural phenomenon that is topic of the paper but rather the intuitive understanding of platforms with a high number of buyers and suppliers. When a marketplace operator also supplies its proprietary data on the marketplace, both *Data Marketplace* and the corresponding vendor category, mostly *Raw Data Vendor*, is ticked.

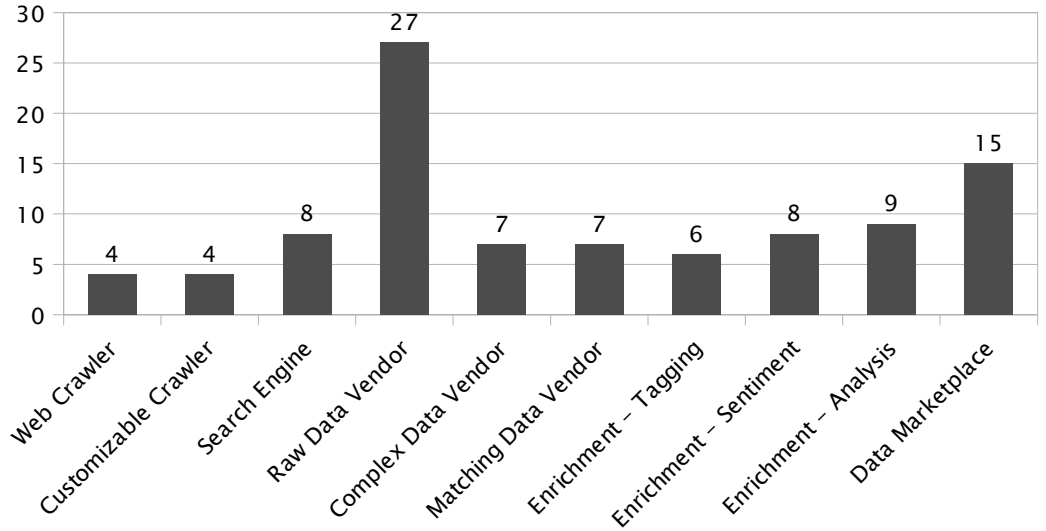

Figure 1: Histogram of Type in Frequency.

Figure 1 details the distribution of Type. *Raw Data Vendor* is the most commonly encountered type in the survey with a share of 37.5%, followed by *Data Marketplace* and *Enrichment – Analysis*, each with 20.8% and 12.5%, respectively. Fifty-two or 72% of the providers classify for a single category, 17 or 23.6% correspond to two categories, and the remaining three respond to three categories. This suggests that every category represents a sensible business model that can stand on its own. It could also be an indicator that most providers prefer to focus on a single offering without spreading their business model too far. The most common combinations of categories are among the enrichment services which make up for 23 counts with only 15 distinct providers.

Domain This dimension describes the area of application or topic of the offered data. Whereas the dimension is not mutually exclusive, the *Any* category is exclusive to classify data vendors that sell a variety of data unrestricted to any domain. *Economic Data* is data on stock markets, company developments, product information like pricing, and on specific economic sectors. *Scientific Data* describes data on environmental, pharmaceutical, medical, or scientific work or research. *Social Media* refers to the capturing of posts, tweets, opinions, and trends on social media. *Geo* is any data relating to maps, landscapes, and the geographical position of businesses or individuals expressed in coordinates. Contact data in the form of address lists, email lists, or customer information is categorized in *Address Data*.[2]

The domain distribution in Figure 2 shows that data without domain restrictions makes up for 29.2% of the data. Only 13.8% of the providers offer more than one domain of data which implies that most data providers specialize in only one domain. This possibly reflects the limitation to only one business model. On the other hand, the results from [22] and [26] clearly suggest a trend

[2]As opposed to [22] and [26] the *Economic Data* and the *Scientific Data* categories are renamed to clarify their content, their meaning remains.

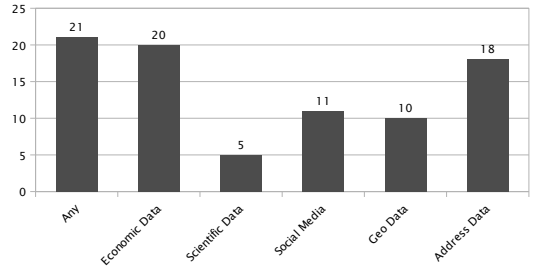

Figure 2: Histogram of Domain in Frequency.

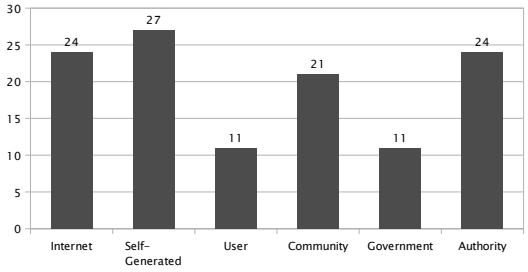

Figure 3: Histogram of Data Origin in Frequency.

towards *Any* data. As such, the results from this survey might be heavily influenced by the group of newcomers. The combinations among the providers with more than one domain are evenly distributed on *Economic / Geo / Address Data*, *Economic / Address Data*, and *Geo / Address Data*.

Data Origin The *Internet* as a data source means that providers manually or automatically collect data from other web pages and either sell the aggregation or further processing of the data. *Self-generated* sources either refer to services that assemble their data privately through patented generation and analysis methods or to services that gather the data from various other data sources not covered in the remaining categories, like news agencies. *User*-generated content means that users of the service need to provide some data input in order to receive the desired results. This category cannot stand on its own because all user input needs to be matched against some proprietary data. A service is categorized as *Community* when the data is supplied by the users like in a marketplace or in a crowdsourcing service or when the users can edit the supplied data. Data from *Governments* is official data collected by highly trustworthy sources like ministries or government agencies and distributed by the provider. *Authority* as a source describes data that is curated by some expert (organization), e. g., the Postal Office on addresses. Only institutional sources are recognized as authoritative, e. g., reputable journals like ”Nature” are not.

The results in Figure 3 show two opposing trends. On the one hand, reliable data from *Authorities* and governmental sources seems to be commonly used. On the other hand, despite their questionable reliability, self- and community generated data is relied upon by an even higher number of providers. It should be noted that ca. one fifth of the 40 providers that have only a single data source relies on self-generated data alone. Only 12 or 16.7% of the vendors use three or more data sources. While data and meta data available on the Internet remain a main data source for all providers due to their relatively effortless exploitation there seems to be a trend towards self-generation. This indicates that individualized data sources become a unique selling point.

Time Frame The currentness of the data and whether it needs to be updated regularly to remain valid is observed in this dimension. The categories are overlapping because different types of data may be offered. *Static/Factual* data are facts that are valid for longer periods of time such as macroeconomic statistics. *Up to Date* data like stock or social media data is only valuable for short periods of several days and needs to be updated regularly.

Figure 4 indicates a provider preference towards *Static* data. Of the surveyed providers, 26.4% sell both data types and only 13.9%, less than a fifth of the remaining providers, sell *Up to Date* data as their only offering. This may be attributable to several constraints of *Up to Date* data: Its collection requires more sophisticated setups and it is often collected without finding a buyer immediately, therefore demanding capacities without being used.

Pricing Model Some services provide their data for *Free*. In *Freemium* models a part of the service can be used for free before paying for a premium account or service. This category cannot stand on its own and is always in combination with the remaining two categories: *Pay-per-Use* or *Flat Rate*. The former counts the number of times a dataset is called via API queries or

access clicks. The latter charges a monthly or annual fee for the data access, sometimes with an amount limit.

The results of the Pricing Model dimension are presented in Figure 5: 54.2% of the providers offer only one pricing model, 25% offer *Freemium* in combination with another model, and 6.9% offer three or more pricing models. With 61.1% *Freemium / Flat Rate* is the most popular combination among the freemium models, followed by 27.8% offering both *Flat Rate* and *Pay-per-Use* in combination with *Freemium* and only 2 providers offering *Freemium* in combination with *Pay-per-Use*. Only one provider, the Microsoft Azure Marketplace, offers all pricing models. For the non-freemium providers, *Flat Rates* still take front rank before *Pay-per-Use* with 59% over 41%, though with a narrower margin. Its clear lead over other pricing models suggests that continuous access to data may take precedence over granular pricing. This is most likely due to provider preference for *Flat Rates* because those provide a higher certainty of revenue [16]. Whether customers also prefer the simpler pricing plans as opposed to individualized ones would be a valuable input to the issue of appropriate pricing models.

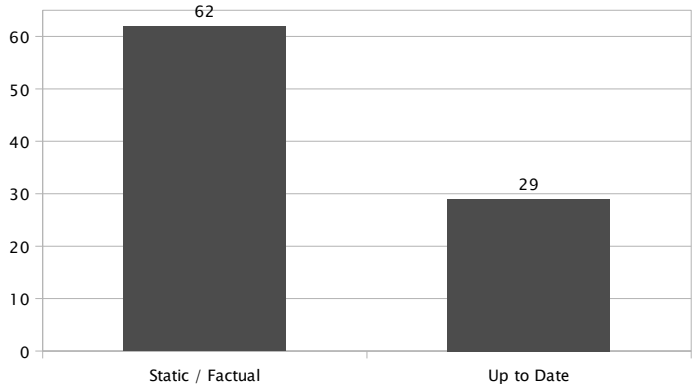

Figure 4: Histogram of Timeframe in Frequency.

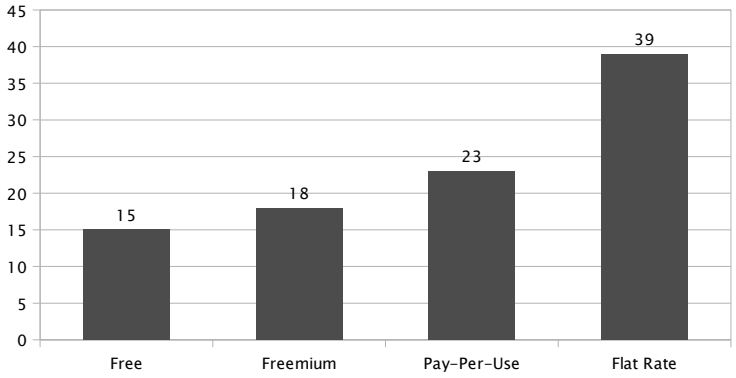

Figure 5: Histogram of Pricing in Frequency.

Data Access The data access dimension determines how the user can display and access the data. Most services offer several options. An *API*, an application programming interface, allows for seamless integration of the data provided into other software applications because it is not bound to a specific platform. *Download* requires no special prerequisites on the customer's side and provides clients with a reliable data access on their desktop. *Specialized Software* developed by the data provider helps examine, analyze, or visualize the data via software clients, mobile apps, or desktop applications. A *Web Interface* allows the customers to directly explore and use the data in the browser.

No clear trend towards a specific data access type can be identified in Figure 6: 18.1% of the providers offer only one data access, 41.7% two, 27.8% three and only 8.3% offer all access types. Since one third of the providers offers three or all data access types, data providers seem to identify a necessity to give customers more flexible data retrieval options. The relatively uniform distribution on *API*, *Download*, and *Web Interface* shows that both average users and more technically versed users are targeted.

Data Output Most services do not rely on a single output format but rather offer a combination of several data display and retrieval options. The exchange formats *XML* and *JSON* are used for structured data. *RDF* represents data in triples and is often used for graphical representation. Tabular data readable with most standard spreadsheet software is grouped in the *CSV/XLS* category. The *Report* category discloses all visualized data formats like PDF, DOC, or JPEG. Offerings that provide data in formats not covered by any of the output categories are treated as 0 values in all categories.

The notion of flexible data access points is somewhat countered by the results in Figure 7. Only 25% of the providers offer more than two data output formats. The most common combination is *CSV / JSON*, closely followed by *CSV / Report*. The high number of *CSV* data could possibly show that data providers aim at a convergence towards the mainstream market.

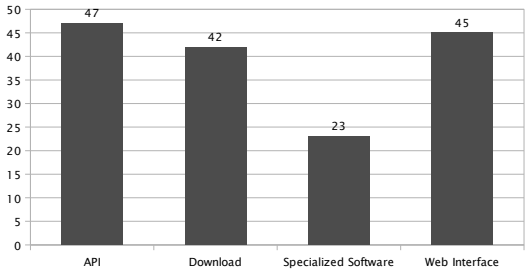

Figure 6: Histogram of Data Access in Frequency.

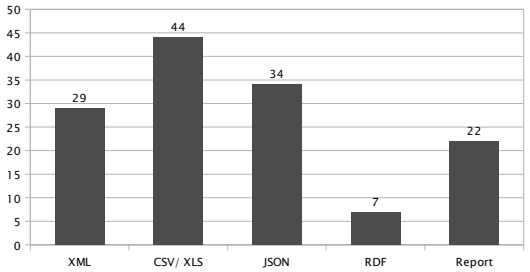

Figure 7: Histogram of Data Output in Frequency.

Data Language This dimension refers to the language of the meta data of the offered data. This includes names of columns and tables as well as localizations of units (e.g., Fahrenheit vs. Celsius). It is not relevant whether the data refers to information in other languages, only the language of the meta data itself is observed. The predominance of *English* and *German* is due to the selection criteria that have been applied. All other languages available have been counted and the three most frequently encountered have been added to the dimension, namely *Spanish*, *French*, and *Portuguese*. The *More* category is used when a data provider offers the data in a further language.

Apparently, a national focus does not mean that providers also translate the meta data which universally remains in English as the primary language for all data-related technology. This is further supported by the fact that only one data provider did not offer English data. Of course, the acquisition of new providers is based on English keywords which heavily skews their distribution, so these results should not be overemphasized.

Target Audience The clients of the providers surveyed are of concern in this dimension. Business-to-Business (B2B) services have other companies as their buyers and are categorized in *Business*. In business-to-consumers (B2C) the service is geared towards private persons interested in certain information and categorized in *Consumer*.

As can be seen in Figure 9, businesses remain the main customers of data providers. Of the 20 providers serving consumers, only 8 target them exclusively. It should be noted that several consumer-oriented providers are excluded from the survey, namely wikis and institutional websites geared towards citizens, which therefore skews the results. Just 18.1% of the providers target both customer types.

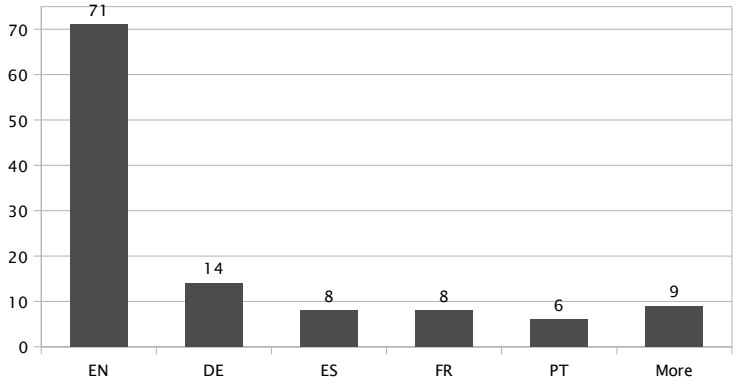

Figure 8: Histogram of Data Language in Frequency.

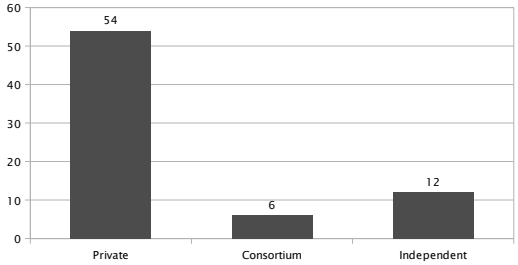

Figure 9: Histogram of Target Audience in Frequency.

Figure 10: Histogram of Ownership in Frequency.

Ownership The new dimension Ownership is introduced to evaluate whether the manifestation structures identified in [27] occur in the data marketplace. Services that control the flow and

price of the data offered are *Private*. Platforms operated by associations of several providers are categorized as a *Consortium*. Services simply providing the infrastructure for the marketplaces are *Independent*. In line with the model definitions, marketplaces where the provider also takes an active part on its own marketplace are not independent but are categorized as consortium marketplaces. Two providers, `Dayta.com` and `eXelate.com` offer separate services: a privately operated data service and an independent marketplace infrastructure where they themselves are not active. Due to the different roles they take on in the dual offerings they are included twice as distinct offerings.

The Ownership distribution is presented in Figure 10. Clearly, nearly all services are privately owned. Of the 15 observed marketplaces, nine are independently operated and 6 are consortium-based. The remaining three independent operators run search engines.

Pre-Purchase Testability The possibility of evaluating the offered services prior to a final purchase is rated in this dimension. These categories are mutually exclusive. With *None*, the buyer has to rely completely on the additional information without any means of previewing the data before buying. *Restricted Functions* means that only some functions of a tool are unlocked for the potential customer to preview. *Restricted Number/Volume* testability allows the customer access to the full functionality of the service but is limited to a fixed number of operations or a timeframe. *Complete access* means that every user can use all functions and features of the final product immediately or after registering.

Figure 11 details the distribution on Pre-Purchase Testability. Seeing that this dimension constitutes a main alleviation to buyers' uncertainty, the number of providers that do not offer any possibility to preview or test the data is surprisingly high. An additional group of providers relies on the assumption that a glimpse of their offering (i. e., *Restricted Functions*) is enough to convince potential customers. Together, they make up for 50% of the providers. The remainder provides at least limited access to the complete offering with the majority giving complete access. When considering that a portion of those are most likely free providers, the results clearly suggest that most providers hesitate to allow access to their data.

Pre-Purchase Information In this mutually exclusive dimension the information on the final product available is relevant. The amount rather than an even more subjective notion of information quality is the determinant for this subjective dimension. With *Barely Any* information, the potential customer has to guess the features of the service offered or – as with most services in this category – has to request more information via email. *Sparse Medial Information* refers to providers that give some information on the general features of their products without technical details or implementation instructions. *Comprehensive Medial Information* refers to services that provide a variety of information from demo videos to fact sheets, screenshots, or customer reviews.

Figure 12 visualizes the Pre-Purchase Information dimension. Two-thirds of the providers give out plenty of information on their precise offering and its functions in the form of videos and demonstrations. Only 26.4% and 6.9% of the services supply sparse or no information, respectively. In combination with the results from Figure 11, these results show that providers prefer to lower the high buyers' uncertainty through information rather than through previews of the data.

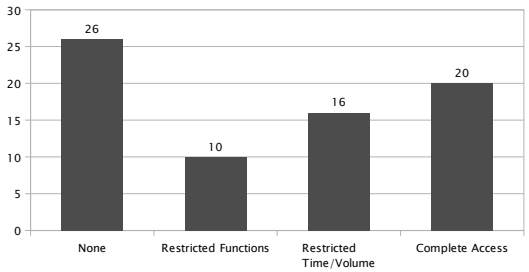

Figure 11: Histogram of Pre-Purchase Testability in Frequency.

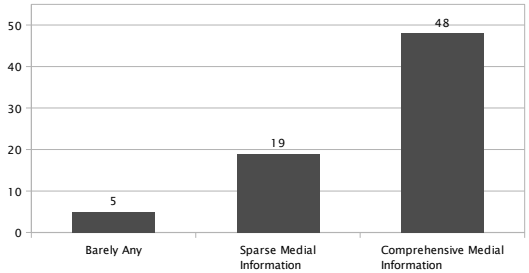

Figure 12: Histogram of Pre-Purchase Information in Frequency.

Trustworthiness This subjective dimension rates services on their trustworthiness, mainly based on the data sources. The detailed disclosure of data generation methods with named, reliable sources points towards a *High* trustworthiness. Services that only provide their general sources or rely on rather debatable sources indicate a *Medium* trustworthiness. Tagged as *Low* are offerings that do not even claim to provide complete and reliable data. Other factors include the level of sophistication of methods of data retrieval (basic crawling service vs. daily crawling with manual checkup) and the reputation of the vendor. The categories are not mutually exclusive to reflect different data qualities within one offering.

Figure 13 reflects the distribution of the Trust dimension. The relatively uniform distribution on *Low* and *Medium* combined and *High* does not allow for meaningful conclusions. Only eight providers check for more than one Trust category and are evenly spread on *Low / Medium* and *Medium / High*.

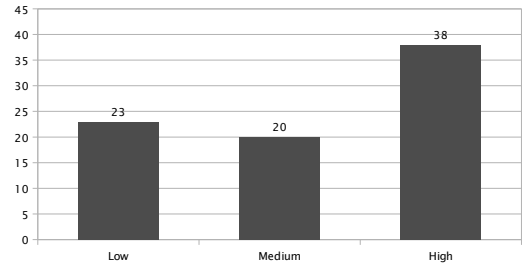

Figure 13: Histogram of Trustworthiness in Frequency.

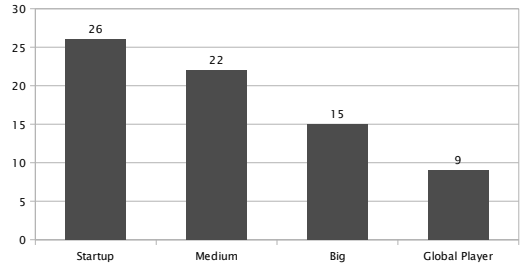

Figure 14: Histogram of Size in Frequency.

Size of Vendor As only the respective websites are evaluated, the estimation of size of a company is subjective and mutually exclusive. It should be noted that the size of the vendor refers to the company behind the concrete project so that a rather small project like `Freebase.com` is still categorized as ”Global Player” because Google operates it. *Startups* have only recently been founded by investors. *Medium* refers to businesses that have left the startup phase and established themselves in the market, usually with one key product. *Big* refers to vendors that have a well-established market position and cover a big market share with a variety of products. *Global Player* refers only to the biggest companies in the internet market such as IBM, Google, or Yahoo.

The results of the mutually exclusive Size dimension are visualized in Figure 14. A little more than a third of the providers are *Startups*. This indicates that the market provides market gaps which can be filled by first-movers. This potential for growth is balanced by the high share of established firms of all sizes which apparently still have sufficient possibilities for development. The combined results suggest a market in motion which has not yet exhausted all innovation potential. It should be noted that 16 of the 35 newly included providers are *Startups* which skews the results especially for this dimension.

Maturity This mutually exclusive dimension is subjective as well and refers to the stage of business development. *Research Projects* are rarely commercialized and refer to trials of projects or proof-of-concept websites. *Beta* projects are in development and sometimes already commercialized. *Medium* offerings provide a sophisticated data or service supply. A *High* maturity refers to a range of different, refined products.

Of the surveyed providers, 59.7% possess a *High* maturity. Additional 19.4% have a *Medium* maturity which indicates a generally high maturity among the offerings. The exact numbers can be found in Figure 15. Combined with the size results in Figure 14 the market could be tentatively characterized as innovative with sophisticated products.

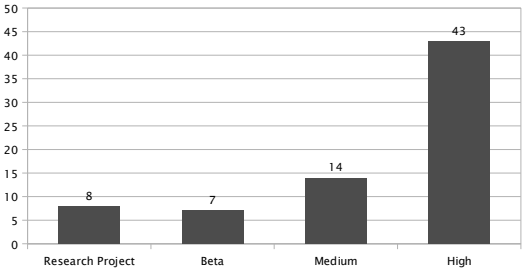

Figure 15: Histogram of Maturity in Frequency.

3.2 Statistics

The computation of marginal tables for multiple response categorical variables was carried out using the MRCV package of the R language, which provides functions for analyzing the association between one single response categorical variable and one multiple response categorical variable, or between two or three multiple response variables. As explained in Section 2.4, the underlying algorithm in particular counts the number of cases that apply to each combination (e. g., Scientific / Raw Data) and the according percentage. These numbers do not sum up to the population size or 100% because the offerings are usually not mutually exclusive. Additionally, the data is tested for simultaneous pairwise marginal independence (SPMI) with bootstrapping and Bonferroni adjustments in R. This means that for every single combination of the two dimensions, the hypothesis of independence between them is tested. If the returned p-value is below the confidence level of $\alpha = 0.05$, it can be assumed that they are independent.

Table 1: Marginal Table Type / Domain.

	Domain of Data											
	Any		Economic		Scientific		Social Media		Geo		Address	
Type	count	%	count	%	count	%	count	%	count	%	count	%
Web Crawler	1	1.39	1	1.39	0	0.00	2	2.78	0	0.00	0	0.00
Custom. Crawler	2	2.78	0	0.00	0	0.00	2	2.78	0	0.00	0	0.00
Search Engine	5	6.94	0	0.00	3	4.17	0	0.00	0	0.00	0	0.00
Raw Data	8	11.11	11	15.28	0	0.00	1	1.39	3	4.17	10	13.89
Complex Data	1	1.39	6	8.33	0	0.00	0	0.00	0	0.00	0	0.00
Matching Data	0	0.00	4	5.56	0	0.00	0	0.00	4	5.56	6	8.33
Enr. – Tagging	0	0.00	2	2.78	0	0.00	4	5.56	0	0.00	0	0.00
Enr. – Sentiment	0	0.00	0	0.00	0	0.00	8	11.11	0	0.00	0	0.00
Enr. – Analysis	1	1.39	2	2.78	0	0.00	6	8.33	1	1.39	1	1.39
Marketplace	8	11.11	0	0.00	2	2.78	0	0.00	3	4.17	3	4.17

$p.boot < 0.0005$ and $p.adj. < 0.0001$ are significant at a confidence level of $\alpha = 0.05$.

Table 1 shows the combinations of Type and Domain among the surveyed providers. The two most common combinations are *Raw Data Vendor / Economic Data* and *Raw Data Vendor / Address Data*. *Raw Data Vendor / Any*, *Marketplace / Any*, and *Enrichment – Sentiment / Social Media* tie for the third place. Those results are somewhat expected since *Raw Data Vendor* and *Economic Data* are among the most often encountered categories. Most data domains are distributed over a variety of different business models with the exception of *Scientific Data* which is distributed via only two distribution channels. Even though this domain may also be covered in the *Any* category (Thomson Reuters, for example, sells a variety of medical and pharmaceutical data) it is evident that *Scientific Data* is not only rarely sold as a standalone product but also through only a limited variety of providers.

The combined results of Type / Origin in Table 2 confirm some intuitive speculations: Enrichment services and crawlers collect their information on the Internet while marketplaces provide mainly community-curated data. One result is somewhat misleading: It appears at a first glance that the majority of *Raw Data Vendors*, the category that most providers match, collects their data themselves which could indicate a demand for specialized, not yet publicly available data. However, only six providers depend on the self-generated raw data alone which means that the true majority aggregates online, federal, and institutional sources which indicates a demand for aggregated, cleaned data.

The independence hypothesis can be rejected for three combinations, which means they are highly correlated. These combinations are: *Matching Data / User*, *Marketplace / Community*, and *Enrichment – Sentiment / Internet*.

Table 2: Marginal Table Type / Origin.

	Origin											
	Internet		Self-Generated		User		Community		Government		Authority	
Type	Count	%	Count	%	Count	%	Count	%	Count	%	Count	%
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

$p.boot < 0.0005$ and $p.adj. < 0.0001$ are significant at a confidence level of $\alpha = 0.05$.

Table 3: Marginal Table Type / Target Audience.

	Target Audience			
	Business		Customer	
Type	Count	%	Count	%
Web Crawler	4	5.56	0	0.00
Custom. Crawler	4	5.56	0	0.00
Search Engine	5	6.94	7	9.72
Raw Data	25	34.72	6	8.33
Complex Data	7	9.72	1	1.39
Matching Data	7	9.72	0	0.00
Enr. – Tagging	6	8.33	0	0.00
Enr. – Sentiment	8	11.11	0	0.00
Enr. – Analysis	9	12.50	0	0.00
Marketplace	13	18.06	8	11.11

$p.boot = 0.0161$ and $p.adj. = 0.0013$ are significant at a confidence level of $\alpha = 0.05$.

As evident from Table 3 which maps the dimensions Type and Target Audience, only a limited number of provider types is concerned with private customers. Only *Search Engines* show an almost even distribution on both audiences. The association is significant at a confidence level of $\alpha = 0.05$ for all methods, with *Search Engine / Customer* as the only significant combination.

In Table 4 one can see that specialized data in the sense of focused on a specific domain is rarely given away for free. With the exception of *Scientific Data* with 80% free distributions virtually none of the other domains are distributed free of charge. The *Any* category shows no clear trend with its even distribution on the pricing models. *Social Media* and *Economic Data* tend to be priced in flat rates which makes sense given that the majority of them needs to be updated regularly. The

Table 4: Marginal Table Pricing / Domain.

	Domain											
	Any		Economic		Scientific		Social Media		Geo Data		Address Data	
Pricing	Count	%	Count	%	Count	%	Count	%	Count	%	Count	%
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

$p.boot < 0.0080$ and $p.adj. = 0.008$ are significant at a confidence level of $\alpha = 0.05$.

hypothesis of independence can be rejected at a confidence level of $\alpha = 0.05$ with *Free / Any* and *Free / Scientific Data* as the two significant combinations.

Table 5: Marginal Table Audience / Pricing Model.

	Pricing Model							
	Free		Freemium		Pay-per-Use		Flat Rate	
Audience	Count	%	Count	%	Count	%	Count	%
Business	8	11.11	18	25.00	23	31.94	39	54.17
Customer	15	20.83	5	6.94	1	1.39	6	8.33

$p.boot < 0.0005$ and $p.adj. < 0.0001$ are significant at a confidence level of $\alpha = 0.05$.

Table 5 shows that less than a third of the offerings geared towards private customers charge for the data. Virtually all of the remaining services offer a *Freemium* model. When bearing in mind that only 9.7% of the providers serve exclusively private customers, it becomes apparent that surveyed providers focus solely on B2B relations. For *Business* customers, fees seem to be the norm. Considering that they are the most common combination, *Freemium* and *Flat Rate* models could represent a strategy to accustom customers to the data offering and make use of lock-in effects. The hypothesis of independence of the two dimensions can be rejected for all combinations between audience and pricing model.

Table 6: Marginal Table Type / Pricing Model.

	Pricing Model							
	Free		Freemium		Pay-per-Use		Flat Rate	
Type	Count	%	Count	%	Count	%	Count	%
Web Crawler	0	0.00	2	2.78	1	1.39	4	5.56
Custom. Crawler	0	0.00	1	1.39	2	2.78	3	4.17
Search Engine	5	6.94	2	2.78	1	1.39	2	2.78
Raw Data	3	4.17	9	12.50	7	9.72	19	26.39
Complex Data	1	1.39	2	2.78	3	4.17	4	5.56
Matching Data	0	0.00	4	5.56	5	6.94	5	6.94
Enr. – Tagging	0	0.00	0	0.00	4	5.56	3	4.17
Enr. – Sentiment	0	0.00	1	1.39	1	1.39	7	9.72
Enr. – Analysis	0	0.00	2	2.78	0	0.00	6	8.33
Marketplace	7	9.72	2	2.78	4	5.56	4	5.56

At a confidence level of $\alpha = 0.05$, $p.boot = 0.0021$ is significant, $p.adj. = 0.1327$ is not.

In Table 6 it can be seen that some types of data providers prefer certain pricing models. Some of the previously identified associations between dimensions provide possible explanations for this: enrichment services mainly sell *Social Media* data from *Internet* sources (which in turn are closely

associated as well) which favors *Flat Rates* (as evident from the Tables 1, 2 and 8). Those cross-associations suggest that similar business models manifest in the same way across dimensions.

The more conservative Bonferroni adjusted $p.adj.$-value is not significant at a confidence level of $\alpha = 0.05$ and no significant combination could be found which indicates that the association between Pricing Model and Type is not strong.

Regarding the *Raw Data Vendors*, the clear trend towards *Flat Rate* and *Freemium* (which is mainly distributed on *Flat Rates*) indicates that a constant supply to data represents an important selling point. *Marketplaces* have the most diverse pricing models with nearly half of the 15 providers offering their data for free while the other half is evenly distributed on *Pay-Per-Use* and *Flat Rate*. The lack of certain results is also due to the methodology: *Web Crawlers* that provide their code free of charge are excluded from the survey due to the lack of proprietary data (and non-profit crawlers could not be found) so no combination of those two categories is observed.

Table 7: Marginal Table Origin / Domain.

	Domain of Data											
	Any		Economic		Scientific		Social Media		Geo		Address	
Origin	Count	%	Count	%	Count	%	Count	%	Count	%	Count	%
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

$p.boot < 0.001$ and $p.adj. < 0.0001$ are significant at a confidence level of $\alpha = 0.05$.

Table 7 shows that some domains draw from a variety of sources whereas others are rather restricted to a specific type of source. For most domains this allocation is natural, i. e., only the *Internet* for *Social Media* or all data sources for *Any*. *Economic Data* is mainly derived from authoritative and individual data sources and only rarely from “freely” available sources. Again, this might either reflect the different data types contained in that category or rather point towards *Self-Generated* data as a distinguishing feature for competitive advantage. The most common combination of Domain / Origin is *Address Data* and *Self-Generating* which implies only a dim transparency on the sourcing process of address data on the internet. *Any* data is mostly coming from *Communities* which could indicate that low participation barriers lead to unrestricted data domains. The hypothesis of independence can be rejected at a confidence level of $\alpha = 0.05$ for all methods with *Social Media / Internet* as the only significant combination.

Table 8: Marginal Table Timeframe / Domain.

	Domain of Data											
	Any		Economic		Scientific		Social Media		Geo		Address	
Timeframe	Count	%	Count	%	Count	%	Count	%	Count	%	Count	%
Static/Factual	20	27.78	19	26.39	5	6.94	3	4.17	10	13.89	18	25.00
Up to Date	7	9.72	11	15.28	0	0.00	11	15.28	1	1.39	2	2.78

$p.boot < 0.001$ and $p.adj. < 0.0001$ are significant at a confidence level of $\alpha = 0.05$.

The results of the combination Domain / Timeframe in Table 8 mostly reflect the inherent time dependence of the data domains. *Economic Data* is the exception with its segmentation into both long- and short-term timeframes. This could either indicate that *Economic Data* possesses varied information validity or simply be due to the different data types in this category, both factual information on economies and daily-changing stock data. The hypothesis of independence can clearly be rejected at a confidence level of $\alpha = 0.05$.

4 Trends

In this section the results from the two preceding surveys will briefly be summarized and checked as to whether a continuance of trends can be recognized. Contradicting results could either be attributed to the adjustment of methodology or indeed to a market change. The section also looks at trends we were able to identify, first in the previous two surveys, then in all three taken together, which even allows for an outline of potential future scenarios.

4.1 Previous Surveys

In an effort to help closing the research gap we had identified regarding data marketplaces, we have conducted two earlier surveys in 2012 [22] and in 2013 [26], where we have investigated offerings by data marketplace platforms and data vendors based on their respective web sites. These investigations have been performed manually through categorization of vendors along the set of dimensions that was described above in Section 3.1.

In those surveys, we found that the amount of offerings with raw and unstructured data has declined; by contrast, processed and high-quality data offerings have meanwhile become increasingly popular. Among the other dimensions, a trend towards more diversification could be observed: The number of different types of vendors participating in the market has increased. Additionally, the data access methods have been continuously expanded, towards sophisticated forms such as Web interfaces and reports, allowing the data to be easily accessed by business staff. Also, many vendors tend to offer multiple interfaces, so that customers can choose which one suits them best. Moreover, the languages of the data and the web sites became more diverse.

Another interesting observation has been the increasing popularity of up-to-date data offerings, which underlines the observation made elsewhere (e.g., [17, 24, 25]) that timeliness is a core factor influencing the value of data. Additionally, data from unconventional data sources could be found more often. This includes data sources such as *community data* (e.g., data from Wikipedia), which was considered considered less trustworthy in the past.

Generally, from the previous two surveys it could be concluded that the market for data and data-related services was subject to constant change and hence is by no means mature yet. This was evident by the number of market participants leaving and entering the market as well as by the changes in their respective business model and core offering.

This third iteration of our surveys has aimed at continuing the research approach pursued so far. The main difference between the latest survey and the previous ones is a change in the provider definition based on previous work by the authors [27]; this is accompanied by a change in the number of surveyed subjects. These changes have been elaborated upon in more detail in the previous sections.

4.2 Global Trends

When looking at the results of the surveys over the course of the last three years, five global trends can be identified:[3]

1. Some provider manifestations seem to make more sense than others: Enrichment providers often cover sentiment analysis and other enrichment services of social media, sourced from the

[3]The *emphasized* writing of categories will be dropped from now on to improve readability.

Internet and sold through flat rates. Another common type is matching data services, which use user- and self-generated data to match addresses, geographical, and economic data. Generally, most providers focus on only one category (73.6%) and limit themselves to only one domain (89%) and one data source (56.9%). This indicates that the providers split themselves into two groups: Hierarchical ("vertical") providers with only a single domain offering and intermediate ("horizontal") platforms where unrestricted data domains can be acquired. Community contributions on marketplaces result in data on a variety of topics.

2. The growing significance of unique data is evident from the increase of self-generated data. Providers who specialize in one domain rarely give their data away for free and usually charge a fee. In light of the fact that the market is mainly a B2B one, this is little surprising. Regarding data origin, a clear move towards self-generation can be identified. In the survey from 2012, Internet sources make up for half of the observed sources, while the survey from 2013 finds an overproportional increase of self-generated, community and user sources. As self-generated data is rarely employed as the only source, those sources represent a point of differentiation among the competition. This development also indicates that providers decrease their efforts in reselling data available on the Internet and move to individualized data sources. Interviews with data providers have also shown that some customers have complex demands which are not satisfied with currently available data [16].

3. The clear advancement of flat rates over pay-per-use is somewhat unexpected when compared to the other surveys where those two types lie level with each other. Providers clearly prefer flat rates due to their steadier revenues and usually combine them with freemium models to reduce uncertainty and take advantage of lock-in effects. Furthermore, pay-per-use models have not (yet) reached the level of sophistication necessary to prevent arbitrage exploitation. To find technical and policy amendments, research has been conducted and has presented in [2, 11]. Customers favor simpler pricing models as well and are not satisfied with granular pricing models that restrict unfocused data exploration. This is supported by the pricing development on the private TV sector: Flat rate models as offered by, e.g., Netflix are far more successful than models where customers have to decide on their willingness to pay for every single movie. Also, providers still have plenty of options for differentiation potential so the pricing competition is not very pronounced [16].

The developments on other data-related markets like private TV can indeed serve as reference points for future directions of this aspect of data markets. Private TV channels have experimented extensively in the past years regarding the pricing of their offer: From full packages to single channel selection to single item offerings like movies or shows - with varying success. However, on-demand services have recently become much more popular and it can be seen that in the wake of this trend, private TV has benefited immensely from the change both in Zeitgeist and market conditions. Overall, there is a trend towards flat-rate-based pricing models for digital media content in general. This is evident when considering success stories such as Spotify for music streaming, Netflix for video streaming, or even Amazon's Kindle Unlimited for eBooks. This trend will continue, such that flat-rates will be even more dominant as a pricing model for data and data-related services.

4. The results from the ownership dimension indicate that hierarchical ("vertical") relations still dominate the data market. The low number of intermediaries shows that the efficiency of the market is still limited and that data products are very differentiated.

5. The occurrence of data access types has changed over the last years, away from APIs which were originally dominating the field; Web exchange formats like JSON and XML gained importance, only to be surpassed by CSV data this year. Although this could be related to the sample, the likewise high number of report formats allows for two possible explanations. Either, as argued in [26], these two results point towards more processed data or, when considering the high number of raw data vendors, this indicates that the providers aim at making the data more available to non-technical users. The development of data format and access options suggests an orientation of the market towards a mainstream market that is also targeting non-technical companies and users: A high number of providers offers several, some even all, access possibilities but limits the number of data formats. The restriction to mostly standard formats like reports or CSV probably

aims at reducing presuppositions on data use. The high number of API accesses indicates that this development does most likely not involve a withdrawal from the initial target group.

With regard to the size of the providers, an interesting observation can be made when looking at the progression across the surveys. Initially, the market consisted mainly of bigger, established companies originating from other soft- and hardware related industries. Over the years, that domination has diminished, as the market became more diverse with providers of different sizes and especially new companies participating. Through the extension of the sample, some of the new entrants are now included and surveyed as well, as evident by the high number of startups in the newcomer group. The combined findings allow for the suggestion that initially the market had rather high entry barriers. This gave advantages to established companies that could raise the necessary investments and quickly establish a relevant market share. Ever since the first survey, the entry barriers have clearly lowered and now allow startups to form and join the market. Since electronic markets are considered to have low entry barriers this is just one possibility but probably the most likely. This development is supported by a growing number of startups that consume data from data markets[4].

The collective entry of startups does not contradict the finding of a growing and maturing market. Quite to the contrary, their development insinuates that the trading of data through intermediaries is now established and investors are willing to fund innovative new concepts. The tendencies in the maturity dimension confirm this. Despite the new providers the direction towards a high maturity is constant throughout all surveys. This means that startups show a high maturity as well and start off with sophisticated business models. All this suggests that the market has settled from its initial launch phase into a more stable but still highly innovative phase where both newcomers as well as established data providers find plenty of potential for development. This phase is accompanied by a high fluctuation of providers that enter and leave the market, also evident in the sample with seven closed services since 2012.

Concerning the standardization of data quality, the trend identified in [26] towards processed data opposes the one in this survey. One explanation for this presumes a market development towards raw data away from processed data. Another explanation suggests different types of data which satisfy different demands. The second explanation is backed by the observation of diversified data sources, specifically towards self-generated and individualized sources. The high number of raw data vendors in this year's iteration of the survey as well as the high number of processed data suppliers last year allow the conclusion that data providers predict the same trends.

When applying the second explanation that presumes two data demands, commoditization gains relevance. A commoditization of individualized data with a high specificity is presumably undesirable for the consumers. As such, an intensification of commoditization for that group is unlikely. In the case of the first group, data of constant quality, a convergence towards commodities would likely accelerate and amplify its exchange. As presented in [27], the more standardized a product is, the lower the costs of implementation are, and the more likely its purchase on a marketplace is. This would entail a more competitive market for that group. The most important indicator for that development will probably be the development of the Ownership dimension. Intermediary platforms will proliferate and represent that development. Due to the fact that competition and commoditization are highly interdependent, their parallel advancement would presumably catalyze the commoditization of data further.

[4] see, for example, http://mlwave.com/ycombinator-2014-data-science-start-ups/ or http://www.kdnuggets.com/2013/05/42-big-data-startups.html.

4.3 Emerging Scenarios

The future development of data markets remains dependent on the resolution to ARROW's information paradox. This paradox states that if a customer wants to evaluate the quality and value of information he needs to examine the information itself before purchasing it which he cannot as he would then have gotten the information for free [1]. Today, more than 50% of the providers offer at most an excerpt of their offering and are very reluctant to provide potential customers with previews of their data. Apparently they are aware of this obstacle and aim to reduce the buyers' uncertainty through information provision on the data. However, the content of data is far more relevant than the functionality of the accompanying API so that it is rather unlikely that pre-purchase information supplied by the seller is sufficient to completely resolve all uncertainty on the buyers' side.

One of the most interesting results are the seemingly opposing notions of a trend towards processed data in [26] and this year's trend towards raw data. The most simple reason for the occurrence of these results would be that the market has changed its direction. This explanation can be traced back to the development and expansion of the market in general in the last two years: Originally dominated by larger companies from other industries, the market is now diversified through a number of startups that entered the market. While this is a plausible suggestion, another suggestion is presented in [16]. In interviews with data providers, two customer demands in data acquisition are distinguished: a first one in which customers expect complete, formatted, and reliable data; and a second one in which customers are not dependent on the quality of the data and rather wish for tendencies and answers to be integrated into the decision making of companies [16].

When extending that idea further, two different scenarios can be developed. In the first scenario, the data is used as a type of manufacturing input. In order to process the acquired data further and use it as a basis for the production of another good, its quality must be extremely high and the access to it must be reliable. Especially the growing importance of data in the medical and pharmaceutical sector supports this notion, as does, for example, the emerging area of 3D printed cars. In the second scenario, the data is considered an add-on and a specialized product that can be spot purchased whenever necessary or be acquired on a regular basis. Its quality is not of crucial importance compared to the importance of its specificity. An example of such a demand could be 3D printing files (other than for cars). In the add-on scenario, customers expect a higher individuality of the product to match their particular wishes while data buyers in the first scenario would more likely expect a constant standard which they can depend on. Examples of the first scenario are the financial data APIs offered by `Xignite.com`, `BloombergPolarLake.com`, or `InteractiveData.com`. The specialized inputs in the second scenario could be some enrichment services like `CrowdSource.com`, crawling services like `80legs.com` or address sellers like `xDayta.com`.

Under the assumption of those two scenarios, the opposing directions can be resolved and explained. Additionally, this explanation is backed by the continually inconclusive trustworthiness dimension which takes shape in both high and low trustworthiness. Also, the origin dimension, which shows an importance of both highly reliable sources like authorities as well as the strong increase in self-generated data, could point towards the second scenario.

Clearly, this explanation is not exhaustive. Several providers like the address validation tools fall into neither category or one would have difficulty deciding for one category like in the case of social media analytics. Some are obviously spot-purchase oriented like `VICO-Research.com` but other like `Gnip.com` could serve customers both as a regular pillar of information in business or be only an add-on information service. Nevertheless, it provides an interesting perspective on the different data types demanded and insinuates that not only high quality data is demanded.

5 Conclusions

In this paper we have reported on the third iteration of our data marketplace study. We have presented its results in graphical as well as statistical form, using the formal foundations we have established elsewhere as a yardstick. Finally, we have compared our recent results to earlier ones obtained in previous years, we has allowed us to identify trends and outline future scenarios.

Concluding the third iteration of the data market survey, one result has been obvious: ARROW's information paradox remains the major obstacle to data trading. The empirical and qualitative data confirm that providers are very reluctant to share information about their data before a business deal. As long as this issue prevails, the pricing of data remains far from its competitive price.

Regarding the next five years, the current trend towards flat-rate-based pricing models is going to change. Even today, a diversification of providers can be observed, as established providers mature and innovative new companies and business models emerge. A trend towards the mainstream market for non-technical companies and subsequently non-technical staff can be observed which involves data formats that are easily accessible through downloads and Web interfaces are becoming more common. The original market, however, remains relevant. One business model that is currently emerging allows consumers to directly sell their personal data (e.g., from fitness trackers) for profit on platforms, `handshake.uk.com` and `datacoup.com` being among the first companies to offer this.

Automated surveillance and analysis of social media data, however, is the most promising business model to become the next "big thing" within the data community. As all social networks continue to grow and expand their offerings themselves, companies become increasingly reliant on observing what happens when it happens. Services that cater those demands like Gnip can offer value to those companies, which they will certainly be willing to pay for. As such, the value of data is likely to become a normal thing and expectations that "all information on the Internet is free" need to adjust (or will fade away anyway). On the other hand, it remains to be seen whether an appropriate exploitation of surveillance data can indeed be made to scale.

For the time being, most business models on the Internet can easily be identified, as most of them embody the virtual translation of previously existing industries like, for example, contact data selling or business partner verification. Contradicting the perception that entrepreneurs entering the data market will always be innovative, most business models so far stick with specific, consolidated business models that promise secure revenue opportunities, an observation that does not apply to the Internet at large. Although the data procurement has moved to the market as evident from the publicly accessible web sites surveyed, "real" intermediaries in the sense of open platforms are still rather rare. Most providers seem to prefer hierarchical relations.

Regarding the commoditization, data products are still highly differentiated and not in direct competition with each other. Data is still a highly individualized good and it is hard to compare different data sets with each other. Nevertheless, the data format is being homogenized with the observed rise of standards like XML and JSON. This leads to easier data handling and processing for customers, and lower overall costs for data integration. A similar progression has been observed in the international freight traffic, which has been substantially simplified through the usage of standardized containers, e.g., on ships and trucks. Furthermore, the high number of raw data vendors indicates that the market moves in that direction and that data will become more of a commodity. The data market will become more competitive and pricing models will become even more relevant. This is especially relevant for static and factual data because the marginal cost for an additional copy of the product are virtually zero, which potentially leads to existence-threatening price competitions. The consequences this might have on the willingness to pay on the consumers' side will be interesting to observe.

As for the future of the data demands mentioned earlier, the demand for data as manufacturing input will remain within private business relationships between large-scale providers. Specialized data on the other hand has the potential to be provided and purchased on intermediary platforms. Statista and their infographic service targeted towards newspapers is a good example as publishers can purchase the specific information and re-use it for their purposes.

Through large Internet companies, such as Google and Facebook, as well as the emergence of the Web 2.0 (also known as the read-write Web), the value of personal data has gained public awareness as these companies generate large amounts of revenue from it. Similarly, such a development is about to take place with regards to personal medical data; this is evident, for instance, by the fact that a German insurance company has started to subsidize the purchase of the Apple Watches [5].

Concluding this trilogy of data market surveys, we have provided a comprehensive overview of the market as well as predicted important future trends. Our focus was mostly on vendors, however, considering also buyers has the potential to contribute further insights and remains an open issue.

References

[1] Kenneth. J. Arrow. Economic welfare and the allocation of resources for invention. In National Bureau of Economic Research, editor, *The rate and direction of inventive activity: Economic and social factors*, pages 609–626. Nber, 1962.

[2] M. Balazinska, B. Howe, and D. Suciu. Data markets in the cloud: An opportunity for the database community. In *Proc. of the VLDB Endowment*, volume 4, page 12, 2011.

[3] C. R. Bilder and T. M. Loughin. Testing for marginal independence between two categorical variables with multiple responses. *Biometrics*, 60(1):241–248, 2004.

[4] C. R. Bilder and T. M. Loughin. *Analysis of Categorical Data with R*. Taylor & Francis, 2014.

[5] Guido Bohsem. Zuschuss für die apple-watch. *Süddeutsche Zeitung*, 2015.

[6] S. A. Chun, S. Shulman, R. Sandoval, and E. Hovy. Government 2.0: Making connections between citizens, data and government. *Information Polity*, 15(1/2):1–9, 2010.

[7] E. Dumbill. Data markets compared, 2012. URL `http://radar.oreilly.com/2012/03/data-markets-survey.html`. Last accessed: 2014-11-24.

[8] H. Gislason. The emerging field of data markets – our competitive landscape, 2011. URL `https://blog.datamarket.com/2011/02/25/the-emerging-field-of-data-markets-our-competitive-landscape/`. Last accessed: 2014-11-24.

[9] Global Open Data Index. Global open data index, 2014. URL `http://global.census.okfn.org`. Last accessed: 2014-10-21.

[10] Robert Kosara. The rise and fall of swivel.com, 2010. URL `https://eagereyes.org/criticism/the-rise-and-fall-of-swivel`. Last accessed: 2014-11-20.

[11] Paraschos Koutris, Prasang Upadhyaya, Magdalena Balazinska, Bill Howe, and Dan Suciu. Query-based data pricing. In *PODS*, pages 167–178, 2012.

[12] N. A. Koziol and C. R. Bilder. *MRCV: Methods for Analyzing Multiple Response Categorical Variables (MRCVs)*, 2014. http://CRAN.R-project.org/package=MRCV.

[13] K. B. Lee, S. Yu, and S. J. Kim. Analysis of pricing strategies for e-business companies providing information goods and services. *Computers & Industrial Engineering*, 51(1):72–78, 2006.

[14] P. Miller. Podcasts, 2012. URL `http://cloudofdata.com/category/podcast/`. Last accessed: 2014-11-24.

[15] Paul Miller. Nurturing the market for data markets, 2012. URL http://cloudofdata.com/2012/01/nurturing-the-market-for-data-markets/. Last accessed: 2014-11-24.

[16] A. Muschalle, F. Stahl, A. Löser, and G. Vossen. Pricing approaches for data markets. In M. Castellanos, U. Dayal, and E. A. Rundensteiner, editors, *Enabling Real-Time Business Intelligence*, pages 129–144. Springer, 2013.

[17] Felix Naumann. *Quality-Driven Query Answering for Integrated Information Systems*, volume 2261 of *Lecture Notes in Computer Science*. Springer, 2002. ISBN 3-540-43349-X.

[18] S. O'Grady. What's holding back the age of data, 2011. URL http://redmonk.com/sogrady/2011/12/08/holding-back-the-age-of-data/. Last accessed: 2014-11-24.

[19] T. O'Reilly. Government as a platform. In D. Lathrop and L. Ruma, editors, *Open Government: Collaboration, Transparency, and Participation in Practice*, pages 11–40. O'Reilly Media, 2010.

[20] F. Pasquale. The dark market for personal data, 2014. URL http://www.nytimes.com/2014/10/17/opinion/the-dark-market-for-personal-data.html?_r=0. Last accessed: 2014-11-24.

[21] H. J. Scholl, M. Janssen, M. A. Wimmer, C. E. Moe, and L. S. Flak. *Electronic Government*. Springer, 2012.

[22] F. Schomm, F. Stahl, and G. Vossen. Marketplaces for data: an initial survey. *ACM SIGMOD Record*, 42(1):15–26, 2013.

[23] H. J. Seltman. *Experimental design and analysis*. 2014. http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf.

[24] C. Shapiro and H. R. Varian. Versioning: The smart way to sell information. *Harvard Business Review*, 77(6):106–114, 1998.

[25] Florian Stahl. *High-Quality Web Information Provisioning and Quality-Based Data Pricing*. PhD thesis, University of Mnster, 2015.

[26] Florian Stahl, Fabian Schomm, and Gottfried Vossen. Data marketplaces: An emerging species. In Hele-Mai Haav, Ahto Kalja, and Tarmo Robal, editors, *Databases and Information Systems VIII*, pages 145–158. IOS Press, 2014.

[27] Lara Vomfell, Florian Stahl, Fabian Schomm, and Gottfried Vossen. A classification framework for data marketplaces. Technical Report 23, ERCIS – European Research Center for Information Systems, Münster, 2015.

[28] World Federation of Exchanges. World federation of exchanges, 2014. URL http://world-exchanges.org/member-exchanges. Last accessed: 2014-10-23.