ERCIS

European
Research
Center for
Information
Systems

Working Paper No. 14

Stahl, F. ■
Schomm, F. ■
Vossen, G. ■

Marketplaces for Data: An Initial Survey

ERCIS – European Research Center for Information Systems

Editors: J. Becker, K. Backhaus, H. L. Grob, B. Hellingrath, T. Hoeren, S. Klein, H. Kuchen, U. Müller-Funk, U. W. Thonemann, G. Vossen

Westfälische Wilhelms-Universität Münster

Working Papers

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

Working Paper No. 14

Marketplaces for Data: An Initial Survey

Florian Stahl, Fabian Schomm, Gottfried Vossen

Abstract

Data is becoming more and more of a commodity, so that it is not surprising that data has reached the status of tradable goods. An increasing number of data providers is recognizing this and is consequently setting up platforms that deserve the term “marketplace” for data. We identify several categories and dimensions of data marketplaces and data vendors and provide a survey of the current situation.

ISSN 1614-7448

cite as: Florian Stahl, Fabian Schomm, Gottfried Vossen: Marketplaces for Data: An initial Survey. In: Working Papers, European Research Center for Information Systems No. 14. Eds.: Becker, J. et al. Münster. 07 2012.

Table of Contents

List of Figures

Figure 1: Number of vendors for each category. 11

Figure 2: Number of vendors for Time Frame. 11

Figure 3: Number of vendors for each domain. 12

Figure 4: Data Origin Distribution..... 13

Figure 5: Trustworthiness Distribution. 14

Figure 6: Number of vendors for each pricing model..... 15

Figure 7: Data Access Distribution..... 16

Figure 8: Number of Vendors per Data Output Category. 17

Figure 9: Language of Web sites and Data. 17

Figure 10: Number of vendors by size. 18

Figure 11: Maturity of Vendors..... 19

List of Tables

1 Introduction

Today information is one of the most crucial driving factors for most business. Only if high quality information is available, correct decisions (i.e., decisions in the interest of company revenues) can be made on a rational and well-founded basis. In accordance with that, it can be observed that evermore suppliers of data, which build the basis of information, emerge. Given the high transaction costs a buyer faces when looking individually at the emerging heterogeneous market for data, we also observe the arrival of *data marketplaces*. In 1998 this term was used by Armstrong and Durfee in [AD98], who modeled trading of information between digital libraries, focusing on the motivation and behavior of participants and identifying factors that affect cooperation in a network. Since then, numerous other marketplaces for data have emerged; this paper identifies several categories and dimensions of data marketplaces as well as vendors and surveys the current state of affairs in this field.

Information intermediaries are as old as the Web itself. Indeed, shortly after the arrival of the Web in the early 1990s a new category of professionals emerged to which a search task could be given, and who would then search the Web correspondingly (for a fee) and return the results found. Thanks to advances in technology, but also to the vast amount of data nowadays available, a modern information marketplace or information intermediary can provide added value in numerous ways, even though one could argue they often re-publish data that is already available on the Web. First, data may be hard to find and scattered across different websites. A data vendor that aggregates these single datasets into a bigger and more refined one performs a service that makes it easier for customers or end-users to find relevant data. A second reason is that datasets from different providers often have different access mechanisms and formats. Therefore, offering one single mechanism to access data in a consistent format can save time and money for customers.

While there has been research on particular data marketplaces such as MS Azure[1] [Mic11] and others (e. g., [MD12]), there is – to our knowledge – to date no comprehensive survey and comparison of multiple data markets and data vendors. Therefore, we have conducted a survey of a total of 46 suppliers of data and data marketplaces of various kinds. The study was conducted from April to July 2012 and aims at providing a taxonomy for classifying data vendors, thus enabling us to derive conclusions regarding what types of vendors currently exist as well as which gaps or even niches might need to be filled. Furthermore, we can give possible reasons for which implications our findings might have and hints for future research.

The remainder of this paper is organized as follows: First, the approach to the survey will be described in Section 2. Then we present our findings, i.e., groupings, categorizations, as well as correlations we have found in Section 3. In Section 4 we describe use cases for all categories to emphasis the relevance of our findings. Section 5 gives an overview of related work that has been conducted in this area. The paper is concluded by summing up our findings in Section 6.

[1] https://datamarket.azure.com

2 Methodology and Approach

In this section, we first elaborate on what we consider to be a data market or data vendor. Then we explain how the survey was conducted, using an iterative approach for both collecting data suppliers and deriving categories (described in Section 2.2). Section 2.3 briefly elaborates on the statistical analysis and Section 2.4 discusses limitations of the method applied.

2.1 Data Marketplaces and Data Vendors

In the context of this work we have analyzed data vendors and data marketplaces. In order to restrict the potentially vast amount of companies, we have focused on companies offering either a platform for trading data (e.g., datamarket.com), raw data in any form (e.g., www.data.gov), or data enrichment tools (e.g., attensity.com). In order to gain a comparable set of data vendors, we have chosen to focus on vendors that offer online Web services. This implies that we have excluded offline products for data cleansing or data fusion and similar tasks.

More precisely, we define a *data marketplace* as platform on which anybody (or at least a great number of potentially registered clients) can upload and maintain data sets. Access to and use of the data is regulated through different licensing models.

A data vendor has data and offers it to others, either for a given fee or free of charge. However, it is not important how vendors obtain this data and many ways are common, e.g., aggregation from freely available sources, generation using proprietary methods or buying from other vendors. It is important to note that a data vendor can offer its data either on its own or through a data marketplace as described above. Vice versa, it is also possible that a data marketplace operator also sells data and thus takes on the role of a vendor.

2.2 Sampling and Basic Analysis

The initial set of vendors consisted of well-known suppliers we found in adjacent research [MSLV12]. From this starting point, keywords were derived that were then used for a broader online search which revealed a more comprehensive set of different products and services. The next step consisted of analyzing the vendors and categorizing them along different dimensions according to the following iterative approach:

1. Draft an initial set of dimensions and categories thereof (columns).
2. Categorize vendors along those dimensions (rows).
3. Expand dimensions (find more categories) or develop new dimensions if necessary.
4. Return to Step 2 and complete previously incomplete rows.

Table 1 shows the final set of dimensions. Initially, we started with dimensions 1 to 7; however, when analyzing the vendors, it became clear that more dimensions were necessary to fully capture the broad range of different vendors in the market. Following our iterative approach, the original list of dimensions was expanded to also include dimension 8 to 12.

Table 1: **Set of dimensions**

#	Dimension	Categories	Question to be answered
1	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?
2	Time Frame	Static/Factual, Up To Date	Is the data static or real-time?
3	Domain	All, Finance/Economy, Bio Medicine, Social Media, Geo Data, Address Data	What is the data about?
4	Trustworthiness	Low, Medium, High	How trustworthy is the source of the data? Rather subjective
5	Pricing model	Free, Freemium, Pay-Per-Use, Flat Rate	Is the offer free, pay-per-use or usable with a flat rate?
6	Data access	API, Download, Specialized Software, Web Interface	What technical means are offered to access the data?
7	Language	English, German, More	What is the language of the website? Does it differ from the language of the data?
8	Data Origin	Internet, Self-Generated, User, Community, Government, Authority	Where does the data come from? Who is the author?
9	Output Format	XML, CSV/XLS, JSON, RDF, Report	In what way is the data formatted for the user?
10	Size of Vendor	Startup, Medium, Big, Global Player	How big is the vendor?
11	Maturity	Research Project, Beta, Medium, High	Is the product still in beta or already established?
12	Target Audience	Business, Customer	Towards whom is the product geared?

In this approach, values are strictly binary. An offering either fulfills the criteria for a certain category or it does not. This is inherent to the data. Therefore, we have chosen not to increase

the granularity for two reasons: First, there is no scientific approach to derive a reliable figure and second the additional insight gained would be negligible in most cases. Furthermore, categories are not mutually exclusive in most cases. This means that one offering can fall into multiple categories, have multiple pricing models, or provide multiple ways for data access. Some dimensions (e.g., maturity), however, are mutually exclusive. Where this is the case, it will be stated explicitly in the dimension description in Section 3.

The facts about the data vendors were gathered by means of a Web search. As every vendor has a website, this publicly available information was used to determine how to categorize each vendor. After having done that with the initial set of vendors, it was checked how many entries a category had to justify its existence. When a category had only few entries, a new Web search for more data suppliers falling into that category was started to make sure no important vendors were omitted. If more companies were found, the list was extended and the new companies were analyzed regarding the other dimensions. This shows how finding companies also was an iterative process. However, if no more companies were found, the category definitions were reconsidered and updated.

2.3 Correlation

In order to find co-occurring categories, we have applied basic association rule mining techniques as described in [ZZ02, HKP11]. Here, it was only investigated whether two categories are correlated, and we have restricted the analysis to a support of 0.3 and a confidence of 0.6. We have excluded *time frame* and *target audience*, both dimensions with only two categories, and the language dimension as they were dominated by one language as these would have been likely to correlate with many other dimensions. Furthermore, even though we set the minimal support to 0.3 we did not find many association rules, and in fact some of them were trivial or not of interest (e.g., 82% of highly mature companies offer CSV / XLS files). Nevertheless, the rules we considered relevant and meaningful are pointed out in the according findings section.

2.4 Limitations

The information we used was taken directly from the website of each vendor. This may limit the accuracy of our findings in some cases, where the description of a product exceeds the actual functionality. Due to resource and time restrictions, it was not possible to verify whether or not every product lives up to its specifications. Random samples, however, indicate that the descriptions match the services provided.

Nevertheless, there are also cases where the information provided on a vendor’s website was not sufficient to categorize all dimensions. This was particularly the case for B2B vendors, which only reveal their pricing models upon request. We chose rather to leave these dimensions out than to speculate about their value. As a result, however, the numbers of these dimensions are minimally skewed.

The market of data vendors and data market places is highly active, i.e., new actors emerge and others disappear, and the market as such is growing rapidly. Therefore, it cannot be guaranteed that this study is fully exhaustive with regard to the number of vendors in the

market. That said, we are confident that – during our observations from April to July 2012 – we have obtained a representative sample that allows for meaningful analysis.

Furthermore, it has to be stated that data trading channels are not necessarily made public. This means that we are aware of the fact that a certain amount of data is traded directly between (large) corporations or within an ecosystem (such as social networks) without the use of intermediaries. It is obvious that it is impossible to investigate those forms of data trading using our Web survey approach.

3 Findings

As stated in the previous section, the following twelve dimensions have been examined: Type, Time Frame, Domain, Data Origin, Trustworthiness, Pricing Model, Data Access, Data Output, Language, Size of Vendor, Maturity, and Target Audience. These are described in more detail next.

3.1 Type

The first dimension *Type* is used to classify vendors based on what their core product is. In order to form a common understanding of the different categories they are explained below:

- *(Focused) Web Crawler:* Services that are specifically designed to crawl a particular Web site or set of Web sites. These are always bound to one domain.
- *Customizable Crawler:* General purpose crawlers that can be set up by the customer to crawl any website and search for arbitrary content.
- *Search Engine:* Services that offer their content via an interface similar to a search engine. Customers specify keywords as input and the search engine produces output relevant to the input.
- *Raw Data Vendor:* This category comprises vendors that offer raw data, most often in forms of tables or lists.
- *Complex Data Vendor*: These vendors offer data that is the result of some kind of analysis process, for example finance service providers that calculate stock indicators and sell access to this data.
- *Matching Data Vendor*: Vendors that offer the matching of input data against some other database. These vendors most often operate in domains where a customer does not want a complete dataset, but rather needs the data they already have corrected or verified, e. g., address data.
- *Enrichment – Tagging:* This category describes services that enrich an input (mostly texts, but others are also possible) through means of tags. This enables customers to make more use of their data.
- *Enrichment – Sentiment:* With the proliferation of social media websites on the internet, a multitude of vendors emerged that specialized on what is commonly referred to as sentiment analysis [PL08]. The core service is a collection of data from social media and the analysis with regard to certain factors.
- *Enrichment – Analysis:* The data offered is enriched with analysis results obtained through various means, i.e., comparisons with historical data or forecasts.
- *Data Market Place:* These services allow customer both, to buy and sell data by providing the infrastructure needed for such transactions.

Figure 1 shows how many vendors fall into which category. It has to be kept in mind, though, that these categories are not mutually exclusive and one vendor can fulfill the criteria of multiple categories.

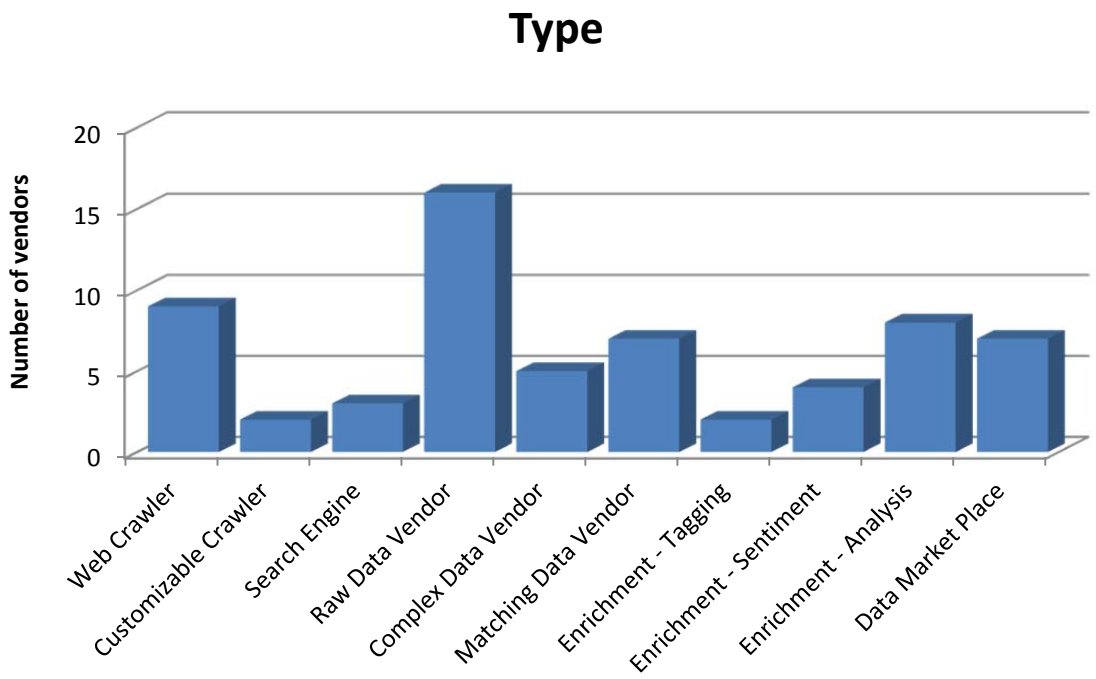

Figure 1: Number of vendors for each category.

Some facts in **Figure 1** are notable. First of all, the number of vendors that offer a customizable crawler is relatively low. The reason for this could be that such an offering serves a relatively small niche market. Customers who want data from a crawler often have a precise understanding of what they want to have crawled, and a specialized offer that is usable out-of-the-box is much easier to put into action.

The next interesting fact is that the enrichment – tagging category is also rather low. The reason for this lies in the methodology used to obtain the list of vendors. As explained in Section 2, we have intentionally excluded offline tools. However, most traditional software solutions that offer tagging functionalities are built upon an offline infrastructure and are therefore not included in this survey.

3.2 Time Frame

The time frame dimension captures the temporal context of the data. We distinguish two categories in this dimension:

- *Static/Factual*: Data is valid and relevant for a long period of time and does not change abruptly, i.e., population numbers, geographical coordinates, etc.
- *Up-To-Date*: Data is important shortly after its creation and loses its relevance quickly, i.e., current stock prices, weather data, or social media entries.

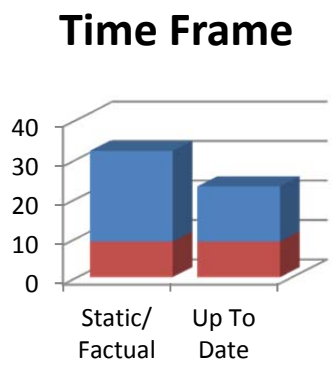

Figure 2: Number of vendors for Time Frame.

As evident in **Figure 2** we found that static data (32 offerings) was offered more often than up-to-date data (23 offerings). Even though both categories are not mutually exclusive, we found that only less than

20% (9 offerings) of the vendors examined offer both static and up to date information. This suggests that generally data vendors specialize in either of the two options.

3.3 Domain

Dimension domain describes what the actual data is about. While most domain names are self-explanatory, domain *any* deserves clarification. This domain was used to classify vendors whose offers were not restricted and could incorporate arbitrary domains. Whilst other domains were not mutual exclusive (i. e., a vendor could supply more than one domain), vendors serving any domain did not count towards explicit domains. The results are shown in **Figure 3**.

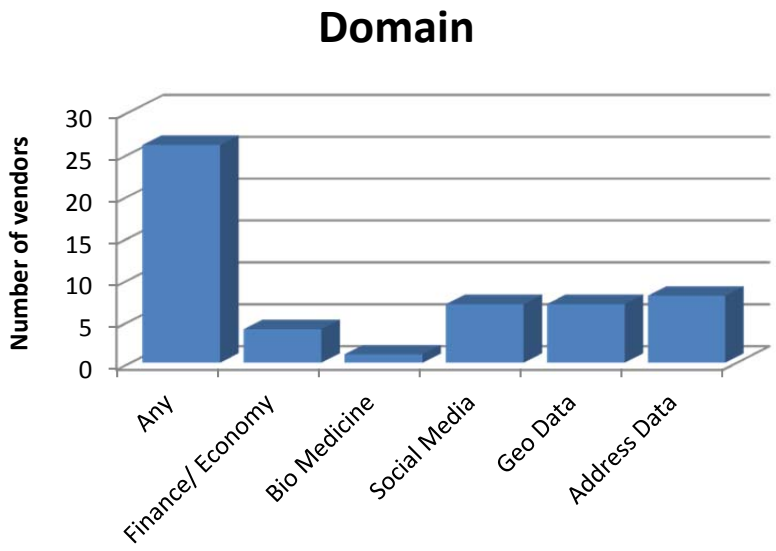

Figure 3: Number of vendors for each domain.

It is obvious that the *any* domain is by far the biggest group. An explanation for this is that data market places, search engines, and customizable crawlers do indeed serve any domain, depending on what customers choose to upload or search for. Given that they account for more than a fourth of all companies under investigation, the peak in *any* is not very surprising. The other domains have a lower number of vendors, because they are more specialized. Furthermore, we have observed that the geo data (7) and address data (8) domains have a significant overlap (6), which can be explained by their obviously close relationship.

3.4 Data Origin

The origin of data describes where it comes from. We have identified six different categories in this dimension:

- *Internet*: The data is pulled directly from a publicly and freely available online resource.
- *Self-Generated*: Vendors have means of generating data on their own, i.e. manual curation of a specific dataset or calculating forecasts based on patented methods.
- *User*: Users have to provide an input before they can obtain any data, i.e. address data offerings that return the address for a given name.

- *Community*: Based on a wiki-like principle, these vendors obtain and maintain their data in a very open fashion. The restrictions as to who can participate are usually rather low.
- *Government*: Governments capture and process huge amounts of data and have recently begun to make this data publicly available.
- *Authority*: Authorities in a domain are entities which are the main provider of data, i.e. the stock market for stock prices or the postal offices for address data.

In our survey the most popular category was *Internet*. Almost 50% of all vendors receive their data from an online source. Another category with a large number of vendors was *Authority:* 32% obtain their data from authoritative sources. The main advantage of these offers is that the data is usually of high correctness, completeness, and credibility. This also holds for the *Government* category, into which fell 15% of vendors. Categories *Self-generated* and *Community* are matched by 15% and 19%, respectively. Lastly, category *user* with 15% is a special case because it cannot stand on its own, i.e., every vendor classified into this category also gathered data from another source. This is inherent to the definition of this category, according to which users submit their data and receive back their data with additional annotations for which a vendor needs additional data sources. These facts are illustrated in **Figure 4**.

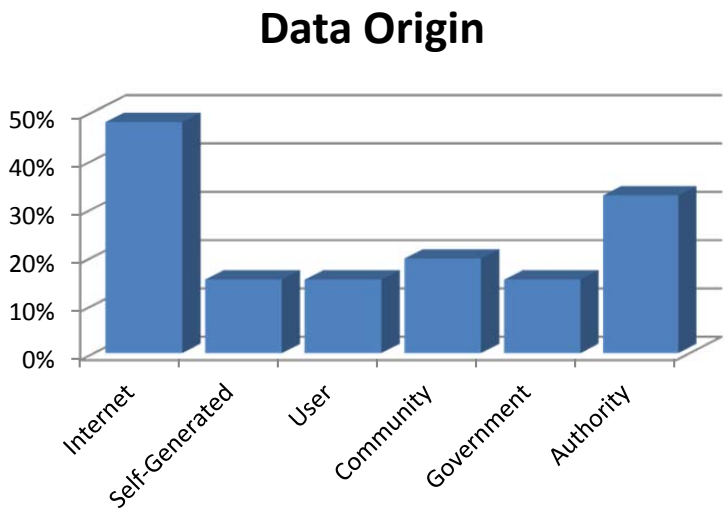

Figure 4: Data Origin Distribution.

3.5 Trustworthiness

This dimension indicates how trustworthy a vendor is. Whether a vendor has been categorized as *low*, *medium*, or *high* is depended on the origin of the data as well as on how it is processed. Data that comes from a community generally has a lower trustworthiness than data that is sourced from an authority. Nevertheless, this dimension is not quantifiable and, thus, the results here could be slightly subjectively biased.

As depicted in **Figure 5**, we have found that 54% of all vendors have a high trustworthiness. Among these are those vendors that carefully select the data they offer in a transparent and comprehensible way. Also, authorities and governments as explained in Section 3.4 all exhibit a

high trustworthiness. Category *medium* is populated by around 33% of all examined vendors. The main indicator for their classification that they *seem* to be trustworthy based on the descriptions, but this could not be verified in any way, e.g., because they do not explicitly state their data sources or explain their analytical methods. The lowest amount of trustworthiness applied to only 22% of all vendors. Typical vendors in this category are those that do not even claim to deliver correct or complete data, like web crawlers or community-supplied websites.

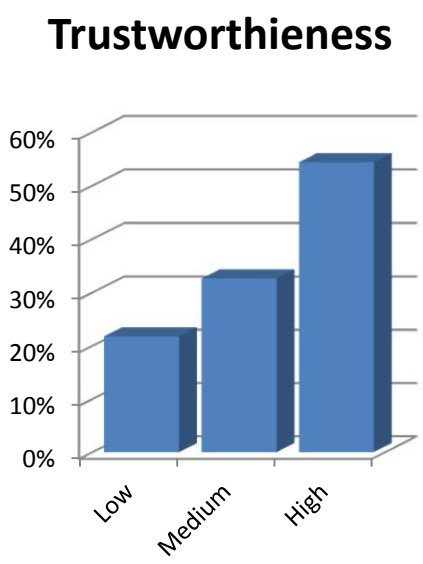

Figure 5: Trustworthiness Distribution.

Notice that the overlap between the three categories stems from the fact that one vendor can offer multiple datasets from different source. In such a case, we have assigned all possible levels of trustworthiness. Furthermore, while it is intuitive that high trustworthiness is good, it is not the case that low trustworthiness is bad by default. There are scenarios in which incomplete data is sufficient for a rough estimation (cf. Section 4), or data with a high trustworthiness is not available (e.g., social media analysis). This leads us to the conclusion that vendors with all levels of trustworthiness are likely to co-exist in the future, because they fulfill different demands.

In regard to a high trustworthiness we could find a correlation to the data source authority. In fact, all data markets that were ranked as highly trustworthy had an authoritative source and 60 per cent of those using authoritative sources were highly trustworthy.

3.6 Pricing Model

Pricing models are very important to understanding how exactly the different vendors set up their business models. Four main pricing models could be identified; the number of vendors for each model is illustrated in **Figure 6**. A verbal explanation of the pricing models is provided by the following list:

- *Free:* These services can be used at no charge. Reasons for offering such a service for free are, among others, that it is only a beta test or research project, the vendor is a public authority funded by taxes, or simply interested in attracting more customers. Vendors in this category do not count towards one of the following categories.
- *Freemium:* As a portmanteau combining *free* and *premium*, this pricing model offers a limited access at no cost with the possibility of an update to a fee-based premium access. Freemium models are always combined with at least one of the following two payment models.
- *Pay-Per-Use:* Customers are billed based on how much they use the service. This manifests mostly in the form of x$ per thousand API calls.
- *Flat Rate:* After paying a fixed amount of money, customers can make unlimited use of the service for a limited time span, mostly a month or a year.

Figure 6: Number of vendors for each pricing model.

3.7 Data Access

Dimension *Data Access* describes through which means end-users receive their data from vendors. The main categories identified and presented in **Figure 7** are:

- *API*: An API (application programming interface) is used to provide a language- and platform-independent programmatic access to data over the Internet.
- *Download*: Traditional download of files is the easiest way to access a set of data, because anyone can use such a service with only a web browser.
- *Specialized Software*: Some vendors have implemented a specialized software client to connect with their web service. While this approach does have downsides (implementation and maintenance expense, dependency issues, etc.), there are some scenarios in which the concept is worthwhile. For example, providing the customer with an easy-to-use graphical user interface as an out-of-the-box solution that needs no further customization, or granting access to real-time streams of data.
- *Web Interface*: In a Web interface, the data is displayed to the customer directly on a website.

The flexibility and modularity of APIs have made these the most popular of all access methods. More than 70% of all vendors offer an API. However, less than 30% of all vendors have an API as their only way to access data. Most vendors offer an API next to other methods. For example, Web interfaces or file downloads are used to give previews of the dataset, to make it easier and more accessible for the customer to see what the actual data looks like. The concept of specialized software does not seem to stand very well on its own. Out of all investigated vendors, only three use specialized software as the only way of data access. The reason for this might be, that this approach lacks flexibility, because customers are restricted in the way they can use the data by the functionality of the provided software. However, most customers that want data also want the possibility to process the data in any imaginable way.

From a theoretical point of view, it seems to be the best approach for a vendor to offer all the aforementioned means of access to his data, because that allows customers to choose their preferred way of access. However, we have not found a single vendor that does so, which is probably due to the high cost associated with creating such a broad offering.

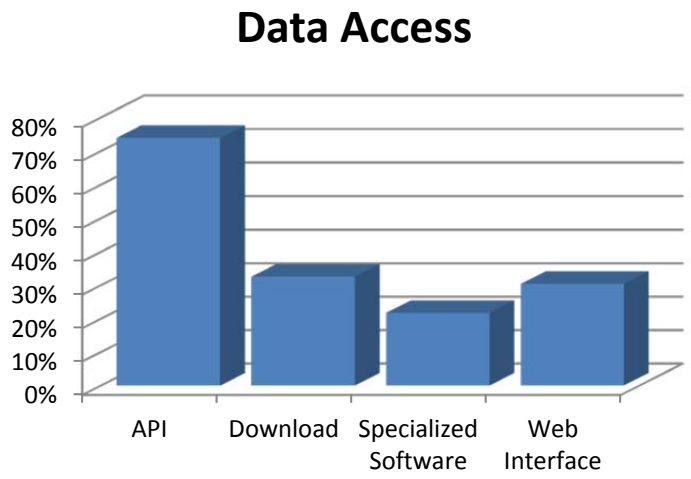

Figure 7: Data Access Distribution.

Furthermore, co-occurrence could be identified between *API* and the data domain *all.* Concretely 65% of vendors serving any domain offered an API and 86% of vendors offering an API offered data of any domain. Also a high connection to the data output formats *XML (94%), CSV / XLS (72%),* and *JSON* (100%) could be found. Per cent values show how many vendors offering an API also offered the given formats. The result, however, is little surprising, as in particular XML and JSON are data exchange formats.

3.8 Data Output

This dimension shows the format in which data can be obtained. To us, the most reasonable set of categories in this dimension is the following:

- *XML*: Being both human- and machine-readable, the Extensible Markup Language is a widely established standard for data transfer and representation.
- *CSV/XLS*: Most structured data is laid out in a tabular way, so it makes sense to wrap it into a table file format. We do not distinguish between CSV and XLS and other table file formats, because the main differences between them, like formatting and embedding, do not apply when you are showing raw data
- *JSON*: The JavaScript Object Notation is similar to XML and is also used as a data transfer format. Data is represented as text in key-values pairs.
- *RDF*: The Resource Description Framework is a method to describe and model information. It uses subject-predicate-object triplets to make statements about resources.
- *Report*: When data is preprocessed, aggregated and prettified in some way, we declared the output as a report. The main difference in this category is that the customer does not have insight into the underlying raw data. Also visual reports in the form of MS Excel spreadsheet classified for this category.

The most popular category in the output dimension shown in **Figure 8** is CSV/XLS. With 22 vendors, almost half of all vendors offer the possibility to receive their data as a raw table. However, only six of those vendors have CSV/XLS as their only output format. Most vendors also offer either an XML (10) or a JSON (6) interface, some even both (3). This is in concordance with the observation from the previous dimension, that API is the most popular way of data access. An API usually produces XML or JSON output.

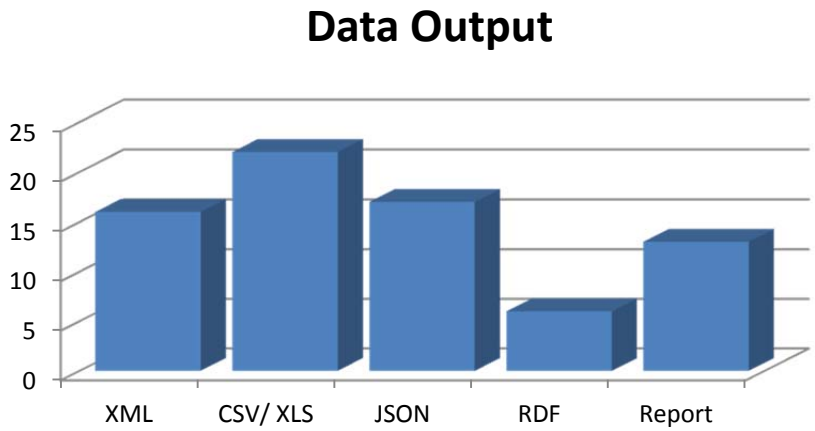

Figure 8: Number of Vendors per Data Output Category.

3.9 Language

The languages we have focused on were English and German, distinguishing between the language of the website and the language of the data offered. Further languages were aggregated into a third category called *more*. A visual representation of the results is shown in **Figure 9**.

Nearly all investigated vendors (98%) run an English-language website. For the majority, English is also the only language available (89%). Only some globally operating companies run a multilingual website (9% German; 7% More). This picture changes when looking at the language of the data itself. We observed that again 98% offered English Language Data, but about 30% offered German data and almost 20% of the vendors also offered data in further languages.

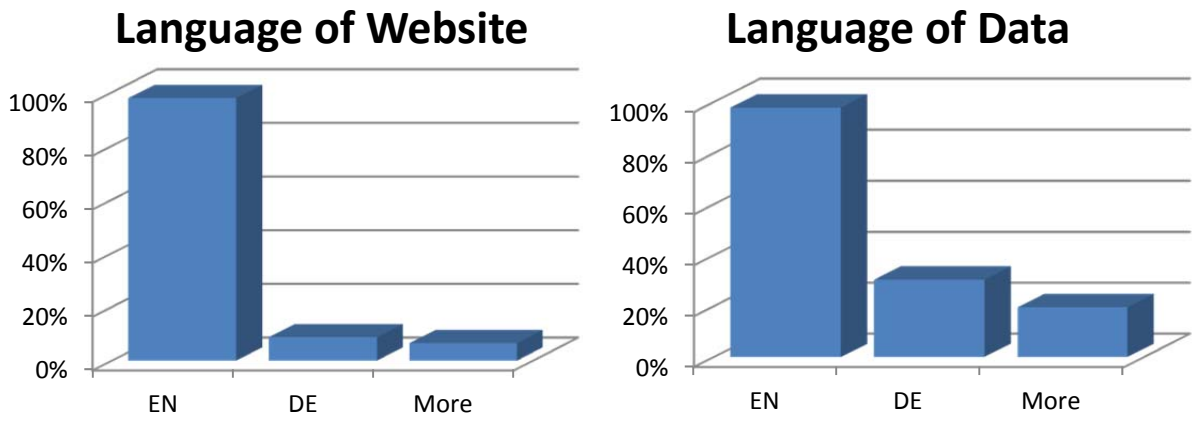

Figure 9: Language of Web sites and Data.

We have seen that English is the dominant language for both websites and data. This is not surprising because the market for data has a global scope and English seems to be the best suited language for that. However, there is also a demand for local data in the corresponding language, which is suggested by the amount of vendors that offer such data. Also, this search does not include the Asian market, so it might be the case that similar offerings exist of which we are unaware.

3.10 Size of Vendor

For the size of a vendor we have created four categories:

- Startup: Companies that are newly created and that have only a small number of people involved are usually referred to as *startups*; examples include Uberblic or QuantBench.. These are often funded by investors, as they do not have a positive cash flow from the very beginning.
- Medium: Leaving the beta stage, gaining experience and maturity, and not being dependent on investors anymore are the key characteristics that set medium-sized companies apart from startups. Examples include eXelate or Spinn3r.
- Big: Companies that are well-established and have more than one product in their offering range are considered big, e.g., Infochimps or Lexis Nexis. While there is no sharp dividing line between medium-sized and big companies, we still felt that separating the two in different groups yields more accuracy for the analysis.
- Global Player: In this category fall only the biggest companies out there, like Yahoo!, Microsoft, IBM, etc.

For this dimension we have made the categories mutually exclusive, although it is obviously the case that a medium-size company might be a startup, or a big company might be a global player (and vice versa in either case). **Figure 10** shows the number of vendors for each size. It can be seen that the number of startups is the lowest. This could indicate that the market for data is not easy to enter. The number of global players also seems rather low, but one has to keep in mind that these vendors have the potential to quickly seize huge market shares, because they usually have experienced people and high capital. The majority of vendors is either medium-sized or big.

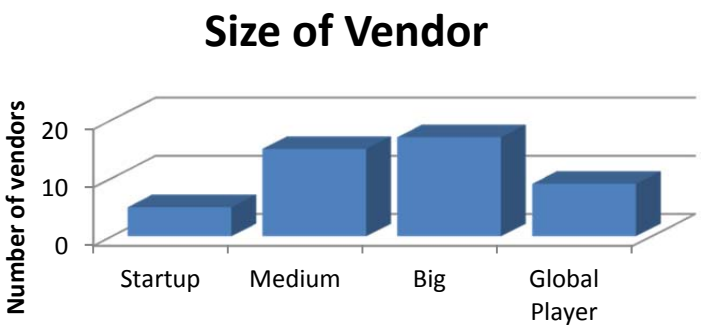

Figure 10: Number of vendors by size.

3.11 Maturity

The maturity of all offerings has been classified into the following four categories, which are mutually exclusive:

- Research Project: These offerings are usually not for profit and can therefore be used free of charge. They are mainly executed as a proof-of-concept.
- Beta: A beta product is still in development and has not been fully launched yet. Nevertheless, we have also seen offerings in beta phase that already demanded a usage-fee.
- Medium: This category classified products that were already out of beta, but were still not as highly developed as other products.
- High: Full-fledged products that are generating a considerable amount of revenue fall into this category.

Evaluating the numbers presented in **Figure 11** has shown that only 3 research projects, 6 betas and 6 medium-matured offerings could be identified. The remaining 31 offerings can all be classified as having a high maturity. This observation can also serve as an explanation to the previous finding of a low number of startups. When there are already established vendors with mature projects, the space for new companies to enter the market is relatively small.

In regard to maturity, again association rules could be found. For instance, 62% of those companies with a high maturity serve all domains of data; 68% offer an API, 76% have a high trustworthiness, and 100% of them are big companies. In most cases the co-occurrence in the other direction is less strong, nevertheless the factors identified are all properties one would expect from a mature company.

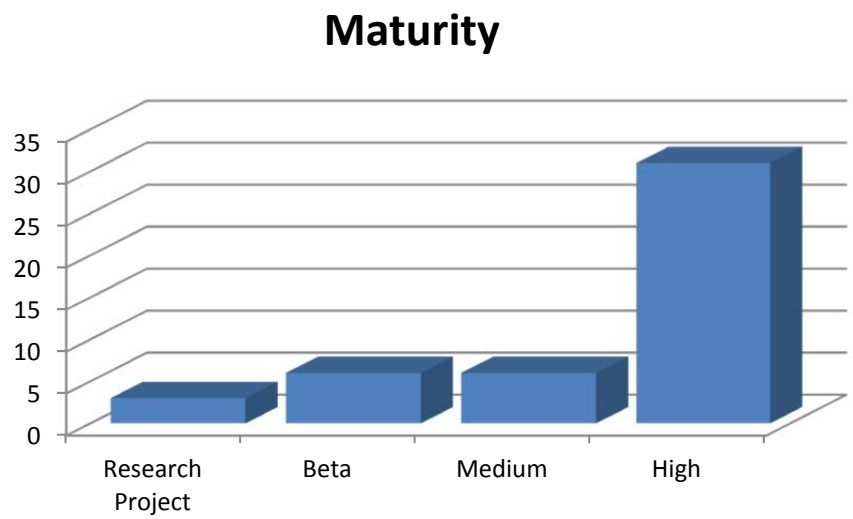

Figure 11: Maturity of Vendors.

3.12 Target Audience

The last dimension is concerned with the target audience. Here, we have investigated towards whom the offering is tailored. As is evident from **Figure 12**, there are only two categories in this dimension, *business* and *customer*. Providing data for another company in a B2B fashion is the most logical application area of data vending. Out of all vendors in this research, 87% offered data in a business context, 41% sold data relevant for end consumers, and 28% had data that could be of use for both groups.

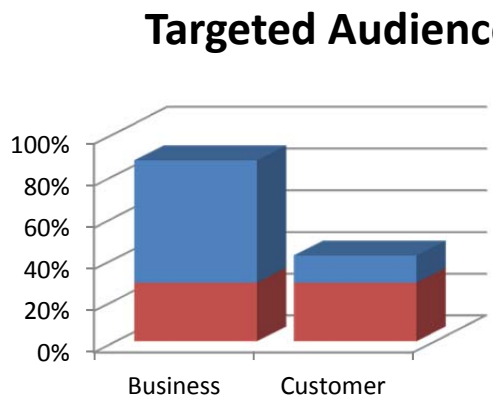

Figure 12: Number of Vendors by Targeted Audience.

4 Use Cases

Having shown the different dimensions and categories covered by various vendors, this section will give further insight into *how* the offered data could be put into use. To this end, we will describe four sample use cases and illustrate core business ideas.

4.1 Start-ups

Consider a startup company that has a great idea for a new kind of application. However, in order to realize the idea, they need access to relevant data. For example, a restaurant recommendation app needs data about the locations and offerings of all restaurants in a given area. In order to obtain such a data set, one can think of various ways. Manually aggregating and curating it would be a tedious task and possibly not worth the effort. Buying the data from a premium data vendor is another option, but given the budget constraints startups usually face, this may as well not be feasible.

The best option would be to have access to relevant data for free. Somebody somewhere probably has already collected parts of the relevant data, and all that is needed is a platform where these different parts are compiled and made accessible through a standardized interface. Among the data providers that we have analyzed in this survey, the following offer free data sets that are easily accessible and ready to use: *Factual, AggData, Windows Azure Marketplace, Infochimps Data Marketplace, DataMarket, Uberblic, CloudMade Data Market Place, Semantics3, Project Nimbus, Kasabi, Freebase, Data.gov US, Data.gov.UK, Data.govt.NZ and The Data Hub*.

The benefits for a company to source data from such a provider are obvious: First, the data is free and thus no additional costs are incurred; secondly, the data can be obtained through a single API, eliminating the need to adapt to heterogeneous data sources. On the other hand, there are also disadvantages and limitations to this approach. Being free, the data may lack quality or completeness. While this might not be so severe for the restaurant example, it could be a potential deal-breaker for other use cases where data quality is mission-critical. A further issue with free data sets is their limited availability. Obviously, not all relevant datasets are available, and even if they are, they may not always free.

4.2 Public institutions

Public institutions generate plenty of data that is potentially interesting to a broader audience, which is not yet made available. Such data could, for example, be about aid programs by the Red Cross, or about government spending on different projects. Since probably nobody is willing to pay for these data they might be hard to sell. Nevertheless, the data could be of use to stakeholders, e. g., for charitable purposes or to increase transparency. With no commercial interest involved, it is difficult to derive a business model that generates enough revenue to pay for the required technical infrastructure and involved labor that is needed to publish the data and make it accessible.

In order to facilitate the publication of free data sets, a platform is needed that allows uploading and maintaining the data free of charge. Such platforms are increasingly called *data*

marketplaces, and some companies already operate them productively, e.g. *Windows Azure Marketplace, Infochimps Data Marketplace, DataMarket, Kasabi and The Data Hub*. These services allow users to publish their data without having to set up their own servers. For every set of data, licenses can be defined that explicitly regulate in which way the data may be used, including free to use licenses.

4.3 Corporations

Today, many established corporations run their own address databases, many times even embedded into a customer relationship management (CRM) solution. Often, the address data contained within such a CRM system has been gathered over years from varying, heterogeneous sources (e. g., via mergers and acquisitions). The result is that the data quality decreases over time, as the data becomes more and more inconsistent. Furthermore, some data entries may expire, resulting in erroneous records that lead to increased cost, for instance when mail cannot be delivered correctly and is hence returned.

In order to cope with this problem, data quality policies need to be enforced. In the case of address data, this is a complicated task, because the data needs to be synchronized with real-world events such as relocations. There are vendors who offer address cleansing as an online Web service, e. g., *AddressDoctor, DQ Global, Experian QAS, InfiniteGravity, Intelligent Search, PitneeyBows and CustomLists*. These vendors maintain their own address database and allow customers to match their own data against it. Additional tasks like duplication detection are often offered as well. Especially organizations that process lots of address data (i.e. retailers) can benefit immensely from using such services.

4.4 Brand Monitoring

The recent rise of social media has made the Internet a valuable source of information for companies. People share their experiences with the products they buy and give recommendations to others. This data is freely available and very valuable to companies that have established brands and sophisticated marketing strategies. However, it is difficult to extract this information, because it is stored in different websites and formats, e. g., Facebook, Twitter or blogs. Furthermore, the vast amount of data available has to be carefully filtered in order to find relevant pieces of information.

Based on this issue, a number of vendors have emerged that are specialized in the field of social media monitoring, such as *MeaningMine, Sysomos, Radian6, Attensity Analyze, VICO Research, Gnip and Lexis Nexis*. The functionality of these services ranges from simple word counting to advanced text mining algorithms and sentiment analyses. This allows a client to gather customer feedback regarding specific products, or assess the success of advertising campaigns, allowing better customer-oriented marketing decisions. However, special care has to be taken when interpreting the results of such automated analyses. Natural language processing is not (yet) able to fully capture the meaning of all linguistic constructs of most natural languages.

Nevertheless, the data that can be obtained through brand monitoring on social media channels – though not perfect – can still be very valuable. For example, an established company that tries

to extend its portfolio with novel products can gain insight into the reaction of targeted customers. This information could also be used to augment traditional approaches of feedback collection, such as questionnaires or surveys. Also, trend identification and prediction methods could be applied, where companies try to find out in what general direction customer preferences head.

5 Related Work

Ge et al. [GRC05] also studied electronic marketplaces but restricted themselves to Web sites such as Askjeeves.com where users can ask questions, which are then answered by other users or experts. Furthermore, they only described five Web sites and focused rather on business models than on surveying marketplace properties. Regarding data markets as we defined in them Section 2.1 – to our knowledge – no similar work has been done thus far.

However, there have been investigations into particular market places. For instance on Kasabi, described as a “web-based information marketplace” [MD12]. On Kasabi data is stored using the Resource Description Framework (RDF) with the goal of bridging the gap between data publishers and application developers by providing a platform that allows hosting of and searching for data. It is designed after the *linked data* paradigm originally outlined by Tim Berners-Lee. The basic idea of *linked data* is to publish data in a structured way that allows for linkage to data sets. An overview of this concept, the technical principles and its applications can be found in [BHB09]. A survey about the current usage of these dataset is given by [MHC10] and actual trends are outlined in [Biz09].

In the course of the *Linked Open Data* (LOD) movement, FactForge emerged as a publicly available service that is meant to “provide an easy point of entry for would-be consumers of Linked Data” [BKO+11]. It was built with the intention to facilitate access to the LOD cloud of data by integrating the major datasets into one view. These datasets include DBPedia, Freebase, Geonames and five more.

A different approach is pursued by the authors of Freebase. They try to create what they call a “collaboratively created graph database for structuring human knowledge” [BEP+08]. The collaboration aspect is inspired by Wikipedia and based on the idea that data quality improves when lots of people refine datasets. They employ a graph database, because it depends less on a rigid schema and is more flexible. The authors even state explicitly that they want to allow conflicting and contradictory types and properties to exist simultaneously in order to “reflect users’ differing opinions and understanding” [BEP+08].

Microsoft’s contribution to the market is called Windows Azure Marketplace [Mic11] and has been launched in 2010. It is designed to make the sharing of data as well as applications an easy process for both consumers and providers of data. The key features are: global reach through a central platform, unified billing and access mechanism, high data quality, and easy integration with other Microsoft products. Unique to Windows Azure Marketplace is the combination of datasets and applications. This allows providers of data not only to sell their raw data, but also bundle it with applications which are designed specifically for that data. Customers can purchase these bundles directly and have a working out-of-the-box solution with no additional implementation effort.

6 Conclusion

In this study we have presented an initial overview of data vendors and marketplaces for data. Utilizing an iterative approach we have derived dimensions along which such data providers can be classified and grouped. We have then presented a survey drawing a preliminary picture of the current data vendor landscape. The practical relevance of our findings has been underlined by an outline of several realistic use cases.

Our survey gives an overview of the current market situation and shows which categories are currently underrepresented and which ones can be particularly interesting for practitioners. However, our findings are also relevant to academics, who can get a feeling for where the market is heading and where potentially more research is needed. In our view, customizable crawling and enrichment are areas that offer value for businesses and consumers and should thus be fostered.

Indeed, a major research question that is currently under investigation by various people is that of appropriate pricing. When data is to become a form of tradable goods, asking the “right” price in the right context is of crucial importance. This is comparable to other commodities such as electricity or gasoline. Yet besides economic aspects, formal questions related to data (and query) pricing are investigated by [BHS11, KUB+12] or by [LM12]. Beyond that, legal, organizational, social, and technical issues deserve considerable further studies.

References

[AD98] Armstrong, A.A.; Durfee, E.H.: **Mixing and memory: emergent cooperation in an information marketplace***, Proceeding of the Third International Conference on Multi Agent Systems*, *1998,* pp.34-41

[BHS11] Balazinska, M.; Howe, B.; Suciu, D.: **Data markets in the cloud: An opportunity for the database community**. PVLDB, 4(12):1482{1485, 2011.

[BEP+08] Bollacker, K.; Evans, C; Paritosh, P; Sturge, T; Taylor, J: **Freebase: a collaboratively created graph database for structuring human knowledge**. In *Proceedings of the ACM SIGMOD international conference on Management of data*, 2008 pp. 1247-1250.

[BHB09] Bizer, C.; Heath, T.; Berners-Lee, T.: **Linked Data - The Story So Far.** *International Journal on Semantic Web and Information Systems (IJSWIS),*2009, vol. 5,no 3, pp. 1-22.

[Biz09] Bizer, C.: **The Emerging Web of Linked Data**, *Intelligent Systems, IEEE* , vol.24, no.5, 2009, pp.87-92.

[BKO+11] Bishop, B.; Kiryakov, A.; Ognyanov, D.; Peikov, I.; Tashev, Z.; Velkov, R.: **FactForge: A fast track to the web of data**, In *Semantic Web*, vol. 2, no 2, 2001, pp.157-166.

[GRC05] Ge, W.; Rothenberger, M.; Chen, E.: A Model for an Electronic Information Marketplace, In Australasian Journal of Information Systems; vol. 13, .no 1, 2005

[HKP11] Han, J.; Kamber, M.; Pei, J.: **Data Mining: Concepts and Techniques**, 3rd ed., Morgan Kaufmann Publishers, 2011

[KUB+12] Koutris, P.; Upadhyaya, P.; Balazinska, M.; Howe, B.; Suciu, D.: Query-based data pricing. Proc. Annual ACM SIGMOD/PODS Conference 2012, pp. 167-178

[LM12] Li, Ch.; Miklau, G.: Pricing Aggregate Queries in a Data Marketplace. Proc. 15th International Workshop on the Web and Databases (WebDB) 2012

[MD12] Möller, K.; Dodds, L.: **The Kasabi Information Marketplace**, 21nd World Wide Web Conference, Lyon, France, 2012.

[Mic11] Microsoft: **Windows Azure Marketplace**, White Paper, http://go.microsoft.com/fwlink/?LinkID=201129&clcid=0x409, June 2011, last accessed: 2012/07/13.

[MHC10] Möller, K; Hausenblas, M; Cyganiak, R; Handschuh, S: **Learning from Linked Open Data Usage: Patterns & Metrics,** *Web Science Conference,* 2010.

[MSLV12] Muschalle, A.; Stahl, F.; Löser, A.; Vossen, G.: **Pricing Approaches for Data Markets**; to appear in 6th International Workshop on Business Intelligence for the Real Time Enterprise (BIRTE), Istanbul, Turkey, 2012.

[PL08] Pang, B.; Lee, L.: **Opinion Mining and Sentiment Analysis**. *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, 2008, pp. 1-135.

[ZZ02] Zhang, C.; Zhang S.: **Association Rule Mining – Models and Algorithms**, *Lecture Notes In Artificial Intelligence,* Springer, Berlin Heidelberg, 2002.

Working Papers, ERCIS

No. 1 Becker, J.; Backhaus, K.; Grob, H. L.; Hoeren, T.; Klein, S.; Kuchen, H.; Müller-Funk, U.; Thonemann, U. W.; Vossen, G.: European Research Center for Information Systems (ERCIS). Gründungsveranstaltung Münster, 12. Oktober 2004. Oktober 2004.

No. 2 Teubner, R. A.: The IT21 Checkup for IT Fitness: Experiences and Empirical Evidence from 4 Years of Evaluation Practice. March 2005.

No. 3 Teubner, R. A.; Mocker, M.: Strategic Information Planning – Insights from an Action Research Project in the Financial Services Industry. June 2005.

No. 4 Vossen, G.; Hagemann, S.: From Version 1.0 to Version 2.0: A Brief History Of the Web. January 2007.

No. 5 Hagemann, S.; Letz, C.; Vossen, G.: Web Service Discovery – Reality Check 2.0. July 2007.

No. 7 Ciechanowicz, P.; Poldner, M.; Kuchen, H.: The Münster Skeleton Library Muesli – A Comprehensive Overview. January 2009.

No. 8 Hagemann, S.; Vossen, G.: Web-Wide Application Customization: The Case of Mashups. April 2010.

No. 9 Majchrzak, T. A.; Jakubiec, A.; Lablans, M.; Ückert, F.: Evaluating Mobile Ambient Assisted Living Devices and Web 2.0 Technology for a Better Social Integration. January 2011.

No. 10 Majchrzak, T. A.; Kuchen, H.: Muggl: The Muenster Generator of Glass-box Test Cases. February 2011.

No. 11 Becker, J.; Beverungen, D.; Delfmann, P.; Räckers, M.: Network e-Volution. November 2011.

No. 12 Teubner R.; Pellengahr A.; Mocker M.: The IT Strategy Divide: Professional Practice and Academic Debate. February 2012.

No. 13 Niehaves B.; Köffer S.; Ortbach K.; Katschewitz S.: Towards an IT Consumerization Theory – A Theory and Practice Review. July 2012.

ERCIS – European Research Center for Information Systems
Westfälische Wilhelms-Universität Münster
Leonardo-Campus 3 ▪ 48149 Münster ▪ Germany
Tel: +49 (0)251 83-38100 ▪ Fax: +49 (0)251 83-38109
info@ercis.org ▪ http://www.ercis.org/

ISSN 1614-7448

Working Paper No. 14

Marketplaces for Data: An Initial Survey