Current Status and the Future Directions of Open Data: Perceptions from the Finnish Industry

Antti Herala
Lappeenranta University of Technology
Skinnarilankatu 34
53850 Lappeenranta, Finland
antti.herala@lut.fi

Jussi Kasurinen
Lappeenranta University of Technology
Skinnarilankatu 34
53850 Lappeenranta, Finland
jussi.kasurinen@lut.fi

Erno Vanhala
University of Tampere
Kalevantie 4
33100 Tampere, Finland
erno.vanhala@staff.uta.fi

ABSTRACT

Open data has been a hot topic of the decade, as even the president of the United States has given input to the deployment of open data practices. Besides politics, open data has been discussed in the scientific literature. So far the big breakthrough has not happened, although several cases have proved that the open data concept works, and provides positive results in several ways. In this article, we study and discuss the perceptions towards open data in business. Our data consisted of 45 survey responses gathered from various Finnish companies. The results indicate that companies want to use open data in different ways — ranging from application development to process efficiency — but at the moment open data in Finland is scattered and most of the companies do not find what they need. These findings are in line with the previous research executed on the field and argue that open data still needs more positive examples of application, but the potential is real, and companies consider it as a useful concept.

CCS Concepts

•**Applied computing** → **Information integration and interoperability;** *Enterprise data management;* IT governance;

Keywords

Open data, Economic value, Finnish industry

1. INTRODUCTION

Open data — data that can be used, reused, and shared by anyone — has been estimated to offer a boost to the economic and social value between three and five trillion dollars annually, as calculated by McKinsey Global in 2013 [15]. They estimated that the value would be unlocked across seven sectors, for example with 210-280 billion in finance and 300-450 billion in the healthcare business. However, unlike some other concepts to increase revenue, the data itself does not hold any value [2] since the data is shared without costs, but the added value stems from the use of data, where anyone can innovate and develop new business from the data. In order to realize this goal, the most critical requirement is that the governments need to release their data sources for its citizens and private sector, and the opened data has to satisfy specific needs in a business context. This mainly requires three aspects: high quality, high update frequency, and ease of reuse [21].

There currently exists multiple business models around open data [1]. These models allow companies to build viable businesses solely based on open data, or enhance their existing products and services with new data [22].

However, in order to create new innovations and develop new businesses around open data it is necessary to know what the companies expect from the current and future open data and what do they want to do with it.

This study aims to answer to a question "*How do companies perceive open data?*". This question is divided into sub-questions: "*How does the company size affect the actions in open data ecosystem?*", "*How do the perceptions differ within open data ecosystem?*", and "*What affects the perceived added value?*". To understand these considerations in the business, our research group conducted a survey of 45 business organizations, collecting information on themes such as interest towards data business in general and application of open data in the current business strategies. Based on our observations, the smaller organizations are more eager to adopt open data into their business, but there also seems to be confusion with the strategy that successful application of open data practices would require. In general, the business needs more success stories and positive example cases to invest in the open data practices.

Rest of the paper is structured as follows: In the next section the related research about similar studies are covered and their relevant findings are described. The third section outlines the research process: protocol, participants, and data analysis. In the fourth section, the results are summarized and discussed further with the statistical significance and implications from the hypotheses. The fifth section offers discussion about the most important findings and implications. We conclude this paper in the sixth section.

2. RELATED RESEARCH

The entire worth of the software industry is approximately

This is a pre-print version of an article. The actual version will be published in ACM DL at http://dl.acm.org/citation.cfm?doid=2994310.2994312. Please use the official version of the paper and publication reference when citing: Antti Herala, Jussi Kasurinen, Erno Vanhala. 2016. Current status and the future directions of open data: perceptions from the Finnish industry. In Proceedings of the 20th International Academic Mindtrek Conference. ACM, New York, NY, USA, 68-77. http://dx.doi.org/10.1145/2994310.2994312

ISBN -.
DOI: http://dx.doi.org/10.1145/2994310.2994312

407 billion USD [17] if taking into account only the companies, which provide software as a product. If every area of industry, which applies some form of a software component is taken into account, there is a question of what product or service industry provider doesn't fit the bill, since almost every new product or service has some form of a software component. In this sense, the application areas of open data are not limited by the area of industry, since almost every business domain has some service, product or analysis method, which could apply open data or open data-related services. Open data has a potential to become a major component for business [21], even to the extent that it should be possible for companies to exist solely on the business based on the open data resources — as has already happened in the U.S. [22].

Gonzalez-Zapata and Heeks [7] identified four different perspectives of open government data (OGD): Bureaucratic, Technological, Political, and Economical. The Chilean case study [7] found that in the analysis of stakeholders of OGD, the private companies and economic perspective were expressed the least as stakeholders from all of the four perspectives. Their results are indicating that companies do not have any specific strategies for the use of open data. However, at the same time, the open data activists from the supply side see the fostered innovation and entrepreneurship as a sufficient economic impact, disregarding the enhancement of existing businesses.

Jaakkola et al. [11] used a survey to determine, how Finnish organizations are interested towards open data and its applications. Jaakkola et al. targeted private and public organizations with a survey in the Satakunta region, receiving 43 responses. They measured the knowledge about open data, the business opportunities, and information related to open data. The study discovered three main results: 1) companies do not have experience about open data, 2) companies can see the business potential of open data but not the business opportunities, and 3) information about open data is lacking. In their article Jaakkola et al. also determine different roles in open data ecosystem that are applied in this article. However, the concept of open data ecosystem is much grander and divided than the roles, as is presented in the literature review by Zuiderwijk et al. in 2014 [23].

3. RESEARCH PROCESS

In order to find out what perceptions companies have concerning open data, a quantitative survey was conducted in the Spring of 2015. The objective of this survey was to determine the corporate opinion towards published open data, and the current methods of publishing. The survey method for data collection was selected in order to engage multiple companies within a short period of time.

Protocol: In order to collect data from the companies, a five-minute online survey was conducted. The survey comprised sixteen multiple choice and open-ended questions and it was constructed for Finnish companies only, so the survey was conducted only in Finnish to ensure that all respondents understood the question items correctly. The survey was structured into three topics: basic information, open data in business, and future contacting. The goal of the survey was to assess, how companies currently perceive open data, how their actions affect their views, and where does the added value come from in their business domain. In addition to the main survey, the industry discussion panel was formed as a professional discussion forum and has been open for registration continuously. The registration form for industry panel contained questions about basic information concerning the companies and networks, as well as their existing interest towards data business. Both the survey instrument and the panel registration form are described in the appendix.

Participants: The survey was sent to an industry discussion panel — in which each participant voluntarily joined per their interest — that consisted of one hundred companies and nine networks of organizations. In order to conduct a business-focused analysis, the networks were left out from the analysis. From each business organization, only one submitted answer was accepted to maintain a balance between the different sized organizations and to ensure that the data does not over-represent the larger organizations of the industry panel. Out of the 100 companies, 45 individual answers were collected. 19 out of the 45 respondents (42.2%) were identified directly as software development companies. 10 respondents (22.2%) were identified as companies working directly with information technology, such as IT consultation or hardware provider. The rest of the companies represent different business domains which by themselves do not have a direct link to information technology, such as machinery and process design, industrial design or security. The companies are identified as Micro ($<$ 10 employees), Small ($<$ 50 employees), Medium ($<$ 250 employees), and Large (250+ employees), according to the European Commission's definition [5]. For the survey respondents, the distribution of sizes was as follows: Micro 62.2%, Small 17.8%, Medium 8.9%, and Large 11.1%.

Data analysis: The hypotheses of this research are divided into two categories: size-related and role-related. The first three hypotheses H1-H3 are used to establish if the company size affects their actions related to the data business and open data ecosystem. Sayogo et al. [20] determined that smaller companies can benefit more from publishing open data, than the larger companies. With the first three hypotheses, the goal is to understand if similar phenomenon in the open data consumption is possible from the company point of view. The hypotheses H4, H5 and H6 take into account the roles in open data ecosystem and try to establish their impact on the perceptions about open data and data business. The article by Jaakkola et al. [11] determined that definite roles exist in the open data ecosystem; in our research, the objective is to test if these roles have an impact on the business of different companies. In the final hypothesis H7, the aim is to find out how do the interests towards specific civil sectors affect the added value from data. As stated by Manyika [15], open data can bring a different amount of economic value to different sectors. This hypothesis brings additional insight about the impact of current trends on the sources of added value of open data.

The following hypotheses were developed and assessed via the survey instrument:

- Hypothesis 1: *The target of interest towards data business is determined by human-based resources.* (Applies registration form Q10 and Q11)
- Hypothesis 2: *Size of the company determines the key source of added value.* (Registration form Q10 and Survey Q10)
- Hypothesis 3: *The availability of human resources increases actions in every role of open data ecosystem.*

(Registration form Q10 and Survey Q11)

- Hypothesis 4: *Roles in open data ecosystem defines the necessary information about open data.* (Survey Q11 and Q12)
- Hypothesis 5: *Role in open data ecosystem determines the sources of added value.* (Survey Q11 and Q10)
- Hypothesis 6: *Role in open data ecosystem defines the interest towards data business.* (Registration form Q11 and Survey Q11)
- Hypothesis 7: *Source of added value correlates directly with the interest towards a specific data topic.* (Survey Q10 and Q6)

The hypotheses H1, H2, and H3 were analyzed using Least Squares-method. The method was selected because the data fills the four recommendations offered by Gray and MacDonell [8]: multiple degrees of freedom, homogeneous dataset without drastic outliers, linear variables, and harmonious data. With the sample size of 45, the analysis cannot be used for predictivity, but it offers insight towards the trends. The hypotheses with non-linear variables, H4, H5, H6, and H7, are analyzed with descriptive analysis and their statistical significance is tested with Pearson's χ^2 test [3].

The organizations were divided into different role groups based on their open data practices. Jaakkola et al. [11] identified five roles, which were defined as follows:

- *Suppliers* are organizations that supply data through an interface or a portal.
- *Aggregators* are organizations, who collect and aggregate data for visualization and analysis.
- *Developers* create applications based on the available data to benefit consumers as well as other organizations in their ecosystem.
- *Enrichers* are actively seeking added value from open data to their existing products and/or services.
- *Enablers* facilitate the opening and usage of open data by offering hosting services, instrumentation, and consultation and advisory services.

The article also presents a sixth role: a consumer. However, the role of the consumer is usually understood as a citizen or citizens in the ecosystem, which is not in the scope of this study. Excluding the supplier and consumer, other roles can be described as intermediaries between the supplier and consumer. The roles and their interactions are presented in Figure 1.

4. RESULTS

In this section, each of the hypothesis is presented separately and reflected to the available data. Before the hypotheses, the statistical significance is addressed and the implications of these results are discussed in the last part of this section.

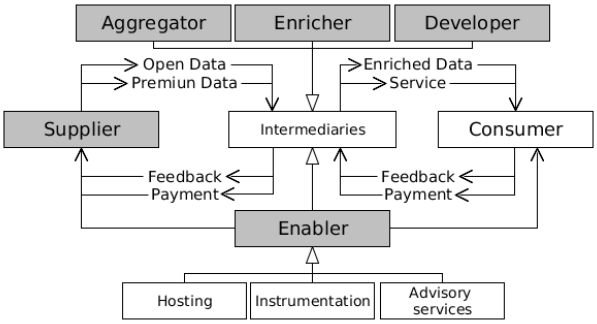

Figure 1: Roles in open data ecosystem, adapted from [11].

Table 1: Pearson's χ^2 test results

Hyp.	Degrees of freedom	χ^2 value	χ^2 critical (5%)
H4	28	9.7467	41.34
H5	16	6.4189	26.3
H6	40	4.6832	55.76
H7	72	26.8800	92.81

4.1 Statistical significance

The hypotheses from H4 to H7 were tested for statistical significance with Pearson's χ^2 test. The test measured if the measured variables correlate with each other; the tests were measured with the limiting value of 5% and the results are presented in Table 1.

Pearson's statistics is used to determine if there is a correlation between two variables: the experimental values are being compared with the critical value [9]. Greenwood and Nikulin continue that with Pearson's χ^2 test there are always two hypotheses: the hypothesis that variables correlate or the null hypothesis that the variables do not correlate. When the χ^2 value is higher than the critical value, the null hypothesis is rejected and similarly if the critical value is not exceeded, the null hypothesis is accepted. The critical value is determined by the level of significance and available degrees of freedom [18]. The level of significance was determined to be the least acceptable level of 0.05 [4]. In this study, our hypotheses were — for every separate hypothesis from H4 to H7 — that the variables in the analysis correlate. However the χ^2 value does not exceed the critical value in any set of variables, which dictates that the variables do not correlate and there cannot be any definite outcomes. In other words, our findings are not predictive, they are descriptive.

4.2 Hypotheses

In this article, we tested seven different hypotheses; the results are summarized in Table 2 and presented further in this subsection. The implications of these results are discussed after the hypotheses.

Hypothesis 1: ***The target of interest towards data business is determined by human-based resources.***

The first hypothesis was tested with the data from registration form items Q10 and Q11. The data analysis indicates that the interest towards any section of data business is not

Table 2: Summary of the hypotheses and results.

Hyp.	Result	Justification
H1	Rejected	The analysis indicates that larger companies are more interested in every section of the data business.
H2	Accepted	The analysis indicates that the smaller companies prefer application development and specialized use of open data, while the larger companies tend to lean more towards data aggregation, data usage in research and development, and working process improvements.
H3	Rejected	Smaller companies are more interested in the roles of developer and enricher than larger organizations, who prefer the roles of aggregator, enabler, and supplier.
H4	Rejected	The interest towards information about the open data is unanimous and not related to these roles in this set of organizations. The most important information are 'lists the about open data' and 'lists about the prospective dataset'.
H5	Rejected	The sources of added value are very similar between different roles and since the data does not correlate, any conclusions from small differences would be meaningless. The most potential sources of added value were data usage in an application and data aggregation.
H6	Rejected	All of the groups are interested in the same aspects of data business. The most interest is pointed towards data-based products and data enhancement.
H7	Inconclusive	All of the sources of added value lean towards specific sectors, but it does not provide enough insight to make conclusive deductions. The most popular data topics are the geographical, traffic, and IT sectors.

determined by the size of the company. The analysis, presented in Figure 2, clearly indicates that the interest towards different sections rise when the size of the company grows. Some of the data business categories are rising faster than others, but with this data sample, statistical significance cannot be found. This leads us to discard the hypothesis H1.

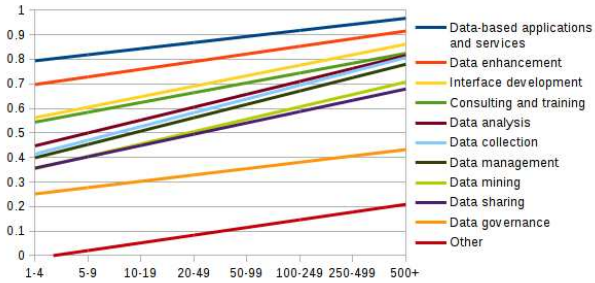

Figure 2: Least Squares-analysis on company size correlation to interest in data business.

Hypothesis 2: ***Size of the company determines the key source of added value.***

This hypothesis is tested with an item from the registration form (Q10) and an item from Survey (Q10). The analysis in Figure 3 shows that using open data in an application is far more popular for small companies than it is for the large ones. Also the category Other is more popular in smaller companies, dictating that smaller companies use open data for specific, specialized goals not listed in this survey. Based on the analysis, larger companies tend to be more interested in the research and development, data combination and intensifying their own working processes. This result leads us to accept hypothesis H2.

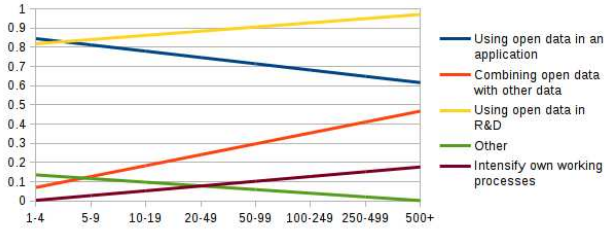

Figure 3: Least Squares-analysis on company size correlation to the source of added value.

Hypothesis 3: ***The availability of human resources increases actions in every role of open data ecosystem.***

In order to find out if the hypothesis is acceptable, questions Q10 from registration form and Q11 from the survey are used. Based on the analysis seen in Figure 4, the role of the aggregator is rising rapidly with the size of the company. Also, the role of enabler and supplier are more popular with larger organizations. Only the roles of developer and enricher — direct manipulators of data — are more popular with smaller companies, which would indicate similar results as in hypothesis H2. This result is rejecting the hypothesis H3.

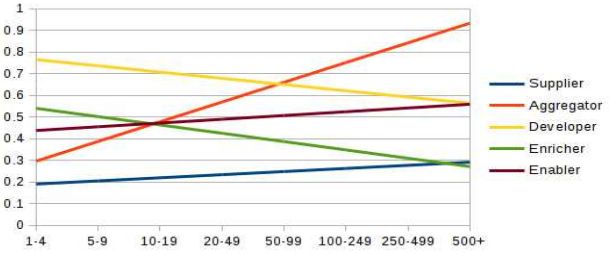

Figure 4: Company size correlation to roles in open data ecosystem analysed with Least Squares-analysis.

Hypothesis 4: Roles in open data ecosystem defines the necessary information about open data.

This hypothesis was tested with the survey items Q11 and Q12. As based on the results illustrated in the Figure 5, the different roles of the open data users correlate between the different groups. In all groups, the application types list about open data and list about prospective dataset are the two most important information sources while open data service-level agreement was the least important factor. Even if there is really no strong correlation between the roles and the usages (see Table 1), there are some minor differences; for example, the enablers are relatively most interested in the applications that apply open data sources. The data would indicate that all companies are fairly interested about two main aspects: what data is currently available and what data will be available in the future. All of the groups are unanimous in their preferences, which leads us to reject hypothesis H4.

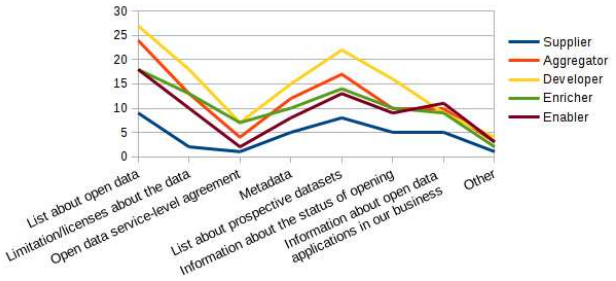

Figure 5: Necessary information about open data for different roles.

Hypothesis 5: ***Role in open data ecosystem determines the sources of added value.***

This hypothesis was tested with survey questions Q11 and Q10. It was expected that different roles would perceive the added value from open data differently, but as can be seen from Figure 6, the ratio of answers were practically unanimous. The ratio was calculated by dividing the number of answers of one role about one source of added value with the sum of all the answers in that group. There is a small rise of added value from open data in applications for the group of developers, but the rise is a lot smaller than initially expected and it takes space from the R&D, which is logical. Added value from combining open data with other data is similar to every role, while it was expected that aggregators would have more interest towards that than the rest of the groups. The differences are significantly small and since the data does not correlate based on Pearson's χ^2 test, drawing definite conclusions from the data is not feasible. However, the data would suggest that we have to reject H5.

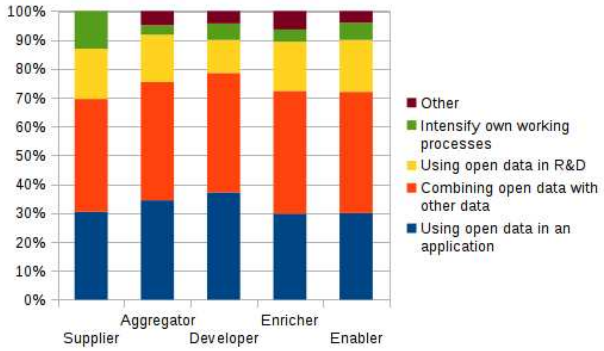

Figure 6: The source of added value for different roles.

Hypothesis 6: ***Role in open data ecosystem defines the interest towards data business.***

For this hypothesis, the questions from registration form and survey were used, registration form Q11 and survey Q11. Looking at Figure 7, it is quite evident that every company, independently from the role, is interested in the same aspects of data business. The ratios in the figure are calculated by summing all of the answers per role and using that as the divisor for the number of answers in each interest. All of the groups raise the data-based applications and services as the most interesting topic. Otherwise, the results seem more uniform, only the interest towards data governance and data collection drop from the others. While the data is not statistically significant (Table 1), the participating companies do seem to be interested in the same fields of open data and data business. This result would suggest rejecting the hypothesis H6.

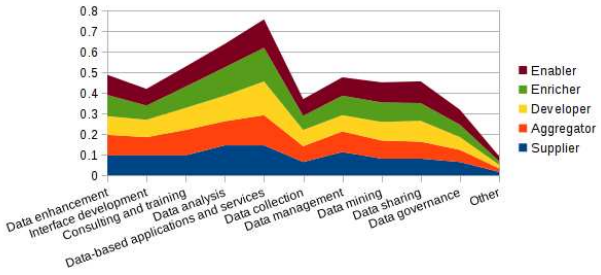

Figure 7: Roles per interest towards data business.

Hypothesis 7: ***Source of added value correlates directly with the interest towards a specific data topic.***

Figure 8 illustrates the division between the application methods with the most common application domains identified from the industry. In all of these domains minimum of ten organizations indicated that they have some business-related open data application in mind. To specify an answer for this hypothesis, survey questions Q10 and Q6 were used. The scale in the figure is the number of answers per data topic.

From Figure 8 the sectors where open data could bring in the most added value are IT and traffic sectors as well as the geographical applications. In every field, the value of combining open data exceeds the interest towards direct

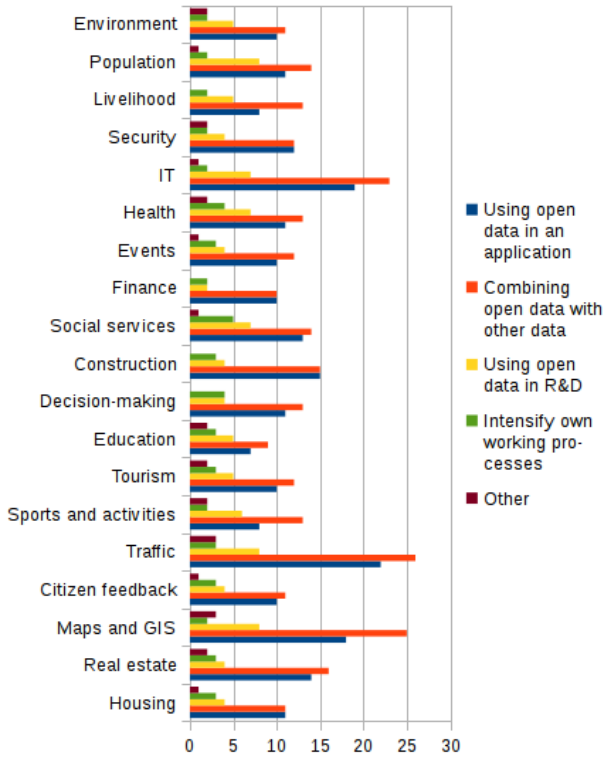

Figure 8: Added value from different data topics

applications. Only fields where the interest towards direct applications are at least on par with data aggregation are housing, construction, finance, and security.

From this data, the acceptance or rejection of the hypothesis H7 is inconclusive. On one hand, the interest peaks definitely towards few fields, but on the other hand, it includes all of the sources of added value. From the data, it is impossible to ascertain that the specific fields lean towards a specific source of added value. In a dataset that is not correlating, according to Pearson's χ^2 test, the differences are not that remarkable. However, the resulting figure does raise interesting insights on the mindset of the companies about what they want to do with the available data. Even with the differences in popularity, the sources of added value have approximately the same shape in every field. The popularity of these topics is similar to the results of the study analyzing the commercial interest towards Spanish open data domains [21], where the geographic and financial data were clearly the most popular topics, 51.1% and 46.8% respectively. In our data traffic and IT are also popular topics while the financial data is not in the spotlight. However, both datasets agree on the popularity of geographic information.

4.3 Implications of the results

As stated earlier, due to the constrained dataset, the results cannot be used to predict and present definite correlations. However, some implications can be drawn from the data and they are discussed in the following. The acceptance of H2 is not that surprising; it offers insight to the fact that smaller companies are trying to create applications based on open data; they are trying to create direct value with products. Some products and even companies can be built essentially on top of open data. Larger companies are more interested in indirect value, such as research and efficiency improvements.

In the hypothesis H3, the implications are fairly similar to H2: smaller companies prefer the role of developer and enricher, which would indicate that smaller companies are trying to benefit directly from open data and transform the freely accessible data to a new or improved product. The rapidly increasing interest towards the role of aggregator by larger firms is in line between the two hypotheses. Larger companies, who are interested in data aggregation and R&D with open data are heavily leaning towards the role of aggregator, disregarding the other roles.

H4 would indicate that what companies really need is information about the sources of open data, where it is and how can it be accessed. The problem with this is the fragmented nature of the open data release since open data in Finland is available in a governmental portal while some cities are providing their data through websites.

The analysis of H6 brings an interesting contrast to the other hypotheses, namely H2 and H3. Those two hypotheses stated that smaller companies are more interested in application development based on open data, but at the same time, H6 states that most of the companies still regard applications as the data-based core business. This shows the difference for a data business as a whole and the open data business: larger companies are regarding open data as something more than just a business resource for direct value, while smaller companies use it as a resource in order to grow.

In H7 the most selected fields were traffic, IT, and maps and GIS (Geographic Information System). Since the geographic data is practically ready for publishing and does not require any anonymization, it is an easy solution for data suppliers to share. The use is also universal since multiple different applications can be improved or innovated based on the available data about locations and other spatial data. The traffic sector has been discussed a lot in the open data community since it could bring multiple benefits to the society. Such like the spatial data, traffic data can be used to satisfy multiple aims, from which the smart city viewpoint is one of the grandest. The interest towards open data in IT sector can be used directly to improve their existing products, benefiting the existing clientele and society, and also to increase the portfolio of services. The analysis also shows the differentiation of the interest towards applications versus the combination of data, where data aggregation is more popular in any field. This would confirm that the added value of open data does not come from the data in itself but through shared value from the social and economical standpoint [13].

What strikes most from the data is the fact that while the sample is not statistically significant, the results are looking very similar, nearly identical in H4, H5, and H6. Even when the companies are all working in their own, respective fields and extending their own portfolio, this sort of uniformity seems impossible. The interest towards specific data topics is also biased towards few topics, which are easy to use. It would seem that while companies already have experience about open data and they see the possibilities of it, they still follow the directions set by data supplier. This would suggest that companies do have the basic knowledge about open data, but they are unable to follow the data suppliers

as much as would be required, so they follow the recommendations made by the publisher. This limits the reuse of open data since the interested parties cannot find the data they need. It also creates a problem, since the most preferred way of communication and learning about the open data possibilities (see Figure 9) is direct conversations with the suppliers and workshops, which would require knowledge and experience in the domain.

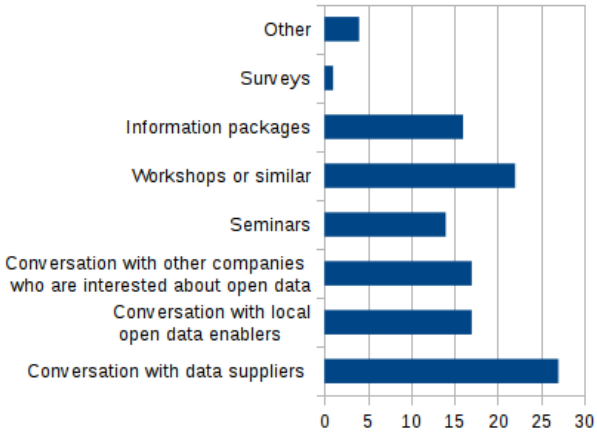

Figure 9: Preferred communications methods from the company perspective (Survey Q13)

The hypotheses of this survey discuss the different applicabilities of the open data. These observations can be summarized as follows:

- Size of a company does not change the interest towards data business, but it changes how the company sees open data. Smaller companies are directing their attentions towards open data based application development, while larger companies prefer indirect applications and benefits of open data.
- Participating in the open data ecosystem does not cause any differentiations about what knowledge is needed about open data, where the added value comes from, nor the views about data business.
- The added value in open data ecosystems origins from the open data based application development for smaller companies and data aggregation for larger firms. It would seem that in every field of society, the interest towards combining data exceeds the interest towards data-based applications.
- Companies do not have the capability to follow the data supplier as much as they are required to in order to utilize open data to the full extent. It seems that companies, in general, do not know what they want, or what is offered.

5. DISCUSSION

Based on the results, it can be argued that companies are currently following the open data trend, but it is not clear for them how and where they could apply the open data effectively. In the case of applicability of the open data, the case study by Gonzalez-Zapata and Heeks [7] found that the companies are not that interested in open data and governments are not that interested in cooperating with private firms. Our findings in this article support some of the findings, but are not as definite; the companies are interested in the possibilities of open data, but they do not know what data could be applied and when. Our data and our findings are closer to the findings of Jaakkola et al. [11]: companies see and understand the business potential in the open data activities, but they are lacking the means such as knowledge and experience to tap into the business opportunities. In addition, the companies lack the communication channels concerning the open data possibilities, as seen in Figure 9. In any case, the research done by Jaakkola concentrates on one provincial region, while our study has a national focus. Comparison between our findings and Jaakkola would suggest, that the results do not differ much between the different Finnish provinces and at least in the Finnish domain, these concerns are valid and need to be addressed.

Blame about the lack of communication can be shifted towards data supplier, who should be able to discuss with companies when they are opening their data. Some governments and municipalities nowadays are opening their data based on the reason that it has to be open [12]; they do not think about the opportunities that can emerge from the data. Therefore there should be more dialogue between data openers and users, whatever the role in the ecosystem or society. Currently, both sides are concentrating on their own data policies and they practically disregard each other in the process. In order to find new economic development from open data, the publisher should know what data is necessary to provide for a company to use it in their product or service. This works both ways: the companies who are interested in the government's data should actively research and discuss with the supplier about what they want and how. This would enable the publisher more freedom and flexibility in their processes and decision about which set of data will be opened.

These issues are not in any way new and they are being actively tackled by organizations such as OGP with other similar issues [16]. The most distressing fact is that these aspects we found in this study have been addressed for years, but still it seems to be a problem. So it begs a question: are the governments doing enough and are they doing the right decisions and actions to further the economic impact and value of open data?

5.1 Validity

First of all, in the survey the sample size of 45 organizations may seem somewhat limited. However, similarly, as in [10], the sample size is small but sufficient if analyzed correctly. In our study, the threat of overfitting the data — over-representing certain sub-groups of participants — was addressed by selecting the organizations to represent different software domains and types of organizations, and only allowing one responded per organization to prevent larger organizations from over-representing themselves in the data. Also related to the number of organizations, a paper by Sackett [19] discusses the conceptualization of signal-to-noise-ratio in statistical research. Their approach to define confidence as based in practicality of observations: *confidence = (signal / noise) * square root of sample size.* In practice, this indicates that the confidence for the result

being non-random weakens if the amount of noise increases while signal decreases. In the Sackett model, the attributes are abstracted, meaning that the noise can be considered to be any uncertainty on the data. In this study, the sample population was first screened with a participation survey and then ensured with the industry discussion board that the operating domain and the intentions of the answers were understood correctly. Since the concept of the Sackett study is that the confidence in the survey data increases the validity of the study, our study addressed this problem by screening the sample for wanted types of organizations. Therefore it can be argued that our signal was very good and noise low, so the overall confidence should be good. Finally, the Pearson's χ^2 test is considered a reliable tool for assessing goodness of fit between two groups (e.g. [3]).

Considering the technical structure of the survey, the responses were collected with established methods, for example by applying the 'like best' (LB) technique with roles established in the prior research, which is common survey data collection strategy (e.g. [6]). In addition, Kitchenham et al. [14] divides comparable survey studies into exploratory studies from which only weak conclusions can be drawn, and confirmatory studies from which strong conclusions can be drawn. This study is an exploratory, observational, and cross-sectional study that explores the phenomenon of applying open data and the expectations of the open data in practice, and provides more information and understanding to both researchers and practitioners to refine their future work into the topic.

The validity can be also questioned with the biased size of participating organizations as well as their individual fields of business. Only 11.1 percent of the respondents were large organizations, while the rest of them were SMEs. This issue of validity was noticed by Verhulst and Caplan, where the smaller organizations were recognized as the main beneficiary because of the equalizing effects of open data. The positive effects of open data are most likely felt by smaller organizations first because they have limited access to data, information, and analytical tools, a trait not shared with larger companies [22]. 42.2 percent of the survey participants are doing business directly in the software development field and 22.2 percent of the companies are doing business in an area closely related to information technology. It causes a bias towards the software development and software business in general. However, this threat to validity would be larger, if the companies were randomly selected and non-screened since the amount of irrelevant noise would also increase. For this study, only companies which are currently doing business with the open data or within the open data ecosystem are participating in the survey since they have tangible viewpoints.

6. CONCLUSION

In this paper, we have presented our results on a national survey regarding the application and adoption practices of open data in Finland. The survey collected 45 responses from several different business domains, such as construction, software development, tourism and even sports event organizers. In general, the results confirmed our prior expectations knowledge based on the prior research and another regional survey on the application of open data in Finland.

Across the different industry domains and company sizes, the survey implicates similar results: the application of open data offers potential for economic value in several sectors, but so far the companies have had difficulties finding a suitable strategy for implementing open data resources. The most important observation was that the companies cannot find the open data they need for their business, even though they are willing to use the data. The scattered supply of open data requires more resources from companies than they want to deploy. In general, the results were in line with the earlier regional studies, and confirm that the observations and recommended actions for promoting open data initiative needs to address these issues on the national level.

In the future, the results of this study can be used to guide the decision-makers and open data suppliers to further optimize their strategies of publishing data. One possible line of inquiry would be to focus on determining the possibilities and most promising venues, which could be used to supply open data more efficiently to citizens and businesses with reasonable costs to the publisher.

7. ACKNOWLEDGMENTS

We are grateful to 6Aika-project for their cooperation. We would also like to thank DIMECC S4Fleet-project and the network of companies associated with it.

8. REFERENCES

[1] F. Ahmadi Zeleti, A. Ojo, and E. Curry. Exploring the economic value of open government data. *Government Information Quarterly*, Feb. 2016.

[2] P. Booth, P. Gaskell, and C. Hughes. The economics of data: quality, value & exchange in web observatories. pages 1309–1316. ACM Press, 2013.

[3] H. Chernoff and E. L. Lehmann. The Use of Maximum Likelihood Estimates in χ^2 Tests for Goodness of Fit. *The Annals of Mathematical Statistics*, 25(3):579–586, Sept. 1954.

[4] M. Cowles and C. Davis. On the origins of the. 05 level of statistical significance. *American Psychologist*, 37(5):553, 1982.

[5] European Commission. What is an SME? - European Commission.

[6] A. Fink and J. Kosecoff. *How to Conduct Surveys: A Step by Step Guide*. Sage Publications, Inc. California, 1985.

[7] F. Gonzalez-Zapata and R. Heeks. The multiple meanings of open government data: Understanding different stakeholders and their perspectives. *Government Information Quarterly*, 32(4):441–452, Oct. 2015.

[8] A. R. Gray and S. G. MacDonell. A comparison of techniques for developing predictive models of software metrics. *Information and Software Technology*, 39(6):425–437, Jan. 1997.

[9] P. E. Greenwood and M. S. Nikulin. *A guide to chi-squared testing*, volume 280. John Wiley & Sons, 1996.

[10] J. Iivari. Why are CASE tools not used? *Communications of the ACM*, 39(10):94–103, Oct. 1996.

[11] H. Jaakkola, T. Mäkinen, and A. Eteläaho. Open Data: Opportunities and Challenges. In *Proceedings of*

the 15th International Conference on Computer Systems and Technologies, CompSysTech '14, pages 25–39, New York, NY, USA, 2014. ACM.

[12] M. Janssen, Y. Charalabidis, and A. Zuiderwijk. Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, 29(4):258–268, Sept. 2012.

[13] T. Jetzek, M. Avital, and N. Bjørn-Andersen. Generating Value from Open Government Data. In *Proceedings of the 34th International Conference on Information Systems. ICIS 2013*, Atlanta, GA, 2013. Association for Information Systems. AIS Electronic Library (AISeL).

[14] B. Kitchenham, S. Pfleeger, L. Pickard, P. Jones, D. Hoaglin, K. El Emam, and J. Rosenberg. Preliminary guidelines for empirical research in software engineering. *IEEE Transactions on Software Engineering*, 28(8):721–734, Aug. 2002.

[15] J. Manyika. *Open data: Unlocking innovation and performance with liquid information*. McKinsey, 2013.

[16] Open Government Partnership. From commitment to action — Annual Report 2015. Technical report, 2015.

[17] C. Pettey. Gartner Says Worldwide Software Market Grew 4.8 Percent in 2013, 2014.

[18] R. L. Plackett. Karl Pearson and the Chi-Squared Test. *International Statistical Review / Revue Internationale de Statistique*, 51(1):59, Apr. 1983.

[19] D. L. Sackett. Why randomized controlled trials fail but needn't: 2. Failure to employ physiological statistics, or the only formula a clinician-trialist is ever likely to need (or understand!). *Canadian Medical Association Journal*, 165(9):1226–1237, 2001.

[20] D. S. Sayogo, J. Zhang, T. A. Pardo, G. K. Tayi, J. Hrdinova, D. F. Andersen, and L. F. Luna-Reyes. Going Beyond Open Data: Challenges and Motivations for Smart Disclosure in Ethical Consumption. *Journal of Theoretical and Applied Electronic Commerce Research*, 9(2):1–16, 2014.

[21] D. Tinholt. The Open Data Economy: Unlocking Economic Value by Opening Government and Public Data. *Capgemini Consulting Analysis*, 2013.

[22] S. Verhulst and R. Caplan. Open data: A twenty-first century asset for small and medium-sized enterprises. Technical report, 2015.

[23] A. Zuiderwijk, M. Janssen, and C. Davis. Innovation with open data: Essential elements of open data ecosystems. *Information Polity*, 19(1, 2):17–33, 2014.

APPENDIX

Survey

Background information

1. I answer to this survey as:
   (a) Company representative/Network representative
2. First name
3. Last name
4. Name of the company/network
5. City

Open data in business

1. (Q6) Which data or programming interfaces are the most interesting to you from the viewpoint of your company's/network's business?
   (a) Housing / Administration / Real estate / Maps and GIS / Citizen feedback / Culture / Law and regulation / Traffic / Sports and activities / Tourism / Counselling / Education / Decision-making / Construction / Social services / Finance / Events / Health / IT / Security / Livelihood / Population / Environment / Other
2. (Q7) Does your company/network have services or products in use or in design, where you apply open data or open interfaces?
   (a) Yes / No
3. (Q8) If yes, what kind of product/service and what data/interface it is using or could use?
4. (Q9) Do you think that using open data could bring added value to your company/network in the future?
   (a) Yes / No
5. (Q10) If yes, where does the added value come from?
   (a) Using open data in an application / Combining open data with other data / Using open data in R&D / Intensify own working processes / Other
6. (Q11) What are the roles of your company/network in open data ecosystem?
   (a) Supplier (provides data) / Aggregator (collects data from different sources and visualizes it) / Developer (develops applications based on the data) / Enricher (combines open data with their own data) / Enabler (raises awareness and interest towards open data) / Other
7. (Q12) What are the most important sets of information for your company/network that the cities should provide about open interfaces and data (top 1-3 most important)?
   (a) List about open data / Limitations or licenses about the data / Open data service-level agreement / Metadata / List about prospective datasets / Information about the status of opening / Information about open data applications in our business / Other
8. (Q13) What are the preferred methods of communication your company/network would be most interested about?
   (a) Conversation with data suppliers / Conversation with local open data enablers / Conversation with other companies who are interested about open data / Seminars / Workshops or similar / Information packages / Surveys / Other
9. Comments about the industry discussion panel?

Future communications

1. I want to be contacted by a representative about utilizing open data.
2. Which city representative you want to have communications with?
   (a) Helsinki / Espoo / Vantaa / Tampere / Turku / Oulu

Registration form (for companies)

Contact information

1. Name
2. Appellation
3. Phone number
4. E-mail address
5. Organization
   (a) Company/Network

Company information

1. Company name
2. Company website
3. Business ID
4. Office locations
5. (Q10) Company size
   (a) 1-4 / 5-9 / 10-19 / 20-49 / 50-99 / 100-249 / 250-499 / 500+
6. (Q11) Interest towards data business
   (a) Data-based applications and services / Data enhancement / Interface development / Consulting and training / Data analysis / Data collection / Data management / Data mining / Data sharing / Data governance / Other
7. Expectations from the panel