



International Conference
for Entrepreneurship,
Innovation and Regional Development

**University-Industry Links:
Coproducting Knowledge,
Innovation & Growth**
31 Aug – 1 Sept 2017
Makedonia Palace Hotel
Thessaloniki, Greece



UNIVERSITY-INDUSTRY COLLABORATION THROUGH BIG DATA AND OPEN SCIENCE

M. Roman¹, J. Liu², T. Nyberg³

¹ Aalto University, Department of Industrial Engineering and Management, Maarintie 8, Espoo, Finland, mona.roman@aalto.fi

² Aalto University, Department of Industrial Engineering and Management, Maarintie 8, Espoo, Finland, jinqwei.liu@aalto.fi

³ Aalto University, Department of Industrial Engineering and Management, Maarintie 8, Espoo, Finland, timo.nyberg@aalto.fi

There exists today little motivation for researchers to share their research data, which is their source of scientific reputation. This paper addresses this major challenge for open science and identifies alternative business models for researchers to open their data repositories for external users. Our research question is “How to open big data in a way that it can be transferred to sustainable business cases that bring value to data owners as well as potential external users?” Our paper is based on a case study of a Finnish foundation and its currently closed database. As key findings, we identify opportunities, barriers and possible solutions for opening this database based on 11 interviews with key stakeholders. Our research contributes to open science literature through illustrating how to open big data in a way that brings value to both data owners and external users.

Keywords

Big Data, Open Science, Open Innovation, Research Data, University-Industry Collaboration

1. Introduction

There have been many initiatives and actions to open government data in Europe to boost open innovation. Open government data means that data produced or commissioned by government or government controlled entities can be freely used, reused and redistributed by anyone [1]. In Europe, there is a pan European network offering advice on implementation of the (Revised) Public Sector Information Directive, best practices [2] and local guides [3]. While the open government data is already well established including the ways of operating and best practices, the open science is still in its infancy. Open science can be defined as a movement to make scientific research, data and dissemination accessible to all levels of an inquiring society [4].

The European Commission is running a pilot under Horizon 2020 called the Open Research Data (ORD) Pilot that aims to improve and maximize access to and re-use of research data generated by Horizon 2002 projects [5]. The project recognizes the different nature of research data from government data, and thus the need to balance openness and protection

of scientific information, commercialization and intellectual property rights (IPR), privacy concerns, security as well as data management and preservation questions [5]. Furthermore, EC has set up OpenAire2020 as a key infrastructure to monitor H2020 research outputs [6]. While the infrastructure and the policies related to open research data are being to emerge, the decision what data to open and how widely still rests on the individual researchers.

Our research takes an explorative approach to investigate how to motivate big data owners to share their data and how to get external stakeholders interested to exploit it through a case study of a database owned by a Finnish foundation, which we refer to as Foundation A. We utilize semi-structured interviews with key stakeholders as a method for primary data collection. The interviews focus on answering two key questions: 1) how to motivate the owners of the database to share their data and 2) what are the needs and wants of external scientists and industry to exploit this database. Based on 11 interviews with database owners, external researchers and open science experts, we identify alternative solutions to open the currently closed database of the Foundation A.

2. Literature review

Prior literature has identified various models for commercialisation of big data. In principle, these can be put under six different business models. They vary from simple data supplier to service provider based on whether the data is internal or external and what is done with the data (ranging from providing data for reuse, analysing and aggregating existing data or even providing services). Revenue model for the different strategies are mostly subscription based but can also entail “freemium”, pay-per-use or advertising [7]. Once an individual or organisation is comfortable with the notion of commercialising data they will start to release more data and try to move on to more complex business models as they generate more revenue. Complex business models are also harder to execute, prompting open collaboration and co-creation. Scientists need to realise that data ecosystems are essential for making most value out of their big data [8].

For universities, public research institutes and government to collaborate effectively with industry it is important to align the goals of these different organisations. They could form a Technology Transfer Alliance (TTA) and internet-based for-profit models owned by universities while giving rights to inventors and assisting with funding and commercialisation [9]. The challenges of open data can be mitigated by a boundary organisation that supports university-industry partnerships [10]. The boundary organisation, in exchange for sponsoring, allows companies to shape the research programme while funding universities that employ high-calibre scientists to do research on them. The key to overcoming scientists’ motivational challenge is through methods like mediated revealing and enabling multiple goals to co-exist that might not advance companies’ agenda but can achieve high scientific impact during a project. To address the reluctance of researchers to share their big data, it may be best for funding agencies to require that data supporting publicly funded publications should also be made publicly available [11].

However, merely sharing is it is not always enough since raw data itself is often not useable. This means data that is made public also needs essential metadata. Success cases for opening data can be considered the Bermuda Principles of Human Genome Project in 1996

and National Science Foundation (NSF) [12]. In the former it became common practice to make human genomic sequence data openly available in trustworthy data repositories. NSF requires its funded project to share its data with other researchers and the implementation of this is specified in a mandatory Data Management Plan. Publications have also started to use different strategies implementing data sharing. This includes research data as independent information object, data paper or enriched publication.

A collaborative way to adopt big data technologies involving researchers is the 3-legged big data environment strategy focusing on integrating developer community, researchers and business user [13]. The strategy is iterative and encourages collaboration between the three parties involved. It is considered helpful in addressing challenges like reusability, manageability, development scalability and maintainability. It is considered good practice for business leaders and researchers to work together to derive business value from big data. This is due to the technical nature of collecting, storing and analysing the data [14]. To ensure certain quality of the data, tasks like filter, cleansing, pruning, matching etc. should be applied at the earliest stage possible.

Much research and literature has been conducted on the commercialisation of big data in general as well as the needs and obvious benefits of opening up research data [15]. Furthermore, research on how to make this accessible has also been addressed [16]. However, with much emphasis on the open data and open science phenomenon further research is needed on how to actually create value and make use of the big data should it become open and accessible [9]. Our study on Foundation A aims at unveiling some results on this.

3. Research methods

Our research is based on a case study [17]. The role of the case study is to illustrate the underlying opportunities, requirements and bottlenecks for opening data repository for external users. Our main criteria for the selection of a case study is that it needs to provide potential to discover real challenges of opening big data, and to identify ways of overcoming these challenges.

We selected Foundation A in Finland as our case study, as it fulfils the above criteria, and it is readily available for us to study. They own a currently closed database with over 20.000 applications for funding, the funding decisions and the progress of funded applications. We invited all board members of the Foundation A (six persons) in the interviews and one old employee being expert in the topic having worked over 15 years with the database to identify opportunities and constraints related to opening the database. We then invited potential external users to the interview based on the convenience sample, the first author knowing them as doing related research to clarify their interest and needs for database. Furthermore, we invited experts in open science to the interviews to provide their view on the possible barriers and ways to solve them to open the database for external users. We conducted altogether 11 interviews in March-April. The interviews were conducted face-to-face, when convenient, otherwise by phone. The interviews lasted approximately 30 minutes. The informants received a memo of the interview notes, which they were able to verify and add further information. The list of interviews and the informants is presented in Table 1.

Table 1 List of interviews

Title of informant(s)	Organisation(s)	Interview date
Chairman of the Board	Foundation A	2017-03-23
Board member	Foundation A	2017-03-20
Vice Executive Director	Foundation A (until 2013)	2017-04-04
Department Vice-Head for Research, Assistant Professor	Aalto University, Department of Industrial Engineering	2017-03-23
Research Scientist	VTT Technical Research Center of Finland	2017-04-05
Assistant Professor	Hanken School of Economics, Entrepreneurship	2017-03-21
Postdoctoral Researcher	University of Jyväskylä, Department of Computer Science and Information Systems	2017-03-27
Legal Counsel	Aalto University, Research and Innovation Services	2017-04-04
Grant Writer	Aalto University, Research and Innovation Services	2017-04-05
Specialist	Aalto University, Research and Innovation Services	2017-04-06
Senior Statistician	Statistics Finland, Research	2017-04-04

We analysed the interviews through comparing the views of the owners, potential users and external experts in regarding the value and future use of the database. We also collected secondary research data regarding the database and open science strategies and the related business models. Having different data sources through triangulation we can validate our findings from interviews [18]. Based on the interviews and secondary data sources we developed a proposal for possible solutions related to the future use of the database. Based on the views of the key stakeholders, we furthermore selected two of the solutions of sustainable business cases that bring value to both data owners and potential users.

4. Case study

Our case is about a currently closed database that is owned by a Foundation A, which was founded in the 1970s. The foundation has supported thousands of Finnish private persons and small companies with granting conditional pay-back funding. The case study consists of interviews with key current and possible future stakeholders of the database. Based on interviews, we identified as opportunities for researchers with database to conduct high-quality academic research, for policy-makers to enhance current policies and for industries to develop or enhance own ideas as well as to search for potential partners. There are though remarkable barriers (legal, IPR and technical) for opening the database for external users. First, the database contains confidential information regarding individual people and companies. Second, there are technical barriers as the database is not developed for external purposes and thus there are no guidelines, no classification of data contents and no separate copies of it, which all are necessary requirements for external use purposes. These barriers lower the motivation of the Foundation A to share their database for external users, as there is risk that the external user would not treat confidential information accordingly and

could accidentally even destroy the whole database, as it is not possible to define different usage groups with different usage rights.

Based on the interviews, we developed alternative solutions to overcome the barriers. First option is to give an access right for selected researchers who sign NDA to utilize the database for research purposes (either free of charge or with yearly subscription fee). Second option is to develop a passive database with separate copy with anonymous data and restricted contents. Third option is to transfer (sell or give) the database to an external stakeholder such as Statistics Finland or National Archive. Figure 1 illustrates our findings from interviews regarding the underlying opportunities of the database, the barriers to open the database for external users and alternative solutions.

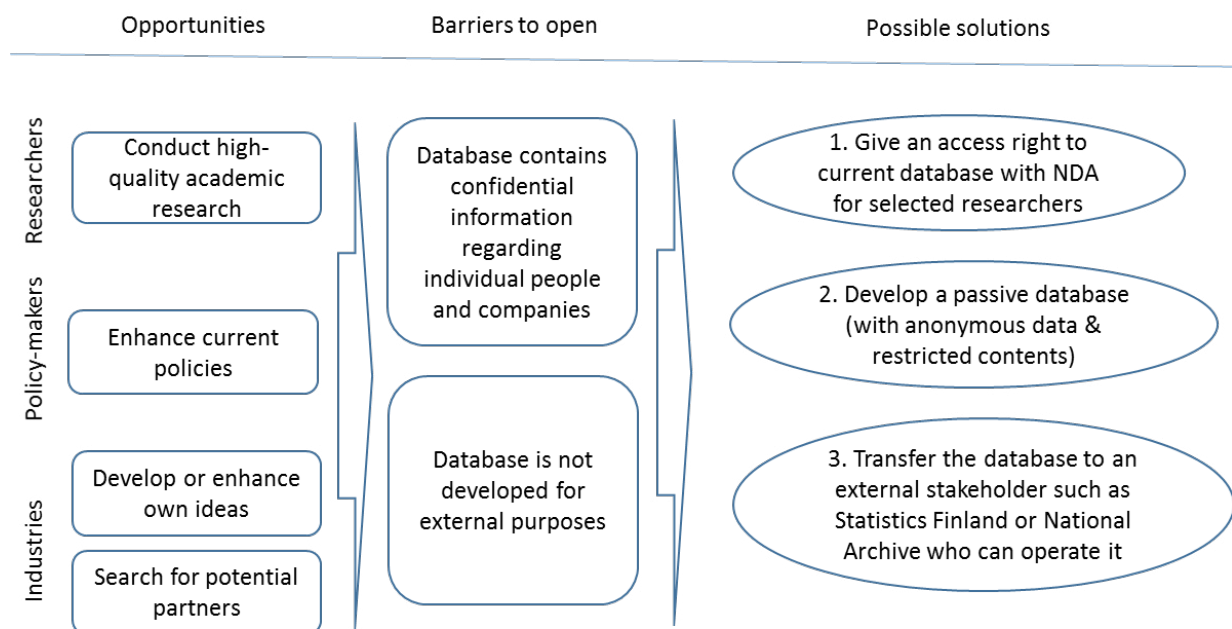


Figure 1 Opportunities, barriers and solutions related to the database of Foundation A

Based on the interviews with potential users of the database, we can conclude that there is initial interest towards the database. The database is seen as comprehensive, unique and large. The real research value of it comes from linking the database with other datasets e.g. from Statistics Finland, to understand the connections of individual people and companies and the influence of this. The willingness to pay for it depends on how well the data suits to researchers' specific research questions and objectives and how easy it is to use the database to collect the data needed. Typically, in commercial databases there is a person working with administrating the database, who researchers can contact for guidance and further information regarding the database. The data is then often anonymous for a researcher, and the administrator handles the raw data that is confidential. Our interviews reveal that when the following conditions are met, fit to the users' specific research questions, easiness-of-use and the administrator support available, the researchers would be ready to pay for the access to the database. If only the first criteria is met, the researchers pay less.

When we compare the user requirements to our possible solutions we can conclude that our first and third solutions are viable. However, as the researchers would not be ready to pay much for pure access, the first solution could only work as an intermediate solution for the Foundation A as long as it has needs for the database. Giving an access right to selected researchers with NDA would support sharing some of the database costs and also offer possibility for the Foundation A to get better understanding about the real value of the database for researchers and other stakeholders. Our second solution about developing a passive database, is not very appealing to the researchers, as it is very relevant for research purposes to link the data to other data sets (on individual level). Thus, there would be no interest to pay for a passive database, and it would not be profitable for the Foundation to do that. Furthermore, this solution is not interesting to policy-makers and industries, as they also need the in-depth information in order to get value of the database. Our third solution to transfer (sell or give) the database to an external provider to administrate is also a viable one. In our interviews, two of the researchers would see this as a good solution of which they would be ready to pay, when they have closely related research projects. In this option the new owner of the database would act as an administrator taking care of the confidentiality issues according to high-quality, standards.

5. Conclusions

Based on our literature review we can conclude that the business models and infrastructures for open research data are emerging. However, as there is still relatively little open research data available, the question remains how to motivate and encourage researchers to start adopting these emerging models and infrastructures. Based on our interviews with open science experts, researchers today have too little knowledge about open science. Some researchers think that open research data means the data can be accessed by anyone whenever and for whatever purpose, similar to government data. However, this is not true as today researchers can themselves decide who can use their data and how. They can restrict the usage to only the members of their research team and then give conditional access to others e.g. to make a joint publication based on the data. In this way, they can also verify that the external user understands the nature of the data and its limitations. They can also define that the research data can be used only in research purposes, and in certain levels.

The key motivation for researchers to open research data is to get an additional publication based on it, as both the journal publication and the data publication are viewed as two individual publications. Opening research data also enables others to get interested in the work of the researcher, read it, cite it and possibly to develop joint collaboration around the research data. The problem today seems to be that the researchers are not aware of these additional possibilities to disseminate their research through publishing the research data. Only one out of the four researchers in our interviews had published open research data.

As the opening of the research data is still its infancy, there are limited possibilities for university-industry collaboration in open science. The current model is to collaborate based on a separate agreement. However, it is important to think ahead about open science and to develop communities around specific fields within participants from universities, industries and society. If the communities become successful, this will get researchers more motivated

to publish research data within them, as the network of potential users is already there. In our case of Foundation A and its database, we found as a long-term solution for it to be combining it to a larger data repository in Finland such as Statistics Finland. In this way, it would be openly available for everyone to exploit it, while the IPR and privacy concerns, would be handled by a database administrator. There are already many researchers doing quantitative research who have found the micro data services of Statistics Finland as a possible source of research data.

Our research contributes to open science literature though illustrating how to open big data in a way that brings value to both data owners and external users. As our research is based on only one case, we recommend future research to examine additional cases in other contexts. It would be important to highlight successful cases from opening research data for both open science and open innovation purpose from different research fields to demonstrate for researchers how to share the data in a sustainable way that brings value to all parties involved, and thus get them motivated to publish their own research data.

Acknowledgement

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 693651, Science2Society, which we are grateful for the financial support. We are also thankful for Foundation A and all interviewees for their important share in the project. Lastly, this article reflects only the authors' view.

References

- 1 Open Government data. 2017. What is open government data. [Cited 6 April 2017]. Available from: <https://opengovernmentdata.org/>
- 2 Share-PSI 2.0. 2016. Best practices for sharing public sector information. [Cited 6 April 2017]. Available from: <https://www.w3.org/2013/share-psi/bp/>
- 3 Share-PSI 2.0. 2016. Guides to Implementation of the (Revised) PSI Directive. [Cited 6 April 2017]. Available from: <https://www.w3.org/2013/share-psi/lq/>.
- 4 European Commission. 2016. H2020 Programme Guidelines on FAIR data management in Horizon 2020. [Cited 6 April 2017]. Available from: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
- 5 Foster. 2016. The portal. [Cited 6 April 2017]. Available from: <https://www.fosteropenscience.eu/about>
- 6 European Commission. 2017. OpenAIRE. [Cited 6 April 2017]. Available from: <https://www.openaire.eu>
- 7 Hartmann PM, Zaki M, Feldmann N, Neely A. 2016. Capturing value from big data - a taxonomy of data-driven business models used by start-up firms. *International Journal of Operations & Production Management*, 36(10): 1382-1406. Available from: <http://www.emeraldinsight.com/doi/abs/10.1108/IJOPM-02-2014-0098>
- 8 Thomas LDW, Leiponen A. 2016. Big data commercialization. *IEEE Engineering Management Review*, 44: 74–90. Available from: <http://ieeexplore.ieee.org/document/7502386/>

- 9 Cervantes M, Meissner D. 2014. Commercialising Public Research under the Open Innovation Model: New Trends. *Foresight Russia*, 8(3): 70–81. Available from: <https://foresight-journal.hse.ru/data/2014/10/18/1100831033/5-Cervantes.pdf>
- 10 Perkmann M, Schildt H. 2015. Open data partnerships between firms and universities: The role of boundary organizations. *Research Policy*, 44(5): 1133–1143. Available from: <http://www.sciencedirect.com/science/article/pii/S0048733314002248>
- 11 Grabowski M, Minor W. 2017. Sharing Big Data. *IUCrJ*, 4: 3–4. Available from: <http://journals.iucr.org/m/issues/2017/01/00/hi0156/hi0156.pdf>
- 12 Pampel H, Dallmeier-Tiessen S. 2014. Open Research Data: From Vision to Practice. *Opening Science*: 213–24. Available from: http://book.openingscience.org/vision/open_research_data.html
- 13 Dhar S, Mazumdar S. 2014. Challenges and best practices for enterprise adoption of Big Data technologies. *2014 IEEE International Technology Management Conference*, XXV(4): 38–44. Available from: <http://jitm.ubalt.edu/XXV-4/article4.pdf>
- 14 Katal A, Wazid M, Goudar RH. 2013. Big data: Issues, challenges, tools and Good practices. *6th International Conference on Contemporary Computing*: 404–409. Available from: <https://pdfs.semanticscholar.org/34ad/09cda075101dc4ce3c04006ff804aca3ebf8.pdf>
- 15 Childs S, McLeod J, Lomas E, Cook G. 2014. Opening research data: issues and opportunities. *Records Management Journal*, 24(2): 142–162. Available from: <http://www.emeraldinsight.com/doi/abs/10.1108/RMJ-01-2014-0005>
- 16 Friesike S, Widenmayer B, Gassmann O, Schildhauer T. 2014. Opening science: towards an agenda of open science in academia and industry. *Journal of Technology Transfer*: 581–601. Available from: <http://dx.doi.org/10.1007/s10961-014-9375-6>
- 17 Eisenhardt K M. 1989. Building theories from case study research. *Academy of Management Review*, 14: 532–550.
- 18 Yin R K. 2009. *Case Study Research: Design and Methods*. 4th ed. Thousand Oaks, CA: Sage Publications.