*NOTE FOR THE READER*

*This is an abridged version of a joint research proposal by communication scholars of the University of Tampere (UTA) and computers studies scholars of the Tampere University of Technology (TUT). We submitted the proposal to the Academy of Finland in 2014. The plan was rejected though it got fairly good grading. Now we are contemplating on how to improve the proposal, and possibly develop it towards an (EU funded?) international effort.*

*If you can spare a few moments during the seminar for this, we would be glad of any comments and suggestions.*

*Ari Heinonen, UTA & Jukka Huhtamäki, TUT*

---

Research proposal (abridged version)

**Deliberation and open data. A critical study on the ecosystem and social impact of open data (DELIBDATA)**

PI: Prof. Ari Heinonen, University of Tampere, Finland; PI subproject: Adj. Prof. Ossi Nykänen, Tampere University of Technology,
Finland

## Abstract

This interdisciplinary research endeavor will look into the societal significance of open data phenomenon in the light of deliberative democracy. Based on empirical case studies in Finland with additional comparative material from Sweden and the UK, we shall 1) examine the nature of open data as a resource for democracy, 2) depict the open data ecosystem formed by data openers (public agencies), journalistic media, and open data activists, and the internal and reciprocal relationships of these social worlds, and 3) model and develop novel network analyses methods.

For the purposes of this project, we specify that the open data ecosystem can be described by studying three groups of social actors: a) public agencies that open their data and make it accessible, b) journalistic media that acquire data, and assess and publish it following the cultural conventions of journalism, and c) open data activists who as citizens not only acquire data and refine it but often also actively push public agencies to open data resources. Our project will study selected cases using a combination of conventional methods (e.g. content analyses and in-depth interviews) and experimental computational network analyses developed during the project. We shall also apply open research approach by inviting key players in the cases to assess and comment our tentative findings.

The four-year project is carried out by a research consortium consisting of communication and technical studies researchers from the University of Tampere (UTA) and Tampere University of Technology (TUT) in Finland. Essential strength of this multidisciplinary consortium is that the partners have already previously carried out successful joint research and education projects in the field.

Integral part of the project is researcher mobility for field work in Sweden, UK and USA. The team will also benefit from the input of its international network of eminent advisors from Sweden, Germany, Belgium and USA.

## Rationale

This interdisciplinary research endeavor will look into the societal significance of open data phenomenon in the light of deliberative democracy. Based on empirical case studies in Finland with additional comparative material from Sweden and the UK, we shall 1) examine the nature of open data as a resource for democracy, 2) depict the open data ecosystem formed by data openers (public agencies), journalistic media, and open data activists, and the internal and reciprocal relationships of these social worlds, and 3) develop novel network analysis methods for open data ecosystem modeling. The key objective of the project is to enhance civic empowerment by forming a comprehensive analytical model of the open data ecosystem.

Our proposal is based on the normative view that enlightened civic debate is both a requirement for and a sign of healthy democracy. Furthermore, we perceive adequate informational resources to be a prerequisite for fruitful public debate. Opening various data reservoirs to this end is important, but open data is not sufficient in itself – interested communities are needed in society to make sense of the opened data and put it into the use of the general public. Therefore, we intend to study *the ecosystem of open data* taking into account various actors in the field, their roles with regards to contributing public debates based on open data, and also their interrelationships. This systemic and contextual approach combining qualitative and quantitative methods is a novelty in deliberation research; instead of examining isolated cases and processes (as is most often the case), we pay attention to the relations of actors and settings in order to model the deliberative process of producing social discourse.

Drawing on the ideals of deliberative system (Mansbridge 2012), we perceive this type of social interaction as something where more or less loosely attached social worlds with their specific and common practices seek truth, establish mutual respect, and strive for generating inclusive, egalitarian decision-making. Obviously, information as such and public discussion do not automatically guarantee healthy democracy; actions, affects and fabricated publicity must be taken into account as well as the fact that public discourse often is strategic, purposively misleading argumentation rather than dialogue. (Fishkin 2009, Mouffe 2005) Nevertheless, we argue that all efforts towards more informed civic debate are worthwhile as they may raise the level argumentation. We intend to focus on how open data possibly could enhance informed citizenship in the spirit of deliberative democracy. In doing this, we acknowledge (but don’t succumb to) the advantages permitted by new communications technology. Our view is that making open data accessible via online channels and this way increasing the transparency of public policies, citizens can become better-informed participants of public deliberations.

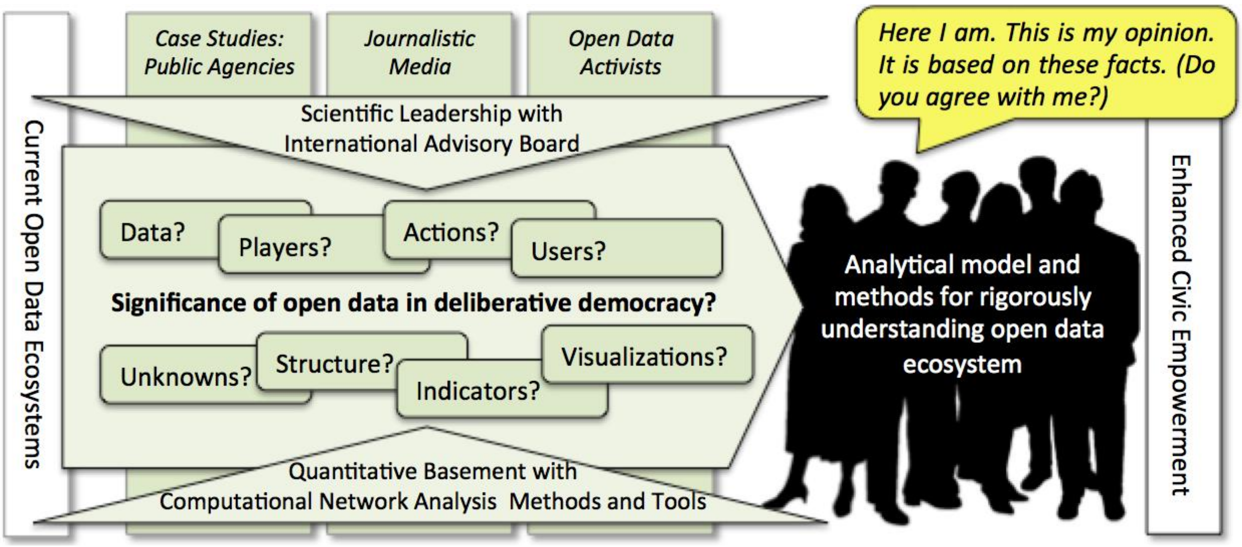
Case Studies: Public Agencies
Journalistic Media
Open Data Activists
Here I am. This is my opinion. It is based on these facts. (Do you agree with me?)
Current Open Data Ecosystems
Scientific Leadership with International Advisory Board
Data?
Players?
Actions?
Users?
Significance of open data in deliberative democracy?
Unknowns?
Structure?
Indicators?
Visualizations?
Quantitative Basement with Computational Network Analysis Methods and Tools
Analytical model and methods for rigorously understanding open data ecosystem
Enhanced Civic Empowerment

*Figure 1. Project outline: "Deliberation and open data"*

We are fully aware that opening data is not without problems. Massive and systematic surveillance and collection of user data done by social media companies and public authorities raises questions about the right to decide which (and whose) data is being collected, opened, and why. There are also gloomy forecasts of internet becoming more as a threat of democracy with its algorithm based, personalized structure (Turow 2011). Some researchers predict that the era of big data may in fact introduce ”new forms of opacity (regarding, for example, the type of information collected about users by state and private entities) even as it promises transparency” (Andrejevic 2013,10).

While acknowledging this, the DELIBDATA project provides an excellent opportunity to learn from the aforementioned challenges related to open data, and to disseminate this hands-on knowledge, in open dialogue, to scientific and stakeholder communities.

By open data we refer to both certain type of data and social dynamics attached to it. Here we make distinction between "my data", "big data" and "open data". We understand "my data" as an individual's rights to access to the data about her and to decide upon its use. Our focus is on open data, that is data which is free from all kinds of restrictions be they copyrights, patents or other kinds of control[1]. It is noteworthy that from the societal point of view this is a narrow definition ignoring that data has impact only through different parties in society that are interested in this particular type of public information resource. To study the open data ecosystem, we take a big data approach, collecting data on the actors and their activities within the open data ecosystem from various sources relevant for the project.

In accordance of our systemic approach, we identify three social worlds (Parasie & Dagiral 2012) that are involved in the process(es) of empowering public knowledge: those that produce and open data (public agencies), those that make sense and diffuse it professionally (journalistic media), and those that involve in the sense-making process on a voluntary basis (open data activists). Our project will study this ecosystem.

*Public agencies* are key actors in our study as "open government" has become an important term underlining the importance of public access to public agencies' data and information and the transparency of governmental actions. One of the origins of the recent discussion about open government is that the US and Great Britain have started the opening of governmental data with projects like www.data.gov and www.data.gov.uk. Finland has a long institutional history in the

[1] http://opendefinition.org/od/, http://opendatahandbook.org/en/what-is-open-data/.

publicity of government information, acknowledged by law since 1951. Access to government information became a policy concern in the mid-1990s. Finland joined in international Open Government Partnership (OGP) in April 2013

In addition, opening data and developing open data ecosystem are part of the Finnish government's ICT-strategy[2], which underlines the topicality of our project.

As *journalistic media* have quintessential role in creating and framing public debate in mediatized societies, they are by definition another set of key actors also in the open data ecosystem. In the ideals of journalism, serving enlightened democracy is perhaps the most cherished one, but journalism can provide services which democracy requires only when adequate machinery of record already exists by which the social world can be accurately described. The practical manifestation in journalistic media of grasping open data is data journalism, which has its roots in investigative journalism and in computer-assisted journalism (CAR) (see e.g. Ettema & Glasser 1998, Meyer 1973). While data journalism offers both new journalistic and business perspectives for the media, there is regrettably little systematic research on the topic. Some how-to guidebooks have been published (Gray et. al. 2012, Bradshaw 2013), training courses organized (for example the members of this research consortia organized together the first-ever university level data journalism course in Finland in 2011) and tools developed, but there is evident need to look into media's data journalism acts from the perspective of empowering public knowledge.

The third social world or group of actors that is relevant for our project is *open data activists*. This reflects a wider development in communications sphere; in the world of increasingly open public information news organization are no longer in sole control of news (Schudson 2010, Anderson et al. 2012). Open data activists acquire and refine data, and publish data visualizations and data-driven applications side by side with professional journalists - and often for the same audiences. Moreover, they make demands for opening data, publish link collections of data and related tools, give tips to useful applications and make tutorials for using open data resources and tools. Social media networking and events such as hackathons function as interfaces of collaboration between journalists and coders sparking ideas for activists-journalists' co-creation, which may result in intriguing open data efforts that have social impact.

The three social worlds, public agencies, journalistic media and open data activists together form *the framework for the open data ecosystem that we shall study in our project*. Our aim is to map the topography of the three actor groups within the entity of open data phenomenon, because we think there is a reason to take seriously warnings of societies being divided in those who are data/information rich and those who are data/information poor (boyd & Crawford 2012). Therefore, we shall analyze the whole ecosystem, also including the end-users of open data (i.e. the public) of each actor group. To achieve this, we intend to develop and make use of novel computational network analysis methods that are currently gaining traction within the computational social science domain[3] and which will reveal both the network structure of the open data ecosystem as well as the role positions of individual performers in each field and in the whole picture. By doing this, we anticipate that it is possible to assess the societal impact of open data phenomenon better than studying each actor group in isolation. Existing research on networks shows that network analysis has a good fit for the project: much is already known about structure of networks (Barabási and Bonabeau 2003), the roles of individual actors in the network (Hansen et al. 2011) ) as well as the role of network structure in information diffusion (Weng et al. 2013). Visual network analysis is a key method for investigating social configurations and in communicating findings to others (cf. Freeman, 2009). Computational network analysis is data-intensive and, moreover, the data is heterogeneous by nature. Therefore, the analysis process often proceeds in an iterative, incremental manner (Telea 2008). Thus, adaptive data modeling methods should be developed to support computational open data ecosystem analysis.

[2] https://wiki.julkict.fi/julkict/avoin-data

[3] http://www.iccss2015.eu/

We are agonizingly well aware that the research growing from this background setting can only draw on scarce previous research as this kind of holistic and cross-disciplinary approach has not yet been applied to open data studies. We take this deficiency in scholarly knowledge as a starting point with the aim of producing results, which not only have academic value but also contribute to social life by revealing new strands in enabling informed public deliberation, and thus enhancing positive impact of open data phenomenon.

Members of the research consortium have been involved in making similar kinds of data-driven investigations on the structure and dynamics of online networks and innovation ecosystems in micro, meso and macro levels. The Somus project (2009-2011, Academy of Finland) studied the phenomena of the self-organizing networks in the societal problem solving and open public data (Näkki et al. 2011). In the Next Media -programme UTA carried out studies about the state of open data in Finland, data journalism abroad, and team-work and data journalistic processes. In 2014-15 UTA conducts PRIANO, a project exploring how the Finns find the changing borders of privacy in the internet. In Tekes-funded innovation research projects Sindi, Reino and SPEED, TUT has studied the ways to use secondary data sources including open data, social media data and big data in general to support innovation ecosystem research. In SPEED, TUT is further looking into ways that open data allows for the creation of new business within the Digile (a SHOK programme) ecosystem. Moreover, methods and supporting tools to collect, manage, refine, curate, analyze and visualize data compiled from various sources have been developed at TUT (Huhtamäki et al. 2012, Still et al. 2013).

* * *

## 4. Research methods and material, support from research environment

As our research setting is based on the case studies of the actions performed by the three actor groups and we aim to analyze the ecosystem they form, our study calls for a repertoire of multiple and multidisciplinary. We intend to gather four types of research material for our analyses in order to be able to answer our research questions:

*1. Actor material: Players and their motives?*

Being interested in understanding what drives and motivates the three actor groups, we shall compile a body of research material using the key players of each actor group as our informants. The material will be gathered using semi-structured in-depth interviews. We anticipate that there will be 5 to 10 persons in each case, of whom some may be interviewed more than once. This qualitative data will reveal commonalities and variations of the reasonings for opening data. In the case of journalistic media, we will use various ethnographic methods such as on site observation to document the work processes of data desks and data newsrooms.

*2. Actions material: What actions? Nature of the data?*

In order to be able to assess the significance of data that is dealt with in each case, we shall apply methods of document and content analyses to each case. Guided by a common but at the same time flexible code book shall analyze a) the data that is being opened and/or refined and diffused, and b) public and internal documents related to each case. The exact research material will vary depending on the case. E.g., in a journalistic media case, a data journalistic article will be analyzed with additional information gathered from reader comments and possible debates in social media and in a public agency case, the nature of the data that is being opened will be analyzed (topic matter, amount, accessibility, etc.) but also the official documentation of the process will be taken into scrutiny.

In the case of open data activists, the nature of actions is revealed by using *group history telling method*. This method (Ryan & al. 2013) serves for self-reflection and action history documentation and creates a collaborative and participative relationship between the researchers and the activists. The aim is to publicly produce a shared understanding of a group experience, with multiple, even disagreeing perspectives of the nature of selected case(s).

*3. User material: Who are users? What is their response?*

Keeping in mind that open data requires its citizen users to become a resource for public deliberations, we intend to engage users of data in each case in our study. To locate the users we shall cooperate with the actors of the cases requesting them to submit us pointers to assumed users. Again, the users will be of different nature in each case. In a public agency case, we shall organize focus group discussion sessions with the help of the public agency in question. In a data journalistic case, we will be able to reach readers with the help of the newsroom. In a case of activists, similar approach will be used. In these cases, we shall use guided (moderated) online discussion as a method.

*4. Network material: How to access and computationally analyze data?*

Accessing data on the actors, their actions and users of open data ecosystem requires interfacing with various technical platforms and communication tools, and aligning and interpreting the resulting dataset with common data models. This results into preliminary understanding of the underlying information, which needs to be iteratively refined both in terms of the technical infrastructure and semantic explanation of the data, with respect to the application and analysis needs. We acknowledge the evolution of the technical platforms, real-world processes, models, scientific information needs and concrete application requirements as the basic characteristics of our work.

These four bodies of research material will be brought as a common property of the entire research group as *working reports*. The reports will be presented, discussed and evaluated in our mid-term seminar with our international advisers. The reports are also excellent raw material for our researchers to write articles for academic journals.

The reports from the case studies are a corpus, which form the basis for our *secondary analyses* in which we shall look into the results through the prism of our main research questions. If necessary, we shall then gather additional material. The material from the research work carried out in Sweden and the UK will be incorporated into our material reservoir at this point allowing us to juxtapose the findings in Finland to those in other countries. The secondary analyses aim at verifying our results of open data ecosystem. We shall use participatory methods and invite all interested actors to publicly discuss and comment the results of the four focus areas. When possible, the research data will be opened for browsing, commenting, and re-use, after taking care of removing all data that could be used for identifying the respondents. Thus, we shall follow the principles of open research currently under further development in Finland[4].

The team is in an excellent position to enjoy of interaction with research communities involved in related studies. At UTA, journalism and media research has a long and sound tradition, and at TUT the work done in the field of computational ecosystem analytics is valuable for us. What is important for this project is that the team has direct contacts with both newsrooms and open data activists. Internationally, the project will benefit from the input of our scholarly network that serves as a critical evaluator of our work at different stages of the project. (See Collaboration in chapter 7.)

TUT will be utilizing the Tampere Center for Scientific Computing (TCSC) computing cluster for computing-intensive processing and analysis of the data representing open data ecosystem collected from different sources. Access to TCSC gives the project a chance to scale up to terabytes data when needed, e.g. when analyzing the Wikipedia corpus. The research material will be at first stored by the team in secure electronic and analogue repositories, and after the

[4] http://avoinsuomi2014.fi/sites/default/files/Riitta_Maijala_Avoin Suomi 2014_09092014.pptx

project, suitable material will be handed over to be stored at the Finnish Data Archive of Social Sciences (FSD) for further use.