End of AY 2016 Report for SIP- Group3

Project Title

Knowledge and Data Management for Policy Making Towards Data Driven Society
データ駆動型社会に向けた政策立案のための知識とデータのマネジメント

Team

GSDM ID	Name	School	Department	Year (e.g. D1)	Leader/ member
14108	Koki Muraoka	Engineering	Chemical System Engineering	D1	Leader
14205	Yuta Yamauchi	Economics	Statistics	D1	Leader
14102	Teruaki Hayashi	Engineering	Systems Innovation	D3	Member
13203	Naoki Nonaka	Engineering	Technology Management for Innovation	D3	Member
13107	Masaaki Imaizumi	Economics	Statistics	D3	Member
14125	Seonwoo Kim	Engineering	Mechanical Engineering	D1	Member
16115	Kensaku Matsunami	Engineering	Chemical System Engineering	M1	Member
16118	Lee Hee Woon	Public Policy	International Public Policy	M1	Member

Objective: Explain what social/global issues that this project tried to address and why the issue is important.

The potential expectation on data-driven society to the extent that generating innovative businesses and creating values by both combining and exchanging data among different domains has been increased. However, even the information about data (e.g., where the data is stored, who the data owners are) and knowledge of data utilization (e.g., what kind of social issues can be solved by the data) are not fully shared. Although Open Data is one of the movements for encouraging data utilization, data driven society may not be achieved only by publishing data on the Web. In order to support decision making, e.g., policy making, by using appropriate data and tools, the management of knowledge and data is attracting social attentions.

In our SIP, we provide the platform (workshops and Web applications) for data owners, users, analysts, and those who want to obtain data in the future, to meet, discuss and share their resources and knowledge. We collect and store not only the information about data, but also the information about tools, skills, and ideas for data processing and utilization. We plan to externalize the social issues through the communication among participants in various domains in GSDM, and solve them by exchanging our skills of data analyses. As the first step, we are tackling on REF data (the social report of Research Excellence Framework of UK cases in 2014) based on the natural language processing obtained in SIP of

Big Data Analysis last year, together with bibliometric approach, where collaborative analysis with experts of policy making is expected.

Method: Explain through what kind of approaches you tried to achieve the objective.

In our SIP, we provide the platform (workshops and Web applications) for data owners, users, analysts, and those who want to obtain data in the future, to meet, discuss and share their resources and knowledge. In our SIP, we plan to externalize the social issues through the communication among participants in various domains in GSDM and solve them by exchanging our skills of data analyses.

Our method takes the following steps.

1. We collect the information about data, tools, ideas and knowledge from GSDM students. The entry form is already available online. Based on the stored information, we provide the platform for sharing knowledge, skills, and datasets for data utilization.
2. We conduct workshops for externalizing hypotheses of social issues and analysis plans introducing the workshop methods (Innovators Marketplace on Data Jackets and Action Planning) produced by Hayashi and Ohsawa lab.
3. Presenting our procedures in conferences.

Outcome: Explain what kind of results you obtained from this project and discuss how it addressed your focal social/global issues.

Our outcome of each step is as follows:

1. Collect the information about data

Based on the results of the previous SIP of Big Data Analysis, we conduct the analysis through the discussion with experts of policy making. Figures in Appendix show the results using REF data.

2. Conduct workshops

We hold workshops for externalizing hypotheses of social issues and analysis plans introducing the workshop methods (Innovators Marketplace on Data Jackets) produced by Ohsawa lab (Hayashi arranged the workshops).

3. Present our progress

We presented the result of our data analysis in several domestic conferences. Details are in Appendix.

Budget: List the budget this project implemented. *About the details, add the appendix.

Purposes	**Expense**
Books	0
Travel fee	208,720
Honorarium	0
Others	25,000
Total	233,720

Appendix

This section contains the details of the activities of SIP.
1. EIC research society
2. Market of Data Research society
3. Workshop in IEL
4. REF data analysis
5. Discussion of Environmental Improvement of Data Platform
6. Details of budget

1. EIC research society:

SIP leader Yamauchi and member Lee participated at Domestic Conference: 2016 EIC (The Institute of Electronics, Information and Communication Engineers, 原:2016 年電子情報通信ソサイエティ大会) which was held in Hokkaido University, Japan. 20th of September, 2016, On the first day of the conference, SIP presented the concept and the progress of the research in lecture session "B-18. Intelligent Environments & Sensor Networks (原: B-18. 知的環境とセンサネットワーク)" with the lecture title " Review on comprehensive evaluation methodology using Bibliometric Information(原:書誌情報を用いた研究インパクトの包括的評価法の検討)." The moderator of the lecture session, Professor Mitsugi from Keio University, gave questions to the contents of the presentation to clarify the research methodology and research development plan afterward. On the second day of the conference, SIP participants attended the lecture session "B-7. Information Network(原: B-7.　情報ネットワーク)" to seek new ideas and fellow research to improve SIP's research. Unfortunately, all of the delivered lectures in the session were unlikely related to SIP's research.

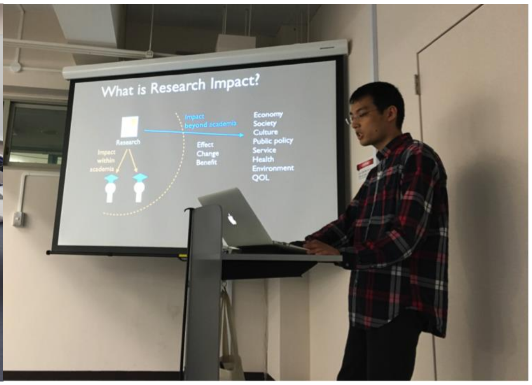
2. Market of Data research society:

SIP Leader Muraoka and member Hayashi and Nonaka participated at Domestic Conference: **IEICE: The Inst. of Electronics, Info. And Communi. Engineers (2016) Tech. The committee** which was held in Shujitsu University, Okayama, Japan. 18th of February, 2017, SIP team presented the outcome and the plan of the research based on SIP activity in AY2016 in lecture session with the lecture title " Evaluation of Multidisciplinary Research Impact(原題:多分野にわたる研究インパクトの評価法の検討)." We were asked about the potential use of other data sources such as authorship and contents of commercial books. The moderator of the lecture session, Professor Osawa from University of Tokyo, commented about the potential side effect of introducing new evaluation metrics to academic research and we discussed possible solutions to this risk.

3. Workshop in IEL:

We held our IEL on 22nd July as a part of our SIP activity. 35 participants joined our IEL in total (17 students from GSDM, three faculty members of GSDM, and 15 external audiences), which was one of the biggest IELs in GSDM. We realized the participants' strong interests about multidisciplinary data utilization and data-driven society. In the first half of this IEL, we presented the result of the on-going data-driven project, which tackles attractive data of research impacts, assessing the quality of studies in the UK higher education institutions (REF data). We reported how classical bibliometric analysis provides unfair evaluation weighing heavily on specific fields. We also suggested that research impact should be fairly evaluated in different ways such as our new framework using REF data. Although some audiences were unfamiliar with this topic, we used a variety of figures to illustrate how to evaluate research impact using data. Through this IEL, we expect that participants understood both the result of this project and the importance of data-driven society.

The second half of the IEL, we conducted a gamified workshop called Innovators Marketplace on Data Jackets (IMDJ) for discussing the social issues and possible combinations of knowledge elements (data, tools, knowledge, skills, and ideas). Some participants submitted us the information about their interests as knowledge elements in advance. We summarized and visualized their information as a scenario map. We set the main theme of this workshop as "Data Utilization for Policy Making and Creation of Business toward Data-driven Society." Participants stated their requirements and created solutions using the elements visualized on the map for about 1 hour. Finally, we got 76 requirements, 49 solutions, and 20 additional information. These outputs are expected to help activate data-driven society. In the future, we are going to create analysis scenarios and conduct actual data analysis.



What is Research Impact?

Research

Impact within academia

Impact beyond academia

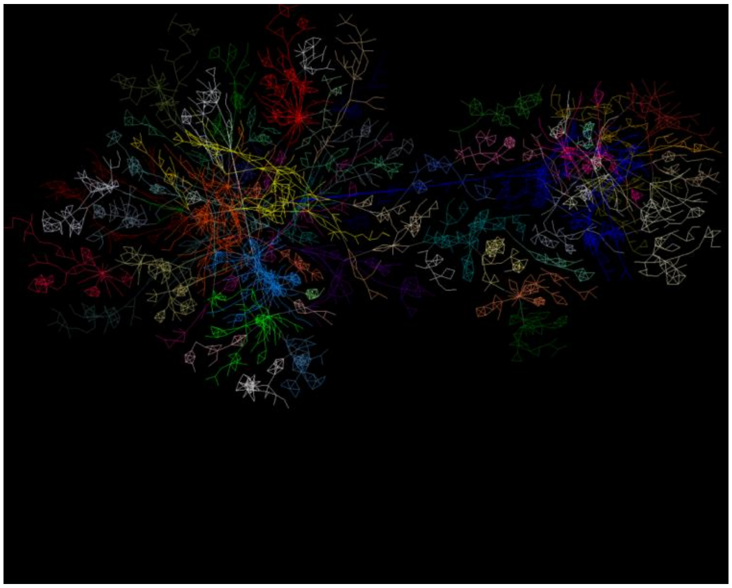
Effect
Change
Benefit

Economy
Society
Culture
Public policy
Service
Health
Environment
QOL

4. REF data analysis:

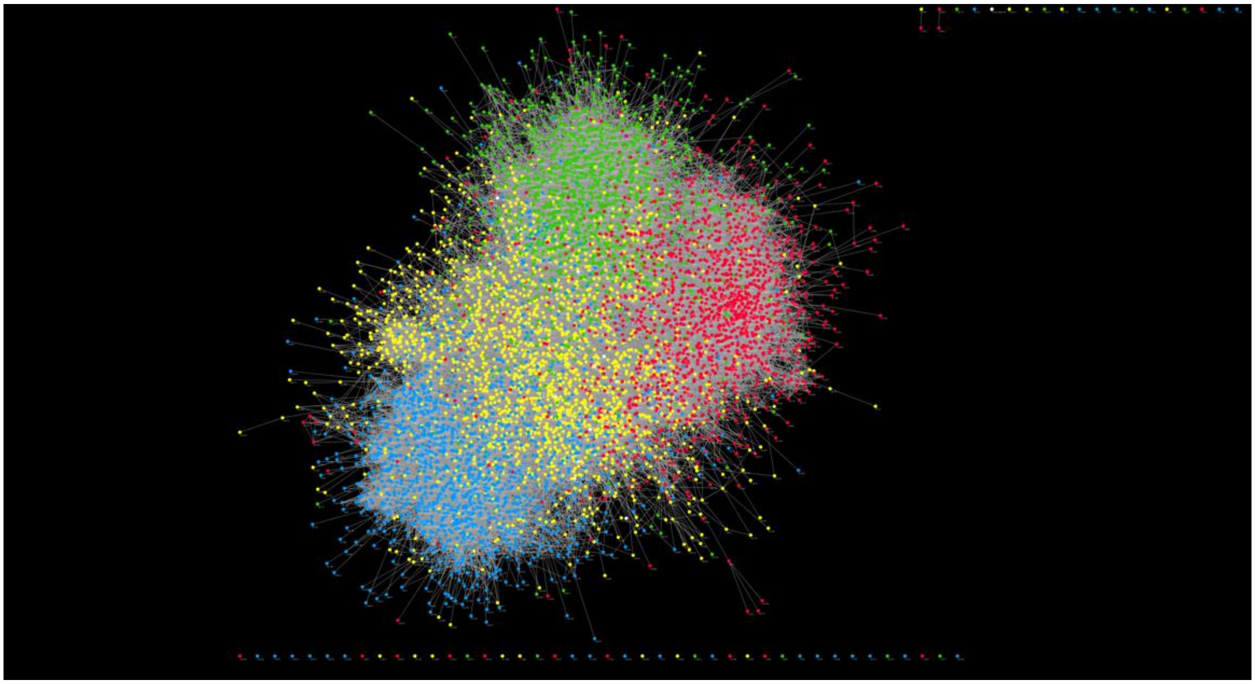
Recently, accurate evaluation of the “impact” caused by the academic studies is required to understand the effects the academic activities give to the society and distribute the budget for studies efficiently. Bibliometric approaches have been proposed and utilized conventionally to survey the impact within the academic society as represented by the counting the number of citations, but currently, broader and more general metrics to describe the impact "beyond academia" that considers social and economic impact are required. Several new descriptors or investigations have been reported to cope this issue, including REF (Research Excellence Framework) data. UK government performed a series of studies at 154 research institutes and collected self-evaluations of 6640 research impacts described in natural language by researchers themselves. The data covers multiple disciplines including physics, medicine, computer science, law, history, and music. We surmised that this data merits further investigation and analyses to find a clue to evaluate the impact of studies that is valid for multiple fields, which is conceived to be difficult to elucidate. You might be thinking that simply impact factor can be the metric to measure the research impact. The definition of impact factor for a journal is the number of citations per number of articles during two years. It simply means that more cited journal has more impact. Sometimes this impact factor is useful, but its fairness is not well established especially considering the studies from different disciplines. We have tackled on REF data based on the natural language processing obtained in SIP of Big Data Analysis last year, together with the bibliometric approach, where collaborative analysis with experts of policy making is expected.

Even if these challenges a lot of methods to evaluate research impact have been proposed. The impact factor is one of them. Other methods including h-index, Excellence in Research, Star Metrics, are classified as traditional bibliometrics approaches. They use the peer-review system as represented by impact factor. Among these bibliometrics methods, an academic landscape system is a state-of-the-art tool using citation network analysis, which is development in this university. Imagine that there are three papers and one of them cites the others. As recognizing that an article as a node and citation relationship as a link, we can represent this bibliographic information as a network. Once the network is constructed, we can apply many algorithms to characterize this network, including clustering to understand the nature of the overview of the academic discipline. The other stream for evaluation of research impact is relatively new methods. Considering commercial income is one of them, but it captures just one aspect of research impact. Payback framework is the most widely accepted way. It considers both academic output and wider societal benefits. Dutch project introduced SIAMPI while Australia proposed RQF, respectively, to capture research impact by different methods. REF is another system by the UK.



As the first step of the analysis, we performed citation network analysis, which is a state-or-art bibliometric approach. For the purpose, we developed a program to obtain whole REF data available online. The data is distributed as HTML page. By using well-established web scribing techniques, we could obtain the information of research papers or proceedings cited in the underpinning research of the research impact. This information can be extracted as so-called DOI, which is a unique code for each research paper. We submitted this DOI code to the web of science, which is the world's largest database for research papers. Base on the DOI data, Web of Science returns bibliometric data including citation relationship. In contains the information of other papers that are cited by the focused paper. This procedure is repeated for the other literature too, and finally, I obtained more than 17 thousand citation information. They were combined and submitted to the academic landscape system to construct citation network. Once the network is constructed, a well-known clustering algorithm can detect groups in the network and do the partitioning, to visualize this colorful citation network like a firework. As a result, we could obtain the citation network of REF data. Each line represents the citation relationship, and each dot is research paper or proceedings. The colors represent the belonging clusters. This citation network of REF data contained more than 3 thousand papers. It means that other 10 thousand papers did not have citation relationship to any of these papers in this network. There were 5 thousand links, which is relatively sparse compared with conventional citation network. And 60 clusters were detected by modularity maximization algorithm. Then, we moved on to the characterization of clusters. The bibliometric data we obtained from the web of science database contained abstracts. By applying text mining techniques, we can extract words that are characteristics of the cluster. By plotting term frequency versus inverse document frequency of the term appeared in the abstracts, we can understand the characteristic terms of the cluster. Because the term frequency represents frequency, and inverse document frequency represents uniqueness, the term appears in the top-right is the important term for the cluster. By looking at the same plot for each cluster, we can put the name for each cluster. The largest cluster was Geriatric disease judging from the term like dementia, and knee pain. It is interesting that the care for old people is the largest cluster among all different research topics. It might reflect the population age distribution. The keywords for the 2nd largest cluster is diabetes, cholesterol and so on. The topic of this cluster is also related to clinical medicine, especially adult disease. Similarly, cluster #3 contains keywords like depression, disorder, anxiety. We concluded that this cluster focuses on mental diseases. Cluster #4 is a little bit different. The keywords were child, language, intervention, parent, and disorder. It seems that this cluster signifies elementary education and disorder. The red cluster in the Figure shown above is the cluster #1, which is the geriatric disease. #2, the yellow one is related to the adult disease. #3, the orange cluster, was the mental disease. And the green one is elementary education and disorder. Surprisingly, top 4 largest clusters somewhat related to medicine. The keywords for cluster #5 were climate change, emission, and uncertainty. Clearly, the name of cluster #5 should be climate change. Cluster #6 is again related to medicine. It would be a genetic disease. Cluster #7 is similar to cluster #5 but more focusing on global warming and resulting sea level. The keyword of cluster #8 is less intuitive, but judging from the intake, salt, vegetable, food, and obesity, the topic would be dietetics. Cluster #9 is again genetic disease, but a little more biology side compared to the previous clinical science one. Cluster #10 lacks the term in the top right, but the topic is typical medical one, especially men. So it must be the disease for men. Cluster #11 contains the different keyword, which is ecosystem service. This concept understands that ecosystem is beneficial to service for people, like cleaning drinking water and decomposition of wastes. And Cluster #12 is clearly smoking cessation. More that 60 percent of the clusters was classified into clinical medicine. 2nd most popular were biological science and environmental science follows. Next was psychology, and a small number of clusters was observed fro education, chemistry, economics, and law. In this citation network, there was no cluster that can be classified as physics, math, computer science, and the others. It is surprising that even if REF data collected all research impact from those research disciplines without bias, but only limited number of discipline could form the citation network. If you naively accept this result, the conclusion would be life science has more impact than environmental science, and economics follows, and its impact is equal to Law. And research impact of physics, history, linguistics is so low that no one cannot detect it. That’s why we have to think that there is a limitation in bibliometric approach. Simply collecting research papers from researchers and evaluating it always has the risk of overestimation or underestimation depending on the research discipline. Let’s go back to the original research question we would like to compare the bibliometrics approach and questionnaire-based approach. So far we have seen citation network analysis-based approach. Next, we went the approach using text mining of questionnaire data.

We extracted the text written in “Underpinning research” section that is essentially the same to “Reference to the research” while the former is written in natural language by the researchers themselves. A standard technique of text mining was applied to compute the similarity between two texts. If two impacts are enough close regarding this metric, we connected two impacts. The resulting network was fully relaxed by spring algorithm, and colored based on the node’s original classification done by REF editors (Green: Environment, Chemistry, Physics, Math, Information, Aero, Civil Engineering; Yellow: Law, Economics,Sociology, Education; Blue: Literature, History, Philosophy, Art, Music; Red: Life science). The obtained network shown below showed a beautiful clustering regarding the original discipline, though the information was not explicitly given.



When you look at the boundary between clusters, you can see several cross-boundary types of research like "The psychology of prayer" suggesting that the method employed correctly performed clustering capturing boundary researches. The most connected impact was the research about immunosuppression that can be categorized as medicine or life science which accords with the observation seen in the previous bibliometric analyses. The center of the analysis having highest betweenness reported platinum in a crater in Greenland that worth million-billion dollars, which is conceived to be a tremendous impact in economic viewpoint. The similar analyses were performed on the rest of REF data, and we are trying to correlate these networks obtained and get an insight of the relation between the world observed in citation network and the real world.

5. Discussion of Environmental Improvement of Data Platform:

We visited the governmental meeting about Environmental Improvement of Data Distribution in Japan (データ流通環境整備検討会) organized by Prime Minister of Japan and His Cabinet. We unofficially showed the chairperson our research activities about the methods and approach for data-driven society. We kept contact with them and exchange information. Followings are the meetings.

meeting	date
the 1st meeting	2016/09/30
the 2nd meeting	2016/10/14
the 3rd meeting	2016/10/28
the 4th meeting	2016/11/11
the 5th meeting	2016/11/25

the 6th meeting	2016/12/09
the 7th meeting	2016/12/22
the 8th meeting	2017/01/27
the final meeting	2017/02/24

6. Details of Budget

Purposes	Expense
Conference fee (IEICE)	25,000
Travel fee (Hokkaido)	118,360
Travel fee (Okayama)	90,360
Total	233,720