Lin, Chao Chen

**Conference Paper**

# Where is the limit of big data? A case study of journalism practices pertaining to datasets of e-Government in Taiwan

20th Biennial Conference of the International Telecommunications Society (ITS): "The Net and the Internet - Emerging Markets and Policies" , Rio de Janeiro, Brazil, 30th-03rd December, 2014

**Provided in Cooperation with:**
International Telecommunications Society (ITS)

*Suggested Citation:* Lin, Chao Chen (2014) : Where is the limit of big data? A case study of journalism practices pertaining to datasets of e-Government in Taiwan, 20th Biennial Conference of the International Telecommunications Society (ITS): "The Net and the Internet - Emerging Markets and Policies" , Rio de Janeiro, Brazil, 30th-03rd December, 2014, International Telecommunications Society (ITS), Calgary

This Version is available at:
https://hdl.handle.net/10419/106854

**Where is the Limit of Big Data?: A Case Study of Journalism Practices Pertaining to Datasets of E-Government in Taiwan**

## I. Introduction

2010, computer scientist, the inventor of the World Wide Web Tim Berners-Lee speaking at the launch of the first government datasets for spending by departments of more than £25,000, he was asked who will analyze them once the geeks have moved on. What's the point? Who's really going to hold government, or anyone else, accountable?

"The responsibility needs to be with the press," Berners-Lee responded firmly. "Journalists need to be data-savvy. He reckoned the future of journalism lies in analyzing data in his speech. In his view, journalists could use software to discover the story lurking in datasets released by governments, local authorities, agencies, or any combination of them – even across national borders. "Data-driven journalism is the future," Berners-Lee insisted( Arthur, 2010).

Berners-Lee has been working to get the civil service and local government to open up their data. Berners-Lee and Professor Nigel Shadbolt are the two key figures behind data.gov.uk, a UK Government project to open up almost all data acquired for official purposes for free re-use. If we want to realize Berners-Lee 's argument, we need to understand the connection between e-Government releasing data and the emerging data- driven journalism.

There is a revolution in the provision of e-Government services to citizen during the past decade. The term e-Government(electric government) is used to describe the legacy of any kind of use of information and communication technology within the public sector. Early applications were focused on building management information systems for planning and monitoring( Bhatnagar, 2004). The practice of this decade in 21th is about big data of government. E-Government is considered as a guiding vision toward modern administration, democracy and transparent for government to public dealing. E-Government could collect, store and make data available online and machine readable. Many developed countries like United States and United Kingdom have invested resources to advance technology and infrastructure to look for a digital future. It means that e-Government is not an option but a must for governments to realize (Baqir and Iyer, 2010).

In order to address the points mentioned above, the developed countries adopted different ICT system and made a technical standard for data format, up to specifying

the procedures, organizational structures, and the systems to maintain data over the years and to foster interoperability among public administrations. People could use both computers and traditional ways to analyze big data and catch the meanings of the content(Lewis, Zamith,&Hermida,2013). It is important that e-Government must make data available to all citizens not just to a minority who could afford to have access to e-service. Even every citizen has the same rights on data, journalists will be the people who concern data much more than ordinary people. That's also the reason why Berners-Lee encourages the press should equip themselves with the tools to analyze data from government(Arthur, 2010).

Based on the data , there are excellent computer-assisted reports in journalism. Tauberer (2012) analyzed the 2002 winner of the Pulitzer Prize for Investigative Reporting, he found the series could not have been told without access to government records *at scale*. The so called computer-assisted reporting has been around for decades, but when it's restricted to newspaper format, it can't realize its full potential. On the web, it can sing with depth, customization, search ability and a long shelf life(Briggs, 2010:255).

During these decades, there was a traditional aspect to the ownership conflict, the what-I –own-you-can't-have notion, also newly applied to information. The issue of who owns electric data recorded in the act of governance had been going on, since the Watergate tapes had been erased in America. The federal district court held that computer records could constitute "public records" and be subject to Federal Records Act and the Freedom of Information Act(FOIA) (Branscomb, 1994: 159). Sharing data is an important aspect. News organization also opened up their contents through the use of application program interface(API). The New York Times, the BBC, NPR and the Guardian have all announced APIs for their data. This means other grammars and organizations can mash up data and news stories for use on their Web sites(Briggs, 2010:261). Western news media companies have developed data journalism to a point where they are able to process news by extensively using digital technology (Thornburg, 2011) and transform large amounts of boring data into interesting news stories. Data exist everywhere. The issue about the usages of digital data is important in popular social media and mass media. What is important here is to note that the value of Facebook is in its data. The so called "data play" is also something that is coming to define contemporary journalism (Beer, 2013). People could realize big issues from big data(Neuman, Guggenheim, & Bae,2014). The emergence of data journalism is opportunities abound for using databases, spreadsheets and other forms of structured or fielded data in news coverage or story development(Briggs, 2010).

The impact of data journalism has also garnered considerable interest in Taiwan. Comparing the obligation to release information between the Taiwanese and the US

governments, the Freedom of Information Act has been in effect in the United States since 1966, whereas a similar law, the *Freedom of Government Information Law*, was passed in Taiwan in 2005. That is, Taiwanese government has the responsibilities to release the data, and whether government-released data are credible sources for all the citizens.

Taiwan is a free and democratic country in Asia. All the media are deregulated 20 years ago. The development of ICT in Taiwan is advanced, too. Even so, we are regretted that the competition of media is heavy and make news outlets sensational. On one hand, the news manufactured by the media companies in Taiwan are cheap and little media would invest and develop data journalism. This reminds me to think what McChesney (1999) described "rich media, poor democracy" in America. The scenario of media is "poor media, poor democracy" in Taiwan. On the other hand, the development of e-Government is a key reason for journalists to make good stories from big data. After all, it is not easy to require all the people to be e-citizens or digital citizens. The problem of digital divide does exist. Under this condition, we need e-Government could induce big data, then, big data could induce data journalism. That is, the quality of data of E-government is important for journalists to use electronic service and find stories. A key factor causing these problems is the data. This study is in the attempt to understand the information distribution conditions of the Taiwanese government and proposes suggestions to enhance the quality of data in order to improve the development of data journalism in Taiwan.

## II. Conceptual Background

### *Big Data and E-Government*

According to International Data Corporation(IDC), approximately 90% of the digital data we encounter today didn't exist two years ago. It is important to understand that big data is not only about the original content stored or being consumed but also about the information around its consumption (Gantz & Reinsel, 2011). "Big data" is a term associated with massive datasets. Whether the data is used to explain any issue, big data clearly has a significant role to play(Malik, 2013). Even people all agree "big data" is a big job to do without hesitation. It is much more important for us to understand the pathologies of big data.

Big Data is notable not because of its size, but because of its relationality to other data. Due to efforts to mine and aggregate data, big Data is fundamentally networked. Its value comes from the patterns that can be derived by making connections between

pieces of data, about an individual, about individuals in relation to others, about groups of people, or simply about the structure of information itself (Boyd & Crawford, 2011). At present, “big data” becomes a commonly seen term. Big data concerns large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, big data is now rapidly expanding in all science and engineering domains ( Wu, Zhu, Wu, & Ding, 2014).

Jacobs(2009)argued that the pathologies of big data are primarily those of analysis. In the digital time, e-Government services are becoming an important part of everyday life for citizens, businesses, and government administration themselves. Such services are numerous and rely on complex infrastructures. The e-Government architectures had several advantages to help maintaining a good level of data base consistency. It also was a good solution to keep a consistent architecture among different applications while also allowing easy information exchanges between databases. However, when IT departments were seen as centralized, changing anything became a complicated issue(Toporkoff, Rannou,Soufron, & levy, 2008:11-12). Toporkoff et al. argued that the 2007 study on e-Government architectures focused on the experiences of the U.K., Germany, France, the U.S. and Hong Kong. Misra(2012)has witnessed a prolific advancement in Indian over the years. It means the policy of e-government is useful for individual rights. After that, citizens could press for information to be made available in a form compatible with the computers.

Some scholars argued critically is the increasing use of ICT in support of surveillance and intelligence gathering in a range of policing methodologies that impact on the daily lives (Hayes, 2009). Sophisticated devices and techniques have greatly enhanced the capacity of government to intrude into the lives of citizens. Many of the new forms of surveillance are well suited to the networked society. Technology now allows the compilation, storage, matching, analysis, and dissemination of personal data at high speed at low cost (Fox,2001). Some policing institute might do this through the information technologies to gather large amounts of data often without their knowledge or permission( Aspland, 2012).

In general speaking, e-Government initiatives worldwide have gained momentum. Government agencies globally are trying to serve citizens through web interfaces. However, this endeavor comes with challenges of its own(Arif, 2012). Arif tried to demonstrate how constructs of customer orientation in e-Government, and the last, demonstrated how the customer orientation concept might be used to suggest improvements to e-government project management. Misra (2012: 62) hypothesized that services supplied would meet their desired level of success through sustained use

by end users. According to Aspland (2012), Misra (2012) and Molinari, et al., (2012), " citizen-centric" is a core of e-Government. " Citizen-centric" meant to denote any approach related with or focused to the co-creative collaboration with the forthcoming users of products and/or services under development. It is especially useful in the field of e-service and of ICT based or supported service(Molinari, et al., 2012:159). For example, the Scottish government has divided the users into general public, blogger/journalists, data analyst, front-end web developer, app developer, web application developer, data scientist. The Canadian government showcases citizen satisfaction as progress(Roy, 2007).To maximise value of data, it is important to serve the needs of all of these users.[1] The big data of e-Government is talking about the same thing. There is little doubt that the quantities of data are indeed large.

For the reason to emphasize the open data of e-Government, some scholars also use the term "open government" to replace e-Government. Open government data is the Big Data concept applied to open government. Open government data differs from "information" or "knowledge", could grant the public access to these sorts of government records, often on paper or an electronic equivalent (Tauberer, 2012). Tauberer believed that there are some mediators to "hack" open data from the government. As his idea, hacking is a good word. What he called "Civic hacking" means using government data in particular has implications far beyond our experience with government. Civic hackers can be programmers, designers, data scientists, good communicators, civic organizers, entrepreneurs, government employees and anyone willing to get his or her hands dirty solving problems (Tauberer. 2014). Journalists act just as mediators.

## *Big Data and Journalism*

In the networked environment wherein almost all texts link to others, and where openness and transparency have become more strongly related. The emerging practice of data journalism serves a related need (O'Sullivan, 2012). Tauberer(2012) argued the value of those government records came from reporters' skills in turning the records. That is what this paper called "data journalism". Even more, the relationship between big data and journalism is clearly becoming closer, and that big data is closely linked to democratic ideals(Lazar, 2012). Data journalism focuses on processing the background data of news reports, and is believed to be a profession that improves the quality of news. Julian Assange, the founder of WikiLeaks, therefore referred to data journalism as "scientific journalism" (Gray, Bounegru, & Chambers, 2012: 22).

The definition of "data journalism" remains unclear. The so called "data

[1] The website is: http://www.scotland.gov.uk/Publications/2013/12/6550/6

journalism" is journalism done with data(Gray, Bounegru, & Chambers, 2012). *The Data Journalism Handbook* (Gray et al., 2012) is a guide for journalists to find ways of telling stories out of the new forms of digital data that are available. It notes the changing nature of journalism, as journalists are faced with new opportunities to use data, thus requiring new sets of skills and knowledge(Beer, 2013).Data journalism is a new form of news presentation where original "data" are converted into "information." Data, in this context, are characterized as those that can be interpreted by computers and digitally stored as numbers, whereas "information" consists of understandable messages obtained from the processing and consolidation of data through software (Wigand, Shipley, & Shipley, 1984). Egawhary and O'Murchu (2012) argued that data journalism is a profession related to numbers and the ability to analyze and examine numbers. Through data journalism, people are able to manage and correctly interpret large quantities of numbers.

At the same time, the adoption of ICT in developing countries presents particular difficulties and opportunities (Shaviko, Villafiorita, Zorerthe, Chemane & Fumo, 2010). Many public sector strategies have acknowledged the strategic value of e-technology. Technology is undoubtedly the backbone of the infrastructure that is required to support electronic government initiatives. Yet there is a danger in placing too much emphasis on the technology aspect of e-service(Asgarkhani, 2012). Comparing to last century, there are more free and functional software to analyze big data. The ICT technology is not the most relevant characteristic of this new data ecosystem. The substance of issue is about big data itself.

The emergence of data journalism brings a big challenge for the education of journalism. Originally, ordinary software could not be employed to process big data. However, following technological advancements, the boundary between using ordinary software and robust computers to process big data has gradually blurred (Manovich, 2011); 5 to 10 years ago, large amounts of information could only be processed by analyzing survey reports , and the majority of journalists and editors based their articles on external data sources. Composing news reports with self-analyzed data was extremely difficult because such endeavors required more advanced computer technology. By contrast, contemporary data processing conditions are entirely different. The prevalence of the Internet has increased the accessibility of free information and tools to process large amounts of data and numbers, and various web-based applications are available which allow people to share and verify information (European Journalism Center, 2010).

The prevalence of data journalism is particularly apparent in the field of print media, which attempts to identify news opportunities from immense data sources such as the Internet. Simon Rogers (2013:10–11), formerly an editor responsible for the

Datablog of *The Guardian*, an English newspaper proposed that data journalism has become the standard of the journalism industry in the past few years, and is the primary method by which *The Guardian* reports its stories.

In the digital time, journalists do try to find a new way to conduct news outlets. This is why journalists should see data as an opportunity to make the world different. By using data, the jobs of journalists shifts its main focus from being the first ones to report to being the ones telling us what a certain development might actually mean(Gray, Bounegru, & Chambers, 2012: 3-4). Data journalism encourages students, journalists, citizen journalists, data analyst and anyone who are engaged in data mining, not only in the media. It is optimistic for us to believe the dominant of public relation will weaker than before.

Data journalism could help a journalist tell a complex story through engaging infographics. It is popular that we use information visualization tools to explore data, generate, refine and test hypothesis(Pousman, Stasko & Mateas, 2007). It means we need to understand a science of data visualization. Data visualization is the sheer quantity of information that can be rapidly interpreted (Ware,2000).Spence(2001)said there are many situations in which data is available, sometimes in very large quantities, and where some human insight into that data is required. Shapiro(2010)stated that many of the talents required for creating good information visualization are widely recognized, there is commonly overlooked in more formal settings. The talent is the art of storytelling. Krebs(2010) believed that finding complex patterns in data and making them visible for further interpretation utilizes the power of computers, along with the power of the human mind. His job is to explore some datasets that reveal interesting insights behind them.

Ware(Ware, 2000: 28) also argued that if the goal of visualization research is to transform data into a perceptually efficient visual format, we must be able to say something about the data that can be exist for us to visualize. Unfortunately, the classification of data is a big issue to do. It is a new craft for the journalists to reveal the information or stories from visualization behind the data. The journalists need to understand more things than the quantity of number such as the context of the data, the meaning from the data….. That is the same professionalism for the journalists to learn about storytelling in the traditional journalism.

## III. Methods and Research Questions

The objectives of this paper contend the following 2 aspects are essential for laying a foundation for the prevalence of data journalism in Taiwan: (1) the publicizing of

government data in accordance with relevant laws, and (2) the data format available for computer software analysis of journalists. Advancements in data journalism can only be achieved when these 2 aspects are satisfied. In the initial stage of research, I employed the 5 star deployment scheme proposed by Tim Berners-Lee (http://5stardata.info/) to comprehensively understand the willingness of the Taiwan government to release data. More stars means better quality of the open data. The criteria of 5 stars to judge the quality of data is below :

| | |
|---|---|
| ★ | make your stuff available on the Web ((e.g., whatever format and PDF) |
| ★★ | make it available as structured data (e.g., Excel, SHP, WMS) |
| ★★★ | use non-proprietary formats (e.g., CSV, JSON、KMZ、KML) |
| ★★★★ | use URIs to denote things, so that people can point at your stuff |
| ★★★★★ | link your data to other data to provide context |

The present study employed 2 research methods, namely a quantitative analysis method and a qualitative in-depth interview method, to evaluate the quality of big data from e-Government in Taiwan. Four graduate students were employed to conduct statistical analysis, in which the disclosed data were classified using a five-star scoring system. According to Tim Berners-Lee, the criteria from 1 to 5 stars，the present study examined 153 central and local governmental units. Excluding the Office of the President and the five government branches, 41 primary units of the five Yuan and 21 primary units of local governments were examined. A total of 81 secondary central and local governmental units were also included in the study scope. According to The Freedom of Government Information Law, the official websites of all 153 government institutions are required to include a governmental information disclosure section. The quality of the disclosed information listed on the websites of the 153 governmental institutions was evaluated and analyzed. This data includes the Office of the President (N=1)and the five Yuan(N=5)，Main central government(N=41), branches of the 41central government(N=85), local government(N=21). This paper will emphasize on the platforms of main central government's websites for the representatives of all branches of central governments in Taiwan.

In the second stage of research, I conducted in-depth interviews with numerous

subjects, including graduate students of journalism, program engineers, who shared their experiences pertaining to their attempts to obtain information from government units. In Taiwan, not all the journalists have abilities to use computer software to analyze data. It is a required craft for graduate students to learn in. That is the reason why most of the users interviewed in the paper are graduates students of journalism institute in Taiwan. All the 8 interviewees have real experience to contacts with open government data by themselves.

| code | background | time |
|---|---|---|
| Interviewee A | journalism | 2014/ 6/16 |
| Interviewee B | journalism | 2014/6/23 |
| Interviewee C | journalism | 2014/6/24 |
| Interviewee D | journalism | 2014/6/24 |
| Interviewee E | journalism | 2014/6/24 |
| Interviewee F | computer science | 2014/8/11 |
| Interviewee G | computer science | 2014/8/11 |
| Interviewee H | professional journalist | 2014/6/16 |

I adopted a qualitative methodology, by which semi-structured interviews are conducted with graduate students of journalism, program designers, and a journalist. In the present study, the author would like to focus the questions:

Q1:According to your practical experience, how the government in Taiwan released the data about the public issues?

Q2: what's the quality of data for the programmers and journalists' experience?

Q3: Did the big data really benefit news story telling?

## IV. Findings

On the websites of Taiwanese government, the disclosed data were presented in different formats; thus, analyses were conducted using percentage. For example, 75% and 25% of the data of the Executive Yuan were in pdf and Excel file format, respectively; the pdf data received a one-start score and the Excel data received a two-star score. In the study, the percentage for which each governmental unit's data accounted was calculated using the five-star scoring system. The percentage for each of the 153 governmental units was then summed and divided by the number of government units, thereby acquiring the needed percentage.

Taiwanese governmental units on all levels must disclose data in accordance with

The Freedom of Government Information Law. The present study analyzed the disclosed data, and statistical percentages of central and local governmental data are discussed as follows: (Table 1)：

**Table1 Analysis of Taiwanese Governmental Data Using the Five-Star Scoring System**

| units | | 0☆ | 1☆ | 2☆ | 3☆ | 4☆ | 5☆ |
|---|---|---|---|---|---|---|---|
| All Governmental Units | 153 | 0.52% | 64.74% | 29.25% | 2.78% | 2.71% | 0% |
| the Office of the President and the five Yuan | 6 | 0% | 73.33% | 25% | 0.83% | 0.83% | 0% |
| Central Government | 41 | 0% | 58.90% | 30.37% | 8.41% | 2.32% | 0% |
| Local Government | 21 | 3.10% | 52.86% | 43.33% | 0.48% | 0.24% | 0% |

## 1. *Big data is not published in a structured form*

The present study determined that the formats in which governmental data are presented are not assistive to all computation. Overall, 64.74% of the disclosed data from the 153 governmental units received one star. (Figure1) In other words, the majority of governmental data is presented as pdf files, which cannot be directly computed with mathematical programs on the website, obstructing data disclosure. Although some progress has been made in providing data in open formats, too many documents in the UK and elsewhere are still exclusively published as pdfs -certainly not an open format in the data world. I t is astonishing how many government agencies still refuse to publish information in a structured form that can be checked and analyzed(Egawhary &O'Murchu, 2012).

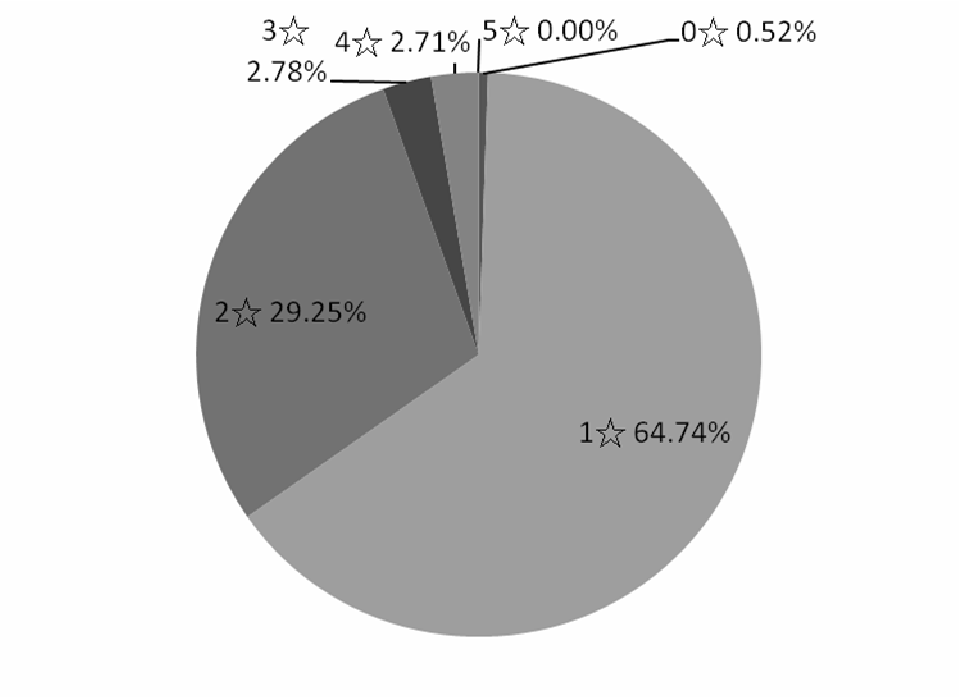


**Figure1: Analysis of the data in percentage of Taiwanese government**

The majority of data disclosed by Taiwan's highest authorities, the Office of the President and the five Yuan(governmental branches in Taiwan), was also presented as pdf files. (Figure2)

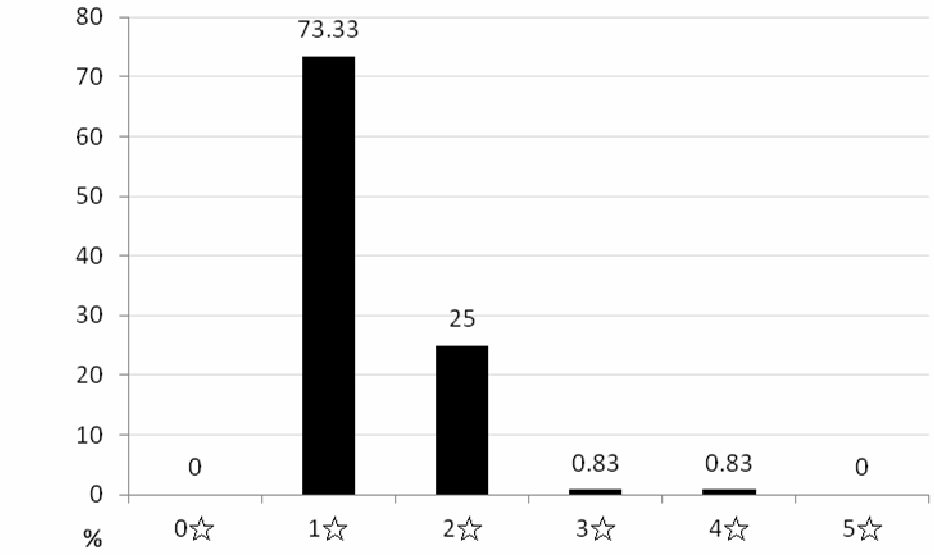


**Figure2：Analysis of the data from the office of the President and the five Yuan**

Certain local governmental units do not use electronic files and only provide data on paper. For example, data is not disclosed on the Changhua County Government's

website, and only a small percentage of the files can be directly analyzed by computers. Among the disclosed data, data that received two or more stars accounted for a mere 34.74%, of which two-star Excel files, three-star files, and four-star files accounted for a respective 29.25%, 2.78%, and 2.71%. In short, public utilization of the Taiwanese government's disclosed data is difficult, and therefore, participation in public affairs is obstructed. Compared to the data disclosure by local governmental units, that of central governmental institutions was more computer-based. Nevertheless, no governmental unit received a five-star score for the quality of disclosed information. Comparisons between central and local governmental units are as follows.(Figure3)

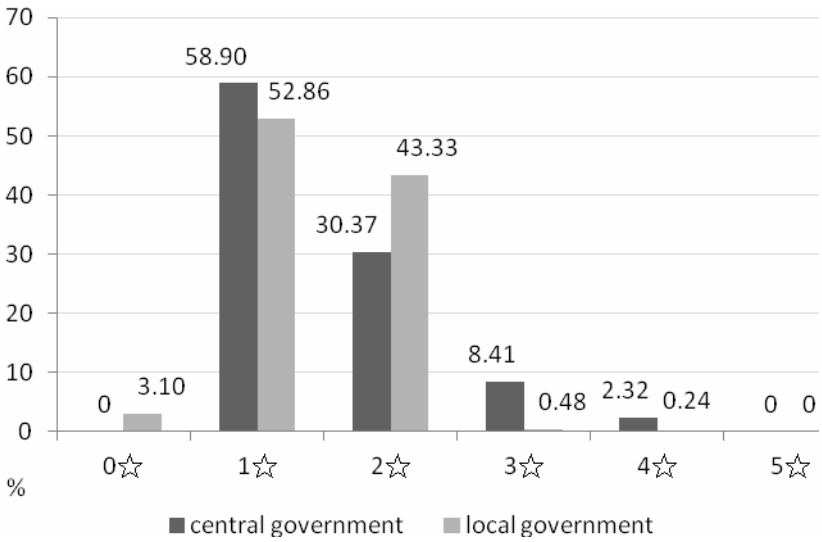


**Figure3 Analysis of the data from Central and Local Governments**

Fielded data refers to information that is often organized into fields in a spreadsheet or database. For example, the *Washington Post* has used databases to tell many stories that are based in its coverage of the federal government. The information was there in public(Briggs, 2010:256-257). In Taiwan, it is hard for journalists to employ data available in a form compatible with the computers. All the interviewees complained the pdf format frustrated their work. Even they have the data, but they could not read the data with the software of computer. They always spent much time to key in the data into a spreadsheet to compute. For example, interviewee E was composing a map to show the correlation between surveillance camera locations and crime incidents; however, this project was hindered because the majority of relevant information was presented as pdf files. It will disrupt the usages for the press to find stories hidden in database. The role of watchdog for the press will be damaged.

### 2. *Big data is not shared with citizens*

The majority of data disclosed by the Taiwanese government is presented as pdf files, which creates inconvenience for users. Interviewee C, who was composing a report on political donations, reported that information provided by Control Yuan was in the form of scanned copies; the data quantity was substantial but was not complete for all years. Therefore, Interviewee C had to screen all paper-based data at the Control Yuan and convert the required information into computer-readable data. When examining disclosed data on the website of the Ministry of the Interior, Interviewee A observed that even though pdf files were provided, data were incomplete. Interviewee A then had to physically collect data at the Ministry of the Interior in person because of lacking data on the official website. However, the government charges a photocopying fee of NT $10 per page. Interviewee A required between 4,000 and 5,000 pages of data, and thus could not afford the cost. Interviewee A stated, “I visited the Ministry of the Interior numerous times, and the civil servants employed there gave me pdf printouts in secret when they realized that I required the information for a school project.” Based on the experiences of the interviewees, although online data disclosure is now common in Taiwan, the majority of disclosed data is incomprehensive and cannot conveniently computerize.

The present study determined that the Taiwanese government is gradually presenting disclosed data through files that can be directly processed by computers. This type of file is classified as two-star and above. The statistical results indicate that a respective 30.37% and 43.33% of the central and local governmental data can be directly utilized using Excel. Three-star data accounts for 8.41% of the central government’s disclosed files, but a mere 0.48% of files disclosed by local governments. Four-star data accounts for a respective 2.32% and 0.24% of data disclosed by the central and local governments. Although a variety of formats are used for data disclosure, disclosed data remains incomprehensive. For example, half of the nine JSON files disclosed by the Soil and Water Conservation Bureau, Council of Agriculture, were pdf files, which cannot be read on the Web browser. The Ministry of National Defense has provided CSV files; however, the quantity of such files is minimal, accounting for only 5% of all data disclosed by the Ministry of National Defense. Furthermore, 60% of all the data they disclosed are pdf files. Interviewee F expressed that the Taiwanese government has only disclosed a minimal amount of data that achieves three or four stars; however, even this information does not benefit users because it is unimportant. Interviewee E stated that while the file names for certain documents are intriguing, the documents may only contain one page, whereas

some documents contains lists of laws and acts or are incomplete. “The government thinks that by presenting these types of documents, they have disclosed information,” Interviewee E commented.

The results of the in-depth interview indicated that overall, collecting government-related data is difficult. The users felt perplexed because the government did not provide requested data when data disclosure had been promised. Interviewee B recounted that when she collected data from three units within the Executive Yuan, two of the units refused her request: “The data I asked for was supposed to be disclosed; nevertheless, the civil servants refused to give me this information.” When one of the units provided pdf files to Interviewee B, the interviewee suggested that Excel files be provided to facilitate computer operations. However, the officials refused to convert the files, stating that the law only mandates the government disclose data, but does not stipulate the form of disclosure. Furthermore, a government official stated, “Only pdf files are provided, because that way, the public does not have convenient access to relevant information.”

However, some governmental units were willing to provide comprehensive electronic data. Certain governmental institutions have established additional websites dedicated to data disclosure; however, data on these websites were scant and a substantial amount of data remains undisclosed. In addition, information from the central and local governments are stored separately. This lack of data integration indicates that users must obtain data from either the central government or the local governments. Having backgrounds in computer science, Interviewees F and G frequently feel discouraged when collecting disclosed data, lamenting that the reference numbers for numerous government departments are not unified. For example, the lack of unification between the reference numbers for data disclosed by the Ministry of the Interior and the Ministry of Transportation and Communications hinders cross-examination and obstructs breakthroughs in data survey.

### 3. *Big data is difficult been transited into a story*

When Berners-Lee made his speech to ask journalists should have some new tools to manage the data, he also remind the press should keep data in perspective, helping people out by really seeing where it all fits together, and what's going on in the country( Arthur, 2010). Indeed, it is a hard work for people to find stories from thousands of, hundred of data. Even the Interviewee F with the background of computer science could make big data into a structured format, it is hard for him to find stories from the data. Interviewee F said his responsibility is to structure the unstructured data from the government. All the work is done. When he showed his

data with visualization, he was told he just visualized the data and there is little stories in it. Interviewee F responded :" I am a computer worker, it is not my job to find the stories." Interviewee G said that the workers with the background of computer science have trained to mine the data, but never been trained to find stories from the data. "It is the jobs for journalists." said Interviewee G. Both of them agreed that the skills needed to make sense of “big data” are much more sophisticated than for simple tables.

This situation tells the different backgrounds between journalism and computer science. On one hand, the workers of journalism are available to find stories but fail to compute. On the other hand, the hackers like Interviewee F and G with are hard to find the meanings or stories from database. It means the importance of cooperation between people of journalism and computer science. But, it is a long way to go in Taiwan.

It is not an easy job for journalists, either. For the graduate students of journalism, it is hard for them to find the stories from the data in specific perspective. A lack of comprehensive data disclosure indicates that the results of statistical analyses on electronic disclosed data may differ from actual conditions. Interviewee D reported that correct addresses are rarely provided in disclosed data. Instead of specifying exact coordinate locations, only relative locations are provided, and thus, manual research is required. Interviewee C stated that the lack of coordinates in Taiwan’s earthquake data created difficulties when identifying specific locations on maps; Taiwan’s earthquake data are more easily found on U.S. websites.

There is another example which could explain how the journalists try to find stories from different types of big data released from government. According to big data on political donations disclosed by the Control Yuan, six profit-seeking enterprises have each contributed one million Taiwanese dollars to President Ma Ying-Jeou. The students of journalism also check the data of all government-contracted cases for these companies were examined based on information disclosed on the Public Construction Commission website. The graduate students found that two of the six companies have undertaken numerous large procurement bids since 2002 in which the total bidding amounts to hundreds of millions of dollars. But the one million political donation from each company and numerous large procurement bids of two of the six companies could not make sense in a news story.

It means that it is hard for journalists to discover the story lurking in two types of big data released by governments. The substantial discrepancy between the donation amount and the companies’ income emphasizes doubt regarding the authenticity of the declarations for the mentioned political donations. Interviewee H, a well-known investigative journalist in Taiwan, stated that he never refers governmental

information when conducting investigative reports. The government requests that political officials must declare political donations. However, many donations remain undeclared; for example, former President Chen Shui-Bian had mysterious overseas accounts that contained hundreds of millions of Taiwanese dollars. This common practice of not declaring political donations and mysterious bank accounts leaves the credibility of the government's disclosed information open to question. Therefore, Interviewee H stated that journalists should not readily believe the authenticity of disclosed information. This situation shows that there is a risk of 'unscientific' usage resulting in lower news quality. This reminds us to consider whether the analysis of huge quantities of data maybe limited in validity and scope((Mahrt & Scharkow,2013).

Interviewee F agreed that a substantial amount of data remains withheld by the Taiwanese government. Despite the existence of files that have received four or five stars, data disclosure is considered poor because the disclosed information is irrelevant, whereas important data remains withheld. Thus, Interviewee F hopes that influential groups will demand for a higher level of governmental data disclosure. Interviewee G stated that the primary task requiring attention is establishing a unit designated to providing a unified information disclosure model for governmental units. This model should include unified reference numbers and a single application method, and should comply with relevant laws. Taiwan currently lacks a unified format for disclosed information, and different government units employ dissimilar formats. As an information engineer, Interviewee G is well adapted to converting files; however, the majority of people who require disclosed information find it difficult to utilize this information.

## V. Conclusion

The objectives of the present study was to examine the condition of governmental data disclosure in Taiwan, where democracy is highly valued. Laws pertaining to data disclosure were established in recent years; therefore, the condition of data disclosure remains unsatisfactory. Excluding the data with nation security and individual privacy, it is a must for a democratic government to release the data by law. The ruling government need to draw the clear line between the dart open or not opened. Furthermore, the government is concerned about how the public may utilize disclosed data; thus, approximately 70% of disclosed data of Taiwanese government is presented as pdf files to deliberately create inconvenience and reduce public motivation for using disclosed data. This phenomenon conflicts with the ideals of a democratic government and is detrimental to government transparency.

Incomprehensive data disclosure indicates that Taiwan's governmental information transparency is subpar, and will affect the supervisory ability of the press and obstruct developing data-driven journalism. Data-driven journalism has motivated younger generations to learn basic computer skills, which can be used to identify news leads from among large quantities of data and achieve information transparency and governmental supervision. Because disclosed data that can be easily read by computers is lacking, substantial improvement in Taiwan's governmental data disclosure is required.

"Big data" is a big issue in Taiwan. It has the potential to change the face of journalism in Taiwan. We still need to know the limitation about big data in Taiwan. It is just a beginning to disclose the data by law in Taiwan. I think it is a good start for us to evaluate the qualities of data in Taiwan. As boyd & Crawford (2012: 16) stated that there is a deep government and industrial drive toward gathering and extracting maximal value from data, we do think there are serious and wide-ranging implications of big data, and what it will mean for future research agendas. In the moment of writing this paper, I feel we need to examine big data with the context of specific issues. It could help us to evaluate the usages of big data and find big fact from big data.

The present study had certain limitations. Research on government transparency and data disclosure was conducted from the perspective of users, and governmental units were not interviewed; thus, possible difficulties faced by governmental units pertaining to data disclosure were not explored. Hence, the researchers will adopt the government's perspective in future research to determine the government's understanding of the users, perception on the value-added application of disclosed data, and definition of information disclosure and privacy. The researcher suggests that information security should be further considered by the government and focused on in future research.

## Reference


Arif,M.(2012). Customer orientation in e-government project management: A case study. In F. Bannister(Ed.), *Case studies in e-government*, (pp.1-21). Reading, U.K.: API.

Arthur,C.(2010). Analyzing data is the future for journalists, says Tim Berners-Lee, Retrieved September,7,2014,from http://www.theguardian.com/media/2010/nov/22/data-analysis-tim-berners-lee ?guni=Data:in%20body%20link

Aspland, D.(2012). The other side of "big brother": CCTV surveillance and intelligence gathering by private police. In In C.G. Reddick(Ed.). *Cases on public information management and e-government adoption*,(pp.80-99). Hershey, Pa: Information Science Reference.

Asgarkhani, M.(2012).The effevtiveness of e-service in local government: A case study. In F. Bannister(Ed.), *Case studies in e-government*, (pp.22-41). ?:API.

Baqir,M.N. &Iyer, L.(2010). E-government Maturity over 10 years : A comparative analysis of e-govermnent maturity in select countries around the world. In Christopher G. Reddick(Ed.). *Comparative E-Government*,(pp.3-22).NY: Spring.

Beer, D.(2013).*Popular culture and new media: The politics of circulation*. NY: 質 PALGRAVE MACMILLAN

Bhatnagar, S.(2004). *E-government from vision to implementation: A practice guide with case studies*. New Delhi: Sage.

Boyd,D. & Crawford, K.(2011). Six Provocations for Big Data. Presented at Oxford Internet Institute's "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society," September 21, 2011. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431

Branscomb,A.W.(1994). *Who own information?: From privacy to public access*. NY: A Division of Harper Collins Publishers.

Briggs, M.(2010). *Journalism next: A practical guide to digital reporting and publishing*. Washington: CQ Press.

boyd,d. & Crawford,K (2012): Critical questions for big data, *Information, Communication & Society*, 15:5, 662-679

Egawhary.E.& O'Murchu,C.(2012). Data journalism. Retrieved 20,Dec.2013, from: http://www.tcij.org/sites/default/files/u4/Data%20Journalism%20Book.pdf

European Journalism Centre(2010).Data-driven journalism: What is there to learn? Retrieved 13,Dec,2013, from: http://mediapusher.eu/datadrivenjournalism/pdf/ddj_paper_final.pdfFox, Julia, R.,

Fox(2001). Someone to watch over us: Back to the panopticon? *Criminal Justuce*, *1*(3), 251-276.

Gantz,J., & Reinsel,D.(2011). Extracting Value from Chaos. Retrieved August 25,2014, from: http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

Gray,J., Bounegru,L., & Chambers, L.(2012). *The data journalism handbook*.Cambridge:O.REILLy.

Hayes, B.(2009). NeoConOpticon. The EU security-industrial complex. Retrieved August 18,2014, from: http://www.statewatch.org/analyses/neoconopticon-report.pdf

Jacobs, A.(2009). The pathologies of big data. *Practice*, *52*(8), 36-44.

Krebs,V.(2010). Your Choices reveal who you are: Mining and visualizing social patterns. In J. Steele & N. Lliinsky(Eds.).*Beautiful visualization: Looking at data through the eyes of experts*, (pp.103-122).Cambridge: O'Reilly.

Mahrt, M. & Scharkow, Ml.( 2013 ). The Value of Big Data in Digital Media Research. *Journal of Broadcasting & Electronic Media*, 57(1), 20-33.

Malik, P.(2013).Governing big data: Principles and practices. *IBM Journal of Research and Development*, 57(3/4), 1-13.

Manovich,L. (2011). Trending: The promises and the challenges of big social data. Retrieved December, 12, 2013, from http://www.manovich.net/DOCS/Manovich_trending_paper.pdf

McChesney,R. (1999).Rich media, poor democracy: Communication politics in dubious times. Urbana and Chicago: University of Illinois Press.

Lazar, N.(2012). The big picture. Retrieved December, 10, 2013, from http://dx.doi.org/10.1080/09332480.2012.668458

Lewis,S., Zamith, R., and Hermida, A.(2013). Content analysis in an era of big data: A hybrid approach to computational and manual methods. *Journal of Broadcasting & Electronic Media*, 57(1), p34-52.

Misra, H.(2012). Citizen-centric service dimensions of Indian rural e-government systems: An evaluation. In C.G. Reddick(Ed.). *Cases on public information management and e-government adoption*,(pp.50-79). Hershey, Pa: Information Science Reference.

Molinari, F.,Wills,C., Koumpis,A., & Moumtzi,V.(2012). A citizen-centric platform to support networking in the area of e-democracy. In C.G. Reddick(Ed.). *Cases on public information management and e-government adoption*,(pp.128-159). Hershey, Pa: Information Science Reference.

Neuman, W., Guggenheim, L., Jang, S. & Bae, S. (2014). The dynamics of public attention- agenda-setting theory meets big data. *Journal of Communication*, 64,193–214.

O'Sullivan, J.(2012).Changing value: The "Good" journalist online. In E. Siapera & A. Veglis(Eds.). The handbook of global online journalism.(pp.39-58). MA: Wiley-Blackwell.

Pousman, Z., Stasko,J.T. & Mateas,M.(2007).Casual information visualization: Depictions of data in everyday life. IEEE *Xplore*, *13*(6), 11451152.

Rogers, S.(2013).*Facts are sacred: The power of data*. London: Faber and Faber Limited.

Roy, J.(2007). E-government in Canada: transition or transformation? In D.F.Norris(Ed.), *Current issues and trends in e-government research*.(pp.44-67).London: Cybertech Publishing.

Shapiro, M.(2010). Once upon a stacked time series. In J. Steele & N. Lliinsky(Eds.).*Beautiful visualization: Looking at data through the eyes of experts*, (pp.15-36).Cambridge: O'Reilly.

Shaviko, P.,Villafiorita, A.,Zorer, A.,Chemane,L., & Fumo,T.(2010). E-government interoperability framework: A case study in a developing country. In Christopher G. Reddick(Ed.). *Comparative E-Government*,(pp.639-662).NY: Spring.

Spence, R.(2001).*Information visualization*. London: Addison-Wesley.

Tauberer, J.(2014). Civic Hacking. In J. Tauberer(Ed.). *Open Government Data: The Book.* http://opengovdata.io/2014/civic-hacking/

Tauberer, J.(2012). Open Government, Big Data, and Mediators. In J. Tauberer(Ed.). *Open Government Data: The Book.* http://opengovdata.io/2014/open-government-big-data-mediators/

Tauberer, J.(2012). Open Government, Big Data, and Mediators. In J. Tauberer(Ed.). *Open Government Data: The Book.* http://opengovdata.io/2014/open-government-big-data-mediators/

Thornburg, R.M.(2011). *Producing online news : Digital skills, stronger stories.* Washington, D.C. : CQ Press

Toporkoff,S., Rannou,H.,Soufron,J., & levy, S.(2008).E[government architectures. In A.R.Shark & Sylviane Toporkoff (Eds.), *Beyond e-government & e-Democracy: A global perspective*,(pp.11-24). Lexington: Public Technology Institute & ITEMS International.

Ware, C.(2000). *Information visualization*. London:Morgan Kaufmann.

Wigand, R.T., Shipley, C., & Shipley, D.(1984). Transborder data flow, informatics, and national policies. *Journal of Communication*. 34(1):153-175.

Wu, X., Zhu, X., Wu, G., & Ding, W.(2014). Data mining with big data. IEEE Transactions on Knowledge and data Engineering , 26(1), January.