JeDEM 7(1): 24-28, 2015
ISSN 2075-9517
http://www.jedem.org

Wazing the Information Super Highway: Linking the World's Open Government Data Resources

Alon Peled

The Hebrew University of Jerusalem, Israel, msapeled@huji.ac.il

Abstract: Government deals daily with massive volumes of data, and yet poor public-sector information sharing remains a critical government challenge that wastes public funds and put lives at risk. Peled's book Traversing Digital Babel *proposes an innovative solution to improve information sharing inside government by using selective incentives to nudge government officials to exchange data in an automated Public Sector Information Exchange (PSIE). Enabled by grants from Google and Yissum, Peled embarked on the first step to creating a testable PSIE prototype: building the world's first fully-automated corpus of metadata about Open Government Data (OGD) released by local and national governmental agencies worldwide. The OGD corpus currently contains metadata about information assets from 24 different countries, and presents rich opportunities for research including examining and comparing government patterns of information release. The corpus could also enable an innovative model of combined free and for-payment government data release.*

Keywords: Open Government Data, Information Sharing, Public Sector, Exchange, Metadata

1. Keynote Speech

Who owns the largest volumes of information in the world? Imagine that the US Library of Congress's entire printed material from the past 214 years is digitized. Let's say that this entire digital data is the length of a bacterium - you have to use a microscope to see it. Now, see how Google's annual processed data from 2007, 2011 and 2013 is much bigger (comparatively the size of a hornet, love bird, and Marlyn bird). So, is Google the largest processor of digital data in the world? Not even close! In 2014, one antenna of one project of one US agency (NASA) produced 25% more data than all of the data that Google processed in 2013. So, is NASA the biggest producer and processor of data? Not even close! The latest US National Security Agency (NSA) data center can store 700% more data than Google processed in 2013!

If government deals daily with such massive volumes of data, are they experts in data integration and sharing? Unfortunately – no! Poor information sharing was the single greatest failure of the US government in the lead-up to the 9/11 attacks. Numerous government agency

 CC: Creative Commons License, 2015.

databases held important information about Mohammed Atta – the ringleader of the 9/11 terrorists. This information was not shared with the CIA.

The Hurricane Katrina relief effort was impeded by poor information sharing. The Katrina investigation committee stated: "Better information would have been an optimal weapon against Katrina." After the 2011 tsunami and the Japanese nuclear power plant accident, a lack of information sharing led to residents being evacuated directly into the radioactive plume. Aborted or unsuccessful information sharing projects in the public sector cost billions of dollars annually. Failure to share information among government agencies is a critical challenge that wastes public funds and threatens lives.

In my new MIT Press book published in 2014, *Traversing Digital Babel*, I explain why governments worldwide have failed for more than two decades to force, beg, or coax large government agencies to share information with each other and with the public. Agencies view information resources as assets and resist releasing them for free. The book shows that government agencies worldwide refuse to publish high quality information on Open Government Data portals because they know that their information is valuable and have no incentive to release it free of charge to other agencies or to the public.

Traversing Digital Babel proposes an alternative, innovative solution to improve information sharing inside government by using selective incentives to nudge government officials to exchange data in an automated Public Sector Information Exchange (PSIE). A PSIE is a data exchange program that enables public sector agencies to open up virtual "data goods" shops and offer their information goods to other agencies for compensation. To trade, agencies inscribe their trading algorithms into intelligent software agents that can find other agents, negotiate a price, sign a contract, and deliver home purchased information goods. Agencies accumulate credits by allowing other agencies to browse or access their data. Agencies that perform well exchange credits for dollars to invest in information products. PSIE solves the biggest information sharing problem inside government—providing motivation for agencies to open up information assets to other agencies. With PSIE, citizens will benefit from faster responses, improved data-quality, and new e-Government information services.

2. From PSIE to the World's Largest Corpus of Metadata about Open Government Data (OGD)

In 2013, I dared and submitted the PSIE concept to a prestigious worldwide grant competition, the *Google Faculty Research Award*. Typically, only computer scientists win this competition. With great (but typical Israeli!) audacity, I wrote to Google: "I am a political scientist but also know how to develop software. Give me the money to build PSIE!" To my delight, Google granted me the award. Around the same time, I won another smaller grant in an internal competition held by Yissum, the Hebrew University's corporation in charge of helping scholars convert their research into products that change lives out there in the real world. Now, I had about $100K and one year to build PSIE. I hired Steven Karas, a young and brilliant software developer part time, purchased two laptops and began the software development work.

 CC: Creative Commons License, 2015.

Google and Yissum demanded that I test the PSIE prototype against *real* government data. But how does a lone Israeli scholar acquire access to official and raw governmental data at all levels of government in dozens of countries and in every language? Google nudged me to look into Open Government Data (OGD) portals. I had a slight ethical problem here. You see, in an article from 2011 (and chapter three of my book), I argued that Open Government Data does not and will not work because public sector agencies have already discovered the monetary value of their data and will not release it to the public free of charge. But the Google offer was tempting so I decided to give OGD a second chance. After all, I only needed enough governmental data to test my PSIE system.

Back then, at the end of 2013, I assumed that someone, somewhere had already created a unified catalog of OGD assets released by local and national governments worldwide. I needed access to such a catalog so I could select and download the test data I needed to build PSIE. But I could not find such a unified catalog of worldwide OGD assets. There were silo-OGD portals such as the "American federal government OGD portal." Business analytics vendors, such as Socrata in the USA and Junar in South America, sold propriety OGD solutions to specific government clients. Even important scholarly projects such as the Open Data Barometer were focused on a small number of OGD publications in many countries. The best of the best, was the European crowdsourcing *ENGAGE project*, but even this project was tightly focused on Europe and held metadata about only 52K OGD assets. I needed a much larger catalog that contained metadata about every OGD information asset released by *any* agency, in *any* language, in *any* country. I promised that PSIE would work at all levels of government, in every country and my test data had to match this promise.

So, we began building the first part of PSIE as a Google-like service to capture and enrich the metadata about all OGD information assets released worldwide. We were only two guys with about $100K and a one-year project deadline so we had to work smart and fast. We made many mistakes but, as Google likes to put it, we "failed fast and learned quickly."

We were successful. One day we stared at this "beast" that we had created—the so-called test data corpus and realized that unexpectedly we had created the world's first, fully-automated, biggest, most comprehensive and richest corpus of OGD metadata. How big is this "biggest?"

Well, today we have almost 400,000 OGD information assets in the corpus from local and national governmental agencies in 24 countries and in 14 different languages. Our metadata about these 400,000 assets contains 4.75GB of metadata about 10TB of data that governments have released on the Web. Right now, we cover 521 catalogs—the famous USA federal OGD portal, data.gov, is just one of these catalogs—worldwide and our software continues to crawl the Web, collect metadata about OGD information assets and increase the corpus-size.

So what can you do with this corpus? For a start, you can search and find OGD information assets that you cannot find with Google. Imagine, for example, that you are searching to find information about the Italian city of Bari. Compare what is retrieved from Google for "Bari" with what is retrieved by a PSIE search. Note also that the metadata in PSIE is already translated into English (originally was published in Italian of course!). Note also that the OGD information about

 CC: Creative Commons License, 2015.

Bari in PSIE originated from multiple OGD portals including Bari itself and the Italian national OGD portal!

Next, you can mine the PSIE corpus to discover fascinating new insights. For example, Professor Karine Nahone and I discovered in PSIE that a tiny group of no more than 70 federal OGD enthusiasts was the real force behind President Obama's OGD vision. These enthusiasts often crossed the boundaries of their own units and convinced other federal units to contribute data to data.gov, the USA federal OGD portal. In another paper, Professor Nahone and I used PSIE data to demonstrate exactly how federal agencies opted to collaborate more closely with informational metadata elements while steering away from complying with the requirement to publish OGD data that could be used to hold them more accountable for the execution of congressionally-mandated programs. At CeDEM 2015, Dr. Shkabatur, Professor Nahone and I will present another PSIE-based paper in the "Open Data, Transparency and Open Innovation" panel. In this paper, we demonstrate that one can even measure the OGD heartbeat of cities that claim to comply with the OGD policy innovation, and separate cities that pay lip service to this innovation from undecided cities that are yet to make up their minds about OGD and cities where OGD has become a way of life. More recently and in another paper, Dr. Shkabatur and I demonstrated that OGD evolution in developing countries such as Moldova, the Philippines and Morocco follows a different historical evolution path than that of developed countries. The PSIE data empowers you to do all that and more!

But why stop there? There is so much data in the PSIE corpus that you can actually visualize it and see it interactively. Having translated all the PSIE metadata into English, we wrote software to analyze the metadata and detect the potential, "hidden" economic uses of the released data—unanticipated even by the public officials who published the data. You can see for yourself that in virtually every OGD country a certain breadth of metacategories must be reached before governments begin *really* releasing a lot more data.

….and then, the PSIE corpus data can also be used to support joint projects with the government itself! NASA hosts an annual Space Apps hackathon. This year's hackathon was held in 147 cities with about 13,000 participating software developers. NASA wanted to include a challenge about Open Data but was unable to do so because NASA itself had some technical problems with publishing its data (these problems are now fixed). So, in the PSIE corpus, we discovered enough NASA data published on OGD portals of other space agencies in other countries, to create a Hackathon challenge. The goal of this challenge was to find a way to transform NASA 's information assets so that they are easier to discover on the Web, to enable citizens, entrepreneurs, and experts working in non-space domains can discover and use them. 45 developers from eight countries—including Egypt and Tunisia—labored for three days and three nights to solve this PSIE-based challenge. NASA liked some of the solutions so much that they used the PSIE challenge—out of 25 challenges—to demonstrate the Return of Investment (ROI) of this competition for NASA!

In fact, PSIE may even hold the keys to solving the toughest "supply" problem. Long before Obama and OGD, government agencies were asked to sell some of their information products to supplement their budgets. The later OGD policy innovation was lukewarmly welcomed by these

 CC: Creative Commons License, 2015.

agencies—how can you be expected to simultaneously sell information and release it for free? But what if, through PSIE, you could support a new balance, a new OGD ecosystem where you sell some data and release other data for free. In Switzerland, for example, a simple map of the country is free and available on the web portal of the national GEO agency. But, if you are a navigator and require more sophisticated and higher-resolution maps that cost more to produce – you can pay 1 euro for such a map. PSIE can support both OGD-for-free and OGD-for-sale. Remember that Stuart Brand said in 1985 that "information wants to be free" (everybody remembers that part) but he also said (the second half of his sentence) that "information wants to be expensive". In our information age, PSIE might hold the keys to creating a new balance between free governmental data and governmental data that is for sale.

3. In Conclusion

I am here all week long and will be delighted to brainstorm with all of you about additional creative uses for the PSIE corpus. In the meanwhile, please remember that in many places OGD is both the law and is also—potentially—very good business.

I do not know if I will ever be given the opportunity to finish building PSIE the way it is described in my MIT Press book. But, for now, I am busy dreaming and creating new ways of employing the PSIE software and corpus to create a new virtuous cycle between the supply and demand sides of the OGD policy innovation. Thank you very much!

Alon Peled

Alon Peled is an associate professor and political scientist at the Hebrew University of Jerusalem, Israel. He is fascinated by the interaction between information and politics in the public sector and innovative information-sharing technologies that facilitate this interaction. In his work and research he draws on rich experience both in academia and in the world of software engineering (where he has worked as principal software engineer and program manager in data warehouse and office automation software projects in the US and Israel).

 CC: Creative Commons License, 2015.