Are They Making Our Privates Public? – Emerging Risks of Governmental Open Data Initiatives

Thomas P. Keenan

Faculty of Environmental Design, University of Calgary
Department of Computer Science
Calgary, Alberta Canada
keenan@ucalgary.ca

Abstract. Governments around the world are opening their data vaults to public (and corporate) access and scrutiny. Notable examples including New York City's NYC Datamine, Philadelphia's OpenData Philly, Europe's Open Data Challenge, and Canada's Open Data Framework, which now spans several cities including Toronto, Vancouver, Edmonton, and Ottawa. Anyone can read government plans, budgets, contractor lists, and in many cases, documents relating to individual citizens. The intention behind these data transparency projects is laudable, but it behooves those interested in privacy to take a careful look at just what information our governments are sharing with the world. There have already been "Open Data Hackathons" which have discovered interesting and unforeseen vulnerabilities, often by combining multiple data sources. There are also commercial ventures using government released data in combination with other sources in ways that were never anticipated, such as genealogical research. We are breaking new ground here and we need to generate new principles to protect privacy in the face of data that is going from "public" to "super-public".

1 Introduction

"The road to hell is paved with good intentions" is a proverb whose origins are lost in history but that becomes more timely every day. Governments around the world are leaping breathlessly onto the "Open Data" bandwagon, driven by a desire to improve citizen services, a fear that officials will be accused of hoarding data, and just plain economics, since enabling third party Open Data applications often costs the government almost nothing. There is also a certain cachet to being "open" with your data -- perhaps because of the warm feelings people have for open source software, the Creative Commons license, etc. In addition, the Wikileaks disclosures of government data have convinced many in government that their confidential data will get out anyway. It looks far less suspicious and sinister if they release it voluntarily and systematically.

What follows are examples of major governmental Open Data initiatives and a demonstration of how each illustrates a type of privacy-related problem.

J. Camenisch et al. (Eds.): Privacy and Identity 2011, IFIP AICT 375, pp. 1–13, 2012.

1.1 New York's *NYC Datamine* – Fat Fingers at the Data Office

As the financial capital of the United States, as well as a media center and home to numerous high tech companies, it is predictable that New York City would take a lead in opening government-collected data to the public. And so it did, with a highly touted and freely available collection of 103 municipal data sets and a promise to add even more. The NYC Datamine project was unveiled on October 6, 2009, only to be greeted with an immediate privacy scandal.

According to Nancy Scola, writing on TechPresident.com, "Discussion on the public Sunlight Labs Google Group revealed that one XLS table listing the city's more than 1,100 women's organizations contained not only the personal email address of the group's contact with the New York City Commission on Women's Issues (CWI), but what he or she was using as the Secret Question and Secret Answer." [1] This relatively minor privacy breach was actually fortuitous because it forced the city to examine all its newly released files to ensure that they did not contain other personally identifiable information.

While this oversight could be dismissed as a "teething problem," it illustrates that even in a major jurisdiction, well aware of privacy issues, mistakes can and will occur. The clear lesson is to carefully review everything that is being released, thoroughly remove personal data, and act swiftly if a mistake is made.

1.2 Philadelphia's *Open Data Philly* – Too Much Sharing

One of principles of North American democracy is that contributors to political campaigns (who often receive a generous tax deduction for their contribution) should be identified. This seems justified, as the public has a right to know who is financing campaigns to ensure that politicians are not being "bought" by generous donors. Some jurisdictions only report contributions over a certain monetary value. The City of Philadelphia, PA, USA has chosen to report all contributions. and make the results "freely viewable and printable, but not available for download" on a website, www.opendataphilly.org.

There, we learn that a certain person contributed $3.00 US to the Communication Workers of America (a trade union). We are also given that person's precise home address as reported on the campaign contribution receipt. In fact, checking some of these addresses against the best known U.S. home address directory database (www.anywho.com, operated by Intellius, Inc. and promoted by AT&T) shows that the information in the contributor database is frequently more complete than what is available in the directory. In many cases a person's address in the contribution database is not even shown in the directory database, perhaps because that person asked to be unlisted for privacy reasons. In an era of home invasions and identity theft, this is a very common choice.

On the other hand, people are highly likely to provide a complete and accurate address for a campaign donation receipt, since it is an official document and they are expecting to file it with their income tax return to claim a tax deduction.

The most reasonable justification for including contributor addresses in the public dataset would appear to be to disambiguate donors who share the same first and last names. There might also be some value for doing a geographic analysis of donation patterns. Despite the claim that this dataset was not downloadable, it was actually easy to download parts of it, such as everyone with a particular surname. Checking the three most common American surnames, Smith, Johnson and Williams, produced only a few apparent duplicates where having the address information may have been helpful. The experiment performed is described below.

Method: The donor files for 2010 were downloaded from OpenDataPhilly as .CSV files, imported into Microsoft Excel and sorted by the contributor name field. "Potential Duplicates" were defined as additional names that were shown identically as a name in the database, but with a different address. Of course it is possible that a person moved or gave a home address on one donation form and an office address for another donation. So the "duplicate" might really be the same person. Therefore, this is a conservative estimate of how much benefit might possibly accrue from having the addresses.

Results: The results were as follows:

Name	Number of Entries	Number of Potential Duplicates
Smith	588	8
Johnson	426	3
Williams	400	4

Conclusions and Observations: There were few cases where providing an address would be of any value in distinguishing people with the same or similar names. It should also be noted that other interesting inferences can be made from having these addresses available. For example, a significant number of contributors (86/588 of Smith, 63/426 of Johnson, 77/400 of Williams) listed the same address, 1719 Spring Garden Street, Philadelphia, PA., 19130. Viewing this location on Google Maps Streetview shows that it is the office of the I.B.E.W. Electricians Union. This makes sense because these donors are apparently receiving their receipts in care of that business at its office address.

The fact that over 15% of the respondents in these three name groups gave the 1719 Spring Garden Street address further illustrates the futility of trying to disambiguate people with identical names based on their addresses. For example, there are two runs of contributions at that address from a "Michael Smith". Are they from the same person? There is no way to tell from this database. So the inclusion of home address in this database actually compromises privacy without adding any real functionality. It was probably just easier for those in charge of making the data public to leave the address in rather than taking it out.

Should we be concerned by the inclusion of address information in this database? There is certainly an issue of informed consent and purpose of use for the data. It is highly doubtful that people realized when they filled out their donation receipts that their addresses would be permanently posted on the Internet for all to see. In fact, this database contains information back to 2005, long before the OpenDataPhilly project even existed. So it is fair to say that, in many cases, citizens have unwittingly disclosed information through their city government that they did not willingly provide to other sources such as the telephone directory company.

As for whether or not home address is protected as Personally Identifiable Information (PII), the law and practice varies widely by jurisdiction. However, the US National Institute of Standards and Technology, in its Special Publication 800-122 explicitly defines "address information such as street address" as being PII. [2]

1.3 Canada's *Open Data Framework* – Are Elephants Feasible?

There is certainly an argument that people who are paid with public money should be willing to have that information placed in the public domain. This would include those who received consulting contracts from a municipal government. Yet, overly aggressive journalists or snoopy citizens could easily make unfair use of some of the data.

A number of Canadian cities have jointed the Open Data Framework with a commitment to making as much civic data as possible freely available to the public. Looking into the consultant expenses of the City of Toronto, one might well ask why a Mr. John Lehnhardt was paid $3,275 for an "Elephant Feasibility Study." We already know that elephants are feasible. Fortunately, this line item is tagged with the label "Toronto Zoo," so we can imagine that it is probably quite valid and justified. Then again, in tough times, Toronto citizens might well ask why Victor Ford & Associates charged their city government $7,500 for a "Mountain Bike Skills Park Site Assessment." In these cases, providing more data might have been desirable, e.g. some justification of the expenses so that they are not mis-interpreted when disclosed in database form.

It should be acknowledged that the designers of Toronto's Open Data initiatives have made some attempts to protect the privacy of the general public. For example, their database on calls to the city's complaint and service request line (reached by dialing 311) is anonymized to show, in general, only a partial (three out of six characters) postal code, e.g. M4V. That narrows the address to a part of the city, but it might be many city blocks. However, there are numerous cases where this field instead contains a precise intersection such as WOLFE AVE & DANFORTH RD, SCARBOROUGH. With enough cases like this, and other databases such as Google Maps and Google Streetview, it is certainly possible that the calls could be traced back to an individual property and hence to the owner.

While these cases would probably only be of interest to a bored journalist or a vindictive neighbor, the principle is clear. Data being placed in the public domain can be used for inappropriate purposes. As more and more data sets are released, the chances go up that someone will find something interesting to analyze and possibly track it back to an individual person.

1.4 Edmonton's Election Results – Did My Wife Vote for Me?

Election results are high on the list of data that clearly belongs in the public domain. Well-meaning Open Data fans even build "real time dashboards" to display election results more graphically to a waiting public. Still, there are privacy issues here, such as reporting very low vote counts.

The City of Edmonton, Alberta, Canada made the results of its 2010 municipal election available on the Internet. [3] From this dataset, we can learn that a mayoral candidate named Robert Ligertwood received 0 votes in the city wide hospital voting. So what? Suppose his wife was hospitalized at the time of the hospital poll and confirms that she voted in it and says she voted for Mr. Ligertwood. Now there is either a vote counting problem (rather unlikely given the auditing procedures for Canadian elections) or a marital honesty problem. In any case, one should not be able to deduce how an individual voted on a secret ballot from publicly released information.

While this seems like (and is) a contrived example, it is indicative of a general problem in data release whereby the reporting of small numbers can be used to make fairly accurate inferences about individuals.

Statistics Canada is the Government of Canada's data collection arm and has legal authority to compel individuals and businesses to provide data, e.g. on census forms. They also have an obligation to protect individual and business privacy. They dealt with the "low number problem" long ago by reporting "not significant" when cells in a database fall below a certain threshold.

The US Government publication cited above [2] also contains guidelines for effectively anonymizing data reports, which include introducing noise and replacing data in a group with the average value. All of these could be applied for the low-scoring candidates in an election.

1.5 Europe's *Open Data Challenge* – Bring on the Lawyers

The Open Knowledge Foundation sponsored the Open Data Challenge, held from April to June 2011 which by all accounts was a huge success. Offering total prize money of 20,000 € brought an impressive 430 entries from citizens of 24 EU member states. The winner, Znasichdani.sk, created by the Slovakian NGO Fair-Play Alliance, allows anyone to enter a name and obtain the value of Slovakian government contracts issued to companies in which that person plays a role. Entering "Vladimír Poór" (a Slovak entrepreneur, and one of the names suggested by the site) pulls up his association with contracts ranging back to 2005 and with a total value of 52 828 130,05 €. The data is obtained by cross-referencing existing governmental databases.

This application clearly touched a nerve, since one of the companies listed in it successfully sued to have certain data removed. According to news reports, "statutory representative, Jarmila Povazanova, of the Strabag construction company demanded in court that the total value of all public contracts of companies Povazanova represents be removed from the site. The Bratislava II district court ordered the NGO to remove that information." [4]

On one level, this demonstrates the value of the database. If a company bothered to sue to have the data removed, it may be assumed that it “had something to hide”. Ironically, the media attention from this lawsuit has served to turn the spotlight on the companies involved and their perhaps too cozy relationship with the Slovak government.

There is also an interesting technical issue here since the Znasichdani.sk application is only a conduit to official governmental databases. Since the offending data was not held within the scope of that application, there is certainly something strange about a court order to delete some data when in fact the data was hosted on the government’s own databases!

2 Should the Rich Have Less Privacy?

The fact that legal action was even taken in the Strabag case demonstrates the often contentious nature of government-held data. Most people agree that those who receive public money should be subjected to scrutiny. Yet, this principle certainly does not apply at all socio-economic levels.

Consider the EBT/Food Stamp subsidy program in the USA, which is intended to help those who cannot afford to buy basic food products.

There is a strong feeling that this program is abused by many recipients. As one online commentator put it, “I saw people come through my line who would buy all kinds of expensive junk food. Name brand foods. Chips, donuts, fruit snacks, microwave popcorn, all kinds of stuff like that. Then they'd pull out their little EBT card, and I would stand there and think ‘Wait... is our government really helping you?’” [5] The same writer notes that some people use food stamps for their groceries, then pay cash for beer and cigarettes, demonstrating that they had money that could have been spent on food.

Technologically, we could easily track and even post the purchase history of food stamp recipients, who are, after all, consuming public funds. Yet as a society we have chosen not to do that. The outrage about EBT spending expressed in the blog posting quoted above remains just a personal rant, not a systematic disclosure. This raises a provocative question about whether or not government database releases are effectively discriminating against certain sectors of society.

A competition, sponsored by the Ethics & Excellence in Journalism Foundation and the John S. and James L. Knight Foundation was held in April 2011 at the WeMedia NYC conference. Companies competed for two $25,000 prizes based on their innovative technology ideas.

One winner was a proposed website called Stable Renters: Public Scoring for Apartments and Landlords (www.stablerenters.com). Among other things, it will allow tenants to identify who really owns their apartment building, as well as searching for health and building code violations. [6] The example shown on their website illustrates a building in Brooklyn, NY with “191 open violations since 2000” and provides the real names of the owner (as opposed to a holding corporation) and manager. It also assigns a grade (in this case “F”) to the property to warn prospective renters. In accepting the prize, site founder Benjamin Sacks said he wanted to “level the playing field” between landlords and renters, since the former already have access to tools, such as credit reports and even confidential blacklists, to evaluate prospective tenants.

Landlords are people, with privacy rights like anyone else. Just as some doctors object to physician rating systems like www.ratemymd.ca and many professors bristle at anonymous student comments on www.ratemyprofessors.com, landlords might well feel that people are posting untrue and unfair comments about them and their properties on the Stable Renters site. With no easy way to have this information corrected, they might well feel like victims. The problem escalates to a higher level when this information is combined with other sources, such as those that might reveal the home address of a landlord.

Of course it is not only the rich whose personal details can be exposed through the release of government data, and it is not all about money. Lives may be endangered. Although it was not planned, the July 2010 Wikileaks-driven release of US military files relating to the war in Afghanistan was reported to disclose the true identities of Afghani translators and informants who cooperated with the US Forces, possible endangering their lives. There was also a situation after the 2006 Katrina hurricane in the US in which data on 16,000 aid recipients was improperly posted on a web site, reportedly endangering some, such as those who were being protected from abusive spouses. [7]

3 Indirect Risks of Releasing Government Data

There is every reason to believe that private companies will use government data to their advantage, both in overt ways (like re-selling it) and for their own internal purposes such as looking up past government contracts or development plans to improve their own commercial fortunes. Indeed, corporate use is part of the reason the data is being released. The New York City MTA (which runs the busses and subways) even makes a virtue of this, with ads in subway cars bragging that “our apps are whiz kid certified.” [8]

What they really mean is that they have allowed independent contractors to create smartphone apps using the MTA’s data. Of course they also saved the cost and annoyance of having their own IT department develop them, and they can disavow responsibility if your bus doesn’t show up at the time shown on your smartphone.

Bus schedules and arrival times don’t compromise personal privacy (except perhaps of errant bus drivers) but genealogy sites most certainly do. Consider the wildly popular genealogy website ancestry.com, which has local versions like ancestry.ca in Canada, ancestry.co.uk in the UK, etc. According to their December 2010 report “more than 6 billion records have been added to the site in the past 14 years. Ancestry users have created more than 20 million family trees containing over 2 billion profiles.” [9] The vast majority of the company’s data comes from government sources, and, in that same report, they note they have recently added US military cadet applications and U.S. penitentiary records.

Clearly, the vast majority of these birth, death, marriage, immigration, travel and military service records were not created with the intention of being part of a for-profit company's genealogical database. The persons mentioned in them were never asked for permission for these records to be released. They have simply been made available to this company.

Aside from the great convenience of the ancestry.com user interface, privacy compromise in genealogical records is also facilitated by the existence of a common data format for their exchange, GEDCOM, developed for the (Mormon) Church of Jesus Christ of Latter Day Saints which has a huge interest in genealogy.

Are breaches of personal privacy occurring on genealogy sites? Almost certainly. According to a posting by a certified genealogist on rootsweb (hosted by Ancestry.com) "In just one file that I downloaded . . . I found more than 200 names of persons born within the last 70 years." She quotes another report that "I was shocked and dismayed to find that someone had copied my entire GEDCOM and put it up on their Web site. While I have no objection to anyone using my dead ancestors, this person had included the living as well . . ." [10].

This author goes on to plead for voluntary restraint in the posting of information relating to living persons:

"We should exercise good manners and respect the privacy of our families -- those generous relatives who have shared information with us or who shared with a cousin of a cousin. Additionally, there is another and growing problem -- identity theft. Why make it easy for cyberthieves to steal your or a loved one's identity?" [10]

In terms of potential commercial misuse of genealogical data, the most commonly cited example is insurance companies who might infer, for example, if all your known relatives died at a young age you might be a poor risk for life insurance. More subtle interactions can also put privacy at risk through informed speculation about genetically-linked medical conditions.

According to a US National Institutes of Health publication, [11] a set of gene mutations referred to as Lynch syndrome is linked to colorectal cancer. Identifying it by genetic testing can be helpful in selecting the best treatement. However, individuals who carry this mutation are also susceptible to other cancers. "An insurance company or potential employer who learns that a person carries the mutations that can cause Lynch syndrome now knows that the person is susceptible not only to colon cancer but also to other cancers as well," comments attorney Andrew Spiegel, chief executive officer of the Colon Cancer Alliance. [11]

Things get even creepier when DNA data is added into the genealogy database, and that is certainly becoming feasible. A 2007 report claimed that ancestry.com was in the process of adding DNA data to its site:

"Ancestry.com intends to launch a DNA testing program to their site by the end of the summer, all for $200 and decrease in personal privacy for your entire family gene pool. Ancestry.com has 24,000 genealogical databases, meaning that your cheek-swab test would be available to anyone with access to their site. Sorenson Genomics is partnering with Ancestry.com on this project." [12] Their DNA testing price has now dropped to $149 USD. Another genetic genealogy site, www.dnaancestryproject.com is also quite sweeping in its scope.

Those who contribute DNA to databases like these, even with the best of intentions, are compromising the privacy (and possibly the insurability, employment prospects, etc.) of themselves and even their family members. As the author at lossofprivacy.com points out, there are very large privacy and confidentiality issues here. “What happens when an insurance company gets a hold of these results and then denies your claims, or even insurance, because you have a possible genetic, pre-existing condition? You might not even have your DNA on file with Ancestry.com, but your sister, brother, mother, father, cousin, etc., might and their results could still tell a lot about you even though you’ve taken the precaution to not have your DNA in their database.” [12]

London-based watchdog Privacy International filed a lawsuit against Ancestry.com relating to their use of DNA data, stating that it believes "that the practice substantially violates UK Data Protection law" as well as the European Union Data Protection Directive.” [13]

4 The Implications of De-anonymnization Techniques

US legal scholar Paul Ohm notes that, contrary to common belief, computer scientists “have demonstrated they can often 'reidentify' or 'deanonymize' individuals hidden in anonymized data with astonishing ease” [14] and argues that anonymity is much less effective at protecting privacy than is commonly believed. “This mistake pervades nearly every information privacy law, regulation, and debate, yet regulators and legal scholars have paid it scant attention,” he writes. Ohm goes on to provide technical and legal suggestions for dealing with the growing ability to deanonymize databases.

There are numerous well known examples of supposedly anonymous data being “deanonymized” by sophisticated analysis. Notable among these are successful attacks on the anonymity of the Tor Network using traffic analysis techniques [15] and the Netflix Prize which offered $1M US for the best algorithm to predict user ratings of movies. Researchers [16] found a way to identify certain individual users in the anonymized dataset released by Netflix, an online video rental company. This resulted in a lawsuit which was settled out of court and the cancellation of plans for a second contest. [17]

Further evidence that the risks of de-anonymization are more than theoretical and definitely apply to governmental data releases come from a recently released paper [18] that reports the results of turning three groups of students loose on supposedly anonymized re-offender data from the UK Ministry of Justice (MoJ.) O’Hara and colleagues found that at least one case of supposedly anonymized data was identifiable in conjunction with information on a local news website. This resulted in the data being sent back to the MoJ for further data redactions.

It is also worth noting that as progress is made in this type of analysis, previously released databases that were considered safe from attack may become vulnerable. This prompted one observer to caution that we should “beware of time travelling robots from the future.”

5 A Carpenter Is Only as Good as His Tools

Releasing public data sets would be a non-event if there were no tools to retrieve, organize, download and analyze them. Many Open Data projects are incorporating intuitive, easy to use interfaces with their data. Freely available databases such as SQL Lite also play a role here.

A clever tool called ScraperWiki collects Ruby, Python and PHP scripts that people have written for various purposes, thus providing convenient information retrieval across many databases. Available scrapers currently include everything from the zodiac signs of Nobel Prize Winners to the staff directory of employees of the Digital Enterprise Research Institute in Galway, Ireland. This information is all generally available in other forms – the role of the Scraper is to present it in a convenient format and its website www.scraperwiki.com alerts users to what is available, inspiring creativity in data searching.

Open Data competitions and contests as well as "Open Data Hackathons," (physical or virtual meetings where people develop tools and uses for public data) have entered the culture in a big way. David Eaves, advisor to the Mayor of Vancouver, Canada has said that "Open Data competitions are the innovation labs of open data, they are important not only because they foster new applications, but because they can expand our horizons and begin to reveal the depths of our imagination, and the potential of the open data opportunity" [19] Of course there are White Hat and Black Hat Hackathons, and we need to be very concerned about malicious uses of public data that might emerge from the latter.

In his writings, Eaves makes an important point which was raised when a version of the present paper was first presented at the *IFIP Summer School on Privacy and Identity*, held in Trento, Italy, Sept 5-9, 2011, and also in other venues. Eaves cautions us not to abandon worthwhile Open Data projects simply because the data released might possibly be used in a way that is harmful, illegal or embarassing to someone. In a blog posting [20] he likens Open Data initiatives to the building of public roads, which of course can be mis-used by speeders and criminals.

"The opportunity," Eaves writes, "of both roads and data, are significant enough that we build them and share them despite the fact that a small number of people may not use them appropriately. Should we be concerned about those who will misuse them? Absolutely. But do we allow a small amount of misuse to stop us from building roads or sharing data? No. We mitigate the concern." [20] Section 7 of the present paper presents some suggestions for such mitigation, as well as references to those who are tackling the privacy risks of Open Data.

6 The Rise of "Super-Public" Data

A great deal of data about individuals has been "public" for a long time. Documents filed in court cases, such as divorce proceedings, can often be accessed, though it might require a trip to the dusty basement of a small town courthouse. With the rise of the Internet and digital document preparation, all that has changed. As Fertik and

Thompson write, “if it happened in the past ten years, it might be online. If it happened in the past five years, it’s probably online. And if it happened in the past two years, it’s almost certainly online.” They add that “anything that is said online may be available forever, no matter how hard anyone tries to delete it” [21]

Private data aggregation companies such as Alpharetta, GA, (US) based ChoicePoint Systems Inc. have collected data on individuals for many years, going far beyond what credit bureaus keep in their files. They reportedly sent employees to hand copy court records such as divorce proceedings to build up their files, which they resold to prospective employers and others for a substantial fee. ChoicePoint was involved in numerous privacy breach and identity theft scandals and in 2006 was ordered to “pay $10 million in civil penalties and $5 million in consumer redress to settle Federal Trade Commission charges that its security and record-handling procedures violated consumers’ privacy rights and (U.S.) federal laws.” [22] The company was purchased in 2008 and is now part of LexisNexis. There is every reason to believe that firms in the data business will freely avail themselves of any data made public by governments and use it as they see fit.

The ease of access to public data, largely driven by open government (and to some extent parallel programs in the private sector,) is qualitatively changing the very nature of public data. What used to be “public” in the sense that you could access it with a great deal of effort such as flying to another city, is now available with a few clicks of a mouse. It seems fair to call this data “super-public” since it is a far cry from what used to be thought of as public data.

7 Conclusions and Recommendations

Open Data initiatives are very much in the spirit of governmental transparency, open source sharing, and the mantra that “data wants to be free.” They are extremely valuable and important, and it is certainly not the intent of this paper to hinder their development.

However, governments do need to think more carefully about the privacy implications of Open Data. They must develop ongoing and effective safeguards to deal with the creativity of the crowd, which may sometimes invade personal privacy. In the long run, this will benefit the Open Government and Transparency movements, since it will build public confidence in these projects and minimize the negative effects of data breaches, which are probably inevitable.

As a minimum, those implementing Open Data projects should:

-Scan files carefully for direct PII that may be included;
-Consider ways in which PII may be revealed indirectly;
-Act promptly to remove or redact databases that are shown to reveal PII, and retain clear legal rights to do so;
-Anticipate the cross-correlation of government data with other databases, public and private;
-Provide a convenient mechanism for users to express privacy concerns and ensure proper follow up;

-Sponsor hackathons before the data is released to try to foresee unanticipated uses;
-Negotiate strong privacy protection on data provided to the private sector;

There are also important roles for NGOs, privacy commissioners and the general public in monitoring the release and use of governmental information, and objecting promptly and loudly when Open Data projects appear to violate commonly understood privacy standards.

A good set of suggestions and list of resources on Open Data Policy can be found at the Civic Commons Wiki [23] but there is much more to be done. There will need to be a thoughtful, evolving balance between data openness and personal privacy and this task will be ongoing as new technologies like facial recognition arrive on the scene.

Acknowledgements. The many helpful comments provided by the Program Committee, reviewers, and participants in the *IFIP Summer School on Privacy and Identity*, held in Trento, Italy, Sept 5-9, 2011 were very useful in developing this paper. Travel support from the conference and the University of Calgary is also gratefully acknowledged.

References

[1] http://techpresident.com/blog-entry/and-then-there-were-102-nycs-datamine-glitch (accessed August 26, 2011)

[2] http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf p. 7 (accessed August 26, 2011)

[3] http://data.edmonton.ca/City-Administration/2010-Municipal-Election-Results-Raw-Data-View/gw7p-ee8r (accessed December 30, 2011)

[4] Zijlstra, T.: A court has ordered Slovak NGO Fair-Play Alliance to take down data from their award winning Open Data application, http://www.epsiplatform.eu/news/news/open_data_challenge_winner_ordered_to_remove_certain_data (accessed August 26, 2011)

[5] http://www.squidoo.com/gov-food-stamps#module63313422 (accessed August 26, 2011)

[6] http://we.ideascale.com/a/dtd/Stable-Renters-Public-Scoring-for-Apartments-and-Landlords/106829-12001 (accessed August 26, 2011)

[7] http://www.scmagazineus.com/hurricane-katrina-evacuees-victims-of-data-breach/article/155121/ (accessed August 26, 2011)

[8] http://techpresident.com/blog-entry/open-data-makes-good-advertising-mta (accessed August 26, 2011)

[9] http://ir.ancestry.com/releasedetail.cfm?ReleaseID=552742 (accessed August 26, 2011)

[10] Gormley, M.V.: Oxymoron: Privacy and the Internet, http://www.rootsweb.ancestry.com/~mistclai/privacy.html (accessed August 26, 2011)

[11] US National Academy of Sciences, The Value of Genetic and Genomic Technologies, Washington, DC (2010), http://www.ncbi.nlm.nih.gov/books/NBK52756/pdf/TOC.pdf, with Spiegel's comments in the online discussion http://www.ncbi.nlm.nih.gov/books/NBK52749/

[12] http://www.lossofprivacy.com/index.php/2007/06/ancestrycom-adding-dna-test-results-to-their-site/ (accessed August 26, 2011)

[13] Privacy International, PI files complaint about online DNA genealogical testing firm, https://www.privacyinternational.org/article/pi-files-complaint-about-online-dna-genealogical-testing-firm (accessed August 26)

[14] Ohm, P.: Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. UCLA Law Review 57, 1701 (2010)

[15] Abbott, T.G., Lai, K.J., Lieberman, M.R., Price, E.C.: Browser-Based Attacks on Tor. In: Borisov, N., Golle, P. (eds.) PET 2007. LNCS, vol. 4776, pp. 184–199. Springer, Heidelberg (2007)

[16] Narayanan, A., Shmatikov, V.: Robust De-Anonymization of Large Sparse Datasets. In: IEEE Symposium on Security and Privacy, Oakland, CA, pp. 111–125 (2008)

[17] Albanescius, C.: Netflix Prize Scrapped Over Privacy Concerns, http://www.pcmag.com/article2/0,2817,2361349,00.asp (accessed December 30, 2011)

[18] O'Hara, K., et al.: Avoiding the Jigsaw Effect: Experiences with Ministry of Justice Reoffending Data, research paper, http://eprints.ecs.soton.ac.uk/23072/8/AVOIDINGTHEJIGSAWEFFECT.pdf (accessed December 30, 2011)

[19] quoted on home page, http://opendatachallenge.org (accessed August 26, 2011)

[20] Eaves, D.: blog posting, http://eaves.ca/2010/10/06/how-governments-misunderstand-the-risks-of-open-data/ (accessed December 28, 2011)

[21] Fertik, M., Thompson, D.: Wild West 2.0: How To Protect and Restore Your Online Reputation on the Untamed Social Frontier. In: AMACOM, New York, pp. 50–53 (2010)

[22] http://www.ftc.gov/opa/2006/01/choicepoint.shtm (accessed August 26, 2011)

[23] http://wiki.civiccommons.org/Open_Data_Policy#Privacy_.26_Security (accessed August 26, 2011)