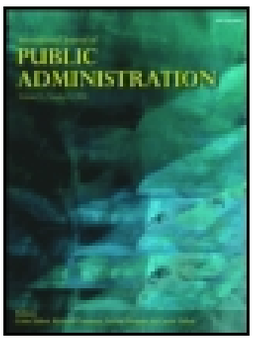

International Journal of Public Administration

ISSN: 0190-0692 (Print) 1532-4265 (Online) Journal homepage: http://www.tandfonline.com/loi/lpad20

Disseminating Public Data as Public Use Files: What Makes a Successful Initiative?

Sergio I. Prada & Lina Martínez

To cite this article: Sergio I. Prada & Lina Martínez (2017): Disseminating Public Data as Public Use Files: What Makes a Successful Initiative?, International Journal of Public Administration, DOI: 10.1080/01900692.2016.1274907

To link to this article: http://dx.doi.org/10.1080/01900692.2016.1274907

Published online: 31 Jan 2017.

INTERNATIONAL JOURNAL OF PUBLIC ADMINISTRATION
http://dx.doi.org/10.1080/01900692.2016.1274907

Routledge
Taylor & Francis Group

Disseminating Public Data as Public Use Files: What Makes a Successful Initiative?

Sergio I. Prada and Lina Martínez

Facultad Ciencias Administrativas y Económicas, & PROESA, Universidad ICESI, Cali, Código, Colombia

ABSTRACT

The current policy emphasis on data-driven decision-making is creating the right incentives for government agencies around the world that have not traditionally disseminated their administrative data to do so. The literature on statistical disclosure control focuses on the technical aspects of a variety of methods designed to protect data confidentiality. There is, however, a void in the literature in regard to what other elements are necessary to create and sustain a successful initiative. This paper examines six case studies of individual-level datasets. It reviews current practice in several domains and summarizes recommendations from expert practitioners including challenges for future initiatives.

KEYWORDS

Case studies; data products; public use files; statistical disclosure control

Introduction

In December 2009, the White House issued the Open Government Directive, requiring statistical agencies to deposit at least three new high value datasets to Data.gov by January 22, 2010 (Orszag, 2009). The Directive emphasized three principles of open government: transparency, participation, and collaboration. The United States has been leading this trend, and since then, many developed nations have joined the effort issuing similar openness data declarations (Huijboom & Van den Broek, 2011). According to Data.gov at least 39 US states, 34 US cities and counties, 41 international countries, and 132 other international regions have developed open data initiatives.

The open data trend benefits both governments and citizens in several ways. First, it enhances transparency and informs citizens about what kinds of data governments are collecting, which translates into more civic engagement and actions to hold governments accountable (Janssen, 2011). Second, it generates mechanisms to improve government performance, leading to better provision of services and outcomes (Virnig & McBean, 2001). Third, it is also very valuable for creating innovative products and services that help businesses and citizens alike (https://www.data.gov/applications). Fourth, it reinforces the right of freedom of information that has been constitutionalized in western countries (Huijboom & Van den Broek, 2011).

However, the implementation of open data initiatives comes with several technical challenges. Before data become available to the public, there are at least three phases that need to be sorted out: i) data generation; ii) collection, aggregation, and processing; and iii) de-identification and privacy protection (Ubaldi, 2013). Despite the growing interest in releasing administrative data to the public, there is no literature on best practices on how to effectively create successful initiatives. This article seeks to contribute to the literate by providing valuable information about several dimensions of the process. This information is useful for both academics and practitioners, especially in an era in which the current policy emphasis on data-driven decision-making is creating the right incentives for federal, national, and regional agencies that have not traditionally disseminated their administrative data to do so.

This article examines six case studies of individual-level datasets sponsored by the US federal government, drawing from the expertise of agencies that have successfully shared administrative data. Statistical agencies in the United States, such as the Census Bureau or the National Center for Health Statistics, are required by law to collect and disseminate statistical information and thus have developed a

CONTACT Sergio I. Prada siprada@icesi.edu.co Facultad Ciencias Administrativas y Económicas, & PROESA, Universidad ICESI, Calle 18 No. 122–135, Cali, Código 760031, Colombia.

systematic approach to also fulfill the requirement of protecting the privacy and confidentiality of individuals included and excluded from data being disseminated.

The intended audience for this article is practitioners interested in creating public use files. While the literature is profuse in statistical and computational methods aimed at de-identifying data or re-identifying data, a void exists in terms of investigating the many other pieces that surround the successful creation of public use files, as well as collecting advice and summarizing recommendations from practitioners. This article is a first step in that direction. For the purpose of this study, successful initiatives are defined as well known within the social sciences in the United States, have been publishing microdata for more than 10 years, and plan to continue doing so.

This article is divided into six sections, of which this introduction is the first. The second section presents a literature review on open government and transparency and the technical challenges of creating public use files. The third section discusses the methodology. The fourth section presents the main results of this investigation, and the fifth closes with a discussion that includes recommendations for a successful PUF initiative and main challenges for future releases.

Background

Government openness and transparency

Government is probably the largest creator and collector of data. However, policy makers and public servants often avoid opening data since new insights might results on critical questions on government performance and decision-making (Janssen, Charalabidis, & Zuiderwijk, 2012). Many coincide that President Obama's inaugural flagship open data declaration was the first step on making attempts on generating mechanisms to create an open and more transparent governments on developed countries (Harper, 2012; McDermott, 2010; Peled, 2011). As a consequence, recent years have experienced trends toward creating new public data sources granting a greater access to information and promotion of government transparency (Bertot, Jaeger, & Grimes, 2010). The creation of an open data culture positively impacts several domains of government performance and its relationship with a wider environment. Even though this is an emerging field, there is some research that documents how open data create an added value not only for government functioning but also for civic participation.

For instance, it has been documented how open data improve government functioning through major accountability, transparency, and democratic control (Anderson, 2009; Shim & Eom, 2009). Open data have been perceived as a powerful tool to rise levels of public trust and perceived responsiveness of government actions. This has been reported in countries like United States, United Kingdom, Australia, Denmark, and Spain where open data programs have been in place (Huijboom & Van den Broek, 2011).

Other authors argue that open data promote social participation and engagement (Davies, 2010). Access to information enables individuals to make better decisions, improving their quality of life and promoting civic engagement in public domains. The expanding use of technologies readily available in portable devises combined with the powerful force of data is creating an environment in which citizens are not only passive consumers of content, but also active contributors (Ubaldi, 2013). There are several examples of the use of public data that have been developed and used by civilians that allowed better decision-making in health care (Blue Button, US), energy consumption (Green Button, US), fuel economy (choosemyplate.com, US), or political control (TheyWorkForYou.com, UK) (Hogge, 2010; Howard, 2012; Thaler & Tucker, 2013). Data.gov lists 192,119 datasets available in topics such as agriculture, business, climate, consumer, ecosystems, education, energy, finance, health, local government, manufacturing, ocean, public safety, and science and research.

Efficiency and effectiveness in government services can be also impacted by the release of data to the public. The interactive nature of open data facilitates two-way interaction between government and civil society, expanding services beyond business hours or the availability of personnel for face-to-face interactions. This facilitates service delivery and increases responsiveness to citizens generating higher trust on government (Gore, 1993; West, 2004). Additional benefits include reductions in workload of public servants (since information requests can be addressed with public data), reduction in paper work, and transaction costs (Tat-Kei, 2002). Using the benefits of open data, Bristol City Council, for instance, is introducing open government data catalogues, and by doing, this is reducing on 15 times the cost of a typical service transaction (Ubaldi, 2013).

Lastly, academia also directly benefits from open data initiatives (Bauer & Kaltenböck, 2011). Academic research depends heavily on data and the cost of data collection, and processing is prohibitively high for

academics, especially for students and young scholars. By the release of open data sources, academics can conduct original research that not only benefit their academic fields, but also provide information to policy-makers to make better informed decisions.

Privacy and confidentiality

Concerns about privacy and confidentiality in governmental efforts to collect and disseminate information are not new. As a review by Anderson and Seltzer (2009) suggests, "The roots of the modern concept of federal statistical confidentiality can be traced directly back to the late nineteenth century." Notwithstanding this history, the literature on statistical disclosure control (SDC) methods is fairly recent by modern standards (Dalenius, 1977 is considered the seminal paper).

As discussed in Prada et al. (2011), the literature on disclosure limitation techniques and their achievements is a new but growing field and has been dominated by statisticians and computer scientists. This literature discusses in detail the merits and flaws of the different techniques—advanced in particular by national statistical agencies in the developed world (e.g., US Census Bureau, Statistics Netherlands)—designed to protect the privacy of the analytic units in the data while retaining as much information as possible so as to avoid distorting the utility of the data.

In 1998, for example, the Journal of Official Statistics devoted an entire issue to statistical disclosure control. Since then, the literature has spread to a range of journals in a variety of fields, examples of which include the American Journal of Epidemiology, Artificial Intelligence in Medicine, IEEE Transactions on Knowledge and Data Engineering, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Journal of the American Medical Informatics Association, and the Journal of the American Statistical Association. The most recent addition to the list is the Journal of Privacy and Confidentiality, supported in part by the Department of Statistics at Carnegie Mellon University.

In addition, every 2 years UNESCO sponsors an international conference (i.e., Privacy in Statistical Databases) that gathers worldwide experts from different disciplines to discuss current issues in the field. Proceedings are published by Springer in the series Lectures Notes in Computer Science (Domingo-Ferrer, 2001; Domingo-Ferrer & Franconi, 2006; Domingo-Ferrer & Magkos, 2010; Domingo-Ferrer & Saygin, 2008; Domingo-Ferrer & Torra, 2004; EC, 1998).

Despite the growth in the technical aspects of SDC, the literature is silent in regard to the resources needed, challenges associated with, and lessons learned from the creation of public use files. While some papers report on surveys conducted on national statistical agencies (Felso, Theeuwes, & Wagner, 2001), on the experience of particular agencies (Longhurst et al., 2007; Mas & Prado, 2007; Pinto, 2005), or on particular datasets (Drechsler & Reiter, 2009), they too tend to concentrate on the SDC methods in use.

O'Rourke et al. (2006) stands as one of the few papers discussing best practices. These authors suggest best practices for the various groups affected by publicly distributed research data. For data users, "it is the responsibility of the data user to investigate the disclosure limitation measures applied to the data and their impact on the desired analyses." For investigators facing demands to share data, they "should begin planning early in the research process," this to avoid mistakes such as disclosing information in papers, posters, and presentations that can be used later by users to re-identify records in the PUF. For data distributors, forming a cross-disciplinary disclosure reviews committee, including statisticians, disclosure experts, programming experts, and researchers with experience in the content area of the data, to provide a full picture of the intended uses of the data and thus prioritize among the available data to be made public. Lastly, O'Rourke et al. (2006) suggest as a best practice for data producers, to formulate SDC solutions that are amenable to replication, in particular automating routines in statistical packages such as SPSS or SAS.

Methodology

Two criteria governed inclusion in the study: (1) a project had to provide individual (person)-level data, and (2) the data had to be of sufficient complexity to provide useful insights regarding methods used for de-identification. A convenience sample methodology is used because there is no registry of PUFs and because it allows to study well-known agencies such as the Census Bureau. It is acknowledged, however, that a non-probability sample is biased and that the results may not be generalized in the statistical sense.

Search starting point was a list of federal organizations in the United States with a proven record of releasing data, among them, Agency for Healthcare Research and Quality, National Center for Health Statistics, Centers for Disease Control and Prevention, National Institute on Aging, Census Bureau, Department of Education (DoE), and Department of Homeland Security (DHS). This was

followed by searching the federal government web sites Data.gov and FedStats.gov on terms such as "public use file." Finally, general web search engines were used (e.g., Google, Bing), again using variations on terms such as "public use file."

An initial set of nine initiatives were identified as potential representative case studies. The initial information analyzed on these initiatives was obtained from project web sites and through phone calls with agency personnel. All nine initiatives were invited to participate. The list of initiatives that accepted to participate is included in Table 1. Although the initial focus was on specific initiatives such as the OASDI benefits and earnings PUF 2004, earlier in the process it was also clear that most of the themes included in the survey applied to all PUF initiatives within a given federal agency, for instance the various PUFs produced by CDC's National Center for Health Statistics (i.e., National Health Interview Survey, National Hospital Discharge Survey, etc.).

Interview guide was designed and developed around five main themes that represent the lifecycle of a PUF (Hundepool et al., 2010; O'Rourke et al., 2006) and based on reviews to the literature on privacy and confidentiality (Prada et al., 2011): background information, creating (and de-identifying) the file, assessing the re-identification risk and analytic utility of the file, accessing the file, and assessing the initiative. Questions were validated with experts. Finally, the interview guide was modified on a case-by-case basis to account for case-specific particularities. Additional information not explicitly included in the guide was gathered by probing on topics as they emerged during interviews

Six interviews were conducted, one with the technical head of each initiative, all by phone with one exception (which was an in-person interview). Interviewees were initially contacted by mail and then scheduled dates and times through telephone communication and/or email. The interview guide was sent to interviewees in advance and informed them at the same time the intention to audiotape the interview. In two occasions, questionnaires were sent back with notes, answers to questions, and useful references to further explore. At the beginning of each interview, a written statement was read on the purpose of the interview as well as a request for authorization to audio-record. All participants consented to being audio-taped. Extensive notes were taken during the interview. Interviews were not transcribe, but audios were marked to access quickly different sections of the interview.

Four of the six case studies are public use files (PUF) and two are non-public use files (NPUF) (SEER and NHES). For the purposes of this article, a PUF is

Table 1. Summary characteristics of selected initiatives.

Initiative name	Sponsor	Data access method(s)	Features
American Community Survey (ACS)	Census Bureau	Web download Onsite at Census Bureau Research Data Centers	Experienced issue with age perturbation causing invalid gender ratios Restricted data available at Research Data Centers: network of 10 centers around US Started: 70s
National Center for Health Statistics (NCHS)	Centers for Disease Control and Prevention	Online analysis tools FTP download Onsite at Research Data Center	Confidentiality Officer Disclosure Review Board Restricted-use files linked to CMS and other agency data Restricted-use access onsite at Research Data Center (RDC) Researchers can submit queries electronically to be executed in the RDC with output returned by e-mail Started: 60s
Old-Age, Survivors, and Disability Insurance, Benefits, and Earnings (OASDI)	Social Security Administration	Web download	Data linked to SSA Master Beneficiary Database Started: 90s
National Household Education Survey (NHES)	National Center for Education Statistics	Web download Certified mail (restricted-use files)	Disclosure Review Board Offsite restricted-use data "lending" Started: 90s
Surveillance Epidemiology and End Results (SEER)	National Cancer Institute	Via CDC's SEER*Stat software DVD Web download Encrypted restricted-use file	Multiple demographic variables in data Linked public-use data Restricted-use data linked to CMS data Restricted-use data provided for offsite use Started: 70s
Healthcare Cost and Utilization Project (HCUP)	Agency for Healthcare Research and Quality	Purchase through HCUP Central Distributor	Largest collection of longitudinal hospital care data in the United States More than 100 clinical and nonclinical variables for each hospital stay Required Web-based data user training course Started: 90s

defined as a dataset characterized by free and unrestricted access to any user. Consequently, NPUFs are files characterized by any type of access restrictions, in particular restrictions that oblige users to reveal their identity and use intentions. Typically, a NPUF demands signed data use agreements (DUAs) before having access to data.

Results

Either by mandate (e.g., NCHS, Census) or by election (e.g., SSA) publicly funded agencies, and other organizations provide PUFs with the objective of disseminating information to increase public understanding of topic areas. Every single initiative for which an interview was completed felt strongly that its PUFs are an excellent way to increase the potential of its data.

These organizations have kept the pledge of privacy and confidentiality given to the public by using both simple and sophisticated statistical disclosure avoidance techniques. All cases included in this study report having a strong record of success in protecting PUFs against disclosure risk. This is due in part to the amount of resources devoted to the task, but also to the development of a tiered system of access according to the level and detail of the data required. The increase in computational power and the explosion of data available for free or at very low cost was frequently mentioned as an important threat to PUF creation moving forward. It is worth noticing that a variation was found among initiatives in some aspects of the PUF creation (i.e., disclosure limitation methods and software) and consensus in others (i.e., statistical weights). Given that all initiatives are successful and their data widely used, variation suggests to practitioners that there are valid alternatives to choose from, while consensus strongly suggest that other courses of action must not be taken.

Disclosure risk standard

Among the four interviewed institutions that provide PUFs without restrictions, it was found that the Confidentiality and Data Access Committee's (CDAC) "Checklist on Disclosure Potential of Proposed Data Releases"[1] is widely used as the tool to assess disclosure risk of proposed data. In contrast, it was found that the CDAC Checklist is not used by the initiatives that do not release PUFs (i.e., the two that release research files subject to DUAs and other restrictions). Among the latter two initiatives, one declared the process to be a "judgment call" based on its knowledge of the data and data users; the other relies on the individual risk analysis made at the source.

None of the initiatives uses a theoretically established risk framework (individual risk methodology, GenMASSC,[2] etc.). Similarly, it was not found a formal definition of a "safe threshold" (beyond the HIPAA "Safe Harbor" method) below which a candidate file could be considered ready for release. A "safe threshold" refers to the percentage of records at risk of re-identification within the file. Decisions are made on a case by case basis.

Disclosure review board

It was found that, following the recommendations in Statistical Policy Working Paper # 22, the PUF case study initiatives have centralized their review of disclosure-limited data products. A review panel, team, or board have been established in each, and a common practice is for a completed Checklist memo to be submitted to that Disclosure Review Board (DRB) or other similar panel for review. Interestingly, DRBs are nonexistent for the initiatives interviewed that do not release PUFs. None of the documents reviewed by DRBs or any of the DRB decisions made are available to researchers, nor is any information released by NPUF institutions on data disclosure avoidance steps taken. Interestingly, in one case in which the PUF includes information from two different agencies, it is required that both DRBs approve the file.

Geographic information

It was found that detailed geographic indicators are generally stripped from the PUFs studied.[3] This is a recommendation that follows CDAC's Checklist suggestions, as geography is a key factor in enabling identification. As the CDAC Checklist states, "while few respondents could likely be identified within a single state, more respondents—especially those with rare and visible reported characteristics—could be identified within a county or other small geographic area." In addition to the direct naming of geographic areas, the Checklist contains alerts on geographic information

[1]Available at http://www.fcsm.gov/committees/cdac/cdac.html.
[2]Generalized Micro-Agglomeration, Substitution, Subsampling, and Calibration.
[3]NHES includes geographic indicators to the Census's level four region. Some Census products present information at the Public Use Microdata Area (PUMA) defined after a minimum threshold population of 100,000.

that may be "implicitly" contained in details concerning sample units of design or variables with a geographic reference.

Geographic indicators are not stripped from the case study initiatives that do not release PUFs. For instance, in the case of SEER, state and county identifiers are available, and there are no minimum population requirements. In the case of HCUP, some states release zip codes at the 3-digit and some even at the 5-digit level.

Statistical weights

The statistical final weight is the number of people in the population that the sampled person represents. PUFs whose data are collected through surveys only release final weights. They do not release the components that make up the final weight, because these may be indicative of geographic areas. They usually do not release the actual primary sample unit (PSU) and strata identifiers because that can be risky as well. Instead, they provide pseudo-strata and pseudo-PSU variables containing less information than the replicates. Replicate and/or bootstrap weights are provided for variance estimation. In the case of the Census ACS, each PUF file contains a weighting factor for each population (PWGTP) and housing unit record (WGTP) to obtain full population estimates. These weights are not adjusted after disclosure limitation methods are applied.

Professional expertise required

It was found that the level of expertise necessary to create PUFs and NPUFs is high as well as interdisciplinary and scarce. Several senior statisticians and mathematicians are involved in the PUF creation process. Additional experts are also involved, including DRB members and other senior staff who review the files before release, data analysis experts such as economists and epidemiologists (depending on the data collected), and SDC experts at the firms contracted to create these PUFs. An interesting finding raised in one of the interviews is the scarcity of professionals with interest or background in SDC methods. The problem seems to be rooted in a lack of interest for the topic in graduate schools around the United States.

As far as internal personnel procedures, it was found strict procedures in some instances. For example, NCHS requires each new employee or contractor to watch a confidentiality video, sign a nondisclosure affidavit (which spells out the particulars of the laws covering their conduct while under contract), and review documents and material dealing with their responsibilities with respect to confidential information while working at or for the agency. On departure from employment, the individual is also required to undergo an equally well-defined exit process.

Identifying risk

According to interviewees, the main criterion used to identify potentially identifiable records in PUF initiatives is whether a record is unique with respect to a combination of key variables. This method is called *k-anonymity* in the SDC literature (Skinner & Elliot, 2002). Typically, demographic indicators such as gender, age, race, education, marital status, number of children, geographical location are used to define combinations. Which variables and the exact nature of the combination(s) used is confidential information. This criterion is shared by both PUF and non-PUF initiatives.

Disclosure limitation methods and software

It was found no preference among the many methods available for limiting data. The agencies included in the interviews use various techniques including coarsening, suppression, top and bottom coding, rounding, random rounding, and data swapping. The decision on which method to use is typically case specific and even variable specific. It also depends on internal deliberation. Similarly, each initiative has its own algorithm for disclosure avoidance[4] and all refrain from using disclosure avoidance software. The reason given is the possibility of reverse engineering. As expected, PUF initiatives use more sophisticated masking techniques than non-PUF initiatives.

Risk of match to other datasets

It was also found that agencies take into consideration other files available to the public (e.g., online) when evaluating the risk of disclosure. This is true for both external datasets and previously released data (e.g., reports, tables) from the same source. These comparative activities are conducted both in-house and/or by outside contractors (particularly data security firms). The degree and level of sophistication of these activities vary by initiative, with Census, NCHS, and NCES exemplifying initiatives that are highly concerned and highly cautious and non-PUF initiatives much less so.

[4]A similar result was found by Felso et al. (2001) in their review.

Re-identification certification

None of the cases studied has a re-identification certification procedure in place. Re-identification refers to the possibility of an intruder being able to identify someone in a PUF and learn information that he/she would not be able to learn otherwise. Re-identification certification tests the vulnerability of PUFs to external sources and should be conducted by a third party.

Data utility

It was found little information on what is done regarding data utility (e.g., an assessment of information loss after applying SDC methods). While the interviewees for the PUF initiatives reported that they do conduct such analyses, these documents are not available to researchers. Although access to these documents was requested, they were not made available. The data utility analyses concentrate on comparisons of means and distribution tests before and after disclosure treatment to determine their effect on pretreatment statistical characteristics of the data. In-house and consultant statisticians also do multivariate tests to study effects on relationships among multiple variables. Despite the limited nature of these tests, all PUF-initiative interviewees highlighted the importance of data utility analyses and, in particular, the coordination of such analyses between statisticians and program directors (topic experts) to avoid unnecessary distortions in the data to be released. These concerns were less pronounced for non-PUF initiatives.

Data access

According to the interviewees, access to data is granted via online query systems, PUFs, licenses, and onsite at Research Data Centers (RDCs). It was found no access method to be preferred over another. Case study institutions have adopted such methods in response to user demands. The degree to which access to detailed information is allowed (of the type that is not available in PUFs) depends on both the pertinence of the research question, and the degree by which an individual (or individuals) can be held accountable for the use of the data. Even though the availability of data on the web is growing at increasing rates, in interviews conducted there were not identified plans to stop producing PUFs.

Confidentiality requirements to users

It was also found that the degree to which PUF initiatives warn users-to-be on confidentiality issues (such as to explicitly avoid actions aimed at re-identifying individuals) varies greatly, from short statements on the webpage where the data are located to the acknowledgment of online DUAs before download. The two NPUF initiatives required users to sign and submit DUAs for agency revision and approval before granting any data access.

Documentation

All initiatives investigated were similar in stress documentation as a key success factor. Descriptions of data fields are provided in data dictionaries. Descriptions cover the content of each field, method of presentation, and disclosure avoidance steps taken to provide confidentiality. However, PUF initiatives are cautious of not revealing unnecessary information, and therefore, the language used is rather generic. The review of the documentation available online for all the initiatives suggested that the topic of what has been done to protect the data is not discussed at length nor is easy to find in documentation.

Communication channels and user feedback

There are three main channels of two-way communication between each of these initiatives and its users: e-mail, phone numbers, and personal communication at national conferences. Regarding diffusion of data releases, the main channels are via listserv and notifications on their respective websites.

As for user feedback, it was found that data are not generally modified in response to user demands, as these typically involve requests for more detailed information. However, when errors are discovered, either by staff or by users, corrections are made, and the PUFs are re-released.

Social media channels such as Facebook and Twitter were not in use at the moment of the interviews. However, the review of websites suggests that these agencies are moving rapidly in that direction. The Centers for Disease Control and Prevention (CDC), the US Census Bureau, and the Social Security Administration can be followed on Facebook and Twitter.

Strong success record

Despite the variety of approaches to guarantee privacy and confidentiality (methods and access) among the case studies studied, it was found a strong overall record of success. Perhaps the only problem cited, in NPUFs, is that occasionally researchers publish papers

with a small-count cell in a table (e.g., less than 10 individuals, less than 3 institutions). In those cases, researchers are asked to immediately take the necessary steps to remove or retrieve such information from where it has been published. According to the interviewees, there has been no reported breach of confidentiality to date in any of the cases studied. This contrasts with the Felso et al. (2001) survey in which three of their respondent agencies had experienced disclosure-related problems in the recent past. It is not possible to know from their paper whether these cases pertain to US agencies, international agencies, or a combination of the two.

Discussion

Citizens, governments, researchers, and businesses are the likely beneficiaries of the open data government movement. Open data initiatives seek to make administrative data publicly available in a way that access and use are guaranteed without having to make a request or pay a fee for access to the government. This is done by publishing datasets directly on the governmental agency's website or on "open data portals" such as Data.gov in the United States or Data.gov.uk in the United Kingdom. Key to the sustainability and credibility of these initiatives is to preserve the confidentiality and privacy of personally identifiable information. However, it is now clear that success in this endeavor involves not only issues of privacy but also utility, that is, reaching out to the public to solicit input and to encourage use of the datasets for the public good.

While publishing datasets directly on websites is not new for statistical agencies or other producers of data, it can be argued that the US 2009 President's Open Government Directive was a key in promoting a movement that is now global, with at least 41 countries starting to make administrative data public. The 2009 policy has been revised and improved by Office of Management and Budget 2013 Open Data Policy entitled "Managing Information as an Asset," pushing further the movement by requiring agencies to collect or create information in a way that supports easy information processing for anyone and dissemination activities; to ensure information stewardship; and to building information systems in a way that maximizes interoperability and information accessibility. As the movement grows, there is a need for a better understanding of the drivers of success.

It was found that to maximize the likelihood of success, PUF initiatives should allow enough time for development, begin with limited scope, gain experience, keep in focus the objective that motivated the project when facing challenges and obstacles, surround themselves by highly skilled SDC experts and program content analysts, allow enough time for rigorous and constant risk analysis, produce comprehensive documentation, provide users with more than raw data to spark and keep the interest in the files, have enough two-way channel communications, and design a tiered system of access to better respond to users requirements.

A main objective of the study was to collect recommendations for PUF development, in particular for data that is highly sensitive (i.e., health status and healthcare utilization). It is worth noting that the question on the creation of PUFs was asked when the source of data is administrative records. Challenges associated with the creation of PUFs may be different depending on the source of data, be it from a survey or from administrative records. Several characteristics make survey data different: First, consent to include information in PUFs (provided confidentiality is protected) is usually explicitly asked and granted by respondents; second, survey data are usually comprised of small samples, where each individual represents hundreds if not thousands of people; third, while both are subject to reporting error, it is easier to track and compare administrative data than oral responses to a survey (e.g., income reported as $25,000 versus payroll records showing $25,550 as income).

For the purpose of this study, successful initiatives were defined as well known within the social sciences and have been publishing microdata for more than 10 years and plan to continue doing so. Although successful, all of these initiatives face important challenges ahead, according to the review and personal interviews. The first challenge is a lack of human capital to strengthen re-identification risk analyses due to budgetary constraints. The re-identification threat is another major problem for PUFs given the increasing pace at which personal information is being collected, stored, and sold or shared in today's society. While it is true that the risk of re-identification is low, because it takes someone highly skilled in SDC methods and computer programming with the time, intention, and money to disclose the identity of a few records, what data producers fear the most is public embarrassment and a possible lack of trust by the public.

Finally, it is important noting that amid the recommendations and suggestions collected in this article, and variation among initiatives in some aspects of the PUF creation (i.e., disclosure limitation methods and software) and consensus in others (i.e., statistical weights) was found. Given that all initiatives are successful in protecting privacy and confidentiality and their data widely used,

variation suggests to practitioners that there are valid alternatives to choose from, while consensus strongly suggest that other courses of action must not be taken.

Acknowledgments

We are grateful to all the individuals who graciously shared their time and expertise and who provided most of the data and information reported here. Mark Boward led the initial stages of this investigation.

References

Anderson, M., & Seltzer, W. (2009). Federal statistical confidentiality and business data: Twentieth century challenges and continuing issues. *Journal of Privacy and Confidentiality*, *1*(1), 7–52.

Anderson, T. B. (2009). E-government as an anti-corruption strategy. *Information Economics and Policy*, *21*, 201–210.

Bauer, F., & Kaltenböck, M. (2011). *Linked open data: The essentials*. Vienna, Austria: Edition mono/monochrom.

Bertot, J. C., Jaeger, P. T., & Grimes, J. M. (2010). Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies. *Government Information Quarterly*, *27*(3), 264–271. doi:10.1016/j.giq.2010.03.001

Dalenius, T. (1977). Toward a methodology for statistical disclosure control. *Statistik Tidskrift*, *15*, 429–444.

Davies, T. (2010). Open data, democracy and public sector reform. *A look at open government data use from data*. MSc Dissertation submitted for examination in Social Science of the Internet, at the University of Oxford, Summer. Retrieved from http://www.opendataimpacts.net/report/wp-content/uploads/2010/08/How-is-open-government-data-being-used-in-practice.pdf

Domingo-Ferrer, J. (2001). *Privacy in statistical databases*. LNCS 2316. Berlin, Germany: Springer-Verlag.

Domingo-Ferrer, J., & Franconi, L. (2006). *Privacy in statistical databases*. LNCS 4302. Berlin, Germany: Springer-Verlag.

Domingo-Ferrer, J., & Magkos, E. (2010). *Privacy in statistical databases*. LNCS 6344. Berlin, Germany: Springer-Verlag.

Domingo-Ferrer, J., & Saygin, Y. (2008). *Privacy in statistical databases*. LNCS 5262. Berlin, Germany: Springer-Verlag.

Domingo-Ferrer, J., & Torra, V. (2004). *Privacy in statistical databases*. LNCS 3050. Berlin, Germany: Springer-Verlag.

Drechsler, J., & Reiter, J. (2009). Disclosure risk and data utility for partially synthetic data: An empirical study using the German IAB establishment survey. *Journal of Official Statistics*, *25*(4), 589–603.

European Communities. (1998). *Statistical data protection '98*. Proceeding by Office of Official Publications, Statistical Office of the European Communities Lisbon.

Felso, F., Theeuwes, J., & Wagner, G. (2001). Disclosure limitation methods in use: Results of a survey. In P. Doyle, J. Lane, J. Theeuwes, & L. Zayats (Eds.), *Confidentiality, disclosure, and data access: Theory and practical application for statistical agencies*. Chapter 2. Washington, DC: Urban Institute.

Gore, A. (1993). *From red tape to results: Creating a government that works better & costs less*. Report of the National Performance Review. U.S. Government Printing Office, Superintendent of Documents, Washington, DC, USA.

Harper, J. (2012). Grading the government's data publication practices. *Cato Institute Policy Analysis*, *711*, 2.

Hogge, B. (2010). Open data study. *A report commissioned by the Transparency and Accountability Initiative*. Retrieved from http://www.soros.org/initiatives/information/focus/communication/articles_publications/publications/open-data-study-20100519

Howard, A. (2012). *What is smart disclosure?* Radar O´reilly. Retrieved January, 2016. http://radar.oreilly.com/2012/04/what-is-smart-disclosure.html

Huijboom, N., & Van den Broek, T. (2011). Open data: An international comparison of strategies. *European Journal of Epractice*, *12*(1), 4–16.

Hundepool, A. et al. (2010). *Handbook on statistical disclosure control*. ESSNet S D C. Retrieved from http://neon.vb.cbs.nl/casc/.%5Ccasc%5Chandbook.htm

Janssen, K. (2011). The influence of the PSI directive on open government data: An overview of recent developments. *Government Information Quarterly*, *28*(4), 446–456. doi:10.1016/j.giq.2011.01.004

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, *29*(4), 258–268. doi:10.1080/10580530.2012.716740

Longhurst, J., et al. (2007, December 17–19). The review of the dissemination of health statistics in England. In *Work session on statistical data confidentiality*. Manchester, UK: Eurostat, Methodologies and Working Papers.

Mas, M., & Prado, C. (2007, December 17–19). Dealing with Confidentiality in Dissemination: The experience of the Basque Statistics Office. In *Work Session on statistical data confidentiality*. Manchester, UK: Eurostat, Methodologies and Working Papers.

McDermott, P. (2010). Building open government. *Government Information Quarterly*, 27(4), 401–413. doi:10.1016/j.giq.2010.07.002

O'Rourke, J. M., Roehrig, S., Heeringa, S. G., Reed, B. G., Birdsall, W. C., Overcashier, M., & Zidar, K. (2006). Solving problems of disclosure risk while retaining key analytic uses of publicly released microdata. *Journal of Empirical Research on Human Research Ethics: JERHRE*, *1*(3), 63–84. doi:10.1525/jer.2006.1.3.63

Orszag, P. (2009, December 8). *Memorandum for the heads of executive departments and agencies*. Retrieved from http://www.whitehouse.gov/open/documents/open-government-directive

Peled, A. (2011). When transparency and collaboration collide: The USA open data program. *Journal of the American Society for Information Science and Technology*, *62*(11), 2085–2094. doi:10.1002/asi.v62.11

Pinto, A. (2005, November 9–11). Statistics and Confidentiality in the Portuguese Case. In *Work Session on statistical data confidentiality*. Geneva, Switzerland: Eurostat, Methodologies and Working Papers.

Prada, S., Gonzalez-Martinez, C., Borton, J., Fernandes-Huessy, J., Holden, C., Hair, E., & Mulcahy, A. T. (2011).

Avoiding disclosure of individually identifiable health information in public use files: A literature review. *SAGE Open*, *1*(3). doi:10.1177/2158244011431279

Shim, D. C., & Eom, T. H. (2009). Anticorruption effects of information and communication technology (ICT) and social capital. *International Review of Administrative Sciences*, *75*, 99–116.

Skinner, C. J., & Elliot, M. J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society Series B*, *64*, 855–867. doi:10.1111/rssb.2002.64.issue-4

Tat-Kei, H. A. (2002). Reinventing local governments and the e-government initiative. *Public Administration Review*, *62* (4), 434–444. doi:10.1111/puar.2002.62.issue-4

Thaler, R. H., & Tucker, W. (2013). Smarter information, smarter consumers. *Harvard Business Review*, *91*(1), 44–54.

Ubaldi, B. (2013). *Open government data: Towards empirical analysis of open government data initiatives*. OECD Working Papers on Public Governance, No. 22. Paris, France: OECD Publishing. Retrieved from http://dx.doi.org/10.1787/5k46bj4f03s7-en

Virnig, B., & McBean, M. (2001). Administrative data for public health surveillance and planning. *Annual Review of Public Health*, *22*(1), 213–230. doi:10.1146/annurev.publhealth.22.1.213

West, D. M. (2004). E-government and the transformation of service delivery and citizen attitudes. *Public Administration Review*, *64*(1), 15–27. doi:10.1111/puar.2004.64.issue-1