A Realistic Look at Open Data

Sharon S. Dawes
Center for Technology in Government at University at Albany/SUNY

The basic assumption of the open data movement is that more intensive and creative use of information and technology can improve policy-making and generate new forms of public value.. These efforts can focus on education, public health, transportation, environmental stewardship, economic development and many other areas. Each application has its own substantive considerations, forms of expertise, and interests. Ironically, information is often treated as a black box in the open data movement. Stakeholders, analytical techniques, and technology tools all receive considerable attention. But information is often seen as a given, used uncritically, and trusted without examination. However, he very kind of data that is now being released from administrative systems as "open data" was collected or created for other purposes. It has undeniable potential value, but it also substantial risks for validity, relevance, and trust..

Government data for policy analysis and evaluation

Government is the major source of open data. While the day to day value of this information comes from its use in specific government programs and services, the societal value of these information resources is derived primarily from unpredicted and flexible uses of the data content by all stakeholders (Bellamy and Taylor, 1998).

Certain sources of government data are commonly used by external analysts. These include government agencies that have the formal responsibility and professional skill to collect, manage, maintain and disseminate data for public use. They represent a long-standing government commitment to collect and provide certain kinds of social, economic, and demographic information to the public. Census, economic, and other formal statistical series are well-understood and readily usable because they apply the standards of social science research in data collection and management. They collect well-defined data on specific topics using well-documented methodologies that follow a logical design. The data files are managed, maintained and preserved according to explicit plans that include formal rules for access, security, and confidentiality.

The explosion in so-called "administrative data" is now attracting great attention for its potential value both inside and outside government. Administrative data reflects the operations of the government programs. The automation of government activities and the advent of electronic government services has brought with it not only online convenience, but also the ability to capture enormous amounts of digital information about services, individuals, organizations, and transactions (Snellen, 2005; Bekkers, 1998; Frissen, 1992). E-mails in government records systems, for example, number literally in the billions. Transactional data reveals the workflow activities of case management systems or customer service exchanges, much of it collected in real-time. Government-deployed sensor networks gather data about transport, air quality and other topics.

The open government movement is making tens of thousands of these administrative data sets available to the public through programs like Data.gov in the US whose purpose is to make more data from federal government agencies readily accessible for external use. Its central web portal provides electronic access to raw, machine-readable information about government finances, program performance, trends, transactions, and decisions. The goal is to allow people and organizations outside government to find, download, analyze, compare, integrate, and combine these datasets with other information in ways that provide value to the public. And this phenomenon is not limited to the federal level. States and municipalities are experiencing similar growth in data holdings and taking advantage of new technologies to gather and analyze data from routine operations.

However, because administrative data are not created with external or unplanned use in mind, they are not managed in the precise and structured way that we have come to expect from the government census agency. They offer new opportunities , but they are also more difficult to use and interpret and therefore more subject to misunderstanding and misuse (Dawes, 1996; Ballou and Tayi, 1999).

Sources of information problems

Information problems stem from a variety of causes that both government information providers and independent analysts need to understand.

Conventional wisdom

A set of common beliefs and unstated assumptions are often substituted for critical consideration of information. In the positivist tradition of the social sciences, for example, quantitative data is automatically preferred or assumed to have better quality than qualitative data (Heinrich, 2012). In her book about the challenges of the government performance movement, Beryl Radin (2006) identifies several common beliefs about information that lead to unrealistic, even false, expectations and results. These include assumptions that needed information is available and sufficient, objectively neutral, understandable, and relevant to the task of evaluation. Left unchallenged, they compromise all forms of program assessment and policy analysis.

Recent open government initiatives like Data.gov, present similar problematic beliefs. They convey an unstated assumption that large, structured "raw" data sets are intrinsically better than processed data, and that data in electronic form suitable for delivery on the Internet is superior to other forms and formats for information. Thus the "low-hanging fruit" of available machine-readable raw datasets receives more attention than better defined and potentially more suitable traditional datasets that reflect some interim processing or cannot easily be posted on the Web. Meijer (2009) argues that increased publication of raw data can have undesirable effects. Instead of providing more transparency, the consumption of this data can actually threaten public trust because the data is removed from shared social experience, takes an overly structured, predominately numerical form, and directs attention to narrow and sometimes irrelevant (but quantifiable) concerns.

Provenance

Much "open data" emerges from activities and contexts that are far different in purpose, context, and time from its eventual use. Taken out of context, the data loses meaning, relevance, and usability. Although the public may be offered thousands of data sets from one convenient web address, these information resources are actually distributed among different government organizations, locations, and custodians. The datasets are defined and collected in different ways by different programs and organizations. They come from a variety of different systems and processes and represent different time frames and geographic units or other essential characteristics. Most come from existing information systems that were designed for specific operational purposes. Few were created with public use in mind.

Metadata is essential to understand this data but unfortunately, it receives little attention in most organizations. An administrative or operational dataset is usually defined at the point of creation in just enough detail to support the people who operate the system or use the data directly. As the underlying data set or system changes over time, corresponding maintenance of metadata is a low priority activity. The idea of fully describing the data for the benefit of some unknown future user is hardly considered at all. Even when metadata exists in reasonably complete form, it often fails to capture contextual knowledge that can have a powerful effect on its quality and usability.

Practices

Research shows that in order to understand data, one needs to understand the processes that produce the data (Dawes, et al., 2004). Data collection, management, access, and dissemination practices all have strong effects on the extent to which datasets are valid, sufficient, or appropriate for policy analysis or any other use (Dawes and Pardo, 2006). Data collection schemes may generate weekly, monthly, annual, or sporadic updates. Data definitions and content could change from one data collection cycle to the next. Some data sets may go through a routine quality assurance (QA) process, others do not. Some quality assurance processes are rigorous, others are superficial. Some data sets are created from scratch, others are byproducts of administrative processes, still others may be composites of multiple data sources, each with their own data management practices.

Data sets may be readily accessible to internal and external users, or require some application or authorization process. They may be actively disseminated without cost or made available only on request or for a fee. Access may be limited to certain subsets of data or limited time periods. In addition, data formats are most likely the ones that are suitable and feasible for the organization that creates and manages the data and may not be flexible enough to suit other users with different capabilities and other interests.

Data quality and fitness for use

Given the practical realities outlined above, we can see that even if government information resources are well-defined and managed, substantial problems for use cannot be avoided. The term "data quality" is generally used to mean "accuracy," but research studies identify multiple aspects of information quality that go well beyond simple accuracy of the data. Wang and Strong (1996) adopt the concept of "fitness for use, "considering both subjective perceptions and objective assessments, all of which have a bearing on the extent to which users are willing and able to use information.

- *Intrinsic quality* most closely matches traditional notions of information quality. It includes accuracy and objectivity, but also involves believability and the reputation of the data source.
- *Contextual quality* refers to the context of the task for which the data will be used. It includes considerations of timeliness, relevancy, completeness, sufficiency, and value-added to the user. Often there are trade-offs among these characteristics, for example, between timeliness and completeness (Ballou and Pazer, 1995).
- *Representational quality* relates to meaning and format. It requires that data be not only concise and consistent in format but also interpretable and easy to understand.
- *Accessibility* comprises ease and means of access as well as access security.

The current emphasis on open data plus the evolving capability of technological tools offer many opportunities to apply "big data" to complex public problems. However, significant challenges remain before most government data can be made suitable for this kind of application. Policies, governance mechanisms, data management protocols, data and technology standards, and a variety of skills and capabilities both inside and outside government are needed if these information-based initiatives are to contribute to better understanding of critical social and economic issues and better policies to address them.

Conclusion

Open data presents both promise and problems. We are more likely to achieve its promised benefits if we take a hard, realistic look at its character. One way to do this is to see the data as one of four linked phenomena (policy, management, technology, and data) embedded in social, organizational, and institutional contexts that have substantial influences on data quality, availability, and usability.

In this view, some of the challenges of government information use can be understood as technical problems addressing information storage, access, inquiry, and display. Another way to understand the challenges are as management problems such as defining the rationale and internal processes of data collection, analysis, management, preservation, and access. The challenges also represent policy problems including examining the balance and priority of internal government needs versus the needs of secondary users, the resources allocated to serve both kinds of uses, as well as traditional information policy concerns with confidentiality, security, and authenticity. These many new sources of government data offer potential value for society – but the value will be realized only if government information policies and practices are better aligned with the needs of external users. Likewise, analysts and other users need to take responsibility for "looking under the hood" of data sources and adjusting their expectations and assumptions to more closely match the realities of data quality and fitness for use.

References

Ballou, Donald P., and Tayi, Giri Kumer. 1999. "Enhancing data quality in data warehouse environments." *Communications of the ACM* 42(1): 73-78.

Bellamy, Christine, & Taylor, John A. (1998). *Governing in the information age.* Philadelphia: Open University Press.

Bekkers, Victor J. J. M. (1998). New forms of steering and the ambivalency of transparency. In I. T. M. Snellen & W. B. H. J. Van der Donk (Eds.) *Public administration in an information age: A handbook* (pp. 341-357). IOS Press.

Dawes, Sharon S., Pardo, Theresa A. 2006. Maximizing knowledge for program evaluation: critical issues and practical challenges of ICT strategies. Proceedings of the 5h International Conference, EGOV. Springer: *Lecture Notes in Computer Science*.

Dawes, Sharon S., Pardo, Theresa A., and Cresswell, Anthony M. 2004. "Designing electronic government information access programs: A holistic approach." *Government Information Quarterly* 21(1): 3-23.

Dawes, Sharon. S. 1996. "Interagency information sharing: Expected benefits, manageable risks." *Journal of Policy Analysis and Management* 15(3): 377−394.

Frissen, Paul H. A. 1992. Informatization in public administration: Introduction. *International Review of Administrative Sciences*, 58, 307-310.

Heinrich, Carolyn J. 2012. "How credible is the evidence, and does it matter? An analysis of the Program Assessment Rating Tool." *Public Administration Review* 72(1): 123–134.

Meijer, Albert. 2009. "Understanding modern transparency." *International Review of Administrative Sciences* 75(2): 255-269.

Radin, Beryl A. 2006. *Challenging the Performance Movement: Accountability Complexity and Democratic Value.* Washington DC: Georgetown University Press.

Wang, Richard Y., and Strong, Diane M. 1996. "Beyond accuracy: What data quality means to data consumers." *Journal of Management Information systems* 12(4): 5-34.