

## Open Data and Fitness for Use: A Realistic Look



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 and economic value. Open data initiatives are focusing on education, public health, transportation, environmental stewardship, economic development, and many other areas. Ironically, this information is often treated as a black box in the open data movement. Stakeholders, analytical techniques, and technology tools all receive considerable attention, but the information itself is often seen as a given, used uncritically and trusted without examination. However, the very kind of data that is now being released as *open data* was actually collected or created for other purposes. It has undeniable potential value, but it also contains substantial risks for validity, relevance, and trust.

**GOVERNMENT DATA FOR POLICY ANALYSIS AND EVALUATION** The explosion in so-called *administrative data*, is attracting great attention for its potential value both inside and outside government. Administrative data reflects the operations of government programs through the operation of automated activities and the advent of electronic government services. Much of this data is collected in real-time as these systems do their regular processing. For example, transactional data reveals the workflow activities of case management systems or the steps and results of customer service exchanges. Government-deployed sensor networks gather data about transport, air quality, and other topics for regulatory purposes. Financial management systems record budgets, grants, contracts, cash flow, and reconciliations.

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.

Certain sources of government data have been used by external analysts for decades. 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 specific kinds of social, economic, and demographic information to the public. The census, economic, and other formal statistics they produce 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.

**eGOVPOLINET** The eGovPoliNet/Crossover Consortium, sponsored by the European Commission FP7 research program, is an expanding international network of research institutions investigating globally important data and technology challenges in policy making. As an NSF-funded consortium member, CTG is investigating how social networks, information, and technology influence policy analysis, decision making, and policy evaluation in different parts of the world. Involvement in this international community enhances our work in the US on the value and use of government data for governance, policy-making, and social and economic benefit.

**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. 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. Emerging open data initiatives 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.

**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 tends to be a low priority activity.

**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.

### **FITNESS FOR USE\***

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

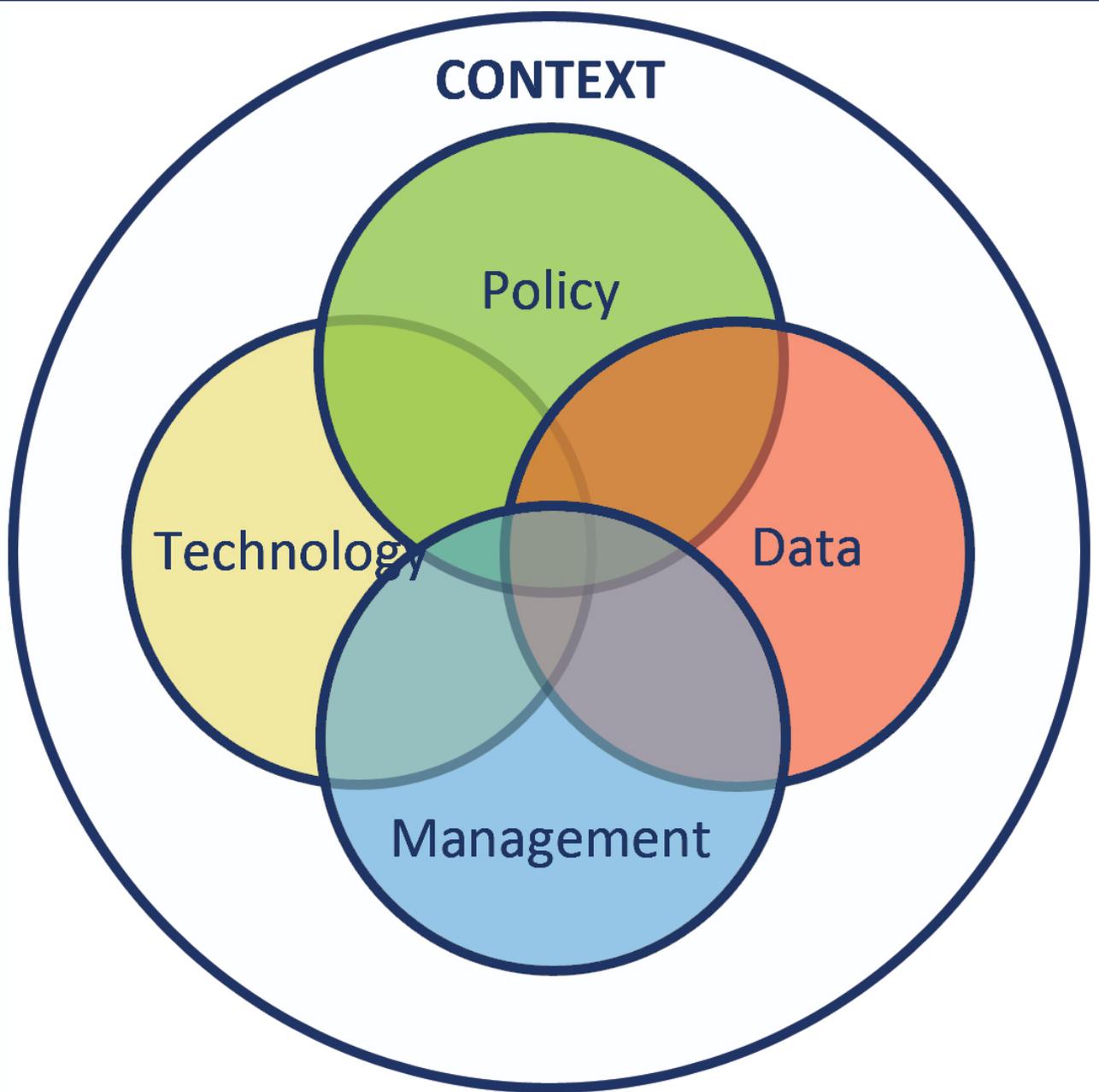
\*Wang & Strong (1996)

**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.

The current emphasis on open data plus the evolving capability of technological tools for analysis 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.

## Conceptual framework for understanding fitness for use of government information



**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 consider data in conjunction with the policies, management practices, and technology tools that create and shape it. Further, we need to understand how this ensemble of considerations is 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.

Sharon Dawes, Senior Fellow Natalie Helbig, Senior Program Associate