# Uncertain risk: assessing open data signals



Anne L. Washington
Steinhardt School, New York University, New York, NY, USA

Received 1 September 2019
Revised 26 January 2020
1 April 2020
Accepted 27 April 2020

## Abstract

Purpose – Open data resources contain few signals for assessing their suitability for data analytics. The purpose of this paper is to characterize the uncertainty experienced by open data consumers with a framework based on economic theory.

Design/methodology/approach – Drawing on information asymmetry theory about market exchanges, this paper investigates the practical challenges faced by data consumers seeking to reuse open data. An inductive qualitative analysis of over 2,900 questions asked between 2013 and 2018 on an internet forum identified how a community of 15,000 open data consumers expressed uncertainty about data sources.

Findings – Open data consumers asked direct questions that expressed uncertainty about the availability, interoperability and interpretation of data resources. Questions focused on future value and some requests were devoted to seeking data that matched known sources. The study proposes a data signal framework that explains uncertainty about open data within the context of control and visibility.

Originality/value – The proposed framework bridges digital government practice to information signaling theory. The empirical evidence substantiates market aspects of open data portals. This paper provided a needed case study of how data consumers experience uncertainty. The study integrates established theories about risk to improve the reuse of open data.

Keywords Analytics, Risk, Uncertainty, Information asymmetry, Open data

Paper type Research paper

## Introduction

Open data initiatives drive public sector analytics yet there is a gap in understanding the risks data consumers face using these sources. Evidence-based policy (Heimstädt and Dobusch, 2018), democratic data labs (Batarseh and Yang, 2020) and government artificial intelligence (Margetts and Dorobantu, 2019) are all possible because of freely available open data. While initial scholarship focused on how to implement digital government to produce open data (Dawes, 2010; Margetts and Hood, 2010), concern has shifted to the consumption of material in data-driven efforts, including analytics (Rempel et al., 2018; Safarov et al., 2017). This article contributes to this growing body of research by focusing on the practical challenges open data consumers face managing the uncertainty of working with open data.

Data consumers who choose to use open data in analytics face many risks. Incorporating a new data source demands an investment of time and an investment in talent. Data science requires substantial technical skills, which can mean expensive talent (Hofman et al., 2017). Data consumers must integrate new sources into existing workflows for analytics projects (Provost and Fawcett, 2013). Data consumers cannot mitigate risk of potential problems with a price mechanism (Ahmadi Zeleti et al., 2016) and therefore must trust that data

This project is based in part on work supported by the National Science Foundation Grant No. #1833119. This would not have been possible without several NSF-supported summer research assistants. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.



1750-6166
DOI 10.1108/TG-09-2019-0086

producers will continue consistent production of data (Mavlanova et al., 2012; Ham et al., 2019). While open data resources are free, they do come at a cost based on uncertainty.

Approaching data portals as mechanisms of exchange leverages economics as a tool for understanding data consumers. Economic theory (Akerlof, 1970; Spence, 2002) suggests that visible cues reduce uncertainty and increase control in exchanges. The exchange between data producers and data consumers can be considered a market for information (Desouza and Awazu, 2003; Ghose, 2009). The economics of information anticipates tensions inherent in exchanging open data sets.

The purpose of this paper is to analyze questions open data consumers ask and subsequently characterize their signals of uncertainty. This is a conceptual and an empirical exploration of signals needed to improve the reuse of open data. Drawing on economic theory to understand risks in material exchanges, this qualitative case study investigates market aspects of data reuse.

We address the following research question:

RQ1. How do data consumers characterize uncertainty about open data?

Data consumers are represented through public discussions on the opendata.stackexchange.com forum, which is part of a larger network of technology-themed question and answering sites. An analysis of nearly 3,000 questions on stack exchange revealed that data consumers expressed uncertainty about prospective tasks associated with availability, interoperability and interpretation of data sources.

The paper is organized as follows: a review of prior research on information asymmetry and the theory of risk; description of the research methods; results of the case study analysis; and implications of open data uncertainty. The study proposes a data signals framework to illustrate the hesitation experienced by data consumers. This qualitative case study provides needed empirical evidence of how data consumers experience uncertainty.

## Theory

Transparency efforts are intended to support innovation, citizen participation and other improved exchanges between governments and publics. Many scholars have noted that the release of digital files is rarely sufficient to achieve transparency (Jaeger and Bertot, 2010; Ruijer et al., 2020). Transparency, commonly understood as clarity (Bannister and Connolly, 2011) or accountability (Lourenço et al., 2017) is ambiguous in its meaning for complex digital objects. Governments themselves, despite substantial investments, still struggle to understand the value of open data to the public (Ohemeng and Ofosu-Adarkwa, 2015; Tempini, 2017) and open data portals are under used (Ruijer and Meijer, 2016; Abella et al., 2019). Previous research on risk and tasks within the context of information economics illuminate new ways of conceptualizing the exchange between data consumers and data producers.

### Information asymmetry

Information asymmetry theory complicates thinking about open data because it reminds us how people behave when they experience uncertainty. Information economics (Akerlof, 1970; Spence, 2002; Stiglitz, 2000) theorizes what happens when information sharing is imperfect. Imperfect information creates an asymmetry in a market exchange. For a market to function through information asymmetries, both sides need signals to resolve uncertainty about the exchange. Online markets demonstrate how information asymmetry functions with digital exchanges (Mavlanova et al., 2012). Transparency is not the whole story of open

government data without considering the information asymmetry faced when estimating the effort to complete data analytic tasks. Imperfect information in markets challenges assumptions about the transparency of open data.

Knowledge markets. Traditional markets rely on revenue, income and price. Desouza and Awazu (2003) argue that knowledge markets are similar to markets for goods and services. Knowledge markets exchange knowledge objects without revenues or costs but they do rely on reputation. Sharing knowledge is part of broader organizational learning goals (Choi and Chandler, 2020). New knowledge objects are more risky than ones with a track record because they cannot be verified in advance (Desouza and Awazu, 2003). Information uncertainty in knowledge markets generally resolves with access to a knowledge object, such as a document or data source.

We adopt the Desouza and Awazu (2003) position in this paper and situate open data sets as knowledge objects within a knowledge market such as an open data portal. Open data portals are internet websites that provide freely available data sets to the public (Gray et al., 2018; Thorsby et al., 2017). Open government data portals are points of exchange between agencies and the public (Zhu and Freeman, 2019). An ability to exchange information is fundamental to open data for public sector analytics. It is through understanding open data as exchanged knowledge that transparency is enacted.

Asymmetry. Markets operate through the disclosure or control of information about products, consumers or producers. A perfect market is one in which everyone knows the best price in addition to characteristics of consumers and producers. Perfect markets are rare and in most situations, products are exchanged with some level of uncertainty (Stiglitz, 2000). If poorly constructed, unreliable or low quality objects are exchanged, the market will remove incentives for exchanging high quality objects at high prices (Akerlof, 1970).

Asymmetric markets are based on uncertainty. The producer and the consumer may be unsure about the quality of the product or each other. When one side knows more than the other, there is an information asymmetry. Information can be valued as a cost by reflecting on what is lost in the exchange. Information can also be valued as a benefit if its utility brings additional value. Information signaling theory (Spence, 2002), a subset of information economics, considers how asymmetry is resolved.

Signals. Signals mitigate market uncertainty by making differences detectable. A professional credential, such as passing the bar for lawyers, is a signal. Signaling devices (Spence, 2002) resolve uncertainty by making value or potential evident. Signals give producers and consumers more control over transactions and serve as visible mechanisms to control uncertainty.

Data consumers are in a vulnerable position when downloading data from anonymous data portals. There is little opportunity to negotiate meaning and value (Abella et al., 2019) in ways that might occur in a financial market. However, open data present both negative and positive impacts. Data may incur extra cost if extra cleaning or interpretation is required. Data may also add value by providing timely information for competitive advantage.

We investigate whether these concepts of uncertainty from information asymmetry theory help to explain problems with data reuse, specifically aspects of visibility and control. Information signaling is used as a theoretical framework for identifying solutions for improving the reuse of open data.

## Risk

Risk is the potential for an adverse outcome that brings unwanted consequences of some magnitude (Maguire and Hardy, 2013). Risk is commonly quantified by measuring the

threat size and the extent of vulnerabilities (NASCIO, 2008). Loukis and Charalabidis (2011) define risk in the digital government context as conditions that "present serious threats" to the successful completion of a project. Governments use standardization, information gathering and behavior modification to manage large-scale risk to communities (Hood et al., 2004). Automated decision-making processes provide subtle nudges to create optimal behaviors and limit adverse outcomes (Yeung, 2016). Risk is assumed to be tangible and manageable through signals.

Risk and signals are intrinsically connected in information asymmetry theory. Uncertainty emerges from an inability to make reliable predictions (Maguire and Hardy, 2013). With sufficient information about specific criteria, a possible loss or risk could be estimated. Conversely, risks could provide rewards for steering through uncharted territory.

The predictive analytics asks data scientists to build models based causality along with domain knowledge (Shmueli and Koppius, 2011). Measurements and conceptualizations must align for data consumers to make comparisons that are relevant and valid (Leonelli, 2015). Data are dynamic objects (Buckland, 1991; Desouza and Awazu, 2003) that require translation and context each time they are reused. For instance, to estimate whether a data set is worth the investment of time, data consumers would need to understand their vulnerability to the uncertainty, the degree of the threat and potential opportunity. Government information, however, may require layers of interpretation to grasp the underlying administrative procedures and regulatory policy (Gray and Simon, 2014). Open data users may experience uncertainty about the translation of public policy into data representations.

Hardy and Maguire (2016) argue that a process view of risk takes three positions in time. A prospective risk predicts potential future setbacks. A real-time risk seeks a solution to something currently occurring. A retrospective risk reflects on something that went wrong. We use Hardy and Maguire (2016) temporal dimensions of risk to analyze data consumer uncertainty.

## Uncertainty in tasks

The tasks involved with using data sources are not straightforward. A data set holds the potential for many comparisons over time. Unlike information quality concerns about a document, data consumers must consider how information in the data interlocks with existing sources across multiple interactions. Are names in order of first and surname, surname and first name? Is the year a calendar year or a fiscal year? To use a data set in analytics, it is necessary to consider how order and meaning integrate with other material. Data consumers must calculate the time necessary they may spend to screen out inappropriate, unwanted or poor quality data. Seeking information is a common task that has been studied over decades from group psychology (McGrath, 1984) to management studies (Campbell, 1988), and to computer human behavior (Li and Belkin, 2008).

Campbell (1988) suggested four types of tasks: simple, decision, judgment and problem. Simple tasks have a single resolution. Decision tasks require a choice between multiple outcomes. Judgment tasks have multiple paths toward multiple outcomes that require a balance between conflicting positions. Problem tasks have one outcome but multiple paths that conflict. Each Campbell (1988) task moves from uncertainty towards a resolution.

McGrath (1984) proposed a task schema based on a four stage process: generate, choose negotiate and execute. Generating tasks include both action-oriented plans and creativity to generate ideas. Choosing tasks involves either solving for one correct answer or selecting preferences. Negotiation resolves conflicts around motives and incentives. Execution

resolves conflicts through victory or improved performance. The McGrath (1984) schema reflects uncertainty at different stages of a process.

This paper, recognizing that tasks are complex and interdependent, leverages multiple systems of analysis to categorize observations. Existing models are used as heuristics to understand how data consumers approach the task of resolving uncertainty about open data. We leverage the Li and Belkin (2008) comparative analysis of Campbell (1988) and McGrath (1984) to organize our thinking about how data consumers seek resolution of their uncertainty. The next section describes the research design that leverages signaling theory to explain observed phenomena.

## Methods

The study was designed to identify the uncertainty of data consumers through an empirical analysis of online forum questions. An inductive line of inquiry supported with computer-assisted tools is a research design used by other researchers (Corrales-Garay et al., 2019; Gioia et al., 2013) using qualitative data (Creswell, 2013; Denzin and Lincoln, 2011). This study examines the written traces of data consumers on an open internet forum.

### Stack exchange

Question and answering (Q&A) communities enable people with varying levels of expertise to solve problems together and leave a record of that solution for others. Scholars of online forums contend that they are collaborative editing spaces for a community to ask, redefine and facilitate an answer (Ibarra et al., 2005). These communities monitor themselves and sustain a network by social engagement and reputation (Faraj et al., 2011).

Stack exchange, a popular online question and answering forum for technologists, was selected for this study. Stack exchange is designed as a place for newcomers and experts to meet. Stack exchange instances can vary greatly from only a few hundred users, seen on stackexchange.com, to tens of thousands of users, such as stackexchange.com. The size of the instance is measured in number of users, answer activity and months as started (Santos et al., 2019). The open data forum is mid-range size for a stack exchange forum (Santos et al., 2019). We selected stack exchange because of its history of openness and its connection to open data policy.

The open data instance, opendata.stackexchange.com, started simultaneously with the launch of major US federal policy in May 2013 (Burwell et al., 2013; White House, 2013). The US federal CIO council recognized the need to increase internal and external conversations stating that it needed "to extend its reach to existing social network and developer communities including Twitter, LinkedIn, GitHub and stack exchange" (US CIO Council, 2013). We believe that the intentional creation of this community of open data consumers alongside policy initiatives makes the material ideal for academic attention.

There are several constraints to using stack exchange. The people typically on stack exchange are familiar with technology and may not represent an average novice open data consumer. Further, the majority of links pointed to sites in the USA although over 200 were tagged for Europe and the UK. Open data culture in the USA might vary from other parts of the world.

Stack exchange allows its posts to be used for research and has a specific channel to discuss the criteria for use. The data explorer feature on stack exchange allows data dumps. Researchers can conduct dynamic search and retrieval to obtain data. The research ethics board at the university associated with the researchers did not declare this as human subjects research but to protect the privacy of individual stack exchange users some small changes have been made to the quotes in this report.

TG

## Research design

Stack exchange posts were included in the study if they contained a link and received at least one answer. This narrowed our examination to posts about specific named data sets. This eliminated many community-building posts about how to use stack exchange. The diversity of language across these short posts was not consistent enough for systematic quantitative content tools. For instance, the columns of a spreadsheet were described with a variety of words including: feature, property, parameters, factor and variable. Because of this linguistic diversity, qualitative coding was the primary method of analysis. The research team worked separately to group similar posts and discussed findings weekly. Dedoose software, version 8.0.35, was used to share findings between the research team. The research took place in three stages.

In the first stage of analysis, we selected criteria and gathered data from stack exchange. Initially, we carefully reviewed and categorized a subset of posts that focused on prediction and modeling. We expanded the data set to the current size and used computational linguistic tools to compare frequency (Anthony, 2006) and establish distance (Richards, 2005) from the material.

In the second stage of analysis, we identified uncertainty factors to infer types of risk identified by data consumers in relation to the literature. The initial review of the posts was in vivo coding, using the natural language of the post (Denzin and Lincoln, 2011). Subsequent reviews (Creswell, 2013) built on axial coding where concepts from theory built on the original set of codes. The factors of uncertainty were categorized using the literature on risk and task completion.

In the third stage of analysis, the uncertainty factors were further analyzed using previous literatures to develop the data signals framework. Taking these observations as entry points, we explain how uncertainty is expressed along the dimensions of visibility of the signal and control of the data consumer.

This study does have limitations as an analysis of a single case study. Written traces of data consumers were analyzed instead of direct observation of uncertainty because of the challenge of finding people who are interested but have reservations about open data.

## Open data uncertainty

The following sections characterize the uncertainty expressed in stack exchange posts and propose a framework for understanding the missing signals inherent in the questions.

## Question themes

This study covers five years, from May 2013 through June 2018. As of June 2018, the open data stack exchange forum had 15,282 users who asked 4,291 questions with 71% of the posts receiving at least one answer. The answered posts containing at least one link totaled approximately 272,104 words, given that some stack exchange posts contained programming code. Three broad themes were established in the analysis of the questions data consumers asked: availability, interoperability and interpretation. Availability (Gebre and Morales, 2020; Ham et al., 2019) considered whether the material existed and was obtainable as open data. Interoperability (Tempini, 2017; Washington, 2016) questions asked about interconnecting open data to existing sources and algorithms. Interpretation (Gray and Silbey, 2014; Leonelli, 2015) reflected on how different groups understood the material within the original context and its new context. Below are example stack exchange posts that reflect these themes.

Availability. Stack exchange posts asked what open data setes existed. The queries ranged from asking whether the set existed at all, to asking whether it was available as open data:

> This is a basic question concerning data availability. I am looking for historic geo-spatial data.
>
> I am looking for a dictionary of English words along with the probability that it might be spelled incorrectly.
>
> I am looking for wind speed and gust speed data for the Northern California region including Sonoma County.

In some cases, the data consumer expected something to be available but was not sure where to locate it. Other queries sought data sets to support models that appeared in the publications or academic literature:

> In his 1927 paper, Udny Yule considers a thought experiment of a simple pendulum hanging from a fixed point . . . Is there a data set available that will illustrate this pendulum example?
>
> Where to find data to explore the “Rescorla-Wagner model”?

Timeliness of the data was a thread interwoven between many stack exchange posts. Data consumers were concerned about when data sets were updated and subsequently released:

> I cannot find information on the currency metric of the data. I need to know when the data was collected for my model.
>
> Historically, it takes on average nine months for the data to release so by the time it is ready you are typically looking at data from a year ago.
>
> What year does the May 2015 data represent? Is there any indication of when the firms filled out the surveys?

These questions reflect the underlying concern of data consumers about the long-term availability and consistency of open data. Analysts who rely on open data risk jeopardizing analytic models that require continuous data flows.

Interoperability. Data consumers on stack exchange asked about interoperability between data sets and also about specific values and taxonomies. Interoperability posts asked how an unknown open data set could be combined with a known set:

> The ID in the corresponding data is not in the same format to the ID from the schedule data, which makes it hard to connect the historical data to the scheduled data. Can someone provide some suggestions?
>
> Is there any dataset labeled with this <link>? I want to predict categories based on the taxonomy.
>
> I need 2013 migration data for validating a model I want to build a matrix indexed by country where A[x,y] is the migration of citizens of country x to country y.

Some data consumers explicitly stated that they did not trust the information provided in the data set. Note that in the cases below, the author hints at issues of timeliness as part of trust. One reason the information is not trusted is because it is not clear whether it has been updated:

> What is the status as of April 2018? Beta? Production? Good to use? I would like to use these but I have no idea how trustworthy they are.

TG

> I would like to get some external verification to see if I am correct. But it seems that some categories are no longer present in this version. I haven't been able to find any documentation on this change.

A surprising aspect of the study were questions that were not about features nor topics. They simply asked for data sets similar to theirs in terms of its predictive ability or types of data. Stack exchange posts seeking similar data sets also discussed generic tasks such as testing or benchmarking:

> I am looking for dataset for NYC similar to the one of <link> for Chicago. The <link> dataset does not contain a lot of information about when and where.

> Is there a collection of regression datasets for benchmarking?

> I'm looking for multivariate-numerical data sets labeled for clustering with 100,000 instances. 10,000 is too small.

The "data like mine" phenomenon may be difficult to standardize, however, data producers should be aware that consumers are interested in this feature. Data analytics relies on the ability to logically interconnect data sources (Hofman et al., 2017). Organizations that run data portals must begin to recognize the need for verifying a micro-level connection between sources.

Interpretation. Questions reflected uncertainty about how to interpret portions of the data set. Some questions centered on the complexities of government regulation or needed to grasp underlying laws or definitions. Others were looking to simply expand their current material with data that had the same type of interpretation:

> Can I put the contents of NCES data on a website? What does this sentence mean <link>?

> What do the summary files in the American Community Survey mean? They are in a text format that looks like CSV but does not have headers.

> I am training my system with frequencies to predict sounds in five different categories. I'm following this <link> and they are using 78 different sounds manually collected. Dataset should have more than 50 data with categories.

> It's difficult to interpret service requests data. If we see that part of a city has more calls for it, does that mean that there are more of them or more people who will make service requests?

In the next stage, we analyzed the risk factors in relation to existing theory to frame our findings in concert with existing literature.

## Uncertainty factors

An inductive analysis of the 2,969 questions on stack exchange identified 14 uncertainty factors. Each uncertainty factor was further analyzed using the task reasoning classifications (Campbell, 1988), task process categories (McGrath, 1984), risk types (Spence, 2002) and risk dimensions (Hardy and Maguire, 2016). The number of stack exchange posts for each factor appears in a table in the Appendix along with representative questions. The counts of each factor were used to indicate their relative frequency for analysis.

The stack exchange questions expressed uncertainty about data for a specific topic, metadata, regulatory interpretation, match for specific rows and columns, licensing, prediction, models or algorithms, academic or published reports, matching "data like mine,"

alternate time/location, reliability, existence of data, unexpected updates and changes, open version of proprietary or private data and file formats

Open data tasks. The review of the literature on tasks considered both a process approach to tasks (McGrath, 1984) and a reasoning approach to tasks (Campbell, 1988). The uncertainty factors were analyzed with both modalities.

Task Process. The stack exchange questions primarily were concerned with generating and choosing actions within a task process. Questions analyzed through task process (McGrath, 1984) were categorized as follows: 1,188 generate 40%, 1,236 choose 42%, 0 negotiate and 545 execute 18%. Negotiations were non-existence given the asymmetry of the open data knowledge market. A smaller portion of the questions asked about the execution of tasks.

Task Reasoning. Most of the questions reflected non-ambiguous tasks or tasks that needed decisions. The stack exchange questions analyzed through the task reasoning framework (Campbell, 1988) were categorized as follows: 1,304 simple tasks 44%, 973 decision tasks 33%, 278 judgment tasks 9% and 414 problem tasks 14%. Few tasks require value judgments or presented deeply complex problem tasks that could have multiple appropriate answers.

Open data risk. The literature on risks considers temporal dimensions of risk (Hardy and Maguire, 2016), in addition to classic positive and negative positions from economic theory (Spence, 2002). This analysis considers both approaches to better understand how data consumers approach the risk of open data. We found that data consumers were seeking to add value to their analytics practices. The questioners at stack exchange wrote posts to seek value more frequently than asking questions to avoid risk.

Risk type. Information economics (Spence, 2002) suggests that uncertainty could lead to positive or negative results. Negative uncertainty is about avoiding risk and controlling for unpleasant situations. A positive view on uncertainty views sees an opportunity for benefit. The open data posts sought value more frequently than avoiding adversity. Stack exchange questions seeking additional value, creativity or competitive edge represented 67% with 1,996 posts. Questions evaluating potential adverse outcomes represented 33% or 973 posts.

Risk dimensions. The literature on risk suggests that timing is an important factor in managing unexpected events (Hardy and Maguire, 2016; Maguire and Hardy, 2013). Risk assessment may require immediate attention, be a preventative measure or reflect on a past situation. The 1,956 stack exchange questions about prospective risk represented 66% of the posts. Real-time risk concerns represented 927 or 31% and retrospective risks represented 86 or 3% of the posts. Few data consumers needed to contend with the risks of events in progress.

In summary, data consumers on stack exchange asked direct questions to generate opportunities that would avoid future risk.

## Data signals framework

This study leverages information signaling theory to understand how visible ways to control risk are reflected in the questions open data consumers asked. Risk can be managed with control over detectable challenges. The 14 uncertainty factors asked on stack exchange are summarized in the following framework, which explains signals for open data. The data signals framework, illustrated in Figure 1, organizes the uncertainty factors by control and visibility.

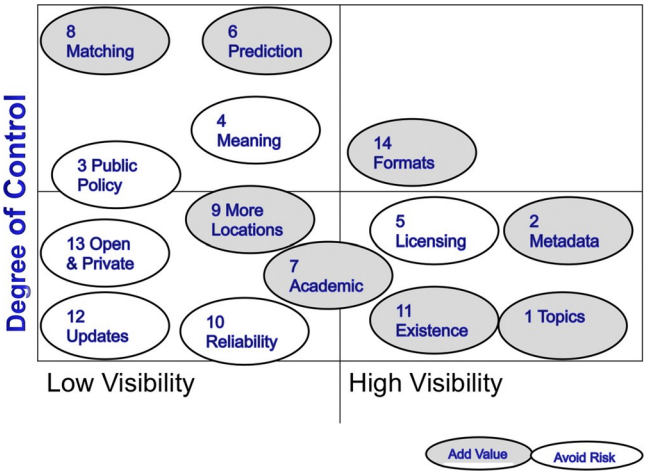


Figure 1.
Data signals framework

This conceptual framework demonstrates the complementary nature of signals about data sets content and the degree of control data consumers need to complete tasks. Dawes (2010) divided the concept of transparency into two prongs: stewardship and usefulness. Usefulness increases public value and promotes innovation while stewardship reduces risk, engenders trust and assures quality (Dawes, 2010). The data signals framework invites governments to reflect on their stewardship responsibilities in making signals visible and useful to data consumers.

Visibility. Many scholars recognize the paradox that transparency can actually obscure information (Ruijer et al., 2020; Wang and Shepherd, 2020). The visibility of information is essential to mitigating risk. The data signals framework positions the uncertainty factors from low to high visibility. High visibility factors can be resolved through publicly available means, such as a search engine or data portal. The information is readily available and can be queried without additional input. Low visibility factors must be resolved through interpersonal connections. An individual citizen may find it difficult to find the few experts inside an organization who could resolve a question about a data set.

Control. Data consumers in the stack exchange forum were seeking control over their uncertainty. Some of the questions reflected actions that were in their control. This could be knowing whether a new data source matched their existing workflow or how they wanted to design a predictive model. Many of their concerns were activities outside of their control such as updates, timing of releases or reliability of data.

The data signal framework maps the data consumer experience, indicating where questions focused on adding value or avoiding risk. Data consumers wanted to avoid risks such as using the wrong license or interpreting a value incorrectly. However, many of the questions sought explicit value such as saving time by reusing data or saving money by discontinuing paid data subscriptions.

A fundamental assumption of public sector analytics is people will be able to appreciate the value of open data once it is released. This contribution addressed that assumption by considering how open data consumers experience uncertainty. The empirical evidence substantiates market aspects of open data portals.

## Conclusion

This research explored why and how data consumers experience risks of uncertainty when considering specific open data resources. The use of widely available open data sets is limited because data analysts may have insufficient signals to assess the material. The findings have implications for public sector analytics that relies on open data sources.

## Implications for public sector analytics

The availability of open data through robust data portals is essential for successful public sector analytics. The first generation of open data were built with limited funding (Jaeger and Bertot, 2010). Government organizations released copious data sets with scant metadata and documentation. It is unlikely that public administration budgets could produce sufficient metadata to address the concerns found in this study and others (Fini et al., 2018). However, governments may be able to tailor their offerings to the needs of specific groups or personas. Personas (Pruitt and Adlin, 2006) are descriptions of typical idealized users and are frequently used in designing online products. Personas also extend previous research on the relevant social groups who use open data (Lassinantti et al., 2019; Washington, 2019). The data signal framework could begin to support that effort.

The questions posed on stack exchange also revealed that few people have the domain expertise to leverage open government assets. Question and answering sites solve many of these concerns because they allow experts and novices to engage in conversations that are available to others. Governments may consider other knowledge sharing technology that encourages and supports data consumers.

## Implications for research

Public administration scholars could benefit from this analysis that indicates incentives for engaging with open government data. Transparency requires healthy and frequent negotiations between data consumers and producers (Zhao and Fan, 2018). Implications for public policy research are also evident. What should the next generation of open data policy consider? What is the best balance between costly metadata and actual reuse? Sociologists of knowledge may be interested in this individual-level evidence of knowledge product culture. The data signal framework reveals the micro foundations of individual practices within open data research.

This study offers multiple directions for additional research. The issue of risk needs further elaboration to better understand how data consumers discern the relationship between data quality and potential value. Future researchers may consider the cost and benefits of additional signals for open data. Additional work is needed to understand how data professionals express similarity between data sets as indicated in the “data like mine” finding. The proposed framework provides exploratory evidence based on US Government data that may need to be refined for use in other contexts such as cities (Ilies and Titah, 2017) and municipalities (Zhu and Freeman, 2019). Finally, future research could address the immediate concern of accommodating multiple and perhaps conflicting approaches to risk.

This study connected digital government practice to information signaling theory. The research intended to understand how and why data consumers act. The results have implications for public administration research on risk and uncertainty in digital government. The illustrations in this evidence contribute new insights into data consumers that could improve the reuse of open data.

TG

References


Abella, A., Ortiz-de-Urbina-Criado, M. and De-Pablos-Heredero, C. (2019), "The process of open data publication and reuse", Journal of the Association for Information Science and Technology, Vol. 70 No. 3, pp. 296-300.

Ahmadi Zeleti, F., Ojo, A. and Curry, E. (2016), "Exploring the economic value of open government data", Government Information Quarterly, Vol. 33 No. 3, pp. 535-551.

Akerlof, G.A. (1970), "The market for 'lemons': quality uncertainty and the market mechanism", The Quarterly Journal of Economics, Vol. 84 No. 3, pp. 488-500.

Anthony, L. (2006), "Developing a freeware, multiplatform corpus analysis toolkit for the technical writing classroom", IEEE Transactions on Professional Communication, pp. 275-286.

Bannister, F. and Connolly, R. (2011), "The trouble with transparency: a critical review of openness in e-Government", Policy and Internet, Vol. 3 No. 1, pp. 1-30.

Batarseh, F.A. and Yang, R. (2020), "Data democracy for you and me (bias, truth, and context) chapter 1", in Batarseh, F.A. and Yang, R. (Eds), Data Democracy, Academic Press, pp. 3-8.

Buckland, M.K. (1991), "Information as thing", Journal of the American Society for Information Science, Vol. 42 No. 5, pp. 351.

Burwell, S. et al. (2013), "Open data policy-managing information as an asset memo M-13-13", Office of Management and Budget, Executive Office of the President (9 May 2013), available at: https://pdfs.semanticscholar.org/766f/6a298eebb16e95a5ac0f0be9947a9d62f013.pdf

Campbell, D.J. (1988), "Task complexity: a review and analysis", The Academy of Management Review, Vol. 13 No. 1, pp. 40-52.

Choi, T. and Chandler, S.M. (2020), "Knowledge vacuum: an organizational learning dynamic of how e-government innovations fail", Government Information Quarterly, Vol. 37 No. 1, pp. 101416.

Corrales-Garay, D., Ortiz-de-Urbina-Criado, M. and Mora-Valentín, E.-M. (2019), "Knowledge areas, themes and future research on open data: a co-word analysis", Government Information Quarterly, Vol. 36 No. 1, pp. 77-87.

Creswell, J.W. (2013), Qualitative Inquiry and Research Design: choosing among Five Approaches, 3rd ed., SAGE Publications, Los Angeles.

Dawes, S.S. (2010), "Stewardship and usefulness: policy principles for information-based transparency", Government Information Quarterly, Vol. 27 No. 4, pp. 377-383.

Denzin, N.K. and Lincoln, Y.S. (2011), The Sage Handbook of Qualitative Research, 4th ed., Sage, Los Angeles, CA.

Desouza, K.C. and Awazu, Y. (2003), "Constructing internal knowledge markets: considerations from mini cases", International Journal of Information Management, Vol. 23 No. 4, pp. 345-353.

Faraj, S., Jarvenpaa, S.L. and Majchrzak, A. (2011), "Knowledge collaboration in online communities", Organization Science, Vol. 22 No. 5, pp. 1224-1239.

Fini, R., Rasmussen, E., Siegel, D. and Wiklund, J. (2018), "Rethinking the commercialization of public science: from entrepreneurial outcomes to societal impacts", Academy of Management Perspectives, Vol. 32 No. 1, pp. 4-20, doi: 10.5465/amp.2017.0206.

Gebre, E.H. and Morales, E. (2020), "How 'accessible' is open data? Analysis of context-related information and users' comments in open datasets", Information and Learning Sciences, Vol. 121 Nos 1/2, pp. 19-36.

Ghose, A. (2009), "Internet exchanges for used goods: an empirical analysis of trade patterns and adverse selection", MIS Quarterly, Vol. 33 No. 2, pp. 263-291.

Gioia, D.A., Corley, K.G. and Hamilton, A.L. (2013), "Seeking qualitative rigor in inductive research notes on the Gioia methodology", Organizational Research Methods, Vol. 16 No. 1, pp. 15-31.

Gray, G.C. and Silbey, S.S. (2014), "Governing inside the organization: interpreting regulation and compliance", American Journal of Sociology, Vol. 120 No. 1, pp. 96-145.

Gray, J., Gerlitz, C. and Bounegru, L. (2018), "Data infrastructure literacy", Big Data and Society, Vol. 5 No. 2.

Ham, J., Koo, Y. and Lee, J.-N. (2019), "Provision and usage of open government data: strategic transformation paths", Industrial Management and Data Systems, Vol. 119 No. 8, pp. 1841-1858.

Hardy, C. and Maguire, S. (2016), "Organizing risk: discourse, power, and 'rectification'", Academy of Management Review, Vol. 41 No. 1, pp. 80-108.

Heimstädt, M. and Dobusch, L. (2018), "Politics of disclosure: organizational transparency as multiactor negotiation", Public Administration Review, Vol. 78 No. 5, pp. 727-738.

Hivon, J. and Titah, R. (2017), "Conceptualizing citizen participation in open data use at the city level", Transforming Government: People, Process and Policy, Vol. 11 No. 1, pp. 99-118.

Hofman, J.M., Sharma, A. and Watts, D.J. (2017), "Prediction and explanation in social systems", Science, Vol. 355 No. 6324, pp. 486-488.

Hood, C., Rothstein, H. and Baldwin, R. (2004), The Government of Risk: Understanding Risk Regulation Regimes, Oxford University Press, Oxford, UK.

Ibarra, H., Kilduff, M. and Tsai, W. (2005), "Zooming in and out: connecting individuals and collectivities at the frontiers of organizational network research", Organization Science, Vol. 16 No. 4, pp. 359-371.

Jaeger, P.T. and Bertot, J.C. (2010), "Transparency and technological change: ensuring equal and sustained public access to government information", Government Information Quarterly, Vol. 27 No. 4, pp. 371-376.

Lassinantti, J., Ståhlbröst, A. and Runardotter, M. (2019), "Relevant social groups for open data use and engagement", Government Information Quarterly, Vol. 36 No. 1, pp. 98-111.

Leonelli, S. (2015), "What counts as scientific data? A relational framework", Philosophy of Science, Vol. 82 No. 5, pp. 810-821.

Li, Y. and Belkin, N.J. (2008), "A faceted approach to conceptualizing tasks in information seeking. special issue on adaptive", Information Processing and Management, Vol. 44 No. 6, pp. 1822-1837.

Loukis, E. and Charalabidis, Y. (2011), "Why do eGovernment projects fail? Risk factors of large information systems projects in the Greek public sector: an international comparison success", International Journal of Electronic Government Research (IJEGR), Vol. 7 No. 2, pp. 59-77.

Lourenço, R.P., Piotrowski, S. and Ingrams, A. (2017), "Open data driven public accountability", Transforming Government: People, Process and Policy, Vol. 11 No. 1, pp. 42-57.

McGrath, J.E. (1984), Groups: interaction and Performance, Prentice-Hall, Englewood Cliffs, NJ.

Maguire, S. and Hardy, C. (2013), "Organizing processes and the construction of risk: a discursive approach", Academy of Management Journal, Vol. 56 No. 1, pp. 231.

Margetts, H.P. and Hood, C. (2010) Paradoxes of Modernization: unintended Consequences of Public Policy Reform, Oxford University Press, Oxford, New York, NY,

Margetts, H. and Dorobantu, C. (2019), "Rethink government with AI", Nature, Vol. 568 No. 7751, pp. 163-165.

Mavlanova, T., Benbunan-Fich, R. and Koufaris, M. (2012), "Signaling theory and information asymmetry in online commerce", Information and Management, Vol. 49 No. 5, pp. 240-247.

National Association of State Chief Information Officers (NASCIO) (2008), Digital States at Risk: Modernizing Legacy Systems, NASCIO National Survey on Legacy Systems and Modernization in the States, Washington DC.

Ohemeng, F.L.K. and Ofosu-Adarkwa, K. (2015), "One way traffic: the open data initiative project and the need for an effective demand side initiative in Ghana", Government Information Quarterly, Vol. 32 No. 4, pp. 419-428.

TG


Provost, F. and Fawcett, T. (2013), "Data science and its relationship to big data and data-driven decision making", Big Data, Vol. 1 No. 1, pp. 51-59.

Pruitt, J. and Adlin, T. (2006), The Persona Lifecycle: Keeping People in Mind Throughout Product Design, Elsevier: Morgan Kaufmann Publishers, an imprint of Elsevier, Amsterdam; Boston,

Rempel, E.S., Barnett, J. and Durrant, H. (2018), "Public engagement with UK government data science: propositions from a literature review of public engagement on new technologies", Government Information Quarterly, Vol. 35 No. 4, pp. 569-578.

Richards, L. (2005), Handling Qualitative Data: A Practical Guide, SAGE Publications, London Thousand Oaks, CA,

Ruijer, E., Détienne, F., Baker, M., Groff, J. and Meijer, A.J. (2020), "The politics of open government data: understanding organizational responses to pressure for more transparency", The American Review of Public Administration, Vol. 50 No. 3, pp. 260-274.

Ruijer, E. and Huff, R.F. (2016), "Breaking through barriers: the impact of organizational culture on open government reform", Transforming Government: People, Process and Policy, Vol. 10 No. 2, pp. 335

Safarov, I., Meijer, A. and Grimmelikhuijsen, S. (2017), "Utilization of open government data: a systematic literature review of types, conditions, effects and users", Information Polity, Vol. 22 No. 1, pp. 1-24.

Santos, T., Walk, S., Kern, R., Strohmaier, M. and Helic, D. (2019), "Activity archetypes in question-and-Answer (Q8A) websites—a study of 50 stack exchange instances", ACM Transactions on Social Computing, Vol. 2 No. 1, pp. 1-23.

Shmueli, G. and Koppius, O.R. (2011), "Predictive analytics in information systems research", MIS Quarterly, Vol. 35 No. 3, pp. 553-572.

Spence, M. (2002), "Signaling in retrospect and the informational structure of markets", American Economic Review, Vol. 92 No. 3, pp. 434-459.

Stiglitz, J.E. (2000), "The contributions of the economics of information to twentieth century economics", The Quarterly Journal of Economics, Vol. 115 No. 4, pp. 1441-1478.

Tempini, N. (2017), "Till data do us part: understanding data-based value creation in data-intensive infrastructures", Information and Organization, Vol. 27 No. 4, pp. 191-210.

Thorsby, J., Stowers, G.N.L., Wolslegel, K. and Tumbuan, E. (2017), "Understanding the content and features of open data portals in American cities", Government Information Quarterly, Vol. 34 No. 1, pp. 53-61.

US CIO Council (2013), "Project open data", available at: https://project-open-data.cio.gov/api-basics/

Wang, V. and Shepherd, D. (2020), "Exploring the extent of openness of open government data – a critique of open government datasets in the UK", Government Information Quarterly, Vol. 37 No. 1, p. 101405.

Washington, A.L. (2016), "The interoperability of US federal government information: interoperability", in Aggarwal, A. (Ed.), Managing Big Data Integration in the Public Sector, IGI Global.

Washington, A.L. (2019), '"Who do you think We are? The data publics in digital government policy", Proceedings of the 52nd HI International Conference on System Sciences, Maui, HI.

White House (2013), "Under the hood of the open data engine", Data.gov, available at: www.data.gov/blog/under-hood-open-data-engine

Yeung, K. (2016), "Hypernudge': big data as a mode of regulation by design", Information, Communication and Society, Vol. 20 No. 1, pp. 118-136.

Zhao, Y. and Fan, B. (2018), "Exploring open government data capacity of government agency: based on the resource-based theory", Government Information Quarterly, Vol. 35 No. 1, pp. 1-12.

Zhu, X. and Freeman, M.A. (2019), "An evaluation of U.S. municipal open data portals: a user interaction framework", Journal of the Association for Information Science and Technology, Vol. 70 No. 1, pp.

Appendix

| Factors | Count | Example stack exchange post |
|---|---|---|
| 1. Topic | 358 | 104492017 I am looking for Clark County Nevada Parcel GIS data. |
| 2. Metadata | 416 | 12013 Is there a list of all US Government agencies and sub agencies and is it available via API? |
| 3. Public policy interpretation | 162 | 66762015 I am looking for a data set of physician notes with annotated PHI as defined in HIPAA |
| 4. MicroMatch – meaning | 143 | 98072016 Can anyone clarify how the 'Rewritten' value in the status description column impacts the data? |
| 5. Licensing | 290 | 53782015 Can I insert a CC BY-SA photo as a figure in another work, for example, an academic paper, that is not itself CC BY-SA? |
| 6. Prediction/models/ algorithms | 86 | 119742017 I am working on a research project for developing predictive models for hydro-/aqua-/aeroponics. Where can I found related data sets for the nutrients and other sensitive parameters? |
| 7. Academic | 124 | 124092018 I came across this interesting paper that make use of the Klink-2 computer science ontology (CSO). Where can I find an API or something like that to access this ontology? |
| 8. Matching (data like mine) | 162 | 13174 I am looking for a data set that has at least one feature that is categorical and that takes more than 1,000 different values |
| 9. Alternate time/ location | 110 | 99522016 How can I obtain public transport data for Hong Kong/ Shenzhen |
| 10. Reliability | 357 | 114372017 How reliable are 250 m soil map data from soilgrids.org? |
| 11. Existence of data | 116 | 5622013 Is there a global database of all products with EAN 13 barcodes? |
| 12. Updates and changes | 157 | 104652017 Does anyone have any concrete knowledge of whether or not there are plans to continue to update the Scorecard each year? |
| 13. Open version of proprietary data | 203 | 117282017 Are there any public sources for LIDAR data for the country of Israel? |
| 14. File formats | 358 | 104282017 I need to motorway map of the UK as data in CSV format |

Table A1. Uncertainty factors

About the author

Dr Anne L. Washington is an Assistant Professor of Data Policy at NYU, a scholar of public-interest technology with an expertise in government data. As a computer scientist trained in organizational ethnography, Professor Washington unites inductive qualitative research methods with technology tools. At the broadest level, her research work considers the ethical impact of technology on society through the lens of digital record keeping. Anne L. Washington can be contacted at: washingtona@acm.org