

Explaining Download Patterns in Open Government Data: Citizen participation or private enterprise?

Authors

Jonathan Bright (corresponding author), Sumin Lee, Helen Margetts, Ning Wang, Scott Hale.

All authors are affiliated with the Oxford Internet Institute, University of Oxford

Abstract

Open data remains one of the most significant current trends in public administration, with hundreds of projects around the world seeking to open up stores of public sector information for future re-use for a wide range of hypothesized benefits, especially in terms of enabling citizen participation in government. However, as open data has grown, a critical literature has also emerged which questions who the true beneficiaries of open data are, as well as highlighting the high costs it places on government. Hence, systematic research on the actual outcomes of open data projects is urgently needed.

This article seeks to contribute to this area. Based on a unique observational dataset scraped from the website data.gov.uk, this article seeks to explain the factors which promote greater levels of downloads of open government data, and thus shed light on the types of values which are actually supported by such programs. We show that downloads of datasets relevant for private sector enterprise far outnumber downloads of data which could enhance citizen participation through stimulating either government transparency or government efficiency. We also show that well updated datasets with high quality metadata are more likely to be

downloaded. We conclude by supporting currently developing calls for prioritization in open government data programs.

Keywords:

open government data, open data, digital government, e-government

Highlights

- We investigate the causes of downloads of open government data
- Data which could be used for enterprise is highly downloaded
- Transparency and service efficiency data is poorly downloaded
- Well updated datasets are also downloaded more
- Our results support the need for prioritization in open data programs

1. Introduction

Open data is one of the most significant current trends in public administration, with over 100 current open government projects around the world investing considerable time and effort in opening up public sector information for free re-use. Supporters of OGD have a wide variety of motivations, but probably the most crucial is enabling new and enhanced forms of citizen participation in government, through providing data (which may well be analysed and processed by third party media or civil society organisations) that enables citizens to evaluate the transparency and efficiency of government. However, the movement, whilst attracting widespread support, has also proved highly controversial, with critics highlighting both the high financial burden it appears to place on government departments and its uncertain political consequences; with some going so far to claim that the rhetoric of government transparency and open innovation has been co-opted into the service of neoliberal deregulation (Longo 2011).

Despite the significance of the movement and the importance of the critiques, "little systematic OD research has been performed to date" (Peled, 2013, p. 187), with both supportive and critical authors often relying on interviews and one-off case studies which do not support generalisation. The major reason for this lies in the openness of open data itself: with no requirement to create an account before accessing data, there are few records of who is using it or for what purpose. Hence, it is hard to systematically answer many of the major questions being raised about open data, and we still know little about what type of values open government data programs really support.

The aim of this article is hence to bring some quantitative evidence to the debate, looking in particular at how to explain variation in levels of downloads of open government datasets. Based on information downloaded and scraped from the website data.gov.uk (the UK's open data portal), we develop and test potential explanations for the observed variation in dataset downloads. In particular, we explore factors relating to enhancing transparency (and

hence trust) in government, enhancing government efficiency, and providing enterprise and profit making opportunities. We also address whether completeness of metadata and frequency of updates to datasets have an impact on their eventual use. Our results challenge some of the key critiques currently live in the open data debate, and also suggest a way forward for the area in terms of financial sustainability and its contribution to democratic accountability.

The article is structured in the following way. In part 2 we define open data, review some of the intense debate around the subject, and hence develop our hypotheses about how variation in dataset downloads can be explained. In part 3, we describe our data collection and operationalization of key variables. In part 4, we test our hypotheses with a variety of regression models. We conclude by highlighting the implications of our findings for both literature and the open data movement.

2. Explaining Downloads of Open Government Data

"Open Data" is a way of thinking about data sharing. Data is "open", broadly speaking, when it is freely and easily accessible in a non-proprietary machine readable format, without any major restrictions placed on its re-use (a variety of more specific definitions of how data can be open exist - see e.g. Bates, 2012). In a government context, which is the major interest of this article, opening data involves allowing access in an open fashion to stores of "public sector information", data which is collected by various branches of government, often in the course of performing routine administrative tasks (see Kalampokis, Tambouris, & Tarabanis, 2011; Wang & Lo, 2016). The idea of opening government data up in some way is one of the most important current trends in public administration, with over 100 examples of "Open Government Data" [OGD] projects around the world (Davies, 2012). Indeed, a commitment to open data now often forms part of wider program of government modernization under recent moves towards "digital era governance" (Margetts & Dunleavy, 2013; Bright & Margetts, 2016).

The OGD movement has attracted considerable research and debate within academic literature. One of the main strands of this debate has concerned the values promoted by the opening of government data in this way (Jetzek 2013). An eclecticism of actors support the idea of OGD, from enthusiast individual “hackers”, private industry, academia and civil society to a variety of people working in government itself, each of which come to the movement from different ideological perspectives (see Bates, 2012 for a discussion). This variety of actors has led to an equal variety of motivations being put forward for the opening of OGD (we address literature in this area more fully below). One of the main questions debated by the literature has concerned the extent to which these different values are actually realized “in practice” by OGD programmes by being reflected in data usage patterns (see in this context Bates, 2013; Longo, 2011).

However, as Peled (2013) puts it, “proponents and critics [of OGD] alike use little systematic quantitative analysis to substantiate their arguments” (p. 187). The aim of this article is to respond to this call, by bringing some quantitative evidence to the debate. Our research question is simple: what explains variance in download patterns of open government data?

Download patterns are interesting and worthy of explanation because they provide a potential indicator of how open data is actually used, and hence offer a means of shedding light on the debate concerning the values that OGD programmes support. Of course, it is worth acknowledging up front that this indicator is by no means perfect. A dataset might be downloaded but not used; meaning the indicator may overstate usage patterns. A downloaded dataset might also be shared onwards by the person downloading it, perhaps through a repurposing of the data or the creation of some kind of service based on that data. In this, case, the indicator will understate usage patterns. Despite both of these caveats, we expect that download statistics will be broadly correlated with usage patterns. Hence, by analyzing

download patterns, and seeking to explain variation in these patterns, we can shed some light on the key debate surrounding OGD values.

In what remains of this section, we explore in more detail existing literature on OGD values, as well as other factors that might help explain download levels, and thus develop the hypotheses which we test in the article. As we have remarked above, open data has been attributed to a considerable variety of potential aims and potential uses (summaries can be found in, *inter alia*, Clarke & Francoli, 2014; Gonzalez-Zapata & Heeks, 2015; Jetzek 2013). From the wide-ranging literature on this subject, we identify three core values that have been attributed to open data: improving government transparency, enhancing government efficiency and stimulating enterprise. Each of these provides a potential reason for individual citizens, NGOs, private companies or other actors to download OGD. We will look at each of these in turn here. We should also note that the first two (transparency and efficiency) speak to a broad overarching value of much work on OGD, which is about enhancing citizen participation by providing citizens with the tools to effectively understand policy contexts and scrutinize government action.

Transparency in government means opening up internal processes and practices with the aim of trying to prevent ‘distortion, inequity, bias and abuse of office’ (Hood, 1991, p. 12). The aim of improving government transparency is embedded as a core principal in a wide variety of OGD programs (Bertot, Jaeger, & Grimes, 2010; Dawes, Vidiassova, & Parkhimovich, 2016; Koczanski & Sabou, 2015): indeed, in a multi-country review, De Blasio and Selva (2016) have argued that of all OGD motivations the transparency aim is the one which has been emphasized the most. Improvements in transparency generated through open data are often held up as a way of improving public trust in government (see e.g., Welch, Hinnant, & Moon, 2005), which is particularly important in the contemporary politics of many countries where levels of trust are often held to be in decline. For example, one of the major driving forces

behind the UK government's open data campaign was a desire to enhance trust following a scandal relating to expenses being claimed by politicians (Bates, 2013; Bates, 2014; Longo, 2011, see also Kelso, 2009), and open data hence became a key plank of the UK government's Transparency Agenda from 2010 onwards (Cabinet Office, 2011; Worthy, 2014, 2015).

Although in theory all types of OGD constitute by definition some sort of contribution to government transparency, in that they must reveal something about the internal working of government, in practice the literature identifies two types of data which are particularly important in this regard as they would enable direct citizen participation in the process of enhancing transparency. First, accounts of government spending can be released. These permit the identification of potential fraudulent use of money. In the UK, the desire was expressed that "an army of armchair auditors" would comb through these accounts (McClean, 2011, p. 7), and existing anecdotal examples do highlight large teams of volunteers at civil society organisations using at least some of this data (Maguire, 2011, p. 523). Second, details of the behavior of government staff can be released, especially in terms of hiring practices, meetings with lobbyists, claimed expenses, etc. Again, these would allow potential identification of corruption on the part of officials, though Bearfield and Bowman (2017) have highlighted that departments are typically less willing to release staff behavior data than they are general accounts and spending data.

If the transparency aim of open data is really being realized in practice, then we would expect datasets related to both government spending data and staff behavior data to be frequently downloaded. We thus develop our first hypothesis:

H1: Datasets that enhance the transparency of government will be downloaded more

A second area concerns enhancing government efficiency. This relates partly to enhancing the potential for citizens to play an active role both in decision making processes and evaluating the results of decisions (Dawes et al., 2016). It also relates to identifying

opportunities for reducing waste in the system: the open government movement has coincided with what is for many countries around the world a time of financial difficulty, hence the emphasizing of its potential as a means of highlighting areas where money could be saved.

Again, many types of OGD might conceivably lend themselves to the improvement of government efficiency. However, we can also highlight two types of dataset which might be particularly useful in this regard. Firstly, there is the publication of policy relevant documents that help to highlight the evidence base upon which governments are making decisions. Enabling citizens to form opinions on decisions which are being taken should subsequently enable them to play a more informed part in the decision making process (though some authors have argued that the expectation of widespread citizen use of this kind of data is unrealistic, because of the technical challenges involved in its interpretation and use - see e.g., Noveck, 2009). Second, there is the publication of service performance statistics (Bertot & Choi, 2013). Typical metrics include things such as service waiting times, educational outcomes and hospital procedure results. Release of this information allows, potentially, citizens to hold both government as a whole and individual government agencies to account, by measuring their performance against recognized statistics. It also potentially allows citizens to make a choice between different local government service providers (for example, different schools), something which highlights the potential connection of open data to the philosophy of new public management (Longo, 2011).

If the efficiency aim of open data is being realized in practice, then we would expect datasets relating to both general policy and service performance to be downloaded more. We thus develop our second hypothesis:

*H2: Datasets that enable the evaluation of the efficiency of government will be
downloaded more*

A final major use category is data for private sector innovation and enterprise (Zeleti, Ojo, & Curry, 2016). One of the major initial arguments for releasing data, especially in the UK context, was that this can act as a stimulus to industry by providing them with some of the raw materials they need to operate services (see Newbery, Bently Cipil, & Pollock, 2008; Pollock, 2008). Indeed, many authors have tended to regard private sector profit making as perhaps ultimately the most important value pursued by OGD program. For example, Gurstein (2011) argues that "the most likely immediate beneficiaries of open data are...the private sector who have the means and the interest" to use the data (however, it is also worth highlighting authors such as Worthy, who have feared that the complications of the open data program may discourage private investors – see Worthy, 2015).

Data relating to enterprise can take a wide variety of forms (for example, releasing data about house purchases and house prices has stimulated the development of services that provide information for those considering purchasing a house, whilst releasing postcode data has stimulated the development of many location based delivery services). Hence offering a precise definition of the types of data which can (and cannot) stimulate private enterprise is difficult. In this article, we hence take a narrow focus on one type of data which can be definitively said to have enterprise value, which is that released by “Public Sector Trading Funds” [PSTFs] (see Bates, 2013, p. 123-124). These funds are branches of the UK government which specialize in selling information to industry and which have a requirement to raise much of their budget through such sales. Hence, data released openly by PSTFs is likely to have high commercial value. Indeed, the open release of data which was previously provided for a fee could be seen as a kind of subsidy to this industry.

If the enterprise aim of open data is being realized in practice, then we would expect datasets which enable innovation and business opportunities to be downloaded more. We thus develop our third hypothesis:

H3: Datasets that enable private sector enterprise will be downloaded more

In addition to the potential uses of open data, it is also worth commenting on other factors that might influence OGD download patterns. We highlight two in particular here.

First there is the quality of metadata associated with the dataset (Zuiderwijk, Janssen, Choenni, & Meijer, 2012). This metadata may aid users in terms of both locating the dataset and understanding its potential uses (Najafabadi & Luna-Reyes, 2017), and therefore aid them in potentially making use of it (Wirtz, Weyerer, & Rösch, 2017). Complete and accurate metadata may also send a more general signal to users about the quality of the dataset (Hernandez-Perez, Rodriguez-Mateos, Martin-Galan, & Garcia-Moreno, 2009; Schuurman, Deshpande, & Allen, 2008; Xiong, Hu, Li, Tang, & Fan, 2011; Zuiderwijk et al., 2012): and the extent datasets appear as high quality is likely to be an important driver of their use (Vetrò et al., 2016). Indeed, research has shown that many initial open data releases lacked descriptions and important contextual information that would allow people to evaluate their quality (Bass et al., 2010; Peled, 2013; Thurston, 2012). These observations lead us to our fourth hypothesis:

H4: Datasets with more complete metadata will be downloaded more

Finally, there is the extent to which datasets are updated. Many pieces of open government data will be dynamic (for example, indicators of air quality, or spending receipts), and will require regular updating. Frequent updates may stimulate more downloads simply by requiring users to return to the dataset when it is refreshed. They may also stimulate downloads by, like complete metadata, sending a signal about quality: a dataset which has not been updated for years is likely to be one in which little effort is being invested (Vetrò et al., 2016; Viscusi, Spahiu, Maurino, & Batini, 2014; Zuiderwijk et al., 2012). Indeed, many authors have worried that published OGD datasets are now "being left to rot" online (Clarke & Margetts,

2014), concerns which are supported by empirical research (Peled, 2011). We thus develop our final hypothesis:

H5: Datasets which are updated more will be downloaded more

3. Data Collection and Variable Operationalization

In this section, we outline our data collection strategy and how we operationalize our variables. This study is based on data drawn from the website data.gov.uk. This website is designed to function as a directory for the UK's open data, focusing especially on data produced by government bodies. The website is the most visible public face of the UK's OGD program, with all departments under significant pressure to publish records on the site, meaning it is unrivalled as a potential information source (although not perfectly complete, it records an estimated 75-85% of all available open data in the UK¹).

Data was collected from the data.gov.uk website in mid-October 2013, through a combination of site usage data provided by data.gov.uk itself and the usage of several automated 'screen scraping' techniques. At the time there were just over 14,000 datasets listed on the website. However more than 5,000 of these were records for stores of data which were not actually open (one of the aims of the site is to record all data which could potentially be made open as a way of encouraging publicity about it and hence stimulating its eventual release). We removed these datasets from consideration. We also decided not to make use of datasets published by national statistics agencies², as these organisations have had a long standing remit to openly publish statistical information about the country upon which the open

¹ Telephone conversation with Antonio Acuña, Head of data.gov.uk, 05/02/2015

² In particular, we removed datasets published by the Office for National Statistics, the National Archives, and the Information Services Division in Scotland.

data movement has not really had an impact. This left us with a total of 7,540 published datasets, which we used as the sampling frame for our study.

The dependent variable used in the study is the number of downloads each dataset received. This information was also published on the data.gov.uk website³. Two caveats should be noted about the variable. First, data.gov.uk does not store datasets themselves, just links to them: and these links may well point at other publicly facing websites. Hence these download statistics will not include people who found a dataset by another means. Second, the statistics were only available for datasets that had been downloaded at least ten times⁴ (with all other datasets recorded as having had zero downloads). For both of these reasons, the number of downloads we record is expected to understate the total downloads of any given dataset. However, we have no reason to expect that this underestimation is unevenly distributed across our independent variables of interest, and hence we do not expect it to bias our analysis.

Our first three hypotheses required that we operationalize the potential use of each dataset, to allow us to determine whether transparency datasets (H1), efficiency datasets (H2) or enterprise datasets (H3) were downloaded more. We achieved this task in two steps. First, to address H1 and H2, we randomly sampled 1,000 datasets from the total available, and two of the authors coded each of these datasets into one of two “use” categories, according to whether they thought the main potential use of the dataset related to either government transparency or government efficiency. Of course, the decision to code datasets into only one usage category is a simplification of the complexity of open data. Many datasets would, in practice, potentially lend themselves to multiple aims: for example, government spending data might enhance transparency, but could also be used to improve government efficiency. Hence,

3 Currently available from <http://data.gov.uk/data/site-usage/dataset> [Accessed 02/05/2017]

4 Telephone conversation with Antonio Acuña, Head of data.gov.uk, 05/02/2015

four sub-categories were also coded, based on our literature review above which outlined different ways in which each value could be realised. In particular, for transparency datasets we coded whether the dataset related to spending data or staff data, whilst for government efficiency datasets we coded for relation to either policy context data or service performance data (a description of each of these categories can be found in Table 1). This four category classification ameliorates the problem of multiple potential uses of data by providing a more granular categorization. One hundred datasets from the sample were also double coded, which allowed us to measure intercoder reliability statistics. We had an 88% agreement for the main use categories (Krippendorff's Alpha = 0.76), and a 73% agreement for the sub-categories (Krippendorff's Alpha = 0.64).

Second, to address H3, we added to our 1,000 randomly sampled datasets a further purposive sample of all of the 42 datasets created by Public Sector Trading Funds [PSTF] that were published on data.gov.uk at the time of data collection and that were also available for free download (hence in total we had 1,042 observations). As highlighted above, identifying in general whether datasets offer any enterprise potential is quite difficult, hence we focus purely on PSTF data where the claim of some kind of enterprise value can be made in a stronger fashion: PSTFs are branches of government which specialize in selling information services, hence the data they have opened up for free is likely to have a high value. Adding these datasets hence allows us to address H3. The data.gov.uk site recorded clearly which branch of government published the dataset, making the identification of datasets published by these funds straightforward. However, as datasets published by these funds were relatively rare, we chose to purposively sample all 42 which appeared in our sample frame, rather than simply relying on them appearing in our random sample, to make sure there was a sufficient volume of observations for analysis.

Table 1: Codebook

Use Category	Use Sub-category	Description	Example Datasets
Transparency	Spending	Data on spending and purchasing decisions, and contracts, the release of much of which is required by law. This does not include however any staff costs (which are in the next category).	“Spend over 25000 in East Midlands Ambulance Service NHS Trust for June 2011”
Transparency	Staff	Data on staff of the department. This can include pay scales, organisational structures (organograms), expenses, gifts received and meetings taken with lobby groups. Can also include election results or details of hiring processes.	“Competition Commission Salaries Data” “Director's Hospitality in Wales Office”
Efficiency	Policy Context	Data on the extent or nature of policy relevant issues for a given department, many of which are collected during the course of departmental work. These are issues which the department itself seeks to manage in some way through its work. They may also include documents on current policy situation, e.g. current tax policy.	“Alcohol Profile: Alcohol-attributable mortality - females” “Special Educational Needs in England”
Efficiency	Performance	Data on the way the department itself carries out its activities, which might be used as a means of assessing performance. This can include quite general things such as waiting times and complaints. They can also be service specific, e.g. deaths during police contact, dispensed prescriptions, locations of public houses, results of surgery.	“Social Landlords Possessions and Evictions” “Improving Access to Psychological Therapies Dataset”

In order to assess H4 (regarding metainformation), we measured the amount of tags associated with the dataset (something which is likely to enhance retrievability) and the length of the dataset description in words. Of course, these are only two of many potential bits of metainformation, nevertheless they are ones which we believe are likely to be of particular importance in the initial decision which a user makes about whether to download a dataset. Finally, in order to assess H5 (on updates), we looked at the data.gov.uk “update history” for each dataset, and measured the amount of updates contained in this history. It should be noted

this history is not a perfect measure, as it does not distinguish between very small updates (for example, fixing a typographical error in the description) and very important ones (for example, uploading a dataset in a new format). However, it nevertheless provides what should be a valid proxy.

In addition to our main independent variables, we also include two control variables. First, data.gov.uk publishes data from a variety of branches of government. However, these different branches of government may have different dynamics (Barry and Bannister 2014). For example, we might expect the datasets published by local government to have a smaller potential user base and hence generate less downloads. Hence, we coded each of our datasets according to the publisher: whether it was central government, local government, the NHS, or another body (our coders had a 97% agreement for this variable, which gave a Krippendorff's alpha of 0.96). Second, we include another control variable that measured the amount of months since the dataset was published at the time our data was collected. Our measurement of total downloads took place at the time of the data collection (October 2013), however not all datasets were uploaded at the same time; and we would naturally expect datasets which had been around longer to have higher aggregate download scores. This variable allows us to control for this potential artefact in the data.

4. Analysis

Initial descriptive statistics for all the variables in our dataset are presented in Table 2. As we purposefully oversampled from Public Sector Trading Funds [PSTF], the table contains descriptive statistics for both our dataset in particular and for more general population estimates which take this oversampling into account. There are a few points worth noting from this table. In our data, the average dataset was downloaded 36 times, though our population estimate is that in general datasets have been downloaded an average of 32 times (which of course is already an indication that the enterprise datasets are downloaded more than average). However,

as will be apparent from the standard deviation, downloads have a significant right skew: the majority of datasets in fact were never downloaded (77% of the datasets in our sample), with a small number being downloaded a lot. Distribution across the category codings is also somewhat uneven: the majority of datasets fell into the “efficiency” category (comprising policy and performance), of which policy relevant datasets were the most common. However, there is a significant amount of datasets present in all of the categories apart from enterprise (albeit slightly fewer in the staff category, supporting the claims made by Bearfield and Bowman, 2017). This shows that, in terms of publication priorities, both transparency and efficiency aims are being emphasized, whilst the aim of stimulating private industry has apparently received much less attention.

It is also worth highlighting how the average values of both the downloads metric and our quality metrics vary amongst the different use categories. This is explored in Figure 1. Enterprise datasets are the most downloaded type of datasets (with approximately 150 downloads per dataset), whilst spending datasets have low levels of activity (just over 12 downloads per dataset). The other three categories are quite even at approximately 40 downloads per dataset. This offers initial support to H3, whilst undermining somewhat H1 and H2. However, of course, it also shows that all use categories have at least some activity.

Table 2: Descriptive statistics

	Total Downloads	Total Tags	Description Length	Total Updates	Months Since Publication	
Mean (data)	36.2	6.1	41	5.5	13.5	
Standard Deviation (data)	193	4.5	34	8.6	3.6	
Mean (population est.)	32.2	6.2	41	5.4	13.4	
Use Category	Transparency		Efficiency		Enterprise	Uncodable
Use Sub-category	Spending	Staff	Policy	Performance		
N (sample)	277	89	380	222	42	32
% (sample)	27%	9%	36%	21%	4%	3%
% (population est.)	28%	9%	38%	22%	0.6%	3%
Total Observations	1042					

In terms of metadata and updates, there is little variation across categories in terms of tags (all categories averaging between 4 and 6 tags per dataset), but spending datasets are notable for having much shorter descriptions than other datasets (the average spending dataset is described in around 25 words, whilst most other datasets are described in 40-50 words). However, perhaps paradoxically, spending datasets are amongst the most updated: approximately 7 updates per dataset have been applied to spending datasets, which is second only to enterprise datasets (9 per dataset). Policy context and service performance datasets, meanwhile, were the least updated.

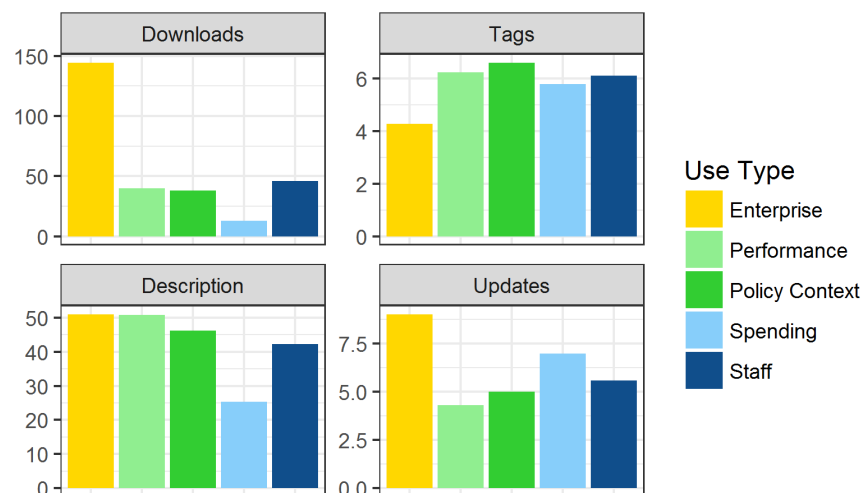


Figure 1: Download and quality measures for different types of dataset. Figures are average values per dataset. Yellow indicates “enterprise” datasets, green indicates “government efficiency” datasets and blue indicates “government transparency” datasets.

We will now move on to analyze the relationship between the number of times a dataset is downloaded and the variables highlighted above. As the majority of datasets in our data were never downloaded, we focus first on logistic regression models that looks at whether or not a dataset has been downloaded at all, putting aside for the time being variation within the 23% of datasets that were downloaded. These models are reported in Table 3 (we would highlight again that, due to the way the data was created, datasets which have zero downloads in our model may in fact have been downloaded up to nine times). Regression coefficients are

exponentiated, which means that they can be interpreted as percentage changes in the probability that a dataset is downloaded. The fit of our models was tested using the technique proposed by Esarey and Pierce (2012), which assesses the extent to which predicted probabilities generated by the model match the observed frequency of events. This method suggested that a logarithmic transformation of the updates variable was appropriate, after which the models were shown to fit well. The R^2 of models was calculated using the technique proposed by Tjur (Tjur, 2009).

A number of findings are apparent from the data (many of these will be intuitive from the findings of Figure 1). Datasets relating to private sector enterprise (PSTF datasets) were by far the most downloaded category of dataset, offering strong support to H3. In all of our regressions, both “efficiency” related datasets and “transparency” related datasets were approximately 90% less likely to be downloaded than these PSTF datasets, meaning little evidence of support for H1 and H2. When these main values are broken into sub-categories (model 1.4), we find that spending data is the least used, whilst service efficiency data is comparatively more used (though a service efficiency dataset is still 85% less likely to be downloaded than an enterprise dataset).

We also find good support for H4 and H5 in the regressions (that the quality of metadata and the amount of updates to a dataset makes a difference). Metainformation (in the form of the number of tags and the length of a description) increases the likelihood of a download: each extra tag increases the download chance by just over 10% in models 1.2-1.4. Updates also make a considerable difference: as the number of updates doubles, the chance of being downloaded increases by approximately 90%.

Table 3: Explaining Dataset Downloads

	Model 1.1	Model 1.2	Model 1.3	Model 1.4
<i>Use Categories</i>				
Transparency	0.13***	0.09***	0.08***	
Efficiency	0.16***	0.11***	0.11***	
Enterprise (reference)	1.00	1.00	1.00	
<i>Sub-categories</i>				
Spending Data				0.06***
Staff Data				0.12***
Policy Data				0.08***
Performance Data				0.15***
Enterprise (reference)				1.00
<i>Metadata and Updates</i>				
Number of Tags		1.12***	1.12***	1.12***
Description Length		1.01***	1.01***	1.01***
Number of Updates (log ₂)		1.90***	2.08***	2.13***
<i>Control Variables</i>				
Months Since Publication			0.93***	0.93***
Local			1.84	1.82
NHS			0.90	0.97
Central (reference)			1.00	1.00
Observations	1,042	1,042	1,042	1,042
Tjur R²	0.04	0.20	0.21	0.22

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

It is also worth commenting on the control variables. The amount of time a dataset had been published for had, somewhat surprisingly, a negative effect. This may be a result of attention building on the open data movement, meaning that initial data releases were comparatively underused compared to later ones. There were no effects in terms of who published the dataset: local publishers were just as likely to produce downloaded datasets as national publishers were.

Finally, it is worth noting the R^2 values. There is a low R^2 for model 1.0, which focused only on our use category data. The value rises considerably for model 1.1 (and remains relatively stable thereafter). This indicates that metadata quality measures and update levels are comparatively more important in explaining download levels than the actual use to which the data could be put.

In Table 4, we shift the focus to linear models which explore variation in download levels amongst the smaller subset of datasets which have been downloaded at least once (hence the N of both models is considerably smaller than those in Table 3). As we describe in Table 2, the distribution of downloads is heavily right skewed (or heavy tailed): we hence chose to log-transform our dependent variable, the absolute number of downloads a dataset receives. This transformation gives a good approximation of normality, and works well in this context because we are looking only at datasets that were downloaded at least 10 times (i.e. there are no 0 values). Residual plots and Tukey tests indicate no problems with model fit for either model. In order to simplify coefficient interpretation, and also to make them comparable with our first set of models, we present exponentiated regression coefficients, which means they can be interpreted as percentage increases in the overall level of downloads.

These models reinforce the findings from Table 3. Datasets relating to transparency are considerably less downloaded than enterprise ones (around 60% less). Datasets relating to efficiency are also downloaded around 40% less. This again lends good support to H3 but undermines H1 and H2. Spending data is again the least downloaded sub-category. In terms of our metadata and updates hypotheses (H4 & H5), tags are no longer significant, but description length and updates continue to be positively associated with increased downloads (doubling the number of updates increases the amount of downloads by just under 40%). Finally, it is worth noting that the publication time variable is also no longer significant.

Table 4: Further Models

	Model 2.1	Model 2.2	Model 2.3	Model 2.4
<i>Use Categories</i>				
Transparency	0.41 *	0.31 *	0.35 *	
Efficiency	0.61 ***	0.63 ***	0.63 ***	
Enterprise (reference)	1.00	1.00	1.00	
<i>Sub-categories</i>				
Spending Data				0.32 ***
Staff Data				0.38 **
Policy Data				0.74
Performance Data				0.53 *
Enterprise (reference)				1.00
<i>Metadata and Updates</i>				
Number of Tags		1.01	1.01	1.01
Description Length		1.01 **	1.01 ***	1.01 ***
Number of Updates (log ₂)		1.39 ***	1.36 ***	1.38 **
<i>Control Variables</i>				
Months Since Publication			1.04	1.04
Local			1.13	1.18
NHS			0.67 *	0.70
Central (reference)			1.00	1.00
Observations	239	239	239	239
Adjusted R²	0.01	0.18	0.19	0.19

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

5. Discussion and Conclusions

In this paper, we have offered an empirically driven exploration of the factors driving attention to downloads of individual datasets on the data.gov.uk website as a means of shedding some light on the debate about OGD usage patterns. We found good evidence that datasets related to enterprise and innovation were highly downloaded, but only weak evidence that datasets relating to either government efficiency or government transparency were being made use of. We also showed that both the metadata attached to datasets and the amount of times they were updated making a difference in terms of downloads. In this section, we will conclude by highlighting the significance of these findings both for the wider academic debate on open data and for the open data movement itself.

The first point to make in this regard is that, on the basis of our data, there is clearly a gap between publication priorities of the UK OGD programme and the values that are realized “in practice” through downloads. We showed that publishing of datasets that might support citizen participation through the evaluation of government transparency and service performance is widespread, however their usage is not. This undermines claims that OGD is really stimulating citizen participation in government (or, for that matter, that it is helping to boost the agenda of new public management). By contrast, the small amount of enterprise datasets that were published received high usage patterns. This makes it clear that the values of an OGD movement cannot be set simply by focusing on publishing priorities: further measures need to be taken to actually stimulate the usage of datasets for these priorities to be realized in practice. However, it is also important to note that all categories of datasets showed *some* levels of usage. This shows, in a sense, that OGD is advancing all of these agendas, albeit perhaps to different degrees.

A second and closely related point our results speak to is whether currently developing calls in the open data movement for a shift towards prioritization are justified (see e.g., Peled & Nahon, 2015). Prioritization can be defined simply as asking departments to publish a limited number of high quality datasets that are carefully curated and well maintained. This reduction in breadth/scope would allow departments to achieve this curation whilst nevertheless keeping spending on opening data itself limited (as a variety of authors have raised concerns about the financial sustainability of open data – see e.g., Biddick & Kash, 2013). Our data show good support for the idea that prioritization would produce benefits: we have shown that frequently updating a dataset and curating its metadata drives downloads. We have also shown that, currently, resources may not be being allocated in the most efficient way: for example, spending data is one of the most frequently published type of dataset, and also one that receives considerable updates, yet spending datasets are amongst the least used. By contrast, service

performance datasets are used much more frequently, yet are updated less than spending datasets. Indeed, almost 80% of datasets observed were never downloaded at all: continuing to invest resources in their maintenance is of questionable value.

Of course, if we accept the need for prioritization, the key question becomes which datasets to prioritize. It is difficult to know before the fact which stores of data are worth opening up: indeed, one of the central tenets of the OGD movement is that surprising uses of data may be discovered by chance and hence data should be published regardless of whether a clear use case is known in advance. In this respect, we would highlight our unexpected finding that the length of time a dataset has been published for does not increase either the likelihood of it being downloaded or the number of downloads it receives: rather, it seems that if downloads are going to happen, they happen quite quickly. This means that departments could pursue a publishing strategy which revolved around publishing and curating a dataset only for a very short amount of time (say, a few months), after which point they should be confident about whether any interest will develop or not. Those working in civil society or journalism might raise the legitimate concern that such a download focussed strategy would prioritize highly downloaded private sector enterprise data over data that might support the goal of citizen participation. However, in this respect, it is worth highlighting again that both transparency and efficiency related datasets do show *some* patterns of downloads. This strategy would hence also allow departments to quickly find out which citizen participation datasets are going to be used, and subsequently maintain a focus on them.

In concluding the article, it is important to highlight some of the limitations to our study, and hence outline areas for further research. One key area here is our limitation to only one country's OGD activities (the UK). Further multi-country research could usefully clarify whether these findings are unique to the UK or reflect more general patterns. There are also limitations in the way our variables are measured. Our coding of the potential use of a dataset

simplifies the complexity of open data by putting each dataset in one category only (and we lacked a robust means of coding for the enterprise value of a dataset). The updates variable reflects updates to the record on data.gov.uk itself, but does not offer us means of distinguishing between minor updates and important ones (and indeed does not reflect the fact that different datasets might require different levels of updates). Furthermore, and perhaps most importantly, the value of downloads themselves is not always apparent: while a dataset cannot be used without being downloaded, a useful application of a dataset might emerge from just one download, whilst many hundreds of people might download a dataset which appears important but turns out to be of little value. As research continues in these areas, we will understand more about how variation in OGD downloads can be explained, and the consequences of this for the wider values of the OGD movement.

References

- Bass, G., Brian, D., Fuchs, M., Schwartz, A., McDermott, P., Miller, E., & Weismann, A. (2010). Letter Encouraging the Administration to Improve Its Open Government Efforts. Retrieved April 7, 2017, from <http://www.pogo.org/our-work/letters/2010/gs-og-20100203.html?print=t>
- Bates, J. (2012). “This is what modern deregulation looks like”: co-optation and contestation in the shaping of the UK’s Open Government Data Initiative. *The Journal of Community Informatics*, 8(2), 1–13.
- Bates, J. (2013). The politics of open government data: A neo-gramscian analysis of the United Kingdom’s Open Government Data initiative. *Policy & Internet*, 5(1), 118–137.
- Bates, J. (2014). The strategic importance of information policy for the contemporary neoliberal state: The case of Open Government Data in the United Kingdom. *Government Information Quarterly*, 31(3), 388–395. <https://doi.org/10.1016/j.giq.2014.02.009>

- Bearfield, D. A., & Bowman, A. O. (2017). Can you find it on the web? An assessment of municipal e-government transparency. *The American Review of Public Administration*, 47(2), 172–188. <https://doi.org/10.1177/0275074015627694>
- Bertot, J. C., & Choi, H. (2013). Big data and e-government: issues, policies, and recommendations. In *Proceedings of the 14th Annual International Conference on Digital Government Research* (p. 1). Quebec, Canada: ACM Press. <https://doi.org/10.1145/2479724.2479730>
- 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. <https://doi.org/10.1016/j.giq.2010.03.001>
- Biddick, M., & Kash, W. (2013, August). Federal Government IT Priorities: Vision Vs. Reality - InformationWeek. *InformationWeek Government*.
- Bright, J., & Margetts, H. (2016). Big Data and Public Policy: Can it succeed where e-participation has failed? *Policy & Internet*, 8(3), 218-224.
- Cabinet Office. (2011). Making Open Data Real: A Public Consultation. Retrieved April 11, 2017, from <http://www.cabinetoffice.gov.uk/resource-library/making-open-data-real-public-consultation>
- Clarke, A., & Margetts, H. (2014). Governments and citizens getting to know each other? Open, closed, and big data in public management reform. *Policy & Internet*, 6(4), 393–417.
- Davies, T. G. (2012). The Promises and Perils of Open Government Data. *Journal of Community Informatics*, 8(2).
- Dawes, S. S., Vidasova, L., & Parkhimovich, O. (2016). Planning and designing open government data programs: An ecosystem approach. *Government Information Quarterly*, 33(1), 15–27. <https://doi.org/10.1016/j.giq.2016.01.003>

- De Blasio, E., & Selva, D. (2016). Why Choose Open Government? Motivations for the Adoption of Open Government Policies in Four European Countries. *Policy & Internet*, 8(3), 225–247. <https://doi.org/10.1002/poi3.118>
- Esarey, J., & Pierce, A. (2012). Assessing fit quality and testing for misspecification in binary-dependent variable models. *Political Analysis*, 20(4), 480–500.
- Gonzalez-Zapata, F., & Heeks, R. (2015). The multiple meanings of open government data: Understanding different stakeholders and their perspectives. *Government Information Quarterly*, 32(4), 441–452. <https://doi.org/10.1016/j.giq.2015.09.001>
- Gurin, J. (2014). Open governments, open data: A new lever for transparency, citizen engagement, and economic growth. *SAIS Review of International Affairs*, 34(1), 71–82.
- Gurstein, M. (2011). Open data: Empowering the empowered or effective data use for everyone? *First Monday*, 16(2).
- Hernandez-Perez, T., Rodriguez-Mateos, D., Martin-Galan, B., & Garcia-Moreno, M. (2009). Use of metadata in Spanish electronic e-government: the challenges of interoperability. *Revista Española de Documentación Científica*, 32(4), 67–91.
- Hood, C. (1991). A public management for all seasons? *Public Administration*, 69(1), 3–19. <https://doi.org/10.1111/j.1467-9299.1991.tb00779.x>
- Kalampokis, E., Tambouris, E., & Tarabanis, K. (2011). A classification scheme for open government data: towards linking decentralised data. *International Journal of Web Engineering and Technology*, 6(3), 266–285. <https://doi.org/10.1504/IJWET.2011.040725>
- Kelso, A. (2009). Parliament on its knees: MPs' expenses and the crisis of transparency at Westminster. *The Political Quarterly*, 80(3), 329–338.
- Koczanski, A., & Sabou, M. (2015). Sustainability Implications of Open Government Data: A Cross-Regional Study. In *Proceedings of the ACM Web Science Conference*. New York,

- USA: ACM Press. <https://doi.org/10.1145/2786451.2786463>
- Longo, J. (2011). # OpenData: Digital-era governance thoroughbred or new public management Trojan horse? *Public Policy and Governance Review*, 2(2), 38–51.
- Maguire, S. (2011). Can data deliver better government? *The Political Quarterly*, 82(4), 522–525. <https://doi.org/10.1111/j.1467-923X.2011.02249.x>
- Margetts, H., & Dunleavy, P. (2013). The second wave of digital-era governance: a quasi-paradigm for government on the Web. *Philosophical Transactions of The Royal Society A*, 371, 1–17. <https://doi.org/10.1098/Rsta.2012.0382>
- McClean, T. (2011). Not with a bang but a whimper: the politics of accountability and open data in the UK. In *APSA 2011 Annual Meeting Paper*.
- Najafabadi, M., & Luna-Reyes, L. (2017). Open Government Data Ecosystems: A Closed-Loop Perspective. In *Proceedings of the 50th Hawaii International Conference on System Sciences*.
- Newbery, D., Bently Cipil, L., & Pollock, R. (2008). *Models of Public Sector Information Provision via Trading Funds Copying and Redistribution*.
- Noveck, B. (2009). *Wiki government: how technology can make government better, democracy stronger, and citizens more powerful*. Brookings Institution Press.
- 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. <https://doi.org/10.1002/asi.21622>
- Peled, A. (2013). Re-Designing open data 2.0. *JeDEM*, 5(2), 187–199.
- Peled, A., & Nahon, K. (2015). Towards Open Data for Political Accountability: Examining the US and UK models. *iConference 2015 Proceedings*, (Idc), 1–12. <https://doi.org/10.2139/ssrn.2546017>
- Pollock, R. (2008). *The economics of public sector information*.

- Schuurman, N., Deshpande, A., & Allen, D. M. (2008). Data integration across borders: a case study of the Abbotsford-Sumas aquifer. *JAWRA Journal of the American Water Resources Association*, 44(4), 921–934. <https://doi.org/10.1111/j.1752-1688.2008.00192.x>
- Thurston, A. (2012). Trustworthy records and open data. *The Journal of Community Informatics*, 8(2).
- Tjur, T. (2009). Coefficients of determination in logistic regression models—A new proposal: The coefficient of discrimination. *The American Statistician*, 63(4), 366–372.
- Vetrò, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., & Morando, F. (2016). Open data quality measurement framework: Definition and application to Open Government Data. *Government Information Quarterly*, 33(2), 325–337. <https://doi.org/10.1016/j.giq.2016.02.001>
- Viscusi, G., Spahiu, B., Maurino, A., & Batini, C. (2014). Compliance with open government data policies: An empirical assessment of Italian local public administrations. *Information Polity*, 19(3,4), 263–275.
- Wang, H.-J., & Lo, J. (2016). Adoption of open government data among government agencies. *Government Information Quarterly*, 33(1), 80–88. <https://doi.org/10.1016/j.giq.2015.11.004>
- Welch, E. W., Hinnant, C. C., & Moon, M. J. (2004). Linking citizen satisfaction with e-government and trust in government. *Journal of Public Administration Research and Theory*, 15(3), 371–391. <https://doi.org/10.1093/jopart/mui021>
- Wirtz, B. W., Weyerer, J. C., & Rösch, M. (2017). Citizen and Open Government: An Empirical Analysis of Antecedents of Open Government Data. *International Journal of Public Administration*, 1–13. <https://doi.org/10.1080/01900692.2016.1263659>
- Worthy, B. (2014). *Making Transparency Stick: The Complex Dynamics of Open Data*.
- Worthy, B. (2015). The impact of open data in the UK: Complex, unpredictable, and political.

Public Administration, 93(3), 788–805. <https://doi.org/10.1111/padm.12166>

Xiong, J., Hu, Y., Li, G., Tang, R., & Fan, Z. (2011). Metadata distribution and consistency techniques for large-scale cluster file systems. *IEEE Transactions on Parallel and Distributed Systems*, 22(5), 803–816.

Zeleti, F. A., Ojo, A., & Curry, E. (2016). Exploring the economic value of open government data. *Government Information Quarterly*, 33(3), 535–551. <https://doi.org/10.1016/j.giq.2016.01.008>

Zuiderwijk, A., Janssen, M., Choenni, S., & Meijer, R. (2012). Socio-technical impediments of open data. *Electronic Journal of E-Government*, 10(2), 156–172.