Articles, special issue: ODSCSD

User Centred Methods for Measuring the Quality of Open Data

Mark Frank
University of Southampton, United Kingdom
Corresponding Author.
mark.frank@soton.ac.uk

Johanna Walker
University of Southampton, United Kingdom
J.C.Walker@soton.ac.uk

A project to identify metrics for assessing the quality of open data based on the needs of small voluntary sector organisations in the UK and India. We used small structured workshops to identify users' key problems and then worked from those problems to understand how open data can help address them and what the key attributes must be for successful use. We then piloted different metrics that could be used to measure the presence of those attributes. This user-centred approach to open data research highlighted some fundamental issues with expanding the use of open data from its enthusiast base.

Current metrics of the quality of open data are mostly based around the production of datasets and technical standards and not around the needs of potential users. Data portals often track progress by reporting the number of datasets that conform to the five stars of linked open data (Berners-Lee, 2006). More sophisticated attempts such as the Open Data Barometer (Davies, 2013) include measuring the progress of open data through the presence of datasets in certain categories, whether they meet legal criteria, and the existence of technical functions such as the ability to download the dataset in bulk. Caplan et al. have aggregated a list of initiatives to measure the value of open data (Caplan et al., 2014). Although there are some metrics that are specific to a sector and take into account the content of the datasets, these are derived 'top–down', for example, by assessing what properties the data needs to conform to regulations. While they provide a valuable perspective and are relatively easy to implement, there is no evidence that these top–down approaches address users' most pressing concerns. As such, they are weakly linked to the impact of open data. The highly technical nature of open data in

Frank, M., Walker, J. (2016). User centred methods for measuring the quality of open data. *The Journal of Community Informatics*, 12 (2),(Special issue on Open Data for Social Change and Sustainable Development), 47-68.

Date submitted: 2015-07-20. Date accepted: 2016-05-09.

practice harbours the potential for citizen users to become disengaged from the process of shaping and constructing relevant quality characteristics. This prompted our research question: What is the nature of open data metrics derived from user requirements and are they viable?

This project explores 'bottom–up' methods for measuring the quality of open data that are grounded in what users need from data to perform core functions; thus producing metrics that are more directly related to the impact of the data. It has two aims – developing a methodology for identifying metrics that are relevant to a specific user context, and identifying and evaluating some metrics for one such group. Our focus is on practical metrics that are either already in use or could reasonably be used in the near term. As such, they need to strike the right balance between being easy to implement and relevant: easy to implement in the sense that they can be used without excessive effort; relevant in that they are closely correlated with key desirable characteristics of the datasets in a given context.

We believe that in exploring the perspective of the non-specialist user in some detail we are starting to address a significant gap in open data research which has, to date, focused on the implementation and strategic implications of open data. We have done this by using a combination of new methods such as structured workshops and role-plays which have their roots in traditional IT methods for investigating user requirements but have been adapted for open data.

The literature on data metrics and open data metrics often refers to metadata. However, the scope of the word 'metadata' has not yet been clearly defined. It can be limited to structured data about a dataset (for example a data portal may have fields for author and such-like) or expanded to include any document which describes the dataset. We have used metadata in the first sense and used supporting documentation to refer to any documents which give information about the dataset.

The paper is structured as follows: First we review data quality, metrics and methodologies literature, and examine the criteria for a good metric before outlining our methods. Then we present the results of the workshops, and derive and pilot our metrics. We conclude with a discussion and implications for the field and further work.

Literature review

Data quality characteristics

The literature reviewed fell into three areas concerned with data quality and the assessment thereof. Information systems and database management literature provides general data quality research that is applicable to open data (Batini et al., 2009; Scannapieco & Catarci, 2002). Linked open data (LOD) research examines machine-readable assessments of data (Behkamal et al., 2014; Erickson et al., 2013). Finally, the socio-technical open data field has looked at data quality and measurement through the frame of barriers to usage of data (Martin et al., 2013; Zuiderwijk et al., 2012).

There is no common standard of definitions for data quality (Scannapieco, Missier & Batini, 2005). Wang (1998) memorably defines it as 'fitness for use' but this lacks measurability and

requires more detail in order to be operationalised. There is not even consensus on the meaning of the terms used to outline the dimensions, for instance, timeliness may be used to refer to the average age of a source or the extent to which the data is appropriately up-to-date (Batini et al., 2009).

Scannepieco and Cartarci (2002) survey six sets of dimensions of data quality, representing a variety of contexts and define *data quality as a set of dimensions including accuracy, completeness, consistency (at format level and at instance level), timeliness, interpretability and accessibility* as the most important factors from a list of 23. Of these, accuracy and completeness were the only factors that were cited in all six sets of dimensions. Even then, few of these are absolute measures and they are often relative to a specific dataset or application. For instance, current data may in fact be too late for a specific application, and therefore not timely (Scannapieco & Catarci, 2002; Scannapieco, Missier & Batini, 2005).

Data quality is also subject to tradeoffs, often between timeliness and another dimension such as accuracy or completeness (Scannapieco, Missier & Batini, 2005). These will vary with the domain, as may the attributes themselves.

Reviewing the LOD literature, Zaveri et al. (2012) identified 26 dimensions of data quality. Compared to the dimensions cited above there was a far greater emphasis on provenance and other trust-based metrics, reflecting the distributed nature of open data. They also focus on *amount-of-data, noting, an appropriate volume of data, in terms of quantity and coverage, should be a main aim of a dataset provider.* Licensing and interlinking are also key attributes of LOD. Accurate metadata is also vital for findability and cataloguing; reflecting the fact that open data is no longer defined within an organisation and thus needs to be discoverable by anyone (Maali, Cyganiak & Peristeras, 2010; Reiche, Hofig & Schieferdecker, 2014). Reiche et al. (2014) propose metadata quality as a characteristic, being the fitness of the metadata to make use of the data it is describing. This speaks to the unpredictable use of open data.

Socio-technical research such as Barry and Bannister (2014) and Zuiderwijk et al. (2012) primarily derive their quality characteristics from interviews and workshops with civil service (publishers) and academia. Discovery is a frequent issue in the literature (Conradie & Choenni 2014; Keen et al., 2013). It identifies interpretability – the users' ability to comprehend the data in a set, and a number of aspects of interoperability, including formats, endurance, varying quality and licensing, reflecting the envisaged use of combining datasets from various sources.

Although data quality assessment is well-established, the added complexities of open data – the autonomy and openness – mean that a new set of data quality areas have started being added to the literature. However, this is becoming divided into two areas – the machine-readablility issues addressed by the LOD field and the more people-oriented issues identified by the socio-technical studies. A more unifying approach may be called for. Additionally, these studies are approached from a publisher or machine-intermediary point of view, and the field of user-derived metrics is in its infancy.

Metrics

To operationalize data quality there must be a way to assess it. It is clear from the preceding section that open data quality characteristics include some that are novel to the general corpus of data quality work. Consequently, new metrics such as tau, 'the percentage of datasets up to date in a data catalogue' (Atz, 2014) are being developed to engage with these attributes alongside those previously identified.

Many metrics are based on the technological structure of LOD such as examining consistency through the ratio of triples using similar properties (Behkamal et al., 2014) or by applying an automated test such as the Flesch-Kincaid Reading Ease. These metrics have the value of automation but cannot be performed ad hoc on all open datasets.

Bizer and Cyganiak (2009) suggest three classifications of metrics for information quality filtering. The first, structured content, could be assessed statistically by analysing the structure. Context-based metrics usually rely on a third-party check, for example, against a list of trusted providers, or metadata analysis. Ratings-based metrics (such as the Five Stars of Linked Open Data) are about the information or information provider, and depend on some subjectivity or skill of the assessor creating the rating, which may often be produced algorithmically.

The above suggests that the creation of metrics must address a number of different dimensions. They pertain not only to the data but its creator, and not solely to its presentation or structure but to its meaning. They may change over time and are only useful in so far as they serve the purpose of the user of the metric. There are few existing metrics that can be applied without tools to any kind of dataset.

Data quality assessment methodologies

Batini et al. (2009) define a data quality methodology as an operational description of a logical process to assess and improve the quality of data. This makes explicit the idea that the attempt to understand data quality is made not in and of itself, but in the service of utilisation. Pipino et al. (2002) state that there is a lack of 'fundamental principles for […] developing useable metrics in practice and note that it is not practicable to create 'one size fits all', but rather these fundamental principles should be sought. Su and Jin (2004) suggest these might be derived in three ways: intuitively, systemically and user-based.

Batini et al. (2009) review 13 methodologies for the assessment of data quality. These indicate a variety of approaches are employed, from questionnaires through subjective and objective metrics to statistical analyses. Most methodologies are appropriate for distributed systems, however; they generally apply to co-operative situations, where most of the parties can be considered to be aware of each other, which cannot be true for open data.

This suggests there is a potential need to create a method for deriving and addressing fundamental principles amongst a variety of open data users that is appropriate for autonomous use in an extremely distributed system. It would of necessity be an audit model – one where only use decisions, and not the data itself, can be improved by users, and not the data itself. Su and Jin (2004) suggest user-based methodologies are problematic as being

most subjective, but they are also robust in that for a specific group of users they identify their exact concerns.

What makes a good metric?

The preceding section suggests that open data requires a definition of data quality that is broad and loosely defined. We want to take into account more than the content of the datasets, e.g. discoverability, and we want to apply our metrics in environments where the aims are not clearly defined. We therefore propose a correspondingly broad and loose definition of a metric as 'An observable characteristic of one or more datasets that acts as a proxy for some other characteristic of interest which is less easy to observe'.

In this paper we will refer to the characteristic of interest as an **attribute** of the data.

Choosing a metric for an attribute is similar to choosing an operational definition for a concept. Like an operational definition a good metric should be **valid** (closely correlated with the attribute of interest) and reliable (gives consistent results over time and between observers). It should measure attributes that matter and should be sufficiently closely tied to the attribute that it is difficult to 'game' the metric. In addition, during the course of the project we identified these desirable characteristics of a metric:

- **Discriminatory**. The metric should be sensitive enough to discriminate between common values of the attribute.
- **Efficient**. The less time and resource required to use it the better. In some contexts poor efficiency can lead to poor validity and reliability. If the aim is to measure a large number of datasets for a large variety of users (e.g. the Open Data Barometer) then poor efficiency may force the assessors to use a small convenience sample which potentially introduces both bias and sampling error.
- **Transferable**. The same metric can be used in a variety of different contexts – in our case a range of different user groups – and across cultural and economic variation.
- **Comparable**. This is an extension of transferability. If a metric is comparable not only is the metric transferable to a wide variety of contexts but the results can be meaningfully compared. Ideally this would result in a universal standard that transcends cultures and applications.

We propose that the ideal metric would rate highly on all of these criteria: an efficient assessment (e.g. automated) that could be quickly run against a large group of datasets with high validity and reliability giving results that are comparable for a wide range of contexts. In practice there is often a trade-off between these criteria. For example, we may accept limited transferability as a cost of increased validity. Metrics lie on a spectrum between the most **subjective**, which involve a high degree of judgement, and the most **objective**, which involve little judgement. Greater objectivity is associated with greater reliability. More objective metrics may also allow for automation (for example, automatically inspecting metadata for the recent updates) which can lead to greater efficiency – although this is not always true. However, it is often hard to find an objective metric that is valid, and a subjective metric with suitable guidance and support for the assessor can have more utility. In this project we

focused on metrics that are towards the objective end of the scale, while noting the importance of subjective metrics as an alternative.

While these are all relevant criteria for a good metric, the quality of a metric depends ultimately on whether it fulfils its purpose. In the context of open data this purpose can be as varied as comparing progress of data providers, estimating the impact of open data, or evaluating the usability of a data portal. In this paper we assume the purpose is to determine the value of a group of open datasets to a defined community of users. The group of datasets is deliberately left loosely defined. It might be as large as all data published by a national government or as small as a portal from a specialist provider.'

There are also cultural considerations in choosing a metric. A metric that is hard to understand or which has an obscure relationship with the attribute it is intended to measure is unlikely to be accepted in practice – even if it is efficient, valid and reliable. We therefore focused on straightforward metrics which have a direct and logical relationship with the attribute they were intended to measure.

Methodology

Our interest was in identifying metrics that reflect the core concerns of users who are not open data specialist or enthusiasts. We needed to uncover the relationships between users' problems, the information that would help them solve those problems, and the data that might supply that information. This depth and subtlety of insight would be very difficult to uncover through large-scale quantitative approachs. Therefore, we used qualitative methods to work closely with two small selected groups of potential users to explore in depth if/how open data could contribute to their work. Our approach differed from many open data events for users (e.g. hackathons) which are aimed at promoting open data and developing a community. Such events are vital to the open data movement, but they have two characteristics which made them unsuitable for our purposes: the participants come because they have some kind of special interest in open data and typically they are presented with some data and then work to find good uses for it. This means the participants are not at all representative of the populations and is a serious skewing of the sample from the point of view of our research. The data to problems approach can be very successful, but the danger is that it produces interesting solutions to relatively minor problems – problems selected because they are amenable to open data solutions – not because they are core concerns of the users. To ensure that we were addressing significant problems we reversed the order – starting with identifying problems that most concerned our users and then trying to discover how open data might help them with those problems. We were determined not to have preconceptions as to what matters about the data and let the users tell us what mattered to them. Only then did we go on to consider suitable metrics.

Selection of user groups

We had prior contact with voluntary sector organisations supporting the homeless in Winchester, UK. These organisations were a suitable group of users for this study because:

- We wanted to start with a well-established open data culture such as the UK, which should minimise confounding variables relating to the early stages of the availability of open data;
- These particular organisations are not 'open data aware'. There is no special requirement for IT or data skills and in this respect they are typical of thousands of voluntary sector organisations; and
- Preliminary discussion with a voluntary sector coordinator established that the UK voluntary housing sector has several real business issues that might be addressed by open data.

We used unstructured interviews and e-mail to identify three such organisations and confirm that they were suitable to participate.

While we wished to restrict the study to a well-defined and limited set of user problems, we were keen to develop metrics that transcended cultural and developmental barriers and were relevant to emerging economies. We identified four organisations in Gujerat, India, who met similar criteria in terms of being relatively small, focused on a specific urban geographical area (Ahmedabad) and working with the homeless. The difference in stage of developmental growth and governmental policies meant we could not expect to exactly mirror the activities of the UK organisations, but they all delivered programmes to support the homeless or poorly housed in Ahmedabad, which we felt was sufficiently similar so as not to affect our methodology.

Identifying key attributes

Each group of users attended two structured workshops (see Table 1) to jointly develop and document the user's story including:

- The problems they have to solve;
- The information they need to solve them;
- How open data can contribute to this information;
- How this data can be found; and
- What attributes of the data are required for using the data in this context and what attributes (if any) are preventing them from using it.

These attributes were interpreted very broadly ranging from technical format and licensing arrangements through to details of the content, availability of support, currency and provenance. The key consideration was to discover, without preconceptions, which attributes are truly significant for the users.

The first workshop was used to identify the most important problems facing the group and what additional information would most help them address those problems.

In the period between the workshops the researchers tried to identify open datasets that could potentially supply at least some of the missing information. To do this:

Table 1 *Workshops and participants*

Workshop location	Workshop type	Number of attendees	Number and type of organisations
Winchester UK	Problem and information need specification (2 hrs)	4	3 2 x temporary shelter 1 x social housing
Winchester UK	Data selection and assessment (2 hrs)	4	3 2 x temporary shelter 1 x social housing
Ahmedabad, India	Problem and information need specification (3 hrs)	6	4 1 x state budget analysis 1 x slum rehousing 1 x migrant workers 1 x basic services for slum dwellers
Ahmedabad, India	Data selection and assessment (3hrs)	4	3 1 x education intervention 1 x basic services for slum dwellers 1 x slum rehousing

- We discarded problems where we knew the required information would not be available as open data (e.g. information about named individuals);
- We searched the relevant government open data portal (data.gov.uk and data.gov.in) using keywords derived from the information the users needed;
- In the UK we searched specialist data portals such as the Shelter Databank, and in India we used exemplar projects such as Transparent Chennai and the Karnataka Learning Partnership as a guide to what might be available in Gujerat and in what form;
- We took advice from specialists such as the Department for Communities and Local Government (DCLG) in the UK and DataMeet in India; and
- We used Google as this sometimes proved more efficient at finding data than the government portal search mechanisms.

As a result we selected a small number of datasets (see Appendix) that came close to providing part of the information that the users had identified in the first workshop.

At the second workshop each group was presented with the selected datasets, asked to review them and decide, through a group discussion, whether and how they could they could be used in practice. This allowed us to identify important dataset attributes at high level. We then asked the participants to select one or two datasets out of those discussed that had the most potential (one dataset in the UK and two in India). They were asked to annotate these datasets with comments on what would make them useful and their annotations were encoded. We then asked the participants to tell the story of a typical situation in which they might use these

datasets. The objective of this 'role-play' was to recreate, to a limited extent, the environment in which our participants would be using the datasets and thus uncover any important requirements derived from their working environment which may not be obvious when focusing on the datasets themselves in a workshop environment. This allowed us to confirm and add details to the key dataset attributes. The output of the two workshops was a list of attributes that the users agreed were important if open data was to be useful. Details of the structure of both workshops are the appendices.

Following the workshops we investigated possible metrics that both help to identify whether the selected attributes are present and are practical to implement. For any given attribute there are potentially an indefinitely large number of ways of measuring it. However, due the approach adopted, in practice there were few candidates for any given attribute.

We wanted to evaluate each metric against our outlined criteria. We did this by piloting the metrics against a sample of datasets relevant to our community. This comprised ten datasets from the UK and five from India. The UK datasets were selected from a list generated for the Open Data Institute Housing Open Data Challenge[1]. We chose these because they had been selected as being relevant to housing by experts independent from our project. From that list we selected the first dataset from each provider to give a cross-section of providers. There was no equivalent list for India so we included the datasets that were used in the second workshop as we knew they were relevant to our users. We piloted the metrics against all the datasets and noted:

- The metric score for each dataset;
- How confident we felt in the score (a measure of objectivity and therefore reliability); and
- How easy it was to make the assessment (a measure of efficiency).

We then assessed the validity and transferability/comparability of the metric on theoretical grounds.

Results

Problems and information needs

The organisations attending the UK and Indian workshops had much in common in terms of their biggest problems and the information that would help them. For example, organisations in both countries struggled with identifying which welfare benefits individuals were entitled to. A full list of problems and information requirements is provided in the appendices.

Attributes

The most important result for this project was that five attributes of datasets were identified as being significant by this group of users. There is no accepted terminology for these

1 http://www.nesta.org.uk/closed-housing-open-data-challenge

attributes (W3C 2015) so we have used terms that we believe are unlikely to cause confusion based on current usage.

- **Discoverability**. Datasets can be discovered via many different routes including general purpose search engines such as Google; government data portals such as data.gov.uk or data.gov.in; specialist intermediaries such the UK Shelter Databank; and word of mouth. As described in the methodology section, the researchers searched for relevant datasets based on the information needs of the participants using a combination of these routes. This proved to be very demanding even with the support of subject matter experts. For example, we were advised by the Ahmedabad Centre for Environmental Planning and Technology University that slum data in Excel format existed but we were unable to locate it using either Google or data.gov.in. The participants commented that it would be a very significant issue had they undertaken the search themselves.

  It's a full time job [tracking down the appropriate data] isn't it? (UK)
  It is an issue with people who have not looked at this data, they would not put in those titles (India)

- **Granularity**. To address some of their most pressing problems the attendees needed information about individual people and potential homes, such as knowing the benefit status of a homeless person or the addresses of landlords that will accept lodgers receiving state benefit.

  It isn't sufficient to know rates of acceptance in Winchester. It has to be number 2 something street. (UK)

  For privacy reasons open data is most unlikely to provide this information which severely limits the utility of open data in this context. For other problems it was useful to have data aggregated at higher levels such as city or district within city. The most useful level varied according to the dataset and specific problem being addressed. For example, generic data on the cost of crime and health services is sufficient for a funding application for additional resources. But data on the cost of specific crimes and treatments is required when making the case for providing permanent housing to an individual client with a particular profile.

  [It's] good for research on aggregate level but in terms of providing service [we] need more detail (India)
  If it is not linked to a specific ward, how useful can it be? It can give you good overview of what is happening in the area but not for an intervention. (India)

- **Immediate intelligibility**. While the attendees were very competent in their field, they often found datasets hard to interpret. An apparently straightforward field such as the number of homeless people in a city, immediately raised questions of interpretation. At one extreme someone might be considered homeless if they are forced to leave their home for temporary reasons such as a flood, at the other extreme they might be someone sleeping the streets who is not known to the local authority. Without further explanation and information about how the data is collected it is impossible to know what the figure means. Similar issues of interpretation arose for almost all the datasets examined. Over half (26 out of 51) of the annotations on the datasets expressed a need for more information. On the other hand, the role-play revealed that participants did not typically have much time to understand datasets and therefore the time to

understand the data is a critical aspect of this attribute. For example, in the UK workshop, the attendees explored using a dataset on costs of health treatments which, while initially hard to understand, was explained by a 58-page supporting document. Despite the presence of the supporting document, this dataset was not useful to the community as it would take too long to use the document.

One doesn't have too much time to read through it (UK)
Is the question that was asked 'where are you getting the water you drink' or 'where is your nearest drinking water'? (India)
What is difference between 'no exclusive room' and 'one room'? (India)
You need the documentary information that supports this. (India)

- **Trusted/authoritative**. This was a particular concern in India where participants were extremely sceptical about the veracity of government data. For example, they assumed data on slums was incomplete because, by law, local government has to support inhabitants of slums and thus there is an incentive not to include slums in the data. In the UK, participants felt it was important that data came from an authoritative source and that they understood how it was collected, particularly if they used that data as part of a funding application.
  For slum surveys, who is collecting the data and who is implementing it? If the same agency collects data about what are the gaps in provisions it may not [unintelligible] to collect, if a third party is doing the survey and paid directly by central government then it can conduct fair, impartial surveys. (India)

- **Linkable to other data**. Both countries identified a need to discover relationships between data items that were not available in the datasets in the published format but must have existed in the raw data. For example, the UK participants needed to compare the cost of their interventions with the cost of crime and health interventions for the homeless. This can be seen as a requirement to have data in the appropriate format. Data presented in PDF or Excel format had been curated by the publishers and selected, in most cases, from a larger set of data. This meant certain choices about what would be displayed in that particular dataset had been made by the publishers, and it was not possible to 're-attach' other data that had been excluded. Technologies such as LOD or HATEOAS could potentially address this requirement but current tools are beyond the scope of these users.

  To use this we need some other variables as well, like this many people are having own house, but not infrastructure, and we need geographical area. (India)
  If you can cross-check, [with certain income data] whether someone has a TV and a fridge, this can verify whether their income is correct. (India)

Metrics

Following the workshops we proposed and assessed metrics for each attribute.

Discoverability

Metrics for discoverability presented significant difficulties. It is not practical to develop a metric that takes into account all possible routes that may be used to discover a dataset. Therefore any proposed metric must be relative not only to the data being sought but also the

route being used. Even within this constrained context, we struggled to identify a useful metric.

We considered the following metric which assumes that the route to discovering the dataset is via a keyword search (which is frequently the case): 'Given a set of keywords to search for a dataset how many alternative datasets are generated and what proportion of the alternatives include the required data'.

This might act as a proxy for how quickly the route leads to the target data. It clearly raises difficulties in choosing appropriate keywords but this need not matter if the result is not sensitive to the precise choice of keywords. To test this we searched for datasets on **housing stock** using different combinations of keywords on data.gov.uk. However, in practice the results appear to be extremely sensitive to the choice of keywords. For example, using *dwelling* as a synonym for housing and supply as a synonym for stock we obtained the results in Table 2:

Table 2 *Sensitivity of different keywords when searching for data on 'housing stock' on data.*gov.*uk*

Keywords	Number of datasets returned
Dwelling stock	102
Housing stock	154
Dwelling supply	17
Housing supply	78

Subjective metrics also present a problem as they require the assessor to put themselves in the shoes of a typical user who may have very different skills and attitudes to the assessor. We therefore focused on an approach based closely on our own problems in discovering appropriate data. All of these tasks proved challenging:

- Identifying the organisation likely to supply appropriate data;
- Finding a data portal or other data search service used by that organization;
- Using the data portal/search service to produce a list of possible candidate datasets that was not too long and which we were confident had included any datasets of interest. A key concern here was that data might be referred to by a synonym (e.g. 'dwelling' instead of 'house' and as a result we might not find it); and
- Examining the list of candidate datasets to see if they contained the data of interest quickly.

In addition:

- We had no way of knowing whether appropriate datasets existed other than finding them and wasted a lot of time looking for data that we never found and may not have existed; and

- When we found datasets in one format (e.g. PDF) it was often challenging to determine if they were available in a more useful format such as Excel.

We constructed a metric based on the availability of solutions to these challenges (with the exception of identifying the organisation for owning the data, for which we were not able to identify any solution). For any given dataset we awarded one point for each of the following:

- The publisher/owner of the data has an open data portal (or similar search mechanism);
- The publisher/owner of that portal publishes an updated, searchable list of datasets;
- The publisher/owner of that portal publishes an updated, searchable list of datasets with synonyms;
- The publisher/owner of that portal publishes a list of datasets which are known to exist but are not currently available. This would limit the time wasted on abortive searches; and
- The dataset is accompanied by a list of alternative formats. Publishing in multiple formats is recommended by the World Wide Web consortium (W3C, 2015).

We piloted this metric against the sample datasets. Many of these features are features of the data portal to which the dataset belongs. All but one of the UK datasets were served by data.gov.uk and none of them included a list of alternative formats. Therefore, the majority of datasets had the same score and discrimination was low. The Indian datasets, and the one UK dataset not served by data.gov.uk, had different scores and we therefore believe the low discrimination was a function of the limited datasets on which we piloted the metric. Discovering the features of the portal was time consuming, but once they were established it took only a few minutes to rate a dataset according to this metric and there was little judgement involved. We therefore found the metric to be reliable and rated its efficiency as medium. It also appears to be transferable and comparable. There is a major issue over validity because it does not measure the difficulty of the initial challenge of identifying an owner. Nevertheless, it is directly related to other challenges in discovering data and therefore has some validity in that sphere.

Granularity

Although the required level of granularity varies according to the problem being addressed, it is always possible to combine data with greater granularity into higher levels with less granularity, while the reverse is generally not possible. This suggests that a metric could be based on the principle that the more granular the data the better. Although there is some potential for doing this automatically, the technology is not currently at the level where we could pilot it. For the foreseeable future, most datasets will require human intervention and subject matter knowledge to recognise different levels of granularity. For a well-defined context it can be straightforward to specify the levels of granularity that are most meaningful for a type of data. An assessor can then assess datasets according to whether they include these levels.

While the effective granularity of a dataset is in theory a function of the relationship between the different fields in the dataset, for the purposes of this metric we only considered fields

independently. We piloted this approach using five levels of geographical granularity on the sample datasets. For the UK datasets we used National (i.e. UK), Country, County, City, Address (i.e. the identified building); for the Indian datasets we used National, State, District, City (or Village), Address. The results were promising. For the UK datasets, in every case it was possible to identify the level almost immediately on opening the sample dataset with very little requirement for personal judgement (in some cases data was presented by local authority and a small level of background knowledge and judgement was required to decide whether this should be classified as county or city level). This was a little harder for the Indian datasets as in several cases a dataset comprised several tables in a PDF document some of which were at state level (which was assumed from the context of the document) and some at district level. Nevertheless, the level of granularity was apparent for individual tables. Thus, the metric had high efficiency and reliability. The metric had such a direct relationship to the attribute of granularity it was hard to doubt its validity. All five levels were found among the datasets suggesting good discrimination. There seems to be little problem in theory applying the metric to other types of granularity and other data although the results would not be comparable.

This approach is simple and direct but limited. It measures how granular a dataset is in a very specific context and requires prior specification of the context. We considered a more sophisticated approach, which is to measure the scope of a dataset to support different levels of granularity in a broader context including being combined with other data. This relies on the fact that some data facilitates aggregation while other data does not. For example, in the UK post codes allow for aggregating data geographically but house names do not. We refer to such linking data as class data as it indicates a class to which the individual can be allocated. Some class data is more generic than other class data in that it is not specific to a domain. The post code can be used almost wherever there is a requirement for geographical aggregation. The residential status of a house (owner-occupied, private rented, public rented, vacant) can be used for aggregation but is context specific. We explored a scale from 1 to 4 where the levels are:

1) Includes aggregated data only e.g. national statistics;
2) Includes individual unit level data but with no generic class data;
3) Includes generic class data; and
4) Includes more than one form of generic class data.

We piloted this scale against the sample datasets. The metric proved to be reliable and efficient; in every case it was possible to classify the dataset on inspection with minimal judgement required. All datasets scored either 1 or 3. This casts some doubt on its discrimination as it suggests it may effectively be a two-level metric. The fact that the key data is generic suggests that the metric would have good transferability and comparability. It is harder to assess the validity. The metric measures the ability of data to participate in aggregation but it is by no means certain that this translates into aggregations that are useful for our community of users. It is a concept which needs further research.

Immediate intelligibility

To assess immediate intelligibility we considered using an automated test of data readability – similar to the Flesch-Kincaid test for document readability (Flesch, 1948) – as a metric. However, current tests of readability are designed for documents not data, and a test would need to be developed. Even then, it is not clear that such a test would be valid. The intelligibility problems that the participants came across were a function of background knowledge rather than the specific words that were used and it is hard to see how an automated readability test would detect this type of problem. We therefore focused on measuring the availability of supporting information.

A simple approach is to rate datasets on the accessibility of supporting information bearing in mind that speed of intelligibility is vital. A possible scale might be (with increasing value):

1) Supporting documentation does not exist;
2) Supporting documentation exists but as a document which has to be found separately from the data;
3) Supporting documentation is found at the same time as the data (e.g. the link to the document is next to the link to the data in the search);
4) Supporting documentation can be immediately accessed from within the dataset but it is not context sensitive. This might be a link to the documentation or text contained within the dataset;
5) Supporting documentation can be accessed immediately from within the dataset and it is context sensitive so that users can directly access information about a specific item of concern. This might be a link to a specific point in the documentation or the text contained within the dataset; thus eliminating the need to search the documentation and speeding up access to the relevant material.

We piloted this against the sample datasets with limited success. Evaluating the level of support involved some subjectivity in many cases e.g. Does a footnote in a spreadsheet count as level 5 support? Does supporting documentation fall into level 2 or 3? Can we be sure there is no supporting documentation because we have failed to find it? The process was efficient in that it was possible to determine the level of support almost immediately upon opening the dataset and there was good discrimination with results including levels 1, 3, 4 and 5. There is no problem in principle in transferring the metric to other domains and the results would be comparable if they are simply interpreted as measuring the speed of availability of supporting documentation.

The biggest issue is validity. The metric raises some issues where datasets are available in multiple formats. Some formats such as LOD and Excel facilitate linking to supporting documentation better than others such as CSV. We intend the metric to refer to the available format of the dataset that has the best links to supporting documentation. However, as discussed under the section on discoverability it may not be easy to determine all the available formats for a given dataset. Also the metric takes no account of the quality of the supporting documentation. A point identified by Reiche et al. (2014) when discussing metadata quality is that it is one thing to quickly locate supporting documentation, but another to understand it and get the required support.

Trustworthiness

Our users trusted (or mistrusted) data for a variety of reasons:

- They know (or don't know) how it was collected and processed;
- It comes from a trusted source;
- It is internally consistent and plausible; and/or
- It is consistent with other external sources.

The first two reasons suggest metrics based on provenance. The second two suggest metrics based on consistency tests. There has been theoretical work on metrics of consistency and plausibility, see for example Prat and Madnick (2008), but this has not resulted in any usable tools or methods. We therefore primarily considered metrics based on provenance. There is extensive literature about systems for tracking provenance (Suen et al., 2013) and standards for exchanging information about provenance (Buneman, Khanna & Tan, 2000; Moreau et al., 2011; Pignotti, Corsar & Edwards, 2011). But this does not suggest metrics that could be implemented in the short term.

We explored a relatively simple approach – evaluating whether the data or supporting documentation answers key questions that are relevant to provenance. Corsar and Edwards (2012) make the case that open data metadata, in addition to common requirements such as date and authors, should:

> *If possible, expand on this with a description of the dataset's provenance. This includes describing the processes involved (e.g. screen scraping, data transformation) the entities used or generated (e.g. the downloaded timetable webpage and the generated timetable spreadsheet), and the agents (e.g. users, agencies, organisations) involved in the creation of the dataset. This record should also include the relationships between them.*

Ram and Liu (2009) propose seven questions (the seven Ws) which can provide the basis for this approach:

1) What is the data?
2) Who author/ organisation which created it?
3) Why was the dataset created?
4) (W)How was it collected - what events lead up to its collection?
5) When was it collected?
6) Where was it collected?
7) Which instruments were used to collect it?

The same approach can be used objectively – simply recording whether the question has been answered – or more subjectively, but potentially with greater validity, by instructing an assessor to judge the quality of the answer. We piloted the objective approach on the sample datasets, awarding from 0 to 7 points to each dataset – one point for each of the 7 Ws for which there was an answer in the dataset or supporting documentation. Despite adopting the objective approach, it proved difficult to judge whether some of the questions had been answered or not. For example, if data refers to occupancy levels in 2012 is that sufficient information to answer the question: When was it collected? And it was time-consuming to inspect the documentation to see if the questions were answered. So we assessed reliability as medium to low and efficiency as medium. Discrimination was good with datasets being scored as low as 1 and as high as 7 (3 was the only level not represented). The metric is not context specific so it can be transferred and there seems no reason why the results should not be compared.

The key concern is over the validity of the metric. In many cases the data scored quite low on the metric but was from a highly trustworthy source such as the UK Office of National Statistics. The metric takes no account of reputation-based trust (Artz & Gil, 2007), where trustworthiness of the data is derived from the trustworthiness of the source. A more sophisticated approach might take this into account.

Linkable to other data

The Five Stars of Linked Open Data is an accepted and easily applicable measure of open data format standards which reflects the user need to be able to discover unanticipated relationships among data. It can be interpreted not just as a technical standard but as a 'soft' standard, for example, making data findable and putting it in context. As this metric has already been studied and used extensively we did not do any further evaluation and accepted that it has high reliability, discrimination, transferability and comparability. We assessed the validity as medium because, while the metric can be a valid measure of the technical scope for exploring new relationships in the right hands, there are several reasons it might fail in practice. The ability to link data, especially using automated methods, depends not only on the technical format but the structure and choice of data in the dataset. Users need to have the skills, time and resources to use the data and make the linkages. Even developers found it challenging when LOD was first introduced by the UK government (Sheridan & Tennison 2010). A more valid metric might reflect the value of presenting data in multiple formats which would allow for users in different contexts to manipulate it in different ways. However, this entails a basis for weighting the value of different formats for different communities of users which would require further research.

Summary

Table 3 summarises our assessment of the proposed metrics. We rate each metric as high, medium or low against the criteria (except that comparability and transferability are combined for conciseness). It is important to bear in mind that these assessments were based on using the metrics on a small sample of datasets relevant to the users we worked with. Nevertheless the results indicate that there is potential for viable metrics for the key attributes for this community based round simple and direct proxies.

Table 3 *Summary of assessment of metrics*

Attribute	Metric	Valid	Reliable	Discriminatory	Transferable/ Comparable	Efficient
Discoverability	5-point scale indicating presence of features which enable discoverability.	Medium	High	Low (for the sample datasets but this may be an exception)	High	Medium Large effort to assess portal – shared over many datasets.
Granularity	Observe whether dataset includes preselected (context-specific) levels	High	High	High	Transferable but not comparable	High
	Levels based on presence of generic class data	Medium	High	Medium	High	High
Intelligibility	Scale for quality of link to supporting information	Medium	Low	High	Transferable but not comparable	High
Trustworthiness	Number of answers to the 7 Ws	Low	Medium	High	High	Medium
Linkable to other data	5 Stars of Open Data	Medium	High	High	High	High

Discussion

The data attributes

Some of the data attributes that emerged from the workshops reflected known concerns with open data. Granularity is a key element of the primary principle of the 8 principles of Open Government Data (Malmud, 2007) and the need to link data is one of the fundamental tenets of the open data movement. On the other hand, the emphasis on being able to comprehend the data quickly was less predictable. Timeliness, which is often at the centre of such metrics, was only mentioned once in the workshops, and this was in the context of how often data was collected, rather than when it was published.

As part of our methodology the study was deliberately limited to a specific set of users. This was an advantage in that the participants agreed about problems, information and attributes; but it also places limits on how widely the conclusions might be applied. For example, large campaigning charities have staff whose sole task is to analyse evidence and who have the time to understand data. They are likely to be less concerned with immediate intelligibility and more concerned with ensuring data is up-to-date so that it can be used in campaigns. Other communities may have different key attributes which would require different metrics.

Nevertheless, there is no reason why the same approach to identifying appropriate metrics for a particular group of users cannot be used in other contexts.

All of the five attributes that emerged were important in both countries. There was a difference in emphasis, possibly because of the relative maturity of open data in the two countries. Although it was still a challenge, discoverability was much easier in the UK where there are several useful portals at both central and local government levels, and there is relatively good coordination between local and central government in the collection and distribution of statistics. In India there is still a lack of effective portals at the local level and there is less coordination between central and local government. For example, while collecting data is often a function of central government, it frequently fails to provide sufficient granularity for local government who have to regulate and administer programmes based on the data. Trustworthiness was a concern in the UK but the participants felt that the reputation of the provider might be sufficient to make the data trustworthy. The Indian participants required a stronger understanding of provenance and possible unreliability before they trusted the data.

Metrics

The aim of the project was to investigate a different approach to open data metrics. The resulting metrics should be considered simply as ideas for discussion, refinement and further research. However, the attributes they measure have been recognised in other literature and therefore there is good reason to suppose they are applicable to a wider user community.

Validity is fundamental to any metric. The temptation to measure something just because it is reliable and efficient is very strong – but should be resisted. Measuring the wrong thing well is worse than measuring the right thing badly. We assessed only one of our metrics as high validity. This was the first of the two measures of granularity – simply inspecting datasets to see if they contained data which met predetermined levels of granularity. This limited the metric to a very specific context and meant the metric had no comparability. Another possible way to increase validity is to take a more subjective approach but at a cost in reliability; for example, by asking an assessor to judge whether a dataset is quickly intelligible rather than seeking an objective proxy for intelligibility. By definition subjective approaches require judgement which may differ from one assessor to another and thus affect reliability. However, subjective approaches do not necessarily increase validity. The assessors of open data are unlikely to be representative users and may struggle to adopt the role of a user with a different skill set, attitude and environment. To some extent this can be mitigated by supplying the assessor with strong guidance, but this requires the resources to develop and test the guidance which may well have to be repeated for different user groups. We hope that our approach has at least focused attention on what really needs to be measured (the attributes) and thus raises the profile of validity.

There are developing technologies and standards which may provide better metrics in the future. Bizer, Heath and Berners-Lee (2009) suggest that a PageRank type algorithm – TrustRank – could eventually emerge for measuring trustworthiness, but this would be dependent on a great many more datasets in any one domain being available. The W3C Working Group on Data on the Web Best Practices recently identified indicating the non-availability of datasets as one of its draft best practices (W3C 2015, sec. Best Practice 21),

and the Sunlight Foundation's fourth principle of Open Government Data recommends a full inventory of available data and helpful context on what is unlikely to be released (The Sunlight Foundation, 2015) which would address some discoverability issues.

Further implications

Several lessons emerged beyond the aims of this project. It was apparent that a lot of the information that participants find critical to solving problems is information about processes, for example, how to recognise and respond to different kinds of 'legal high'. This parallels Heald's distinction between process and event transparency (Heald, 2011). This kind of information on how or why is not typically available through open data.

The workshops suggest that more research is needed into what constitutes data literacy, and what skills it might comprise to increase the impact of open data. Our users were as competent as anyone could reasonably expect: technically (they included experienced Internet and Excel users); in their knowledge of the subject matter; and also in their understanding the significance of data. Yet, they struggled to interpret aspects of every dataset that was presented to them. Furthermore, the knowledge they needed to interpret the data was specific and not necessarily applicable to other datasets. This suggests that there is scope for more work to be done on the best way to provide context to any given dataset, which would go some way to removing this onerous requirement from the user.

The study was limited to one small set of users in two different environments and similar user-oriented research needs to be done for a wider variety of user groups. Each group is likely to have its own key problems, information needs and attributes, and it is only by conducting a range of similar studies will it be possible to determine the scope of any conclusions. The methodology needs to refined and could be expanded. We learned several lessons which are noted in the appendices. It would be fruitful to ask users to find their data (as opposed to doing it for them as we did) and to get their feedback on the metrics. In addition we strongly believe that our approach of starting with the needs of users who are not open data enthusiasts needs to be used more widely – not just for developing metrics but for gaining a greater understanding of the limitations of open data and how it should move forward if it is go beyond the domain of specialists.

Acknowledgments

The funding for this work has been provided through the World Wide Web Foundation 'Open Data for Development Fund' to support the 'Open Government Partnership Open Data Working Group' work, through grant 107722 from Canada's International Development Research Centre (web.idrc.ca). Find out more at http://www.opengovpartnership.org/groups/opendata.

References

Artz, D., & Gil, Y. (2007). A survey of trust in computer science and the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, *5* (2), 58–71.

Atz, U. (2014). The Tau of Data: A New Metric to Assess the Timeliness of Data in Catalogues. In P. Parycek & N. Edelmann (Eds.), *CeDEM14 Conference for E-Democracy and Open Government* (Vol. 22, pp. 147–162). Krems, Austria.

Barry, E., & Bannister, F. (2014). Barriers to open data release: A view from the top. *Information Polity*, *19* (1), 129–152.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for Data Quality Assessment and Improvement. *ACM Comput. Surv.*, *41*(3), 16:1–16:52. http://doi.org/10.1145/1541880.1541883

Behkamal, B., Kahani, M., Bagheri, E., & Jeremic, Z. (2014). A Metrics-driven Approach for Quality Assessment of Linked Open Data. *Journal of Theoretical and Applied Electronic Commerce Research*, *9* (2), 64–79. http://doi.org/10.4067/S0718-18762014000200006

Berners-Lee, T. (2006, July). Linked Data: Design Issues. Retrieved 2 December 2014, from http://www.w3.org/DesignIssues/LinkedData.html

Bizer, C., & Cyganiak, R. (2009). Quality-driven information filtering using the WIQA policy framework. *Web Semantics: Science, Services and Agents on the World Wide Web*, *7* (1), 1–10.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, (Special Issue on Linked Data).

Buneman, P., Khanna, S., & Tan, W.-C. (2000). Data Provenance: Some Basic Issues. In S. Kapoor & S. Prasad (Eds.), (pp. 87–93). Springer Berlin Heidelberg.

Caplan, R., Davies, T., Wadud, A., Verhulst, S., Alonso, J., & Farhan, H. (2014). Towards common methods for assessing open data: workshop report & draft framework. Retrieved from http://opendataresearch.org/content/2014/709/towards-common-methods-assessing-open-data-workshop-report-draft-framework

Conradie, P., & Choenni, S. (2014). On the barriers for local government releasing open data. *Government Information Quarterly*. http://doi.org/10.1016/j.giq.2014.01.003

Corsar, D., & Edwards, P. (2012). Enhancing Open Data with Provenance. *Digital Futures*. Aberdeen.

Davies, T. (2013). *Open Data Barometer*.

Erickson, J. S., Viswanathan, A., Shinavier, J., Shi, Y., & Hendler, J. A. (2013). Open Government Data: A Data Analytics Approach. *IEEE Intelligent Systems*, *28*(5).

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3).

Heald, D. (2011). When Transparency meets Surveillance: External Monitoring of Country Public Finances (pp. 19–20). Newark, New Jersey.

Keen, J., Calinescu, R., Paige, R., & Rooksby, J. (2013). Big Data + Politics = Open Data : The Case of Healthcare Data in England. *Policy and the Internet*, *5*(2), 228–243.

Maali, F., Cyganiak, R., & Peristeras, V. (2010). Enabling interoperability of government data catalogues (pp. 339–350). Springer.

Malmud, C. (2007). The Annotated 8 principles of Open Government Data.

Martin, S., Foulonneau, M., Turki, S., Ihadjadene, M., Paris, U., & Tudor, P. R. C. H. (2013). Risk Analysis to Overcome Barriers to Open Data. *Electronic Journal of e-Government*, *11* (1), 348–359.

Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., … den Bussche, J. Van. (2011). The Open Provenance Model core specification (v1.1). *Future Generation Computer Systems*, *27* (6), 743–756. http://doi.org/10.1016/j.future.2010.07.005

Pignotti, E., Corsar, D., & Edwards, P. (2011). Provenance Principles for Open Data. Nottingham, UK.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, *45* (4), 211–218.

Prat, N., & Madnick, S. (2008). Measuring Data Believability: A Provenance Approach (pp. 393–393). http://doi.org/10.1109/HICSS.2008.243

Ram, S., & Liu, J. (2009). A New Perspective on Semantics of Data Provenance (Vol. 526). Washington DC.

Reiche, K. J., Hofig, E., & Schieferdecker, I. (2014). Assessment and Visualization of Metadata Quality for Open Government Data. In P. Parycek & N. Edelmann (Eds.), *CeDEM14 Conference for E-Democracy and Open Government*. Krems, Austria.

Scannapieco, M., & Catarci, T. (2002). Data quality under a computer science perspective. *Archivi & Computer*, *2*, 1–15.

Scannapieco, M., Missier, P., & Batini, C. (2005). Data Quality at a Glance. *Datenbak-Spektrum*, *14*, 6–14.

Sheridan, J., & Tennison, J. (2010). Linking UK Government Data. Raleigh, N.C.

Su, Y., & Jin, Z. (2004). A Methodology For Information Quality Assessment In The Designing And Manufacturing Processes Of Mechanical Products.

Suen, C. H., Ko, R. K. L., Tan, Y. S., Jagadpramana, P., & Lee, B. S. (2013). S2Logger: End-to-End Data Tracking Mechanism for Cloud Data Provenance (pp. 594–602). Los Angeles, CA, USA. http://doi.org/10.1109/TrustCom.2013.73

The Sunlight Foundation. (2015). Open Data Policy Guidelines. Retrieved from http://sunlightfoundation.com/opendataguidelines/

W3C. (2015, February). Data on the Web Best Practices. Retrieved March 4, 2015, from http://www.w3.org/TR/dwbp/

Wang, R. Y. (1998). A product perspective on total data quality management. *Communications of the ACM*, *41* (2), 58–65.

Zaveri, A. (2012). Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. Quality assessment methodologies for linked open data. *Semantic Web Journal. Submitted On*, *12*, 14.

Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., & Alibaks, R. S. (2012). Socio-technical Impediments of Open Data. *Electronic Journal of e-Government*, *10* (2), 156–172.