The Tau of Data: A New Metric to Assess the Timeliness of Data in Catalogues

Ulrich Atz

Open Data Institute, 65 Clifton Street, EC2A 4JE, London, UK
ulrich.atz@theodi.org

Abstract: *We review existing studies that assess the timeliness of data in catalogues and propose a new metric: tau, the percentage of datasets up-to-date in a data catalogue. Obsolete data will stifle innovation, whereas spotlighting timeliness can foster efficiency and support the sustainability of the open data ecosystem, for example, by encouraging automated publication of data.We validate the tau in three case studies: the World Bank catalogue, the UK data catalogue (data.gov.uk) and the London Datastore. For the World Bank and London we find that roughly half of the datasets are up-to-date, whereas data.gov.uk performs worse. However, there are considerable caveats when it comes to missing and undocumented metadata. The tau of data is easy to implement, can be readily interpreted and be generalised with further parameters across all data catalogues.*

Keywords: *open data, timeliness, data quality, data catalogue, datastore*

Acknowledgement: *I wish to thank Dr Tom Heath, Head of Research, Dr Jeni Tennison, Technical Director, at the Open Data Institute and four anonymous reviewers for valuable feedback and suggestions.*

The research presented in this paper was supported in part by the OpenDataMonitor project, with support from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n°611988.

The Tau of Data: A New Metric to Assess the Timeliness of Data in Catalogues

Introduction

Governments and institutions often publish open data as part of a collection. A minimum requirement for these data catalogues are discoverable and up-to-date datasets.[1] To the best of our knowledge, there is no rigorous quantitative analysis on the timeliness of data in catalogues because of the varied (and arguably messy) landscape of open data portals. We chose a case study approach and propose a new metric that may allow for comparisons in the future.

The timeliness of data matters for several reasons, for example:

- Businesses and startups using open data want to trust the publisher that the data will remain available and up-to-date. Obsolete data will stifle *innovation*.
- A measure of timeliness will put the spotlight on the update cycle. Automating this process can lead to gains in *efficiency* in publishing, analysis and re-use.
- Timely data being produced more efficiently is a necessary, though perhaps not sufficient, condition for a *sustainable* open data ecosystem.

Findings

Here are some of the general findings:

1. **Missing timeliness**. More evidence points towards the hypothesis that many datasets are *not updated* with a regular schedule or at all.
2. **Poor metadata**. Ironically, the data about open data seems to be incomplete, undocumented or hard to find. On the plus side, there is enough metadata available to make this statement.
3. A **new metric tau** (τ) to assess the **timeliness** of data. The London Datastore scores "ok" with 0.52 (i.e., slightly more than half of the datasets are updated according to schedule.) For our case studies this could easily be improved by releasing *monthly* datasets on a more regular basis.

Summary of the three case studies

The **World Bank** updates its data catalogues on an irregular schedule. There are 102 datasets that have revision dates and are set to be updated. Overall slightly less than half of the datasets were updated according to schedule (τ = 0.46). The number of missing dates is relatively large, which is a substantial caveat.

[1] The interested reader can find an extensive, global list of data catalogues at http://datacatalogs.org

The **UK data catalogue** has an irregular release cycle. Even worse, only around 25% (4,000) of datasets include data on update frequency. This may be one of the reason why it performs so poorly on the τ with 0.25. The UK data catalogue updated almost ¾ of its datasets in 2013.

The **London Datastore** hosts around 550 datasets. They were released with stark differences for releases in some months over the last three years. More importantly, the updates are not concentrated in recent months, which suggests a poor update cycle. The τ = 0.52 is optimistic because its metadata update variable possibly includes minor updates.

On the timeliness of data

What is an up-to-date dataset? This is not a trivial question and is a function of the forecast update frequency. A dataset that is only released annually will probably only be updated once a year. Yet knowing the timeliness is important and Lindman, Rossi and Tuunainen (2013) write in their *Open Data Services: Research Agenda* that "from the services perspective, [...], the most critical questions revolve around achieving sufficient timeliness of the data."

Fast-paced communications streams like Twitter are an indication of the trends in data. Implicitely this may also increase the pressure to improve the timeliness of data. Tinati *et al.* (2012, 2013) and Gurin (2014) allude to the changing pace, as well as how the publication of data is improving efficiency between government departments, councils and local authorities. We are not aware of any studies that look at the relative importance of timeliness compared to, for instance, quality or relevance. Ultimately, all are part of an exemplar publication of open data.[2]

In "Annex A: Improving data on Whitehall" of the *Whitehall Monitor 2013*, Bouchal, Stephen and Bull (2013), urge publishers, among other suggestions, to "explain the update cycle" and "clearly signpost periodicity". They also argue that "the evidence suggests that there are improvements in [data quality], but there is still a long way to go."

Furthermore, a dataset should always contain current data. Some datasets such as the UK census may be released according to their pre-defined schedule, but are too far behind users' need. Here we will not discuss the questions of what is current data and focus on the timeliness of data catalogues.

Methodology

The varied landscape of open data portals prohibits a simple quantitative analysis. (Despite the limited number of data portal software such as CKAN.) Some have tried by looking at the Socrata metadata, though face numerous caveats (Levine, 2013).

We chose a case study approach by looking at three case studies: the World Bank, the UK data catalogue and the London Datastore. The three cases were selected because we have existing relationships with the publishers and they represent different regional levels (international, national and local, respectively). Maali, Cyganiak and Peristeras (2010) selected seven data catalogues in a similar fashion.

Yin (2009) argues that case selection is crucial and we were careful to choose cases that allow for analytical generalisation (as opposed to statistical generalisation from surveys).

An additional difficulty is that an uneven release cycle can stem from

[2] On how to publish open data, compare further: https://certificates.theodi.org

- datasets that differ substantially in their update cycle; and
- “waves” of updating datasets unrelated to the availability at the source.

Without additional information we cannot distinguish between the two explanations. Even if we know how often datasets have to be updated, without a standardised metric the answer will only be suggestive. We therefore devised an unambiguous metric, the tau of data (see next section). However, "garbage in, garbage out"[3], its usefulness relies on the underlying quality of the metadata. In our case studies the amount of missing metadata poses substantial reason for concern for the reliability of individual metrics. However, this is unrelated to the construct validity which we believe to be high because of the relativly simple nature of the metric.

The tau of data

We propose a new metric for measuring the timeliness of data. The *tau* (τ) can be interpreted as *the percentage of datasets up-to-date in a data catalogue*. Before we move on to its definition, a concept of timeliness.

$$timeliness = \mathbf{I}\left(\frac{update\ frequency}{today - last\ substantial\ update}\right)$$

Here, *timeliness*, is simply an indicator (1 or 0) whether the dataset's last substantial update was a longer time ago than an anticipated release based on the reported update frequency. **`I()`** is the indicator function[4] and takes 1 if the ratio is bigger than one and 0 otherwise. For example, a dataset with an annual cycle and an update in 2013, would yield 1. A dataset with a monthly cycle and a last major update in October would result in a 0 (based on Dec 2013).

By *substantial* we mean a new release of the data. Minor updates, for example if someone discovers a typo in the title and corrects it, should not appear as an update. The τ of a data catalogue is the average across datasets (indicated by the subscript *i).*

$$\tau = \frac{1}{N}\sum_{i=1}^{N}\mathbf{I}\left(\frac{update\ frequency_i \cdot \lambda + \delta}{today - last\ substantial\ update_i}\right)$$

N is the number of datasets in the catalogue. We can make this more flexible by introducing two parameters in a linear form, delta (δ) and lambda (λ): the "leeway" of days we allow the data catalogue for updating. The δ is a fixed number of days applicable to all datasets, for example one day for processing. λ, on the other hand, is relative to the update frequency. For example, we may allow for a 10% increase for data cleaning, which for an annual dataset implies 1.2 months and for a monthly dataset 3 days in tolerance.[5]

A τ of 0 means the catalogue has no up-to-date datasets. A τ of 1 means all datasets are up-to-date. Datasets with missing metadata are omitted; if the percentage of missing information is substantial (indicative > 5%), the researcher has to take additional care in interpreting the results.

[3] http://www.worldwidewords.org/qa/qa-gar1.htm, accessed 2013-12-09.

[4] http://turing.une.edu.au/~stat354/notes/node16.html, accessed 2013-12-09.

[5] We have explored a few different values to see how much the tau changes in this instance. It mostly affects the scores for monthly publications around the magnitude of 10%.

Table 1: Proposed benchmarks for different levels of tau

τ (tau)	**timeliness of data**
0.9 - 1	exemplar
0.7 - 0.9	standard
0.5 - 0.7	ok
0.25 - 0.5	poor
0 - 0.25	obsolete

By design the tau of data is limited to a binary, up-to-date or not, classification. In its extreme case this means that a data catalogue that is one day late is recorded in the same way as one that fails to update the datasets completely. However, we deem this extreme case very unlikely and argue that benefits of simplicity outweigh a more complicated approach.

To implement the τ, you need to record two variables: the *last substantial update* and a *standardised update frequency* for all datasets (preferably in days; in our analysis we found a wide range of values used). We recommend the standard set of update frequencies defined by Dublin Core (for an overview see Kurtz, 2013).

Validating the tau of data in three case studies[6]

1. The World Bank data catalogue

The original metadata[7] contains 162 catalogues. For the columns update frequency and last revision date information for around 15% are missing. Missing data are treated as missing at random and are removed.

Table 2: Last updates (revision dates) by year in the World Bank Catalogue

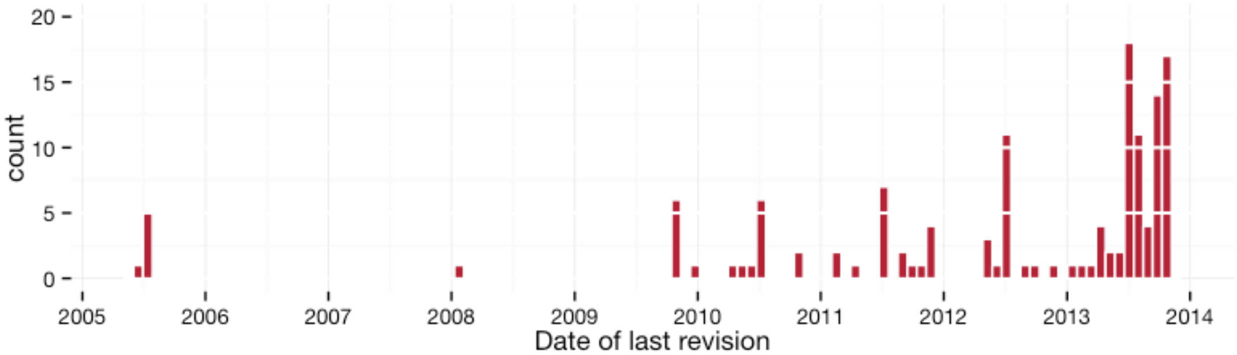
2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**
6	0	0	1	6	12	18	18	75

We can see that the World Bank updated more than half of its data catalogues in 2013. The histogram in figure 1 exhibits the full distribution.

[6] The R code and workspace for the analysis can be found on GitHub: https://github.com/theodi/R-projects/tree/master/data-portal-analysis

[7] http://datacatalog.worldbank.org, accessed 2013-10-15.

Figure 1: World Bank data catalogue last revision date

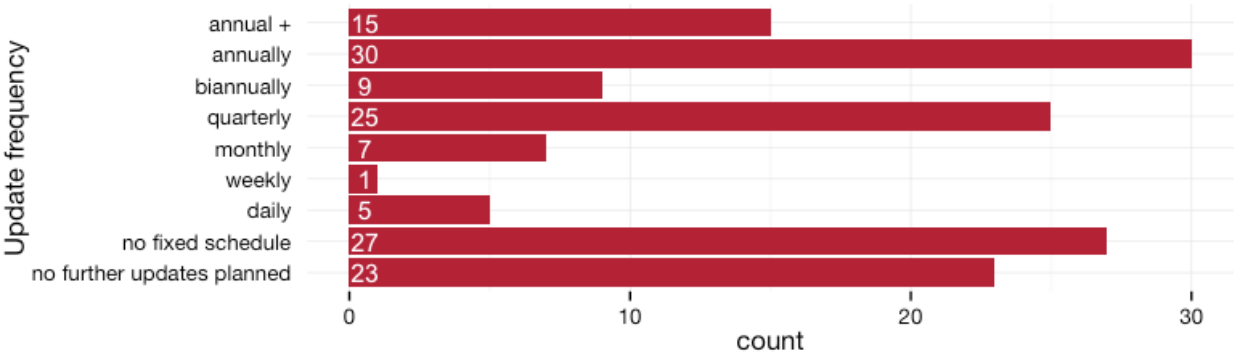


The 2005 figures are an artefact because in the original data they are dated as 1905.

It is also clear that the update cycle has clear spikes in certain months and is not uniform over the years.

What happens if we take the update frequency into account? Not all datasets have to be updated within the last year. Below we can see that some update frequencies are longer than a year or some releases are not even planned to be updated. If we disregard these particular cases, we may bias our metric.

Figure 2: World Bank data catalogue update frequency



The **overall τ = 0.46**, which means slightly less than half of the datasets are updated according to schedule.

Table 3: The World Bank's tau breaks down as follows

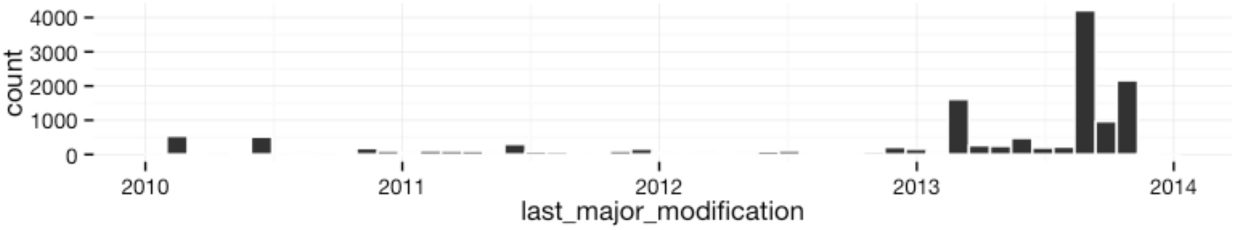
update frequency	**τ**	**count**
daily	0.00	5
weekly	1.00	1
monthly	0.00	7
quarterly	0.80	25
biannually	0.33	9
annually	0.33	30
annual +	0.33	15
no fixed schedule	0.59	27
overall	*0.46*	*119*

To account for a small delay in publishing we added one day to the update frequency (the δ). Here, and in the other two case studies, we allow a 10% in relative delay (the λ). Furthermore, we assume "no fixed schedule" to be two years, which is generous. We set "annual +" to mean a thousand days.

2. The UK data catalogue (data.gov.uk)

The UK data catalogue, data.gov.uk, hosts more than 16,000 datasets, although at least 4,000 of them are currently unpublished.[8] According to the variable *last_major_modification*, which excludes minor revisions, most datasets were updated recently. Almost ¾ of them were updated in 2013.

Figure 3: The UK data catalogue, histogram of last major modification

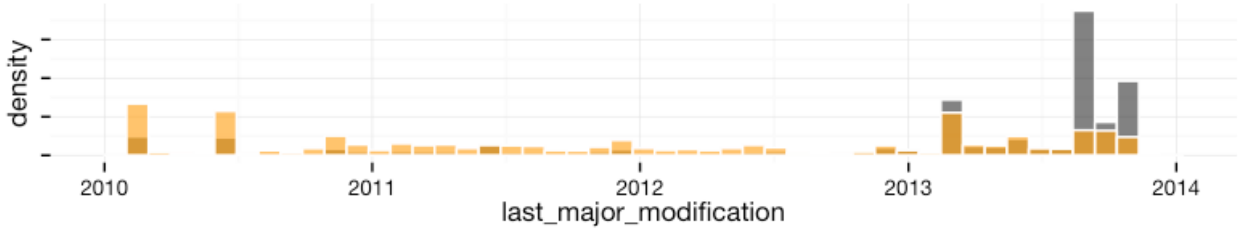


However, there is a substantial problem with missing data for *update_frequency*. This is one reason why the UK data catalogue does not perform well. According to data.gov.uk there is a wider issue of educating publishers on what metadata to include.

If we compare the distribution of all datasets with the one that omits missing *update_frequency* (only around 4,000 remain!), we see a different pattern. The updates are no longer concentrated in recent months.

[8] The metadata in its raw form is available here: http://data.gov.uk/data/dumps/

Figure 4: The UK data catalogue, histogram of the last major modification. Grey bars include datasets with missing update frequency, orange bars exclude them.



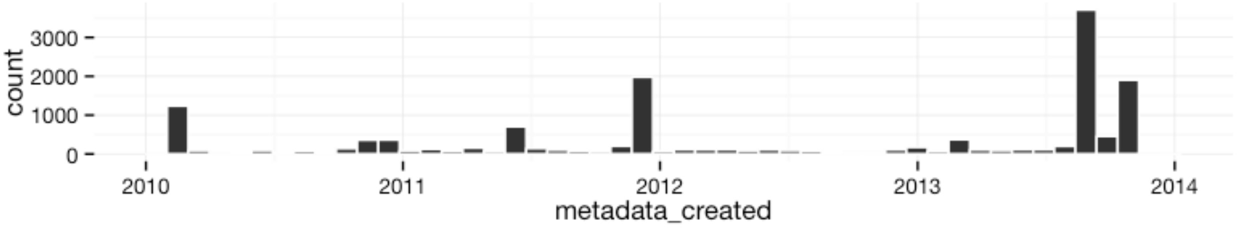
The **overall τ = 0.25** which is a poor figure and below the other two case studies. However, as mentioned above almost ¾ of the update frequency data are missing.

Table 4: The UK data catalogue's tau breaks down as follows

update frequency	τ	count
daily	0.00	45
weekly	0.00	12
monthly	0.06	1445
quarterly	0.27	638
biannually	0.22	228
annually (and various)	0.38	1464
every 2 years	0.06	17
every 10 years	1.00	129
overall	*0.25*	*3978*

Given the strong pattern using all datasets, we might be inclined to assume the UK data catalogue does much better than the τ would suggest. The fact is, though, we cannot know without data. The distribution of *metadata_created* also has a spike in September 2013 (see figure 5). This implies many datasets were added recently and that they may bias the *last_major_modification* variable. The release cycle is also highly irregular.

Figure 5: The UK data catalogue, histogram of metadata created

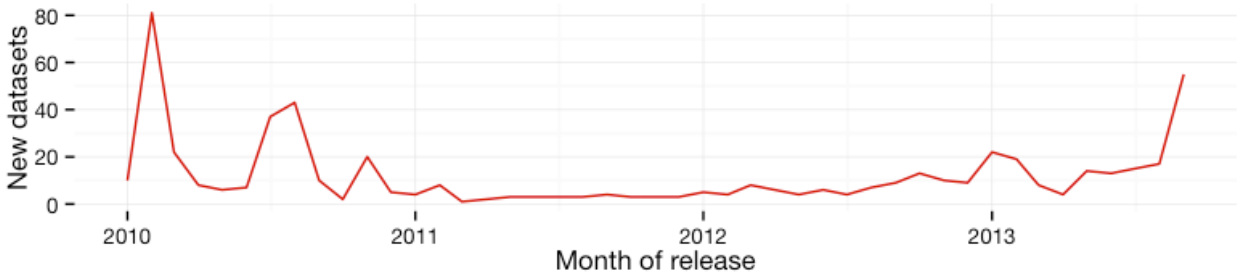


3. The London Datastore

At the time of analysis the London Datastore[9] hosts 537 datasets. They were published with the following pattern since January 2010.

[9] http://data.london.gov.uk, accessed 2013-10-15.

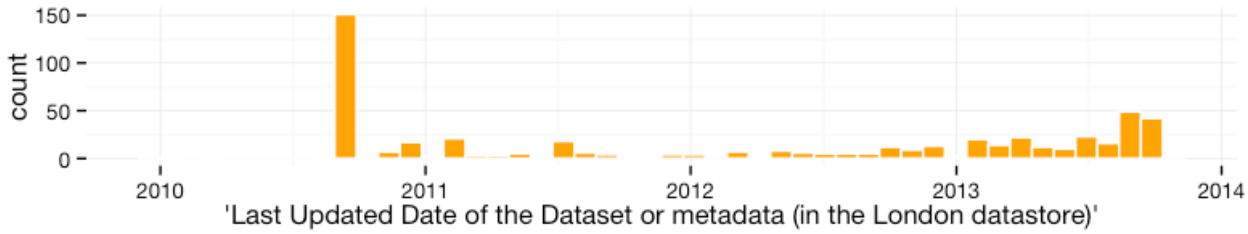
Figure 6: The London Datastore, new data releases per month



The big spikes at the beginning are months were the London Datastore released many similar datasets. For example, in August 2010 the Department for Education released a series of datasets. Or in October 2013 the London Fire and Emergency Planning Authority (LFEPA) added around a dozen datasets to the datastore.

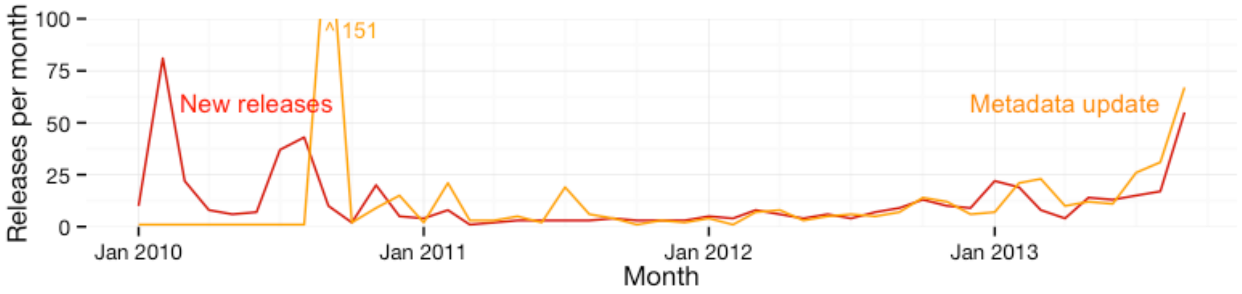
The more relevant variable, however, is called *metadata update*. The metadata update is the "last updated date of the dataset or metadata (in the London Datastore)". As we can see in figure 7, for the London Datastore the month of September 2010 is a large outlier. We do not have a better explanation than a general update of the early releases.

Figure 7: The London Datastore, metadata updates histogram



Otherwise the metadata updates slightly trail the release figures. They are *not*, as you might expect for an up-to-date catalogue, particularly concentrated in recent months. Below are figure 6 and 7 combined in one graphic.

Figure 8: The London Datastore, new data releases and updates combined



The **overall τ = 0.52**, which suggests, as with the World Bank, around half of the datasets are updated according to schedule. Some uncertainty persists as around 20% miss a measure of update frequency. However, the field "last updated date of the dataset or metadata (in the London Datastore)" is more general than needed.

Table 5: The London Datastore's tau breaks down as follows

update frequency	τ	count
daily	0.00	2
weekly	0.00	2
monthly	0.51	37
quarterly	0.49	57
biannually	0.20	10
annually (and various)	0.47	216
every 2 years	1.00	1
every 4 years	1.00	7
every 10 years	1.00	29
overall	*0.52*	*361*

Future research

The timeliness of data will remain a critical question because the demand for quality data will only increase. Thus, more research is needed in several areas.

A promising research question would establish different practices in data catalogues when it comes to updating datasets For example, arguably the biggest area of "dark matter" comes from deleted datasets. To update, a publisher uploads a new dataset and deletes the previous one. Where or how is this reflected in the metadata? At least in the UK data catalogue this scenario seems to be "very, very rare"[10], but practices differ across data catalogues.

Future research should further assess the state of the metadata in catalogues and how to encourage use of standards, e.g. the uptake of the Dublin Core. There is a critical need that publishers are educated in leading practices of publishing data.

Another project could either look at dates within datasets or inspect the date ranges a dataset covers. Comparing these statistics against the last publication may uncover new ways and shortcomings of measuring timeliness.

In the future we also hope to see research that analyses larger samples of catalogues' tau. For example, how does tau vary over time? Are there differences in tau that are a function of geography, size or sector? Where can we find exemplar cases?

Summary

In this paper we addressed the need for up-to-date datasets in catalogues and proposed a new metric: tau. Three case studies validate the feasibility of implementing it. Moreover, the three cases represent different regional levels, yet all of them achieve a less than optimal score and fall short in their publication of metadata.

Thus, much improvement is possible. Timeliness is the third of the eight criteria of open government data and needed "to preserve the value of the data". [11] Measuring timeliness can put a

[10] Personal email communication with a government official on 2013-11-28.

[11] http://opengovdata.org, accessed 2014-02-10.

spotlight on this criterion and therefore may foster efficiency and support the sustainability of the open data ecosystem, for example, by encouraging automated publication of data.

Building trust is difficult for a publisher and can easily be lost by neglecting to keep its data catalogue up-to-date. Third parties such as entrepreneurs are less likely to create start-ups and services on top of open data if they cannot rely on the longevity or timeliness of open data.

Standards are important for numerous reasons (see, for example, Jisc Digital Media, 2013). A standardised metric on timeliness, or any other characteristic, has also the potential to enable broader, more influential research.

References

Bouchal, P., Stephen, J., & Bull, D. (2013). *Whitehall Monitor 2013*. Institute for Government. Retrieved from http://www.instituteforgovernment.org.uk/publications/whitehall-monitor-2013

Gurin, J. (2014). *Open data now: the secret to hot startups, smart investing, savvy marketing, and fast innovation*. McGraw-Hill.

Kurtz, M. (2013). Dublin Core, DSpace, and a Brief Analysis of Three University Repositories. *Information Technology and Libraries*, *29*(1), 40–46. doi:10.6017/ital.v29i1.3157

Levine, T. (2013). *Updating of data catalogs* (code available on GitHub https://github.com/tlevine/socrata-analysis). Retrieved from http://thomaslevine.com/!/data-updatedness

Lindman, J., Rossi, M., & Tuunainen, V. K. (2013). Open Data Services: Research Agenda (pp. 1239–1246). IEEE. doi:10.1109/HICSS.2013.430

Maali, F., Cyganiak, R., & Peristeras, V. (2010). Enabling Interoperability of Government Data Catalogues. In M. A. Wimmer, J.-L. Chappelet, M. Janssen, & H. J. Scholl (Eds.), *Electronic Government* (pp. 339–350). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-14799-9_29

Jisc Digital Media. (2013). Metadata Standards and Interoperability. Guide. Retrieved December 20, 2013, from http://www.jiscdigitalmedia.ac.uk/guide/metadata-standards-and-interoperability

Tinati, R., Carr, L., Halford, S., & Pope, C. (2012). Exploring the Impact of Adopting Open Data in the UK Government. *Digital Futures 2012*. Retrieved from http://eprints.soton.ac.uk/344808

Tinati, R., Carr, L., Halford, S., & Pope, C. (2013). Exploring the Use of #OpenData in UK Open Government Data Community. *Digital Economy 2013*. Retrieved from http://eprints.soton.ac.uk/358942

Yin, R. K. (2009). *Case Study Research: Design and Methods*. SAGE.

About the Author

Ulrich Atz, Head of Statistics at the Open Data Institute (ODI)

Ulrich holds a Diploma in Economics from the University of Mannheim and a M.Sc. in Social Research Methods from the London School of Economics. He has a broad background that blends modern statistical techniques with practical uses of data. At the ODI, Ulrich leads research projects, consults startups and governments on the business case of open data, and helps out with training courses. He regularly holds presentations and keynotes.