Data quality proposals

Title: Everyone wants ‘good’ data - but what does this even mean?

Track: Action track, **Action Area:** Measurement and evaluation

Session objective: This hands-on workshop discusses current definitions of data quality, where they overlap and where they diverge, and seeks to provide a shared understanding of data quality.

Abstract:

Everyone wants ‘good’ data - but what does this even mean?

Some years ago, open data was heralded to unlock tremendous value to the public that would otherwise remain closed within information, locked away. So when the open data movement demanded “Openness By Default”, many data publishers followed the call by releasing vast amounts of data in its existing form. In many cases, open data catalogues or portals are also manually fed and this makes the overall open data management and publication approach weak and prone to multiple errors.

These emerging open data infrastructures often resemble a ‘Tower of Babel’: more information is produced, but it usually hard to use, incomplete, out of date or encoded in different languages and forms, preventing data publishers and their publics from communicating with one another. What makes data usable under these circumstances? How can we close the information chain loop? The short answer: by providing ‘good quality’ open data.

But there is no agreed upon definition what constitutes ‘good’ data quality. Research shows many different interpretations and ways of measuring data quality. And since people use data for different purposes, certain data qualities matter more to a user group than others. This workshop will invite usability researchers, interface designers, data publishers, and data scientists to exchange viewpoints on data quality and outline practical next steps to improve data quality.

Format:
Workshop

Format details:

Prior to the session moderators will gather different definitions of data quality by consulting open data measurement practitioners, data portal providers who run their own usability tests, or committees such as the W3C. In a blogpost we will synthesise the different quality standards and definitions that currently exist.

During the IODC session moderators will shortly introduce the session goals and present different definitions of data quality (10 min). These may draw from frameworks such as:

1. Integrating data dictionaries with datasets
2. Data Management Plans
3. Data Quality Statements
4. Data Use Analytics as a Quality Measure
5. Data Validation/Cleaning (eg, goodtables.io)

In break-out sessions (40 min), participants will group around different quality definitions and discuss their overlaps and differences. Moderators will facilitate a group discussion where session members think through how different data standard definitions apply to ‘real-life’ datasets, and what definitions help to improve the quality of these example datasets. The session will conclude with a summary of learnings from this exercise, which will be synthesised in a blog post series on data quality aspects, and targeted recommendations how different user groups and publishers can enhance data quality.

Moderators:

- Steven De Costa (Link Digital & CKAN Association)
- Carlos Iglesias (Web Foundation)
- Danny Lämmerhirt (Open Knowledge International)
- Kristin Auld (Office of Environment and Heritage, NSW Australia)
- Irum Maqsood (Government of Canada)

Title: Spotting dirty data - easy tips for users to make the most of your data

Track: Action track, **Action Area:** Measurement and evaluation

Session objective: This workshop invites data users, trainers, tool developers, and data publishers to explore what makes data quality “good”, by sharing common problems participants are facing with data, and by learning what different tools and methods can help address these problems.

Abstract:

Data has a problem - different user requirements make it very difficult for data publishers to release data that meets everyone’s needs. To match data user needs with the informational landscape out there, various information-processing technologies and methods are currently practiced - addressing different user needs. What do these approaches have in common? What issues around data quality do they not address? What are the most important facts data users can learn from these approaches, to incorporate them more routinely in their work with data?

In this workshop data users and publishers develop a shared understanding of data quality, exchange common problems they are facing when working with data, and explore tools and resources that exist to help them clean / improve the quality of their data. The session shall increase the appetite of data users for these methods and show advantages of each approach.

Format:
Workshop

Format details:

In a brief introduction, workshop moderators will introduce the session concept, discussing problems arising from dirty data, and showcasing different methods to solve data quality issues (10 min). Moderators will gather questions and data quality issues, that participants would like to be addressed.

In a breakout session (40 min), participants group together with each moderator around a selection of quality issues. Each group will pick 3 problems to solve, and moderators will showcase their approaches to clean data. In short data-dive sessions, we invite participants to issues within the data, and develop a shared understanding when data is usable. Participants will be guided different strategies, methods, and technical tools that can help improve different aspects of data quality.

In a concluding plenary round (10 min) participants will share their experiences, and how well each data processing approach meets their requirements to make data usable.

After the session moderators will gather testimonials from session participants about the usefulness of each data processing approach, and will report the workshop findings in a series of blog posts. These blog posts reflect what participants learned during the workshop, and how data processing approaches help them in their work. The blog posts will also describe advantages of different data processing approaches, using real-world examples how these processes have been used.

Workshop moderators:

- Katelyn Rogers (School of Data)
- Serah Rono (Open Knowledge International)

Requirements: Projector, sticky notes, pens, flip chart, desks with charging stations for participants' laptops.