How data quality is assessed depends on what the data is used for, how and why it was derived, and how it fits with other information it is being used with — standards of practice that are likely different between different stakeholders within a collaborative information management system. Data quality, thus, is not just as simple as asking whether the data match reality, but requires questions about whether the data fits intended use, is aligned with the problem at hand, matches expected meaning, and can be reliable for the current task. Paying close attention to the completeness and accuracy of data will be required in order to fulfil the citizen rights provisions of the GDPR. Furthermore, ensuring data quality will also support the duty to ensure that data is interoperable, thus upholding duties under the right to data portability.

Guiding Questions

How can data quality be assured in a way that meets a diversity of stakeholder perspectives?

How is data determined as fit for the purpose of the collaboration?

How do you maintain data quality over time?

What procedures do you have in place to ensure the completeness and accuracy of gathered and stored data?

Further Information

Data quality is critical for good evidence-based decisions in disaster risk management. Poor data quality can threaten the validity and generalisability of conclusions drawn. This requires the reporting of both general quality features as well as analysis specific quality features that make transparent what happens locally, including methods and assumptions behind the local data-cleaning necessary to make the data shareable more generally. Quality is also dependent upon if the data is fit for specific uses, not just if it is representative or accurate in-and-of itself. As data is stored and used, data stewards should cumulatively include their data quality assessments to help future users better understand when and how the data might be of high or appropriate quality for their needs.

Especially in collaborative settings, data quality must speak to all stages of research, collaboration, and storage, including the development of tasks, protocols, guidance, calls for more, collection and processing practices, and establishing frameworks for analysis. Known key issues around quality include: completeness, uniqueness, timeliness, validity, accuracy,

consistency longitudinal concordance, breadth, data element presence, density, and prediction. Data element agreement, and data source agreement are also considered individual data quality assessment tools.

What has become clear to those who study situations of data quality and collaboration is that, while automated or limited vocabularies for data quality are helpful because it is not feasible and practical to do careful meaning checking of all entries, they will not solve all problems. One way to work with this is to error-check along the way: noting that something seems off, even if one is not exactly sure why, can become helpful for future users as they navigate the data and better determine how it can be made of quality.

Examples

Making databases talk: Guha Sapir and Below (2002) report how disasters may be classified as different types by different databases. This occurs particularly frequently for associated disasters or secondary disasters. For example, a flood, which was a consequence of a windstorm, may be recorded as one or the other. Verification that two different disaster types occurring in the same country on the same day are indeed the same event is only possible by checking the exact location. This could not be done electronically, as the data provided by one database (NatCat) did not include sub-national location.

Matching good data together: During the 2007 Wildfires in San Diego, one of the main concerns for the emergency responders and the public was road conditions: what roads were open, which ones were under threat of fire or smoke, and which ones needed to be held open for response actions. Consequently, those mapping the fires started to incorporate road closers and traffic data into their maps, drawing from official, traditionally reliable sources, such as CalTrans. However, this data was being updated at irregular frequencies. Sometimes the updates would come twice an hour, sometimes 3 or 4 hours would go between updates. This other data going on the maps was being updated in more regular intervals. As a result, when the maps were updated with the other information, map users assumed — incorrectly — that the road data was also being updated. Because of these incorrect assumptions potentially putting lives at risk, the mapmakers decided the road data, however accurate to itself, was not of an appropriate quality for their maps and stopped including it.

Resources

Antelio, M. Esteves, M.G.P., Scheider, D., De Souza, J.M. (2012). Qualitocracy: A data quality colalborative frameowrk applied to citizen science. 2012 IEEE Conference on Systems, Mans, and Cybernetics. [Link]

Guha-Sapir, D., Below, R. (2002) The quality and accuracy of disaster data: A comparative analyse of 3 global data sets. *Disaster Management Facility, World Bank,* Working Paper 191.

Kahn, M.G., Brown, J.S., Chun, A.T., et al. (2015). Transparent Reporting of Data Quality in Distributed Data Networks. *EGEMs,* 3(1), 1052. [Link]

Petersen, K. (2017). Visualizing Risk: Drawing Together and Pushing Apart with Sociotechnical Practices. Journal of Contingencies and Crisis Mangement. 25 (1), 39-50. [Link]

Twidale, M., Marty, P. (1999) An Investigation into Data Quality and Collaboration. *Technical Report ISRN UIUCLIS CSCW*. [Link]