Data Quality for Use: A Linked Data approach

Erwin Folmer (Kadaster, University of Twente) & Wouter Beek (Triply & Kadaster)

Introduction

Quality has a long standing history, mainly from product engineering (such as automotive) and is a broad concept. Literature and practitioners have a tendency to focus on intrinsic quality, which is more or less demarcated and measurable. But in addition to intrinsic (product) quality, there are two other notions of quality. Firstly, there is process quality, which deals with organizational aspects such as maintenance processes. Secondly, there is quality in practice, i.e., quality as observed/experienced within the use of a concrete application.

Organizations such as Kadaster, have a tendency to focus on intrinsic quality, but this focus can be questioned. Juran, one of the quality guru's, defined quality as "Fitness for use". Following this definition, one would expected a focus on quality in practice. Unfortunately this is less demarcated, especially when compared to the focus on intrinsic quality. One explanations for the lack of focus on quality in practice is that this notation makes quality situation-dependent. For example, when quality in practice is applied to datasets, this means that a dataset can have a high quality in one usage scenario, yet a low quality in another usage scenario.

In this paper we look into this more situational notion of quality in practice.

Quality related Problems for Datasets

Within the practice of being one of the main suppliers of open governmental data in the Netherlands, Kadaster has identified the following two main quality-related problems for its datasets.

(Spatial) datasets cannot be found

One main quality issue is that open governmental datasets, while published under an open license, cannot be easily found by developers and other potential users. As a result, open governmental datasets are not currently used to their full potential. Let us take the Key Register Topography (in Dutch abbreviated as BRT) as an example case: it is the authoritative dataset about topography in The Netherlands, and contains many object types, such as schools, churches, castles, and many others. A user who searches the Dutch National Geospatial Register (in Dutch abbreviated as NGR) for schools, churches, or castles will not find the BRT, even though it is one of the most comprehensive dataset for churches in The Netherlands. The reason for this is that the object types that are present in the BRT dataset are not mentioned in the metadata description of the BRT. In general, concepts that occur within geospatial datasets are currently not (automatically) part of the dataset metadata.

Because of the above described issues, a user will only find the BRT dataset if she searches for the title of the dataset (e.g., "BRT"). This means that Kadaster datasets are typically found by people who are already aware of their existence, but not by people who are searching for concepts that appear within Kadaster datasets. What is more, one could argue that many potential users of Kadaster data will not start their search at the National Geospatial Register for Datasets at all, but will be searching from a

generic search engine and/or will use a voice assistant for information about schools, churches, castles, etc.

This situation is not unique for Kadaster. The idea that geospatial datasets should be exposed through, and searched for on, a special geospatial platform is more or less the INSPIRE vision and approach.

It is our belief that a new user, unaware of the name of the dataset but with a specific and articulable need for (geospatial) data, will start his search using a popular web browser or a personal voice assistant. A more advanced user may use a dedicated, dataset-specific search engine like Google Dataset Search or a national data portal (e.g., https://data.gov.uk or https://data.overheid.nl). Ideally, when a user searches for churches in The Netherlands, he will find the BRT among the top results of his search operation. From these search results, the user will dive directly into the NGR page that specifically describes the BRT dataset. While we recognize that there are many different users with different competencies and capabilities, we believe that what we describe above will be the 'happy flow' that a large number of users that are not being served today will follow.

Unfortunately, the current situation is very far removed from what the 'happy majority flow' that we describe above. Many of the open datasets published by Kadaster today are not found at all in popular and generic search engines. Even in search engines that specifically focus on datasets, like Google Dataset Search, authoritative Kadaster datasets like the BRT cannot be found. we find several outdated copies of our data (by commercial organizations, or universities), but not the authoritative source. From user perspective a big quality issue.

We here enumerate the three sub-problems that can be distinguished with respect to the findability of spatial datasets:

1. Metadata descriptions for datasets often contain insufficient detail (e.g., the BRT cannot be found when searching for churches in The Netherlands).
2. Governmental agencies focus on search from within dedicated portals, but users use generic search engines.
3. Spatial datasets published in dedicated portals are often not findable through generic search engines.

Fitness for Use is Unclear

When we apply the definition of fitness for use, we need to know the use case in order to make the quality assessment to find out if the datasets is "fit" for this intended usage. However in the context of our role as publisher of open geospatial data, we most of the time do not know the usage of open data. What is more, the stated purpose of open data is that new users that are currently unknown to the data publisher should be able to use data in different contexts and in originally unanticipated ways.

However, if a data supplier does not know how their (open) data is being used, then it logically follows that they cannot define fitness for use (and therefore the practical quality) of a dataset either. Indeed, when a data supplier assigns quality statements or labels to its datasets, its potential users may misinterpret these static quality indicators as fitness for all use cases, but the latter may not be correct.

Solutions

We present three solutions for the issues identified in the previous section.

Attitude change

We need an attitude change (mind shift) from quality as a static concept that is determined by data publishing organizations, to a dynamic concept that is ultimately determined by the data consumer. Quality is always situational: for a certain user, within a certain use case, working from within a certain context. While it may still be useful to formulate and implement generic data quality metrics, such generic metrics can never capture dataset quality in its totality.

In practice we often notice that dataset owners hold on to a Boolean notion of dataset quality, resulting in two unrealistic 'all or nothing' attitudes. One extreme attitude is that the dataset already has good quality: the dataset is published in a governmental (geospatial) data portal and fulfills the currently formulated quality requirements. The fact that the dataset is not often used in practice is sometimes lamented, but is not recognized as a dataset quality problem. The other extreme attitude is that the dataset does not *yet* have good quality. The fact that the dataset is not often used by others is by design: the users must wait for a new version, a new data model, or a data cleaning initiative. Only once those have been completed will the dataset be ready for use.

Our notion of practical dataset quality opposes both views. A dataset publisher may believe that their dataset has good quality, but if a dataset is not often used then this is an indicator that the dataset may not be fit for use. Similarly, a dataset publisher may believe that their dataset does not *yet* have good quality, but a data consumer may disagree with this, and may already be satisfied with the dataset as it currently is.

Quality Dashboards

We need transparency, and the first step is publishing quality dashboards, which many organizations – including Kadaster – have been doing for quite some time. In the early days, custom dashboard applications were developed within the organizations. Since this is a relatively expensive process, such an approach is only feasible when the intrinsic notion data quality is used.

In recent years we have noticed an increasing need to change and redesign quality dashboard. This reflects a change in the notion of quality that is embraced by the organization: one that is based on a changing practical need. With this more fluid notion of quality for use, it becomes more economical to use standard Business Intelligence (BI) tools like Tableau to create quality dashboards.

Another generic trend is that static reports (often in the form of PDF documents) are slowly being replaced by interactive dashboards. In the near future this will be merged with quality dashboards, into one integrated dashboard, containing a viewer, data model, quality, use case descriptions etc.

Transparency

In the absence of a static notion of intrinsic quality, it does not make sense to advertise dataset quality in absolute terms. Instead, we want datasets to anticipate the fact that users will make use of the dataset in different and potentially unanticipated ways.

In order to achieve the latter, a dataset must seek to transparently communicate its potential for use. A dataset must communicate its potential for use in a multi-faceted and pluriform way, so that individual users are able to determine for themselves whether the dataset is fit for their use.

In the context of Kadaster, Linked Data is used in order to express the multi-faceted potential for data use. Linked Data offers a wide variety of off-the-shelf metadata vocabularies that can be utilized for this purpose. Furthermore, the open-endedness of Linked Data allows new metadata aspects and new vocabularies to be formulated, not replacing but augmenting existing initiatives.
Examples of Linked Data vocabularies that are used to express data quality aspects at Kadaster include:

- Dublin Core: allows generic dataset properties like creator and creation data to be specified.
- DCAT: allows more detailed dataset properties to be specified, including the temporal range covered by a dataset, the spatial range, the update frequency, and the accuracy of its measures.
- PROV: allows a detailed specification of how the dataset was created, curated, and published; including the specific sequence of data operations that was taken.
- Schema.org: allows an increasingly large number of metadata properties to be communicated in a format that is processed by a large number of search engines, web crawlers, and other web-based tools.
- OGP: similar to Schema.org, but mostly focused on metadata that can be used in social media platforms.
- BRT: in addition to the above existing vocabularies, Kadaster datasets introduce their own Linked Data vocabulary. For example, the BRT vocabulary describes the types of objects it contains, including schools, churches, and castles.

Since Linked Data is a web-native metadata paradigm, descriptions of data quality for use can be published online, as part of regular web pages (using JSON-LD snippets). Furthermore, popular web search engines like Google actively look for and index such metadata properties. This allows a wider range of users to determine *for themselves* whether a dataset that Kadaster publishes on the web is fit for *their* use.

Kadaster is currently experimenting with exposing its dataset metadata using the above Linked Data vocabularies. Early results already show that the Linked Data approach allows Kadaster datasets to become better findable on the generic web, i.e., outside of (spatial) data portals.

Conclusion

In this paper we propose a shift of focus in the quality domain. In addition to a different attitude and more quality dashboards, we propose is to put more effort in publishing metadata that follows modern web standards. This results in a higher level of transparency and fosters insight into data practical quality for use. The result will be that in the future more people will be able to find datasets on the web, and can make a quality assessment that is more tailored towards their specific use case.