Raising the quality of your city's data by opening up

Bart Rosseau[I], Pieter Colpaert[II], Ruben Verborgh[II], Erik Mannens[II] and Rik Van de Walle[II]

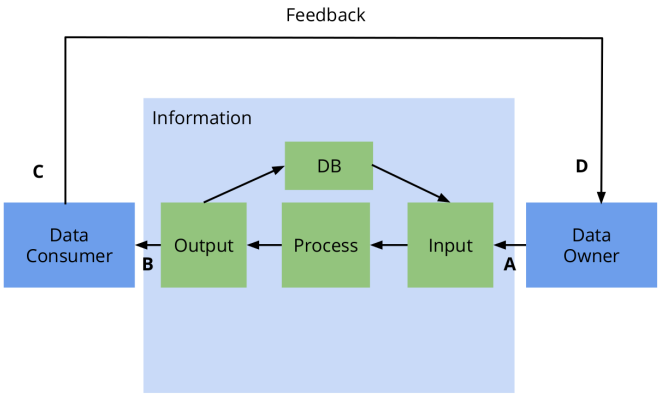
[I] The city of Ghent: bart.rosseau@gent.be
[II] MMLab - UGent - iMinds: {firstname}.{lastname}@ugent.be

abstract - In this paper we describe the effects of opening up data within a local government on the data quality. Two definitions for data quality are introduced and from these definitions we describe an interplay between three indicators: awareness, reuse and data quality. Applying this theory to a case study in the city of Ghent, we conclude that Open Data is not a goal in itself, but a part of a chain where reuse empowers data-owners.

Introduction

Redman and Orr have defined data quality in different ways. Redman [1], on the one hand, uses this definition: *"Data are of high quality if they are fit for their intended uses in operations, decision making and planning" (Joseph M. Juran)*. It is a very pragmatic definition, where the data quality is a measure for how easy the data can be used within certain predefined use cases. When measuring the data quality within information systems, data quality will need to be defined as a collection of measurements such as accuracy, completeness, currency, ease of interpretation, etc.

Orr [2], on the other hand, uses a more abstract definition: *"Data quality is the measure of agreement between the data view presented by an information system and the that same data in the real-world."*. Orr applied this on the Feedback-Control System (FCS) model.



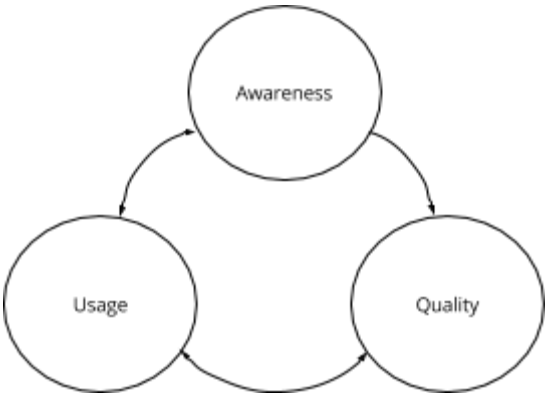


Figure 1. The Feedback-Control System (FCS) applied to data management (left)
Figure 2. Three parameters that influence each other (right)

FCS models how organisations communicate to the outside world. When applied to data management, as illustrated in figure 1, there are certain processes that can be identified: creating the data (A), reusing the data (B), providing feedback (C) and processing feedback (D). From this model, Orr assumes that "*one certain way to improve the data quality is to increase its use*". An Open Data policy stimulates the reuse of data, yet we have noticed that an Open Data is not a magic tool for raising the data quality within a local government. We thus introduce a third parameters which needs to be realised over time: the awareness. This is the awareness within a government that the data created is of high importance and needs to be maintained properly. The awareness will affect A, the quality of the data that is being created (e.g., more time is going to be spend to creating good input forms), and D, processing feedback on the data; the data quality will affect B, the reuse of the data, and the data reuse will affect C, the amount of feedback produced. We thus get three parameters influencing each other: awareness, the quality of the data and the reuse of the data. This results in figure 2: raising awareness, usage and quality go hand in hand.

Case study: city of Ghent

If you build it, he will come
~ Field of dreams

It is *April 14th, 2011*. In the city of Ghent a group of digital creatives gathers in a bar, to attend the third event in a series of 10 of GentM[1]. This recurring event focuses on the digital trends that affect our daily life in the city. The theme was 'Open Data'. The vice mayor responsible for ICT took a leap of faith and committed the city of Ghent to start opening up their data, assured by the promise of the creation of apps for the city and a boost for the creative economy.

[1] http://gentm.be

Today, we can reassess the benefits of Open Data for a local government. So far we noticed that the biggest short term benefits are internal. Using Open Data as a tool to talk about goals, means and co-creation allowed for an extra dimension to audit our organisation and its processes. When you are talking Open Data with other enthusiasts, similar stories arise. How you have to convince departments into opening up their data, how you have to assure them that this will not result in huge cyber attacks or gross misrepresentations of the intended purpose. Yet, beyond these 'war stories' there is another underestimated aspect: the Open Data pioneer will be the first person to have an overview of what kind of data the organisation manages and uses.

It was a ranger, Butch. A lone ranger.
~ The Lone Ranger

Starting from this unique position, the Open Data pioneer can spark some processes which will benefit the organisation. The overview of available datasets alone is already very interesting. Studying the meta-data can teach you where datasets are overlapping and thus where there is room for optimisation in the government structure.

In Ghent we noticed that different departments had some form of a list of streets, with different attributes, according to their goal. The environmental department used this to list the trees, the monuments department annotated the list with explanations of the historical figures featured in the street names, the public works department used this to gauge the quality of the roads. Yet, they were working on different versions of the list of streets, each of these versions were less and less used and the quality decreased with the decrease in usage. As an Open Data pioneer, it was my job to create the awareness amongst these services (cfr. figure 2).

This insight (backed up by evidence) started a process where we renewed the efforts towards (master) data management, and we raise the awareness about key referral data. Together with Digipolis , our IT partner, we also want to embed these insights into all future data projects.

On the outside looking in…

How much efficiency can be gained by distributing the data through some form of an internal data management system? No more worries about updates, because everybody can work on the same list and by assigning responsibilities to certain datasets there is a heightened sense of ownership. Furthermore, it contributes to the continuity of data management and avoids sudden gaps of knowledge when a person who manages a certain dataset leaves the organisation. By talking about opening up datasets, the added value beyond the core purpose becomes apparent. Data as a by-product of a process will be re-evaluated, and sometimes redeemed, like the once hidden value of trash before we

started recycling.

To make this reality, all the efforts to engage the coding community to use your Open Data should also happen within your organisation. At least to raise the *awareness* of the value of data and the importance of data quality. This is where the feedback loop is key. By showing results of third party developments, preferably in a fancy app, a lot of civil servants see their data used and useful in a way that is previously seen as not possible for a government: “trendy, interactive, mobile or user-friendly”, terms that are not easily associated with governments.

You are not alone…

For different reasons, the public accessible toilets in Ghent are managed by different departments, each working with their own ‘siloed’ applications. It took some effort to combine those into a printed map of the public toilets in Ghent. Some considerable editorial effort was necessary to clean up the data, so the right parameters could be attributed consistently to the various sources.

This is a great case where *open standards* and predefined vocabularies are a solution. By implementing the standards in the various ‘mother’ systems, a new, functional and reusable master-dataset emerges, where each department is more aware of what the other is doing, where efforts can be aligned and the ‘map’ is created in a more time-efficient way. This was initiated by the question to add the extra temporary public toilets that are installed for the 10 day Ghent Festival. The program of this massive event is available as Open Data. The various developers asked for extra information to enhance the usefulness of their Open Data based apps. To include the set of public toilets we discovered how data-management can be useful to streamline the interdepartmental cooperation, and yield more efficiency in the preparation of this huge festival.

By applying standardisation to datasets, sharing identifiers and vocabularies managed by different people, their contribution in the data value chain becomes visible. Each can be ‘king in their castle’, without losing synergies.

It’s* data *Jim, but not as we know it.
~ Star Trek

Talking about data and its usage in processes, adds a new element to the cooperation between government organisations and governments as such. When approached on a data level, we are talking about the essentials. This can defuse some tricky negotiations: it is easier to talk about who is responsible for which column in your spreadsheet, than to determine if your tab sheet is better than the other’s table in a word document.

What we described so far are quite simple cases, where we shared the first insights we

gained by opening up basic datasets. These efforts fueled by Open Data resulted in a heightened awareness on different levels about the importance of data quality, but even more about the opportunities to work more efficiently, reduction in IT spending and overall a smarter organisation. If we continue to reach the next maturity levels in our Open Data policy we can take this several steps further: how will a good URI strategy influence our BI environment? How will crowdsourcing influence our feedback loops?

Exciting times are ahead. We need to grow beyond the 'accidental' benefits. It should be part of an overall data and information management effort. We will need to attract some data professionals into the government.

Conclusion

In this paper, two definitions for data quality were discussed. The definition by Juran, on the one hand, is goal-oriented: the data quality should be seen as a collection of indicators (e.g., accuracy, timeliness, currency or ease of interpretation) which are assessed keeping certain use cases in mind. Orr's more high level definition, on the other hand, is based on the agreement with the real world. This agreement with the real-world will improve, as deducted from the Feedback-Control Systems' model, when the awareness of the data owner and the reuse is raised.

The case study of the city of Ghent shows that Open Data is not a goal in itself, but a part of a chain where re-usage empowers data-owners, who are more aware of the importance of the data they manage. In this way, Open Data becomes a tool to optimise different aspects of an organisation. In times where (local) governments face an ever increasing complexity and severe budget cuts force governments to re-evaluate every aspect of their operation, better data for decision support becomes a powerful tool.

References

[1] Redman, T. C. (2001). *Data quality: The field guide*. Boston: Digital Press.

[2] Orr, K. (1998, 12). Data quality and systems theory. *Communications of the ACM, 41*(2),

66-71. doi: 10.1145/269012.269023