From data quality to language quality: the challenges of the fitness for use principle

Laurence Dierickx

Abstract

Automated news productions can be delivered as it to the public or provide automatic drafts for journalists to enrich with background information or their expertise. This second option is here explored through two case studies conducted in French-speaking Belgium, where automated contents were approached as preliminary drafts for journalists. The first project aimed to provide automatic monitoring of air quality in Brussels to feed into a broader investigation into the causes and consequences of air pollutants. The second aimed to automate stock market data to support time-consuming daily routines, allowing journalists to provide more insights into the broader socio-political context on which stock markets also depend.

In data science, as in journalism, it is commonly accepted that poor quality data lead to poor quality information. Moreover, the quality of the language of the automated contents is strongly related to the field of application: one cannot write in the same way on the atmospheric pollutants as on the stock markets, assuming the use of specific lexical fields and reference expressions. These two aspects should not be considered separately as they are strongly linked. Also, taken upstream, the data define the texts and their level of variability.

Although these two case studies took place in different socio-professional contexts, they are reflected in their complementarity since the quality problems were not the same from one experience to another. They depended as much on the data source (free public open data in the first case, licensed paid data in the second case) as on the journalistic requirements, which can be understood under the lenses of the ISO 9000 where quality is tackled in the ability of the system to fit users' needs. Several original assessment methods have been designed to address quality challenges to prevent or correct problems. An interdisciplinary perspective grounded them at the crossing of journalism studies, data science, human-computer interaction (HCI), and natural language processing.

These qualitative and quantitative assessment methods enabled us to grasp the materialities that characterize processes at work in the context of automated news production -from the material condition (the data) to the final material (the content), passing by the translation material (the process that consists of transforming structured data to texts, charts, or any other visual representation)- by providing evidence-based elements to feed the analysis. Although they have not been without limitations, they have also made it possible to approach the phenomenon of information automation as a socio-technical process in action within which journalists, programmers and computer linguists interact.