Why news automation fails

Laurence Dierickx
PhD student, Université Libre de Bruxelles (ReSIC)
laurence.dierickx@ulb.ac.be

ABSTRACT

This paper examines the causes of failures of automated news production. If technical factors have to be considered, such as bad data quality, the human factor remains essential, including within its social dimension. The proposed approach here is both theoretical and empirical. It is highlighted with the results of a case study conducted over one year. The project consisted of developing an automated news application about air quality within a Belgian newsroom, in order to provide real time information to the audiences as well as raw material for an investigative purpose.

KEYWORDS

automated news, data quality, data literacy, innovation, uses

1 INTRODUCTION

Since the beginning of its short history, automated news production is unanimously acclaimed for its qualities of accuracy, speed and high scale production (Graefe 2016). Those technologies enabled to extend media coverage to areas which had a limited or a non-existent coverage before, as well as to deal with large amounts of data that it would not be humanly possible to otherwise treat (Leppänen & alli 2017). Both an object and tool of journalism, automated news production can be used as a final copy for the audiences or to support journalists in their daily routines (Latar 20015, Hansen & alli 2017).

It should be wrong to think that automation technologies never fail. This paper examines the different types of possible failures, in the light of the few examples of failures reported by online news media and of an empirical research about an automated information system, which aimed to support a wider investigative project about air quality within a French-speaking Belgian newsroom. It is framed by a multidisciplinary theoretical backdrop which demonstrates common preoccupations between research fields.

2 GARBAGE IN, GARBAGE OUT

On September 6th 2008, Google News published an article from the South Florida Sentinel about the bankruptcy of a company. As the story had no date attached, Google News tagged it with the current date by default (McCallum 2012). In July 2015, Wordsmith, the software developed by Automated Insights, reported that the individual share of Netflix fell 71 percent and that the company did not reach the expectations of analysts. In reality, Netflix's share price had more than doubled. The error came from a "7-1" split which was not taken into account and became "71" (LeCompte 2015). On June 22th 2017, "Quakebot", an automated program reporting about earthquakes in California, reported in *The Los Angeles Times*, an earthquake of magnitude 6.8 which happened… but in 1925. The bug was related to a "Unix epoch time", which led the software to reinterpret the year. Panic movements were observed on the stock markets, in the first two cases, and on social networks, in the third. Even If errors occur only pretty rarely, the fact is they still sometimes do happen. Their causes have to be found among various and complex factors which have ideally to be tackled at the beginning of the design process, because it is always better to prevent than to cure. Moreover, accurate and reliable information can only rely on accurate and reliable data: as Essa has pointed out (quoted by Flew & al., 2010), journalism and information technology are both concerned. This is probably the reason why failures are not so common within the field of automated news where economics and sports data are mostly covered: when data are bought to data providers, there is normally a guarantee of accuracy and reliability.

2.1 Bad data quality

The data quality literacy considers that only quality data will provide quality information (Batini & al. 2009). It is particularly true within the context of news automation, where data feed information systems which produce texts or other kind of visual representations. Data quality issues are also a matter of confidence: if data cannot be trusted, generated information cannot be trusted too (Haug & al. 2011, Golab 2013). As for any kind of data driven journalism approach, the first golden rule is to fact-check the data source.

Dealing with messy data is a common issue in datajournalism, whether it is automated or not. The concept of data quality is difficult to tackle due its multidimensional character but researchers widely admit that it cannot be solely reduced to the "entity, attribute, value" model as it implies both decisional and operational processes (Redman 1996). Moreover, total data quality cannot exist because there is no absolute reference to verify the correctness of data (Boydens 2012) . Besides, poor data can coexist with correct data without generating any error or data can be error free but without presenting the expected meaning for the user (Moody & al. 2003, Wang & al. 2006, Madnick & Zhu 2006, Batini & al. 2009, Boydens 2012) .

Another particularity of the data quality concept lies in its multidimensional approach, which can be related to different aspects such as reliability, objectivity, believability, relevancy, accessibility, interpretability, semantic integrity or physical integrity (Brodie 1980, Strong & al. 1997, Haug & al. 2009) which can be connected to journalistic requirements. Furthermore, data collected from empirical observations are subject to change over time. For this reason, their representation will always be the result of a moving reality (Boydens 2012). Making the situation even more complex, data quality is not only about the values, it will also depend on how data are distributed: proprietary formats and the non-use of standards are two other symptoms of non-quality (Janssen & al. 2012).

The ISO 9000 standard defines the concept of quality as the ability to satisfy explicit or implicit needs within a particular application domain (Boydens & Van Hooland 2011). In this perspective, data quality indicators taking into account journalistic requirements

Table 1: Formal and empirical indicators for assessing data quality

Axis	Assessment types
Documentary	Terms of use: do I have the right to use it? Unique identifier (can be added within a new column) Availability of metadata and conformity with metadata (contact the data provider in case of doubt)
Encoding	No encoding problems (can be solved with changing the encoding to UTF-8 for accentuated characters) No HTML overload (cleaning is required) No duplicate data (analysis and cleaning required)
Normative	Use of standards (e.g. e-mail, date, address, geolocation, ...): standardization is a common unit and facilitates programming tasks
Semiotic	No orthographical incoherence (analysis and cleaning are required) Explicit labelling (rename if not)
Journalistic	Accuracy: no anomalies observed in values, syntactic correctness (requires understanding of data and abnormal values, can lead to design specific rules if data are nevertheless considered as relevant) Currentness (last update explicitly mentioned) Reliability: even a primary source does not guarantee it (requires fact-checking) no missing values, appropriate amount of data (requires the knowledge of the application domain, through asking an expert for instance)

such as reliability, accuracy and currentness (Clerwall 2014) can be a way to deal with messy data: the sooner problems are identified, the sooner they can be solved. Within the context of an empirical research, where an automated news system called "Bxl'air bot" (dedicated to the reporting about air quality in Brussels) was developed, a conceptual framework was designed to assess both formal and empirical aspects of data quality in the aim to answer both technical and journalistic challenges (Dierickx 2017). Table 1 summarizes this method, where questions can simply be answered on a Boolean mode (yes or no). It also gives clues about how to solve a problem once it is identified. Despite its benefits, this method remains insufficient, especially when data are provided in real-time and when they are available in open data. Considered as an opportunity for journalists, open data do not often meet their promises because of a lack of relevance and of their relative quality regarding the journalistic uses (Stoneman 2015).

2.2 Data life cycle not well identified

For the needs of a research project aiming to study the uses of automated news by journalists, I have developed an automated information system which consists of a web application that can be both considered as an object of journalism, which provided real time news, and as a tool for a wider investigative project about air quality in Brussels, conducted over one year within the newsroom of the Belgian magazine *Alter Echos*[1]. For public authorities, the definition of open data does not always meet the 5-star deployment scheme as defined by Tim Berners-Lee. Open data about the measurements of pollutant rates in Brussels illustrate it well. If data are available and labelled as open data, they are widespread on six different web pages within overloaded HTML tables.

The purpose of the web application was to report a sensitive situation, with the use of natural language generation and of graphical representations. As there were no mentions about the updates on the webpages, the implementation of the conceptual framework to assess the formal and the empirical data quality rapidly appeared to be insufficient to ensure the journalistic requirements of accuracy, correctness and currentness. Air quality data are constantly evolving over time: moving averages only become fixed averages after twenty-four hours. In order to keep a track-record of the data, the system needed the daily fixed averages: how to ensure to retrieve the right data at the right time? The answer was found with the analysis of the data life cycle, a concept derived from data management. Its modeling aims to optimize the flow management within an information system, from data creation to archiving and destruction (Reid 2007, Fox 2014).

Due to a lack of information publicly available, interviews with the data producer have enabled the definition of the data life cycle, with by allowing to the detection of a shift of a half hour from the collect to the diffusion of data. Fixed averages were not published after twenty-four hours but within thirty-two hours, with the possibility to be changed over the following days if abnormal values were found by the data producer. For instance, it is impossible to get a high ozone level during the winter because ozone formation requires high temperatures. Figure 1 consists in a starting point for the modeling of a data life cycle, which will vary from an application domain to another regarding its particularities. Skipping this stage would have produced inaccuracies but it was definitely not the last cause of potential failures observed within this experiment.

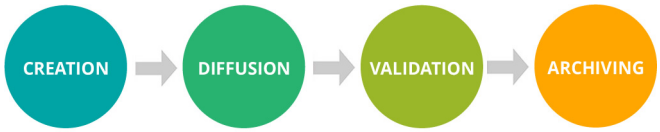

Figure 1: Conceptual modeling of a data life cycle.

2.3 Lack of error prevention

Prevention is always better than cure but all possible cases cannot be anticipated without a deep understanding of the data, which is also useful when the time comes to deal with missing or NULL values as it is likely to lead to misleading: remember the Google News case when the missing value of the date was replaced by the current one. If reasons can be sometimes found, who is able to predict that the data flow could stop because of a power failure, as it was observed during the "Bxl'air bot" experiment? Even reliable

[1] "Bxl'air bot", URL : http://www.bxlairbot.be

data sources have their Achilles' heel. This issue is common to any information system: in all cases, the presence of NULL values has to be seriously investigated as the interpretations will always rely on human judgements, which are potentially not error or bias free. A NULL element does not mean the value does not exist. It can exist but without being known or it can be equal to zero. It can also be optional or inapplicable for the attribute (Fox & al. 1994, Hainaut 2012). NULL values can lead to different interpretations and a false value can be quickly propagated by copying, making the question of discovering the truth tricky (Dong & al. 2013). To deal with NULL values require a good knowledge not only of the application domain but also of the way data are processed, that it why solutions will vary from one experience to another.
Within the case study "Bxl'air bot", the expertise once again came from the data provider. NULL values were not often met and when it was the case, it was for two reasons: either the value existed but was not published until few days later, or the value did not exist due to technical failures. The potentiality of the presence of NULL errors was deal with specific programming rules in order to inform, for instance, that the value is not yet known. When it was observed, it had also to be fact-checked to avoid misinterpretations. This illustrates the necessity of a human monitoring of the data as well as the process' results.

3 THE FORGOTTEN HUMAN FACTOR

From a process side, an error can be repeated for as long it is not detected. The good news is that once it is made, it will never occur twice, according to Helen Vogt, Head of Innovation at Norwegian News Agency (NTB)[2]. Of course, an algorithmic process (or a software) can fail but most of the time, it will be predictable or avoidable. Reasons of failures will be found among a variety of exogenous factors such as unrealistic or unarticulated project goals, inaccurate estimates of resources, badly defined system requirements, unmanaged risks, inability to handle the project's complexity, sloppy development practices or poor project management (Charette 2005). All of those reasons are converging to a common point: the human factor behind any information system. All of those reasons converge to a common point: the human factor behind any information system. Technologies are not purely mechanical because they are, first of all, the result of social associations (Latour 2005).
Within the context of automated news production, there will always be humans to define a text structure or a data analysis and to translate it into pieces of computational code. By doing so, they translate human intentions into technical requirements through a mediation game between the newsroom and the technical staff. Sometimes, intentions will not be expressed explicitly. This could lead to bias due to the lack of knowledge of what an editorial process is. Technologists and journalists are evolving in two distinct social worlds with distinct occupational norms and values, leading to different views about the nature of journalism and its processes amid technological innovation (Lewis & Usher 2013). This distortion cannot be considered as a direct cause of failure but it should be taken into account because of the possible misunderstandings between each other. It supposes that both share common principles or frameworks, even if questions related to the choice and the assessment of data are constitutive of a journalistic process and that those related to the validation and standardization belong to programmers (Hansen & al. 2017, Linden 2017). According to the innovation diffusion theory, one of the attributes for evaluating an innovation consists in the compatibility with sociocultural values and beliefs of potential adopters (Rogers 2003). What happens when the journalist is also a programmer? It might be a way to facilitate dialogues as well as a reciprocal understanding but those professional profiles remain pretty rare, especially within European newsrooms.

3.1 Biased analysis

Any data analysis activity, which consists in giving significance to the data values, encompasses a part of subjectivity, meaning that numbers are as subjective as words (Espeland & Stevens 2008, Desrosières 2008). It involves complex choices which formalizes the act of counting or measuring or categorizing (Stray 2016). By doing so, data can contain truth but also multiple truths and that can be used to construct a variety of different arguments and conclusions (Diakopoulos 2013). Data-mining algorithms are best at discovering new connections between multiple variables with very high statistical significance due to the huge amount of data being analyzed. Irrelevant questions can be asked and lead to biased answers (Latar 2015). Taken outside of their context or in an isolated way, data analysis can blur reality or give a different meaning. For instance, results of an analysis based on a given time-scale could be different on another time-scale. The deep understanding of the application domain is a pre-requisite. According to McCallum (2012), if you don't understand data, where they are coming from and what they are representing, your conclusions can be biased..
Data analysis does not give rise to an immutable knowledge, especially when the description of reality is a part of a work-in-progress storytelling that could lead to different or even to divergent interpretations as it is the case with air quality data. That is why the authority of numbers can be considered as potentially ephemeral. That is also why data analysis implies a part of human subjectivity, whether automated or not: as in any editorial process, it is always a matter of human choices made upstream of the system. Failures can potentially come from non-relevant or biased interpretations, proving that the black box concept goes far beyond the only technological concerns and should be well discussed before any computational implementation.
Besides, subjectivity can also be totally assumed as it was the case within the experiment "Bxl'air bot". As the web application was designed on a collaborative mode in order to fit the needs of the newsroom, one journalist asked to take into account data from all measurement stations spread all over Brussels area. According to the data producer, from a scientific point of view, one measurement station could not be considered as representative of the urban environment. But from a journalistic point of view, this station, localized in the center of the city, is a place of huge car traffic which was important to consider due the negative effects of pollutants on health. Readers were warned about this voluntary bias which, according to them, gave the project more value.

[2] "Your next recruit might be a robot", Helen Vogt, GEN Summit, Vienna, 2016-06-15.

3.2 Out of human control

A lack of data monitoring and of maintenance of the code are likely to lead to system failures. It is not a secret in computational science, where it is considered that technology is the result of human actions (Orlikowski 1992, MacKenzie 2006). Technology can fail to fulfill its missions, because its code is fragile as it is not safe from software errors, bugs, viruses and other failures (McCosker & Milne 2014). Besides, the error can also be human. Regarding the case of the erroneous report about Netflix, the data analyst should have reflected the stock split (LeCompte 2015). About the case Quakebot, a human fact-checking could have probably avoided the bug.

All kinds of errors cannot always be prevented and in many cases, the only way to detect it will be a human fact-checking. The few examples of failures of automated news systems quoted above illustrate clearly that everything cannot be fully automated. Within the experiment of "Bxl'air bot", maintenance activities did not take more than one hour per month but there were crucial to verify if the right data were retrieved. On at least at five occasions in one year, cells within the table have changed, due to the moving format of a web page, with the necessity to review the scraping process. The monitoring of the values of the retrieved data was another aspect to take care, as values were susceptible to be readjusted with time. And it happened, at least five times. The attention was also focused on an eventual evolution of the European norms about the measurement of air pollutants – a norm is always arbitrary – as well on a measurement station that was out of order at the beginning at the experiment. If something had changed on one of those aspects, the automated web application would have been adapted. But that did not happen.

4 NOT BEING USED

Last but not least, when a news automation system is specifically designed to support journalists within their investigative or daily routines, its biggest symbolic failure consists in not being used. We have here to make a difference between the term "use", which is related to the object, and the term "practice", which is related to the human and covers the fields of the uses and of the attitudes directly or indirectly connected to the object (Jouët 1993).

Considering the ISO 9000 standard and its "fitness for use" principle, the object-tool "Bxl'air bot" was first prototyped outside of the newsroom in order to provide a first material to work with. Its first version was very basic: it provided a real-time report in natural language generation, graphs aiming to follow visually the changing rates of air pollutants, a small range of data analysis (overruns counter days, monthly averages, maximum and minimum observed by measurement station) and an automated Twitter feed. The confrontation with the users, which consists of an inclusive process which can be considered as a first form of use (Akrich 1993), enabled the addition of new functionalities such as map of Brussels fed with real-time data, contextual information about the evolution of the overruns over years and a newsletter to alert susbscribers about overruns.

Because the demands of the newsroom were not always explicit, the role of the researcher was subjective as a translation or a mediation activity cannot be considered as formally neutral or objective. The stabilization of the object occurred within a three months period. Nine months later, the experiment closed and it was time to analyze what the six journalists of the newsroom did with the automated web application as well as what they did not and why.

It is important here to underline that uses cannot be mechanically deducted from the design (Akrich 1993).Uses formation is a long and complex process in which many factors will interact such as the representation of the object or its symbolism (Flichy 2001, Musso 2009), the professional background of the journalist, or the cultural and social values commonly or individually shared. Nevertheless, two main observations finally emerged to explain why the automation system was not used by the whole newsroom, despite its involvement within the investigative project.

4.1 Negative representations

Within the experiment "Bxl'air bot", it was verified that a negative representation had led to non-uses. At least three reasons could explain it. First of all, the metaphor of the "robot journalist" is commonly used to name automated news production systems. It carries a mental image where the machine is placed on the same level as a human journalist or worse, where the machine is likely to take over from professionals. Even if its use appears as a tempting shortcut because it facilitates representations, it can be perceived, by journalists, as a threat to their occupation or to their professional identity. "It's an interesting tool to process data, but that does not replace the journalists to contextualize the facts". "The mega threat of the robot that will steal my job is always there", said two journalists. Secondly, journalists have a long love-hate relationship with information technologies. Taken at its best, that means that technologies are good assistants. Taken at its worst, that means that technologies are likely to affect working conditions. As it was highlighted by a journalist: *Using tools such as social networks are blocking me. Emails are time consuming. I am not against technologies but I am an 'old school' journalist even if I am not already 40*". The third reason is to be found among a data driven approach which is not usual inside the studied newsroom. A journalist said that numbers scary her and that she found difficult to deal with data. It requires skills that she does not have time to learn. Even if a machine does the job, it is perceived as it will never be sufficient to make a good paper. The human input is still considered as remaining the biggest part.

4.2 Lack of interest

Some journalists did not see any interest in using the functionalities of the tool for their article. Others did not visit the web application either because of a lack of interest in air quality issues or in data. However, they have all underlined that the "bot" really came to life when the data collected were processed by human journalists. Despite these findings, the newsroom was unanimous about the idea of restarting the experience but within another application domain, closer to some personal center of interests of some journalists. This observation can be connected to the innovation diffusion theory, which emphasizes on the role of the perceived relative advantage within an innovation process adoption (Rogers 2003). A journalist explained that she tried to collect the data by herself but that she

gave up because it was too time-consuming and that the information provided "Bxl'air bot" gave novel insights to feed her articles.

5 DISCUSSION AND CONCLUSION

Reasons why automated news production can fail are both human and technical. If it rarely fails, it has already occurred and it was not consequence-free, especially in the field of financial information where markets are particularly sensitive. Even if errors can be quickly detected, they cannot always be prevented. A formal approach of data does not appear sufficient to prevent failures, neither the multidimensional characters of data quality, which includes empirical considerations. It illustrates the necessity of a human control on the data as well as on contents it produces. Furthermore, human intentions are hidden beyond any technical process.

The way it will be translated is potentially subjective. That's why automated news technologies cannot be considered as purely mechanical. Those considerations allowed the role of a shared framework between those who are developing automated news production and journalists to arise. From a journalistic point of view, the understanding of how technologies are running can enable journalists to dig digger into it and, in some case, to find good stories behind the data. It also participates in a form of "computational thinking", as promoted by a few scholars (Linden 2017), which could allow journalists to make the bridge with technologists. Thinking in terms of process is not so far from traditional journalistic routines. Here, the editorial process is a kind of remix of something that existed before automation. Choices always have to be made in order to answer two questions: what to say and how to say it? From a constructivist point of view, this process cannot be totally bias-free.

Failures are not only technical: they are also social. The different reasons explored in this paper show that automated news production cannot be reduced solely to the data aspects, which remain important because they are feeding the systems. The place of the human journalist in the socio-technical chain appears as crucial viewed through the lenses of the hypothesis of a fruitful man-machine collaboration. If the way might seem long for the ones who do not feel attracted by information technologies or by data driven journalism, there are signs which enable us to think that doors are open regarding the perceived advantages when automation is designed to support journalists. In any cases, technical or social, fitness for use should be the key even it remains insufficient to guarantee the effectiveness of the different kind of uses.

REFERENCES

[1] Akrich, M. De la sociologie des techniques à une sociologie des usages. In *Techniques & Culture* (1990), no. 16, pp. 83–110.

[2] Akrich, M. Les objets techniques et leurs utilisateurs : de la conception à l'action. *Raisons pratiques 4* (1993), 35–57.

[3] Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. Methodologies for data quality assessment and improvement. *ACM Computing Surveys (CSUR) 41*, 3 (2009), 16.

[4] Boydens, I. L'océan des données et le canal des normes. *Les Annales des Mines*, 67 (juillet 2012), 22–29.

[5] Boydens, I., and van Hooland, S. Hermeneutics applied to the quality of empirical databases. *Journal of Documentation 67*, 2 (2011), 279–289.

[6] Brodie, M. L. Data quality in information systems. *Information & Management 3*, 6 (1980), 245–258.

[7] Clerwall, C. Enter the robot journalist. *Journalism Practice 8*, 5 (2014), 519–531.

[8] Daniel, A., Flew, T., and Spurgeon, C. The promise of computational journalism. In *Media, Democracy and Change: Refereed Proceedings of the Australian and New Zealand Communications Association Annual Conference* (Canberra, 2010), Australia and New Zealand Communication Association, pp. 1–19.

[9] Desrosières, A. *L'argument statistique : gouverner par les nombres.* Collection Sciences sociales. Presses de l'école des mines, 2008.

[10] Diakopoulos, N. The rethoric of data. *Tow Center for Digital Journalism* (July 2013). URL: https://towcenter.org/the-rhetoric-of-data/.

[11] Dierickx, L. News bot for the newsroom: how building data quality indicators can support journalistic projects relying on real-time open data. In *Global Investigative Journalism Conference 2017 Academic Track* (2017), Investigative Journalism Education Consortium.

[12] Dong, X. L., Berti-Equille, L., and Srivastava, D. Data fusion: resolving conflicts from multiple sources. In *Handbook of Data Quality*. Springer, 2013, pp. 293–318.

[13] Espeland, W. N., and Stevens, M. L. A sociology of quantification. *European Journal of Sociology/Archives Européennes de Sociologie 49*, 3 (2008), 401–436.

[14] Flichy, P. La place de l'imaginaire dans l'action technique. *Réseaux*, 5 (2001), 52–73.

[15] Fox, C., Levitin, A., and Redman, T. The notion of data and its quality dimensions. *Information processing & management 30*, 1 (1994), 9–19.

[16] Fox, P. Data life cycle: Introduction, definitions and considerations, 2014.

[17] Golab, L., and Johnson, T. Data stream warehousing. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data* (2013), ACM, pp. 949–952.

[18] Graefe, A. Guide to automated journalism. *Tow Center for Digital Journalism* (January 2016). URL: https://www.gitbook.com/book/towcenter/guide-to-automated-journalism/details.

[19] Hainaut, J.-L. *Bases de données-2e éd.: Concepts, utilisation et développement.* Dunod, 2012.

[20] Hansen, M., Roca-Sales, M., Keegan, J. M., and King, G. Artificial intelligence: Practice and implications for journalism.

[21] Haug, A., Stentoft Arlbjorn, J., and Pedersen, A. A classification model of ERP system data quality. *Industrial Management & Data Systems 109*, 8 (2009), 1053–1068.

[22] Haug, A., Zachariassen, F., and Van Liempd, D. The costs of poor data quality. *Journal of Industrial Engineering and Management 4*, 2 (2011), 168–193.

[23] Jouët, J., and Sfez, L. Usages et pratiques des nouveaux outils de communication. *Dictionnaire critique de la communication 1* (1993), 371–376.

[24] Latar, N. L. The robot journalist in the age of social physics: the end of human journalism? In *The New World of Transitioned Media*. Springer, 2015, pp. 65–80.

[25] Latour, B. *Reassembling the social: an introduction to Actor-Network-Theory.* Clarendon Lectures in Management Studies. Oxford University Press, 2005.

[26] Lecompte, C. Automation in the newsroom. *Nieman Reports 3*, 69 (2015), 32–45.

[27] Leppänen, L., Munezero, M., Sirén-Heikel, S., Granroth-Wilding, M., and Toivonen, H. Finding and expressing news from structured data. In *Proceedings of the 21st International Academic Mindtrek Conference* (2017), ACM, pp. 174–183.

[28] Lewis, S. C., and Usher, N. Open source and journalism: toward new frameworks for imagining news innovation. *Media Culture Society 35*, 5 (2013), 4–9.

[29] Linden, C.-G. Algorithms for journalism. *The Journal of Media Innovations 4*, 1 (2017), 60–76.

[30] Mackenzie, A. *Cutting code: software and sociality.* Digital formations. Peter Lang, 2006.

[31] Madnick, S., and Zhu, H. Improving data quality through effective use of data semantics. *Data & Knowledge Engineering 59*, 2 (2006), 460–475.

[32] McCallum, E. Q. *Bad data handbook: cleaning up the data so you can get back to work.* O'Reilly Media, 2012.

[33] McCosker, A., and Milne, E. Coding labour. *Cultural Studies Review 20*, 1 (2014), 4.

[34] Moody, D. L., and Shanks, G. G. Improving the quality of data models: empirical validation of a quality management framework. *Information systems 28*, 6 (2003), 619–650.

[35] Musso, P. Usages et imaginaires des TIC: la fiction des frictions. *L'évolution des cultures numériques: de la mutation du lien social à l'organisation du travail, Limoges, FYP éd* (2009).

[36] Orlikowski, W. J. The duality of technology: rethinking the concept of technology in organizations. *Organization science 3*, 3 (1992), 398–427.

[37] Redman, T. C. *Data quality for the information age.* Artech House Telecommunications Library. Artech House, 1996.

[38] Reid, R., Fraser-King, G., and Schwaderer, W. *Data lifecycles: managing data for strategic advantage.* Wiley, 2007.

[39] Rogers, E. *Diffusion of Innovations, 5th Edition.* Free Press, 2003.

[40] Stoneman, J. Does open data need journalism? *Reuters Institute for the Study of Journalism, Oxford University* (2015).

[41] Stray, J. *The curious journalist's guide to data.* Tow Center for Digital Journalism, 2016.

[42] Strong, D. M., Lee, Y. W., and Wang, R. Y. Data quality in context. *Communications of the ACM 40*, 5 (1997), 103–110.

[43] Wang, R. Y., Ziad, M., and Lee, Y. W. *Data quality*, vol. 23. Springer Science & Business Media, 2006.