Ethics of data journalism

NODA 2015 Helsinki, 21 April 2016

Heikki Kuutti, University of Jyväskylä, Finland

There have been few discussions about ethical requirements and responsibilities of data in journalism. For instance, national press councils do not particularly discuss data and its impact on journalism ethics.

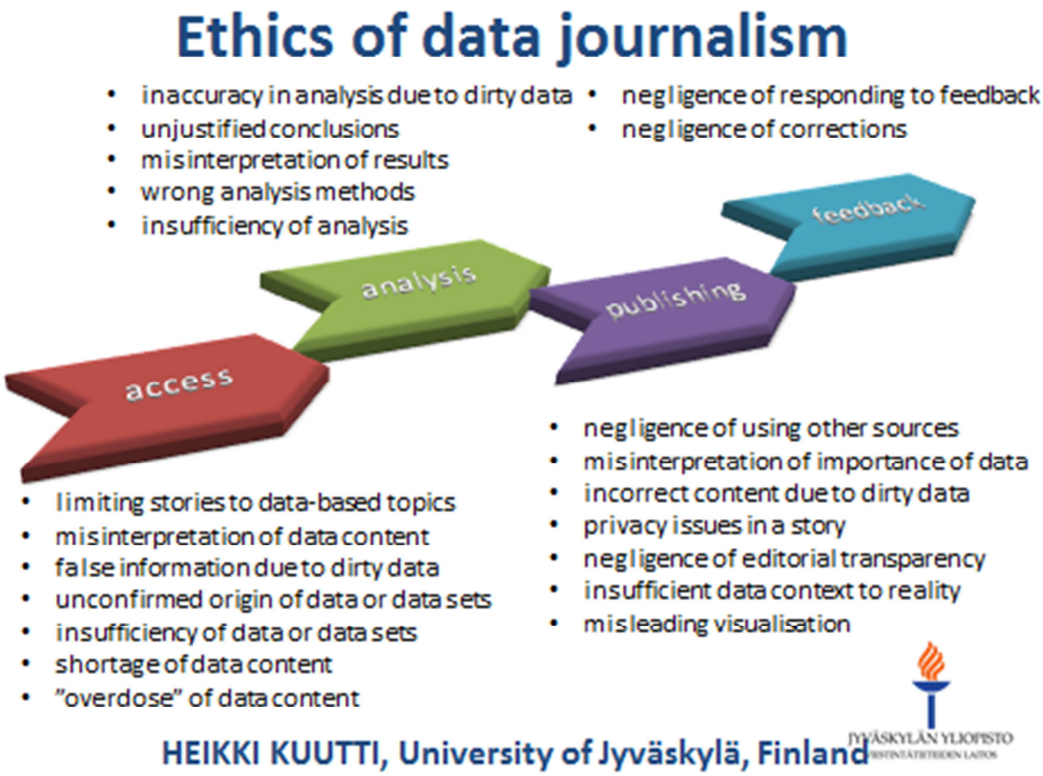
Data when mentioned in national ethical codes usually refer to information and its accuracy requirements which should be taken into consideration in information gathering and publishing.

Lack of ethical complaints may be explained by the difficulty to separate the importance of data from the ethics in data journalism as a whole and compared with other journalistic activities in producing a data story.

Ethical problems in a data story are difficult to discern, unlike misspelled interviewee´s quotes or obviously false information in a story. Whenever suspicions of a journalist´s analysis or interpretations are raised, the search to corroborate incorrect information in a data story is a time-consuming exercise, which also requires the ability to reanalyse the original data.

In this presentation, I have divided ethical problems into four phases in the journalistic working process: access to data, data analysis, publishing a data story and asking the public for feedback.

Each of these phases contains several ethical point-of-views, some of which are common and have their own implementation during the working process. For instance “dirty data” in the access phase refers to the content of data and data sets, in the analysis phase possible errors in analysing data material and in the publishing phase ethical claims to inform the public about possible problems caused by dirty data. The same variations are included in a journalist´s interpretation depending of its object: what does dirt in data mean in the contexts of access, analysis or publishing.



Access

Dirt can be in data itself and in larger data sets. Data is dirty if it is incorrect, stored in a varying format or if it is out of date. Dirt in data sets refers to data bases in which relevant data or whole data fields may be missing, data are stored in incorrect data fields or the same data are stored more than once.

Penalty-free access to data may lead to its limitless use and invite journalists to take advantage of data hastily and carelessly without considering restrictions of data in a story. In addition, this kind of “data blindness” may marginalize information gathering only to data-based materials and those (less important) story ideas where data is available. Consequently, “short-sighted” journalists spend their working time with computers instead of meeting and talking with people or exploring other sources for their stories.

Analysis

It is an incorrect hypothesis that data cannot be untruthful or computer analyses cannot make mistakes. Journalists should be as sceptical with data as with human sources. ‘Interviewing’ data needs rough questions

to data material: how the analysis is carried out in detail, what kinds of results have been eliminated of the story and what has been the interpretive logic behind journalists´ conclusions.

Publishing

Invasion of privacy is one of the major ethical problems in publishing a data story. Data may be confidential or connected to the wrong person. Data as a part of larger data sets may contain less important additional information that is actually not needed for a particular story. However, this kind of “data overdose” offers journalists a kind of “peeping tom situation”. For instance a gun license data set may include personal information about reasons of license cancellations, such as the gun owner´s mental problems.

Besides “data overdose” situations, journalists may be exposed to “over interpretations” when publishing their results. Data should not be forced to tell more than it is able to do and journalists should ask themselves do data tell precisely what they should do, is analyse comprehensive enough and in context with reality. Journalists are eager to search causalities between variables but may not take into account the not-so-visible “third element” that could explain exceptional results.

Another ethical problem in publishing relates to incorrect interpretations when journalists write their stories solely based on data and without using other sources or requesting outside experts to comment analyse results. If data is the only source material in a story, the journalist may be tempted to ‘over report’ analysis results and to neglect some important and relevant issues in the topic.

In a “cautious data story” the question is how much the journalist can “stretch” data analysis if data are not 100 percent clean. Dirty data is relative and certain amount of dirty data does not necessarily pollute the whole data set. However, even small error threats caused by dirty data should be taken into account in analysis results and explained in the story.

Measured results alone do not tell anything about whether those numbers are higher or lower than they should be, or situationa best or worst. Presenting them within a historical context, by person or by date helps make the numbers more meaningful. If journalist tells readers how things affect them, they should also have a sense of the bigger picture

Journalists need to play the “devil´s advocate” with the data and come up with arguments against any initial finding.

Visualization creates an impression among the public that the information in the background is reliable. To avoid ethical problems, visualization and data should have a clear connection between each other. Journalists should be aware of what kind of visualisation is most suitable to what kind of presentation. Unethical visualization may cause confusion among the public and lead to false interpretations.

Feedback

In order to convince the media public of the analysis results, journalists´ activities should be made as transparent as possible. A data story should contain a description of the used data and include its origin, content and character. Important information to the public would also be analysis methods used in the story. Interpretations and conclusions of journalists should contain some kind of “disclaimer” and invitation to the public to inform journalists about possible errors or absences in data material.