ContentCheck: Content Management Techniques and Tools for Fact-checking

by Ioana Manolescu (Inria Saclay and Ecole Polytechnique, France)

Data journalism and journalistic fact-checking make up a vibrant area of applied research. Content management models and algorithms have the potential to tremendously multiply their power, and have already started to do so.

The immense value of big data has recently been acknowledged by the media industry, with the coining of the term "data journalism" to refer to journalistic work inspired by data sources. While data is a natural ingredient of all reporting, the increasing volumes of available digital data as well as its increasing complexity lead to a qualitative jump, where technical skills for working with data are stringently needed in journalism teams.

An ongoing collaborative research programme focused on novel content management techniques applied to data journalism and fact-checking has recently been initiated by: Inria, the LIMSI lab (Université Paris Saclay, CNRS and Université Paris Sud), Université Rennes 1, Université Lyon 1 and the "Les Décodeurs" fact-checking team of Le Monde, France's leading newspaper [L1]. Here, content is broadly interpreted to denote structured data, text, and knowledge bases, as well as information describing the social context of data being produced and exchanged between various actors. The project, called ContentCheck [L2], is sponsored by ANR, the French National Research Agency, and is set to run until 2019.

The project goals are twofold:

- First, in an area rich with sensational fake news, debunked myths, social media rumors, and ambitious start-ups, we aim at *a comprehensive analysis of the areas of computer science* from which data journalism and fact-checking can draw ideas and techniques. Questions we seek to answer include: what kinds of content are involved? What is their life cycle? What types of processing are frequently applied (or needed!) in journalistic endeavours? Can we identify a blueprint architecture for the ideal journalistic content management system (JCMS)?
- Second, we seek to *advance and improve the technical tools available for such journalistic tasks*, by proposing specific high-level models, languages, and algorithms applied to data of interest to journalists.

Prior to the start of the project, interviews with mainstream media journalists from Le Monde, The Washington Post and the Financial Times have highlighted severe limitations of their JCMSs These are typically restricted to archiving published articles, and providing full-text or category-based searches on them. No support is available for storing or processing external data that the journalists work with on a daily basis; newsrooms rely on ad-hoc tools such as shared documents and repositories on the web, and copied files to and fro as soon as processing was required. The overhead, lost productivity, privacy and reliability weaknesses of this approach are readily evident.

The main findings made in our project to date include:

- Data journalism and fact-checking involve a wide range of content management and processing tasks, as well as human-intensive tasks performed individually (e.g., a journalist or an external expert of a given field whose input is solicited - For instance, ClimateFeedback is an effort to analyse media articles about climate change by climate

scientists. See, for example, [L3]), or collectively (e.g., readers can help flag fake news in a crowd-sourcing scenario, while a large consortium of journalists may work on a large news story with international implications. The International Consortium of Investigative Journalism is at the origin of the Panama Papers [L4] disclosure concerning tax avoidance through tax havens).

- Content management tasks include the usual CRUD (create, read, update, delete) data cycle, applied both to static content (e.g., web articles or government PDF reports) and dynamic content such as provided by social media streams.
- Stream monitoring and stream filtering tools are highly desirable, as journalists need help to identify, in the daily avalanche of online information, the subset worth focusing on for further analysis.
- Time information attached to all forms of content is highly valuable. It is important to keep track of the time-changing roles (e.g., elected positions) held by public figures, and also to record the time when statements were made, and (when applicable) the time such statements were referring to (e.g., when was a certain politician's spouse employed, and when did the politician share this information).

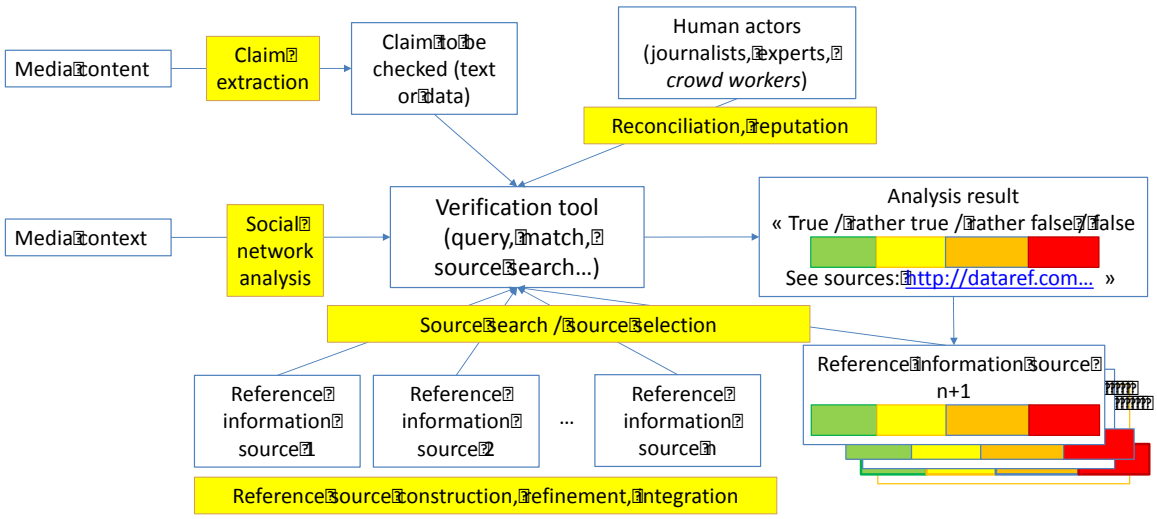

Figure 1: Flow of information in fact-checking tasks, and relevant research problems from the content/data/information management area.

Work to design a single unified architecture for a content management tool dedicated to fact-checking is still ongoing as we write. The overall vision we currently base our analysis on is outlined in Figure 1. Claims are made through various media, and (importantly) in a context, in which one can find the claim's authors, their institutions, friend and organisational affiliations etc. Claims are fact-checked against some reference information, typically supplied by trustworthy institutions, such as statistics national institutes (INSEE in France, the Office for National Statistics in the UK) or trusted experts, such as well-established scientists working on a specific topic. Claims are checked by human users (journalists, scientists, or concerned citizens), possibly with the help of some automated tools. The output of a fact-checking task is a claim analysis, which states parts of the claims that are true, mostly true, mostly false etc., together with references to the trustworthy sources used for the check. Fact-checking outputs, then, can be archived and used as further reference sources.

Scientific outputs of the project so far include a light-weight data integration platform for data journalism [1], an analysis of EU Parliament votes highlighting unusual (unexpected) correlations between the voting patterns of different political groups [2], and a linked open data extractor out of INSEE spreadsheets [3]. Our project website is available at: https://team.inria.fr/cedar/contentcheck.

Links:
[L1] http://www.lemonde.fr/les-decodeurs
[L2] https://team.inria.fr/cedar/contentcheck
[L3] https://climatefeedback.org/evaluation/scientists-explain-what-new-york-magazine-article-on-the-uninhabitable-earth-gets-wrong-david-wallace-wells/
[L4] https://panamapapers.icij.org/

References:
[1] Raphaël Bonaque, Tien Cao, Bogdan Cautis, François Goasdoué, Javier Letelier, Ioana Manolescu, Oscar Mendoza, Swen Ribeiro, Xavier Tannier, Michaël Thomazo. "Mixed-instance querying: a lightweight integration architecture for data journalism" (demonstration), PVLDB Conference, 2016
[2] Adnene Belfodil, Sylvie Cazalens, Philippe Lamarre and Marc Plantevit. "Flash points: Discovering exceptional pairwise behaviors in vote or rating data", ECML-PKDD Conference, 2017
[3] Tien Duc Cao, Ioana Manolescu, Xavier Tannier. "Extracting Linked Data from statistic spreadsheets", Semantic Big Data Workshop 2017.

Please contact:
Ioana Manolescu
Inria Saclay and Ecole Polytechnique, France
ioana.manolescu@inria.fr