Open Data in Data Journalists' Workflow

Uldis Bojārs and Edgars Celms

Institute of Mathematics and Computer Science
University of Latvia, Riga, Latvia
{uldis.bojars, edgars.celms}@lumii.lv

Introduction

In order to fully realise the potential of Open Data we need technologies and tools that help users to discover, transform and re-use this data.

Data analysis, processing and management is no longer just a task of programmers or analysts who prepare the data sets or build tools to publish them online. For the Open Data to be truly useful a wide range of people should be able to process and make sense of the data. These users may have a need to discover, process and visualise the data, and to publish the results (which may include producing derived datasets). Yet they do not have necessary skills to build data processing tools themselves and need tools that would help in their work.

Data-Driven Journalism

Data-driven journalism is a journalistic process based on analyzing and filtering large data sets for the purpose of creating a new story [1]. It can be viewed as a workflow where the data is cleaned, transformed, visualized and a story is formed based on the results of this data exploration [2].

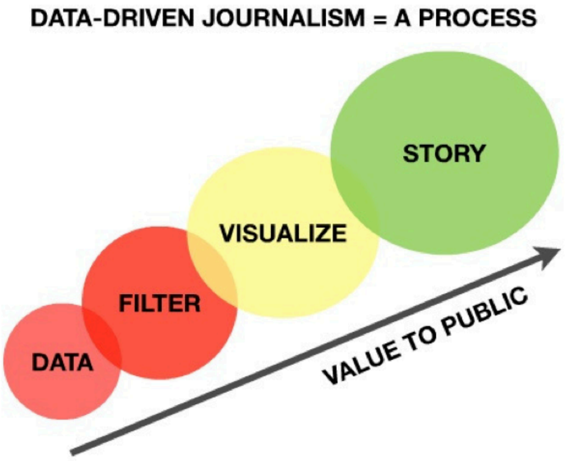

Figure 1. The data-driven journalism process [2].

Data-driven journalism is one of the potential uses of open data published on the Web. Most journalists are not programmers and will not be creating tools for themselves therefore they need to be provided with tools that support all steps of the data journalism workflow.

One can envision a tool or a set of tools that can be combined together to create workflows covering the whole process of working with data (Figure 1) while keeping track of data provenance and enabling us to repeat the workflow when necessary (e.g., when a new set of data becomes available).

Data: Discovery and Publishing

Two steps that "wrap around" the process of data exploration are (1) data discovery and (2) publishing of results.

Data discovery: before the data can be processed we need to get the data to work with. There is a growing number of open data sources available on the Web and the user needs to be able to find the right combination of datasets, possible coming form multiple sources.

Publishing of results: once the story is complete it can be published. However, the value is not just in the story but also in the data. This can be (a) the resulting data that appear in the story (e.g., as a chart); (b) the intermediate data created in the process; or (c) provenance data describing the transformations that the data went through. By publishing the machine-readable data associated with a story we can facilitate further data reuse and the creation of new stories.

Workflows for the analysis of open data need to include (1) data discovery functionality that helps users find and start using open data relevant to their research, and (2) data publishing functionality that supports users in publishing the resulting data.

In an example scenario a user might use data discovery interface to select a number of data sources to work with. This interface would take information about the data needed (e.g., the type of data, time-frame and keywords) and return a ranked list of suitable data sorces. User's work on the data (such as cleaning and transformation) and its results would be recorded by the tool framework. This information would complement the final result -- the story and data included in it -- and can be of value to the user, their organisation and the open data community at large.

Conclusion

In this paper we refer to data-driven journalism as a use case for open data on the Web and emphasize the need for tools and technologies that support it. Two important processes that data exploration starts with and completes with are data discovery and publishing of the results. Data journalism tools need to support these processes. By publishing the data created as a part of the process we can make stories richer and create new data that can, in turn, be reused.

References

1. *Data driven journalism*. Wikipedia article. Retrieved on March 1, 2013. http://en.wikipedia.org/wiki/Data_driven_journalism
2. Lorenz, Mirko. *Data driven journalism: What is there to learn?* Conference materials, based on presentations of participants, August 24, 2010, Amsterdam, The Netherlands. http://mediapusher.eu/datadrivenjournalism/pdf/ddj_paper_final.pdf