The View on Open Data and Data Journalism: Cases, Educational Resources and Current Trends

Irina Radchenko[1(✉)] and Anna Sakoyan[2]

[1] ITMO University, St. Petersburg, Moscow, Russia
iradche@gmail.com
[2] Russian Analytical Publications Polit.ru, Moscow, Russia
ansakoy@gmail.com

Abstract. This article describes trends of open data development and a new discipline, which was formed largely due to the fact that the data have become available and open on the Internet. The authors provide a brief overview of the main directions in the development of open data and data journalism: educational projects, interaction with the community of developers using data management platforms, development of business community on open data basis. The article also discusses Russian educational projects dealing with open data and data journalism.

Keywords: Open data · Open government data · Data journalism · Open educational resources · Data expeditions · Open government

1 Introduction

There are two main features of the approach to the publication of data on the Internet in the form of open data. First, data should be freely available on the Internet, without publisher's control; second, this data must be in a reusable form. For this reason, open data should be accompanied by appropriate open licenses. It should be also submitted online in a machine-readable format. It should be machine-readable formats that allow implementing automatic processing and creating a variety of useful services based on open data.

Data journalism is one of really powerful directions within this movement [1]. The availability of data and the opportunity to create data-driven products provide a serious advantage for the informational agencies, which is specially vital now when their monopoly on information distribution is starting to crush as a consequence of the growing popularity of social networks. Moreover, the very fact of using data in media reporting is usually more explicit and visible to the audience than in many other data based products. This makes a huge contribution to the promotion of the open data movement.

© Springer International Publishing Switzerland 2014
D.I. Ignatov et al. (Eds.): AIST 2014, CCIS 436, pp. 47–54, 2014.
DOI: 10.1007/978-3-319-12580-0_4

2 Open Data for Education, Developers Community and Online Services

Open data have become an important tool in the development of new forms of interaction between government and civil society.

However, there is no particular value in open data, when it is not used by anyone. Open data becomes important only if there is a community of skilled users around them. One of the key components of this community is business, which comes as a stimulus for the creation of new technologies based on the new approach to open publishing data on the Internet. It is the business prospects that produced necessary conditions for the active development of data driven startups in the overseas countries, which was also encouraged by a considerable amount of governmental support at early stages of development. After all, the government is also interested in the promotion of the open data concept, because it allows for the development of new ways of dealing with civil society. This implies involving citizens in active cooperation with the government, as well as the growing transparency and accountability on the part of the authorities [2].

The practice of funding startups is widely spread in the US and UK, the two frontrunners in the area of open government and open data. This trend became even more explicit in the late 2013 and early 2014 when the federal open data portals of the US and UK changed their designs to be friendlier to their target audience of open data users. Now, there is a need for good business models of data based initiatives, as well as show cases of working and profitable projects in order to make the developers' community more motivated and the businesses more interested. It is necessary for theoretical speculations to be backed up by practical implementations, which demonstrate the prospects of using open data in analytical research and program products.

In the United States, Open Data 500 Project [3] is represented, among all, by a considerable list of startups based on open data. In the United Kingdom, The Open Data Institute has also launched a special program to support open data startups [4]. Also, steps are taken to provide the conditions for open data users to share their feedback, in order to improve the quality of the data sets that are being published. For instance, the UK federal open data portal has forums where users can discuss which directions still need more data openness, as well as the quality of the data sets that have been already published, the details of how they are published and processed, and the methods and tools for their use, etc. The portal also has blogs by the portal developers team, representatives of governmental agencies, as well as companies dealing with open data and conducting analytical research. There is also a section discussing linked open data at the portal. It provides some basic information on the topic and has a special sequence of posts tagged 'Linked Data' [5, 6]. Finally, there is a separate section on geodata, in particular the European INSPIRE Directive (Infrastructure for Spatial Information in the European Union).

In order to make working with data sets easier, developers of federal, as well as local, data portals use CKAN (The Comprehensive Knowledge Archive Network) data-publishing platform. The numerous examples include official open data portals of the UK, the US, Canada, Australia, Brazil, Spain, Slovakia, Romania, Norway, Austria, Sweden, the Netherlands, Italy, Germany, Iceland, Argentina, France, Switzerland, etc.

CKAN owes its popularity to its open-source code provided with the option of paid support by Open Knowledge Foundation.

During 2013, the popularity of CKAN was growing rapidly, as many governmental portals switched to this platform from Socrata or self-developed platforms.

Although some are still using alternative platforms, such as Socrata, Microsoft, or Koema, CKAN has already started conquering the globe. Moreover, Microsoft has provided options for CKAN deployment in its cloud (Windows Azure platform [7]).

The more portals are built using CKAN, the more comfortable it becomes for users, because it is always easier to deal with a service with a familiar interface.

It is also important that CKAN's API has single documentation [8], which means that the access to open data sets is universal and, consequently, easy to use for software developers.

In a number of states, governments went even further and openly published the source code of their services and applications, as well as documentation at GitHub [9].

Another direction, in which the open data community is developing, is educational. Apparently, it is impossible to promote the use of open data without explaining what open data is, why it is important and how to work with it.

To this end, activist communities and official institutions in different countries come up with special courses aimed at teaching the basics of working with open data. The educational direction is developed, among all, by The Open Data Institute, Open Knowledge Foundation, Knight Foundation, and Sunlight Foundation. The UN has also been interested in the promotion of the new approach. For instance, FAO UN launched a series of webinars on linked open data in late 2012 [10], and in addition to it the concept of Linked Open Data was presented in Russian as well [11]. Not long time ago, The UN launched its own online course on open data processing [12]. World Bank is also engaged in educational activities in the field of open data [13]. In September 2013 School of Open Data, organized by NGO Infoculture, was launched in Russia [14].

In some countries, a lot of efforts are aimed specifically at forming communities of software developers. These activities may include holding hackathons and organizing civil activists (The Code for America Brigades [15]), as well as launching contests for developers of open data applications. An important event in this respect is the yearly worldwide hackathon on the Open Data Day under the auspices of Open Knowledge Foundation. It should be noted that in 2014, it was the first time Russia has participated in this international event [16].

There are a lot of news subscriptions that can be used to regularly receive updates on recent open data developments. This format is actively employed by Open Knowledge Foundation, Code for America project [17], Citizens for Open Access to Civic Information and Data project [18], Map and GIS Services at the University of Toronto [19]). Special Twitter hashtags are also often used for spreading news across the community. These tags are helpful when it comes to clustering particular topics.

One more aspect of development of the open data community is building free and often online-based tools and services for data processing. An outstanding example of such products is OpenRefine, a handy tool for working with raw data. It allows cleaning datasets using different clustering methods, the method of deduplication, sorting, filtering, regular expressions and a programming language GREL (General Refine Expression Language).

It is also necessary to mention such popular tools as Google spreadsheets, Quandl [20], Raw [21], R programming language, numerous free online services for data visualization, and so on.

3 Data and Journalism

Data journalism is a specific direction in the Open Data movement, as it appears to hold an intermediary position between the data based businesses and socially explicit use of data. Like any other business at least partially based on data analysis, data journalism fits in certain profit-making business models. On the other hand, like the projects in the area of education and civil activism, it is at liberty to demonstrate the underlying mechanisms that lead to the resulting product. Moreover, unlike many spheres, in which statistics and data analysis have traditionally played a huge role (like finance or business analysis), data journalism has developed in the area traditionally associated with humanities. Although there was some amount of work with databases and numbers, still until recently there simply has not been enough material available to make it a separate profession A similar process of adapting digital techniques to traditionally non-digital spheres is now underway in many more disciplines and arts, such as literary criticism [22]. The specific of data journalism is that, unlike scholarly research, it is addressed to a vast general public audience, which potentially makes it a significant tool for promotion of open data by presenting examples of their application.

The meaning of the term data journalism seems to be generally clear, while when it comes to the details, it turns out that its application is actually defined very vaguely [23]. Even though there are numerous attempts of reflection, there is still no common understanding of what exactly it means. There are though several terms, sometimes with overlapping meanings, which are aimed to reflect certain subdirections of this trend. Here are some of them. *Data driven journalism* [24] is normally about producing a digital story which relies heavily on big data sets shown through the lens of interactive visualization. *Database journalism* [25], in its turn, is first and foremost about creating data sets and databases, which could be further used by data driven journalism. There is also *computational journalism* [26] that applies computational techniques and the elements of machine/statistical learning to creating a data driven story. A very loose term *analytic journalism* [27] might also be used in the relation to data journalism as it is sometimes based on data analysis and results in a story describing the conclusions and findings.

The obvious tendency is that more and more processes in many areas, journalism included, are becoming automated. For instance, in the recent years, there have been quite a number of examples of applying the techniques for generating human-readable texts based on data analysis to creating media articles. Several developers, including Narrative Science [28] and Automated Insights [29] have been successfully producing such tools for several years now. They may be rather expensive and not too popular among the media for now, especially if we look at the global scope, but still they are already in use there is a chance that they will become more available and therefore more powerful in the future, both in terms of the resulting product and the skills required from journalists.

That was a somewhat extreme example of digitalization applied to the traditionally human-generated content. But there are also much more common practices that have been already adopted by many leading media. Data based reporting is now becoming one of the necessary parts of a media genre system, while big newspapers create interactive visualizations or at least static infographics aimed at providing their audience with a helpful tool of exploring datasets and extracting relevant information (for examples, Kathimerini [30], New York Times [31], The Guardian [32]).

Again, the popularity of such methods seems to be different across the world at the moment. Many media companies still hesitate to invest in developing specialized data units within their staff, because hiring ready-made specialists or training the existing ones seems too expensive. Still, as the world trend to using data (including big data) in journalism is becoming stronger and more and more people are trained in working with data at least on the basic level, there is a growing need for publishers to consider the introduction of new methods. It is becoming the matter not only of fashion, but of efficiency and competitiveness as well.

In this respect, the Russian media represents an instructive example. The open data movement in the country became widely discussed only after the government resolved to start opening official data in 2013. Before that openness was mostly discussed by activists, but remained somewhat abstract and unnecessary for the majority. However, now that the data are being published, this subject evokes much broader discussion and more individuals and companies seek to make some practical use of it. Media are no exception. There are more and more examples of quality data based stories and visualizations produced by the Russian media companies (for examples, RIA Novosti [33, 34], Lenta.ru [35]).

So the process towards developing, exploring and adopting data based techniques by the media is underway. In order to be competitive, media have no choice but to become more proficient in this area. This, in turn, requires certain conditions, including: the availability of quality data, the developed tools for extracting data, the skilled specialists that can work with data and produce clear and informative data based stories, as well as visualizations. To a certain degree, it can be achieved just by the efforts of data enthusiasts who enjoy learning new techniques on their own and then apply them to their job. But this is by no means enough, unless the publishers realize the need for direct investment in this particular section.

4 Data Journalism: Educational Initiatives

Again, broadly speaking, there are some reasons to suggest that the environment necessary for making use of open data will develop naturally on its own. But it is a complex phenomenon, which is moved by a combination of top-down and bottom-up initiatives. The more people have some experience in working with data, the better they can understand the meaning and the quality of data driven stories provided by the media. The more competent the audience is, the stronger is the stimulus for the media to try hard to produce a good product. On the other hand, the less journalists are scared of learning something technical (which is a rather common fear today), the better

chances they have to take interest in using data. These are two examples of what might create a healthy environment, in which the media managers would elaborate their strategies.

Speaking of the environment, one of the key factors here is broadly available, ideally open educational initiative by individual activists, not-for-profits, governments, etc. These provide not just information about what new trends exist, but also allow acquiring the basic skills, overcoming fears and understanding the meaning of what is going on. For some it may remain the only source of this kind of knowledge; others may take a more profound interest and continue learning to reach more advanced levels. All in all, educational activity in this area seems vital to maintain a healthy society.

Such initiatives have been introduced all over the world, both on national and international levels. In Russia, one of such early educational projects is DataDriven-Journalism.RU, an open resource created by a tiny team of enthusiasts in April 2013. Inspired by such internationally renowned examples as Open Knowledge foundation's School of Data [36] and Peer-to-Peer University [37], it aspires to provide a Russian-language learning opportunity. The resource as it is has a blog that discusses the relevant subjects and storage of tutorials and how-tos, both translated and original, that can be used at any time by anyone. It is also used as a central platform for so-called data-expeditions, interactive peer-learning projects aimed at developing skills of data processing and exploring certain topics. These expeditions take place online and are free of charge, so anyone who is interested can join [38]. This is an example of a bottom-up initiative that struggles to contribute to a well-balanced development of a data-driven environment.

5 Conclusion

Open Data concept is a step forward in the transformation of the Web of Documents to the Web of Data. They allow building on the basis of its new services, to conduct analytical research and empower investigative journalism, and because of this Open Data provide some freedom of information for the understanding of the processes. Open Data give new opportunities for cooperation between the government, citizens and businesses, while transforming the information environment of the Internet. In order to translate these powerful new opportunities into life, organizations (and individuals) in various countries come up with Open Educational Resources introducing the skills necessary to work with open data. Thus, new approaches and technologies become available to a wide range of stakeholders, and the information actually leads to democratization of the society and the transition to the real information society.

Acknowledgements. We take this opportunity to express our profound gratitude to many people who have helped us with their support, assistance or inspiring example. In particular, we would like to thank Ivan Begtin, Director at NGO Infoculture, for his support, encouragement and making a huge contribution to the development of the environment, in which our work takes place. We would also like to thank Lucy Chambers, Project Coordinator at OKF, whose Data MOOC at School of Data last spring was a great source of inspiration for us. Another person

whose research, as well as example and encouragement have been precious for us is Vanessa Gennarelli, Learning Lead at P2P University. One of our most fruitful experiments owes a lot to the cooperation and initiative on the part of Alexey Sidorenko, the head of the project Teplitsa of Social Technologies, and his crew.

This work was partially supported by Government of Russian Federation, Grant 074-U01.

References

1. Gray, J., Chambers, L., Bounegru, L.: The Data Journalism Handbook. O'Reilly Media, Sebastopol (2012). http://datajournalismhandbook.org/
2. Ridgway, J., Smith, A.: Open data, official statistics and statistics education: threats, and opportunities for collaboration. In: Proceedings of the Joint IASE-IAOS Satellite Conference "Statistics Education for Progress", Macao, China, 22–24 August 2013
3. Open Data 500 GovLab. http://www.opendata500.com/
4. Open Data Institute. Startups. http://theodi.org/start-ups
5. Bizer, Christian, Heath, Tom, Berners-Lee, Tim: Linked data - the story so far. Int. J. Semant. Web Inf. Syst. **5**(3), 1–22 (2009). doi:10.4018/jswis.2009081901
6. Hendler, J., Holm, J., Musialek, C., Thomas, G.: US Government linked open data: semantic.data.gov. IEEE Intell. Syst. **27**(3), 25–31 (2012)
7. Harness Open Data with CKAN, OData and Windows Azure. http://msdn.microsoft.com/en-us/magazine/dn520247.aspx
8. CKAN. API Guide. http://docs.ckan.org/en/latest/api/index.html
9. Github and Government. https://government.github.com/community/
10. New Free Webinars @ AIMS on Linked Open Data. http://aims.fao.org/linked-open-data-webinars-at-aims
11. Sixth LOD@AIMS Webinar with Irina Radchenko on "Introduction to the Linked Open Data" (Russian). http://aims.fao.org/linked-open-data-webinars-at-aims/irina-radchenko/eng
12. Open Government Data for Citizen Engagement. http://www.unpan.org/ELearning/OnlineTrainingCentre/OpenGovernmentDataforCitizenEngagement/tabid/1751/language/en-US/Default.aspx
13. Open Data Government Toolkit. http://data.worldbank.org/open-government-data-toolkit
14. Russian Open Data School. http://opendataschool.ru/
15. The Code for America Brigade. http://brigade.codeforamerica.org/
16. Open Data day in Moscow. Open Knowledge Foundation Russia. http://ru.okfn.org/2014/02/24/odd14msk/
17. Code for America. http://codeforamerica.org/
18. Citizens for Open Access to Civic Information and Data. http://civicaccess.ca/
19. Map and GIS Services at the University of Toronto. http://mdl.library.utoronto.ca/map-gis-home
20. Quandl. http://www.quandl.com/
21. Raw. http://raw.densitydesign.org/
22. Acerbi, A., Lampos, V., Garnett, P., Bentley, R.A.: The expression of emotions in 20th century books. PLOS ONE **8**(3), e59030 (2013). http://www.plosone.org/article/info:doi/10.1371/journal.pone.0059030
23. Data journalism. http://en.wikipedia.org/wiki/Data_journalism
24. Data-driven journalism. http://en.wikipedia.org/wiki/Data-driven_journalism
25. Database journalism. http://en.wikipedia.org/wiki/Database_journalism

26. Computation + Journalism. A study of Computation and Journalism and how they impact each other. http://www.computation-and-journalism.com/main/
27. Analytic journalism. http://en.wikipedia.org/wiki/Analytic_journalism
28. Narrative Science. http://narrativescience.com/
29. Let Your Data Tell Its Story. http://automatedinsights.com/
30. Kathimerini. http://www.kathimerini.gr/infographics
31. The New York Times: The Year in Interactive Storytelling (2013). http://www.nytimes.com/newsgraphics/2013/12/30/year-in-interactive-storytelling/
32. The Guardian. DataBlog. http://www.theguardian.com/news/datablog
33. RIA Novosti. Infographics. http://en.ria.ru/infographics/
34. RIA Novosti. Infographics. Russian version. http://ria.ru/infografika/
35. Island of Lost Souls. http://lenta.ru/articles/2013/09/02/dushi/
36. School of Data. http://schoolofdata.org/
37. Peer 2 Peer University. https://p2pu.org/en/
38. Sakoyan, A., Radchenko, I.: Data Expeditions and Data Journalism project as OER in Russian. http://ukwebfocus.wordpress.com/2014/03/15/data-expeditions-and-data-journalism-project-as-oer-in-russian/