INDEPENDENCE THROUGH DATA JOURNALISM

Avner Kantor
Sheizaf Rafaeli
University of Haifa

Introduction

Data journalism (DJ) fosters audience independence. It encourages content exploration through visualizations, storytelling, and direct access to data sources (Amit-Danhi, 2020). DJ assists the audience to be well informed, cognitively active, and contribute to the public sphere (Baack, 2015). Achieving this objective indicates the audience's independence (Entman, 1993).

How independent is the audience of DJ? This question is answered based on the level of engagement of the audience. A low level of engagement prevents the audience from helping each other to interpret the dominant meaning and to identify its deficiencies. When the audience is engaged, they are able to reframe the message from a new perspective, challenging their understanding of the message. By analyzing the audience engagement levels and comparing different types of journalism we gain insight into audience behavior.

Methodology

The study relies on data from the Guardian, one of the oldest and prestigious newspapers in the world. Its articles are read by the elite from all over the world and affect the international agenda. The Guardian has published DJ online since 2009 (Barr et al., 2019). With the use of its API, we have attempted to collect all DJ articles and comments on the Guardian website. The collection of data is limited to the period 2012-2016 because this is a time when the Guardian enabled widespread commenting and replies. To eliminate the bias of news type, only articles classified as news by the Guardian were selected. The final dataset contained 142 (38%) DJ articles and 233 (62%) traditional journalism articles which lacked several of the characteristics of DJ, including visualization and direct access to data sources.

We assessed engagement by analyzing audience-generated content in the newspaper's comment sections. Engagement can take many forms, such as page views, reading time, clicking, liking, sharing, and posting. In a preliminary study, we found that articles with a high level of exposure tend to receive more comments and replies. In addition,

we found that the Guardian DJ is less exposed to the audience. In order to reduce exposure bias, we used the replies ratio as a measure of audience engagement (Ksiazek et al., 2016). The replies ratio is calculated by dividing the number of replies by the number of comments for every article.
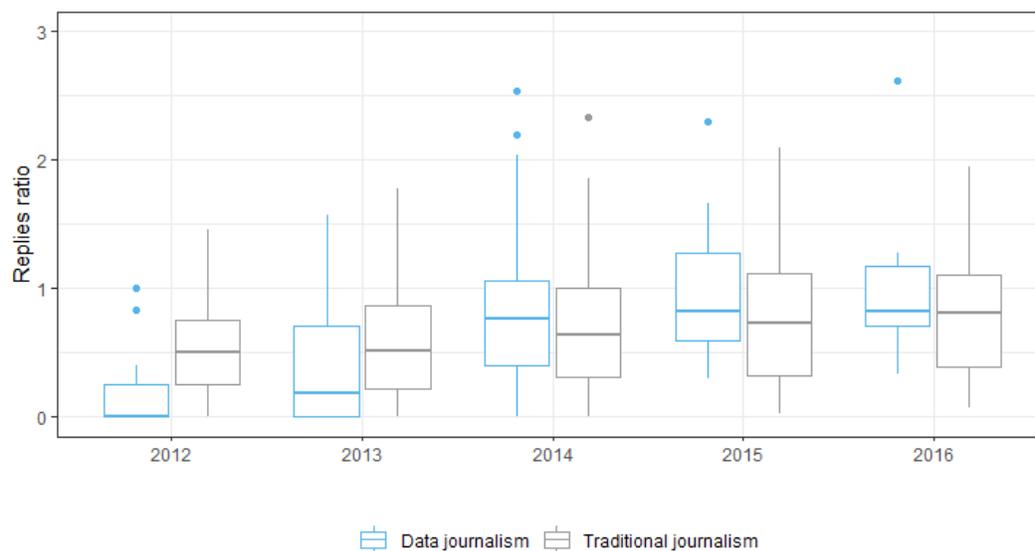
Results

Table 1 shows that DJ replies ratio increased by .92 during the study period, while the traditional journalism replies ratio increased only by .24. Figure 1 shows that between 2012 and 2013, the DJ replies ratio was lower than traditional journalism. According to a t-test, there was no significant difference between DJ and traditional journalism from 2014 to 2016.

Table 1

Descriptive Statistics of Articles Replies Ratio

Year	Data journalism					Traditional journalism				
	N	Min	Max	Mean	SD	N	Min	Max	Mean	SD
2012	17	.00	1.00	.17	.31	21	.00	1.46	.54	.41
2013	39	.00	1.57	.41	.50	73	.00	3.67	.63	.57
2014	62	.00	2.54	.79	.56	75	.00	2.33	.70	.48
2015	18	.30	2.30	.95	.52	44	.03	2.10	.76	.53
2016	6	.33	2.62	1.09	.81	20	.07	1.95	.78	.51

Figure 1

Distribution of Articles Replies Ratio by Journalism Type and Year

Discussion
There was a significant correlation between DJ replies ratio and time during the research period. The replies ratio increased yearly by .27 (R^2 is .2, p<.001). This trend may be explained by the steep learning curve of DJ, which requires a special set of skills, in addition to the willingness to analyze and discuss interpretations with others (Baack, 2018). On the basis of the trend, we can conclude that over time, the audience became familiar with DJ and adopted it. Alternatively, the trend may be attributed to changes in the Guardian's comment moderation policy (Gardiner et al., 2016). Over the years, The Guardian has blocked approximately 2% of its comments. Procedures and policies may change over time, which may influence the response rate.

DJ expects more from the audience than traditional journalists. Investigating data sources and evaluating visualizations requires an increased degree of mental and cognitive effort. Despite this, there seems to be no difference in audience engagement between 2014 and 2016. Based on these results, the difficulties that accompany independence do not reduce audience engagement. This phenomenon may be attributed to a number of factors, including the specific characteristics of the Guardian's audience that are willing to contribute more to the public sphere. By understanding the factors that drive audience engagement, technology developers may be able to implement them on other platforms and improve democratic processes on the web.

Further research is required to examine the independence of the audience. An examination of audience-generated content may reveal their interpretations and framing (Hullman et al., 2015). Interpretations can be assessed by characteristics such as the use of evidence, the formation of rational arguments, and the forming of opposing opinions to the dominant meaning (McInnis et al., 2020). In addition, an examination of different DJ newspapers and a variety of approaches will be able to clarify the factors affecting the audience (Appelgren, 2017). Results could contribute to the realization of the internet's original promise to make the general public independent by providing free access to information.

* Research code and data are available at https://github.com/avnerkantor/guardian.

Bibliography
Amit-Danhi, E. (2020). TMI: Information rhetoric types in digital political visualizations. *Paper Presented at AoIR 2020: The 21st Annual Conference of the Association of Internet Researchers. Virtual Event: AoIR.* https://doi.org/10.5210/spir.v2020i0.11156
Appelgren, E. (2017). An Illusion of Interactivity. *Journalism Practice*, *0*(0), 1–18. https://doi.org/10.1080/17512786.2017.1299032
Baack, S. (2015). Datafication and empowerment: How the open data movement re-articulates notions of democracy, participation, and journalism. *Big Data & Society*, *2*(2), 1–11. https://doi.org/10.1177/2053951715594634
Baack, S. (2018). Practically engaged. *Digital Journalism*, *6*(6), 673–692. https://doi.org/10.1080/21670811.2017.1375382
Barr, C., Chalabi, M., & Evershed, N. (2019, March 23). A decade of the Datablog: "There's a human story behind every data point." *The Guardian*. https://www.theguardian.com/membership/datablog/2019/mar/23/a-decade-of-the-

datablog-theres-a-human-story-behind-every-data-point

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, *43*(4), 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

Gardiner, B., Mansfield, M., Anderson, I., Ulmanu, M., Louter, D., & Holder, J. (2016). The dark side of Guardian comments. *The Guardian*. https://www.theguardian.com/technology/2016/apr/12/the-dark-side-of-guardian-comments

Hullman, J., Diakopoulos, N., Momeni, E., & Adar, E. (2015). Content, context, and critique: Commenting on a data visualization blog. *CSCW 2015 - Proceedings of the 2015 ACM International Conference on Computer-Supported Cooperative Work and Social Computing*, 1170–1175. https://doi.org/10.1145/2675133.2675207

Ksiazek, T. B., Peer, L., & Lessard, K. (2016). User engagement with online news: Conceptualizing interactivity and exploring the relationship between online news videos and user comments. *New Media & Society*, *18*(3), 502–520. https://doi.org/10.1177/1461444814545073

McInnis, B. J., Sun, L., Shin, J., & Dow, S. P. (2020). Rare, but Valuable: Understanding Data-centered Talk in News Website Comment Sections. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2). https://doi.org/10.1145/3415245