Data Collection, Analysis, and Small-Data Visualization: A Qualitative Interpretation of Novel Fandom Behaviors on Social Media Platforms

Accepted version: As part of

SAGE researchmethods cases

SAGE Research Methods: Doing Research Online

Recommended citation:

Chen, Z. T., (2022). Data collection, analysis, and small-data visualization: a qualitative interpretation of novel fandom behaviors on social media platforms. In Sage Research Methods: Doing Research Online. SAGE Publications, Ltd., https://doi.org/10.4135/9781529602098

Abstract:

This case study focuses on the data visualization method in the context of online fandom activities and the digital platform economy. In media and communication studies, fandom and celebrity studies go hand in hand which not only powers up the economy via its entertainment and news industries but also keeps audiences informed, entertained, and, sometimes, educated. In order to better understand novel fandom activities afforded by new digital technologies, my role in this project was to use the data visualization method and conduct data collection of fan-created texts alongside the digital contents they consumed or the celebrities they followed. This case study discusses the process of data collection and visualization of a small body of fan-created media texts, the underlying design thinking, and the challenges I experienced regarding the amount of data, data cleaning of idiosyncratic and vernacular fandom expressions, as well as the new modality afforded by new digital technologies. This case study also critically reflects on how "small data" can be used creatively under the dominating "big data" discourse in academia and how data visualization can be used to inform and compensate qualitative interpretations and ethnographies, which are widely used in fandom and audience studies.

Learning Outcomes

By the end of this case, you should:

- Understand how affordable small data can be collected and used for fandom and media analysis
- Learn to use open-access software Gephi to visualize low-cost and readily available "small data" for research projects
- Understand the importance of data collection prior to data visualization and how it engages with other complementary methods

Project Overview and Context

As part of a larger project approved by and conducted in two Sino-British collaborative universities, this data collection process focuses on a particular video-sharing social media platform Bilibili and its fans' transnational prosumption behavior in China. Prosumption, here is a combination between production and consumption which emphasizes the agency of Chinese young prosumers who are previously known as passive audiences. The findings were to contribute to the analysis of the phenomenon and strategy termed "value co-creation" where consumers are encouraged, educated, and socialized to produce para-texts around a particular media text, product, or celebrity in an online platform economy.

This case is a first step toward the understanding of what fans do on a daily basis with or without corporate or industry influence, using data visualization tools and methods. After fan-created texts are collected, sorted, and analyzed, scholars and businesses alike can make full use of such data critically and/or analytically.

Using this method, projects as such can produce stand-alone research articles or can be expanded to work collaboratively with people using other methods to generate further insights.

This project focuses on a particular technology, danmu (in Chinese) or danmaku (in Japanese), literally bullet curtain or screen overlaid on the video being watched, where users and fans could "shoot" bullet and subtitle- like texts and comments, flying quickly across the screen via a chatbox. It is a technology developed and adopted by a number of video streaming sites, such as Nico Nico, AcFun, Bilibili, and many more in East Asia (Chen, 2020, Chen, 2021). The danmaku technology not only creates a shared watching experience but also provides an additional modality on top of the original media content that is being consumed. This function can be opted out by users who are annoyed by it. However, most users would keep it on when watching these videos alone, and therefore, it has gained popularity among tech-savvy geeks and Otaku users in the animation, comic, and game (ACG) subculture.

The technology of danmaku also boosts these streaming sites and makes them the hotbed for subculture in ACG and spoof videos. Some of the witty-naughty jokes and gags supported by danmaku went viral and became Internet memes and slangs widely used by Chinese youth. Therefore, through studying danmaku interactions, it becomes a gateway to understand Chinese youth culture and the Chinese media market. Chinese content creators and aggregators such as Baidu, Alibaba, and Tencent and commercial cinemas adopted the danmaku technology for their own websites and offline film-going experience.

In a western context, even though danmaku has not been widely used, there are players available (e.g., DPlayer, see more in Dwyer, 2017), which supports such a co-watching experience where comments and subtitles can be collected and analyzed. This method can be applied to studies of websites and platforms that support a comment and review function, such as e-commerce sites, recommendation applications, among other social media platforms.

Section Summary

- Users of social media platforms foster vernacular cultural and fandom groups online, which demonstrates agentic prosumption behaviors worthy of exploration.
- Data visualization and the qualitative interpretation of such data produced by fans can help better understand the meaning-making processes of new media fandom.
- The case of fan-generated para-texts or derivative texts, that is, danmaku provides with researchers a new source and modality subject to scholarly investigation.

Research Design

This section explains how the research case was initially designed to answer its designated research questions. These include what types of texts are being created by fans using this new technology, what it enables and disables, and how to interpret the meaning of such texts. In the following sections, I will explain how particular decisions regarding the research project were made based on a balance between the research questions raised and the options and tools available.

Pilot Studies

Watching and commenting on videos together with users and fans online can be an emotional experience. Observing and reflecting on such experiences is a discovery process. Existing research has well documented how media analysis can be applied for different purposes regarding media consumption and reception studies. Given the novelty of the technology, namely, danmaku, we need to regard it as an additional source of data collection for media analysis. This is because this new medium affects the way in which users and fans communicate with each other. Using this very technology, fans' communication can be stored, extracted, and categorized for further analysis. Given the primary aim of this project is to understand users' and fans' agency—not who they are, but what they can do/say—the data collection process is particularly open-ended and explorative.

Case Selection

At the outset, I had to decide the sample size of the danmaku texts that I was going to collect. However, each and every video on Bilibili can store up to 3000 danmaku comments which make the whole website a potential corpus, a database ready for linguistic analysis. Informed by existing and prior research (Chen, 2020, Chen, 2021) and my initial observation of the website design, it became obvious that users and fans diverge and converge among different interest groups. On the streaming site of Bilibili, such groups are organized in the form of genres, channels, and labels. That is, fans create and recreate digital contents and manually label them and upload them to designated channels for like-minded people to consume, share, comment, and re- produce derivative works.

Small Data Versus Big Data

As discussed previously in the introduction, in media and communication studies, big data analysis has become trendy in recent years, especially for studies using quantitative methods, not to mention in disciplines such as behavioral economics and information systems (Jiang et al., 2016). However, ongoing debates continue on "big data" versus "small data." Numerous authors have argued that small data would continue to inform research in the sciences, social sciences, and humanities because of their utility in answering targeted enquires (Faraway & Augustin, 2018; Xu et al., 2020). In addition, epistemologically big data is made up of small data, where researchers have to strike a balance between benefits and costs associated with acquisition, computation, and privacy (Faraway & Augustin, 2018; Floridi, 2012).

As critical researchers, we always need to remind ourselves what specific data to identify and collect to produce situated knowledge, to question the unquestioned norms, especially in the context of social media platforms. This is because the unquestioned norms such as how a tech giant collects, manages, and sells its data can have a paramount impact on the well-being and data security of social media users and consumers, which may create and recreate inequalities along the lines of class, gender, race, and sexuality (Jarrett, 2017). This is especially the case when such tech giants are moving their services from the public to the private spheres of peoples' lives, including journalism, ecommerce, food delivery, entertainment, and dating, to name but a few. Therefore, by collecting and analyzing small data created by consumers of vernacular cultures and fandom groups, it helps democratize the diversified subjectivities, experiences, and effects afforded by such new digital technologies (Welles, 2014).

For ease of analysis, I decided to start with genres I am more familiar with where relevant works were highly ranked and recommended on Bilibili and other similar platforms that host such subcultural texts. In my first article on alternative space building and exploration by Chinese Otaku fans (Chen, 2021), I used the animated series Yuri On Ice!!! as a case study, a slash/homoerotic genre, as Bilibili is dubbed by its users as the largest homosocial dating platform in a comic way. In another article on verbal identity performance and work (Chen, 2020), I focused on a feel-good and harem genre, where one male protagonist develops romantic encounters with multiple females and/or male characters. In particular, I used an animated series, a comic adaptation, entitled My Youth Love Story is Problematic, as the primary example. These two shows are well-known in both the Japanese and Chinese Otaku fandom. It would be interesting to see how Chinese fans construct their own identities with their unique fannish activities in a transnational context.

Section Summary

- Danmaku as a kind of computer-mediated communication can be stored, extracted, and categorized for further media analysis. Given its novelty, this study adopts an explorative approach which begins with a pilot study.
- Case selection is based on initial participant observation where fans diverge along with their interest groups and coverage within their vernacular cultures marked by channels, labels, and genres.

- Justify your case selection and sample size with the appropriate data size. This is not a test about the number but a qualitative assessment where you balance your choice between big and small data. Studies about subcultural groups can (and should) rely on small but meaningful data.

Research Practicalities

Once the genres and shows were decided, I then collected the danmaku comments from each episode of the animated series for later data cleaning and initial analysis. Normally, social media platforms and video-sharing and streaming sites have open application programming interfaces (APIs) for data scraping and collection. These include Twitter, Weibo, YY, AcFun, Bilibili, to name but a few. However, most platforms have changed their business strategies for commercial purposes or privacy compliance reasons, which makes collecting such publicly available data difficult. Given the limited option that only Bilibili maintains an open API, I chose Bilibili for data collection. Danmaku can be scrapped using web browsers such as Google Chrome. You can also use third-party websites or add ons to extract danmaku comments and save them in one go in an XML format, extensible markup language, a computer language used in text formatting. For other ecommerce and social media platforms, including Amazon, e-Bay, and Tmall (owned by Alibaba), research teams could build your own workstation for data crawling and scraping processes. A reminder on research ethics would be that data obtained in this manner may contain user information and should go through a de-identification process to anonymize the real identity of the users.

Time and Cost

If you have research funding, you may want to hire professionals and research assistants to help you do data scraping and data cleaning, to make the data ready for analysis. However, so far Bilibili's open API allows data scraping in one go which is easy to follow.

Tools and Steps

If you are a Mac user, simply visit a Bilibili video using Safari and press F12 to switch to the developer mode. Press Option+Command+I to enter the console and search "cid" and press enter to locate the ID. For PC and Linux users, you could use a number of third-party websites and simply input the URL of the video and download the XML file containing all danmaku comments.

Other sources to consider:

- https://github.com/Privoce/all-in-danmaku-extension/
- http://www.ibilibili.com/bilitools/danmu/?u=https://www.bilibili.com/video/av1786711

N.B. Please be advised that these sites are from third-party providers and may not be accessible or useful subject to Bilibili's API policy.

Section Summary

- Data collection on social media platforms can be challenging given constraints on ownership, copyright, and ethical issues. Evaluate your choice based on time, cost, and institutional requirements.
- Data collection from different browsers and operating systems may vary. Go for practical options that support an open API with greater compatibility and sustainability.

Method in Action

Once the XML files containing the danmaku are collected, open them in excels, and separate the text data into different columns in a spreadsheet. The reason why I chose the excel spreadsheet was because it is more widely accessible than Tableau, for example. The challenge for researchers and students is that the danmaku comments collected are very short and may only contain some Chinese or kanji characters. Danmaku comments are full of vernacular terms used by subcultural groups which contain foreign languages, emojis, and typos. Sometimes, typos are deliberately used to bypass certain keyword-based censorship screenings.

At the outset, the cleaning and rearranging of the data is key. This includes "combining like terms" as in mathematics, which is especially useful for Internet slangs and conventional memes, such as, high energy alert (indicating a climax or a terrifying scene as a spoiler), 2333 loops (laughing with hands banging on the door), Orz (sigh and kneeling down), just to name but a few.

The next step, which is also the key in mapping such danmaku comments with the media content being consumed and prosumed, aka, recreating through providing commentary, review, and derivative works.

Design Thinking for Mapping and Visualization

This mapping and visualizing process is the backbone of the design thinking underlining this research project. As a previous web designer, who has worked with different assembly languages and content management systems, among other tools, I become to understand one thing that such websites and their designers want to achieve is "what you see is what you get" (WYSIWYG). In practical terms, this means we want to label different items we put in specific sections within a certain webpage and to give corresponding instructions using these computer languages so that search engines can discover them easily. This is also called design thinking for search engine optimization (SEO). For example, if you include a keyword or a phrase that summarizes the whole section of your webpage and want it to be put at a salient position, you apply the Title or Heading 1 instruction to it. In doing so, the term will be first picked up by search engines such as Google or Bing before titles that are applied with Heading 2.

In the case of danmaku comments, such importance or priority is based on the density of the danmaku, that is, the frequency of the actual comments posted through repetition. Such density is contributed by the following measurement,

including who posted it (users ID), when (a timestamp), and how many times by how many users. These items are listed in the XML you collected and can be further organized and categorized based on their density.

Tools and Layouts

For data visualization purposes, I used the free software Gephi and applied Yifan Hu Proportional layout as a dropdown option, which provides improved speed and efficiency compared to other layouts (Hu, 2005). This layout allows you to produce a graph that shows the general pattern of all the danmaku at hand (Figure 1). In addition, it shows the priority of the repetitive themes you gathered and prepared via data cleaning in the previous step.

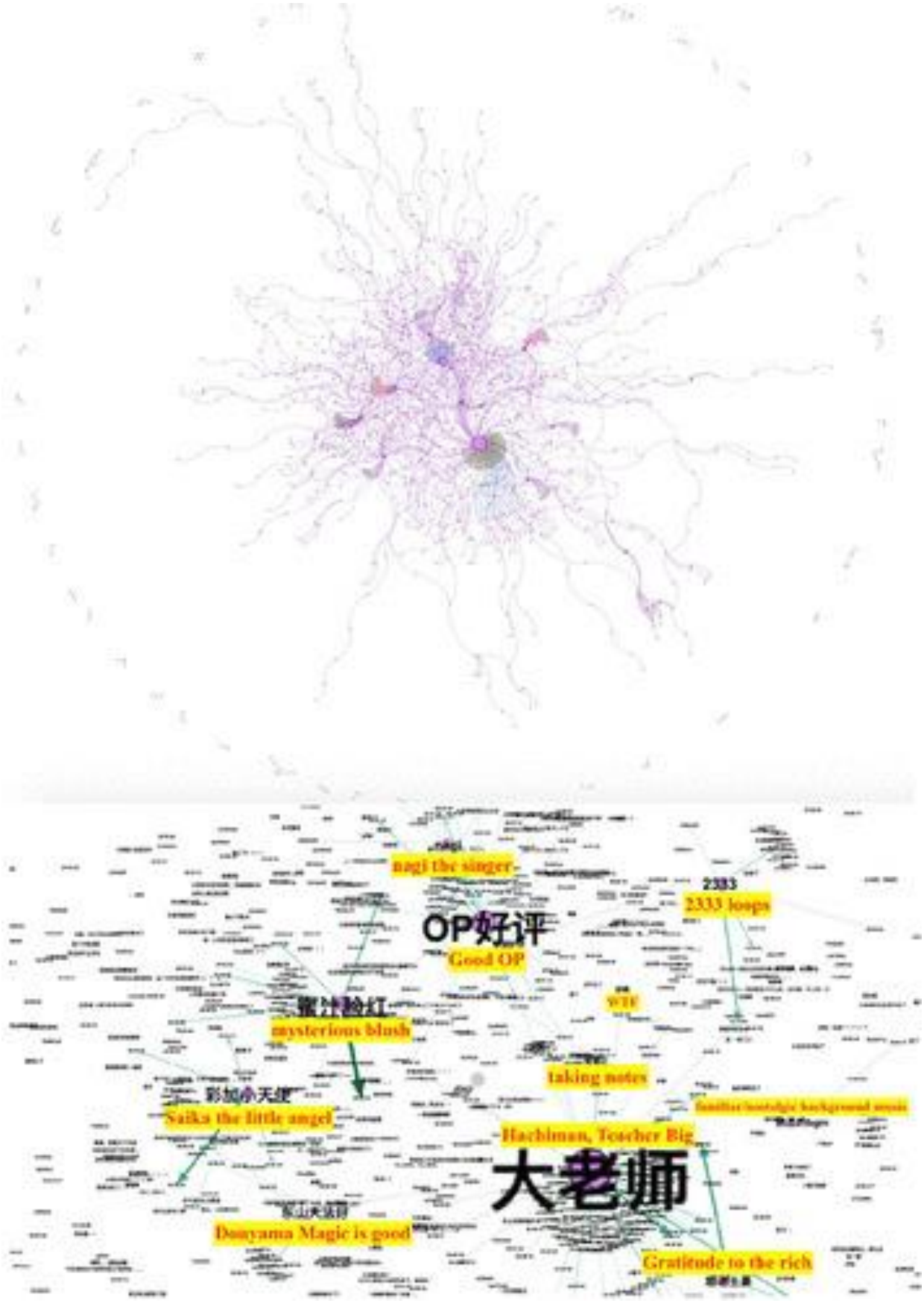

Figure 1. The visual semantic web based on danmaku in Episode 1.

Section Summary

- After data collection, it is important to do a pilot data cleaning exercise to understand your data better. Insight from participant observation would be helpful at this stage as it helps rearrange some of the similar data sets and generate patterns.
- If the initial data collected still appears messy, try to adopt a design thinking approach where you think like a designer who anticipates and guides a user on how to engage with a particular platform and how to create user-generated data.
- After cleaning the data, you need to choose a suitable visualization tool to present your data. Gephi is open-access and powerful visualization tool widely used in published scientific papers. This means you can refer to relevant and useful studies following similar approaches.

Practical Lessons Learned

This explorative case study is a process of discovery and refinement. It started by observing diversified fandom activities on social media platforms to understand the reason and the way in which fans create a plethora of behavioral and media texts. At times, it is frustrating to test various tools, frameworks, and approaches for multiple times to make meaningful interpretations and analyses. There are a number of challenges and lessons learned in this repetitive yet rewarding process:

1. Run a pilot study of your chosen genre, celebrity, game among various media texts to get an estimated forecast of your data size, time, and cost. If you are working within a team for coursework or dissertation, you can choose to collaborate with teammates to narrow down your focus and edit down your sample size. This may affect your final research findings in terms of scale and generalizability. However, it is useful to take into consideration of your research timeline and progression.

2. After case selection, make sure your data cleaning is done consistently across different data sets from different episodes, be it drama, animation, live streaming, or e-commerce platforms. This will make your data analysis efficient and effective.

3. Chose convenient and affordable tools, especially open-sourced ones where support and QA community has been well developed. Make sure the tools you selected excel at processing the source texts you collected, be it Chinese, Japanese, or English. I was using a data visualization tool that needs a subscription. It makes my research difficult when moving from one institution to another.

4. Be explorative and creative. Conventional textual and/or thematic analysis is useful in helping you get pointers in analyzing your collected data. Use visualization tools creatively to present a large amount of data in novel ways, for example, to not only identify repetitive patterns but also identify abnormal patterns which are also worth scholarly attention.

Section Summary

- The choice between small and big data will always be a difficult one. You have to stop collecting more data at some stage. This is especially when you are a student doing a research project for a term paper. Justify your focus and make it match with your sample size.
- Choose convenient and affordable tools. Open-source tools are normally supported by a dedicated community where you can learn from others' experiences and failures.
- Data visualization backed up by qualitative interpretations focuses on both repetitive patterns and abnormalities in your data.

Conclusion

Given the rich texture of fandom activities and the complexity of social media platforms that power up and host such computer-mediated communication texts, using data visualization alone can be somewhat limited. This is especially the case when you aim to do an analysis of the agency of fans in an online platform economy. The rationale for chasing the trend of "bigger" data, and the emphasis on language use (and its associated textual analysis method) is understandable, but in isolation, data visualization or textual interpretations can tell us only one side of the story.

Having a set of good research questions is crucial in entering a field of qualitative research that involves human beings and their behaviors. A solid literature review and some initial observations are important in getting one foot in your chosen field. Once your direction and focus of your study are narrowed down, make sure you get access to research tools that can help you collect information from different open source or paid sources. Even though your stand-alone case study may only apply to one particular genre, show, or platform, try to draw connections through design thinking between these platforms. That is, what are the similarities (see SEO, discussed above) and what are the differences between these platforms. Such holistic and comparative exercises will be of great benefit for your research. That is often where creativity and originality come from.

The data visualization method discussed here can be a good contribution to a mixed-method study, adding valuable insights and/or working as triangulation. It can also provide a bigger picture of the data collected when the project is still taking shape. Findings from such human behavior-based research also have the potentials to forecast changes in market demand and consumption behavior. It also can go beyond the textual level of the data since it adds onto what conventional methods would normally contribute, yet enquires further in terms of the modality, architecture, and design mechanism underlining such prosumption arrangements and practices.

Section Summary

- Data visualization and digital ethnography are not competing with each other in analyzing fandom and user behaviors. They can make each other stronger and can also inform and forecast market research.

- Danmaku users and communities are just one of many techno-cultures that are fostered and shaped by a particular technology. What is important is the design thinking behind that can help students work on user-generated content on other social media platforms.
- Interdisciplinary research is not only something desirable but necessary in the future to understand complex modalities and behaviors that are computer-mediated.

Classroom Discussion Questions

1. Why is it important to observe and understand the subcultural group before data collection and data visualization?

2. Why is it important to use small data for vernacular fandom/culture studies?

3. Why is it important to conduct a pilot study before collecting the targeted data set?

Further Reading

Cruz, A. G. B., Seo, Y., & Binay, I. (2021). Cultural globalization from the periphery: Translation practices of English-speaking K-pop fans. *Journal of Consumer Culture*, 21(3), 638–659. 10.1177/1469540519846215

Zhang, L.-T., & Cassany, D. (2020). Making sense of danmu: Coherence in massive anonymous chats on Bilibili.com. *Discourse Studies*, 22(4), 483–502. 10.1177/1461445620940051

Zhang, L., & Cassany, D. (2021). 'The murderer is him ✓': Multimodal humor in danmu video comments. *Internet Pragmatics*, 4(2), 272–294. 10.1075/ip.00038.zha

Web Resources

AoIR—The Association of Internet Researchers publishes ethical guidelines to assist researchers in making ethical decisions in their research: https://aoir.org/ethics/

The QUT Digital Media Research Centre (DMRC) is a global leader in digital humanities and social science research with a focus on communication, media, and the law. https://research.qut.edu.au/dmrc/the-centre/

MediaLAB Amsterdam: https://medialabamsterdam.com/toolkit/

The Oxford Internet Institute is a multidisciplinary research and teaching department of the University of Oxford, dedicated to the social science of the Internet. https://www.oii.ox.ac.uk/about/

References

Chen, Z. T. (2020). Slice of life in a live and wired masquerade: Playful prosumption as identity work and performance in an identity college Bilibili. *Global Media and China*, 5(3), 319–337. 10.1177/ 2059436420952026

Chen, Z. T. (2021). Poetic prosumption of animation, comic, game and novel in a post-socialist China: A case of a popular video-sharing social media Bilibili as heterotopia. *Journal of Consumer Culture*, 21(2), 257–277. 10.1177/1469540518787574

Dwyer, T. (2017). Hecklevision, barrage cinema and bullet screens: An intercultural analysis. Participations: *Journal of Audience & Reception Studies*, 14(2), 571–589.

Faraway, J. J., & Augustin, N. H. (2018). When small data beats big data. *Statistics & Probability Letters,* 136(487), 142–145. 10.1016/j.spl.2018.02.031

Floridi, L. (2012). Big data and their epistemological challenge. *Philosophy & Technology*, 25(4), 435–437. 10.1007/s13347-012-0093-4

Hu, Y. F. (2005). Efficient and high quality force-directed graph drawing. *The Mathematica Journal*, 10, 37–71.

Jarrett, K. (2017). *Feminism, labour and digital media: The digital housewife*. Routledge.

Jiang, Z. J., Wang, W., Tan, B. C. Y., & Yu, J. (2016). The determinants and impacts of aesthetics in users' first interaction with websites. *Journal of Management Information Systems*, 33(1), 229–259. 10.1080/ 07421222.2016.1172443

Kitchin, R., & Lauriault, T. P. (2015). Small data in the era of big data. *GeoJournal*, 80(4), 463–475. 10.1007/s10708-014-9601-7

Welles, B. F. (2014). On minorities and outliers: The case for making big data small. *Big Data & Society*. 10.1177/2053951714540613

Xu, F., Nash, N., & Whitmarsh, L. (2020). Big data or small data? A methodological review of sustainable tourism. *Journal of Sustainable Tourism*, 28(2), 144–163. 10.1080/09669582.2019.1631318