Gratis, Libre, or Something Else? Regulations and Misassumptions Related to Working with Publicly Available Text Data

Jana Diesner, Chieh-Li Chin
The iSchool/ Graduate School of Library and information Science (GSLIS)
University of Illinois Urbana Champaign
501 E Daniel Street
Champaign, IL, 61820, USA
E-mail: jdiesner@illinois.edu, cchin6@illinois.edu

Abstract

Raw, marked up, and annotated language resources have enabled significant progress with science and applications. Continuing to innovate requires access to user generated and professionally produced, publicly available content, such as data from online production communities, social networking platforms, customer review sites, discussion forums, and expert blogs. However, researchers do not always have a comprehensive or correct understanding of what types of online data are permitted to be collected and used in what ways. This paper aims to clarify this point. The way in which a dataset is "open" is not defined by its accessibility, but by its copyright agreement, license, and possibly other regulations. In other words, the fact that a dataset is visible free of charge and without logging in to a service does not necessarily mean that the data can also be collected, analyzed, modified, or redistributed. The open software movement had introduced the distinction between free as in "free speech" (freedom from restriction, "libre") versus free as in "free beer" (freedom from cost, "gratis"). A possible risk or misassumption related to working with publicly available text data is to mistake gratis data for libre when some online content is really just free to look at. We summarize approaches to responsible and rule-compliant research with respect to "open data".

Keywords: open source data, user and professionally generated online content, gratis versus libre text data, ethics, data repositories

1. Introduction and Problem Statement

Raw, marked up, and annotated text corpora available to the research communities in Natural Language Processing (NLP), Computational Linguistics (CL), the digital humanities, and computational social science have enabled major progress and breakthroughs in these and other areas. Continuing to innovate requires access to contemporary text data that were generated by people using common information and communication technologies (ICT), such as data from online production communities (e.g., Wikipedia and GitHub), social networking platforms, customer review sites, discussion forums, and expert blogs. One problem with work in this area is that researchers do not always have a comprehensive or correct understanding of what types of user or professionally created web content are permitted to be collected and used in what ways (Kosinski, Matz, Gosling, Popov, & Stillwell, 2015; Vitak, Shilton, & Ashktorab, 2016; Zevenbergen et al., 2015; Zimmer, 2010). This paper aims to clarify this point. We focus on risks for researchers who gather and utilize content from publicly available sites rather than on privacy risks for people who make their information available online.

1.1 Benefits of Working with Publicly Available Text Data

On the beneficial side, working with data at any scale that were generated by people who use ICTs and who interact with others and with information within these infrastructures allows for considering both the content and structure of social interactions (Lazer et al., 2009) and for re-evaluating theories that are based on data generated in offline or non-ICT-facilitated environments (Diesner, 2015; Kleinberg, 2008). Research based on contemporary interaction and text has promoted the emergence and advancement of the fields of network science, web science and internet science (Tiropanis, Hall, Crowcroft, Contractor, & Tassiulas, 2015).

Recognizing these benefits, some members of the scholarly community and their funders have been advocating for open access to data, code, knowledge and publications (Hodgson et al., 2014). Corresponding legal and technical solutions have been developed. Examples include copyright licenses by the Creative Commons[1] and open source licenses for software (for an overview see Opensource.org), as well as repositories that enable reliable and persistent access to publications, e.g., PubMed[2] for biomedical literature, as well as to domain specific and general science data (for an overview see "Recommended Data Repositories," 2016).

1.2 Risks of Working with Publicly Available Text Data

On the controversial side, scholars and practitioners might have an unclear or incomplete understanding and different conceptualizations of what "open source data" means and what this meaning implies for their practical, day-to-day work (Diesner & Chin, 2016; Vitak et al., 2016; Zevenbergen et al., 2015). Reasons for this effect include changing norms and regulations over time, and insufficient training on this topic.

Ethicists and privacy scholars have long argued that

[1] https://creativecommons.org
[2] http://www.ncbi.nlm.nih.gov/pubmed

working with user created, publicly available data can involve privacy risks for the individuals who generated and publish these data (Daries et al., 2014; Hoffman & Bruening, 2015; Lane, Stodden, Bender, & Nissenbaum, 2014). Several check points and risk mitigation mechanisms have been put in place, such as updates to Institutional Review Board (IRBs) processes. However, data from online sources might not be subject to review by an IRB if the researchers did not interact with the subjects and the data were already publicly available. Furthermore, collecting and using data from online sources may conflict with other types of regulations, including copyright, terms of service, established cultures in research communities, and personal values (Diesner & Chin, 2015; Kosinski et al., 2015; Zevenbergen et al., 2015). Deviating from these norms and rules may entail risks for researchers, their institutions and scientific communities, and the reputation of science (Zimmer, 2010).

In the remainder of this paper, we first briefly review classic types of sources for text corpora and related regulations. We then clarify what "open source data" means in theoretical and practical terms, and discuss potential reasons for confusion. Finally, we outline possible approaches to the responsible and ethical conduct of research that involves publicly available text data.

2. Background: Sources and Related Regulations for Working with Text Corpora

Some of the resources that have been widely used in the NLP and CL communities were prepared for and released as part of competitions and associated professional meetings, such as the "Text Retrieval Conference" (TREC)[3], "Automated Content Extraction" (ACE)[4], and the "Message Understanding Conference" (MUC)[5]. These data and related evaluation metrics have been serving as acknowledged standards and benchmarks for developing and assessing new computational solutions. Much of this work has been initiated and supported by US-based, federal funding agencies, such as the National Institute of Standards and Technology (NIST). Some of these data are now administered, maintained and distributed by the Linguistics Data Consortium (LDC)[6].

Furthermore, long-standing academe-based initiatives and collaborations have resulted in curated repositories, codebooks, lexicons, and annotations for domain-specific text coding purposes, such as the Human Relations Area Files (HRAF)[7] for the field of cultural anthropology, or the former Kansas Event Data System (KEDS)[8] for political science (Gerner, Schrodt, Francisco, & Weddle, 1994; Schrodt, Yilmaz, Gerner, & Hermreck, 2008).

More recently, private-public partnerships have resulted in the release of large scale archives of digitized text data, such as the HathiTrust[9] (Christenson, 2011; Wilkin, 2009). Some of these data are annotated for various types of textual features, e.g., entities and relations in the "Global Database of Events, Language, and Tone" (GDELT)[10] (Leetaru & Schrodt, 2013).

Most of the mentioned as well as other data sources that are commonly used for NLP and CL purposes include copyright statements, license agreements, or terms of service statements that determine how the data can or must be obtained, managed and used. However, for the wide range of human generated, publicly available content in the form of unstructured (e.g., blog entries) and semi-structured (e.g., Wikipedia articles) text data as well as mixed data (e.g., text and images) that are not behind a pay wall or a login wall, researchers might have a less clearly defined understanding of ethical and rule-compliant practices for data acquisition and utilization.

3. Regulations for Working with Publicly Available Text Data

For the purpose of this paper, "publicly available" means that the data are not behind a pay wall or a login wall, and can be accessed by anybody with a web-enabled device. Furthermore, we divide "publicly available text data" (short PATD) into two groups. First, data provided by ordinary users who utilize ICTs to generate, post or publish information ("user generated web content"), which includes a wide range of social media data. Second, data generated by companies and professional or paid staff, such as online newspaper articles ("professionally produced web content").

In reality, things can be more complex: Some webpages provide both types of information, e.g., Amazon features product descriptions from commercial providers and user reviews of these products, and newspaper websites provide articles written by journalists which users can comment on. Other webpages display snippets of content that originates from other sites and providers; sometimes justifying this practice with the fair use portion of the copyright law.

The ways in which one can engage with either type of PATD are governed by multiple sets regulations, including (1) personal values and ethics, (2) norms and rules that may differ by institution, sector and country (e.g., IRBs or the "Health Insurance Portability and Accountability Act" (HIPAA), (3) copyright law (including fair use), (4) privacy regulations, (5) security regulations, (6) terms of service, and (7) technical solutions (for a brief overview see Diesner & Chin, 2016).

Understanding and implementing these rules can be complicated. Educating instructors and students on these topics may lag behind technical feasibility and reality. Some regulations keep emerging and are later adjusted;

[3] http://trec.nist.gov/

[4] http://www.itl.nist.gov/iad/mig/tests/ace/

[5] http://www-nlpir.nist.gov/related_projects/muc/index.html

[6] https://www.ldc.upenn.edu/

[7] http://hraf.yale.edu/

[8] http://www.aaas.org/page/kansas-event-data-system-keds-project

[9] https://www.hathitrust.org/

[10] http://www.gdeltproject.org/

making them moving targets. Some rules are explicit, while others are more tacit, such as personal values and expected culture in scientific communities. Also, some explicit rules, such as terms of service, might be difficult to translate into practical solutions. The resulting lack of clarity as well as instances of research that received controversial reactions (Kramer, Guillory, & Hancock, 2014; Zimmer, 2010) have stirred debates about responsible and ethical ways for collecting and using PATD (Vitak et al., 2016).

3.1 What are "Open Source Data"?

The way in which a dataset is "open" is not defined by its accessibility, but by its copyright agreement, license, and possibly other regulations. In other words, the fact that a dataset is visible free of charge and without logging in to a service does not necessarily mean that the data can also be collected, analyzed, modified, or redistributed (Zevenbergen et al., 2015; Zimmer, 2010).

The open software movement has introduced the distinction between free as in "free speech" (freedom to use, modify and redistribute information with little restriction, "libre") versus free as in "free beer" (i.e. freedom from cost, "gratis") (Lessig, 2004; Stallman, 2002). The risk with PATD is that gratis might be mistaken for libre when the data really just are gratis (to look at). This misassumption may due to a variety of reasons, such as insufficient expertise, evolving norms, or prior work (performed under different regulations) that has set an example.

That being said, some PATD truly are in the public domain (libre) because they have an open source license. For example, articles, talk pages, and structured meta-data from Wikipedia [11] are released under the Creative Commons Attribution-ShareAlike License[12], which allows people to copy, distribute, adapt and transmit the work as long as they attribute the work and publish any derivations under the same, similar or a compatible license. Another example is WordNet (Fellbaum, 1998), a widely used lexical database of terms and their relationships, which is provided under its own open source license[13]. Also, some text data provided by several US-based federal agencies are in the public domain as the content "was prepared by employees of the United States Government as part of their official duties and, therefore, is not subject to copyright"[14]. An example are transcripts of congressional hearings, which are available through the website of the General Publishing Office (GPO)[15].

However, a wide range of social media data (user generated web content), including posts on many product and film review sites, as well as regular media data (professionally produced web content), including the online presence of classic print media, are gratis for personal use but not libre. In either case, the terms of use for these data are typically defined by the owner of the website. Users who provide content on these sites agree to these terms as part of the process of releasing their work on them. In fact, much of the publicly available online content, especially (social) media data, are protected by terms of service. These terms are often presented as browse-wrap agreements at the bottom of a webpage. Via these agreements, content providers often grant webpage visitors the right to access and making personal, non-commercial use of the data. Overall, rules for interacting with online content can make their permitted use comparable to reading notes on a traditional bulletin board or looking through a store window (gratis).

4. Approaches to Responsible Research with Publicly Available Text Data

Rule-compliant research can be achieved in several ways. First, considering applicable agreements requires awareness and acknowledgement of their existence, and an understanding of their actionable meaning. This applies to both terms of service and other regulations that may apply, such as the "Fair Information Practice Principles" (FIPPs) [16] or the "Health Insurance Portability and Accountability Act" (HIPAA)[17]. Mastering this step is mainly a matter to education and experience.

Second, some data providers offer technical solutions that explicate or implement the sites' data access and sharing, e.g., mainly robot.txt files and APIs. Considering such technical solutions requires a certain level of proficiency.

Third, researchers can contact data providers to obtain permission for data gathering and use under certain conditions. This solution is limited in its scalability as it involves a certain amount of administrative overhead for both sides.

Fourth, while user generated content is still a fairly recent phenomenon and data source, and related policies and regulations are still being developed, some companies have emerged that act as brokers of data between (corporate) content providers and end users, e.g., Crimson Hexagon[18] and BrandWatch[19]. In exchange for a fee, such services typically offer their customers increased data access (fire hose) over public APIs (garden hose) as well as data analytics computed over the raw material. The revenue from these for-pay models is typically not directly shared with users who generated the content, but might be invested in sustaining and improving platforms, services, features, and user experience, for example.

Fifth, we suggest that a novel and alternative solution

[11] https://www.wikipedia.org/

[12] https://en.wikipedia.org/wiki/Wikipedia:Text_of_Creative_Commons_Attribution-ShareAlike_3.0_Unported_License

[13] https://wordnet.princeton.edu/wordnet/license/

[14] http://www.ntsb.gov/about/Policies/Pages/Policies.aspx

[15] https://www.gpo.gov/fdsys/browse/collection.action?collectionCode=CHRG

[16] http://www.nist.gov/nstic/NSTIC-FIPPs.pdf

[17] http://www.hhs.gov/hipaa/index.html

[18] http://www.crimsonhexagon.com/

[19] https://www.brandwatch.com/

would be to enable content generating users to opt in to having their data being freely (libre) used under certain conditions, e.g. demanding that de-identification is performed. This opt-in choice could be provided as part of the process of posting content online.

4.1 Consequences of Using Gratis but not Libre Text Data on Reproducibility

Finally, once a researcher has obtained user or professionally created text data from an online source, another issue with these data may arise. Research should be reproducible, which has already become increasingly challenging with dynamic data and tools (Stodden, Leisch, & Peng, 2014). Federal funders encourage the free (libre) sharing of data and code to enable the reproducibility of work and maximizing the benefits of investing tax payers' dollars. Multiple funding agencies have started to require data management plans as part of proposals submissions. In these plans, researchers are asked - among other criteria - to specify how they intend to provide the outcomes of their work after project completion. Analogously, university libraries, among other stakeholders, have started to create, curate and administer data repositories where researchers can upload and search for data. However, if the data are proprietary or protected in other ways, for example by copyright or terms of service, making them available might not be an option for researchers. For example, some social media data can be obtained in a permitted and lawful manner, such as tweets via the Twitter API[20] or information from certain Facebook pages through their API[21] (both services have increasingly reduced the data that ordinary people can obtain through the APIs, e.g., Twitter in terms of the time window into the past, and Facebook with respect to access to peoples' personal pages). Researchers have annotated such data for a variety of text characteristics, e.g., sentiment, opinions and factuality, often with the goal of building prediction models (McAuley & Leskovec, 2013; Pang & Lee, 2008). However, sharing (redistributing) the annotated (modified) data may not be permitted. Only providing pointers or unique key identifiers that link annotations to the original source can be one technical solution to this issue. Finally, prediction models built based on annotating such data may also be subject to inherited licenses and agreements, even though the original data cannot be reconstructed from these models.

5. Conclusion

In summary, the process of working with user and professionally generated, publicly available text data can be regulated by a multitude of rules and norms. Developing the awareness, knowledge and skills to responsibly consider these rules and account for grey zones is a challenging and evolving issue. One common risk is to mistake gratis data (access free of charge) as libre (collect and use with little or no restriction).

We believe that a vibrant dialogue between academe, the private sector and policy makers is needed to move ahead with establishing best practices and rules that enable the advancement of science, respect peoples' privacy, and offer incentives for commercial activities.

[20] https://dev.twitter.com/overview/documentation

[21] https://developers.facebook.com/

6. Acknowledgements

This work is supported in part by the FORD Foundation, grant 0155-0370, and a faculty fellowship from the National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana Champaign. We also thank the reviewers for their comments.

7. Bibliographical References

Christenson, H. (2011). HathiTrust. *Library Resources & Technical Services, 55*(2), 93-102.

Daries, J. P., Reich, J., Waldo, J., Young, E. M., Whittinghill, J., Ho, A. D., . . . Chuang, I. (2014). Privacy, anonymity, and big data in the social sciences. *Communications of the ACM, 57*(9), 56-63.

Diesner, J. (2015). Small decisions with big impact on data analytics. *Big Data & Society, 2*(2).

Diesner, J., & Chin, C. L. (2016). *Seeing the forest for the trees: considering applicable types of regulations for the responsible collection and analysis of human centered data.* Paper presented at the Human-Centered Data Science (HCDS) Workshop at 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW), San Fancisco, CA.

Diesner, J., & Chin, J. C. (2015). *Usable Ethics: Practical considerations for responsibly conducting research with social trace data*. Paper presented at the Beyond IRBs: Ethical Review Processes for Big Data Research, Washington, DC.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Gerner, D. J., Schrodt, P. A., Francisco, R. A., & Weddle, J. L. (1994). Machine coding of event data using regional and international sources. *International Studies Quarterly, 38*(1), 91-119.

Hodgson, C., Suber, P., Kiley, R., Kaufman, R., Goodrich, J., Eve, M. P., . . . Sutton, C. (2014). Open access infrastructure: where we are and where we need to go. *Information Standards Quarterly, 26*(2), 1-14.

Hoffman, D., & Bruening, P. (2015). *Rethinking privacy: Fair information practice principles reinterpreted*. Paper presented at the 37th Annual International Data Protection and Privacy Commissioners' Conference.

Kleinberg, J. (2008). The convergence of social and technological networks. *Communications of the ACM, 51*(11), 66-72.

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines.

American Psychologist, *70*(6), 543-556.

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788-8790.

Lane, J., Stodden, V., Bender, S., & Nissenbaum, H. (2014). *Privacy, Big Data, and the Public Good: Frameworks for Engagement*. New York, NY, USA: Cambridge University Press.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., . . . Gutmann, M. (2009). Life in the network: the coming age of computational social science. *Science*, *323*(5915), 721-723.

Leetaru, K., & Schrodt, P. A. (2013). *GDELT: Global data on events, location, and tone, 1979–2012.* Paper presented at the ISA Annual Convention, San Francisco, California, USA.

Lessig, L. (2004). *Free Culture: How Big Media Uses Technology and the Law to Lock Down Culture and Control Creativity*: Penguin.

McAuley, J., & Leskovec, J. (2013). *Hidden factors and hidden topics: understanding rating dimensions with review text (RecSys)*. Paper presented at the Proceedings of the 7th ACM conference on Recommender Systems, New York, NY.

Opensource.org. Open Source Licenses by Category. Retrieved from https://opensource.org/licenses/category

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1-135. doi:10.1561/1500000001

Recommended Data Repositories. (2016). Retrieved from http://www.nature.com/sdata/data-policies/repositories#general

Schrodt, P. A., Yilmaz, O., Gerner, D. J., & Hermreck, D. (2008). *Coding sub-state actors using the CAMEO (Conflict and Mediation Event Observations) actor coding dramework*. Paper presented at the Annual Meeting of the International Studies Association, San Francisco, CA.

Stallman, R. (2002). *Free software, free society: Selected essays of Richard M. Stallman*: Lulu. com.

Stodden, V., Leisch, F., & Peng, R. D. (2014). *Implementing Reproducible Research*: CRC Press.

Tiropanis, T., Hall, W., Crowcroft, J., Contractor, N., & Tassiulas, L. (2015). Network science, web science, and internet science. *Communications of the ACM*, *58*(8), 76-82.

Vitak, J., Shilton, K., & Ashktorab, Z. (2016). *Beyond the Belmont Principles: Ethical Challenges, Practices, and Beliefs in the Online Data Research Community*. Paper presented at the 9th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2016) San Francisco, CA.

Wilkin, J. (2009). BackTalk: HathiTrust and the Google Deal. *Library Journal*.

Zevenbergen, B., Mittelstadt, B., Véliz, C., Detweiler, C., Cath, C., Savulescu, J., & Whittaker, M. (2015). Philosophy Meets Internet Engineering: Ethics in Networked Systems Research. (GTC Workshop Outcomes Paper): Oxford Internet Institute, University of Oxford

Zimmer, M. (2010). “But the data is already public”: on the ethics of research in Facebook. *Ethics and Information Technology*, *12*(4), 313-325.