A US Perspective on Selected Legal and Ethical Issues Affecting the Development of Language Resources and Related Technology

Denise DiPersio
Linguistic Data Consortium, University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA 19104 USA
E-mail: dipersio@ldc.upenn.edu

Abstract

This paper surveys three issues of potential interest to the language research community from a US legal and ethical perspective. Unlike the European Union, there is no single *data protection* law in the United States. The various US laws and regulations touching on data protection and data privacy are examined here, along with some feedback from US survey respondents about sharing their personal data and a brief discussion of related ethical concerns. *Public sector data*, i.e., data developed by government agencies in the normal course of their work, can be a useful source for language resource development. US government entities at the federal and local level have launched open data initiatives over the last few years. Examples of the kinds of public sector information potentially available to the language research community is presented. Finally, w*eb data*, the principal source material for language resources, may be subject to use constraints. The major issues affecting that material -- copyright rights and website terms of use – are reviewed.

Keywords: data protection, privacy, human subjects research, open data

1. Introduction[1]

Developing language resources and associated human language technologies is a challenging process. It demands a clear understanding of the problem to be solved, the identification of the required data and a plan for using the data to develop, test and deploy the ultimate output (e.g., system or tool). Much of the data used in this process may have been created for other purposes and thus subject to limits or restrictions on its re-use. Or, the data in question may have been developed for the specific language technology task at hand; in that case, the collection protocol and use plan must take into account relevant legal, regulatory and ethical constraints. This paper examines three issues affecting language resource and technology development from a US perspective: data protection, the use of public sector information and web data collection.

With respect to **data protection**, the principal issues of concern are privacy and confidentiality, particularly as they relate to personal information. There is no single data protection law or regulation in the United States. As a result, issues touching on data protection are handled in sector-specific legislation and in some cases, through federal regulation. Those laws and regulations are discussed below. The recent disclosure of the Facebook-Cambridge Analytic data breach is discussed in the context of this regulatory framework.

Public sector information can be useful for language resource development. Under US federal, state and local open data initiatives, data created by government organizations in the course of their normal activities is available for re-use, usually without restriction. However, it may not always be a good fit for research tasks.

Web data collection is by far the principal way language resources are developed today. Easy access to web sites using automatic crawlers means that a large amount of data can be harvested in a short period of time with little or no capital outlay. But there may be hidden costs. How US copyright law and website terms of use can affect web-based data collection is discussed below.

2. Data Protection

The fact that there is no single data protection law in the United States has been explained as reflecting a bias in favor of scientific and technical progress in general, and for the industries involved, in particular. (Jones, 2017)

This is not to say that Americans are unconcerned about the use of data-based technologies to process their personal information. Advocates for personal privacy and for protections against the mechanization of society have been quite vocal over time. Nevertheless, the US regulatory framework assumes that personal data will be collected and processed and relies on government and industry to implement measures that protect such information against misuse.

Against this backdrop, we review laws that relate to data protection and privacy and assess their effectiveness. In the specific area of collecting data from human subjects for research purposes, we examine the federal « Common Rule », a regulatory scheme that was established to protect humans from research abuses. We close this section with a discussion of ethical matters relating to personal data protection.

2.1 US Data Protection Legislation

In the US, *data protection* and *privacy* are often used interchangeably. (Determann, 2016) Thus relevant laws and regulations affecting the treatment of personal information often contain both words in their titles. Following is a list of the principal federal statutes relating to data protection:[2]

The Fair Credit Reporting Act of 1970 (right to inspect, dispute and correct personal credit information held by consumer reporting agencies)

[1] This paper does not provide legal advice and nothing in this paper should be construed to constitute legal advice.

[2] This paper does not examine data protection/privacy laws enacted by US states, which in some cases offer more security for personal information than the federal counterpart.

Privacy Act of 1974 (practices federal agencies must follow with respect to information about individuals in agency records)
Cable Communications Policy Act of 1984 (personal information cannot be disclosed by cable TV operators without consent)
Electronic Communications Privacy Act of 1986 (extended telephone wiretapping guidelines to email and electronically stored information, covers government and private organizations)
Video Privacy Protection Act of 1988 (prohibits disclosure without consent of personal information in video rental records)
Computer Matching and Privacy Protection Act of 1988 (regulates matching records across government systems)
Driver's Policy Protection Act of 1994 (prohibits disclosure of driver's license, motor vehicle registration and related records)
Health Insurance Portability and Accountability Act of 1996 (HIPAA) (regulations protecting personal health information)
Communications Decency Act (immunity for online platform providers against unlawful uploaded content)
Family Educational Rights and Privacy Act (protects student education records)
Children's Online Privacy Protection Act of 1998 (relates to online services directed to children under 13 years of age)
Right to Financial Privacy Act of 1998 (procedures for government requests for information from personal financial records)
Graham-Leach-Bliley Act of 1999 (requires financial institutions to explain information-sharing practices and to safeguard sensitive data)
Telephone Records and Privacy Protection Act of 2006 (criminal penalties for fraudulent acquisition or unauthorized disclosure of phone records)
Genetic Information Nondiscrimination Act of 2008 (prohibits genetic information discrimination in health coverage and employment)

2.1.1 Is the US Framework Effective?

The list above demonstrates the disparate nature of US data protection legislation. Missing is a single definition of « personal » information that can be applied across the board; each law must be consulted for specifics.
Another shortcoming of the approach is that the laws apply to different actors, in some cases, to government agencies only, in others, to the private sector and in others, to both.
The sheer number of laws and their related regulations makes it a daunting task for individuals to navigate the system. Parsing the required notice from a bank, a health care provider, social media site or mobile phone company may end up leaving individuals more confused than informed, or resigned to the fact that their personal data is a commodity over which their control is slipping. Various government agencies, such as the Federal Trade Commission, serve as clearinghouses for information and as advocates under certain circumstances, but individuals bear the principal burden of responsibility to guard their personal information in an atmosphere privileging access over protection.
Would privacy be better served by a single edict that enumerates conditions on the use, and automated processing, of personal data? One argument in favor of the US approach is its flexibility. As seen above, the cluster of laws span four decades and reflect circumstances raised by new technologies over time. Applying a single framework to scattered agencies and private actors could be unwieldy. And, the ultimate result might not differ much from the current state, with regulations implemented for specific cases.
The ability of aggrieved persons, individually and through class actions, to seek redress for statutory violations through US courts can be viewed as a way to target specific violations in a relatively timely manner. Such claims carry the potential for monetary fines and damages for noncompliance and for deterring future bad behavior.
But the effectiveness of any data protection legislative scheme must be tempered by the order of magnitude by which personal data is collected and processed, on the one hand, and the large number of data breaches on the other hand. The Privacy Rights Clearinghouse reports that as of May 2018, over 10 billion records have been breached from more than 8000 data breaches reported since 2005.[3] Many breaches are never reported or only disclosed long after the fact. (Holtfreter, 2015) It may be that no legal system can adequately handle the task of protecting personal data under these circumstances.

2.1.2 How Do Americans Feel About Privacy?

Americans appear to be conflicted about issues surrounding the protection of their personal data. Despite laws that require that they be informed about how their personal data is collected and used, many Americans are uncertain about how to manage their personal information. At the same time, they are not averse to sharing that information under certain circumstances.
In a survey conducted in 2016 by the US Pew Research Center probing feelings about privacy and sharing personal information, respondents showed a lack of confidence in how their personal data is protected by public and private organizations. (Rainie, 2016) Over 90% of those surveyed agreed or strongly agreed « that consumers have lost control of how personal information is collected and used by companies .» (Ibid.) Yet many of the same respondents supported US government surveillance efforts with respect to international terrorism and indicated that they would support stronger measures. But other respondents were concerned about government surveillance of US citizens, and a number reported that they had taken steps to reduce their online presence in the face of potential surveillance. (Ibid.) And others are willing to share their personal information and risk surveillance if they believe they will receive value in return. (Rainie & Duggan, 2015)
Unfortunately, « a human in the loop » is not really part of the data privacy conversation in the United States although that may change somewhat as a result of the Facebook-Cambridge Analytica case discussed below and the implementation of the European General Data Protection Regulation (GDPR). In the meantime, individuals continue

[3]Privacy Rights Clearinghouse, Data Breaches, https://www.privacyrights.org/data-breaches accessed 30 May 2018.

to surrender a portion of their personhood for the sake of technology.[4]

2.1.3 Facebook-Cambridge Analytica

In March 2018, media reports in The Guardian and the New York Times stated that data from over 50 million Facebook users had been obtained by the political consultant firm Cambridge Analytica and used in 2016 in the US Presidential election and the UK campaign leading up to the Brexit vote. Shortly after, Facebook stated that roughly 87 million user profiles were involved. The information was obtained from a personality quiz application for Facebook's Open Graph platform developed by a Cambridge University researcher in 2013. Under the platform's terms of use at that time, those conducting the survey could collect information from the respondents and their Facebook friends, the latter without the consent of those friends. Facebook claims to have changed the platform terms in 2014, no longer allowing access to friends' information without the friends' consent. But the application developer shared the app and collected data with Cambridge Analytica who then used it to develop profiles for targeting political ads to certain users. Whether Facebook asked Cambridge Analytica to delete the information and Cambridge Analytica certified that it had done so is not clear. Cambridge Analytica has said that it did not use any of the data for ads relating to the US presidential election.

This was not the first time that Facebook's privacy policies were in the news. In 2011 it settled charges that it deceived its users about the privacy of their information by entering into a consent decree with the US Federal Trade Commission (FTC), one of the agencies charged with protecting US consumers. Among other things, Facebook is required to undergo third-party audits of its privacy practices until 2031. Violations of the order are subject to fines of US40,000 per violation per day. After the Cambridge Analytica disclosure, the FTC opened an investigation to determine whether Facebook violated the 2011 order.

This situation exposes the weaknesses of the US approach to data protection and privacy. The original terms of use on the Open Graph platform did not violate any privacy law; the US Privacy Act of 1974 only applies to federal government agencies. Making private companies accountable under a single privacy law combined with a specific opt-in user procedure may have mitigated some damage here.

But recall US attitudes about the tradeoff between giving up personal information in return for access to services. (Rainie & Duggan, 2015) Although there was an initial « delete Facebook » movement after the scandal was exposed, it appears that most users remain.

The reaction of privacy experts, among others, was in the form of a reality check. Users who voluntarily provide their personal information to internet-based entities, particularly social media, should understand that none of that information is ever « private »; at most, it may be less public under certain circumstances.

2.2 Collecting Data from Humans for Research: The Federal Common Rule

The collection of data from human subjects for research purposes is governed in the United States by the Common Rule, a statutory and regulatory system that grew out of past abuses in human scientific experiments. The National Research Act, which covered biomedical and behavioral research, was passed in 1974. Along with a series of regulations, that law created a National Commission which eventually issued the 1979 Belmont Report, the seminal document defining ethical principles and guidelines for protecting human subjects participating in biomedical and behavioral research.

The Belmont Report identified three fundamental principles for all human subjects research: respect for persons, beneficence and justice. Respect for persons means preserving individual autonomy, obtaining participants' informed consent and being truthful with participants about the study. The idea of beneficence is to do nothing to harm subjects; the goal should be to maximize research benefits and to minimize harm to individuals. Finally, justice refers to procedures that are administered fairly, are reasonable and are not exploitative. For instance, if subjects are compensated for their participation, that compensation should be fair. The Common Rule represents the implementation of those ideas.

The Common Rule is administered across US universities by Institutional Review Boards (IRB) that review research protocols for compliance. The thrust of the regulations and IRBs (including their composition) is geared to medical research (e.g., clinical studies). Behavioral research was never defined nor were social scientists directly solicited for input in the early stages of IRB development. (Schrag, 2010) Some IRBs continue to scrutinize benign social science studies as they would a drug trial, which can result in over-restrictive conditions on using and sharing research data. [5]

However, most language-related human subjects studies are considered minimal risk studies that involve no risk greater than the activities of daily life. This means that requests for such protocols can usually be handled by IRBs under expedited review. The protocol must include a method for obtaining a subject's consent to the study; an option for the subject to leave the study for any reason and to withdraw the data contributed to the study; and a process for protecting personal information from disclosure (usually through anonymization) unless the subject agrees that certain personal data can be shared. For studies conducted by the Linguistic Data Consortium (LDC), we also obtain participants' consent to include their data in a corpus that will be distributed and shared with others.

[4]One scholar explains this situation in the following terms: « the political question that developed was not how to *protect* a concept of personhood from computation but how to *use* the computational system to protect a notion of personhood. » (Jones, 2017: 227) (original emphasis)

[5]New regulations making it easier to conduct minimal risk social science studies that represent several years of review by the US government and comments from the community were set to be effective in January 2018. That date has been delayed to July 2018 and may be postponed further.

2.3 Using Personal Data for Research: Is De-identification Enough?

The principal way the language research community handles the problem of personal data in a corpus is to anonymize or de-identify it using random numbers or strings. This has been thought to be an effective way to allow research to proceed while preserving privacy and confidentiality. To the extent that was and is still true for some kinds of data, it may be inadequate to deal with the growing volumes of personal data.

The advent of speaker and facial recognition systems can make it more likely that individuals who are subjects of audio and video data collections can be identified. Also, it is well known that it is possible to identify individuals in an « anonymized » database using only a few data points. For example, three fields from US Census summary data – 5-digit zip code, gender and date of birth – reported unique characteristics for 87% of the US population. (Sweeney, 2000) Patients in a hospital data set were re-identified by crosslinking common characteristics in the hospital set and a local voter registration list. (Ibid.) In a study by MIT researchers, the dates and locations of four purchases were sufficient to identify 90% of 1.1 million credit card holders in a data set covering a three-month purchasing history. (de Montjoye, et al., 2015) In the last instance, the researchers concluded, «our findings highlight the need to reform our data protection mechanisms beyond PII [personal identifying information] and anonymity and toward a more quantitative assessment of the likelihood of reidentification." (Ibid.: 539)

2.4 Algorithm Bias

Another effect of personal data processing is how algorithms developed from such data can be applied in potentially harmful ways. Internal biases built into algorithms used, for example, to choose applicants for job interviews, to make parole decisions and to grant loan applications are examples of problematic outcomes, however unintended. (Jones, 2017; Knight, 2017)

Potential solutions include code audits that would focus on the principal variable(s) involved in a decision and how removing that variable affects the decision. (Cramer, 2017) In the context of government agency decisions based on algorithms, suggestions include providing an explanation for the logic behind an algorithm so that a particular decision can be understood. (Ibid.)

Transparency is a key concept here. Convincing organizations to disclose something generally considered to be proprietary is a challenge. Self-review may be a good first step. But ultimately, some framework that covers accountability, transparency and oversight – perhaps as part of a federal law or regulation – may hold the best promise for lasting results.

2.5 The AI Factor

Although the notion of artificial intelligence is not new to science, the current marriage between natural language processing and artificial intelligence to develop a new generation of interactive robots raises issues that implicate privacy. One commentator has described the issues as touching on three areas: surveillance, access and social presence. (Calo, 2012)

Surveillance includes, of course, traditional spying activities by drones. But it also covers law enforcement-type investigational techniques that are possible because of the robot's size, technical capabilities and mobility. For example, observing a residence from some distance using advanced technology would not necessarily require a court-issued warrant, thereby reducing an individual's privacy zone. (Ibid.: 190-191)

Smart home technology and smart assistants provide access to private spaces. In those spaces, robots record daily activities, communicate with humans in the vicinity and almost always transmit that data to a space (sometimes via the Internet) where it is stored and processed for various purposes. This raises the risk that others can access such data through various means. Hackers can penetrate data storage systems, and law enforcement might simply request the information from the service provider. (Ibid.: 193)

The social ramifications of robot development that diminish a sense of privacy are less obvious. They relate to how humans interact with their robot « companions ». Studies indicate that humans tend to assign social robots human characteristics (e.g., by naming them) and are aware of their presence even when alone. (Ibid.: 195-196) To the extent all humans need times of solitude, the ubiquitous robot reduces those opportunities.

How robot activities and human interactions with robots will affect current US privacy laws remains to be seen. Those laws were designed with more traditional data keeping in mind, where some data is (usually) voluntarily provided by an individual in an identifiable transaction – through a bank account, renting a video, obtaining a driver's license, using a credit card. Data from robot transactions and interactions is generated less obviously and not always through an individual's deliberate action. This will likely require some rethinking from regulators and more importantly, an adjustment of individual expectations about personal privacy.

3. Public Sector Information

US federal and state governments generate volumes of data as part of their routine activities. In 2013, the US Executive declared open, machine-readable data « the new default for government information.»[6] The vehicle for that initiative is the website DATA.GOV, which contains over 200,000 data sets.[7] All of the data on the site is represented as containing public information only. Data from US federal sources is available at no cost and without restriction. The site also contains data from non-federal sources that are freely available as well, but may be subject to license constraints which appear on the particular data set page(s). Similarly, many US states have implemented open data portals granting public access to state government-generated data sets. In 2014, twenty-four states had

[6]Open Government Initiative, https://obamawhitehouse.archives.gov/open, accessed 1 March 2018.

[7]DATA.GOV, https://www.data.gov/ accessed 1 March 2018.

established such portals and ten states had adopted open data policies, the aims of which typically relate to government transparency and accountability, encouraging civic engagement and promoting data use and related innovation. (Drees & Castro, 2014) In Pennsylvania, LDC's home state, close to 100 data sets are available at no cost and without restriction from opendataPA.[8]

Municipalities are moving into the open data space as well. LDC's hometown, Philadelphia, sponsors OpenDataPhilly, which is the City's official data portal and also contains data sets from regional organizations.[9] The City of New York likewise makes all data generated by city agencies available at no cost under its Open Data Law enacted in 2012, which it claims is one of the most « robust » open data policies worldwide.[10]

This is good news for researchers generally, but not necessarily for the language research community. Lacking from available data bases are substantial speech data or annotated text. A search of DATA.GOV for data sets with «speech » yields a few corpora developed by the National Aeronautics and Space Administration for speech acquisition and automatic speech recognition.[11].

The availability of open data developed by government agencies is a good thing in general, but its usefulness to the language resources and technologies communities is not clear.

4. Web Data Collection

Web data plays an important role in developing language resources and infrastructures. The principal US legal issues affecting its use are copyright rights and contractual issues relating to website terms of use.

4.1 Copyright

Despite the fact that it is easily accessible, much of the data available on the web is subject to copyright. Short of obtaining explicit permission from the copyright holder – something that can be difficult when large volumes of data from many sites are involved – one can attempt to rely on the fair use exception to US copyright law.[12] Fair use requires a fact-specific analysis addressing the purpose and character of the use; the nature of the copyrighted work; the amount of the work used in relation to the entire work; and the effect of the use on the work's potential market. [13] Over the last several years, US courts have viewed the exception favorably in cases involving digital data. Thus, fair use has been found to cover full-text searchable databases.[14] The crux of those rulings revolves around the transformative nature of the use which "serves a new and different function from the original work and is a substitute for it,"[15] even if the original work is not changed. Using a whole work for such transformative purposes is generally allowed and such uses have been found not to threaten the work's potential market. All of this is good news for developing language resources and human language technologies.

4.2 Terms of Use

The positive trend of legal interpretations for fair use with respect to digital data must be tempered by the terms of use or terms of service (sometimes referred to as "browser wrap" terms) that are typically included on most websites. Those terms often restrict how material can be used, how it can be shared and whether it can be modified, among other things. Terms for sites hosting third party data, including social media sites, may contain additional restrictions. Most terms state that accessing the site constitutes assent to the terms. The extent to which such terms are enforceable is not clear.[16] There are some web services that track terms of use and service and their modifications.[17] Users should consult qualified legal advice about specific sites and provisions in terms of use or terms of service.

5. Conclusion

This survey of three issues of interest to the language research community – data protection, public sector information and web data collection – is meant to provide a US perspective. As seen above, personal data is subject to a variety of laws, including special regulations for human subjects research. In addition, the growth of technologies using algorithms based on personal data and

[8]OpendataPA, https://data.pa.gov/ accessed 01 March 2018.

[9]OpenDataPhilly, https://www.opendataphilly.org accessed 01 March 2018.

[10]NYC OpenData, https://opendata.cityofnewyork.us/ accessed 01 March 2018.

[11]Superior Speech Acquisition and Robust Automatic Speech Recognition for Integrated Spacesuit Audio Systems Project, https://catalog.data.gov/dataset/superior-speech-acquisition-and-robust-automatic-speech-recognition-for-integrated-spacesu-b2430 accessed 01 March 2018.

[12]17 USC 107.

[13]17 USC 107 (1)-(4).

[14]Authors Guild v. Google, 894 F.3d 202 (2d Cir. 2015) (searchable database of digital copyrighted books); Fox News Network v. TVEYES, Inc., 43 F.Supp.3d 379 (S.D.N.Y. 2014) (searchable database of broadcast news); Authors Guild v. HathiTrust, 755 F.3d 87 (2d Cir. 2014) (full text searchable database); A.V. v. iParadigms, LLC, 562 F.3d 630 (4th Cir. 2009) (student paper plagiarism database).

[15]Authors Guild v. HathiTrust, 755 F.3d at 96.

[16]Electronic Frontier Foundation, The Clicks That Bind: Way Users "Agree" to Online Terms of Service, https://www.eff.org/wp/clicks-bind-ways-users-agree-online-terms-service accessed 4 March 2018.

[17]Terms of Service: Didn't Read, https://tosdr.org/ accessed 4 March 2018; TOSBack, https://tosback.org/ accessed 4 March 2018.

the increasing use of interactive robots suggest the need to rethink the US preference for technical solutions to data protection and privacy. Government open data initiatives have the prospect of providing to the community no cost public sector data without restriction. Finally, one should be aware of potential constraints on collecting and using web data, particularly with respect to site terms of use.

6. Bibliographical References

Calo, M. Ryan (2012). Robots and Privacy. In P. Lin, K. Abney & G. A. Bekey (Eds.), *Robot Ethics, The Ethical and Social Implications of Robotics*. Cambridge, Massachusetts: The MIT Press, pp. 187-201.

CNBC.com. Facebook-Cambridge Analytica : A timeline of the data hacking scandal. Available at: https://www.cnbc.com/2018/04/10/facebook-cambridge-analytica-a-timeline-of-the-data-hijacking-scandal.html.

Cramer, Katie. (Oct. 10, 2017). How Can We Reveal Bias in Computer Algorithms? *The Regulatory Review*. Available at: https://www.theregreview.org/2017/10/10/cramer-bias-computer-algorithms/.

DATA.GOV. Available at: https://www.data.gov/.

deMontjoye, Y., Radaelli, L., Singh, V.K. and Pentland, A. R. (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science* 347 (6221): 536-539.

Determann, L. (2016). Adequacy of data protection in the USA: myths and facts. *International Data Privacy Law*, 6(3): 244-250.

DiPersio, D. (2014). Linguistic Fieldwork and IRB Human Subjects Proposals. *Language & Linguistics Compass*, 11: 505-511.

Drees, L. and Castro, D. (2014). State Open Data Policies and Portals. *Center for Data Innovation*, available at http://www2.datainnovation.org/2014-open-data.pdf.

Electronic Frontier Foundation. *The Clicks That Bind: Way Users "Agree" to Online Terms of Service*. Available at: https://www.eff.org/wp/clicks-bind-ways-users-agree-online-terms-service.

Federal Trade Commission. Facebook Settles FTC Charges That It Deceived Consumers By Failing To Keep Privacy Promises. Available at: https://www.ftc.gov/news-events/press-releases/2011/11/facebook-settles-ftc-charges-it-deceived-consumers-failing-keep.

Gellert, R. (2015). Data protection: a risk regulation? Between the risk management of everything and the precautionary alternative. *International Data Privacy Law*, 5(1): 3-19.

Holtfreter, R.E. (2015). Data breach trends in the United States. *Journal of Financial Crime*, 22(2): 242-260.

Jones, M.L. (2017). The right to a human in the loop: Political constructions of computer automation and personhood. Social Studies of Science, 47(2): 216-239.

Knight, Will. (July 12, 2017). Biased Algorithms Are everywhere, and No One Seems to Care. *MIT Technology Review*. Available at: https://www.technologyreview.com/s/608248/biased-algorithms-are-everywhere-and-no-one-seems-to-care/

Kuner, C., Cate, F. Millard, C. and Svantesson, D. (2012). The challenge of 'big data' for data protection. *International Data Privacy Law*, 2(2): 47-49.

NYC OpenData. Available at: https://opendata.cityofnewyork.us/.

OpendataPA. Available at: https://data. pa.gov/.

OpenDataPhilly. Available at: https://www.opendataphilly.org.

Open Government Initiative. Available at: https://obamawhitehouse.archives.gov/open.

Privacy Rights Clearinghouse, Data Breaches. Available at: https://www.privacyrights.org/data-breaches.

Protection of Human Subjects. 2009. 45 CFR 46.116.

Rainie, L. (2016). *The state of privacy in post-Snowden America*. Pew Research Center. Available at: http://www.pewresearch.org/fact-tank/2016/09/21/the-state-of-privacy.

Rainie, L. and Duggan, M. (2015). *Privacy Information and Sharing*. Pew Research Center. Available at: http:://www.pewinternet.org/2016/01/14/2016/Privacy-and-Information-Sharing/.

Schrag, Z. (2010). *Ethical Imperialism: Institutional Review Boards and the Social Sciences, 1965-2009*. Baltimore, Maryland: The Johns Hopkins University Press.

Sweeney, L. (2000). Simple Demographics Often Identify People Uniquely. Pittsburgh, Pennsylvania: Carnegie Mellon University, Data Privacy Working Paper3.

Terms of Service: Didn't Read. Available at: https://tosdr.org/.

The Belmont Report – Ethical Principles and Guidelines for the protection of human subjects of research. Available at: http://ohsr.od.nih.gov/guidelines/belmont.html.

TOSBack. Available at: https://tosback.org/.

7. Legal Case References

Authors Guild v. Google, 894 F.3d 202 (2d Cir. 2015).

Authors Guild v. HathiTrust, 755 F.3d 87 (2d Cir. 2014).

A.V. v. iParadigms, LLC, 562 F.3d 630 (4th Cir. 2009).

Fox News Network v. TVEYES, Inc., 43 F.Supp.3d 379 (S.D.N.Y. 2014).