COLEG GWYDDORAU'R AMGYLCHEDD A PHEIRIANNEG
COLLEGE OF ENVIRONMENTAL SCIENCES AND ENGINEERING

PRIFYSGOL BANGOR UNIVERSITY

Ethical Concerns Surrounding Web Scraping & Internet Data

Cameron C. Gray

May 8, 2021

Note: The information in this paper is up-to-date as of the date listed above. The larger the difference between this date and the date of use, the more likely that best practice and amendments to law have taken place. If the date of this document is more than six months in the past, please seek an updated version.

Abstract

...

Introduction

This paper aims to frame future deliberations of the College of Environmental Sciences and Engineering Research Ethics committee. It will present the current state of law, technical concerns and best practices with the use of internet-based data collection, online surveys, and the use of social media in research projects. This work is not intended to be a replacement for the deliberations of the committee. This paper can be shared with researcher to help inform their proposals before presentation to the committee.

The internet is, unfortunately, a massively fractured entity in terms of regulation, acceptable standards, and data ethics. The nature of the technology and the development of the internet has meant that control and regulation have evolved in the same way as most of international law. There is a set of generally accepted principles, but these are superseded by law from sovereign states.

As a result this paper will focus on five main jurisdictions which cover the vast majority of cases the ethics committee would need to consider. These jurisdictions are United Kingdom, European Union, United States, Russian Federation/China, and general/international. The Russian Federation and China have been grouped because while they are distinct legal systems, the governments of those territories have followed a consistent approach to policing and regulating the internet/internet use.

Legal Environments

Applying laws to a decentralised, essentially privately owned, communications system such as the internet presents challenges for national legislatures the world over. As such most legal environments are rushing to catch up with the evolving nature of modern communications. This means that there are a mixture of bespoke internet-era laws specifically covering the technology appropriately, and systems applying laws applicable to prior generations of technology.

PRIFYSGOL BANGOR
BANGOR, GWYNEDD,
LL57 2DG, DU

+44 (01248) 351151
WWW.BANGOR.AC.UK

BANGOR UNIVERSITY
BANGOR, GWYNEDD,
LL57 2DG, UK

+44 (01248) 351151

Applicable Jurisdictions

Most legal issues will be found in two distinct jurisdictions; the location of viewer/visitor/user and the location the material is served from. In IT terms, these locations are where the data is said to be 'at rest'. Depending on these two locations the data 'in flight' may transit many other territories and legal systems. Most of these territories are not interested in prosecuting cases that the data would otherwise apply to.

China, however, routinely examines data that transits its so-called 'Great Firewall' [15]. The state is mainly looking for materials and data that could 'destabilise the People's Republic', also known as criticism, democratic movements, and unsanctioned communications. If any data is deemed to have breached these standards, China has been known to prosecute its own citizens for allowing the material and other unknown users in absetia [22].

Cybercrime Law

The key piece of UK legislation, the Computer Misuse Act [11], was gained Royal Assent in 1990. Prior to this prosecutions were attempted under laws such as the Wireless Telegraphy Act [7] - meant to regulate the use of radio, the Forgery and Counterfeiting Act [9] or the Theft Act [8]. The United States has similarly updated their prevailing law with the introduction of the Computer Fraud and Abuse Act [26] in 1986. This law has been amended multiple times since its original enacting. Most recently by the USA PATRIOT Act in 2002 and the Identity Theft Enforcement and Restitution Act of 2008.

EU law surrounds a main cyber-crime directive, the Attacks Against Information Systems Directive (AAISD) [3]. The Directive has now been directly implemented into member states national law as required by the Lisbon Treaty (2009). To date the requirements of the AAISD have not been included those being repealed as part of Brexit. Crimes within the Russian Federation are prosecuted under Section 28 of the Criminal Code [24]. The Russian Federation has passed sweeping new cyber-crime and internet policing laws in the past year (2019). The majority of these changes are to delineate the difference between state-sponsored computer crime and private computer crime. Most commentators consider these changes to be condoning more cyber-crime than preventing it [16], [23]. China has condensed all of their cyber-crime law into a single omnibus law which became effective in 2017 [1].

Copyright Law

Copyright is automatic for any natural person (as opposed to legal person) creating an intellectual or creative work. As such, facts or underlying theory cannot be copyrighted – for example the base content of an encyclopaedia. However, there is a second type of copyright known as "copyright in assembly or collection". This type deems that the intellectual work is in the collection and presentation of the facts. This has been tested in court, in the precedent a telephone directory was held to be a copyrighted in assembly work [6].

Most legal systems hold that copyright claims/suits can be raised in any jurisdiction where the infringing material can be accessed. However, damages are limited to losses in that jurisdiction [20]. This is enshrined in an international treaty, the Berne Convention [27].

Most jurisdictions [4], [10], [25] allow the following exemptions to standard copyright protections; criticism, news reporting, teaching, scholarship, and research purposes. This is known in almost all legal systems as the "Fair Use Doctrine". For almost all research purposes, the concept of fair use will protect the researchers and institution from claims for reproduction of copyrighted works. However, the fair use doctrine *does not* universally cover paraphrasing or modification of the original work. In this scenario, the persons involve create a *Derivative Work*.

There are specialist concerns when dealing with online resources, especially surrounding user-generated content such as on social media, comments on items and on forums.

Sites will often have additional agreements, either linked in the footer or displayed when creating an account, that users must enter into. Collectively these are known as "Terms of Use" or "Terms of Service". These agreements fall under standard contract law of whichever jurisdiction they are subject to. For these agreements and given the international nature of the internet; they will ordinarily state which jurisdiction is to be used. If the agreement does not specifically stipulate, then the jurisdiction where the owner or the user is domiciled will normally be primary.

The current arrangement of terms of service to use these sites and the wording of copyright statutes mean that the site owners/operators do not hold copyright over this content. The user themselves retains the copyright over their contributions. Fair use does still apply to these contributions, as the user has explicitly 'published' the work by making it available on the site.

Data Protection/User Privacy/Data Processing Consent Law

Privacy and Data Protection law is some of the most varied with some jurisdictions, such as the EU and California, setting out strong protections for individuals and others allowing greater unchecked exploitation of privacy and personal data. There is one key element of guidance that applies almost universally to all Data Protection law: *it is a myth that Data Protection law prevents any kind of use or processing.* These laws place a responsibility on those that are processing the data to obtain appropriate permissions from the owner of the data (or the subject of the data if different), rather than prohibiting them from conducting the processing at all [21].

Despite Brexit, the UK's Data Protection Act (as amended 2018) [13] remains fully compliant with current EU data protection regulations including the General Data Protection Regulations (GDPR) [5]. Even if the legislation did not match, the pervasive nature of the GDPR covers all EU citizens irrespective of legal jurisdiction of the data processing. However, due to Brexit citizens of England, Scotland and Wales are no longer covered by the EU GDPR. The Northern Ireland Protocol means that until such time as either party withdraws or the Protocol is superseded by another instrument, NI citizens and operations are still bound by EU regulations.

The issue is somewhat clouded by Statutory Instrument 2019/419 [14] which creates what is being referred to as *UK GDPR*. This Statutory Instrument requires organisations, data controllers, data processors and Data Protection Officers to amend their procedures, practices and published materials to reflect the UK's new independent status. In particular, Article 30 records, privacy notices, DPIAs (data protection impact assessments), DSARs (data subject access requests) and documentation covering international data flows. The new UK GDPR regime came into force 1st Janaury 2021. The Instrument *does not* rescind any additional requirements that the EU legislation places on any entity dealing with EU citizen data.

Researchers at Bangor University are only impacted by these statutes, instruments and regulations when dealing with Personally Identifiable Information of Natural Persons. There are important definitions that must be respected in this legislation and guidance.

- Natural Persons: Almost all jurisdictions distinguish between legal persons (those able to form contracts predominately including corporate entities) and natural persons. These Natural Persons are defined as a living human being. There are however, special categories dealing with minors and those that are vulnerable or incapable of executing their own affairs.
- Personally Identifiable Information: For a piece of data to be classed as Personal for the purposes of law, it must be clearly linked to a single specific Natural Person. This link must be through additional data that is;
  a) already in the public domain,
  b) already in the possession of the data controller,
  c) or could be reasonably expected to be available in either the public domain or within the scope of the data controller.

In addition, UK data protection legislation has also recognised more sensitive categories of Personally Identifiable Information:-

- Private Data: These data points are most often identifiers for various services whether governmental, public or private sector-provided. Most interpretations also recognise key life-stage dates in this category. Some of this Private Data is routinely available, however it only becomes dangerous when multiple points are linked to provide a profile.
- Sensitive Data: Most sensitive data includes life choices, memberships in organisations, religion, political affiliations and medical records. These data points are not routinely available unless disclosed by the subject.
- Protected Characteristics: These items are a special subset of Sensitive Data. These are defined by the Equality Act 2010 [12]. These data points are:
  age, disability, gender reassignment, marriage and civil partnership, pregnancy and maternity, race, religion or belief, sex, and sexual orientation.
  The legislation makes it illegal for all discrimination on the basis of any of these data points. As many are used as a basis of demographic sample design, or inclusion/exclusion criteria; this category of data can be problematic.

Collection of, use, or processing of Personally Identifiable Information is not restricted in current UK law. However, data subjects must provide clear, positive permission for anyone to process any data. This permission must be for a clearly identified purpose and cannot be used for any other purpose unless explicitly re-authorised.

The Different Internet(s)

While these terms are bandied around by mainstream media and often repeated in research, most of the time the names of the 'different' internet areas are misused. All of these are actually a proper subset of the internet.

The World Wide Web

The World Wide Web (WWW) has been used synonymously for the internet since the early stages of consumer access. However, it only represents a small fraction of legitimate traffic on the internet. The WWW covers all services and material that can be accessed using a standard internet browser application, such as Chrome, Edge, Safari and Internet Explorer. This type of traffic amounts to 25.2% of all internet usage worldwide [19]. Almost universally these are traditional web sites, as defined by the common usage. However, since the rise of the mobile internet and wider adoption of smartphones, this definition extends to cover the usage of certain native apps interacting with the API of their corresponding web site. Almost all of these services and this content is indexed by popular search engines and is designed to be easily accessible. This is both through the listings and through standard domain names (e.g. www.bangor.ac.uk).

The Deep Web

The Deep Web is not a separate entity, and shares the same protocols and tools to access the services. Most often the phrase 'deep web' will be used to represent web sites, again accessed via Chrome, Edge, Safari and Internet Explorer, that are not listed in traditional search engines and are not accessed using a well-known domain name. There are some 'split root' domain names used, but most often these services are accessed directly using IP addresses (the native addresses for computers on a modern network). These can look either like `012.123.234.012` (a version 4 address) or like `2001:0DB8:DEAD:BEEF:CAFE:BABE:3231:01CE` (a version 6 address). As these services/sites are not listed in search engines, you must find or know the address needed before hand. There are specialist search engines (such as Torch, DuckDuckGo and ParaZite) that claim to index part or all of the Deep Web, but these are largely unreliable for accessing genuinely hidden services.

The Dark Web

The Dark Web exists as an overlay on the traditional internet. It requires specific clients and access methods. The most well known of these is TOR (The Onion Router). There are other less well known arrangements of VPNs (Virtual Private Networks) and overlays for very specific subcultures and communities. These communities are very close-knit, and often require invitations to fully join. Some services using TOR can be accessed via .onion URLs, but most use node addresses which fluctuate to avoid detection and geolocation. There are some 'community projects' that act as directories for the contents of the Dark Web, such as The Hidden Wiki. Even with these shortcuts, the vast majority of the content in the Dark Web is involved in crime [2]. A 2016 King's College study [18] claimed to find 57% of the content they were able to access was associated with some form of organised or semi-organised crime. The largest group is associated with the illicit/illegal drugs trade.

Potential for Harm

While the process for establishing potential risks and harms to participants is well understood and clear, the legal environment means that there is potential harms to researchers as well.

Harms to Researchers - Data

Bangor University's Research Data Management Policy [17], does not assume institutional responsibility for adhering to applicable regulation and law. The consequence is that researchers are considered for the purposes of the laws as the responsible parties or Data Controllers. As a matter of ethical approval and oversight, the relevant Ethics Committee has a duty of care to ensure that management plans, processes and data requested meet the tests under appropriate law. This is to minimise the risk of prosecution and harm to the researchers themselves.

Harms to Researchers - Copyright

Academics and career researchers are (or at least should be) aware of the broad points of copyright in a plagiarism and referencing context. However, these principles can be less clear when dealing with reproduction, paraphrasing and reporting of online sources. This is usually because there is no standard or established mechanism for appropriate attribution and rights clearances.

Special care needs to be taken when dealing with user-contributed materials (as opposed to materials published by the site owner) as copyright still applies. Depending on the question being investigated seeking permission from each user may influence the research being conducted, or place the researcher(s) in harms way by revealing their identity and motive for seeking permission.

Harms to Researchers - Access Restrictions

Communities and sub-cultures associated with criminal enterprises are often access restricted requiring anything from payment to passing 'tests' to gain entry. These tests are normally based around committing a similar crime, or providing evidence of the same, to prove the newcomer is one of the group. Obviously, no reasonable researcher would agree to commit a crime, however providing falsified 'evidence' of the same can carry risks too.

Harms to Researchers - Physical Violence/Harassment/Retaliation

As with any form of infiltration of a closed community, those that are discovered can be subject to differing forms of retaliation. Anonymity is a central pillar of why these activities are conducted on the Dark Web, therefore anything that threatens to identify actors is seized upon. With most situations, especially where the researcher has taken as many precautions as possible, this will be limited to online threats and harassment. This can be hurtful and worrying, but the risk of escalation is usually low. However, as the level of criminality increases the risk of the retaliation becoming more physical increases too. This would be a particular concern where members of the group are being profiled or outed. The details do not need to be published for them to be perceived as a threat. Many of these cases of violent retribution are handled as 'in the line of duty' where law enforcement personnel are involved. Alternatively, these are seen as 'gangland' murders where criminal elements have turned on each other. There are very few reported cases of innocents being caught up in these conflicts, however this risk is increased if a researcher may be in a position to expose identities and dealings.

References

[1] Central People's Council and State Council. "Cybersecurity law of the people's republic of china (translated)." Original Chinese at http://www.npc.gov.cn/npc/xinwen/2016-11/07/content_2001605.htm. (2016), [Online]. Available: https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-cybersecurity-law-peoples-republic-china/ (visited on 05/22/2020).

[2] G. Davies, "Shining a light on policing of the dark web: An analysis of uk investigatory powers," *The Journal of Criminal Law*, vol. 84, no. 5, pp. 407–426, 2020. DOI: 10.1177/0022018320952557.

[3] European Commission, "Attacks against information systems directive," *Official Journal*, 2012/40/EU, 2013. [Online]. Available: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:218:0008:0014:EN:PDF.

[4] ——, "Copyrght in the digital single market directive," *Official Journal*, 2019/790/EU, 2013. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2019.130.01.0092.01.ENG.

[5] ——, "Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation)," *Official Journal*, 2016/679/EU, 2016. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv%5C%3AOJ.L_.2016.119.01.0001.01.ENG.

[6] *Feist Publications, Incorporated v. Rural Telephone Service Company, Incorporated*. Supreme Court of the United States, 1991.

[7] H. M. Government, *Wireless Telegraphy Act (Repealed)*. London: HMSO, 1967, c. 72. [Online]. Available: https://www.legislation.gov.uk/ukpga/1967/72/contents.

[8] ——, *Theft Act*. London: HMSO, 1978, c. 31. [Online]. Available: https://www.legislation.gov.uk/ukpga/1978/31/contents.

[9] ——, *Forgery and Counterfeiting Act*. London: HMSO, 1981, c. 45. [Online]. Available: https://www.legislation.gov.uk/ukpga/1981/45/contents.

[10] ——, *Copyright, Designs and Patents Act*. London: HMSO, 1988, c. 48. [Online]. Available: https://www.legislation.gov.uk/ukpga/1988/48/contents.

[11] ——, *Computer Misuse Act*. London: HMSO, 1990, c. 18. [Online]. Available: https://www.legislation.gov.uk/ukpga/1990/18/contents.

[12] ——, *Equality Act*. London: HMSO, 2010, c. 15. [Online]. Available: https://www.legislation.gov.uk/ukpga/2010/15/contents.

[13] ——, *Data Protection Act*. London: HMSO, 2018, c. 12. [Online]. Available: https://www.legislation.gov.uk/ukpga/2018/12/contents.

[14] ——, *The Data Protection, Privacy and Electronic Communications (Amendments etc) (EU Exit) Regulations 2019*. London: HMSO, 2019, 2019/419. [Online]. Available: https://www.legislation.gov.uk/uksi/2019/419/contents.

[15] S. Kalathil, *Beyond the great firewall: How China became a global information power*. Center for International Media Assistance, 2017.

[16] D. Kundaliya. "Russia's new cyber laws will fuel online crime, claims report." (2019), [Online]. Available: https://www.computing.co.uk/news/3080270/russia-cyber-crime (visited on 05/22/2020).

[17] Library & Archives Service, *Research data management policy*, Institutional Policy, Bangor University, Mar. 2019. [Online]. Available: https://www.bangor.ac.uk/library/documents/RDM/BU%5C20RDM%5C%20Policy_redraft_March2019.pdf.

[18] D. Moore and T. Rid, "Cryptopolitik and the darknet," *Survival*, vol. 58, no. 1, pp. 7–38, 2016. DOI: 10.1080/00396338.2016.1142085.

[19] NCTA - The Internet & Television Association. "Report: Where does the majority of internet traffic come from?" (2019), [Online]. Available: https://www.ncta.com/whats-new/report-where-does-the-majority-of-internet-traffic-come (visited on 04/27/2021).

[20] *Peter Pinckney v KDG Mediatech AG*. European Courts of Justice, 2013.

[21] J. R. Reidenberg, "Resolving conflicting international data privacy rules in cyberspace," *Stanford Law Review*, pp. 1315–1371, 2000.

[22] V. C. Rogers, "The history of chinese cybersecurity: Current effects on chinese society economy, and foreign relations," 2016.

[23] P. Scott, V. Soldatkin, and Reuters. "Russia's new cyber laws will fuel online crime, claims report." (2019), [Online]. Available: https://www.reuters.com/article/us-russia-internet-bill/russia-enacts-sovereign-internet-law-free-speech-activists-cry-foul-idUSKBN1XB4TF (visited on 05/22/2020).

[24] State Duma and Federation Council. "The criminal code of the russian federation." (1996), [Online]. Available: https://www.wipo.int/edocs/lexdocs/laws/en/ru/ru080en.pdf (visited on 05/22/2020).

[25] U.S. House, 94th Congress, 2541, *Copyright Act*, 17 USC § 1-8, 10-12. 1976. [Online]. Available: https://www.govinfo.gov/link/statute/90/2541?link-type=pdf.

[26] U.S. House, 99th Congress, 4718, *Computer Fraud and Abuse Act*, 18 USC § 1030. 1986. [Online]. Available: https://www.govinfo.gov/content/pkg/STATUTE-100/pdf/STATUTE-100-Pg1213.pdf.

[27] WIPO, *Berne Convention for the Protection of Literary and Artistic Works (ammended)*. Geneva, Switerland, 1971. [Online]. Available: https://www.wipo.int/treaties/en/ip/berne/.