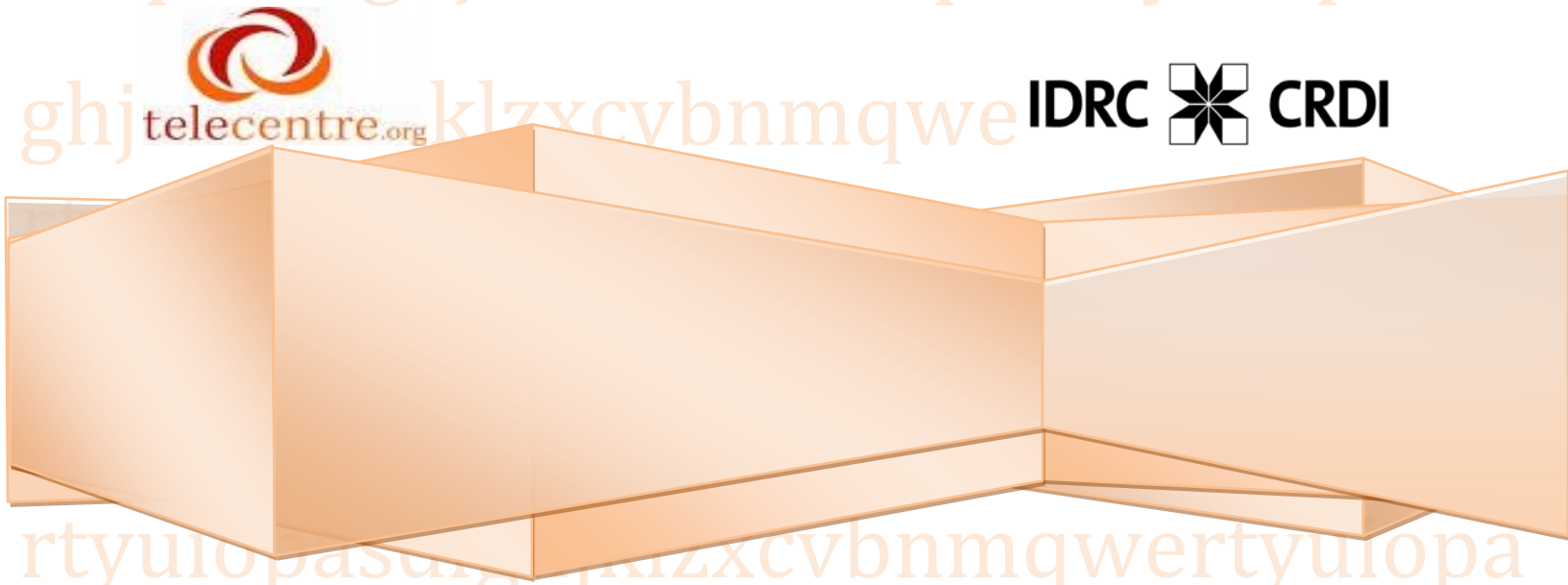
Building a Sustainable Framework for Open Access to Research Data Through Information and Communication Technologies

A Research Paper prepared for

Telecentre.org and the International Development Research Centre (IDRC) Canada

By

Gideon Emcee Christian



Acknowledgment:

The author acknowledges the funding for this research which was provided by the Canadian government through the International Development Research Centre (IDRC), Canada. The author is also grateful to Michael Clarke, Director ICT4D-IDRC, Frank Tulus, Program Officer with telecentre.org for supervising the project, Reiner Mauer of GESIS - Leibniz-Institute for the Social Sciences, Germany, Neda Gharani of Coriell Institute for Medical Research, Ellen Wright Clayton of Center for Biomedical Ethics and Society, Nashville, Tennessee, and Lisa D. Brooks of National Human Genome Research Institute, NIH.

Gideon Emcee Christian LL.M
International Development Research Centre (IDRC)
150 Kent Street
Ottawa, ON, Canada
K1P 0B2
+1-613-265-1300
gchristian@idrc.ca
www.idrc.ca

December 2009

Table of Content	

Executive Summary

The growth in information and communication technology (ICT) has brought about increased pace in information and knowledge exchange. This increased pace is being fuelled in large part by the open exchange of information. The pressure for open access to research data is gaining momentum in virtually every field of human endeavour. Data is the life blood of science and quite unsurprisingly data repositories are rapidly becoming an essential component of the infrastructure of the global science system. Improved access to data will transform the way research is conducted. It will create new opportunities and avenues for improved efficiency in dealing with social, economic and scientific challenges facing humanity.

Researchers, government, research funders, academic institutions, commercial entities as well as private individuals will continue to require greater access to research data from diverse sources. Such access will enable them to explore, experiment, test, create new knowledge and products. In addition, access to research data can result in more innovation with existing knowledge and products, ultimately contributing to our increased understanding of society. Access to research data is vital to the development of science and the human society. This underscores the need to utilize the opportunity provided by information and communication technologies (ICTs) in providing efficient, timely and cost effective access to research data. ICT is the bedrock of open access. It is central to research to the extent that it enables researchers to perform fundamental and applied research, build partnerships and international consortia, conduct experiments, manage data and communicate findings and results to colleagues and the general public in a timely and efficient way.

Despite the admitted benefits of open access to research data, the concept is still bugged by series of factors both legal and ethical which must be resolved in other to derive the maximum benefit arising from open access to data. This resolution will require the development of a sustainable framework to facilitate access to and use of research data by researchers, academics institutions, private individuals and other users.

As a starting point, this research paper examined the legal and ethical issues affecting open access to research data. Specifically, this research examined the diverse intellectual property rights regime relating to data in different legal jurisdictions, while acknowledging that these legal rules in most cases serves to restrict rather than enhance access to data. Afterwards, the research further examined various frameworks for enhancing open access to research data. Such frameworks included the open data contract, Open Content Licenses and the Open Data Commons. The pros and cons of each of these frameworks were also discussed in this paper. Nonetheless, the applicability of each framework in any particular case would depend on the nature and circumstances the case.

Ethical concerns relating to access to data in an open access environment usually arises in the form of privacy and confidentiality of personally identifiable data or information. This also presents another serious challenge to the concept of open data. While privacy has often been defined in terms of having control over “the extent, timing, and circumstances of sharing oneself (physically, behaviourally, or intellectually) with others", confidentiality relates to the treatment of information that an individual has disclosed (in the course of a research or survey) in a relationship of trust and with the expectation that it will not be disclosed or released to others in ways that are inconsistent with the understanding of the original disclosure without permission, or in a way that will be prejudicial to the individual. In the field of research, privacy and confidentiality ensures that information obtained by researchers about their research subject is not improperly divulged.

While on the one hand privacy advocates argue for restricted access to individualized data, open access advocates, on the other hand, continue to emphasize the need for greater access to research data. The latter view is based on the notion that greater and more in-depth access to research data increase the utility of the data whereas restricted access to certain elements of data will inadvertently reduce the utility of such data. Hence, open access to research data will inevitably require the development of a

sustainable framework capable of reconciling the conflicting interest between privacy and open access.[1] The core challenge in developing this framework is the ability to balance the *risk* of open data access with the *utility* associated with it. This research examined the framework adopted by the dbGaP Project in reconciling the privacy and confidentiality issues associated with the research data emanating from its project.

A notable observation made in the course of this research is the fact that while a growing number of pure science databases are adopting full open access policy, this has not been the case in other fields such as social sciences. One likely explanation for this trend is the fact that most of the pure science databases surveyed are usually the product of publicly funded collaborative effort (which is more common in pure science than in social science) between two or more organizations or institutions. These institutions have from the onset of the collaborative research adopted open access as their guiding principle. Hence open access databases are usually set up for the purpose of freely disseminating results or data from the research.

Additionally, it was observed that funders of such collaborative research played a crucial role in enhancing open access to the outcome of the research. This is particularly evident from the fact that most of the pure science open access databases examined in the course of this research contained data from research funded by the National Institute of Health (NIH) in the United States – an Institution that has adopted open access as a policy for its research funding. It suffices to state though that the research presented in this paper is exploratory in nature. The objective of this paper is to provide an overview of the various issues relating to open data, while at same time highlighting areas that should warrant further research.

[1] Such framework should also be useful in overcoming other challenges associated with open access to data such as costs and legal barriers.

I. Introduction

Robert Merton was famous for articulating the norm of "the Republic of Science". The norm has sometimes been redacted into five key words viz communalism, universalism, disinterestedness, originality and scientism. The concept of communalism emphasizes the corporative or collaborative nature of scientific inquiry.[2] Acquisition of scientific knowledge is a cumulative process that depends on the researcher's continuing ability to collect and share essential data.[3] It is now self-evident that production and accumulation of reliable scientific knowledge is essentially a collective as opposed to an individual endeavour, and the more open the process, the better for the scientific community and humanity in general, especially in the developing world where significant barrier to knowledge resources persists.

Economically developed countries spend substantial amount of public resources in research. These public investments generate large amount of data. Although some of the data generated from such research may not be of relevance in developing countries because of their subject matter and/or geographic focus, those aspect of the research data that have broad applicability as a global public good could be utilized in the developing countries if they are made openly accessible. Additionally, experience has shown that data results unconnected to developing world issues could be assembled into unexpected new results which are of much relevance to developing countries.

The research strategy developed by Rita Colwell is a case in point.[4] Utilizing large sets of data from difference continents on sea life, earth observation, historical epidemiology, DNA analyses, and social anthropology, she was able to demonstrate global disease

---

[2] Paul A. David *The Economic Logic of "Open Science" and the Balance between Private Property Rights and the Public Domain in Scientific Data and Information : A Primer* in "The Role of the Public Domain in Scientific and Technical Data and Information" Proceedings of a Symposium, Washington, DC: NAP, 2003

3 Committee on Issues in the Transborder Flow of Scientific Data, National Research Council "Bits of Power: Issues in Global Access to Scientific Data", Washington D.C. NAP (1997)

[4] Colwell, Rita (2002), "A Global Thirst for Safe Water: The Case of Cholera", Abel Wolman Lecture at the National Academy of Sciences, http://www.nsf.gov/news/speeches/colwell/rc02abelwolman/index.htm

patterns that, without the use of ICT tools and access to all the diverse data, would have remained invisible. This thus shows that even data collected for research not relevant to the developing world could be utilized for other research effort applicable to the region.

Digital media have become a dominant means to create, shape and exchange information. Advancement in information and communication technology (ICT) has made possible improvement in the quantity and quality of research data with the resultant rise in online data repositories/archives. Data is the life blood of science and quite unsurprisingly data repositories are rapidly becoming an essential component of the infrastructure of the global science system. Improved access to data will transform the way research is conducted, including research carried out by researchers in the South. It will create new opportunities and avenues for improved efficiency in dealing with social, economic and scientific challenges facing humanity.

Researchers, government, research funders, academic institutions, commercial entities as well as private individuals will continue to require greater access to research data from diverse sources. Such access will enable them to explore, experiment, test, create new knowledge and products, as well as to innovate existing ones, and ultimately to increased understanding of our society.[5] Access to research data is very crucial to the development of science and the human society. This underscores the need to utilize the opportunity provided by information and communication technologies (ICTs) in providing efficient, timely and cost effective access to research data. ICT is the bedrock of open access. It is central to research to the extent that it enables researchers to perform fundamental and applied research, build partnerships and international consortia, conduct experiments, manage data and communicate findings and results to colleagues and the general public.

The ICT revolution has also lead to increased collaborative research and sharing of research data in various fields of science. This effort is greasing the wheel of innovation and presumably, increased access to data through ICTs will accelerate innovation and

[5] Onsrud and Campell "Big Opportunities in Access to "Small Science" Data" *Data Science Journal Vol. 6 June 2007*

discovery thus creating new opportunities. With this change in data collection, management and dissemination has also come a new complex set of issues or challenges which must be resolved in other to derive the maximum benefits arising from the collaborative use of research data. This resolution will require the development of a sustainable framework to facilitate access to and use of research data by researchers, academics institutions, private individuals and other users.

II. Research problems

The three main research problems relating to open access to data is analysed in this paper:

i Ethical issue

The ethical issue that arises in relation to access to data often relates to privacy and confidentiality of information contained in data archives. Many research projects will usually involve the collection of personal data on human subject such as medical or genetic information, information relating to consumer behaviour etc. Although access to such information in many cases may be beneficial for further research, unrestricted access to sensitive data could be prejudicial to individuals, organizations or national interests as the case may be. Hence in advocating open access to research data, there is a need to strike a balance between the divergent interest of the research subject(s) and users seeking access to research data. Therefore, to what extent should the rules relating to privacy and confidentiality justifiably serve to restrict access to data?

ii. Legal issue

One of the primary issues that arise in the legal context of access to data relates to ownership rights in data. The collection, management and use of research data occurs within a legal context. Put simply, data is surrounded by legal rules.[6] These rules determine when ownership or propriety interest in data arises as well as their transferability. It suffices to state though that these rules differ across jurisdictions. In

[6] A Fitzgerald, K Pappalardo and A Austin, "Practical Data Management: A Legal and Policy Guide" (2008) <eprints.qut.edu.au/archive/00014923/01/Microsoft_Word_-_Practical_Data_Management_-_A_Legal_and_Policy_Guide_doc.pdf>. [Fitzgerald et al "Data Management"]

United States, the rule relating to proprietary right in data was discussed by the Supreme Court in *Feist Publications v. Rural Telephone Service.*[7] The U.S. Supreme Court held in that case that raw data, fact or information is not copyrightable (though collection of same is). A different set of rule is applied in Europe. The European Union through its directive on legal protection of databases created legal right in databases.[8] The effect of the rule is to confer proprietary right on the owner of a database with the consequential right to restrict the copying, use or dissemination of the content of the database.

The proprietary status of data is very important within the context of data access as much as it also posses a dilemma. If raw data are unprotected by any proprietary right, researchers or data custodians will be less willing to publish their raw data for fear of appropriation without attribution (as was the case in *Feist v. Rural*). On the other hand, if property right is vested in data (as is the case in EU), whereas data custodians may be willing to publish their data, copyright protection in such data will restrict subsequent use without the consent of the data author or owner. In this case, the process of negotiating right to use, copy or disseminate the data will be costly, time consuming and inefficient.

Hence there is need for a framework that will foster open access to data while at the same time balancing the rights of data authors with the interest of data users. This research will explore various frameworks for open access to data as well as the limitation of these frameworks in granting open access to data. It will also seek to raise issues for further research in this area.

iii. Open data and utility

Various empirical researches have established a positive correlation between open access to research publications and increased citation. If we accept that fact that the effort to

[7] 499 U.S. 340 (1991) <http://caselaw.lp.findlaw.com/scripts/getcase.pl?court=US&vol=499&invol=340> [*Feist v. Rural*]

[8] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31996L0009:EN:HTML> [E.U. Database Directive].

publish data openly will result in grater benefit to the scientific community, particularly to the social science research community and the developing country research community, what needs to be examined is whether same positive correlation also applies in the case of open data. In addition to examining the ethical and legal issues related to data access, this research, will seek evidence that points to the positive correlation between open access to data and data utility.

III. Research Objective

This research aims to explore the state of open data from the legal, ethical and utility perspectives. It is intended to provide a broad overview of the issues and potential framework for overcoming the challenges that can prevent increased access to research data. The research also aims to identify areas for further research in relation to open data frameworks.

IV. Methodology

Data Collection and Analysis

This research is carried out as an exploratory research and its main purpose is to contribute to the existing theories and knowledge around the potential benefits of open access to data, and the plausible frameworks that can lead to sustainable access to research data. The research draws heavily from an extensive review of existing literature, exploration of online databases as well as telephone interviews to obtain relevant data to explain the issues raises in the research. The literatures reviewed are basically scholarly articles on open access to research data. The contents are analysed in terms of their relevance to the legal, ethical and data utility issues examined in this research. The online databases are selected through online search. The databases are classified in terms of their subject matter (e.g. pure science, social science etc) and their access policies (e.g. open, closed or a combination of both). In the case of the telephone interviews, the format

adopted was the use of a semi-structured questionnaire. The advantage of this format was that it gave room for flexibility in the course of the interview. The results from the data collection are organized and presented as narrative case studies, and the conclusion draws on the findings to generate informed questions for further research.

Theoretical assumptions

This research is premised on the basis of the following theoretical assumptions (i) that open licensing models provides a better framework for sustainable access to research data; (ii) that the growing demand for open access to research data should be balanced with privacy and confidentiality of data subjects, and (iii) open access to data can increase data utility more widely for researchers in developed and developing countries.

V. Research Findings

What is open data?

Recently, there has been a growing philosophy which advocates that certain data should be made freely available to everyone without legal or technical restrictions to access. This philosophy has been coined into various terms such of "open access to data", "open data", "open research" etc. Although this philosophy is not novel, it seemed at one point in history to have gone into oblivion only to be resurrected by the rise in the open access movement. The open data phenomenon has today gained notoriety with regards to access to research data as well as government data. An intelligent discussion on open data or open access to data can hardly be made without reference to the concept of openness.

The term openness, or open, is often applied as a descriptive adjective appended in front of a variety of structures such as *open* source software, *open* education, *open* knowledge, *open* science, etc.[9] Smith et al conceives openness as a way of organizing social activities that favours universal over restricted access, universal over restricted participation, and

[9] Smith et al *Open ICT4D* (2008) IDRC

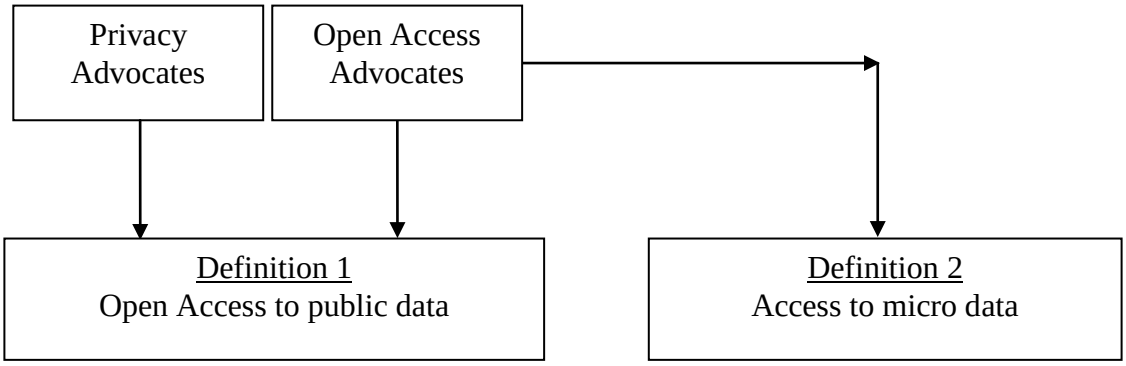
collaborative over centralized production.[10] Whether analyzed from social, scientific or whatever perspective or discipline, one common trend that runs with the concept of openness is the idea of removal of restriction, universal access, collaborative participation etc.

Going further from here, it suffices to state that the concept of open data can be viewed (and will in this research be viewed) from two (but not mutually exclusive) perspectives. First, the concept of open data can be viewed from the perspective of removal of access barriers to publicly available data. This idea relates to the removal of legal, financial and technical blockades such as copyright/licensing restrictions, access fees and lack of interoperability. Open data advocates argue that these restrictions serve to clog the wheel of scientific progress and are against communal good.

Another way we can look at the term is from the perspective of level of access. This second idea transcends beyond mere removal of access barriers to data but further transcends to in-depth access to other aspect of an openly accessible data e.g. part of an open data that may have been redacted or restricted from public access. This second perspective is characterized by not just access but the ***key point here is level or depth of access or in another word access to micro-data***. It is here that we find a serious clash of interest between open access advocates and privacy advocates.

[10] *Ibid*

Figure 1: The Open Data Debate



Whereas open access advocates and privacy advocates have met a consensus in terms of access to public data, this consensus breaks down when it comes to access to micro data. While open access advocates continue to clamour for greater access and rightly because greater access to data increases the usefulness and the quality of subsequent research, privacy advocates are worried (and rightly too) about privacy issues that arise from unrestricted access to data – especially sensitive data. Hence it is argued in this research that a sustainable framework for data access is achieved where these diverging interests reconcile.

Issues arising from increased open access to research data

There are many issues that arise in relation to collection, management and use of research data. Fitzgerald et al has identified these issues to include copyright, moral rights, patents, privacy and confidentiality.[11] Although I have briefly discussed some of these issues above, I will go further to discuss them in details below.

a. Privacy and confidentiality

Privacy in relation to access to data encompasses an individual's right to be free from excessive intrusion as well as the right to determine to what extent information relating to

[11] Fitzgerald et al "Data Management" *supra* note 5 at 2

him/her are shared with or withheld from others.[12] Thus, privacy has often been defined in terms of having control over "the extent, timing, and circumstances of sharing oneself (physically, behaviourally, or intellectually) with others."[13] Confidentiality relates to the treatment of information that an individual has disclosed (in the course of a research or survey) in a relationship of trust and with the expectation that it will not be disclosed or released to others in ways that are inconsistent with the understanding of the original disclosure without permission, or in a way that will be prejudicial to the individual. In the field of research, the principle ensures that information obtained by researchers about their research subject is not improperly divulged.

Privacy and confidentiality are very important considerations in the collection, management and dissemination of research data. Breach of privacy and confidentiality in relation to sensitive data about an individual could have serious consequences or repercussion.[14] Picture an open data regime whereby an employer, prospective employer or insurer could easily dig out health data about an employee, prospective employee or insurance applicant. Such disclosure (where it reveals adverse medical condition) could result in job loss, lack of offer or denial of insurance coverage as the case may be. No less worrisome is the risk of identity theft.[15] The seriousness of breach of privacy and confidentiality relating to personal data was evident in the U.S. case of *Remsburg v. Docusearch Inc.*, [16] where a stalker exploiting an online data service, acquired personal information about his victim. Using the information, he was able to trace her to her place of employment where he fatally shot her. Although this was a worse case scenario, it highlighted the seriousness of privacy and confidentiality concerns in data access.

---

[12] *See* Duncan et al. "Report of the Committee on National Statistics' Panel on Confidentiality and Data Access" , 1993. *Private Lives and Public Policies*, Washington, DC: National Academy Press, p. 23.

[13] IRB Guidebook, Part III.D, Department of Health and Human Services, Office for Human Research Protections.

[14] A research has revealed that 87% of Americans could be identified based simply on their birth date, gender and zip code. See Brian Bergstein "Research explores data mining, privacy" in *USA Today* 6/18/2006

[15] Lane, Julia and Schur, Claudia, "Balancing Access to Health Data and Privacy: A Review of the Issues and Approaches for the Future", (September 13, 2009). Available at SSRN: http://ssrn.com/abstract=1472736

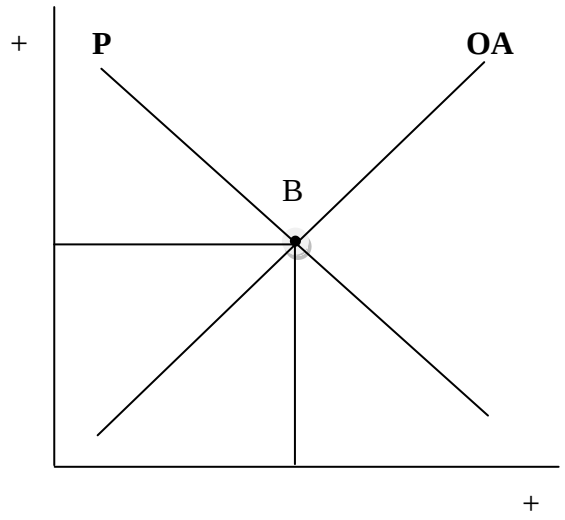
[16] 149 N.H. 148, 816 A.2d 1001

While privacy advocates argue, on the one hand, for restricted access to individualized data, open access advocates, on the other hand, continue to emphasis on the need for greater access to research data. They argue that greater and more in-depth access to research data increases the utility of the data while restricted access to certain elements of data will inadvertently reduce the utility of such data. As indicated by Lane & Schur:

> The more information that is provided and the more researchers that have access to the data, the greater the value of the analytical work that can be undertaken. In addition, the more transparent the access, the more likely it is that a body of knowledge will be developed around the dataset, expanding knowledge about the underlying data quality, the correct uses of the data, and the important data gaps. Finally, data access is essential to ensure that analytical work is generalizable and replicable, which is the essence of scientific endeavour.[17]

Open data and privacy advocates both represent two sides of a coin and bring intelligible arguments to the open data debate. Hence, the current argument admits too great a loss of data utility, on the one hand, and too great a risk to privacy and confidentiality on the other hand. The inverse relationship between open access and privacy is illustrated in the diagram below:

Figure 2: Inverse Relationship between Open Access to Data and Data Privacy



[17] Lane & Claudia, *supra* note 15

The diagram above illustrates the inverse relationship between open access (OA) and privacy (P). A framework characterized by increased privacy can result in decreased access and loss of data utility due to highly restrictive access. Conversely, a framework characterized by increased open access can result in loss of privacy and confidentiality along with the consequential risks stated earlier. Conceptually, a sustainable framework for data access will be represented by point B. The closer one is to point B in the diagram, the greater the likelihood that the conflicting interest between privacy and open access is reconciled. The core challenge in developing this framework is the ability to determine the balance point between the *risk* of open data access and the *utility* associated with it.

Hence going back to my earlier definition of open data, a sustainable framework is one which grants virtually unrestricted access to data that fall within the first definition, while at same time placing a measure of restriction to data within second definition - a measure of restriction which will leave the data open to access for research under terms and conditions reasonably sufficient to protect the privacy and confidentiality associated with the data. This is the type of framework that was adopted by the **d**ata**b**ase of **G**enotypes **a**nd **P**henotypes (dbGaP) – an open access genotype research database.

dbGaP is an open access data repository that was developed to archive and distribute the result of studies relating to the relationship between genotype and phenotype e.g. the genome-wide association studies, medical sequencing, molecular diagnostic assays etc. The nature of these studies involves accumulation of huge database of highly sensitive personal information. Whereas the project has the goal of making these data freely and widely accessible for further research, it is also faced with the conflicting obligation of upholding the privacy and confidentiality of the research subjects. To balance this conflicting interest, two levels of access were developed – open and controlled access. The open access allows for broad release of non-sensitive data, such as summaries of studies and the contents of measured variables as well as original study document text.

The controlled access is utilized in the release of individual-level genotype and phenotype data that have been de-identified.[18]

The generalized data in the database resides in public domain and are made available without any restriction. However where a prospective user needs to access individual-level dataset, additional measures are taken to protects the privacy and confidentiality of the data subjects. To access such data, the prospective user is required to file a request for data access. The request will state the specific dataset requested, and a brief description of the proposed research for which the data is requested. The proposed users will also give an assurance that the data will only be used for the proposed research, that data confidentiality will be respected, that no attempt will be made to identify individual study participants from whom the data was obtained, and that conclusion derived from the research will remain in the public domain without licensing requirement.[19] The problem with this precautionary measure is that it will usually result in transactions costs and delay arising from filling the necessary documentation for access. That notwithstanding, it provides a middle ground for the conflicting interests of privacy and open data advocates.

Another way of dealing with the privacy and confidentiality issues in access to research data is through the use of a prior informed consent. This will require that the researcher(s) obtain from prospective research participants a written consent to have their data made accessible in public archives and for use in further research. The need for a prior informed consent is usually given careful thought when the research involves human subjects; involve collection of private or personally identifiable information; and the data from the research is intended to be made publicly accessible, or accessible for wider distribution. In the case of the HapMap Project discussed below, individuals who agreed to participate in the Project were required to sign a consent form that granted permission for the DNA samples from the research to be made available for further research. With

[18] dbGaP Request Procedures to Access Individual-Level Data < https://dbgap.ncbi.nlm.nih.gov/aa/dbgap_request_process.pdf>
[19] dbGaP Overview < http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/about.html>

full consent having been granted by the participants, it was thus possible for the HapMap database to operate with a full open access policy.[20]

It should be noted though that there are noticeable differences between social science research and natural and health science research. These differences cut across the nature of the research, as well as information collected in the course of each research. Health science research in most cases entails collection of highly personal information such as DNA and other medical information which could provide further additional information about the research subject or which could be used to identify the research subject. Hence, the extent to which ethical issues arise in social and health science researches differs. It is often believed that ethical issues arise more in health science research than in social science research. Impliedly then, it should be expected that the framework for managing ethical concerns in each case should also differ depending on the field of sciences involved. These differences and the extent to which they affect the development of a viable framework in each case are an issue for further research which is not covered in this paper.

b. Copyright

Research data may be subject to copyright protection if they meet the proprietary requirement for such protection which, depending on the legal jurisdiction, may be based on the originality or the so-called 'sweat of the brow' doctrine. Where copyright is established in a literary work, moral rights will accrue if the author is a person (as opposed to a corporation). This right will vest on the author of the work the right to proper attribution and integrity of the work.

Issue relating to patent will arise where a collection of research data gives rise to or forms part of a novel process which may result in an invention, such data may give rise to patent rights. This is usually the case with respect to genomic databases. The issues of

[20] Even where the participants grant full consent to have the data made publicly accessible

privacy and confidentiality will arise where a database or dataset contains personal information the disclosure of which is regulated by law or the data are disclosed on the understanding that they will be held in confidence, such data must be protected against unauthorized access.

Where a database or dataset is protected by copyright or patent, it is not a proper subject for open access, otherwise legal liability may accrue for copyright violation. This though does not in any way imply that such database can not be made openly accessible. However, there are legal frameworks that could be developed to free such data from the shackles of closed or restricted access. Before examining this framework, it will be proper to examine the conventional intellectual property rights (IPR) regime as reflected in copyright and patents.

Copyright is a collection of exclusive legal rights that attach to a literary work when it is created. It is an aspect of intellectual property law that seeks to invest authors with monopoly right or control over their creative work. These exclusive rights include the right to authorize the copying and distribution or dissemination of the work. The general principle of copyright law is that copyright protects the *material form* in which ideas, facts and information are presented as opposed to the ideas, facts and information themselves.

The status of factual data is a complex legal subject. This complexity is the product of different legal standards applied in the determination of copyrightability in different jurisdiction. Since the concept of open access implies universal access devoid of any jurisdictional barrier, it is necessary to examine the various test of copyrightability across jurisdictions. The two basic tests applied in this regard are originality and ‘sweat of the brow’.

The basic concept of the originality standard is based on the rationale that the purpose of copyright law is to promote and protect creative works. Creativity in this case is judged by originality. It is on the basis of this principle that raw data are held ineligible for

copyright protection as they are presumed to be devoid of any creative effort. In contrast, the doctrine of 'sweat of the brow' dispenses with the requirement of creativity by protecting the labour and sweat of the compiler, without the use of his vision and aptitude. Hence mere mechanical and automatic task devoid of any creativity is copyrightable under this doctrine.[21]

The doctrine of sweat of the brow held sway in the United States until the Supreme Court decision in *Fiest v. Rural.*[22] The issue for determination in that case was whether a telephone company under a statutory duty to compile a database of all its customers for free distribution has proprietary interest or copyright in the database. The U.S. Supreme Court overruled the lower court decision which was based on the sweat of the brow doctrine, noting that the purpose of the copyright law was not to reward the efforts of persons collecting data or information, but rather to promote the progress of science and art.

It suffices to state that the U.S. Supreme Court was not here crafting a new rule. On the contrary, the court was actually re-stating an existing rule, which in effect, postulated that the prerequisite for copyright is originality (even though the originality threshold need not be high). The mere fact that considerable time, money or effort has been spent in collecting data is not relevant to the subsistence of copyright.

While mere data or fact cannot pass the test of originality as they are obvious facts, a compilation of data or fact may attract copyright protection where the compilation or collection is done in such a way as to give rise to creativity (e.g., the creative choice of what data to include or exclude, the order and style of presentation etc.)[23] However, even in cases where the compilation or presentation is protected by copyright and if the compilation relates to public domain data, the copyright protection will not extend to

[21] Hailshree Saksena "Doctrine of "Sweat of the Brow" <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1398303>
[22] Supra note
[23] *See* Justice O'Connor in *Feist v. Rural supra* note 6

deny the public from the use of the public domain data. Copyright should protect only the creative form of the presentation, but not the public domain data in the compilation.[24]

Protection of data in the European Union

The legal approach to data right takes a different approach in Europe. In 1996, the European Union enacted a directive on exclusive property protection of databases and compilation of information.[25] The E.U. Directive defined database as "a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means". While acknowledging that copyright protection remains the exclusive form of right for database authors, the Directive went further to state that "in the absence of a harmonized system of unfair-competition legislation or of case-law, *other measures* are required in addition to prevent the unauthorized extraction and/or re-utilization of the contents of a database"[26]

Consequently, the E.U. Directive adopts two approaches to data protection: the conventional copyright protection (based on originality) and the *"sui generis'* protection. Under the conventional copyright approach, the Directive vest copyright in "databases which, by reason of the selection or arrangement of their contents, constitute the author's own intellectual creation".[27] In this case, raw data are not copyrightable. Such data can be copied and re-used freely where accessible. However, a compilation of data which is a product of intellectual creation will be entitled to protection. The Directive tends to lean towards the standard of originality as found in the American jurisdiction. This is the only criteria for copyright protection under the Directive. The Directive goes further to expressly states that no other criteria shall be applied to determine their eligibility for that protection thus implying that the "sweat of the brow" standard does not apply to the test of copyrightability under the E.U. Directive.

[24] *Assessment Technologies v. Wiredata*, 350 F.3rd 640 <http://altlaw.org/v1/cases/1129733>
[25] See E.U. Database Directive *supra* note 7
[26] Article 6 E.U. Database Directive *ibid*
[27] Article 3.1

Another method of data protection adopted by the E.U. Directive is the *sui generis* approach. Suffice it to state that this approach is not founded on a real intellectual property right. The criterion for protection is economic as opposed to qualitative conditions.[28] Whereas the copyright standard discussed above protects the intellectual effort of the database creator, the *sui generis* approach is meant to protect the economic effort or interest of the database creators.[29] Additionally, the *sui generis* rule only provides for protection against unauthorized copying or use in whole or substantial part of the database. Hence the rule does not prevent non-substantial use or what have been termed as "fair use" or "fair dealing".

The principles of fair use or non-substantial copying will apply as an exception to the rules discussed above. However, these exceptions still fell short of solving the problem of data access. In the case of copyright protected database, copyright could be an obstacle for reproducing the data extracted from the database without prior authorization or for integrating the protected data into another database.[30] Where the *sui generis* rule applies, there will be a breach of the database producer's right in the case of a whole or substantial copying of the database. This is especially so even where the nature of the scientific research warrants substantial copying of the relevant data. What amounts to substantial or quantitative copying is a question of fact to be determined by the circumstances of each case.

The framework represented in the two scenarios above will inevitably require the user(s) (or prospective user(s)) to enter into contractual licensing with the right holder in other to avoid liabilities that may arise from a breach of the proprietary right in the data or database. However, if each user of the database were to enter into contract individually with the database author each time they require access to the database, the progress of scientific research will be retarded and transaction cost for data access will be high. There is need for a licensing framework that will accelerate access to research data.

---

[28] Bertrand Warusfel "Legal protection of databases in Europe and public scientific research" <www.epip.eu/papers/20031124/200411.../EPIP%20Warusfel.ppt>. [Warusfel, "Protection of databases"].
[29] Article 7.1
[30] Warusfel, "Protection of databases" *supra* note 27

Frameworks for open data

As has been stated earlier, copyright comes into existence once a copyrightable dataset is created. The author thereon is automatically invested with all rights arising from the dataset. Unlike patent, there is no formal registration required for copyright. Suffice it to state that it is not in every case that the author of a copyrightable data or database would want to exercise or rigidly assert his/her legal right especially in relation to copying and subsequent use of the dataset. In fact, the rightholder may have the intention of making the data freely available to other researchers. This is usually the case in open collaborative research projects. There are various frameworks that could be used to circumvent rigid IPR rule thus making research data openly accessible over the Internet. The following are some of these frameworks:

a. Open data contract:

Although a copyright holder is entitled to exercise all the right that comes with copyright ownership, the right holder is equally entitled to grant permission or license to others to exercise some or all of these rights. Such permission could take the form of a contractual license. There are two types of licenses: exclusive and non-exclusive licenses. An exclusive license permits the licensee to the exclusion of any other person (including the copyright holder himself) to exercise the rights. Exclusive licensing is contrary to the idea and philosophy of open access because it seeks to restrict rather than promote access. A non-exclusive license on the other hand provides for the right to exercise one or more of the copyright owner's right but not to the exclusion of the copyright owner or other prospective licensees.[31]

Hence the term 'open data contract' is used in this research to refer to a non-exclusive contractual framework, whereby a right-holder in data or database grant access to the use or re-use of the data, but subject to the acceptance of terms and conditions precedent to the access. In the cases of online databases, the terms of the contracts or access are usually indicated on a web page (in most cases in the form of a click-wrap agreement). In

[31] Fitzgerald et al, *OAK Law Project Report No. 1* (2006) p44

a clickwrap agreement, the user is required to accept the terms and conditions presented before them on the website (usually by ticking a box which states "I agree to the terms…"), before access is granted to the data. This sort of framework was applied in the International HapMap Project where it was used in enabling open access to research from the project while at same time pre-empting parasitic patenting – a situation whereby data obtained from an open access database is used to file a patent applications that block other users' access to or use of the same data.[32]

International HapMap Project: The International HapMap Project - a scientific research project which was lunched in 2002 was a multi-country effort to identify and catalogue genetic similarities and differences in human beings. Using the information in the HapMap, researchers will be able to find genes that affect health, disease, and individual responses to medications and environmental factors. The Project was a collaborative effort among scientists and funding agencies from Japan, the United Kingdom, Canada, China, Nigeria, and the United States. In terms of access to output, the intention of the project was to make data generated from the research openly accessible to other researchers.[33] Such researchers were also encouraged to publish results based on combining HapMap data with data from other projects, particularly in efforts to find genes affecting a disease or a drug response.

The project follows the data release principle of a "community resource project" defined as "a research project specifically devised and implemented to create a set of data, reagents or other materials whose primary utility will be as a resource for the broad scientific community."[34] To ensure full access to its data, the project adopts an access policy based on acceptance of the terms of its access contract. The terms and conditions of access were designed in such a way as to ensure that the data generated by the project will continue to remain available to all users. The terms provides that users may "access

[32] See Rebecca S. Eisenberg, *Genomics in the Public Domain: Strategy and Policy*, 1 Nature Review Genetics 70, 73 (2000). The author describes the danger of private firms utilizing public data to enhance their own private data thus enabling them to file patent application.
[33] International HapMap Project <http://www.hapmap.org/thehapmap.html.en>
[34] International HapMap Project, Registration for access to the HapMap Project Genotype Database <http://www.hapmap.org/cgi-perl/registration>

and conduct queries of the Genotype Database and copy, extract, distribute or otherwise use copies of the whole or any part of the Genotype Database's data as [they] receive it, in any medium and for all (including for commercial) purposes" provided that they do not "restrict the access to, or the use which may be made by others of, Genotype database or the data it contains."[35] In addition, users may 'disclose data obtained as a result of …access to and use of the Genotype Database only to other parties who have first confirmed…in writing that they too are licensees under the terms of the International HapMap Project Public Access License and so are bound by equivalent terms and conditions."[36] This is somewhat similar to the share-alike condition found in the Creative Commons licenses.

Therefore, the users must agree not to reduce others' access to the data and to share the data only with others who have made the same agreement. Clicking the "I accept" button binds the user or researchers and their employers to the terms of the license. Rather than having to click the "I Accept" button each time the user visits the site, the user is required to complete a registration on the first visit by choosing a user name and password. During the course of the registration the user is presented with opportunity to accept the term of the access contract via a clickwrap. By entering the username and password on subsequent visits, the user thus re-confirmed his/her acceptance of the terms and conditions of access.

Although the clickwrap policy was designed to pre-empt parasitic patenting, there were doubt in some quarters as to its ability to accomplish that purpose. According to Prof. Opderbeck "Nothing in the Patent Act would suggest that a patent could be invalidated because some of the underlying data was derived from a database in violation of the database's term of use. Thus, it is unlikely that the clickwrap license provided in the HapMap project provided the HapMap Consortium any meaningful remedy once a patent has been filed."[37] Notwithstanding this criticism, there was no reported case of violation

[35] ibid
[36] ibid
[37] David W. Opderbeck, *The Penguin's Genome, or Coase and Open Source Biotechnology*, 18 HARV. J.L. & TECH. 167, 199 (2004).

of the access policy or parasitic patenting.[38] However, the criticism raises an area of concern to open access namely - the challenge posed by parasitic patenting on open access to data.

Although the objective of the licensing conditions in open data contract may be to keep the data open and prevent other users from taking any step to keep the data out of public domain, such conditions may also act as an obstacle to full data integration especially in the case of integration between databases that have different contractual terms and conditions.

Open Content Licenses

Although the Internet provides an efficient means for effective distribution of information, merely uploading research data onto the public Internet (even in the absence of any warning or condition restricting use) does not in anyway derogate from or call into question the author's exclusive right, and neither does it at same time provide any information as to the authors intention to make the data available for open access. Such act will likely contribute to confusion and uncertainty, because potential users are left in doubt as to whether or not they are permitted to use the data (and the extent of their permissible use).

Open content license is a contractual agreement under which data, otherwise protected by proprietary rights, are made available openly for use or re-use subject to terms and conditions specified by the rightholder. Some of these conditions are discussed below under Creative Commons – a popular form of open content license. Open content licensing provides a very flexible framework for copyright holders because it makes their data openly accessible over the Internet for use or re-use. This framework has the potential to drastically reduce the transaction time and cost involved in negotiating access

[38] Donna M. Gitter, *Resolving the Open Source Paradox in Biotechnology: A Proposal for a Revised Open Source Policy for Publicly Funded Genomic Databases,* Houston Law Review 43, 1476-1521. In December 10, 2004 the Consortium announced the end of its licensing policy. Thereafter, all of the Consortium's data was made publicly available without restriction. This change in policy was apparently borne out of the fact that various scientific advances had diminished the fear of parasitic patenting that was likely to result from unrestricted and unconditional access to the data. See NIH New Release "International HapMap Consortium Widens Data Access" <http://www.genome.gov/12514423>

right since each user does not need to seek individual permission to use or access the data or dataset. The license would usually among other things indicate the conditions for use as well as restrictions, if any. The open content licensing model has, in the last few years, been increasingly utilized in granting access to copyright protected materials over the Internet.

Also related to the open content licensing is the Open Database License (ODbL), which is a license agreement intended to allow users to freely share, modify, and use a database subject to terms and conditions (if any). It is a license for database users to act in a certain way in return for right of access to the database. The ODbL is much more relevant in the European Union jurisdiction where database rights apply.

Benefits of Open Content Licensing

Open content licensing is an appropriate access framework for organizations and individuals that owns copyright or broad copyright license in research data, and want to make the data openly accessible for use or re-use.[39] Some of the benefits of open content licensing identified by the Open Knowledge Foundation includes: (i) it allows others to circulate the license work freely and widely; (ii) not forcing others to seek permission every time they wish to use or circulate a copy of the licensed material, which can be time consuming; (iii) encouraging others to continuously add value to the work; and (iv) encouraging others to create new works based on or derived from the original work.[40]

Types of Open Content Licenses

Various standard-form open content licenses exist which could be utilized in granting access to digital contents including research data. The GNU Free Documentation License (GNU FDL) is a standard-form copyleft license designed by the Free Software

[39] See Fitzgerald et al "Data Management" *supra* Note 5

[40] Open Knowledge Foundation, *A Guide to Open Licensing,* <http://www.opendefinition.org/guide?action=print>

Foundation for the GNU Project.[41] The license is designed along the copyleft principle which means that derivative work based on the licensed source material must be made available to other users either on same or similar terms.[42] The license is suitable for works whose purpose is to create a set of instructions or a reference material.

Another prominent model of open content license is the Creative Commons license. The Creative Common license is a set of non-exclusive licenses (represented by distinct and identifiable graphic icons) which allows a right holder to grant permission in advance to the world at large with respect to the copyright material. The license seeks to strike a balance between the traditional copyright standard of "all rights reserved" and the public domain "no right reserved". The license allows a right-holder to dictate how others use his work, and for what purpose they may use it. Rather than having to grant permission to myriads of users on various occasions, the Creative Commons license allows a copyright holder to grant a one-time only permission. The license which comes in the form of a standard set of icons representing the licensing conditions is attached to the copyright material(s) so that it moves with it.[43] Various conditions imposed on Creative Commons license includes:

> **Attribution (BY)** – This condition applied in most Creative Commons licenses. It gives the user the right to copy and distribute the work subject to proper attribution to the right holder.
>
> **Non-Commercial (NC) –** restricts the use of the work for any commercial purpose.
>
> **No Derivative (ND) –**The user right is limited to copying, distributing, displaying or performance of the original work. No derivative works based on the original is permitted.

[41] The GNU Project was a software development project which was intended to design "*a sufficient body of free software [...] to get along without any software that is not free." See* The GNU Manifesto <http://www.gnu.org/gnu/manifesto.html>

[42] *See* GNU Free Documentation License <http://www.gnu.org/copyleft/fdl.html>

[43] Fitzgerald et al "Data Management" *supra* note 5

Share-Alike (SA) – the user may make and distribute derivative work but only on the condition that the derivative work is subject to licensing condition(s) identical or similar to the license that governs the original work.

The copyright holder is free to attach any of these conditions or a likely combination of the conditions to the license.[44] Additionally, works licensed under Creative Commons are protected by and subject to applicable copyright law. Hence, a right holder can only use the Creative Commons license in relation to rights conferred on him, but he/she may not use it to affect legal right conferred on others such as those relating to fair use or fair dealing. Creative Commons license has thus provided a very flexible licensing framework for data sharing. The framework was applied in the Universal Protein Resource (UniProt) project – a collaborative scientific research project that provides freely accessible resources on protein sequence data.

Universal Protein Resource (UniProt) project: UniProt project - the world's largest collection of protein database is a collaborative project between the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR) and funded by the National Institute of Health (NIH). The aim of the project is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.[45]

To advance its purpose of providing free access to arrays of data on protein sequence, the project adopts the Creative Commons Attribution-No Derivatives License.[46] This means that the user is free to copy, distribute, display and make commercial use of the databases subject to "attribution". The "No Derivative" element of the license implies that the user may not alter, transform, or build upon the work in the database without first seeking permission to that effect.

[44] For example, the free online encyclopedia Wikipedia uses the Creative Common Attribution and Share-Alike license. Note though that some of the conditions are incompatible and hence may not be combined for e.g. No Derivative and Share Alike.
[45] <http://www.uniprot.org/>
[46] <http://creativecommons.org/licenses/by-nd/3.0/>

c. Open Data Commons

Although the Creative Commons license and open content contracts/licensing has been applied in effecting open access to research data, recently the Science Commons issued a recommendation advising against further use of the contractual framework or license for informational database.[47] With regards to the application of open data contracts, Science Commons views such application as not only a threat to innovation and productivity, but it also restricts scientific freedom.[48] According to Nguyen, licenses and contracts not only impede research but also enable the data provider to exercise "remote control" over downstream user of data, dictating not only what research can be done, and by whom, but also what data can be published or disclosed.[49]

Open content contract and license are both subsistent on copyright laws. This means their validity and applicability are determined by the applicable copyright law, and as we have seen, the standard for copyright protection differs from one legal jurisdiction to another. A dataset may attract protection in Europe but lose the same protection in the American jurisdiction. In the case of a valid contract and license, the contractual terms and conditions may carry legal weight in jurisdictions that provide copyright or *sui generis* protection while in another jurisdiction without any protection, such terms and conditions are at best mere cosmetic surplusage. Access contract or license granting or restricting right of use is of no effect where the data or dataset is not protected by any law. In this case, there is no proprietary right over the data and hence the maxim *nemo dat quod non hebet* (no one [can] give what one does not have) applies. Hence data not protected by copyright or *sui generis* laws should be assumed to be in the public domain and hence unworthy candidates for contractual or Creative Commons licenses.

---

[47] Science Commons "Database and Creative Commons" <http://sciencecommons.org/resources/faq/databases/>

[48] Thinh Nguyen *Freedom to Research: Keeping Scientific Data Open, Accessible, and Interoperable* <http://sciencecommons.org/wp-content/uploads/freedom-to-research.pdf.>

[49] ibid

Another criticism of the Creative Commons license especially in the area of pure science is the fact that life scientists need to integrate data across disciplines to comprehend diseases and discover cures. Such data integration will only be possible where all the databases share identical licenses. For example, a dataset subject to a Non-Derivative term may not be integrated with another dataset subject to Share-Alike term since it will be difficult to comply with both terms simultaneously.

To overcome these problems, a more radical form of open data concept has been conceived – open data commons. This concept requires rightsholders to dedicate their works to the public domain for the benefit of the public, and relinquish all rights in the work, whether copyright or *sui generi* rights. Once these rights have been relinquished, the rightsholder has no further legal right in the work – not even the right to attribution. Prominent forms of open data commons licenses are the Public Domain Dedication and License (PDDL) by the Open Data Commons (ODC), Science Commons Data Protocol by Science Commons as well as the Creative Commons Zero (CC0) licence. According to Nguyen, the principle behind the Common Data Protocol is that the solution to the problem of data access is "to return data to the public domain by relinquishing all rights, of whatever origin or scope, that would otherwise restrict the ability to do research."[50]

The beauty of this concept lies on the fact that it resolves the problems associated with traditional Creative Commons requirement for attribution which was a basic feature of all CC licenses. It also eliminates the ability of the data provider to exercise "remote control" over downstream use of data. In addition, it eliminates the problem of data integration caused by different licensing options under the Creative Commons license. The major problem with the concept, however, lies on its requirement that data providers divest themselves of all rights to the data. The "no rights reserved" element of the concept, which is its main attraction, also seems to be its major undoing. Apparently, many data providers are genuinely concerned about protecting the integrity of their data or projects. Most importantly, even when data providers are disposed to granting free access to their data, they are, in most cases, equally concerned about proper attribution.

[50] ibid

The open data commons does not posses any contractual or license condition to enforce either of this. Hence, in the world of open data commons, the protection of integrity of work as well as attribution will be subsistent on the community norms and ethics as opposed to contractual or license obligations. To what extent this concept will be accepted by the scientific community is a proper subject for further empirical research.

Factual Trends in Open Data

Notwithstanding, the benefits of open data in scientific research and innovation especially in performing new analysis, critical validation etc, it was observed in the course of this research that the culture of open data or data sharing is still bugged by the difficulty of acceptance in many fields of endeavour. Most data archives today still cling to the old-discredited, closed-access or restrictive access policies with its limited benefits. Below are some examples of the limited, but growing number of open data archives or repositories explored in the course of this research.[51] The observations made are annotated below.

[51] Telephone interviews were also conducted with staff of some of the archives or repositories.

	Archive or Database	Organisation (s)	Level of access	Type of Data	URL
1	GenBank	National Center for Biotechnology Information *(*NCBI*)*	Full Open Access	Pure science	http://www.ncbi.nlm.nih.gov/Genbank/
2	The protein data bank	Research Collaboratory for Structural Bioinformatics (RCSB)	Full Open Access	Pure science	http://www.rcsb.org/pdb/home/home.do
3	Universal Protein Resource (UniProt)	European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR).	Full Open Access	Pure science	http://www.uniprot.org/
4	HapMap	National Human Genome Research Institute (NHGRI)	Full Open Access	Pure science	http://hapmap.ncbi.nlm.nih.gov/
5	dbGaP	National Center for Biotechnology Information (NCBI)	Full Open Access (restricted access to data containing personal information)	Pure science	http://www.ncbi.nlm.nih.gov/
6	Afrobarometer	Department of Political Science, Michigan State University,	Full Open Access (Data release delayed for 1 year)	Social science	http://www.afrobarometer.org/data.html
7	Inter-University Consortium for Political and Social Research (ICPSR)	Institute for Social Research at the University of Michigan	Open Access and Restricted Access	Social science and others	http://www.icpsr.umich.edu/icpsrweb/ICPSR/
8	ZACAT - GESIS Online Study Catalogue	Leibniz-Institute for Social Science	Open Access and Restricted Access	Social science and others	http://zacat.gesis.org/webview/index.jsp
9	Economic and Social Data Service (ESDS)	Economic and Social Data Service	Restricted Access	Social science and others	http://www.esds.ac.uk/international/access/dataset_overview.asp#desc_CRONOS
10	Ontario Data Documentation, Extraction Service and Infrastructure Initiative (odesi)	Ontario Council of University Libraries	Restricted Access	Social science	http://www.odesi.ca
11	ADPSS Socio Data	Dept. of sociology and Social research, University of Milan, Italy	Restricted Access	Social science	http://www.sociologiadip.unimib.it/sociodata/eng/index.php?w=home
12	UK Data Archives	UK Data Archives	Restricted Access	Social sciences and humanities	http://www.data-archive.ac.uk/Introduction.asp

It could be seen from the table above that while a growing number of pure science databases are adopting full open access policy, this has not been the case in other discipline, such as in the social sciences. One likely explanation for this trend is the fact that most of the pure science databases surveyed are usually the product of publicly funded collaborative effort (which is more common in pure science than in social science) between two or more organizations or institutions. These institutions have, from the onset of the collaborative research, adopted open access as their guiding principle. Hence, the databases are usually set up for the purpose of freely disseminating results or data from the research.

Additionally, it was observed that funders of such collaborative research played a crucial role in enhancing open access to the outcome of the research through their research funding policies usually embedded in their funding agreements. Most of the natural and health science open access databases in the table above contain data from researches funded by the National Institute of Health (NIH) in the United States – an Institution that has adopted open access as a core term for the its research funding. Hence funding agreements will usually contain a clause to the effect that the result or outcome of the research will be made available in open access.

The only social science database in the table above which operates a full open access policy is the Afrobarometer – a public opinion research group, although in the later case, the data are usually delayed for one year before release in the public domain. The reason for the delay was to give its researchers priority in terms of publishing research work based on the data.[52] Hence, where the research funder(s) adopts an open access policy with regards to the research it/they funds, there is the likelihood that the most of the data emanating from the research will remain in the public domain.

Nevertheless, the situation is usually different where the individual researcher(s) are given the leeway to decide on access policy with regards to their research. A case

[52] Until recently, the data was delayed for two years before release. However, following a request from the core funders, the delay period was reduced. This goes further to show the impact funding institutions could have in effecting open access to research data.

illustrative of this is the ZACAT - GESIS Online Study Catalogue.[53] The database has four different categories of access and researchers wishing to deposit data are allowed to choose the category in which they would like to have their data deposited. Below are the various categories of access provided by the repository, followed by the number of data available under each category:

Access Class	**Description**
Access class O	Full open access, free access for everybody
Access class A	Free for academic research & teaching
Access class B	Free for academic research & teaching, (if results won't be published; in case of publication, permission from depositor necessary).
Access class C	Only released for academic research & teaching after the data depositor's written authorization.

ZACAT - GESIS Online Study Catalogue

While majority of the data depositors would want access to their data limited to academic research and teaching (class A), very few on the other hand are willing to have their data made fully open in the public domain (class O). It should be noted that while the data in

[53] This is a social science data repository run by the Leibniz-Institute for Social Science. <http://zacat.gesis.org/webview/index.jsp>.

this archive are deposited by individual academics, government, independent research organizations, most of the data come from publicly funded research. If this is the case, why is it that only a handful of these publicly funded research data are made available under Class O? It was discovered that the reason for is basically because there is no mandatory requirement for such data to be made available in open access.[54] This further substantiates my earlier assertion that research funding bodies play an important role in making data from their funded projects openly accessible. Passing such responsibility to individual researchers will not be in public interest.

The Relationship between Openness and Utility

According to Thomas Kuhn, scientific revolution or paradigm shift occurs when a sufficient body of data accumulates to overthrow the dominant theories humans use to frame reality.[55] In the field of research, this revolution have been witnessed in the open access movement and its sets of principles designed to ensure that scientific data remains open, accessible and interoperable.

As we have seen above, there are various frameworks that could be utilised in granting open access to research data. Some of these frameworks carry with them certain restrictions or conditions (precedent or subsequent). For example, the use of a clickwrap agreement requires a data user to accept the terms of access and dissemination before being able to access the data. The terms in the agreement may limit the ability of the researcher to re-use or disseminate the data (e.g. non-derivative or share-alike conditions).

The concept of openness does not necessarily imply that terms of use cannot be attached as condition for access. What is required is that the condition(s) of access should not carry financial obligation or be so onerous as to make access and re-use practically difficult or frustrating. Many research studies have documented the direct relationship or

[54] Telephone interview with Reiner Mauer GESIS of Leibniz-Institute for the Social Sciences on 24th September, 2009

[55] Thomas Samuel Kuhn *The Structure of Scientific Revolutions* University of Chicago Press, 1962.

correlation between open access to research publications and citations.[56] That open access results in increased citations is now obvious from various empirical studies. The issue that remains to be researched is weather this correlation could also be found in the case of open access to data - does greater openness result in greater utility with respect to research data?

The problem in answering this question lies in the difficulty in developing an appropriate framework for measuring the true impact or utility of open data. Suffice it to state from the onset that most of the databases explored in the course of this research do not have adequate framework in place for measuring the resulting use of data obtained from their archives or repository. Hence, this is an aspect of open data that needs to be researched further.

Some of the various methods currently used in weighing such utility are limited in its effectiveness to measure the resulting use on further research. These methods include tracking the number of data downloads, online visits to the archives, or publications from the data obtained from the archive or database. Logically, it could be argued that an increase in any of these could lead to greater utility - but not necessarily. Although it is much easier to determine the number of visits to or downloads from an internet archives or databases, using either of this as a yardstick to measure utility could be very misleading. The higher number of visits or downloads from a database or archive, although quite impressive, may not necessarily give a true picture of the utility of the data to the visitor or downloader. The number of visits or downloads gives us little (if any) insight as to what use was made of the data.

[56] James A. Evans and Jacob Reimer "Open Access and Global Participation in Science" (2009) *Science* Vol. 323. no. 5917, p. 1025 < http://www.sciencemag.org/cgi/content/abstract/323/5917/1025>. Gunther Eysenbach "Citation Advantage of Open Access Articles" PLoS Biology (2006) 4:5 p 0692 – 0698 <http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.0040157>. Steve Hitchcock "The effect of open access and downloads ('hits') on citation impact: a bibliography of studies" <http://opcit.eprints.org/oacitation-biblio.html>.

However, the position is different when it comes to the number of research publications emanating from data which were obtained from an open database or archive. In the latter case, the user has found the data useful, analyzed them and utilized them thus adding to the data utility. Therefore, subsequent publications seem to be a more appropriate indicator than the others identified above.[57]

The problems with this framework though is the fact that it is often very difficult for open access databases or archives to keep an accurate record of the number of publications resulting from the use of their data. It was observed in the course of this research that even though most of the open access data archives surveyed (e.g. Afrobarometer) have a policy requiring users of their data to deposit a copy of their publications with the archive, there is hardly a compliance with the policy. Furthermore, the archives generally have no means for strictly enforcing this policy requirement. The only possible and fairly accurate method of keeping track of such publication would be individualized online search of journals and publications. This was the method that was adopted by the HapMap project for keeping track of publications generated from the use of its data samples.[58] Further exploration of the HapMap project in this regards provides a clue as to the likely relationship or correlation between open data and utility.

Although at its inception the HapMap project was conceived on the principle of openness, with regards to the dissemination of the project's research data, it did not adopt full openness at this initial stage. Rather it adopted a defensive (openness) strategy, whereby users are obliged to accept the terms of a clickwrap agreement before they could access the data. The terms of the clickwrap agreement required the user to agree not to prevent others from using the data and to share the data only with those who had agreed to the condition. Curiously, the idea behind these conditions was not to limit openness but rather to "ensure that all these important resources were kept in the public domain".[59]

[57] It should be noted that data utility is not limited to further research publication alone. In the case of pure science research, data could be utilized for other purpose not related to publications such as drug development.

[58] Searches are usually conducted in the PubMed to keep track of research publications based on the HapMap project data.

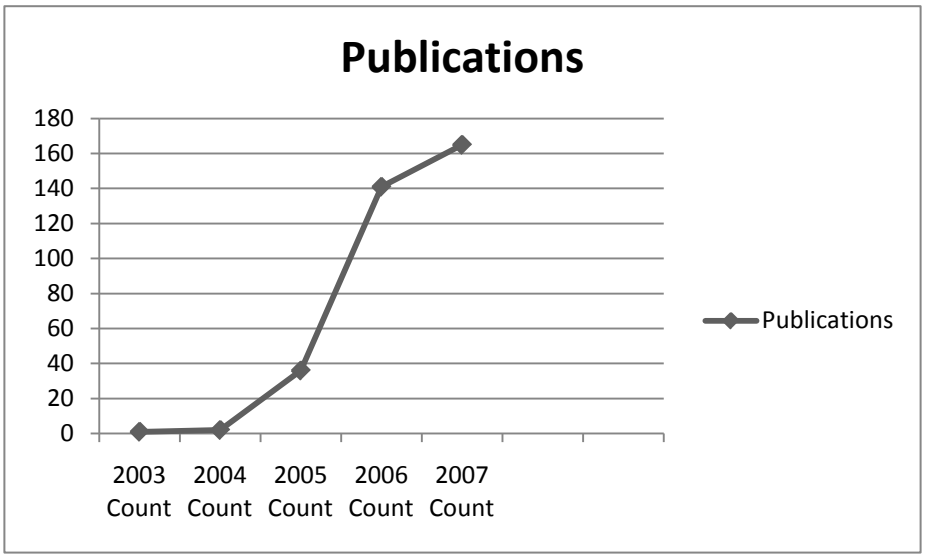
[59] Press Releases: 13th December 2004: "International HapMap Consortium Releases All Data to the Public

There were initial concerns that other users might combine the Project's genotype data with their own data to generate patentable invention, thus using the patent to prevent other from using the Hapmap data.

Even though the motive behind the clickwrap agreement was to keep the data in the public domain – which was in line with open access principle, the agreement was also counter-productive because it restricted the level of openness by making it practically impossible (legally) for data from the HapMap Project to be integrated into major public databases. Hence, the Hapmap database could only be integrated with other databases that carry condition(s) similar to the HapMap database. The effect of this level of openness was limited integration and use of its data. As will be shown later in the graph below, this period was also characterized by limited data utility – the number of research papers resulting from the project's data were very limited.

In December 2004, the Project developed the view that the problem of parasitic patenting was no longer obvious. That being the case, the click-wrap agreement was discontinued thus giving way to full open access. This development resulted in ability to integrate the data with other genomic databases, as well as greater access to the Hapmap database. Comparing the two periods under consideration i.e. the pre-clickwrap and its aftermath will provide useful insight on the relationship between openness and data utility.

HapMap Will Help Identify Genetic Contributions to Disease" <http://www.sanger.ac.uk/Info/Press/2004/041213.shtml> .



The graph above provides information relating to the number of (tracked) research publications based on the data from the HapMap Project. The statistics shows that the number of research publication emanating from the Project during the period when the clickwrap agreement was in force (between 2003 and 2004) was very low. However, when the clickwrap agreement was discontinued in 2004 thus resulting in full openness, the result was a sudden rise in the number of research publications based on the data from the HapMap project. While this is only one case example, the evidence from this study seemed to reasonably suggest that full open access to data will likely result in greater utility than closed access. Notwithstanding, the extent to which full open access impacts on data utility remains an area for further comprehensive research.

Conclusion

The general realization of the importance of open access to data has resulted in the concept of open data gaining prominence in relation to research and innovation. If the full benefit of open data is to be effectively harnessed, however, there is need to resolve the myriads of issues which affects or restricts the ability of various fields of sciences to share data freely in an open access environment.

This research paper has explored various frameworks which could be utilized in providing open access to research data, thus freeing data from the shackles of rigid intellectual property rights which seeks to restrict access to data. Ethical issues relating to open data especially in the area of privacy and confidentiality was also discussed. But there is much more that needs to be done. The adoption of the concept of open data will depends on the extent to which issues relating to open access to research data are resolved. Hence, this research has sought to identify issues relating to open data which merits further research such as the degree to which ethical issues relating to data access differ in various fields of research, as well as the extent to which these differences affect the development of a viable framework for data access.

Secondly, although there are various frameworks for granting open access to research data, each with its pros and cons, there is a need for further research aimed at establishing practical lessons and best practices that can assist individuals and organizations contemplating or willing to adopt the open data concept, particularly for choosing the appropriate framework that suits their particular circumstances. Another important area that also merit further research relates to the investigation of the relationship between open access and data usage. Although various research has established a positive correlation between open access (generally) and usage, the extent to which such positive correlation exists in respect to open data and usage or utility is still to be established. Such research will no doubt go a long way in advancing the argument for open access to data.

Open data scholarship requires a great deal of future development, and this research paper simply presents a starting point – identification of future research agenda. Based on the findings and practical considerations for future research stated above, the following research questions would be relevant for future research: What is the yardstick for determining the most appropriate framework for open access to data? To what extent should ethical concerns appropriately restrict access to research data? What is the empirical correlation between openness and data utility? Related to this is the need for

future research to address the potential utility of open access to data from the perspective of researchers from the developing world.

Literatures

Paul A. David *The Economic Logic of "Open Science" and the Balance between Private Property Rights and the Public Domain in Scientific Data and Information : A Primer* in "The Role of the Public Domain in Scientific and Technical Data and Information" Proceedings of a Symposium, Washington, DC: NAP, 2003

Committee on Issues in the Transborder Flow of Scientific Data, National Research Council "Bits of Power: Issues in Global Access to Scientific Data", Washington D.C. NAP (1997)

Colwell, Rita (2002), "A Global Thirst for Safe Water: The Case of Cholera", Abel Wolman Lecture at the National Academy of Sciences, http://www.nsf.gov/news/speeches/colwell/rc02abelwolman/index.htm

Onsrud and Campell "Big Opportunities in Access to "Small Science" Data" *Data Science Journal Vol. 6 June 2007*

A Fitzgerald, K Pappalardo and A Austin, "Practical Data Management: A Legal and Policy Guide" (2008) <eprints.qut.edu.au/archive/00014923/01/Microsoft_Word_-_Practical_Data_Management_-_A_Legal_and_Policy_Guide_doc.pdf>.

Feist Publications v. Rural Telephone Service 499 U.S. 340 (1991) <http://caselaw.lp.findlaw.com/scripts/getcase.pl?court=US&vol=499&invol=340>

Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31996L0009:EN:HTML>

Smith et al *Open ICT4D* (2008) IDRC Digital Library

Duncan et al. "Report of the Committee on National Statistics' Panel on Confidentiality and Data Access" , 1993. *Private Lives and Public Policies*, Washington, DC: National Academy Press, p. 23.

IRB Guidebook, Part III.D, Department of Health and Human Services, Office for Human Research Protections.

Brian Bergstein "Research explores data mining, privacy" in *USA Today* 6/18/2006

Lane, Julia and Schur, Claudia, Balancing Access to Health Data and Privacy: A Review of the Issues and Approaches for the Future (September 13, 2009). Available at SSRN: http://ssrn.com/abstract=1472736

Hailshree Saksena "Doctrine of "Sweat of the Brow" <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1398303>

Assessment Technologies v. Wiredata, 350 F.3rd 640 <http://altlaw.org/v1/cases/1129733>

Bertrand Warusfel "Legal protection of databases in Europe and public scientific research" <www.epip.eu/papers/20031124/200411.../EPIP%20Warusfel.ppt>

Fitzgerald et al, *OAK Law Project Report No. 1* (2006) p44

Rebecca S. Eisenberg, *Genomics in the Public Domain: Strategy and Policy*, 1 Nature Review Genetics 70, 73 (2000).

David W. Opderbeck, *The Penguin's Genome, or Coase and Open Source Biotechnology*, 18
HARV. J.L. & TECH. 167, 199 (2004).

Donna M. Gitter, *Resolving the Open Source Paradox in Biotechnology: A Proposal for a Revised Open Source Policy for Publicly Funded Genomic Databases*, Houston Law Review 43, 1476-1521.

Open Knowledge Foundation, *A Guide to Open Licensing*,
<http://www.opendefinition.org/guide?action=print>

Science Commons "Database and Creative Commons"
<http://sciencecommons.org/resources/faq/databases/>

Thinh Nguyen *Freedom to Research: Keeping Scientific Data Open, Accessible, and*

Thomas Samuel Kuhn *The Structure of Scientific Revolutions* University of Chicago Press, 1962.

James A. Evans and Jacob Reimer "Open Access and Global Participation in Science" (2009) *Science* Vol. 323. no. 5917, p. 1025 <
http://www.sciencemag.org/cgi/content/abstract/323/5917/1025>.

Gunther Eysenbach "Citation Advantage of Open Access Articles" PLoS Biology (2006) 4:5 p 0692 – 0698
<http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.0040157>.

Steve Hitchcock "The effect of open access and downloads ('hits') on citation impact: a bibliography of studies" <http://opcit.eprints.org/oacitation-biblio.html>.