ISSN 1574-1796

An International Journal on Grey Literature

Summer 2019, Volume 15, Number 2

'OPEN ACCESS TO GREY RESEARCH DATA'

GreyNet

www.textrelease.com Grey Literature Network Service www.greynet.org

The Grey Journal

An International Journal on Grey Literature

COLOPHON

CIP
The Grey Journal (TGJ) : An international journal on grey literature / Dominic Farace (Journal Editor); Jerry Frantzen (Technical Editor) ; GreyNet International, Grey Literature Network Service. - Amsterdam: TextRelease, Volume 15, Number 2, Summer 2019. – TIB (DE), DANS-KNAW (NL), CVTISR (SK), EBSCO (USA), ISTI CNR (IT), KISTI (KR), NIS IAEA (AT), NTK (CZ), and the University of Florida (USA) are Corporate Authors and Associate Members of GreyNet International.
ISSN 1574-1796 (Print)
ISSN 1574-180X (PDF)
ISSN 1574-9320 (CD-Rom)

Subscription Rates:
€240 institutional, including postage & handling

Contact Address:
Back Issues, Document Delivery, Advertising, Inserts, Subscriptions:

TextRelease
Javastraat 194-HS
1095 CP Amsterdam
Netherlands
T/F +31 (0) 20 331.2420
info@textrelease.com
http://www.textrelease.com/glpublications.html

About TGJ

The Grey Journal is a flagship journal for the international grey literature community. It crosses continents, disciplines, and sectors both public and private.
The Grey Journal not only deals with the topic of grey literature but is itself a document type classified as grey literature. It is akin to other grey serial publications, such as conference proceedings, reports, working papers, etc.

The Grey Journal is geared to Colleges and Schools of Library and Information Studies, as well as, information professionals, who produce, publish, process, manage, disseminate, and use grey literature e.g. researchers, editors, librarians, documentalists, archivists, journalists, intermediaries, etc.

About GreyNet

The Grey Literature Network Services was established in order to facilitate dialog, research, and communication between persons and organizations in the field of grey literature. GreyNet further seeks to identify and distribute information on and about grey literature in networked environments. Its main activities include the International Conference Series on Grey Literature, the creation and maintenance of web-based resources, a moderated Listserv, and The Grey Journal. GreyNet is also engaged in the development of distance learning courses for graduate and post-graduate students, as well as workshops and seminars for practitioners.

Full-Text License Agreement
In 2004, TextRelease entered into an electronic licensing relationship with EBSCO Publishing, the world's most prolific aggregator of full text journals, magazines and other sources. The full text of articles in The Grey Journal (TGJ) can be found in *Library, Information Science & Technology Abstracts* (LISTA) full-text database.

Contents
'Open Access to Grey Research Data'

INFORMATION SUPPORT OF SLOVAK SCIENCE

SCIENTIFIC LIBRARY AND INFORMATION SERVICES

- technology and selected areas of natural and economic sciences
- electronic information sources and remote access
- depository library of OECD, EBRD and WIPO

SUPPORT IN MANAGEMENT AND EVALUATION OF SCIENCE

- Central Registry of Publication Activities
- Central Registry of Art Works and Performance
- Central Registry of Theses and Dissertations and Antiplagiarism system
- Central information portal for research, development and innovation - CIP RDI >>>
- Slovak Current Research Information System

SUPPORT OF TECHNOLOGY TRANSFER

- Technology Transfer Centre at SCSTI
- PATLIB centre

POPULARISATION OF SCIENCE AND TECHNOLOGY

- National Centre for Popularisation of Science and Technology in Society

IMPLEMENTATION OF PROJECTS

- National Information System Promoting Research and Development in Slovakia – Access to electronic information resources - NISPEZ
- Infrastructure for Research and Development - the Data Centre for Research and Development - DC VaV
- National Infrastructure for Supporting Technology Transfer in Slovakia - NITT SK
- Fostering Continuous Research and Technology Application - FORT
- Boosting innovation through capacity building and networking of science centres in the SEE region - SEE Science

www.cvtisr.sk
Lamačská cesta 8/A, Bratislava

EDITOR'S NOTE

An upcoming seminar in the GreyForum Series will look at grey literature not only as an important vehicle for the circular economy but will also demonstrate its part in the information industry alongside other industries embedded in a circular economy. Grey Literature as a vehicle is foremost responsible for the knowledge transfer required to raise awareness to the circular economy. And, it is this that is required to bring about needed change in the mindset required for its implementation. Industries like that of information appear to have remained until now on the periphery.

Grey literature encompasses a range of diverse types of multidisciplinary and sustained information resources that allow for innovative solutions to barriers incorporated in a linear, ownership based information industry. Grey Literature is open access compliant, exploits the reuse and enhancement of existing resources, and seeks to preserve the knowledge contained therein for both science and society.

This first of its kind forum on the circular economy and grey literature is organized by GreyNet International together with the Amsterdam based platform gocirularnow.com. Further details are provided On the News Front section of this journal issue.

Dominic Farace,
Journal Editor

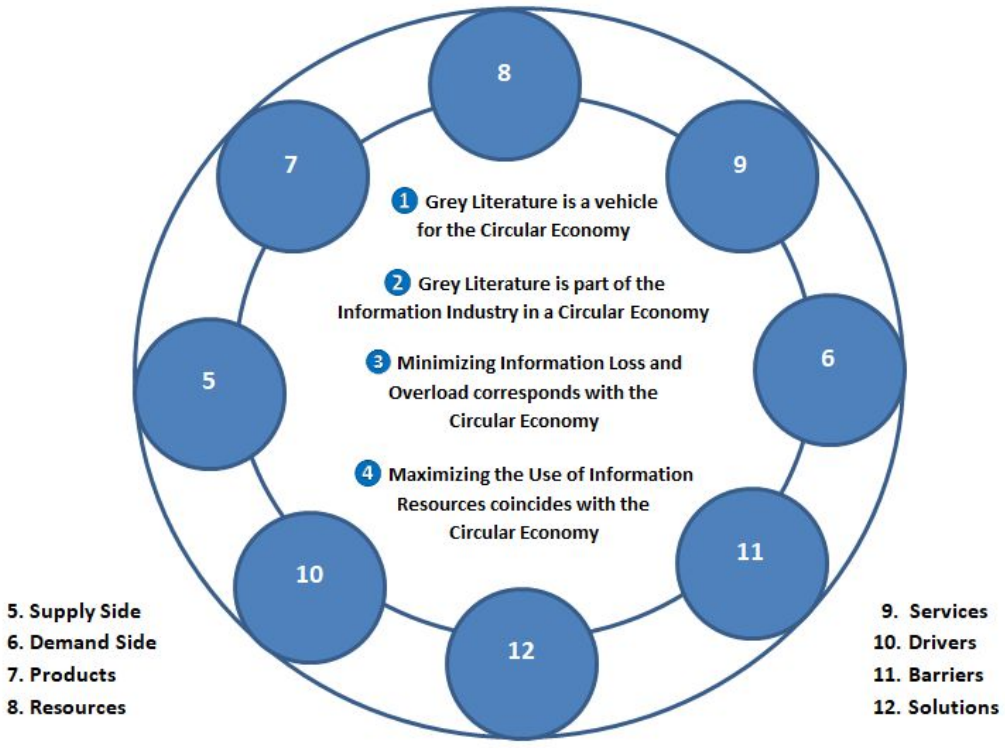


Figure: Grey Literature and the Circular Economy

INIS

www.iaea.org/inis

The International Nuclear Information System

The world's leading source of nuclear information since 1970

IAEA
International Atomic Energy Agency

When is 'grey' too 'grey'? A case of grey data*

Dobrica Savić, Nuclear Information Section,
International Atomic Energy Agency, NIS-IAEA, United Nations

Abstract
Conformity to facts, accuracy, habitual truthfulness, authenticity, information source reliability, and security have become important concerns. Trustworthiness of news and information, and of grey and other literature types has become of interest to the public, as well as to many information science and technology researchers. Starting with a definition of grey literature, and continuing with white, dark and grey data, this paper concentrates mainly on grey data as an emerging grey literature data type and its various 'shades' of trust. Special attention is given to data in the context of grey systems theory, anonymous data, and unstructured and unmanaged data. Based on a review of relevant literature and current practices, trustworthiness of grey data is analysed and elaborated. Guidelines and warning signs of grey data trustworthiness are identified, and conclusions offered.
Keywords: grey literature, grey data

Why are we concerned about the greying of grey data?

Recent research by the European Broadcasting Union (EBU) on misinformation shows that only 59% of people in the European Union (EU) believe what they hear on the radio, 51% believe the television news, and only 47% believe what they read (Financial Times, 2018). Widespread fake news, misinformation, disinformation, spam emails, computer bots, botnets, web spiders, crawlers, and viruses erode our trust in the information and data we encounter in our daily lives, making trustworthiness a concern.

To further illustrate the concern of trustworthiness, consider that 269 billion emails are sent and received each day, of which 60% is spam. 56% of all internet traffic is from automated sources — hacking tools, scrapers and spammers, bots, and other malicious programs. Therefore, conformity to facts, accuracy, habitual truthfulness, authenticity, information source reliability, and security are of increasing importance.

Another factor impacting trust is the amount of data surrounding us. 2.5 exabytes of data are produced every day, the equivalent of 250,000 Libraries of Congress and 90% of all the data in the world that has been generated over the last two years. 13 million text messages are sent every minute, 4.4 million videos are watched on YouTube every minute and 1.7 megabytes of new information are created every second for each human being on the planet. Although the amount of information and data[1] around us is enormous, 99.5% of all data created is not currently being analysed and used. Still, we are hungry for information, as demonstrated by over 6.6 billion Google queries daily, 15% of which have never before been searched.

Uncovering deception and estimating the veracity of information and data is difficult now and will be even more so in the future.

Grey literature

Various definitions of grey literature exist. The 12th International Conference on Grey Literature (GL12), held in Prague in 2010, defined it as "manifold document types produced on all levels of government, academics, business and industry in print and electronic formats that are protected by intellectual property rights, of sufficient quality to be collected and preserved by library holdings or institutional repositories, but not controlled by commercial publishers, i. e., where publishing is not the primary activity of the producing body" (Farace, D. and Schoepfel, J., 2010).

Adams (2016) adds another twist to the definition of grey literature by proposing to look at it from the perspective of traditional publishing, which includes a peer-review process. Accordingly, grey

* First published in the GL20 Conference Proceedings, February 2019.
[1] Data is 'facts or figures from which conclusions can be drawn'. Information is 'data that have been recorded, classified, organized, related, or interpreted within a framework so that meaning emerges'. www.statcan.gc.ca

literature is regarded as "the diverse and heterogeneous body of material that is made public outside, and not subject to traditional academic peer-review processes" (Adams at al., 2016).

The current definition faces some challenges, such as multiple types of originators — humans and machines — volume, and the speed of grey literature creation. It also faces the possibility of becoming obsolete due to its inability to differentiate between grey literature and other types of literature. Therefore, the following new definition was proposed: "***grey literature is any recorded, referable and sustainable data or information resource of current or future value, made publicly available without a traditional peer-review process"*** (Savić, 2017).

This definition considers all major elements of the grey literature concept. Namely, long term preservation, sustainability, usability, and value, while acknowledging the lack of a traditional peer-review process for regular 'white' literature.

As Figure 1 shows, there are many types or forms of grey literature, although this paper mainly deals with grey data sets. The GreyNet website lists over 150 document types including databases, data sets, data sheets, data papers, satellite data, and product data.

Bibliographies	Rejected manuscripts	Publications from NGOs and consulting firms
Discussion papers	Un-submitted manuscripts	Videos
Newsletters	Conference abstracts	Wiki articles
PowerPoint presentations	Book chapters	Emails
Program evaluation reports	Personal correspondence	Blogs and social media
Technical notes	Newsletters	Data sets
Publications from governmental agencies	Informal communications	Committee reports
Reports to funding agencies	Census data	Working papers
Unpublished reports	Pre-prints	Company reports
Dissertations	Standards	Catalogues
Policy documents	Patents	Speeches
	Webinars	Reports on websites

Figure 1: Types of grey literature

There are many new sources of data, such as the Internet of Things (IoT), Machine to Machine communication (M2M), self-driven cars, robots, sensors, security systems, surveillance cameras, and many other systems or apps using AI and machine learning. The estimated number of currently connected electronic devices creating specific data varies by billions. Data produced by these devices is highly contextual and software dependent, making it hard to collect and process, and even harder to make sense of and preserve for future use.

White data

White data comes from the wider concept of 'white literature' (Jeffery, 2006), which is regarded as peer-reviewed and published literature, usually in the form of articles and books. It is often referred to, especially when talking about data, as open data. In other words, "freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control" (Wikipedia).

The International Open Data Charter 2015 defines open data as digital data that is made available with the technical and legal characteristics necessary for it to be freely used, reused, and redistributed by anyone, anytime, anywhere. It promotes the following principles:

- Open by default
- Timely and comprehensive
- Accessible and usable
- Comparable and interoperable
- For improved governance and citizen engagement
- For inclusive development and innovation

The European Union regards open (government) data as the information collected, produced or paid for by public bodies and can be freely used, modified, and shared by anyone for any purpose.

The US Federal Open Data Policy of 2013 refers to open data as publicly available data structured in a way that enables the data to be fully discoverable and usable by end users. It is consistent with the principles of public, accessible, described, reusable, complete, and timely.

Many other countries such as Russia, China, and Japan also have their well-developed and defined national legislation and regulations regarding open data, particularly government and publicly funded data. It is interesting to note that the Russian Open Government Data (OGD) Recommendations of 2014, include requirements for licensing, mandatory procedures for data publication, rules for data publishing, data formats (CSV, XML, JSON, RDF), metadata format, and other technical requirements. Japanese regulations encourage the use of public information for both commercial and non-commercial purposes.

Grey data

Grey data represents a type of grey literature that maintains its basic facets, such as that it is recorded, that it is referable, sustainable, valuable, publicly available, and without a traditional peer-review. Just as with grey literature, it is this last characteristic that causes some data to be regarded as grey data. It is data that is useful and valuable, but not vetted by peer-review or other existing governance mechanisms.

Grey data is also an umbrella term that describes the vast array of data that organizations collect and use. It is often critical to an organization's ability to innovate, enhance, and execute its core mission and it is usually collected for mandatory or compliance purposes, such as HR, budget and finance, contracts, procurement, facility management, registered library users, database subscriptions, and information collection maintenance. Besides being important for operational and internal management, grey data is usually collected and managed for legal and regulatory purposes.

Many organizations that offer products and services collect data on users, product sales, and penetration of services not only for the purpose of production but also, importantly, for marketing.

Dark data

In contrast to white data, dark data is barely visible. It represents the information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes, such as analytics, business relationships and direct monetizing.

Like dark matter in physics, dark data often comprises an organizations universe of information assets. Thus, organizations often retain dark data for compliance purposes only, without much further use. However, storing and securing data typically incurs more expense and sometimes represents considerably greater risk than value. Dark data is also recognized as a potential for revenue. According to Garner research, by 2020, 10% of organizations will have a business unit to make their data commercially available.

Data mining can be used to get value out of dark data as a data analysis process of sorting through large data sets to identify patterns, establish relationships, and solve problems. Data mining tools could even be helpful in allowing enterprises to predict future trends based on their historical data. However, some data archaeology might need to be performed to reuse preserved historical data by recovering information stored in formats that have become obsolete.

The table below offers an overview of major differences between white, grey and dark data.

Facet	White	Grey	Dark
Recorded	x	x	X
Valuable	x	x	X
Referable	x	x	X
Sustainable	x	x	
Used	x	x	
Public	x	x	
Peer reviewed	x		

Figure 2: Differences between white, grey and dark data

Grey data diversity

There is a great variety of approaches to grey data. A brief overview of the four most interesting and important approaches will be presented here. It includes:

- Data in the context of Grey System Theory
- Anonymous data as defined by the EU
- Unstructured data
- Unmanaged (risky) data

The Grey System Theory was made popular by Julong Deng (1982). He successfully developed a methodology which focused on the study of problems involving small samples and poor information, which is very often a common situation in decision- making, irrelevant of the field (e.g. economy, finance, politics and others).

In their book on grey data analysis, Sifeng Liu, Jeffrey Forrest, and Yingjie Yang (2017), present the fundamental methods, models and techniques for the practical application of grey data analysis. They also specifically talk about various types of data, e.g., black, white, and grey. For them 'black' indicates unknown, 'white' indicates completely known, and 'grey' indicates partially known and partially unknown information. More specifically, grey data represents small samples and poor information which is often only partially known, incomplete, inaccurate, and inadequate.

	Situation		
Concept	Black	Grey	White
From information	Unknown	Incomplete	Completely known
From appearance	Dark	Blurred	Clear
From processes	New	Changing	Old
From properties	Chaotic	Multivariate	Order
From methods	Negation	Change for better	Confirmation
From attitude	Letting go	Tolerant	Rigorous
From the outcomes	No solution	Multi-solutions	Unique solution

Figure 3: Black, white, and grey data according to the Grey System Theory

The European Union recognizes anonymous data as a type of grey data and presents it as a legal term used in the EU General Data Protection Regulation (GDPR). The reuse of personal data (processed for purposes beyond its original collection) is a key concern for the EU data protection law. GDPR applies only to information concerning an identified or identifiable natural person. Therefore, anonymized data is no longer considered to be personal and is thus outside the scope of GDPR, but there are problems with the techniques used to make data anonymous. The use of direct and indirect identifiers (quasi-identifiers) such as age, gender, education, employment status, economic activity, marital status, mother tongue, and ethnic background, is of considerable concern.

GDPR regards pseudonymous data also as personal data, e.g. data which uses assigned IDs but where the research team has a key that can be used to connect the data to research participants. A process of well-planned and well-performed 'de-identification', or removal/editing of identifying information in a dataset to prevent the identification of specific cases is legally required. It should be carried out in such a manner that 'de-anonymization' is not possible, i.e. re-identification of data that is classified as anonymous by combining the data with information from other sources.

The European Union also insists on the principle of minimization. In other words, only the minimum amount of personal data necessary to accomplish a task/research should be collected, making most of the data, in fact, grey data.

Unstructured data represents any data that does not have a recognizable structure. It is data that, due to its non-existing structure, is not fit for use in a classical relational database. For example, text documents, email messages, PowerPoint presentations, survey responses, transcripts, posts from blogs, social media, images, AV files, machine and log files, and sensor data.

Due to the remarkable development of information technology tools such as AI, machine learning, deep learning, natural language processing, data mining, and predictive analytics, unstructured data is being analysed, categorized, classified, and efficiently stored. It should be noted that the

line between structured and semi-structured data is very thin. By simply adding metadata tags to the data content, unstructured data can become semi-structured, or even fully-structured data.

Unmanaged or risky data represents, according to some estimates, almost 30% of corporate storage space. Another 30% of storage is filled with active data, while 40% of the data is inert and needs to be kept for archival or regulatory purposes. Out of the 30% of unmanaged data, 15% represents dark storage allocated but unused, 10% is orphaned data that should have been discarded long ago, and 5% is personal data that should not have been placed on corporate servers at all. To decrease the amount of unmanaged data, organizations need to implement well-established data governance policies that include relevant standards, life-cycle management guidelines, and compliance instructions, and quality control measures need to be put in place. Unmanaged data also represents considerable risk for organizations due to data clutter, liability issues, and increased security breaches, as well as the cost of maintenance, backups, disaster recovery, servers, space, electricity, and staff involvement.

Synthetic data is the newest and probably the most interesting part of the grey data spectrum. It is a specific set of data that is artificially manufactured, rather than directly measured and collected from real-world situations. Synthetic data is usually anonymized (stripped of identifying aspects such as names, emails, social security numbers and addresses) and created based on user-specified parameters resembling the properties of data from the real-world. It is an important tool to augment machine learning algorithms when real data is too expensive to collect, inaccessible due to privacy concerns, or incomplete.

AI systems that can learn from real data can also create data sets resembling authentic data. With further developments in information technology, it is expected that the gap between synthetic data and real data will diminish. Waymo LLC, a subsidiary of Alphabet Inc., tested its autonomous vehicles by driving 8 million miles on real roads and another 5 billion on simulated roadways — real proof of the power of synthetic data in practical life.

Unsettling grey data

In exploring grey data, it is also necessary to mention some of the challenges that can be unsettling Two areas of concern are the data itself and its basic purpose.

Concerns about the data itself come about because data is unverifiable, available for use but without any guarantee of its truthfulness. The danger in this lack of basic verification can result in inaccurate, or simply fake data, as is often demonstrated in news feeds. The structure of grey data is also often unclear, as well as its format and the tools required for analysis, making its use difficult. Encryption, an often misleading sign of trustworthiness, combined with redundancy, makes the use of grey data unsettling.

The purpose of data creation and its existence is another concern in ensuring the trustworthiness and usability of grey data. A good rule of thumb is to verify the source of the data and to be wary of questionable sources that may have clandestine reasons for creating the data in the first place, such as misinformation, defaming, or other hidden intents.

A final note regarding the unsettling use of grey data concerns the currently popular and widespread use of **F.A.I.R**. principles (**F**indable **A**ccessible **I**nteroperable **R**eusable). These principles are all valid and should be promoted and used; however, in my opinion, the most important one, **T**rustworthiness, is missing. **T**rustworthiness needs to be established, rigorously checked and followed.

Conclusions

Conformity to facts, accuracy, habitual truthfulness, authenticity, information source reliability, and security have become important concerns. The trustworthiness of news and information, of grey and other literature types, and of grey data has become a public concern. The increasing amount of grey data being created impacts the way we process, disseminate, manage, and use this type of information; consequently, demanding greater trustworthiness.

Processing needs to be well-thought out and present from the beginning of grey data creation. Ad-hoc or post-processing can no longer be regarded as efficient. Environmental and technical, economic and financial, social and organizational constraints need to be taken into consideration for long-term grey data sustainability and usefulness. Its usability requires adequate IT tools, the

availability of qualified human resources, and the protection of intellectual property and personal privacy.

To secure the future use and maintain the value of grey literature, intensive training, wide-spread cooperation, and proper management are needed. Only a small percent of businesses extracts the full value from the data they hold. The use of new IT tools such as AI, could improve its value, improve business results, bring measurable efficiency gains, and increase the quality of products and services. However, this will only happen if grey data reaches the required level of users' trust.

REFERENCES

Adams et al., 2016. Shades of Grey: Guidelines for Working with the Grey Literature in Systematic Reviews for Management and Organizational Studies. International Journal of Management Reviews. 2016. (http://onlinel ibrary.wiley.com/doi/10.1111/ijmr.12102/full).

Farace, D. J. and Schoepfel, J. (Eds.), 2010. Grey Literature in Library and Information Studies. De Gruyter Saur, Germany.

Financial Times, 2018. Matters of Fact. News in the Digital Age, November 2018. Published by Financial Times in collaboration with Google.

GL12, 2010. Twelfth International Conference on Grey Literature. National Technical Library, Prague, Czech Republic. December 6-7, 2010. www.textrelease.com/gl12conference.html

Jeffery, K. 2006. Open Access: An Introduction. ERCIM News 64, January 2006.

Liu, S., **Yang**, Y., **Forrest,** J. 2017. Grey Data Analysis: Methods, Models and Applications. Springer.

Savić, D., 2017. Rethinking the Role of Grey Literature in the Fourth Industrial Revolution. 10th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology. Available from: http://nrgl.techlib.cz/index.php/Proceedings. ISSN 2336-5021. Also published by TGJ (The Grey Journal) Special Winter Issue, Volume 14, 2018.

Legal Issues Surrounding the Collection, Use and Access to Grey Data in the University Setting: How Data Policies Reflect the Political Will of Organizations*

Tomas A. Lipinski, School of Information Studies; University of Wisconsin-Milwaukee
Kathrine A. Henderson; LibSource, A LAC-Group Company, United States

Abstract:
Grey literature is defined as works that are of sufficient importance to be collected and preserved by the library or its affiliated institutional repository. These works are disseminated through channels other than commercial publishing and are generally protected by intellectual property. Intellectual property schemes offer less protection to grey literature's frequent companion, grey data, even though the collector/researcher and his/her home institution may nonetheless consider the data valuable and proprietary (Schöpfel and Lipinski, 2012). Grey data gives rise to a number of legal and ethical considerations often addressed through university policy. This study of university data repository policies is divided into four parts.
Part I considers a definitional framework as defined by Post, Raile and Raile (2010) to demonstrate how political will might be operationalized in the context of developing university data repository policies.
Part II describes the legal issues surrounding collection, use, and access to grey data. The authors identify several intellectual property schemes and other proprietary doctrines that may apply. Part II also addresses ownership and access rights to data including contract, statutory or regulatory schemes that require access.
Part III examines the grey/open data polices set by institutional repositories. The analysis includes Terms of Use/Terms of Access of six institutions in the United States as reflected in Park and Wolfram (2017) and Park (2018). The analytical framework employed follows Lipinski and Copeland (2015) and Lipinski and Kritikos (2018).
Part IV considers the tension that results from the need for universities to raise revenue and the public mission/role of the university in society in the same manner as Rooksby (2016). Finally, best practices are forwarded.
Keywords*: grey data; political will; institutional policy; copyright; trade secret; terms and conditions; misappropriation; best practice*

Introduction

As the definition of grey literature expands it is natural now to also frame questions in the context of grey data. In 2017 Dobrica Savić formulated the following definition of grey literature to include data: "GL [Grey Literature] is any recorded and sustainable data or information resources of current or future value, made publicly available without a traditional per-review process" (Savić, 2017, p. 111). Grey data are works of sufficient importance to be collected and preserved by the library or its affiliated institutional repository. Using a methodology and analysis informed by policy review of Lipinski and Kritikos (2018), this paper reviews several university policies regarding data repositories asking whether the policies express or reflect a "political will" of an institution towards its grey data.

Part I Defining Political Will

Neither individual nor collective political will is particularly well defined in the literature. The concept, as commonly understood, is associated with the success or failure of a government policy. If a policy is enacted, then one credits the 'political will' of those who were in a position of power to directly or indirectly influence the outcome. Conversely, one might say that when a policy initiative failed to move forward is due to the absence of 'political will.' Hammergren characterized political will as "the slipperiest concept in the policy lexicon, the sina qua non of policy success which is never defined except in its absence" (Post et al., 2010, p. 654).

In their 2010 analysis, Post, Raile, and Raile express concern over this casual usage of the term "political will" noting that its ambiguity makes it an "ideal for achieving political aims and for labeling political failures when the diagnosis is unclear." Further they note that the concept is far

* First published in the GL20 Conference Proceedings, February 2019.

too important to "be abandoned to the hollow of political rhetoric." And with this in mind, they created a systematic, definitional approach, which allows political will to be an "empirically useful and actionable concept with an overarching goal of building political will for effective public policies" (Post et al., 2010, p. 654).

Their approach exams political actors in nation-states; however, in this instance, it is used to help gain an understanding of how the political will of a university is reflected in its policies, specifically, its data repository policies. This systematic approach defines "political will" into successive components:

1. A sufficient set of decision makers
2. With a common understanding of a particular problem on the formal agenda
3. Is committed to supporting
4. A commonly perceived, potentially effective policy solution (Post et al., 2010, p. 659).

A Sufficient Set of Decision Makers

There are a number of stakeholders who have an interest in creating policies, which govern the collection and use of and access to research data. These may include, but are not limited to:

- Boards of Regents
- President/Chancellor or Provost
- Faculty Senates
- General or IP Counsel
- Units tasked with hosting repositories
- Corporate partners or other funders
- Faculty/Students who have conducted the research

These stakeholders may share an interest in establishing data repository policies, but may not have decisional authority. Key decision makers must have authority, capacity, and legitimacy to develop and implement policy to exercise the political will of the university. Furthermore, these decision makers must also have the power to support or block the approval, implementation, or enforcement of said policy.

In the analysis of data repository policies (found in Part III) the units tasked with hosting repositories were primarily university libraries frequently in partnership with university technology units and/or social science research units. For example, "[t]he Purdue University Research Repository (PURR) is a core research facility provided by the Purdue University Libraries, the Office of the Executive Vice President for Research and Partnerships and Information Technology at Purdue (ITaP)." It is not clear which, if any, decision makers from the three collaborating units drove the establishment of specific policies at Purdue or whether other stakeholders were involved and if similar circumstances occurred at any of the other universities included in the analysis. However, because the policies exist, it is reasonable to conclude that a subset of decision makers held a common understanding of the problems surrounding the housing and re-use of research data such that the decision was taken to support the establishment of a data repository and, subsequently, its governing policies. The decision makers then vested units hosting the repository with the authority to administer and execute the policy.

A Common Understanding of Particular Problem

Is it unlikely that decision makers will have the collective political will to approve, implement, and enforce a policy if their efforts are aimed toward different problems, which require different policy solutions. It is critical; therefore, that decision makers agree that there is a problem to be solved in the first place; second, there is agreement about the nature of the problem; third, agreement that the nature of the problem is such that it requires a policy solution.

The clearest indication that decision makers have converged on a common understanding of a problem is likely to be demonstrated through public discussions of the problem in which decision makers will begin to use similar terminology and frames. (Post et al., 2010, p. 663). The authors' analysis in subsequent sections illustrates a common understanding of a problem that universities face individually and collectively when it comes to the collection and use of and access to research data. Customary provisions include:

- Requirements for attribution or citations which reflect expectations of scholarly practice;
- Provisions for metadata to meet library needs;
- Warranty disclaimers or damage waivers to shift risk away from the University; and
- Overarching requirements stating that accessing the platform or using the data equals assent to the terms and conditions set-forth in the repository policy.

Commitment to Support

A third element is also critical and represents the point at which the individual decision maker's preferences come to fruition. Each decision maker will likely need to take into consideration his or her constituents' desires and gauge how much leverage these constituents might have relative to a proposed solution. Externalities such as legal or contractual obligations must also be part of any proposed solution.

For example, private partners or government funding entities may obligate decision makers to enact a certain policy provision over another. There may be contractual agreements with corporate partners, which require data to be held in secret or the data must be shared because the research was funded with public dollars. There might also be concerns about the continuation of the positive relationship with funders going forward. Too, there may be constituents who wish to block provisions or policies if they believe the policy or provision is not in the best interest of a specific group or the university.

Until policy is set it is difficult, if not impossible to determine if decisions makers are genuinely committed to a particular policy solution. However, indications would include allocation of a variety of resources and a willingness to apply effective sanctions. The decisions makers' intentions may be ascertained to some degree through observations at meetings, public statements and the like ([Post et al., p. 663]).

A Commonly Perceived, Potentially Effective Policy Solution

The final component of the definition does not require that all decisions makers agree on a specific solution, rather there must be agreement regarding the nature of the policy that is needed. While there need not be agreement on every element of the policy there must consensus on what issues the policy is intended to resolve and the general parameters of how that resolution is to be executed and by whom. This is necessary before there is the political will to achieve a desirable policy solution. Further, the requirement that the solution must also be potentially effective is not meant to 'predict' whether or not a policy will actually be effective based on a set of evaluative standards, but rather to avoid disingenuous, quick-fix solutions that may come about through a political will of avoidance or that stems from manipulation or coercion ([Post et al., p. 666]).

Returning to the data repository policy problem, potential policy solutions include: on-going allocation of library resources toward collection and storage of research data and IT resources for platform development and access control; asserting and enforcing intellectual property rights and creating user agreements. Because universities rely heavily on federal grants to fund research, potential policy solutions would also need to reflect Office of Management and Budget guidelines per the authority granted in 31 U.S.C. § 503 regarding post award requirements for public access to research data established initially by President Obama in his January 21, 2009 Presidential Memorandum on Transparency and Open Government. Currently, the federal government has the right to "[o]btain, reproduce, publish, or otherwise use the data produced under a Federal award" and to authorize others to do the same (2 C.F.R. § 200.315(d)(1) and (2)). Other than exceptions for trade secret and "[p]ersonnel and medial information" (2 C.F.R. § 200.315(e)(3)(i) and (ii)), the guidelines reach a broad array of grey material defining research data as "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings" (2 C.F.R. § 200.315(e)(3)).

Part II Legal Issues Surrounding Collection, Use, and Access to Grey Data.

Lipinski and Kritikos (2018) discuss the ownership of scholarly work product (grey literature) within the context of copyright, in specific, journal literature and open access repositories. However, the analysis of grey data involves additional legal schemes, as protection when extent, may not be found in copyright alone.

Legal Control and the Exercise of Political Will

If legal control over grey data is sought by the institution the legal basis for that property right could rest in several legal doctrines: copyright, trade secret or misappropriation. If none of theses concepts apply then the legal control exerted by the institution and reflected in its data repository policy would contractual in nature.

Copyright

Raw data or facts are not protected by copyright. Under *Feist Publications v. Rural Telephone Service Co*. (1991), facts lack the necessary originality or creativity to qualify for copyright protection. Numerical and descriptive characteristics of a group of research subjects consists of facts, as does the documentation of laboratory experiments and attendant findings. Much of grey data falls into this categorization. Other grey data may have some elements of creativity to it such as lab notes or filed observations consisting of analytic commentary of the researcher and the descriptive observations (facts). Likewise, grey data such as interviews and oral histories would arguably be protected by copyright. Such works, a blend of fact and expressive elements would be considered as works of "thin copyright" under the doctrine of fair use (17 U.S.C. § 512). Unprotected as well as "thin" elements would be subject to re-use by subsequent researchers. In *Maxtone-Graham v. Burtchaell* (1986), the Second Circuit concluded that the reuse and analysis of interview excerpts from a differing perspective constituted fair use. "Maxtone-Graham's book was essentially factual in nature, and, as the district court correctly noted, subsequent authors may rely more heavily on such works" (*Maxtone-Graham v. Burtchaell*, 1986, p. 557). Such works, while arguably protected by copyright, would be subject to robust fair use for reuse.

A compilation is a protected "work formed by the collection and assembling of preexisting materials or of data that are selected, coordinated, or arranged in such a way that the resulting work as a whole constitutes an original work of authorship" (17 U.S.C. § 101). In applying this concept to grey data, if raw data has been refined and manipulated so that the data have been "selected, coordinated, or arranged" into a data set in a meaningful, original and useful manner the data set could meet the very low originality threshold articulated in *Feist Publications, Inc. v. Rural Telephone Service Co.* (1991). The data set would be protected by copyright as compilation. Mere numerical or alphabetical order is of insufficient creativity. The "originality" requirement for copyright protection means that the selection, coordination or arrangement of the facts must have some minimum of originality or creativity. Under United States law though the compilation would be protected the underlying facts contained within such a manipulated data set remain unprotected and are not transformed into protected content merely because the data now resides within a data set subject to compilation copyright. As a result, reproducing the entire compilation data set as a precursor to extracting the unprotected elements would be a fair use. *Ticketmaster Corp. v. Tickets.com, Inc.* (2003) involved the reproduction of concert time and pricing data from a website and *Assessment Technologies of WI, LLC. v. Wiredata, Inc.* (2003) involved the reproduction of public domain real estate data. The court in each decision commented that the extraction of both protected and unprotected content so that the unprotected content could be mined or identified, i.e., the ticketing venue, date, time, place, etc. and identifying characteristics and tax assessment of each property, constitutes a fair use. It could also be argued that reproducing all or a substantial part of a data set to identify and harvest, or mine the unprotected elements is also a fair use. Courts view such uses and transformative because the data-mining or full-text searching serves a "new and different function" and is a "quintessentially transformative" purpose (*Authors Guild, Inc. v. HathiTrust*, 2014, p. 97).

In conclusion, raw data (facts) are not protected by copyright. Regarding works of thin copyright, fair use would support reuse of the facts as well as the thin elements (especially if an alternative analysis were applied). A data set protected as a compilation would be subject to fair use when for example it is reproduced in its entirety so the unprotected elements can be mined and subsets of data identified from entire set.

Trade Secret

Trade secret may protect a variety of content that does not possess the requisite originality for copyright protection but nonetheless is valuable. Trade secrets could include a formula, a process, method or technique, a collection of data not protected by a compilation copyright, such as a research subject list, ratings or scores, characteristics or attribute list. If data is related to an identifiable data subject privacy issues arise but those issues are not covered in this paper. Before trade secret protection applies several criteria must exist. First the owner must "derive independent economic value, actual or potential" from the secret (Wis. Stat. § 134.90(1)(c)). It must not be readily known or ascertainable, i.e., it is a secret. Third, that benefit must derive from its not being generally known to, and not being readily ascertainable by proper means to others. Fourth, the information must have economic value to the party appropriating the secret. Finally, the party claiming trade secret protection must undertake efforts that are reasonable under the circumstances to maintain its secrecy. (Wis. Stat. § 134.90(1)(c)) It is unlikely that trade secret protection would apply to the grey data under discussion here. While the subject matter of some grey data may fall under the scope of trade secret the institutional policies (terms of service) reviewed here suggest an opposite strategy. The purpose of the repositories reviewed here (see Table 3) is to make the data accessible and encourage its reuse. The key element, as the legal phrasing suggests, is missing. The content of the data repositories is not treated as a secret but is accessible to other researchers.

Misappropriation

A third doctrine that can offer legal control over grey data is misappropriation. The doctrine can apply to facts. The U.S. Supreme Court first articulated the tort (legal harm) of misappropriation in the *International News Service v. Associated Press* (1918) decision. International News Service (INS) took notes from Associated Press (AP) stories displayed on bulletin boards displaying early editions of East Coast newspapers, wrote their own news stories, then sold and transmitted paraphrased versions of the stories to newspapers on the West Coast. As only the facts were reused there was no copyright infringement. The Court nonetheless concluded that INS harmed the AP and created a new remedy in the law. The Court articulated the elements of the new tort of misappropriation: an intangible asset, created through expenditure of effort and with the expectation of profit, that was taken with comparatively little effort and used in a commercial setting, causing economic harm to the possessor of the intangible asset. Applying these concepts to grey data it could be said that although the institution or initial collector of the data is likely not undertaking the endeavor solely with monetization in mind there is likely an expectation that the data is of value, producing measurable results such as future external funding, publication, scholarly accolade or ranking, citation counts, etc. Likewise, the user of data is not necessarily seeking to commercialize the data either but desires similar academic rewards.

The Second Circuit articulated the modern concept of misappropriation in *National Basketball Association v. Motorola, Inc.* (1997), limiting its application to circumstances where the information represents so-called "hot news" or current, time-sensitive information There are several elements: the information is gathered at a cost, is time-sensitive, the use constitutes free-riding with the defendant exerting little time or effort to obtain the information, the use is in direct competition with the plaintiff where the free-riding reduces the incentive to produce the product or service. This could apply to grey data in areas of cutting-edge research where multiple institutions are working on the same problem, vying for the next break-through and competing for limited external funds to further the next iteration of discovery. The fact that the grey data is open and available does not matter; it could still be subject to a claim of misappropriation. For example, *Pollstar v. Gigmania, Ltd*. (2000) (concert information) and *Fred Wehrenberg Circuit of Theatres, Inc. v. Moviefone, Inc*. (1999) (movie listings and show times) both involved circumstances where the appropriated information resided on open websites as does the grey data in the institutional repositories reviewed here. However, the misappropriation claim failed in both cases. What was lacking there, as with the repositories here was a disincentive to continue creating the information if the appropriation was allowed. The initiators of the concert and movie websites would continue to create and post prices, times, concert performers and movie titles even when an aggregating competitor website continues to harvest and reuse the information. If it could be said that being "scooped" or "beaten to the punch" by a rival research team at another institution would indeed

trigger a disincentive to engage in such projects in the future, then the circumstances may satisfy the legal requirements for a claim of misappropriation. On the other hand, much research in the "academy" is collaborative, often with teams of researchers working on facets of the same problem at different institutions. The teams may or may not be in competition with each other but if in competition would likely not abandon such efforts in the future because another team is working on the same problem. While grey data could be protected depending on the circumstances by the doctrine of misappropriation in the policies reviewed here, there does not appear the political will to exert such control or claim. Rather the policies reflect an understanding of a different particular problem and an effective policy solution depending on the institution.

Table 1. Summary of Legal Doctrines Applying to Grey Data

Legal Doctrine	Requirements for protection	Example	Application to Grey Data
Copyright	An original, work of authorship, fixed in a tangible medium.	Statistics, numerical and descriptive characteristics of data subjects, results of experiments, etc.	No. Not applicable to "raw" data, grey or otherwise.
Copyright	Same.	Interviews, oral histories or field notes and observations that include commentary or analysis.	Possible. Mixed works of fact and original expression "thin" copyright. Fair use applies to reuse of "thin" copyright works.
Copyright	A collection and assembling of data that are selected, coordinated, or arranged with sufficient to constitutes an original work of authorship.	Manipulated or refined data sets "selected, coordinated, or arranged" in original and useful manner.	Possible. May be subject to protection as a compilation copyright. Fair use for purposes of extraction or mining applies.
Trade Secret	Information with independent economic value derived from not being generally known with efforts to maintain secrecy, not readily ascertainable by proper means, others would obtain economic value from disclosure or use.	Processes or method (viable or failed), formulas, data such as ratings, scores, attributes and characteristics of data or data subjects.	Possible but not available if the data is open; accessible, reusable, etc., i.e., not kept secret.
Misappropriation	Information collected at a cost, time-sensitive, the use by others constitutes "free-riding" in direct competition where the use reduces the incentive to produce the product or service.	Time-sensitive data.	Possible. Researcher(s) must be in scholarly competition with each other such that the reuse of the grey data a disincentivizes future research.

As summarized in Table 1, the authors conclude that the legal basis for control over access and use of institutional grey data is not tenable under copyright unless the data set is selected, coordinated, or arranged with the originality required to constitute a compilation. Other works of grey data may contain creative element and be protected but fair use would offer robust rights of reuse. While some grey data may qualify as a trade secret the open nature of grey data repositories forecloses qualification as a trade secret. Under certain circumstances—where a competition for time-sensitive grey data exists—concepts of misappropriation may apply provided such use by the other researcher(s) disincentivizes the desire for continued data collection or generation by the institution claiming misappropriation.

Terms of Service as an Expression of Political Will in Property Rights

If it is the intent of the institution to exercise control over its data, the legal mechanism supporting the control must then be found in principles of contract law so that the terms of service represent a binding contract elucidating the terms and conditions of the access and use. The policies reviewed here appear to evidence this approach. For example, Michigan uses a click-to-agree mechanism known as click-wrap. Minnesota, Virginia Tech and Purdue each use a browse-wrap license, indicating that use or downloading or accessing the data equates to consent to the terms of use. Illinois uses less clear language stating that "users…consent to the collection and storing of Data." Only Harvard disavows contract formation. Stating that its Community Norms are "not a binding contractual agreement" and "does not create legal obligation to follow these policies."

End User License Agreement are used as this allows the institutional to achieve private ordering of the rights and obligations of the parties (Casamiquela, 2002). "Parties to a contract may limit their right to take action they previously had been free to take" (*Universal Gym Equipment, Inc. v. ERWA Exercise Equipment Ltd.*, 1987, p. 1550).

As the ownership issues that would arise under copyright, trade secret or misappropriation do not appear operable, here contract is used to delineate the contours of access and reuse of the grey data. This suggests that each institution except Harvard perceived a policy problem and sought a policy remedy that could be imposed through binding terms and conditions. Even the Harvard policy reflects a political will of sorts, expressing an institutional desire to claim no rights or exert no conditions on the use of the data whatsoever, though the researcher adding data to the repository possesses the option to adopt "custom terms of use."

Except for Purdue, a consistent condition of data access and use in the policies reviewed here is a requirement of attribution and proper citation. In the U.S. such right is not considered part of the bundle of exclusive rights of the copyright holder. (17 U.S.C. § 106). Attribution is not a property right as is copyright, trade secret or misappropriation. Rather the concept of attribution is viewed as a moral right. "*Droit moral,* or moral right, is generally summarized as including the right of an artist to have his work attributed to him in the form in which he created it to prevent mutilation or deformation of the work." *Museum Boutique Intercontinental, Ltd. v. Picasso*, 1995, p. 157 at n. 3) Moral rights are fully developed in the legal traditions of European countries but have limited application in U.S. law. "American copyright law, as presently written, does not recognize moral rights or provide a cause of action for their violation, since the law seeks to vindicate the economic, rather than the personal, rights of authors" (*Gilliam v. American Broadcasting Companies, Inc.*, 1976, p. 24). 17 U.S.C. § 106A offers moral rights to the creators of limited categories of visual art: "painting, drawing, print, or sculpture, and still photographs" (17 U.S.C. § 101). The attribution requirement found in five of the policies reviewed here demonstrates that one problem of open data repositories is data provenance. Reuse of the data should require the subsequent researcher to indicate the source of the data. Requiring proper attribution reflects the political will of the institution in identifying a "particular problem" and adopting a solution enforced by contractual obligation.

Part III Analysis of U.S. Universities Data Use Policies and Terms of Use

In this part, the authors analyze the available Terms of Use/Terms of Access of 6 institutions in the United States as reflected in Park and Wolfram (2017) and Park (2018). The analytical framework employed follows Lipinski and Copeland (2015) and Lipinski and Kritikos (2018). Of the fourteen institutions of higher learning reviewed in Park and Wolfram (2017) and Park (2018), eight were located in the United States and six were located in the United Kingdom (Edinburgh, Essex, Glasgow, Leeds, Nottingham and Southampton). Those located overseas were not included in this study. Of the remaining 8 U.S. institutions, Iowa Research Online (University of Iowa) was eliminated because the documentation surrounding the repository, the submission form for example, makes clear that it is not a data repository. Rather it is an Open Access Institutional Repository: "Iowa Research Online (IRO) is a dynamic repository created to increase the impact of research and creative scholarship at the University of Iowa. IRO offers the ability to view publications by department, academic discipline or author." Likewise, the "Terms of Use for Arch" (Northwestern University) prohibit users from using "a facsimile of the published version of an article that may be posted in Arch." Other statements on the arch.library.norhtwestern.edu/

website confirm that Arch is indeed "an open access repository for the research and scholarly output of Northwestern University." As a result, further review of Arch (Northwestern University) did not occur. However, both Iowa Research Online and Arch Open Access repositories align with the assessment of iSchool Open Access repositories undertaken in Lipinski and Kritikos (2018). Further review of the data repository websites of the remaining 6 institutions confirmed that the content of each repository indeed consists of grey data (see table, 2).

Table 2. Institutional Data (Grey) Repository Terms of Use

Institution	Repository Name	Repository URL	Description	Documentation Reviewed
University of Illinois at Urbana Champaign	Illinois Data Bank	https://databank.illinois.edu/	The Illinois Data Bank is a public access repository for publishing research data from the University of Illinois at Urbana-Champaign.	Access and Use Policy
University of Michigan	ICPSR (Inter-university Consortium for Political and Social Research)	https://www.icpsr.umich.edu/icpsrweb/	ICPSR encourages and facilitates research and instruction in the social sciences and related areas by acquiring, developing, archiving, and disseminating data and documentation relevant to a wide spectrum of disciplines, and by conducting related instructional programs.	What are ICPSR's terms of Use.
Harvard University	Harvard Dataverse Network	https://dataverse.harvard.edu/	The Harvard Dataverse Network, housed at the Institute for Quantitative Social Science (IQSS) at Harvard, hosts the world's largest collection of social science research. Go here to find data or create a dataverse of your own to share your social science data and get a formal persistent citation.	Dataverse Terms
University of Minnesota	DRUM (Data Repository for the University of Minnesota	https://www.lib.umn.edu/datamanagement/drum	Got data? We're here to help you manage, share, and preserve your research data. In addition to our Data Repository for the U of M curation services, the Libraries will help you navigate available campus resources throughout the data lifecycle…	DRUM Terms of Use
Virginia Tech (Virginia Polytechnic Institute and State University)	VTechData	https://data.lib.vt.edu/	Welcome to VTechData, the data repository of Virginia Tech! Virginia Tech's Data Repository is a platform for highlighting, preserving and providing access to the work generated by the Virginia Tech Community.	VTechData Deposit Agreement, Publication Agreement and Publishing Requirements
Purdue University	PURR (*Purdue University* Research Repository)	https://purr.purdue.edu/	The *Purdue University* Research Repository (*PURR*) provides an online, collaborative working space and data-sharing platform to support Purdue researchers and their collaborators.	PURR Terms of Use

The authors next analyze the terms of use for the remaining six U.S. institutions using a methodology (content analysis) employed in Lipinski and Copeland (2015) and Lipinski and Kritikos (2018). A discussion of this analysis follows.

Table 3. Analysis of Data Repository Terms of Use

	Illinois Data Bank	ICPSR (Michigan)	Harvard Dataverse	DRUM (Minnesota)	VTechData	PURR (Purdue)
Institutional Responsibility	University Library	unit within the Institute for Social Research	Harvard University Library and Harvard University IT	University Libraries	University Libraries (Research & Informatics Division).	Purdue University Libraries, Office of the Executive VP for Research/ Partnerships and IT at Purdue (ITaP)."
Purpose	facilitating access and use… dissemination	terms of use vary depending upon how we acquired the study	facilitates reuse and extensibility of research data	expectation that data will be re-used	improve the discoverability and access of research data… is discoverable and accessible	collaboration, publish and archive datasets for long-term access and reuse
Contract Formation	users… consent'	click on the 'I Agree' continuing you signify your agreement	not a binding agreement, downloading does not create a legal obligation	using or downloading	depositing (uploading) you agree	by accessing you accept, terms effective immediately upon posting, continued use equaling acceptance
Privacy of Data Subjects	none	for research or statistical purposes, not investigation of specific research subject, no use identity discovered inadvertently	none	discovered inadvertently	agree to comply with federal law including privacy, IRB approval	HIPPA, "storage of personally identifiable data [PID] or sensitive information scrubbed, FERPA IRB approval
Use Restrictions	none?	no redistribution NACJD, NCHS click "I Accept" re use identity discovered inadvertently	default is CC0.	none	Visibility Level, right to temporarily restrict access.	non-commercial education and research, circumventing data quota, no spam, solicitations, criminal and civil violations, breach of contract
Privacy of Users	Google Analytics to track users	None	Guestbook to track users (who/what)	none	Administrators maintain the privacy of users	user responsible ID/password security, cookies, opt out newsletters / non-essential communication
Attribution to Data Authors	Descriptive Metadata: CC BY	Akin to CC BY?: reference recommended bibliographic citation	proper credit given via citation	appropriate attribution	minimum level of metadata, citation and digital object identifier	none
Permissions	CC0, CC BY or non-standard license	Akin to CC SA?: include terms of use	CC0 or custom Terms	contact authors directly	permission of the original data creator	none

Copyright	creators retain copyright, fair use applies	None	none	none	permission for data copyright	none
Formats	convert proprietary	None	none	none	migrate or transform to maintain its accessibility	none
Disclaimers	not responsible	original collector / funding agency no responsibility	none	no warranty, fitness for any purpose	warranty?: contact Administrators issues or errors	AS IS, merchantability, fitness for a particular purpose, accuracy, adequacy or completeness, uninterrupted or error-free, viruses or harmful components
Damage Waiver	none	None	none	actual, incidental or consequential	none	consequential, indirect, punitive, special or incidental
Choice of Law and Choice of Forum	none	None	none	none	none	federal and Indiana law, courts in Tippecanoe County
Staff Assistance	none	None	none	finding, accessing, and downloading	contact Administrators uploading, organizing, and describing the data	none
Sanctions	none	revoke agreement, return data, deny future access, revocation of tenure and termination, institutional suspension / damages	none	none	right to remove violation policy federal, state, or local laws	termination spam / commercial advertisements, terminate or suspend without notice any reason violation of terms, unlawful or harmful to others

The institutional responsibility in all but one institution (Michigan) resides or is shared by the university library. In Minnesota the library is the sole responsible party. At both Illinois and Virginia Tech, the library shares this responsibility with a research unit (Research Data Service and Research & Informatics Division, respectively), at Harvard the library shares this responsibility with the campus IT unit (Information Technology organization) and at Purdue the library shares responsibility with a unit that combines research and IT (Research Partnerships and Information Technology).

Michigan uses a click-to-agree mechanism to evidence the assent of data users to its terms of use while Minnesota and Purdue both equate use with contractual assent. Virginia Tech equates use with assent also but not by users of the data repository rather its Publication Agreement governs terms binding the researcher contributing content to the databank. Harvard disavows contractual standards indicating that its "Community Norms" are not binding, that using the data (downloading) does not create a "legal obligation to follow these policies." Illinois states that use equals consent, not for agreement with the terms of use but for the collection of user data through

analytic tools (see Table 3). It is likely that a "sufficient set of decision makers" were involved in all instances. At least the entity or entities administering the repository and a stakeholder with higher (approval) authority would have been involved. As most policies involved contractual obligations by users the Office of General Counsel, Legal Affairs or similar entity would also have been involved.

Michigan, Minnesota, Virginia Tech and Purdue have provisions relating to the privacy of data subjects should the identity of the subject, as phrased by both Michigan and Minnesota, be "discovered inadvertently." As the Virginia Tech agreement is directed at the contributor of data, its terms likewise reference federal and IRB protocols for maintaining the privacy of data subjects. Oddly, Illinois and Harvard make no mention of research subject privacy. In fact, a substantial portion of the Illinois terms relate to copyright and the privacy of users and not the privacy of research subjects. Illinois along with Harvard employ tracking technologies on use of the repository data by others. Virginia Tech states that its Administrators are responsible for maintaining the privacy of users, while Purdue on the other hand admonishes users that the responsibility to maintain the confidentiality of an ID and password falls on the user ("you are responsible"). Purdue also uses "tracking cookies."

Use restrictions are absent from the Illinois terms of use. Likewise, Virginia Tech does not contain any use restrictions, but this is logical as its terms are directed at the depositor and publisher of data not at third party users of the data. As discussed above Michigan prohibits redistribution of National Archive of Criminal Justice Data (NACJD) and requires a further click-to-agree relating to data subject privacy for use of National Center for Health Statistics (NCHS). The Minnesota terms of use are the shortest (14 lines) and do not contain use restrictions while the Purdue terms are almost three and a half pages in length and employ more conventional contractual provisions as discussed in Lipinski (2013), than any reviewed here. Purdue also provides the most elaborate list of prohibited conduct or uses. In addition to prohibitions on spam, solicitations, provisions related to criminal and civil liability including intellectual property infringement and computer crimes (citing 17 U.S.C. § 512 and 18 U.S.C. § 1030, respectively), users are also prohibited from posting or transmitting "any unlawful, threatening, libelous, harassing, defamatory, vulgar, obscene, pornographic, profane or otherwise objectionable content." While many of the categories listed have legal meaning "pornographic" and "otherwise objectionable" are void of legal meaning. Furthermore, much of what might be considered vulgar or profane speech is not unlawful but protected by the First Amendment and the ability of a government entity such as a public university to regulate such speech often depends on the context of the speech. Finally, it is not clear if Purdue reviews all the data submitted to determine if it passes muster under its list of prohibitions. The nature of each policy's terms and conditions reflect the political will of the institution. For example, Illinois addressing copyright ownership issues, Virginia Tech on contributor obligations and use restrictions in others or the absence of restrictions (Harvard) reveals the perceived problem and policy solution at each institution.

Several institutions require proper attribution, i.e., Michigan ("you agree to reference the recommended bibliographic citation"), Harvard ("good scientific practices… proper credit is given via citation") and Minnesota ("appropriate attribution") while it is implied by Virginia Tech ("minimum level of metadata… such that it receives a citation and digital object identifier…") and by Illinois depending on the terms of use employed by the depositor of the data (CC BY license). Likewise, the issue of permissions is provided through use of Creative Commons licensing or by non-standard or custom terms by Illinois (CC0, CC BY or non-standard) and Harvard (CC0 or custom). It is logical that a CC0 license would be employed as much of the grey data is not copyrightable. Michigan implies a CC SA (share alike) when it states that "you must include all accompanying files with the data including the terms of use." Virginia Tech makes references to permissions but again as its agreements govern those who deposit data the charge is to ensure that those uploading data "have the permission of the original data creator." Only Minnesota tells re-users of data to "contact authors directly." As just mentioned Virginia Tech requires that proper permissions be obtained for depositing "data containing information under copyright." The Illinois terms of use contain extensive discussion of copyright and its application to the data deposited stating that the "Creators of Datasets" retain the copyright, assuming there is copyright, but that "the fair use doctrine" applies. It is possible that the data, if factual, might qualify as a compilation copyright. Only two institutions note that data repository staff may migrate data to new formats:

Illinois ("proprietary formats may be converted") and Virginia Tech ("migrate or transform the Published data to maintain its accessibility").

Several institutions have terms of use provisions relating to risk-shifting strategies as discussed in Lipinski (2014). For example, all but Harvard have some type of warranty disclaimer relating to the integrity of the data. This is understandable; Harvard acknowledges that the Community Norms do not "create any legal obligations" nor present "contractual agreement" thus disclaiming any warranty would be legally pointless. Illinois is "not responsible" for the files or the terms a depositor might chose to accompany their data. Michigan "bear[s] no responsibility." Minnesota disclaims a specific warranty: fitness for any purpose. As with the lengthy list of prohibited uses Purdue offers the most elaborate set of disclaimers including that the data are provided "as is" and disclaiming warranties of merchantability and fitness for a particular purpose, as well as a general disclaimer of "accuracy, adequacy or completeness" or that the service will "operate [] uninterrupted or error-free" or be "free of viruses or other harmful components." Finally, Purdue provides that content "do[es] not reflect the stances" or "imply endorsement of Purdue University." The inclusion of this provision suggests the Purdue may have experienced a "particular problem" or controversy regarding the subject matter of some research and the data the research generated. In contrast, Virginia Tech instead of disclaiming warranties appears to imply a warranty of accuracy or at least would take corrective actions by stating that "Depositors and Publishers can contact VTechData Administrators with any issues or errors they encounter in interactions with the VTechData platform…" Another legal risk-shifting device is to disclaim damages should harm from the data result and a court finds the institution responsible. Minnesota and Purdue adopt this strategy. Only Purdue includes a common contractual provision known as a choice of law and choice of forum provision. These clauses usually indicate that the law to be applied in interpreting the terms of use and the court making the application and interpretation is the home jurisdiction of the institution. As a result, the law of Indiana applies to the Purdue agreement and the courts in Tippecanoe County, where West Lafayette (the home of Purdue University) is located, have jurisdiction. Minnesota and Virginia Tech offer staff assistance in "finding, accessing or downloading data" and in "uploading, organizing and describing the data," respectively but Virginia Tech does not aid in "using, analyzing or understanding' the data. Three of the six terms of service provide for some level of sanctions if the terms are not followed.

Virginia Tech may remove data if the content is in violation of law or institutional policy. Purdue again provides two typical remedies of termination and suspension with broad power ("sole discretion") to terminate or suspend without notice for any reason. Michigan in addition to termination ("revoke" and "deny all future access") includes the untypical sanction of "revocation of tenure and termination" as well institutional sanctions ("suspension of all research grants"). Disavowing staff assistance or including repercussions for data misuse illegality also suggest "particular problem[s]" that may have arisen.

Review of the terms of use of the six institutions reveal similarities and differences. Similarities in most or all were found in the responsible party (library alone or with research and/or IT services), purpose (access and/or reuse), contract formation (click or use equals assent), privacy of data subjects (two were silent), requirements relating to attribution or proper citation and disclaimers of warranty or accuracy (all but Harvard disclaimed one or more warranty).

Differences or less similarity (terms of use were silent on the topic) was found among provisions relating to use restrictions, privacy of (re)-users of the data (providing for the ability to track users of data, one vesting responsibility with the institution, another placing responsibility with the user and two were silent), permissions (half using CC or license or embedded terms of use, one silent, two placing the responsibility of the user to contact the "author directly" or to have the "permission of the original data creator"), formats (only two anticipated format migration, the remaining 4 were silent), damage waiver (only two disclaimed damages), choice of law/forum (only one), staff assistance (one providing help in using the data, another offering assistance with "issues or errors") and sanctions of uses not conforming to the terms of use (half contained a sanction provision: right to remove nonconforming content, terminate or suspend access and the most harsh revocation of tenure /termination and denial of institution-wide access).

Part IV Conflict of Roles

Grey data, unlike grey literature, including grey media, does not easily lend itself to copyright protection and so other legal schemes must be employed by the university in order to retain its value. Protecting grey data, just like protecting its grey literature and media counterparts, is subject to the same sorts of difficulties that arise whenever higher education seeks to protect its intellectual outputs. As Rooksby notes, "The concept of higher education as a special sector, set apart from others, reflects the bargain that society strikes with higher education regarding its intellectual capital: storing, creating, and disseminating new knowledge is a difficult but important undertaking, one that must be free from state-imposed restrictions and directives, lest the pursuit of truth suffer. In short, higher education deserves special treatment…precisely because of the intellectual benefits to society that flow from higher education's principle undertakings: research, teaching, and service" (Rooksby, 2016, p. 7). Data policies should, and in some circumstances, must require that the intellectual benefits inherent to grey data flow into society; however, there is a competing narrative: "Higher education is designed to generate scientific discoveries and artistic contributions. These creations [and by extension, the data which emerges from these endeavors] unquestionably further the public good, but so can the ownership and restriction of them." Thus, the question becomes "to what extent must higher education fruits be owned or claimed in order for the public to benefit most?" (Rooksby, 2016, p. 9).

It is a fair point that the economic reality is such that most universities must find additional revenue streams. Exercising intellectual property rights—copyright, patent, trademark or other proprietary rights—is lucrative and helps to ensure institutional longevity; however, it also requires institutions to defend their property rights by controlling access and litigating against infringement, trademark dilution, misappropriation and so forth. Universities have an interest in exercising and protecting intellectual property rights. Universities must take steps to protect trade secrets and the privacy of individuals. The legal infrastructure supports these efforts in offering regulatory exception to the data disclosure rules. (2 C.F.R. § 200.315(e)(3)(i) and (ii)).

The conflict between the institution's educational purpose—research, teaching and service--and the need to protect and monetize intellectual property may influence data repository policies. As mentioned previously, the purpose of the repositories reviewed here is to make the data accessible and encourage its reuse. In addition, some policies show concern for the creator and embrace open data initiatives indicated in permissions provisions. Half of the universities are using creative commons or license or embedded terms of use, one is silent and two place the responsibility of the user to contact the "author directly" or seek "permission of the original data creator." This may be an indication that these universities are tempering their impulse to "restrict or claim the intellectual fruits of their faculty students, and donors in favor of serving as a "knowledge commons" where "resources are generally available for all to use and consume at a minimal cost" (Rooksby, 2016, p. 268).

Conclusion and Recommendations

Providing for storage, access, and use of research data is an ideal opportunity for the university to serve its primary social purpose to create and disseminate knowledge through one of its primary endeavors, research. Establishing a data repository with effective policies is complex, requiring decision makers to: consider legal issues including intellectual property rights and contractual agreements; address funding requirements; and determine how to best control access and manage associated risks. All of which is further complicated by the desire to provide this access in ways that allow and encourage collaboration between and among scholars. While there is no predictable path to establishing an effective data repository policy, among other things, the authors recommend that decision makers consider the legal issues outlined herein and determine the level of risk the university will tolerate and how to best mitigate risk to meet that level of tolerance. In addition, universities should consider how to balance the public purpose of higher education and the protection of intellectual property to be housed in its data repository.

Universities:

- Should adopt terms and conditions of use.
    - The terms and conditions should employ a click-to-agree mechanism rather than a use-equals-assent formulation.

- Warranties regarding the reliability of the data should be disclaimed.
- Should protect repository content by exercising applicable intellectual property rights.
    - Unless the grey data contains copyrightable content or data sets subject to protection as a compilation a CC0 license could be used. Authors recognize that CC0 is not ideal as it places the work in the public domain and there are no attribution requirements.
    - If attribution is desirable use a CC BY license or a provision requiring proper citation/attribution.
- Should encompass other legal obligations including
    - Privacy of research subjects
    - Publication of publicly funded research, and
    - Contractual agreements with private partners.
- Should reflect the educational mission of the university.

References

Casamiquela, R.J. (2002). Contractual Assent and Enforceability in Cyberspace. *Berkeley Technology Law Journal*, 17(1), 475-495. (ISSN: 1086-3818).
Lipinski, T.A. (2011). A Functional Approach to Understanding and Applying Fair Use. *Annual Review of Information Science and Technology*, 45(1), 525-621. (ISSN: 1550-8382).
Lipinski, T.A. (2013). *The Librarian's Legal Companion for Licensing Information Resources and Services*. Chicago, Illinois: Neal-Schuman Publishers, Inc./ALA Editions. (ISBN: 9781555706104).
Lipinski, T.A. (2014), Click Here to Cloud: End User Issues in Cloud Computing Terms of Service Agreements. In John N. Gathegi, J.N., Tonta, Y., Kurbanoğlu, S., Al, U. & Taşkın, Z. (Eds.), *Challenges of Information Management Beyond the Cloud*, 92-111. Berlin/Heidelberg, Germany: Springer Berlin/Heidelberg. (ISBN: 3-662-44412-7).
Lipinski, T. A., & Copeland, A. J. (2015). Is the licensing of grey literature using the full palette of "contractual" colors?: A comparative analysis of grey literature terms of use. *The Grey Literature, An International Journal on Grey Literature*, 11(2), 69-87. (ISSN 1574-1796).
Lipinski T.A. and Chamberlain Kritikos, K. (2018). How open-access policies affect access to grey literature in university digital repositories: A case study of iSchools. *The Grey Literature, An International Journal on Grey Literature*, 14(1), 6-20. (ISSN 1574-1796).
Park, H., & Wolfram, D. (2017). An examination of research data sharing and re-use: Implications for data citation practice. *Scientometrics*, 111(1), 443-461. (ISSN 1588- 2861).
Park, H. (2018) Unpublished list of data repositories prepared by Hyoungjoo Park; on file with the authors.
Post, L. A., Raile, A. N. W., & Raile, E. D. (2010). Defining political will. In *Politics & Policy*, 38(4), 653-676. (ISSN 1747-1346).
Rooksby, J. H. (2016). The branding of the American mind : How universities capture, manage, and monetize intellectual property and why it matters. Baltimore, Maryland: Johns Hopkins University Press. (ISBN 9781421420806 1421420805).
Savić, D. (2017). Impact of Emerging Information Technologies on Grey Literature, *Nineteenth International Conference on Grey Literature: Public Awareness and Access to Grey Literature*, Rome, Italy, October 23-24, 2017, pp. 110-114. (ISSN: 1385-2308).
Schöpfel, J., & Lipinski, T. A. (2012). Legal aspects of grey literature. *The Grey Literature, An International Journal on Grey Literature*, 8(3), 137-153. ISSN 1574-1796).

Case References

Assessment Technologies of WI, LLC. v. Wiredata, Inc., 350 F.3d 640 (7th Cir. 2003).
Authors Guild, Inc. v. HathiTrust, 755 F.3d 87 (2d Cir. 2014).
Feist Publications v. Rural Telephone Service Co., 499 U.S. 340 (1991).
Fred Wehrenberg Circuit of Theatres, Inc. v. Moviefone, Inc., 73 F. Supp. 2d 1044 (E. D. Mo. 1999).
Gilliam v. American Broadcasting Companies, Inc., 538 F.2d 14, 24 (2d Cir. 1976).
International News Service v. Associated Press, 248 U.S. 215 (1918).
Maxtone-Graham v. Burtchaell, 803 F.2d 1253 (2d Cir. 1986), cert. denied 481 U.S. 1059 (1987).
Museum Boutique Intercontinental, Ltd. v. Picasso, 880 F. Supp. 153 (S.D.N.Y., 1995).
National Basketball Association v. Motorola, Inc., 105 F.3d 841 (2nd Cir. 1997).
Pollstar v. Gigmania, Ltd., 170 F. Supp. 2d 974 (E.D. Cal. 2000).
Ticketmaster Corp. v. Tickets.com, Inc., 2003 WL 21406289 (C.D. Calif.) (unpublished).
Universal Gym Equipment, Inc. v. ERWA Exercise Equipment Ltd., 827 F.2d 1542 (Fed. Cir. 1987).

Statutes and Regulations

17 U.S.C. § 101
17 U.S.C. § 106A
17 U.S.C. § 107
17 U.S.C. § 512
18 U.S.C. § 1030
2 C.F.R. § 200.315

On Open Access to Research Data: Experiences and reflections from DANS*

Emilie Kraaikamp, Marjan Grootveld, Hella Hollander, and Dirk Roorda
Data Archiving and Networked Services, DANS-KNAW, Netherlands

Abstract
This paper addresses the question of what open access means in relation to research data. To that end we (1) briefly discuss the definition of open access and the most important expected benefits; (2) highlight the publication - data distinction and the challenges regarding research data; (3) discuss research data and open access in terms of its focus, licensing, and current practices at DANS in publishing data sets open access. Use case illustrate challenges and positive developments. As a result, we see that open access for research data is not clearly defined and that the focus lies on publications, but in general it should come down to accessibility with minimal restrictions for reuse. While publications may need a different business model, making research data open access faces three challenges: awareness is needed for the importance of data and the value for publishing them; ownership should not comprise research data being open access; data management is an essential process in publishing data and needs to be part of a research project. We feel that open access regarding research data logically focusses on reuse. Therefore, it seems logical to combine open access with the FAIR principles. In terms of licensing, it is best not to license data at all, but to apply the Creative Commons CC0 tool. However, citing is a requirement in the initial definition of open access, which would make it relevant to use a Creative Commons licence. CC-BY is in that respect the best match with regards to the open access definition.
While there is no clarity yet about what open access for research data exactly means, we see there is a gradual transition to making data open by using the Creative Commons means. This transition does still need careful guidance as publicly sharing research data is not yet common practice. Some extra reuse conditions may then help in at least making the data accessible. In summary, open access for research data relates more to reuse and less to accessibility and it is here where the requirements still need to be defined and put in place.

Introduction

The Dutch government promotes open access to all research publications and research data that have been funded with public money. All publicly funded research should be published open access by 2024, according to state secretary Sander Dekker in 2013[1].
DANS supports this target. DANS, short for Data Archiving and Networked Services[2], is a joint institute of the Royal Netherlands Academy of Arts and Sciences and the Netherlands Organisation for Scientific Research. DANS encourages researchers to make their digital research data findable, accessible, interoperable and reusable, and monitors the progress of open access in the Netherlands. Now, in 2018, open access research publications amount to more than 50% of the total Dutch output[3]. The statistics for research *data* are not precisely known.

In this article we are concerned with the question of what open access means in relation to research data. To that end we will (1) briefly discuss the definition of open access and the most important expected benefits; (2) highlight the publication - data distinction and the challenges regarding data; (3) discuss research data and open access in terms of its focus, licensing, and current practices at DANS in publishing data sets open access.

We have collected a few use cases from our practice as a research data repository that make the point that using resources goes further than merely accessing resources.

---

* First published in the GL20 Conference Proceedings, February 2019.
[1] Dekker, Sander. *Brief van de* staatssecretaris *van onderwijs, cultuur en wetenschap*. 2013, 31 288 nr. 354, The Hague, https://www.tweedekamer.nl/kamerstukken/brieven_regering/detail?did=2013D45933&id=2013Z22375 .
[2] https://dans.knaw.nl/en/about/
[3] https://www.openaccess.nl/en/in-the-netherlands/current-situation

Open access

Definitions

Open access is a term central to the international academic open access movement that "seeks free and open online access to academic information"[4]. In 2002 and 2003 three important developments mark the beginning of this movement that are central today: the Budapest Open Access Initiative[5], the Bethesda Statement on Open Access Publishing[6], and the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities[7]. Each provide a definition of open access along the same lines stating the need for accessibility and limited conditions for reuse, including citation.

In the search of a definition for open access, it is probably best to look at the Berlin Declaration since this is the most elaborate and refers to data as well. It reads:

"Open access contributions include original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material.

The author(s) and right holder(s) of such contributions grant(s) to all users a free, irrevocable, worldwide, right of access to, and a licence to copy, use, distribute, transmit and display the work publicly and to make and distribute derivative works, in any digital medium for any responsible purpose, subject to proper attribution of authorship (community standards, will continue to provide the mechanism for enforcement of proper attribution and responsible use of the published work, as they do now), as well as the right to make small numbers of printed copies for their personal use.

A complete version of the work and all supplemental materials, including a copy of the permission as stated above, in an appropriate standard electronic format is deposited (and thus published) in at least one online repository using suitable technical standards (such as the Open Archive definitions) that is supported and maintained by an academic institution, scholarly society, government agency, or other well-established organization that seeks to enable open access, unrestricted distribution, interoperability, and long-term archiving."

While data are mentioned, the general focus of open access is on publications. This is clear in the request regarding open access by the State Secretary in 2013 and later in 2017[8]. For publications the aim is open access, while for data the aim is "optimal reuse". Furthermore, the European Commission - a major research funder - currently refers to the Berlin declaration for a definition for publications and does not present a definition with regards to research data. While there may not be a clear definition of open access for data, it is clear that it should be along the lines of open access for publications.

As part of the monitoring process on open access, DANS has measured the type of access used for publications found through NARCIS[9]. NARCIS is one of the core services from DANS, a portal offering access to scientific information including publications. Metadata about access is delivered to NARCIS via the contributing institutions. In order to check if the measurement is correct, the results have been set against the information on access that can be found through Unpaywall[10] based on the Digital Object Identifier (DOI) of the publications. A visualisation of this can be seen in fig. 1. The outcome is quite remarkable as it seems to show that for many publications two types of access exist.

[4] https://www.openaccess.nl/en/what-is-open-access
[5] https://www.budapestopenaccessinitiative.org/read
[6] http://legacy.earlham.edu/~peters/fos/bethesda.htm
[7] https://openaccess.mpg.de/Berlin-Declaration
[8] Dekker, Sander. *Letter to the house of representatives, progress of open science.* 2017, The Hague, https://www.government.nl/documents/letters/2017/01/19/letter-to-the-house-of-representatives-on-the-progress-of-open-science .
[9] NARCIS https://www.narcis.nl/about/Language/en
[10] https://unpaywall.org

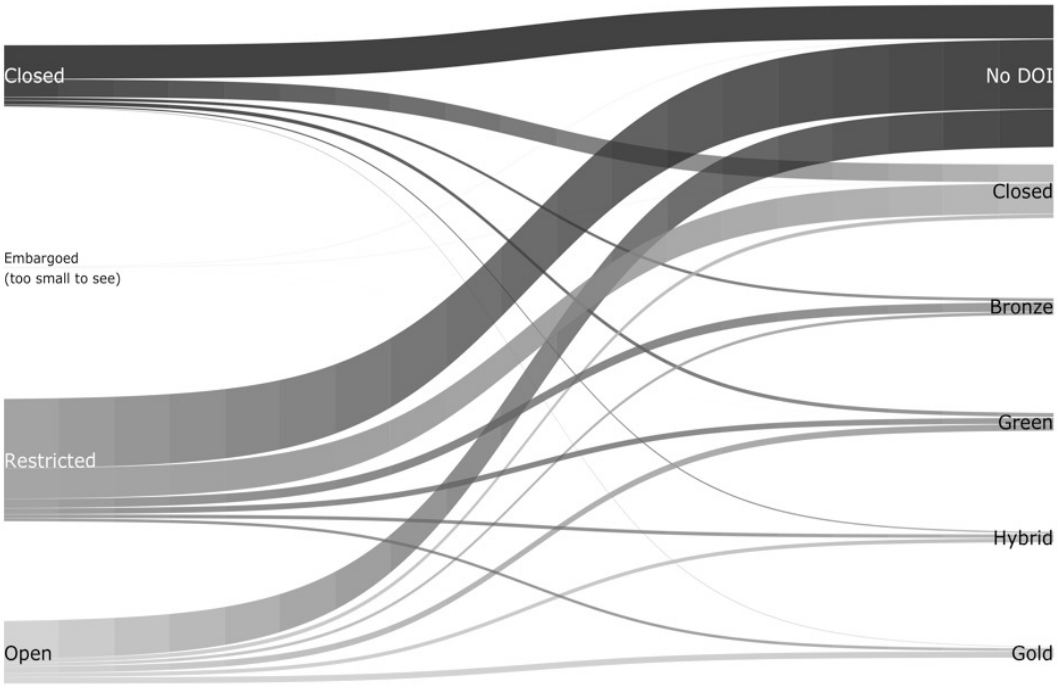


Fig. 1: Access categories for publications. Left axis: situation according to local repositories in the Netherlands (via NARCIS). Right axis: situation according to Unpaywall.
Figure courtesy of Emil Bode, DANS.

The reason that there are often different results for one publication is not due to different interpretations. These outcomes are probably caused by one of the following: either the publisher or the institution provides incorrect metadata, or there is somehow a deliberate choice to publish publications under two different access categories. The exact reasons are still unclear, but this result can be seen as an example of how open access is still in development.

Benefits of open research data

Better communication of research results is the most prominent advantage of Open Access. If reading is free and unencumbered, articles will enjoy an increased and more varied readership. The same applies to research data. This will increase the circulation and development of knowledge, which is a central goal of scholarship. And this may lead to more real-life applications of scientific output, which in turn both drives new research and helps to address societal challenges. For example, "We are moving rapidly towards the day when an epidemiologist in New York can instantly access anonymised patient data from Shanghai, and work with drug researchers in Basel to find new medicines" (RDA 2014). There are also disadvantages, because authors may have to pay their publishers additional fees (see below), and open access journals currently do not have the highest academic reputation. But despite such disadvantages, the benefits are increasingly seen as decisive[11].

Publications versus research data

From the start the notion of "Open Access" has mainly focused on research publications such as journal articles: ideally, they would be openly available to everyone, citizen scientists included. A crucial and highly debated aspect of this ambition concerns the traditional business model of journal publishers, which is based on subscription fees for getting access to a journal. Going for open access means: turn the business model on its head. However, when we shift the focus to research data, the challenge is not so much in the business model. Here, the challenges are threefold: awareness, ownership, and data management.

[11] http://www.openaccess.nl/en/what-is-open-access/pros-and-cons

Awareness

The tradition to publish research data is much less established than the tradition to publish journal articles. That data are valuable in their own right – apart from providing the foundation for claims made in publications – still needs a lot of advocacy.

There isn't even consensus on the definition of "research data"; it may vary per discipline and per research funder. In this paper we understand research data as any object or evidence that is needed to underpin research[12]. Think of observations, facts, artefacts, images, recordings, texts, annotations, collected by fieldwork, surveys, and measurements to be examined and considered as a basis for reasoning, discussion, or calculation. And the outcomes of those analytical processes can be used as data again in other research.

It is a paradoxical state of affairs: data are very close to the central workflows of science, they are much less encumbered with copyrights, and yet they are far less eagerly shared, although a new wave of Open Science[13] is trying to change that.

Ownership

There is a strong trend towards the position that research data should be published under an Open Access regime[14], unless they are sensitive because of privacy, commercial interests, or security. However, researchers often feel that creating such a level playing field is a meagre reward for their efforts to capture or generate the data. By their effort, they morally "own" their data and want to reap the benefits of them.

Embargo periods, during which only the data creators can use the data, are considered part of the solution. Proper citation of re-used data is another part. Although correctly referring to existing publications is a long-standing tradition in academia, data citation has not kept up with that. Also, here a paradox is lurking: just when the practice to measure researchers by the impact of their publications is increasingly seen as unproductive[15], making data citable is being advocated as an incentive for researchers to publish their data.

Data management

Where a publication is set for sharing, research data still need a processing step. Data management is essential and can be intensive. The sheer size of data, their rich variety in data types and in file formats, and their diverse provenance, make it much harder to handle data properly than publications. After selecting the relevant data, it needs to be assessed whether it has sensitive - e.g., personal-related or commercial - data inside, or derivatives of licensed other works. This will restrict the ways the data can be licensed. Quite differently, factual data are not considered copyrightable, so the basis for licensing is different in that case. If data are sensitive, one may choose to make two versions of the data, an anonymised version for public use and a version requiring limited access, which again requires a considerable processing step.

Furthermore, research data need to be made presentable - consider correct files names, ordering etc. - and properly documented. Also, file formats need to be used which can be sustained for the long term; to this end the DANS data archive works with a set of preferred formats[16]. Another important aspect of data management takes its effect after publishing the data. This applies to data that can only be accessed under certain conditions. For these data access control is needed.

These complexities of data have not received the same attention from scientific publishers as publications, and hence common practices such as long-term access, citation, and peer review are

[12] http://sedataglossary.shoutwiki.com/wiki/Research_data

[13] https://www.openaccess.nl/en/news-and-events

[14] For example, see Open Access and Data Management, EU Horizon2020 Open Research Data Pilot. http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm

[15] See for example Carpenter 2016 for an evaluation of publication metrics, leading to the conclusion: *While publication metrics can provide compelling narratives, no single metric is sufficient for measuring performance, quality, or impact by an author. Publication data is but a single chapter in an author's academic and research story. Publication data alone does not provide a full narrative of an author's impact or influence, nor is it necessarily predictive of meaningful health outcomes that may have resulted from an author's research. Other sources include awarded grants, honors and awards, patents, intellectual property, outreach efforts, teaching activities, professional organization efforts, journal editorship, advisory board activities, mentoring efforts, and community engagement activities, to name a few.*

[16] https://dans.knaw.nl/en/deposit/information-about-depositing-data/before-depositing/file-formats?set_language=en

much less supported. The question is: who will be responsible for the infrastructure needed for academic publication of data: the publishers, research funding bodies or scientific institutions? There is an opportunity here to adopt open access before business models that lean the other way become dominant.

Access and licensing

Access literally refers to the possibility of entering something or somewhere. When you access something, you come in close proximity and you are allowed to interact with it, according to certain rules. If access is restricted, you may have to pay for it, or you need to be part of a certain workforce. And once you have gained access, restrictions may apply: only for a certain duration, only a certain interaction is allowed. If access does not cost money, it is called free; if access is not limited to certain people, it is called open; and if it costs money to access something, it is generally not called open.

When researchers want open access to resources, it is because they want to use them in a scientific workflow. That means that other researchers must have the same access to those resources, for performing replication, and that other institutions of society and the public in general must have access to them in order to make policies, check facts, educate themselves. And, ideally, businesses should have access to them in order to make money out of it.

Use case Thesaurus Linguae Graecae

(By Ernst Boogert, Protestant Theological University Amsterdam)

A good example of how restricted data ownership still works is the Thesaurus Linguae Graecae (TLG)[17], a database of more than 10,000 (ancient) Greek texts used by 2,000+ universities and institutions around the world. These texts have been gathered from 1972 onwards, were sold on CD-ROMs till 2008, and are currently accessible via a solid website only including advanced search tools and analytics[18]. It does not need any explanation that this quite complete collection of ancient Greek texts has academic value in itself. However, as soon as researchers want to conduct analyses on these texts themselves that are beyond the capabilities of the website, they are confronted with the fact that it is forbidden to download any portion of the texts or to run a custom script on them[19], even though it is permitted to conduct "text processing and text manipulation activities which are clearly identifiable as scholarly research."

The way how the data are presented on the website and the explicit prohibition to download materials, make any custom analysis on the bare data impossible. The main reason appears to be that the TLG made these texts available "at substantial cost" and that they are therefore copyrighted[20]. This poses the question whether ancient texts can be copyrighted at all. Nevertheless, the fact stands that the TLG provides texts that are of much interest for academic research, without providing the access needed. Researchers are therefore forced to use open alternatives like Perseus, despite the fact that the number of available texts is much lower[21].

[17] Thesaurus Linguae Graecae http://stephanus.tlg.uci.edu

[18] See for the history of the TLG: http://stephanus.tlg.uci.edu/history.php .

[19] "Downloading and/or commercial use or publication of these materials without authorization is strictly prohibited." http://stephanus.tlg.uci.edu/copyright.php . Further explanation of this restriction can be found in the TLG Site License Agreement: http://stephanus.tlg.uci.edu/site.php and http://stephanus.tlg.uci.edu/individual.php (especially article V).

[20] See the Site License Agreement, article I.

[21] See https://scaife.perseus.org . Perseus has currently only 1,178 texts available, which is much lower compared to the 10,000+ in the TLG. See for the most recent numbers the homepage of both websites.

From Open Access to FAIR

Handling research data is a complex configuration of rights and responsibilities, of freedoms and opportunities, of expectations and realisations. The first concepts of Open Access were focused on the accessibility, because traditional publishing made resources findable, but not accessible.

Now that more and more publications are open access, and data receive more attention, the concept of open access - or even Open Science - is widened to include anything that one can do with resources when and after accessing them. In other words, the need for clear usage licences grows. The Berlin declaration already contains the elements of a licence.

The scientific community is in the process of standardising the treatment of research data by means of the FAIR principles: it is not enough to provide open access to data as a one-off activity, but they should also remain Findable, Accessible, Interoperable and Reusable (Wilkinson 2016). To put it bluntly, opening up nonsensical or undocumented data has little value, and this is why the notions of Open and FAIR increasingly are used in tandem. The FAIR standardisation effort can be seen as a follow up on the Berlin declaration; at the same time, DANS and other trustworthy digital repositories (i.e. CoreTrustSeal certified[22]) with a long-term mission have always demanded that data should only be deposited with explicit contextual information, to enable interpretation and reuse.

Creative Commons: CC0 and open licences

Creative Commons[23] developed ways to share works, including data, more freely and more effectively. This is a reaction to copyright law which reserves the rights of the author of a work at the cost of access restrictions and usage limitations for its consumers. In the scientific community that balance is different. Freer, more effective sharing can be done by putting a work in the public domain using the CC0 tool[24], or by licensing it using one of six licences[25]. They all allow that a work may always be copied, distributed and displayed. They all require that the author is credited. The re-user must indicate the changes made (if allowed) and the re-user is not allowed to impose any additional restrictions (such as placing a work behind a paywall). Below we list them in order of increasing openness (see also fig.2).

All these are popular means within the research community and beyond for sharing both publications and data, because they all allow for copying and redistribution in any medium or format, provided that one credits the author appropriately, which is what scholars do per academic ethics.

- **CC-BY** (Attribution): Allows further: making derivative works, including for commercial purposes.
- **CC-BY-SA** (Attribution ShareAlike) Allows further: making derivative works, including for commercial purposes, provided that you distribute derivatives under the same licence.
- **CC-BY-ND** (Attribution-NoDerivatives) Allows further: redistribution for commercial purposes. Prohibits: making derivative works.
- **CC-BY-NC** (Attribution-NonCommercial) Allows further: making derivative works, but not for commercial purposes. Prohibits: redistribution for commercial purposes.
- **CC-BY-NC-SA** (Attribution-NonCommercial-ShareAlike) Allows further: making derivative works, but not for commercial purposes, provided that you distribute derivatives under the same licence.
- **CC-BY-NC-ND** (Attribution-NonCommercial-NoDerivatives): Prohibits: redistribution for commercial purposes and making derivative works.

[22] https://www.coretrustseal.org
[23] https://creativecommons.org
[24] https://creativecommons.org/share-your-work/public-domain/
[25] https://creativecommons.org/share-your-work/licensing-types-examples/

As indicated in fig. 2[26], CC0, CC-BY and CC-BY-SA are considered good options to distribute work (relatively) without restrictions. This is because none, or very limited conditions apply, conditions that do not restrict users in creating and sharing new work. Licences with conditions non-commercial or non-derivatives, or both, are considered as less open. While allowing access and free distribution these conditions restrict further reuse. Additionally, it may often be difficult to define what is considered as commercial, and what as a derivative, this uncertainty may act as another barrier for reuse. On the other hand, the existence of the NC licences may get works in the open that otherwise would not be published at all. The use case of the Hebrew Bible illustrates this.

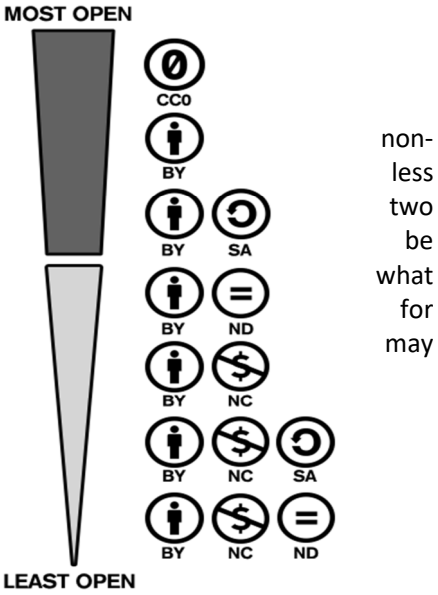


Fig. 2: Creative Commons diagram of CC0 and licences: from least to most open. Source: Creative Commons.

Use case Hebrew Bible

Large parts of humanities' research and data occur in the grey areas between subjective, creative expression (which is protected) and objective registration of facts (which is unencumbered). A case in point is the Hebrew Bible. More concretely, consider data set *Hebrew Text Database ETCBC4b* (Peursen 2015). It contains the plain text of the Hebrew Bible according to a transcription of a specific manuscript. This text has also been published as a printed book by the German Bible Society, together with a critical apparatus which comments on variants with other manuscripts and unclear stretches of text. The book is copyrighted by the German Bible Society. The creators of the *data set* are researchers of the ETCBC[27]. They have added linguistic annotations to every single word of the text. These form a large body of work, which took several people several decades. In fact, it corresponds to a big part of the research output of the ETCBC over the last forty years. In line with the practices of Open Science, the leader of the ETCBC, prof. Wido van Peursen, needs to publish all this as openly as possible (Roorda 2018). But the German Bible Society, who has invested in the text, but also in the linguistic annotations, claimed authority over the publishing of text *and* annotations. For many years, this was one of the reasons that this data set was not openly available. For example, an earlier version, the *ETCBC3* (Talstra 2011), also archived at DANS, is only available upon request.

The way out of this impasse went as follows: in 2014, after completing a project that resulted in the website SHEBANQ[28] that gives access to the text and linguistic annotations of the Hebrew Bible, the ETCBC and DANS observed (1) that academic rules demanded open publication of this work, (2) that there was no legal basis for copyright on the plain text of the Hebrew Bible, (3) that the linguistic annotations were the intellectual property of the university, not the German Bible Society, (4) relations with the German Bible Society had always been friendly and should remain so. At that point they decided to license the whole data set with a standard CC-BY-NC licence. A year later the German Bible Society formally endorsed the new status quo, observing that it has not damaged its business, and that the choice of licence was compatible with its interests. Thanks to this accepted solution, the ETCBC has experienced a boost in visibility, as several parties are now using its data and building new end user apps on top of it, all properly and openly licensed.

[26] https://creativecommons.org/share-your-work/public-domain/freeworks/
[27] http://etcbc.nl
[28] https://shebanq.ancient-data.org

Looking at Creative Commons' CC0 and licences, what fits the open access movement best when it comes to research data? The Berlin statement nearly matches the CC-BY licence except for the conditions to use a digital medium, and the restriction that you may only make a limited number of copies for personal use.

CC0 is the most easy, clear way to distribute both research data and publications. This is because all rights are waived were they would have applied and no conditions apply. CC0 is particularly fitting for data since data in the public domain should not be copyrighted, hence remains unencumbered. However, the lack of any condition might be a downside in the scientific community where citing is standard practice. This condition is essential according to the Berlin declaration on open access, but does this need to be part of a licence? In science, it is common practice to cite sources as the provenance of data is highly valued.
While research data in the public domain can essentially become orphaned, it is unlikely that this will be the case.

CC-BY on the other hand is better suitable for publications than for research data. By applying CC-BY one is actually stating that copyrights apply to a work (which usually is the case with publications), because by using this type of licence the owner grants permission to do some of the actions protected by copyright, permission that otherwise should have been asked for on a case by case basis. This licence is therefore not as suitable for data, because in the case of open data any explicit restriction may render it useless for proper scholarly use. This holds for all CC licences.

Open access at DANS

From its inception, DANS has acted as a proponent of open access, with the slogan "Open if possible, protected where necessary". One of the core services of DANS is a long-term data archive (EASY)[29], focused on permanent access to digital research data from mainly the humanities and social sciences. The developments here show that open access is becoming more and more the norm for research data, but it is a gradual transition.

In the data archive, data sets are deposited by individual researchers as well as institutes. In both cases the depositing party is entitled to decide under what conditions the data should be archived. The most open option has always been set as the default choice. In recent years, the Creative Commons CC0 tool and the Creative Commons licences have become popular means amongst researchers and funders to apply to their data sets. December 2014, DANS implemented the CC0 tool, which is currently set as the default choice. On November 13 this year, the amount of data sets under CC0 already totals 4333. The Creative Commons licences are currently being implemented and a steady rise for the use of the CC-BY licence is expected.

DANS policy is to offer customers the best suitable option with regards to open access. CC0 is regarded as best choice followed by CC-BY and CC-BY-SA. As shown above, these are the top three most open access options that Creative Commons offers. DANS is promoting CC0 as best fit, but deliberately offers these alternatives since experience shows that CC0 is often too big a step to stake yet. DANS often experiences that CC0 is a well-known concept, but that the details are not fully understood or that depositors feel there is no need to opt for this type of openness. This often results in depositors choosing a less open option. It is clear that we are in a transition phase that will need time. DANS experienced that guidance in this transition is very important. For example, it helps to inform depositors that while applying CC0 their data is still subject to the academic code of conduct for proper accreditation. Without proper guidance, an opportunity for open access data may be missed. At this moment, customers may not always choose CC0 where it is best choice, but any open access alternative is seen as a positive outcome.

[29] EASY https://easy.dans.knaw.nl/ui/home

Use case: Unearthing and opening up of archaeological data

When the e-depot for Dutch Archaeology was set up at DANS more than 10 years ago, restricted access was one of the requirements of the archaeological depositors who were afraid of mis-use of their data. The access category "archaeologist only" was created as a specific restricted access category for archaeological data in EASY. This category and the already existing "restricted access" were widely used for archaeological data.

Nowadays these categories are quickly getting smaller: the data which were restricted to professional archaeologists before, or only opened up to individuals, are now accessible for registered EASY users or open for everyone/CC0. This counts for almost 80% of the over 38.000 archaeological data sets. Sharing of data gets more common and trust is growing. Data can still be protected by an embargo period or archived under "restricted access" if archaeologists see reason do so. DANS will no longer offer group access to archaeologists, this feature will disappear in the new ingest system for depositors. Open access will be the norm.

Dutch archaeologists increasingly trust others who want to use their data, not only in the Netherlands but also abroad. In 2016, the ARIADNE portal was launched which provides search access across the integrated data collections of European archaeological data providers (Hollander 2017). The most important outcome of two international surveys held in advance was that researchers expected a data portal with the capability to search and "mine" distributed digital archives for relevant data. Sharing data is no longer threatening academic careers or commercial interests, in the contrary. Archaeologists deposit and share their completed research results to boost their work's visibility and findability.

Presently, over 3,000 archaeological datasets are CC0 available via EASY and the use of Creative Commons licences has already taken a flight. An example is the data of the project Portable Antiquities of the Netherlands (PAN)[30] being deposited in EASY this year. PAN provides an online database of finds of metal artefacts by private metal detectors. DANS takes care of the long-term archiving of this collection of already reported 41,000 objects. The exact location and the personal details are not made public, but the descriptive information, images and reference types are all open access available under the CC-BY-NC-SA licence.

Conclusions

Open access is central to the Open Access movement that started over fifteen years ago. There has always been a focus on publications when it comes to open access, while the principle should apply to data as well. There is no definition of open access for research data, but in a very broad sense it could read: accessible with minimal conditions for reuse. The benefits of open access, for both publications and data, are obvious: when science is open for everyone, better science and faster societal benefits are the result.

When it comes to publications, open access has its effect on the traditional business model, a model that is currently changing. For data there are three challenges in working towards open access: awareness, ownership and data management. Awareness needs to be raised among researchers, to see the value of data and the relevance in publishing data. Ownership can form an obstacle for publishing data as researchers understandingly wish to benefit further from their own work. Embargo periods and data citation, a practice that is taking more and more hold, could be a solution to this. The final challenge is in data management. Where there is a long-standing tradition of publishing publication, this is still developing for data. Data need processing before they can be published and may need access control.

What does open access mean for research data? We cannot provide a concrete answer.

We argue that for data, the focus point is on reuse, instead of on accessibility. In this respect both the value and licensing of data becomes relevant. The FAIR principles are an important means by which data can be made fit for reuse and it seems logical that they should be interlinked with the concept of open access. The CC0 and licences by Creative Commons are popular and suitable means to enhance reuse. In particular, CC0, CC-BY and CC-BY-SA are regarded as best options in term of reuse since these option place minimal restrictions on reuse. We argue that CC0 is best for data since open data should not be licensed. On the other hand, as citation is a standard

[30] https://www.portable-antiquities.nl/pan/#/public/about

requirement in science (and defined as such in the Berlin declaration) it is argued that CC-BY is then a good fit for publications and perhaps we need something similar for data. However, one can wonder if we need a licence to bind researchers to a common academic practice.

Since its inception DANS has always been a propagator of open access. Four years ago, DANS implemented CC0 in the digital data archive EASY and is now implementing the Creative Commons licences. We realise that open access for data is a gradual transition. While we stimulate archiving data CC0, all openly archived data are currently seen as a success.

In conclusion, open access for research data is much more a matter of reuse than of accessibility. Therefore, it seems suitable that the FAIR principles and the CC0 tool or Creative Commons licences, are part of defining this concept. The current challenges together with the transition we currently experience are all steps in the way to finding the meaning of open access for research data.

References

Carpenter 2016. Carpenter, Christopher R. *Using Publication Metrics to Highlight Academic Productivity and Research Impact*. Acad Emerg Med. 2014 Oct; 21(10): 1160–1172. Doi: 10.1111/acem.12482 Author's copy: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4987709/

Hollander, H. 2017 Saving Treasures of the World Heritage at the Digital Archive DANS, Internet Archaeology 43. https://doi.org/10.11141/ia.43.9

Peursen 2015. Peursen, W.T. van (Eep Talstra Centre for Bible And Computing, VU University Amsterdam); Sikkel, C. (Eep Talstra Centre for Bible And Computing, VU University Amsterdam); Roorda, D. (Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Sciences) (2015): *Hebrew Text Database ETCBC4b*. DANS. https://doi.org/10.17026/dans-z6y-skyh

RDA 2014. *The data harvest - How sharing research data can yield knowledge, jobs and growth*. *An RDA Europe report*. 2014 https://www.rd-alliance.org/sites/default/files/attachment/The%20Data%20Harvest%20Final.pdf

Roorda 2018. Roorda, D. *Coding the Hebrew Bible*. Research Data Journal for the Humanities and Social Sciences, 2018, doi:10.1163/24523666-01000011.

Talstra 2011. Talstra, E. (ETCBC, VU Amsterdam) (2011): *Text Database of the Hebrew Bible ETCBC3*. DANS. https://doi.org/10.17026/dans-x8h-y2bv

Wilkinson 2016. Wilkinson et al. *The FAIR Guiding Principles for scientific data management and stewardship*. Scientific Data, volume 3, 2016, doi:10.1038/sdata.2016.18

Open Data engages Citation and Reuse: A Follow-up Study on Enhanced Publication*

Dominic Farace, Jerry Frantzen, GreyNet International, Netherlands
Joachim Schöpfel, University of Lille, France

Introduction

In 2011, GreyNet embarked on an Enhanced Publications Project (EPP) in order to link its collection of full text conference papers with accompanying research data. The initial phase in the study dealt with the design and implementation of an online questionnaire among authors, who were published in the International Conference Series on Grey Literature[1]. From 2012 onwards, subsequent phases in the project dealt with the acquisition, submission, indexing, and archiving of GreyNet's collection of published datasets now housed in the DANS EASY[2] data archive.

In 2017, GreyNet's Enhanced Publications Project was further broadened to include a Data Papers Project[3], where emphasis focused on describing the data and methods applied in gathering it rather than analyzing it. As such, the data paper signals data sharing and in this way promotes both data citation and the potential reuse of research data including any limitations on its potential reuse. This is in line with the FAIR Guiding Principles[4] for scientific data management and stewardship.

Available results from the Data Papers Project presented last year at GL19 concludes where this study commences. Here, we now seek to demonstrate the reuse of survey data collected in 2011 combined with survey data that was newly collected via an online questionnaire carried in May/June of 2018. The results of this study were expected to demonstrate an increased willingness among GreyNet authors to share their research data – this in part due to GreyNet's program of enhanced publication embedded in its workflow over the past seven years. The study sought to provide an example of the reuse and further comparison of the results of survey data, which can be incorporated in GreyNet's program of training and instruction. However, statistics on data citation and referencing are less likely expected to provide as yet indicative results.

Enhanced Publication

In 2011 the guiding principle for enhanced publications was that they inherently contribute to the review process of grey literature as well as the replication of research and improved visibility of research results in the scholarly communication chain. However, in 2018 and in light of the FAIR Principles there is one modification. Emphasis moves from replication to reuse. Further, GreyNet's original project on enhanced publications in 2011 as with this follow-up study in 2018 is intended not to be seen solely as a descriptive case study, but rather as a use case – one which can serve other communities of practice in the field of grey literature.

This study may no doubt be of interest to those who contributed over the years their full texts, research data, and metadata to GreyNet's collection of enhanced publications. However, to show the value of this use case beyond our own community, we are now required to demonstrate how GreyNet's open data engages citation and reuse.

FAIR Data Principles

In follow-up to the initial implementation of the Enhanced Publications Project in 2012, emphasis now has come to rest on the publication process of grey literature in which the individual enhanced publication becomes the beneficiary.

Two current developments in the field of information have considerable impact on grey literature – one being the FAIR Data principles in which data is to be findable, accessible, interoperable, and reusable; and the other, which deals with the publication of data papers defined as "Scholarly publications of a searchable metadata document describing a particular online accessible dataset or a group of datasets published in accordance to standard academic practices. As such, data papers represent a scholarly communication approach to data sharing"[5]. It is by way of a data paper that the FAIR Data principles are implemented. The FAIR Data Principles formulated in 2014

* First published in the GL20 Conference Proceedings, February 2019.

are related to the data and/or datasets deposited in archives. Prior to FAIR was the Data Seal of Approval[6], which was conferred upon the archive and not the individual data or dataset.

Now, in demonstrating how the data from GreyNet's collection of conference papers are findable, we can say that they have been deposited and are preserved in a national data archive. In demonstrating that GreyNet's research data is openly accessible, we can point out that the creators have waived their rights via Creative Commons Zero (CC0) thus allowing optimal access. And, in demonstrating that GreyNet's data are interoperable, we can refer to the rich metadata attributed and linked to the data and datasets including ORCiD and DOI persistent identifiers. However, to demonstrate the potential for how GreyNet's deposited data and datasets can be reusable, research was needed. And this project was born.

Author Survey on Open Data

This study sought to demonstrate the reuse of survey data collected in 2011 combined with survey data that was newly collected via an online questionnaire. In order to do so, a selection of questions from the 2011 Survey was joined with newly formulated questions in constructing the 2018 Questionnaire**.**

Survey Population

The selection used to define the population of the 2018 survey is much in line with the selection used in 2011. It was assumed that in this way that the data collected would allow for insight into changing attitudes and practices within GreyNet's research community and as such would be of more interest to other grey literature communities.

Survey Population	First Authors	Survey Recipients	Survey Respondents	Survey Results %
2011	162	95	50	52,6%
2018	115	94	44	46,8%

As shown in the chart above, the population of the 2018 survey was selected from among 115 first authors in the International Conference Series on Grey Literature. The selection comprised the respondents to GreyNet's 2011 Survey on Enhanced Publications along with first authors in the GL-Conference series from 2012 to 2018. Once the survey population was further reduced, either because an author's email address was currently unavailable, the author had retired or had since moved to another field, the questionnaire was then sent out to the remaining 94 authors/researchers via personalized emails. The final results of this study rest on the responses of 44 survey respondents, which accounts for a near 47% response rate.

Survey Questionnaire

The data collected based on the responses of those forty-four authors/researchers was to ten questions, two of which were open-ended. Six of the questions were taken from the questionnaire carried out in 2011, which dealt with the author's own empirical research data, its availability, the formats in which it appears, and the author's willingness to archive it and make it openly accessible. The four additional questions deal with the respondent's citation and reference to data, their use of data journals in carrying out search and retrieval, and whether they (co)authored a data paper or data article. The research data – long tail[7] in contrast to big data – was collected via SurveyMonkey between May 18th and June 15th 2018, where it remains stored along with a copy in .ods format[8] in the DANS EASY Archive.

Comparison of Survey Results 2011-2018

In the following tables, responses to six of the 2018 survey questions are shown compared with the responses to the same questions that appeared in the 2011 questionnaire. It should be noted that the numerical order of the questions follows that of the 2018 questionnaire.

Q1. Does one or more of your conference papers in the GL-Series base its findings on empirical or statistical data?

2011	Question 1	2018	Question 1
Yes	60%	Yes	70%
No	40%	No	30%
	100%		100%

Q2. If so, would these data and/or datasets still be available in part or whole for archiving purposes?

2011	Question 2	2018	Question 2
Yes	54,1%	Yes	60%
No	45,9%	No	20%
Other	0%	Other	20%
	100%		100%

Q3. Would you be willing to submit data, datasets, or subsets to DANS that would in turn be linked to their existing metadata records?

2011	Question 5	2018	Question 3
Yes	48,9%	Yes	51,2%
No	6,7%	No	9,3%
	Uncertain		Other
	44,4%		39,5%
	100%		100%

Q4. If so, would you prefer that GreyNet entered your (retrospective) data and/or datasets in DANS, or would you prefer to do this directly?

2011	Question 6	2018	Question 4
GreyNet	44,7%	GreyNet	41,5%
Self	18,4%	Self	41,5%
	No Preference		**Other**
	36,9%		17%
	100%		100%

Q5. What kind of data and data formats have you used/are using in your research?

2011	Question 9	2018	Question 5
	Specific		**Specific**
	44%		75,9%
	General		**General**
	38%		18,9%
	N/A		**N/A**
	18%		5,2%
	100%		100%

Q10. Please enter your name, email address, and any other comments or recommendations that would be of benefit to this survey

2011	Question 10	2018	Question 10
Answered	84%	Answered	73%
Skipped	16%	Skipped	27%
Total 50 Respondents	100%	Total 44 Respondents	100%

A comparison of the results of the six questions that were repeated in the two questionnaires held seven years apart indicate that papers in the GL-Conference Series have since increased in the amount of empirical data they contain and that these data are available for archiving purposes. However, the results indicate that only a small increase is shown in the number of respondents willing to submit their data for archiving purposes. A marked increase shows that the respondents would prefer to archive their own data rather than that GreyNet do so on their behalf. The kinds of data formats, which the respondents use in their research has significantly increased since the earlier survey. And, finally it can be observed that the number of respondents willing to provide contact details has decreased since the 2011 questionnaire.

Results based on Responses to the new Questions in the Survey

Question 6

Have you ever cited or referenced data(sets) in one or more of your publications?

Answered: 43 | Skipped: 1
Yes 44%
No 44%
Other 12%

Question 7

When doing research, have you or a colleague ever reused data?

Answered: 40 | Skipped: 4
Yes 50%
No 50%
Other -

Question 8

Do you at times include data papers or data journals in your browse and search strategy?

Answered: 43 | Skipped: 1
Yes 39%
No 56%
Other 5%

Question 9

Have you ever cited or (co)authored a data paper or data article?

Answered: 42 | Skipped: 2
Yes 24%
No 67%
Other 9%

The results of the two questions dealing with the citation and reuse of data by the respondents clearly indicate an even yes-no response rate to both questions. However, the responses to the questions as to whether data papers are included in the authors browse and search strategy and whether they have authored or coauthored a data paper/article are significantly non-affirmative.

Analysis of the Survey Data 2011-2018

An analysis of the survey data demonstrate significant change in the responses to three of the questions in 2011 compared with the same questions repeated in 2018. More data and/or datasets are available in part or whole for archiving purposes 54% → 60% (p =.005), more authors prefer entering their data and/or datasets directly in the DANS Archive 18% → 42% (p =.05), and the specificity of the data formats listed by the respondents increased significantly 44% → 76% (p=.1). As to the other questions repeated in the survey, little or no change was observed except in the final open-ended question requesting contact details.

Some Conclusions and Further Comments

This follow-up study demonstrates a community of practice moving further to open data. There is evidence of more data awareness and data literacy among GreyNet's authors and researchers. However, one must not overlook the fact that not all papers in the GL-Conference Series are based on empirical data and not all data can be shared for reasons of confidentiality, embargo, licensing, (lack of) policy directives, or sheer hesitance by the author/researcher.

Furthermore, this study demonstrates that research data can be published prior to the research paper, allowing for more immediate citation, reuse, and usage statistics. Also, a data paper[9] focusing on the research data that includes persistent identifiers such as ORCiDs, DOI's and other hyperlinks, can further add to the increase in citation, reuse, and usage statistics. Finally, the results of this study have already been incorporated in GreyNet's series of workshops[10] and training on data and data papers offered both within and outside its community of practice.

References

[1] Farace, D. et al. (2012). Linking full-text Grey Literature to underlying research and post-publication data: An Enhanced Publications Project 2011-2012. – In: The Grey Journal, Volume 8, Issue 3, 2012. – pp. 181-189. – ISSN 1574-1796

[2] GreyNet's collection of published datasets in the DANS EASY data archive, https://easy.dans.knaw.nl/ui/?wicket:bookmarkablePage=:nl.knaw.dans.easy.web.search.pages.PublicSearchResultPage&q=greynet

[3] Farace, D., Frantzen, J. and Smith, P.L. (2018). Data Papers are Witness to Trusted Resources in Grey Literature: A Project Use Case. – In: The Grey Journal, Volume 14, Issue 1, 2018. – pp. 31-36. – ISSN 1574-1796

[4] FAIR-Data Principles https://www.force11.org/group/fairgroup/fairprinciples

[5] https://en.wikipedia.org/wiki/Data_publishing#Paper

[6] https://www.datasealofapproval.org/en/

[7] Long tail of research data https://www.radar-projekt.org/display/RE/Glossar#Glossar-Longtailofresearchdata

[8] https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:110917/tab/2

[9] Farace, D. and Schöpfel, J. (2018). Data from "Open Data engages Citation and Reuse: A Follow-up Study on Enhanced Publication". – In: The Grey Journal, Volume 14, Issue 3, 2018. – pp. 149-150. – ISSN 1574-1796

[10] http://greynet.org/greyforumseries.html

The Q-Codes: Metadata, Research Data, and Desiderata, Oh My! Improving Access to Grey Literature*

Melissa P. Resnick,
University of Texas Health Science Center at Houston, Houston, TX, USA

Ashwin Ittoo, HEC Management School, University of Liège, Belgium

Marc Jamoulle and Marc Vanmeerbeek,
Department of General Practice, University of Liège, Belgium

Frank S. Shamenek, Consultant, New York, USA

Chiehwen Ed Hsu, College of Management, National CK University, Tainan, Taiwan

Robert Vander Stichele, Heymans Institute of Pharmacology, University of Ghent, Belgium

Julien Grosjean and Stefan Darmoni,
Department of Information and Medical Informatics (D2IM), University of Rouen, France

Elena Cardillo, Institute of Informatics and Telematics, National Research Council, Italy

Miguel Pizzanelli, Department of Family and Community Medicine,
University of the Republic of Uruguay

***Abstract**:*

***Problem/Goal:** In GL19's "Indexing grey literature in General Practice: Family Medicine in the Era of Semantic Web," Jamoulle and colleagues (Jamoulle et al., 2018) propose the use of a relatively new terminology (3CGP) to allow for the indexing and retrieval of (GP/FM) knowledge which otherwise would be lost, or difficult to locate. Though designed to meet Cimino's (Cimino, 1998) twelve desiderata for the design of a controlled healthcare vocabulary, Jamoulle and colleagues (Jamoulle et al., 2018) acknowledge that a detailed requirement by requirement evaluation of 3CGP was not performed. The goal of this paper is to evaluate the Q-Codes component of the 3CGP terminology, in detail, with each of Cimino's twelve desiderata.*

***Research Method/Procedure:** In our work, we will focus on qualitative analysis, whereby our taxonomy, the Q-Codes, and in particular, its vocabulary satisfies a standard set of desiderata. Qualitative analysis provides a simple and yet effective way to assess the Q-Codes taxonomy's quality. We will briefly describe each of the desiderata and discuss how our taxonomy satisfies each one of them (or not).*

***Anticipated Results of the Research:** The qualitative evaluation is intended as an initial stage, which focuses on the Q-Codes taxonomy's contents, namely, its vocabulary (e.g. terms and definitions). Our aim with the qualitative evaluation is to investigate whether our proposed taxonomy, and in particular its vocabulary, satisfies a set of desiderata. This will enable us to determine whether the knowledge acquisition and (part of) the conceptualization steps of our ontology development process have been performed correctly. We consider that validating our vocabulary against a set of well-defined desiderata is paramount before evaluating other aspects of the taxonomy (such as the relations). As a set of desiderata, we chose that proposed by Cimino in his seminal study entitled "Desiderata for controlled medical vocabularies in the twenty-first century" (Cimino, 1998). These desiderata ensure that our taxonomy can be successfully deployed and exploited in actual GM/FM applications / activities, such as indexing grey literature. The desiderata define a set of (desired) characteristics that (ideally all) standard medical vocabularies should satisfy. Thus, these desiderata help in alleviating inter-operability issues, with the use of common standards ensuring the efficient integration of our taxonomy with other medical vocabularies and resources (taxonomies, ontologies). From the results of this study, improvements can be made to the Q-Codes component of, and thus, the 3CGP terminology. This, in turn, improves the ability to index the grey literature with the 3CGP terminology, providing greater access to needed information.*

***Indication of costs related to the project:** This project has not been funded.*

* First published in the GL20 Conference Proceedings, February 2019.

Introduction

In recent years, grey literature has become more important, especially in research areas, such as General Practice/Family Medicine (GP/FM) (Jamoulle, Grosjean, et al., 2017). Grey literature has different meanings to different people. Thus, there are many definitions for grey literature. Three of these definitions will be presented below.

Denda (Denda, 2002) notes that "grey literature is a body of information that is often not identified through standard acquisitions procedures or retrieved through research tools such as indexes, catalogs, or databases." Some state that grey literature is "material that is difficult to catalogue" (Mahood, Van Eerd, & Irvin, 2014; Tillett & Newbold, 2006). For others, grey literature is defined as: "that which is produced on all levels of government, academics, business and industry in print and electronic formats, but which is not controlled by commercial publishers" (Bellefontaine & Lee, 2014; New York Academy of Medicine, 2018; Paez, 2017; Pappas & Williams, 2011). This third and last definition will be used to define grey literature for this research.

Given this definition, grey literature can encompass many types of materials including, but not limited to, dissertations, conference proceedings, reports, book chapters, magazine articles, newsletters, blogs, wikis, conference abstracts, and preprints (Mahood et al., 2014; TextRelease, n.d.-b). Mahood and colleagues (Mahood et al., 2014) point out that GreyNet provides an extensive list of materials considered grey literature on their website (TextRelease, n.d.-a). However, most of these resources are difficult to find, as they lack bibliographic information, such as author/publisher or volume/issue/page numbers (Mahood et al., 2014) and are not even indexed (Denda, 2002; Mahood et al., 2014). The lack of bibliographic information or metadata, especially indexing terms, can often lead to loss of information. The use of a terminology, one type of metadata, can assist in improving this situation.

During the GL19 conference, a relatively new terminology, Core Content Classification in General Practice Family Medicine (3CGP), was proposed to allow for the indexing and retrieval of GP/FM knowledge, which otherwise would be lost or difficult to locate (Jamoulle et al., 2018). This terminology is composed of two components: (i) the International Classification of Primary Care (ICPC-2), used for clinical concepts, and (ii) the Q-Codes taxonomy, used for contextual concepts (Jamoulle et al., 2018). The remainder of this paper is concerned with the Q-Codes taxonomy component of 3CGP.

The Q-Codes taxonomy is comprised of eight top-level categories, as shown in Table 1 (Jamoulle et al., 2018, 2017). Each of the eight top-level categories has a simple hierarchy of up to three levels. These eight simple hierarchies together form the taxonomy (Jamoulle, Grosjean, et al., 2017). In building this taxonomy, Jamoulle and colleagues (Jamoulle et al., 2018) have attempted to meet the twelve desiderata (guidelines) proposed by Cimino in 1998 (Cimino, 1998).

Table 1. Q-Codes top-level categories overview

Q-Code top-level category	Label	Examples of covered topics
QC	Patient's category	age, gender issues, abuse
QD	Family doctor's issue	communication, clinical prevention, medico legal issues
QE	Medical ethics	bioethics, professional ethics, info ethics
QH	Planetary health	environmental health, biological hazards, nuclear hazards
QP	Patient issue	patient safety, patient centeredness, quality of health care
QR	Research	research methods, research tools, epidemiology of primary care
QS	Structure of practice	primary care setting, primary care provider, practice relationship
QT	Knowledge management	teaching, training, knowledge dissemination

Over the past decades, terminologies have been constructed for several reasons including, but not limited to: capturing clinical findings, natural language processing, indexing medical records, indexing medical literature, and representing medical knowledge (Cimino, 1998). Terminology users tried to use various standard terminologies, but found this difficult, as no one standard terminology was appropriate for all of their needs (Cimino, 1998). Thus, users began to express a need for requirements for terminologies.

By the early 1990s, terminology researchers began writing about various requirements believed useful in building terminologies (Cimino, 1998). As stated by Cimino (Cimino, 1998), researchers have gone past discussing only the definitions in a vocabulary and started discussing the "deeper representational aspects" or other components of a vocabulary. In his seminal paper titled "Desiderata for Controlled Medical Vocabularies in the Twenty-First Century", Cimino (Cimino, 1998) presents these requirements or desiderata, which include: Vocabulary Content, Concept Orientation, Concept Permanence, Non-Semantic Concept Identifiers, Polyhierarchy, Formal Definitions, Rejection of "Not Elsewhere Classified" Terms, Multiple Granularities, Multiple Consistent Views, Context Representation, Graceful Evolution, and Recognized Redundancy.

Our motivations for selecting this set of desiderata are as follows. First, the desiderata have been articulated and formulated based on actual requirements of medical informatics application developers and end-users. Thus, these desiderata ensure that the Q-Codes taxonomy can be successfully deployed and exploited in actual GM/FM applications/activities, such as indexing of the grey literature. Second, the desiderata define a set of desired characteristics that ideally all standard medical vocabularies should satisfy. Thus, these desiderata help in alleviating inter-operability issues with the use of common standards, ensuring the efficient integration of this taxonomy with other medical vocabularies and resources. These desiderata will be described in more detail below.

Goal

Though designed to meet Cimino's (Cimino, 1998) twelve desiderata for the design of a controlled healthcare vocabulary, Jamoulle and colleagues (Jamoulle et al., 2018) acknowledge that a detailed requirement by requirement evaluation of 3CGP was not performed. The goal of this paper is to evaluate the Q-Codes component of the 3CGP terminology, in detail, with each of the twelve desiderata proposed by Cimino (Cimino, 1998).

Methods

In our work, we focused on qualitative analysis, whereby our taxonomy, the Q-Codes, and in particular, its vocabulary, satisfies a standard set of desiderata. Qualitative analysis provides a simple, and yet, effective way to assess the Q-Codes taxonomy's quality.

A copy of version 2.5 of the Q-Codes taxonomy was downloaded from: www.3cgp.docpatient.net/ and used for analysis. The twelve desiderata were obtained from Cimino's 1998 paper titled "Desiderata for controlled medical vocabularies in the twenty-first century" (Cimino, 1998).

The Q-Codes taxonomy was evaluated desideratum by desideratum. This analysis was performed as follows: (1) each desideratum was read, and (2) the taxonomy was examined for the presence or absence of the desideratum. In the next section, we will briefly describe each of the desiderata and discuss how the taxonomy satisfies each one of them (or not).

Results and Discussion

Desideratum 1: Content

The main issue to address concerning the Content Desideratum is the need for a systematic, explicit, and reproducible method for expanding content (Cimino, 1998).

Cimino (Cimino, 1998) suggests two main approaches for increasing content. In the first approach, all atomic units (e.g. single-word terms) of a domain terminology, for example GM/FM, are enumerated. Users are then allowed to combine them in order to compose more complex multi-word terms (Côté & Robboy, 1980). The main benefit of this approach is that by allowing compositional extensibility, it facilitates domain-coverage (D. A. Evans, Rothwell, Monarch, Lefferts, & Cote, 1991). However, this approach suffers from several limitations. First, identifying all domain-

specific atomic units is a nontrivial endeavor. Second, the composition of individual atomic units should preserve the semantics. In other words, the meaning of the atomic unit and of the resulting more complex unit should not be distorted with the composition. In addition, some compositions may yield illogical units, for example, combining the two terms "AIDS" and "flu" into "AIDS flu." Specific grammatical rules need to be defined to prevent such combinations.

In the second approach, contents (e.g. terms) are added as they are encountered in the various data sources (e.g. conference abstracts). Compared to the previous approach, one does not attempt to systematically anticipate and enumerate all possible terminological combinations, for example, by listing all the possible types of fractures (simple, complex, hair-line) for each possible bone. Instead, one would add terms corresponding to the most common/frequent ones (e.g. found more frequently in the data sources analyzed), and then add more complex terms as, and when, they are needed or encountered in the data sources. The main benefit of this approach is that it avoids the unnecessary generation of large numbers of terms occurring through combinatorial explosion and the enumeration of nonsensical combinations.

In the Q-Codes, the second approach has been adopted. Terms (simple and complex) are added as, and when, they are needed, and when they are encountered in the relevant data sources, such as conference abstracts.

Desideratum 2: Concept Orientation

According to most, if not all, researchers in medical informatics and in knowledge representation in general, the fundamental unit of symbolic processing is the concept. A concept can be defined as a mental representation of an object within a specific domain; an object refers to anything perceived or conceived (ISO 9000:2015). Thus, in a given specific domain, a concept embodies and conveys a precise meaning (D. A. Evans, Cimino, Hersh, Huff, & Bell, 1994; David A. Evans, 1988; Lindberg, Humphreys, & McCray, 1993; Rassinoux, Miller, Baud, & Scherrer, 1996; Volot et al., 1993). Concepts are lexically realized as terms (simple one-word terms or more complex multi-word terms). The collection of terms in a given domain is part of its vocabulary. Concept orientation means that terms must correspond to at least one meaning (nonvagueness) and no more than one meaning (unambiguity). Concerning the issue of unambiguity, some authors, such as Moorman and colleagues (Moorman, van Ginneken, van der Lei, & van Bemmel, 1994) argue that ambiguity can be allowed as long as the unequivocal meaning is preserved based on the term's usage in a given context.

A total of 182 single- and multi-word terms comprise the Q-Codes (Jamoulle, Grosjean, et al., 2017). Each one of these 182 terms was given only one definition (Jamoulle & Resnick, 2016). This, in turn, meets both the nonvagueness and the unambiguity criteria for Concept Orientation.

Desideratum 3: Concept Permanence

The desideratum of Concept Permanence follows directly from that of Concept Orientation (desideratum 2 above). Concept Permanence requires that once a concept has been created, its meaning is immutable, i.e. it cannot be changed or violated (Cimino, 1998). This condition of semantic immutability holds even if the concept's preferred name changes or if the concept is marked as inactive, deprecated or archaic (Cimino, 1998). For instance, consider a concept with the name "pacemaker". In this case, the concept's meaning does not change even if it is renamed to "implantable pacemaker". Conversely, consider a concept with name "non-A non-B hepatitis". Here, one cannot simply rename it to "hepatitis C" as "non-A non-B hepatitis" is not a synonym of "hepatitis C". In this situation, we cannot assert with certainty that someone with "non-A non-B hepatitis" is definitely suffering from "hepatitis C". Thus, such a renaming entails an alteration in the meaning.

In addition to semantic immutability, Concept Permanence also demands that concepts are not deleted if they are inactive or deprecated. Instead, they should be flagged as such.

Since the latest version of the Q-Codes taxonomy was recently completed (Jamoulle, Grosjean, et al., 2017), none of the current terms have become old or inactive. As the taxonomy evolves from one version to the next, extensive records will be kept so that older or inactive terms will be flagged and not removed. These records will help to ensure that the Q-Codes will continue to meet the Concept Permanence desideratum.

Desideratum 4: Non-Semantic Concept Identifier

Each concept should be assigned a unique identifier. In the simplest case, the concept's name also serves as its unique identifier. However, the main drawback of this strategy is that it hinders subsequent modifications to the concept's name.

Another approach is to assign a hierarchical code which reflects the position of the term in the hierarchy (Cimino, 1998). One advantage to this approach is that, with some knowledge of the hierarchy, the codes can become readable by humans, and thus, hierarchical relationships can be understood (Cimino, 1998).

A second advantage is that a hierarchical code can be used to search for all members of a particular class. For instance, searching for "QC1" can allow a user to find all of the terms that belong to "QC1 age group" (e.g. "QC11 infant", "QC12 child", "QC13 adolescent", etc.). However, the difficulty with this method of searching for terms in the same class arises when a term appears in more than one class or place in the hierarchy (Cimino, 1998).

During the creation of the Q-Codes, each term was assigned a hierarchical code (e.g. "QR31 qualitative study"). With these hierarchical codes, one can see the relationships between the terms. For example, "QR3 research method" is the broader term encompassing the narrower term "QR31 qualitative study".

Although these hierarchical codes are easy for humans to understand and use, there is a major limitation to systems such as this. As noted by Cimino (Cimino, 1998), hierarchical coding systems can "run out of room." In fact, the hierarchical coding system utilized by the Q-Codes has currently "run out of room."

In the hierarchical coding system employed by the Q-Codes, only nine separate terms are possible for the first level under any top-level category. For example, at the first level under the top-level category "QR research", nine separate terms labeled QR1, QR2, QR3 to, and including, QR9 are possible. Next, terms at the first level (e.g. QR1) can have only nine terms (e.g. QR11, QR12, QR13, etc.). This is also the case for any term on the second and successive levels. However, there is no limit to the number of levels for each top-level category.

This limitation can be solved in at least three ways. The first solution is to expand the numbering at each level of the hierarchical code. For example, the first level terms can be labeled QR01 to and including QR99, thus, allowing for 99 terms on this level. However, in theory, this delays the point in time at which the expanded hierarchical coding system will, again, "run out of room."

A second solution is to represent the hierarchies with links between parents and children (Cimino, 1996b). For example, there would be a link between "research method" (the parent) and "qualitative study" (the child), instead of using the unique identifiers "QR3" and "QR31" respectively.

A third solution involves providing tree addresses for each term, like MeSH and the Gabrielli Nomenclature (Cimino, 1996b). Cimino (Cimino, 1996b) points out that tree addresses provide arbitrary "length and breadth." In future versions of the Q-Codes taxonomy, efforts will be made to expand the numbering at each level of the hierarchical code.

Desideratum 5: Polyhierarchy

Hierarchical structures can have at least two forms. One such form is a polyhierarchical structure. A polyhierarchical structure refers to a tree structure in which a term has more than one parent or broader term (American Society for Indexing, n.d.). This means that some of the terms appear in more than one place in the hierarchy (Coletti & Bleich, 2001).

According to Cimino (Cimino, 1998), there seems to be universal consensus that medical informatics resources (such as vocabularies) should have a hierarchical structure. This facilitates searching and locating concepts either by traversing the tree-like hierarchy, or by grouping similar concepts together.

Such a hierarchical organization is also useful for disambiguation. For instance, if a concept named "cell" is located under "anatomic entity", then one can infer that this concept has a different intended meaning than if it appears under "power source" (Cimino, 1998). Here, the parent

concepts ("anatomical entity" and "power source") help in making the meaning of child concept ("cell") unambiguous (c.f.: desideratum 2). The majority of current standard vocabularies are strict hierarchies (Cimino, 1998). In this case, each child concept can have only one concept as its parent.

The main strength of strict hierarchies is that they are more amenable for computational purposes. A structure in which each child has a unique parent is far more efficient and easier to process than one where multiple parents are allowed. On the other hand, polyhierarchies might provide a more realistic and accurate conceptualization of a domain, as in the case of the concept "hepatorenal syndrome", which needs two parents, "liver diseases" and "renal disease".

Concerning the Q-codes, we adopted a strict hierarchy, i.e. a single parent per child concept. For example, the concept "QR31 qualitative study" has the concept "QR3 research method" as its only parent. We favored this arrangement over a polyhierarchy as this structure complements that of the International Classification of Primary Care (ICPC) (Jamoulle et al., 2018).

Desideratum 6: Formal Definitions

According to Cimino (Cimino, 1998), many researchers in medical informatics and knowledge representation have formulated the requirement that controlled vocabularies should include Formal Definitions. It should, however, be mentioned that a formal definition here does not mean that the concepts should necessarily be expressed in a particular formalism, such as First Order Logic or RDF'S/OWL. Instead, according to this desideratum, the Formal Definition of a concept is expressed as the different relationships in which the concept participates with other concepts. For example, the concept "hay fever" participates in hyponymy ("a type of") relationship with the concept "fever". This relationship is also referred to as "parent-child", "super-class/sub-class" or "generalization-specialization". The same concept "hay fever" participates with the concept "allergen" in a "caused by" relationship.

Concerning the Q-codes, we have in total 172 relationships, including eight top-level categories with a total of 44 first-level children. These 44 first-level terms have a total of 109 children. These 109 second-level terms have a total of 21 third-level children. For example, the concept "QR31 qualitative study" participates with the concept "QR3 research method" in the "is-a" relationship (e.g. "QR31 qualitative study" "is-a" "QR3 research method"). However, it is important for these relationships to be in a form which can be manipulated symbolically (i.e., with a computer), as opposed to narratives like those seen in a dictionary (Cimino, 1998). In the case of the Q-codes, these relationships are, indeed, represented symbolically, (i.e. in the form of "is-a" links, or "parent-child" relationships), thus, making it easier to manipulate them by a computer.

Desideratum 7: Reject "Not Elsewhere Classified"

This desideratum discourages the use of the category "Not Elsewhere Classified" (called "rag-bag" in this case) to represent terms that cannot be classified under any other category, i.e. these terms cannot be classified elsewhere. The main motivation for such a rag-bag category is that no vocabulary can guarantee domain completeness at any one time (Cimino, 1998). Thus, the rag-bag category facilitates the representation of terms, which cannot be classified elsewhere given the current state of the taxonomy. According to (Cimino, 1998), one critical issue with having a rag-bag category is that terms classified under this category lack a formal definition (c.f. desideratum 6). Here, the terms in this category can only be defined via exclusion, based on knowledge of the rest of the terms in the taxonomy. Furthermore, as the vocabulary evolves, the meaning of these "Not Elsewhere Classified" or rag-bag terms could change accordingly. This gives rise to the phenomenon of semantic drift, which hinders the analysis of historical data.

In the case of the Q-Codes, the rag-bag category serves one major purpose, sorting of conference abstracts for the discovery and classification of terms. To assist in this process, the rag-bag category is comprised of four subcategories: (i) unable to code, unclear; (ii) acronym; (iii) out of scope of Family Medicine; and (iv) consider new code. Abstracts in the "unable to code, unclear" subcategory, contain no discernible terms related to contextual concepts in GP/FM.

An abstract/title containing an abbreviation or acronym is placed in the "acronym" subcategory. However, these abbreviations/acronyms are often unclear and not well defined.

Some conference abstracts contain concepts unrelated to GP/FM, and thus, they are placed in the "out of scope of Family Medicine" subcategory. These abstracts usually discuss hospital-based studies.

Finally, the abstracts in the "consider possible new code" subcategory contain newly discovered terms or concepts that are currently not present in the taxonomy. These new terms or concepts are defined and examined for relationships to the terms already present in the Q-Codes taxonomy. Finally, these new terms are added to the taxonomy at the appropriate level, as defined by their relationships to other terms in the taxonomy.

Thus, this category and its subcategories have been useful in discovering terms during the creation of the Q-Codes taxonomy. Currently, however, the rag-bag category is useful for suggesting possible terms as additions to the future versions of the taxonomy. The rag-bag category does not hold any terms at each successive release of a version of the Q-Codes. Therefore, this category has no significance for the various end-users and applications of the Q-Codes taxonomy.

Desideratum 8: Multiple Granularities

When a vocabulary is being constructed for a particular application, there is implicitly a preconception about the level of granularity ("details") at which the concepts should be expressed. Granularity can be at a single level or at multiple levels. As the name suggests, with single granularity, all concepts are presented along a single level. Conversely, multiple granularity allows concepts to be represented with progressively finer-grained precision: "Diabetes Mellitus", "Type II Diabetes Mellitus", and "Insulin-Dependent Type II Diabetes Mellitus" (Cimino, 1998). This desideratum asserts that terminologies with multiple granularity should be preferred over those with single granularity. The main issue with vocabularies that attempt to operate at a single level of granularity is their inadequacy for applications requiring finer-grained information. In addition, they will also be considered too overwhelming and cumbersome in applications requiring coarser-grained information (Cimino, 1998).

The Q-codes are a multi-granular taxonomy, as they contain up to three possible levels for each top-level category. In turn, this multi-level granularity allows the Q-Codes taxonomy to be used for many purposes, including indexing the grey literature. Finally, this granularity provides the user with the ability to choose general or specific terms according to their needs.

Desideratum 9: Multiple Consistent Views

According to this desideratum, if a vocabulary is intended for use in multiple applications, then there is a need to provide multiple, consistent views of the vocabulary, as dictated by the various applications (Cimino, 1998; van Ginneken, van der Lei, & Moorman, 1992). For example, if a vocabulary with multiple granularity (c.f. desideratum 8) is to be used in an application that requires coarse-grained concepts, such as "Diabetes Mellitus", then finer-grained concepts, such as "Insulin-Dependent Type II Diabetes Mellitus", could be collapsed into the coarser-grained concept and marked as a synonym. An alternative approach to providing multiple consistent views is to enable users to show/hide specific levels depending on their needs. In a more extreme case, an application may restrict the user to only one level of the hierarchy, while hiding the remaining levels.

In the case of the Q-Codes, users are able to display the different levels of the taxonomy according to their needs (see www.hetop.eu/q). First, the user sees the eight top-level categories. After choosing one of these categories (e.g. "QC patient category"), the second level, with all of its terms, becomes visible, revealing in this case QC1, QC2, QC3 up to and including QC6. The user can continue to climb down to the next level. At any level, the user can choose a particular term to view its definition, links to literature and other information, and relationships to other terms in the taxonomy. Thus, the user can see any one level while hiding the other levels, and, for any one term, he/she can see the associated broader and narrower terms.

Desideratum 10: Representing Context

Traditionally, vocabularies have been created without consideration for the specific contexts in which they are intended to be used. While this strategy helps in reducing implicit assumptions about the vocabulary and enables it to "stand alone", it also leads to confusion when determining

whether the vocabulary concepts can be used in specific contexts (Cimino, 1998). Thus, this desideratum asserts that vocabularies should contain explicit information about the contextual usage of concepts, i.e. explaining how/when these concepts should or should not be used (Rector, Glowinski, Nowlan, & Rossi-Mori, 1995).

In the Q-codes, we are provided a glimpse of its purpose by the titles given to some of the terms, such as "QS4 primary care provider" and "QS41 family doctor". Thus, these titles indicate that the Q-Codes taxonomy is used for General Practice or Family Medicine. However, little is provided in the definitions of the terms as to how and when to use them.

The purpose and use of the Q-Codes taxonomy has been further documented in the literature (Jamoulle et al., 2018, 2017; Jamoulle & Resnick, 2016). As a "stand alone" taxonomy, the Q-Codes have been used for e-learning courses in GP/FM (Jamoulle, Grosjean, et al., 2017; Jamoulle & Resnick, 2016). In combination with other terminologies, such as ICPC-2, it has been proposed that the Q-Codes assist with indexing and retrieval of grey literature about GP/FM (Jamoulle et al., 2018, 2017; Jamoulle & Resnick, 2016). Beyond this, not much, if anything, has been provided in this literature about how and when individual terms should and should not be used. In the future, notes (annotations) will be added to the definitions, describing the proper use of the terms.

Desideratum 11: Graceful Evolution

All vocabularies are bound to evolve over time. However, experience shows that in most cases, changes are brought about for the convenience of the vocabulary's creators, and such changes tend to be problematic for the users (Cimino, 1996a). Thus, this desideratum states that those parties responsible for maintaining the vocabulary should ensure its graceful evolution. This can be achieved by assuring that all changes as well as the reason/request for these changes be properly documented and logged.

The Q-codes are currently in their infancy, being used for indexing grey literature only for one to two years. It is expected that they will continue to grow and evolve. We are currently exploring methods to document all changes and reasons/requests for changes, as they continue this growth and evolution.

Desideratum 12: Recognizing Redundancy

Redundancy is a phenomenon whereby the same information can be stated in multiple different ways (Cimino, 1998). For instance, consider a case of "pneumonia in the lower lobe of the left lung" in a patient. Now, this information is to be entered in an electronic patient record, which is based on medical vocabulary. However, if the vocabulary does not have a corresponding concept for "pneumonia in the lower lobe of the left lung", then the user may code it under the concept "Pneumonia" and include an additional label/modifier "left lower lobe". If at some later time, the concept "Left Lower Lobe Pneumonia" is indeed added to the vocabulary, then there will be two ways to code the same concept in the vocabulary: the old way, under "Pneumonia" and with another term indicating location; and the new way, directly under the concept "Left Lower Lobe Pneumonia". Such redundancies are to be avoided, but are inevitable when the vocabulary evolves. Thus, this desideratum asserts that a mechanism to recognize redundancies should be put in place. The two desiderata of Formal Definitions and Representing Context (c.f.: desiderata 6 and 10 respectively) can help in detecting redundancies.

At the present time, there are no redundant terms in the Q-codes. However, by satisfying desideratum 6 (Formal Definitions), one measure has been put in place to detect the presence of redundancies. As mentioned above, efforts will be made to institute desideratum 10 (Representing Context), further increasing the chances that redundancies are detected.

Conclusion

The analysis of the Q-Codes taxonomy demonstrates that it meets eleven of the twelve desiderata. For the remaining desideratum, Representing Context, notes or annotations will be added to the definitions of the terms, describing how and when they should be used. In addition, extensive records, noting any future changes, will be kept to guarantee that the Q-Codes continue to meet these desiderata, as they grow and evolve.

From these results, slight improvements can be made to the Q-Codes component of, and thus, the 3CGP terminology. This, in turn, improves the ability to index the grey literature with this terminology, providing greater access to and preventing loss of needed information.

For future work, the ICPC-2 component of 3CGP will be evaluated for the presence of the twelve desiderata. From this evaluation, any forthcoming recommendations will be provided for improvements to 3CGP, and thus, to indexing and retrieval of grey literature, leading to future research in GP/FM.

References

American Society for Indexing. (n.d.). About Taxonomies & Controlled Vocabularies [A Special Interest Group of the American Society for Indexing]. Retrieved November 14, 2018, from http://www.taxonomies-sig.org/about.htm

Bellefontaine, S. P., & Lee, C. M. (2014). Between Black and White: Examining Grey Literature in Meta-analyses of Psychological Research. *Journal of Child and Family Studies*, *23*(8), 1378–1388. https://doi.org/10.1007/s10826-013-9795-1

Cimino, J. J. (1996a). Formal descriptions and adaptive mechanisms for changes in controlled medical vocabularies. M*ethods of Information in Medicine*, *35*(3), 202–210.

Cimino, J. J. (1996b). Review paper: coding systems in health care. *Methods of Information in Medicine*, *35*(4–5), 273–284.

Cimino, J. J. (1998). Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of Information in Medicine*, *37*(4–5), 394–403.

Coletti, M. H., & Bleich, H. L. (2001). Medical subject headings used to search the biomedical literature. *Journal of the American Medical Informatics Association: JAMIA*, *8*(4), 317–323.

Côté, R. A., & Robboy, S. (1980). Progress in medical information management. Systematized nomenclature of medicine (SNOMED). *JAMA*, *243*(8), 756–762.

Denda, K. (2002). Fugitive Literature in the Cross Hairs. *Collection Management*, *27*(2), 75–86. https://doi.org/10.1300/J105v27n02_07

Evans, D. A., Cimino, J. J., Hersh, W. R., Huff, S. M., & Bell, D. S. (1994). Toward a medical-concept representation language. The Canon Group. *Journal of the American Medical Informatics Association: JAMIA*, *1*(3), 207–217.

Evans, D. A., Rothwell, D. J., Monarch, I. A., Lefferts, R. G., & Cote, R. A. (1991). Toward representations for medical concepts. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, *11*(4 Suppl), S102-108.

Evans, David A. (1988). Pragmatically-Structured, Lexical-Semantic Knowledge Bases for Unified Medical Language Systems. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 169–173.

Jamoulle, M., Cardillo, E., Ittoo, A., Vander Stichele, R., Resnick, M. P., Grosjean, J., … Vanmeerbeek, M. (2018). Indexing grey literature in General Practice: Family Medicine in the Era of Semantic Web. In D. Farace & J. Frantzen (Eds.), *GL19 Conference Proceedings* (Vol. 19). National Research Council of Italy Piazzale Aldo Moro 7, Rome: TextRelease,. Retrieved from http://www.textrelease.com/images/GL19_Jamoulle_et_al.pdf

Jamoulle, M., Grosjean, J., Resnick, M., Ittoo, A., Treuherz, A., Vander Stichele, R., … Vanmeerbeek, M. (2017). A Terminology in General Practice/Family Medicine to Represent Non-Clinical Aspects for Various Usages: The Q-Codes. *Studies in Health Technology and Informatics*, *235*, 471–475.

Jamoulle, M., & Resnick, M. P. (2016). *General Practice / Family Medicine multilingual terminology – English version*. Strépy-Bracquegnies, Belgium: Le livre en papier. Retrieved from http://www.publier-un-livre.com/fr/le-livre-en-papier/349-general-practice-family-medicine-multilingual-terminology-english-version

Lindberg, D. A., Humphreys, B. L., & McCray, A. T. (1993). The Unified Medical Language System. *Methods of Information in Medicine*, *32*(4), 281–291.

Mahood, Q., Van Eerd, D., & Irvin, E. (2014). Searching for grey literature for systematic reviews: challenges and benefits. *Research Synthesis Methods*, *5*(3), 221–234. https://doi.org/10.1002/jrsm.1106

Moorman, P. W., van Ginneken, A. M., van der Lei, J., & van Bemmel, J. H. (1994). A model for structured data entry based on explicit descriptional knowledge. *Methods of Information in Medicine*, *33*(5), 454–463.

New York Academy of Medicine. (2018, October). What is Grey Literature? | Grey Literature Database. Retrieved November 13, 2018, from http://www.greylit.org/about

Paez, A. (2017). Grey literature: An important resource in systematic reviews. *Journal of Evidence-Based Medicine*. https://doi.org/10.1111/jebm.12265

Pappas, C., & Williams, I. (2011). Grey Literature: Its Emerging Importance. *Journal of Hospital Librarianship*, *11*(3), 228–234. https://doi.org/10.1080/15323269.2011.587100

Rassinoux, A. M., Miller, R. A., Baud, R. H., & Scherrer, J. R. (1996). Modeling principles for QMR medical findings. *Proceedings: A Conference of the American Medical Informatics Association. AMIA Fall Symposium*, 264–268.

Rector, A. L., Glowinski, A. J., Nowlan, W. A., & Rossi-Mori, A. (1995). Medical-concept models and medical records: an approach based on GALEN and PEN&PAD. *Journal of the American Medical Informatics Association: JAMIA*, *2*(1), 19–35.

TextRelease. (n.d.-a). Grey Literature - GreySource, A Selection of Web-based Resources in Grey Literature [GreyNet International 2018]. Retrieved November 13, 2018, from http://www.greynet.org/greysourceindex/documenttypes.html

TextRelease. (n.d.-b). GreyNet International, Grey Literature Network Service [GreyNet International 2018]. Retrieved November 13, 2018, from http://www.greynet.org/home.html

Tillett, S., & Newbold, E. (2006). Grey literature at The British Library: revealing a hidden resource. *Interlending & Document Supply*, *34*(2), 70–73. https://doi.org/10.1108/02641610610669769

van Ginneken, A. M., van der Lei, J., & Moorman, P. W. (1992). Towards unambiguous representation of patient data. *Proceedings. Symposium on Computer Applications in Medical Care*, 69–73.

Volot, F., Zweigenbaum, P., Bachimont, B., Ben Said, M., Bouaud, J., Fieschi, M., & Boisvieux, J. F. (1993). Structuration and acquisition of medical knowledge. Using UMLS in the conceptual graph formalism. *Proceedings. Symposium on Computer Applications in Medical Care*, 710–714.

When the Virtual Becomes Reality: An Environmental Scan of the Presence of Virtual Reality and Artificial Intelligence in Health and Cancer Care Environments*

Marcus Vaska, Alberta Health Services, Canada

Abstract:

Grey literature has long been associated with technological enhancements, recognizing the power that informational communication, namely, social media, plays in generating interest in blogs, Twitter feeds, and other instantaneous knowledge exchange platforms. The ability of these programs to generate and identify specific data patterns[1] from a single posting has led to increasing interest in two aspects of machine learning in health care, namely Artificial Intelligence (AI) and Virtual Reality (VR).[1] AI "mimics elements of human cognition by computational means"[1], whereas VR enhances this cognition by allowing users to interact with a "three-dimensional, computer generated environment"[2], manipulating objects and scenarios in an artificial world[2]. Introduced as a form of grey literature via Second Life[3], a popular role-playing online world launched in 2003, VR and AI have had a visible presence in numerous sectors, including healthcare.[4]

In 2011, IBM created Watson, a supercomputer considered to be one of the most revolutionary breakthrough's in artificial intelligence[4]. To test this claim, Watson appeared on an episode of Jeopardy, one of the longest-running game shows in the United States, in a friendly competition match between two of the winningest contestants in the show's 50 year history[4]. Watson's emphatic victory over the human contestants drew increasing interest to other applications of artificial intelligence and virtual reality, specifically in the field of healthcare.

While the first use of AI and VR in medicine is believed to have occurred in the 1990s for interpreting electrocardiograms[4], the invention of cloud networking in 2006[4] is considered the first proven use of AI and VR in the modern era focusing on healthcare. Although the arguments for AI and VI in clinical settings are plentiful, ranging from enhancing imaging and increased processing speed in electronic medical record (EMR) applications[4], the scenario is less clear-cut within the environment of cancer care. At the 2016 International Symposium of Biomedical Imaging in Prague, Czech Republic, a joint team of scientists and engineers claimed that the use of artificial intelligence resulted in a "92% accuracy [rate of detection] in breast tissue cancer cells[5]." However, a column authored in 2017 disputed a claim by IBM that Watson was the new revolution to cancer care[6].

This paper will aim to shed light on how artificial intelligence and virtual reality is viewed in both health and cancer care fields via a two-fold environmental scan approach, namely an anonymous survey polling staff working at two cancer care facilities in Calgary, Alberta, Canada, asking respondents to comment on any papers they have ever encountered in their own practice/research discussing AI or VR. This practice will be supplemented with a comprehensive search through the academic literature to achieve a hoped-for grand total of fifty unique papers. Each of these papers will be analyzed via the use of Altmetrics, "a single research output [that] can be talked about across dozens of different platforms"[7], a methodology introduced by Schopfel and Prost at GL 18, to determine how these perceived core papers are being shared via the use of social media.

Artificial Intelligence (AI) in Healthcare: A Brief History

The use of machine learning in healthcare has a long and rich history, originating during World War II when a neurophysiologist and mathematician joined forces to create "a simple neural network made of electrical circuits", i.e. the brain[4]. Over the next fifty years, further enhancements and improvement in machine learning led to the formation of two entities, namely artificial intelligence, "mimic[ing] elements of human cognition by computational means"[1], and virtual reality, allowing users to interact with a "three-dimensional, computer generated environment"[2], manipulating objects and scenarios in an artificial world.[2] In today's technological society, AI can be grouped into one of the following six categories, all of which are applicable to healthcare: knowledge cataloguing, retrieval (i.e. search engines); game theory strategy; semantic analysis; natural language processing (i.e. text-speech translation); task planning (i.e. GPS navigation); systems.[1]

* First published in the GL20 Conference Proceedings, February 2019.

Since the dawn of the new millennium, health services have relied extensively on AI, including enhanced image analysis, bioinformatics, triage, diagnostic testing, and electronic health records.[1]

While some traditionalists claim that the role of AI in medicine is uncertain, evidence regarding the successful use of AI in fields such as radiology, pathology, and dermatology, particularly with regards to the speed at which images are processed, cannot be denied.[4] Miller and Brown argue that no physician can ever be 100% confident in his/her diagnosis for each individual patient, however "combining machines plus physicians reliably enhances system performance."[4] With enhancements in the electronic patient medical record, including the establishment of health data cooperatives, where patients attain greater control of their own health information, AI is seen as reducing medical errors, and enhancing the care management for patients with chronic diseases.[4] These chronic diseases include not only physical ailments, such as cancer, diabetes, and congestive heart failure, but also the mental stigma associated with these conditions which can often lead to depression.[4]

Despite the launch of artificial neural networks in 1943, the application of this artificial technology in the medical field did not occur until 50 years later, namely in electrocardiograms, as a means to diagnose heart attacks, and predicting "intensive care unit length of stay following cardiac surgery."[4] With the initiation of Health Data Cooperatives (HDCs) in recent years, aimed at providing patients with more invested interest and control over their own treatment journey, AI holds promise in electronic medical records, purporting the ability to accurately predict one's diagnosis or further course of treatment required before it actually occurs.[4] Nevertheless, despite this seemingly vast intellectual superiority of machines over the human brain, potentially replacing a seasoned health professional with years of medical experience, concerns remain about the regular use of AI in healthcare. Miller and Brown provide two sound reasons why a traditional medical education still holds value today: there is no replacement for physically holding a medical device, and while artificial machines may be able to think logically, they are certainly not able to function empathetically, when faced with difficult, potentially life-threatening medical decisions.

Artificial Intelligence (AI) & Virtual Reality (VR) in Cancer Care: A Retrospective

Logically defined, virtual reality is a form of near-reality, which in technical terms refers to "presenting our senses with a computer-generated virtual environment that we can explore in some fashion."[2] The person that engages in a virtual world is completely emerged in the here and now, and can move objects and/or perform specified actions at will.[2] VR's foray into the medical world is thought to have been for practical reason, as, in the case of surgery, it allows the surgeon "to take virtual risks in order to gain real world experience."[2] While artificial intelligence can trace its roots back over more than half a century, virtual reality (VR) did not become recognized as a legitimate education format until after the new millennium. Seen primarily as "a technology which allows a user to interact with a computer-simulated environment, "[3] whether real or imagined, it is often associated with forms of enhancing the user experience in computer games. Attributed to Howard Rheingold, whose two works pioneered the concept to lay audience in the early 1990s, VR's integration in the public health domain, more specifically the cancer care environment, has introduced new treatment paradigms that may previously have been viewed as unrealistic. In 2007, the American Cancer Society became actively involved with Second Life, a popular virtual world platform, raising $75,000 for cancer research in its inaugural year of partnership.[3] Seven years later, IBM instructed *Watson*, the supercomputer that prevailed in a friendly Jeopardy competition in 2011, to offer its foreshadowing capabilities to suggest appropriate and best treatment options for cancer patients.[6] In their article detailing the launch of Watson into the cancer care environment, Ross and Swelitz gathered interviews from doctors, IBM executives, and AI experts around the world, who, despite being widely dispersed geographically, came to several dismaying conclusions. While *Watson* was viewed by IBM as the answer for cancer care, Ross and Swelitz report that no published papers were published prior to *Watson* embarking into healthcare on the effects of technology on both physicians and patients. With criticism ranging from *Watson* leveling bias on patients in foreign hospitals to apparent lack of concern regarding patients who do not fit standard treatment regimens, a cancer specialist echoes the feelings of many when stating "*Watson* for oncology is in [the] toddler stage, and we have to wait and actively engage…"[6]

While case studies of the use of AI and VR in cancer care have appeared regularly over the past few years, many in the medical field credit a radiologist, Dr. Judy Yee from the University of California San Francisco (UCSF) with pioneering the technology to make a vision become reality.[12] In 1997, Yee first learned of computed tomography colonoscopy (CTC) while attending the annual meeting of the Society of Gastrointestinal Radiology. With a background in science and mathematics, Yee believed that by making the colonoscopy easier on the patient, more patients would undergo recommended screening, thereby preventing a potential colorectal cancer diagnosis.[12] With the launch of the virtual colonoscopy in 2016, enhanced 3-D images of the "polyps, lesions, and other precancerous anomalies…[are] far less invasive and easier to interpret"[12] compared to traditional scans. Since news of the virtual colonoscopy first broke in the UCSF Summer 2016 newsletter, Yee has embarked on a new project which she refers to as virtual holography CTC. Via the combination of a laser stylus and stereoscopic optical technology, Yee is able to "grab the portion of the scan she wants to examine in more detail and interact with it in three-dimensional space."[12] Over a career that has spanned 20 years, Yee has written more than 100 articles and 22 book chapters, research that has led to continuous improvements in making a colonoscopy a less intimating experience for patients. These include a lower radiation dose during scans, higher quality images detecting polyps and cancers, and reducing the amount of laxatives patients must take before a procedure. However, as Yee explains, when interviewed for the UCSF Summer 2016 newsletter, one of Yee's primary reasons for embarking on the 3-D virtual colonoscopy journey is to create a less invasive experience for the patient and reduce the chance of procedural complications. Currently, a standard colonoscopy lasts approximately 30 minutes, requires sedation, and carries considerable risk of perforation, bleeding, and infection. Yee's methodology challenges and refines the notion of a traditional colonoscopy, as her method "doesn't require sedation, and the low-radiation dose CT scan takes just 20 seconds."[12]

In 2016, Cancer Research UK launched a £100 research challenge to solicit ideas from the global research community on how to address fundamental concerns in cancer prevention, diagnosis, and treatment.[8] Due to the substantial monetary incentive as well as the prestige of gaining recognition from one of the largest cancer research organizations in the world, fifty-seven teams of scientists from twenty-five different countries submitted an application. After careful scrutiny and deliberation, nine teams were selected as finalists. Three of the shortlisted teams were posited with "developing a 'Google Street View' for cancer."[8] Each team approached this challenge in unique ways, while showcasing the power of combining the latest enhancement in technology with medical knowledge expertise. The first team, captained by Greg Hannon, a professor at Cambridge, built a prototype with the ability to "walk around inside 3D virtual tumours."[8] Team number two, led by Ehud Shapiro, professor of computer science at the Weizmann Institute of Science in Israel, proposed a form of VR that would track a patient throughout the cancer journey, from diagnosis to end of treatment and/or death. Despite the tremendous data computing power that would be required to undertake such an initiative, Shapiro and his team believed that they have the ability to "produce the most comprehensive view of a tumor, stretching from its earliest origins to when it begins to spread and beyond."[8] Last but not least, team three, navigated by Dr. Josephine Bunch, professor of physics at the National Physics Laboratory in London, suggested the use of mass spectrometry, combined with next-generation genetic analysis "to generate molecular maps of tumors" thus offering hope for more accurate diagnosis and monitoring.

At the 2016 International Symposium of Biomedical Imaging, held in Prague, Czech Republic, scientists and engineers from Harvard joined forces to successfully utilize AI to distinguish healthy and diseased breast tissue cells. Whereas the diagnosis of healthy or cancerous tissue was traditionally conducted by pathologists reviewing biopsy samples under a microscope, a tedious and tiring task subject to human error, the Harvard conglomerate proved that with the help of AI, an accurate diagnosis rose as high as 92%.[9] This statistic, emphasizing comparable performance by machines vs. humans echoed Jeroen van der Laak, chair of the Prague Symposium's words, "it is a clear indication that artificial intelligence is going to shape the way we deal with histopathological images in years to come."[9] Led by Dr. Andrew Beck, the Harvard team was very particular in choosing metastatic breast cancer cells in lymph node biopsies with which to demonstrate AI's power and accuracy. According to the Centers for Disease Control and Prevention, breast cancer is the second deadliest type of cancer for women, which can often metastasize and spread to other parts of the body.[9] The methodology utilized by Beck's team involved a 'deep learning' approach to

instruct a computer to recognize breast cancer cells. Beck's team loaded thousands of images into the computer, focusing on any images where "the computer was prone to make a mistake in cancer identification, [retraining] the computer using greater numbers of more difficult examples."[9] Nevertheless, despite this apparent diagnostic breakthrough, Beck cautions both optimists and critics that the findings produced by his team was only one scenario, and AI has not yet evolved to the stage where it is capable of diagnosing and identifying rare forms of cancer. As Beck states, AI machines are mere robots that can be "routinely thrown off by an artifact in the biopsy image…humans will be needed to continuously teach the robots."[9] Following his team's success at the Prague Symposium, Beck, together with Aditya Khosla from the MIT Computer Science and Artificial Intelligence Laboratory, launched PathAI, a company focused on creating tools to more closely integrate AI in diagnosing cancer and other diseases "in order to provide faster, more accurate and reproducible results."[5]

While the above case studies have centered on the use of virtual reality among adult cancer patients, Dr. James Hu, assistant professor of medicine at the University of Southern California, along with serving as co-director of the Adolescent and Young Adult (AYA) cancer program believes that a younger demographic may be more in-tune and open to exploring virtual reality possibilities when it comes to cancer care. In an interview conducted in January 2017 with HemOnc Today, Hu explains that the virtual reality content provided to patients between the ages of 15-39 at the AYA is meant to offer a calming respite from reflecting on one's cancer diagnosis. As such, painting, entertainment, and educational programs, involve "scuba diving, fling, thrill rides and high impact adventures, [such as] being chased by dinosaurs."[10] While Hu admits that the benefits of virtual reality with the AYA program have not yet fully been explored, he is adamant that the methods utilized by the AYA are intended to combat fatigue, along with increasing stamina and endurance.[10] In addition, education and awareness are additional key aspects, centered on "side effects of treatment, financial resources, patient experiences, and orientation to medical services."[10]

One of the earliest documented instances of the use of virtual reality in cancer clinical trials occurred in Texas at the M.D. Anderson Cancer Center from February 2011-May2016.[11] According to principal investigator Dr. Susan Peterson, the goal and intended outcome of the trial was to "evaluate a virtual reality-based intervention for training health care providers who are not genetics specialists to effectively communicate with and counsel patients regarding cancer genetics."[11] Restricted to a study population of either genetic counselors or genetic counseling students, participants were required to complete pre and post questionnaires prior to and upon the conclusion of viewing a series of virtual reality scenarios. In addition, following an initial physiological measurement, whereby heart rate and level of perspiration during the virtual reality scenario are assessed, the genetic health care participants will subsequently undergo a genetic counseling session with a virtual patient.[11] Due to the very specific study population chosen, the trial lasted for more than 5 years, during which the proof of concept for the use of VR as a training mechanism for hereditary cancer risk and genetic testing was continuously being assessed.[11] While no results from the study have been posted to date, participant recruitment is no longer taking place, thus lending belief that this unusual prototype virtual reality application will gain greater interest in the scientific and medical fields over the coming years.

AI & VR Connections with Grey Literature

In his seminal paper discussing the impact of emerging information technologies on grey literature, Dobrica Savić posited that disruptive technology may in fact be a leading factor resulting in a less than ideal uptake and association regarding the influence of artificial intelligence and virtual reality in grey literature material. Officially defined by Clayton M. Christen in 1997, disruptive technology "often has performance problems, because it is new, appeals to a limited audience and may not yet have a proven practical application."[17] (p. 77) At the GL 19 Conference in Rome (October 2017), Savić introduced four technologies which are believed to have caused substantial disruption among grey literature information management: artificial intelligence and machine learning, virtual and augmented reality, internet of things, and big data.[17] It is elements of the first two items on this list, namely artificial intelligence and virtual reality, that will be discussed in this section.

While artificial intelligence and virtual reality has not yet officially been recognized as grey literature document types in the Grey Literature Network Service, several challenges existed by forms of grey literature, including being invisible or difficult to identify, access, and acquire content is similar to what is occurring in the virtual world.[3] In fact, VR is seen as sharing four key traits common across grey literature: a specified community where the VR has been implemented, a prolonged user experience reliant on referrals, a set time period during which a VR pilot is conducted, and a "presence that has both longevity and persistence."[3] While grey literature has arguably a much richer epidemiology compared to AI & VR, early instances of the application of this new technology in healthcare were noted almost entirely in scholarly publications, avoiding grey literature documents types. One of the fundamental challenges that has been believed to have caused disconnect between recognition of AI and VR as a form of grey literature and thus encourage wider adoption in healthcare is that "the human mind is not able to tell the difference between computer-generated images and the real world."[17] (p. 79) Further, any instances regarding the measurement of the impact of such research on healthcare usually involved journal impact factors or other sophisticated citation indexes, which could only be generated via the use of bibliographic databases and citation managers.[7] Despite the introduction of web-based bibliography citation managers, in particular Mendeley, which troll the web, looking for any relevant papers pertaining to the subject matter at hand, they do not regularly capture elements of how this information is disseminated via social media to the greater community (both academic and the general public). This dilemma has been mitigated with the rise in 2010 of a new form of research, referred to as "scholarly metrics or social media metrics, and most often defined as altmetrics."[7] (p. 5) This paper will thus focus on representative samples of core published papers depicting artificial intelligence and virtual reality in both cancer care and public health which are disseminated to a wider audience with the aid of social media, focusing on 4 grey literature document types: blogs, Facebook pages, news outlets, and tweets.

As discussed in their GL 18 paper, Schopfel and Prost remind readers that the altmetric concept has existed for less than a decade, being first introduced in September 2010, via a tweet from Jason Priem, a librarian at the University of North Carolina, Chapel Hill. In this tweet, which has been archived for posterity and remains accessible to this day, Priem states: "I like the term #articlelevelmatrics, but it fails to imply *diversity of measures*. Lately, I'm liking #altmetrics."[13] Less than one month after this inaugural introduction of this term, Priem and his colleagues launched a manifesto devoted to explaining the altmetric concept.[14] While this document delves into considerable detail about how the altmetric score of a particular publication is calculated, which is beyond the scope of this paper, is does present a unique perspective of research analysis by traditional impact features of measuring an author's contribution and disseminating this much quicker than previously possible due to tapping into several social media subsets, namely twitter feeds, blog posts, Facebook pages, and various online news outlets. To date, the Altmetric Manifesto has received twenty-two comments and 489 drawbacks.[14] Despite this low number, Priem cautions that the potential power of Altmetrics is still in its infancy, and yet one of the core traits that this new tracking mechanisms shares with grey literature is speed of output, speed of analysis: "altmetrics are fast, using public APIs to gather data in days or weeks."[14]

The commonly accepted definition of altmetrics described above may certainly be relatable to the scientific community, however, it may not be as applicable to researchers and librarians. Schopfel and Prost thus argue for a modified definition attributed to librarians, to showcase how the use of altmetrics can help inform collection development decisions: "altmetrics are attention data from the social web that can help librarians understand which articles, journals, books, datasets, or other scholarly outputs are being discussed, shared, recommended, saved, or otherwise used online."[15] Today, Altmetric LLP, headquartered in London, England has created a multitude of tools and data for users falling into one of five broad-level categories: publishers, institutions, researchers, funders, and research and development.[16] This all-encompassing view allowed the impact of a particular research endeavour to reach not only scientists and physicians, but also the general public, since the Altmetric bookmarklet API is an open-source freely available tool, with an ever-expanding database of more than 9 000 000 research outputs and counting.[16] With each record for which an Altmetric score exists, perusers are able to see the Altmetric Attention Score, which is intended to serve as a representative sample of the amount of attention a particular paper has received on social media. In addition, via the use of tabs, it is possible to easily ascertain the

impact of the article via news outlets, blog posts, twitter feeds, Facebook Pages, and the number of readers that currently have the said paper saved in their online Mendeley accounts. Further, twitter demographics are represented both geographically as well as via population, enabling one to see and compare level of interest in a particular publication between members of the public, practitioners, scientists, and science communicators. Further, since each paper that has been Altmetricized, so to speak, contains a unique URL, it is easy to save the attention score for subsequent referral, in addition to setting up an e-mail alert that will notify the requestor of any increases in the attention score. This ability allows for opportunities to discuss high-trending "flavour of the month" articles, potentially leading to further discussions." [See Figure 1 for a sample of one of the records that was used in this paper for further analysis].

Figure 1: Sample Altmetric Record – use of AI or VR in cancer care

The First Environmental Scan: Polling Cancer Care Clinicians

The author's interest in the subject matter of this paper was peaked during a Grand Rounds presentation at the Tom Baker Cancer Centre in September 2017, lead by a junior oncologist practicing at this facility. Despite the seemingly endless opportunities artificial intelligence and virtual reality bring to the cancer care environment, namely in providing assistance during surgery as well as the ability to enhance patient treatment regimens, the Grand Rounds presenter cautioned that considerable debate exists over whether or not an artificial entity can replace or match human cognition, which has resulted in smaller uptake and awareness of this new technology than was hoped for. Further discussions on artificial intelligence and virtual reality with a colleague from the grey literature community resolved the author's decision to conduct an environmental scan, not only of the existing literature on AI and VR in a cancer care and public health domain, but more importantly seek to understand awareness of these technological marvels among oncologists working at the Tom Baker Cancer Centre in Calgary, Alberta, Canada.

Following approval by the executive director at the Tom Baker Cancer Centre, as well as follow-up through the ARECCI Ethics Guideline and Screening Tool, which determined lowest possible risk to participants as a result of participating in this venture, a brief opened-ended one-question survey was sent to all oncologists (~110) working at or affiliated with this cancer care facility on February 21, 2018. The question posed was as follows:

> *Please provide the citations of any papers that you have ever encountered in your practice/research which discuss the use of virtual reality or artificial intelligence (either specifically in cancer or more generally in public health)*

The survey remained open until April 30, 2018. Despite the extraordinarily lengthy working hours demanded of oncologists working at the Tom Baker, along with several other competing factors, 12 responses were received. While some respondents openly admitted that they were not particularly familiar with the subject matter at hand, others willingly shared either complete citations to papers and/or grey literature information (conference, webinars, proceedings, etc.), where additional papers were referenced. All told, responses from this survey culminated in 13 unique papers being identified that foretold of the influence of AI and VR in not only the cancer field, but healthcare in general.

The Second Environmental Scan: Literature Search

During the period that the survey remained open, a second environmental scan, via a comprehensive literature search, was undertaken in the medical databases, applying a combination of search terms pertaining to artificial intelligence, virtual reality, cancer, and public health, published between 2007-2018. This search identified an additional thirty-nine potential papers to consider for analysis (see Figure 2, Search Planning Document, for a breakdown of the terms used and resources consulted)

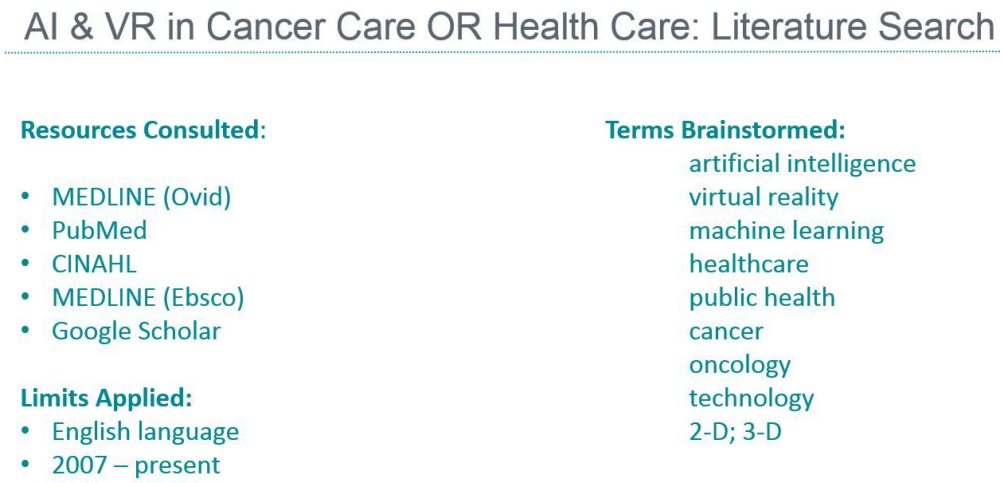


Figure 2: Search Planning Document

Results, Analysis, & Discussion

As discussed by Schöpfel and Prost, initial metric analysis focused almost entirely on publications in mainstream academic journals, where documenting an author's publication impact, documented in impact factors and h-indexes was the norm[7]. Due to increasing use of social media to relay information, scholarly output expanded from a focus on academia represented by journal articles, conference proceedings, and theses to "oral presentations, performances, artifacts, exhibitions, online events, multimedia…and other forms of intellectual property"[7] (pg. 5)

Following the input received from the survey respondents, together with results gleamed from the literature search, fifty-two papers were selected for an Altmetric analysis. Altmetric attention scores for each paper were obtained via the Altmetric it! bookmarklet, available from www.altmetric.com. These papers were colour-coded as either focusing on cancer care (yellow), or a more general discussion of public health (green). Since altmetric scores can fluctuate unexpectedly according to how a particular topic is portrayed in the news, each of the 52 citations were run through Altmetric it! at 3 separate intervals (from April 13, 2018 – June 22, 2018, with at least seven days separating each run). Using this methodology, ten cancer care papers and forty-two public health papers were identified. This ratio further supported findings from the literature indicating that inclusion of AI and VR in the public health sphere is far more pertinent than in a cancer care environment. Thus, to make the analysis more equitable, five papers from the healthcare sector and five papers from the cancer field were chosen, according to the following Altmetric attention score criteria: an altmetric attention score of at least 3, a dedicated Twitter presence (feeds and followers), and an optional additional social media element (either a news outlet, blog, or Facebook page).

AI in Healthcare: Five Selected Papers for Analysis

The five papers selected for analysis (see Figure 3 below) portraying the use of AI in healthcare comprised a publication date range of 2014-2017, and focused on the following topics: acute pain management, chronic pain, weight-related disorders, and a general overview regarding the presence of virtual reality in medical practice. The majority of papers chosen were published between 2016-2017, a result of proceedings from a 2016 conference on Machine Learning in Healthcare, which was specifically mentioned by one of the survey respondents.

- Garret, B., Taverner, T., Masinde, W., Gromala, D., Shaw, C., & Negraeff, M. (2014). A rapid evidence assessment of immersive virtual reality as an adjunct therapy in acute pain management in clinical practice. *Clinical Journal of Pain, 30*(12): 1089-1098.
- Keller, M., Park, H., Cunningham, M., Fouladian, J., Chen, M., & Spiegel, B. (2017). Public perceptions regarding use of virtual reality in health care: a social media content analysis using Facebook. *Journal of Medical Internet Research, 19*(12): e419.
- Miller, D., & Brown, E. (2017). Artificial intelligence in medical practice: the question to the Answer? *American Journal of Medicine, 131*(2): 129-133.
- Wiederhold, B., Gao, K., Sulea, C., & Wiederhold, M. (2014). Virtual reality as a distraction technique in chronic pain patients. *CyberPsychology, Behavior, & Social Networking, 17*(6): 346-352.
- Wiederhold, B., Riva, G., & Gutierrez-Maldonado, J. (2016). Virtual reality in the assessment and treatment of weight-related disorders. *CyberPsychology, Behavior, & Social Networking, 19*(2): 67-73.

Figure 3: AI in Healthcare: Five Selected Papers for Analysis

Following a cumulative approach, where social media elements (twitter feeds, blog posts, news outlets and Facebook pages) were gleamed from all five papers, geographic, demographic, and social media analyses were conducted, with the geographic and demographic totals identified from the twitter feeds.

Geography

The five selected papers (as per figure 3 above) were, as of June 22, 2018 followed by individuals from 14 different countries, including Canada, the United States, Brazil, Venezuela, United Kingdom, Spain, Italy, Turkey, the Netherlands, Burma, India, the Philippines, Australia, and New Zealand.

Demographics

Altmetrics assigns followers of its attention score ratings to one of four categories: members of the public, practitioners (comprising physicians and other healthcare professionals), scientists, and science communicators (including journalists, bloggers, and editors). The five selected AI in healthcare papers were thus followed by 131 members of the public, twenty-three practitioners, twelve scientists, and five science communicators.

Grey Literature Document Types

90% of the entire social media output produced by the five AI in healthcare papers derived from tweets, of which there were a total of 244. This was followed by news outlets (7%; 20); Facebook pages (2%; 5), and blog posts (1%; 2). Please see Figure 4 below for a pictorial representation of this data.

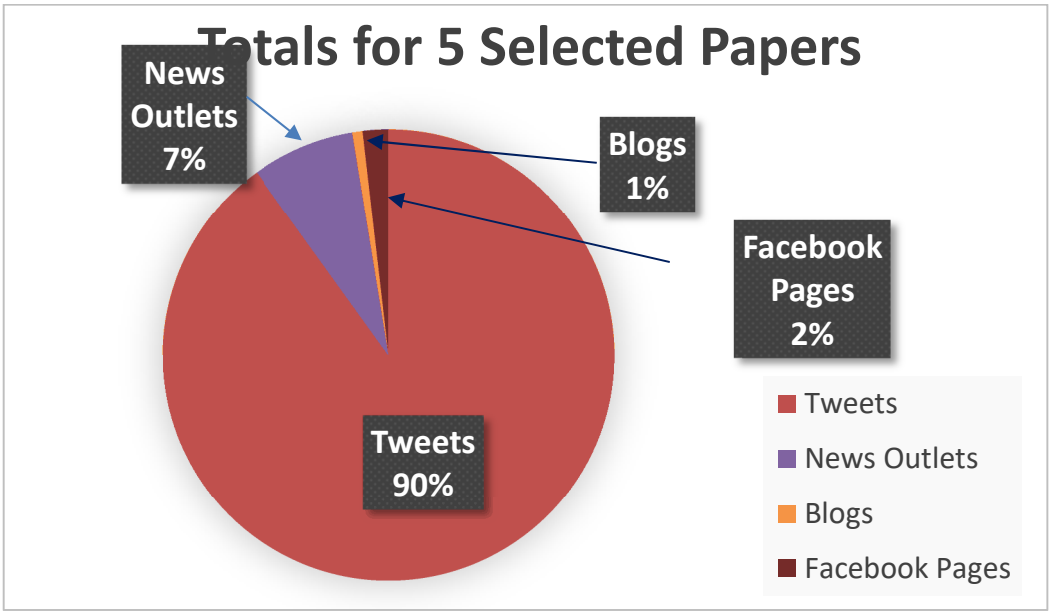


Figure 4: AI in Healthcare: Grey Literature Document Types

VR in Cancer Care: Five Selected Papers for Analysis

- Chirico. A., Lucidi, F., De Laurentis, M., Milanese, C., Napoli, A., & Giordano, A. (2015).Virtual reality in health system: beyond entertainment. A mini-review on the efficacy of VR during cancer treatment. *Journal of Cellular Physiology, 231*(2): 275-287.
- Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L., & Aerts, H. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer* [Epub ahead of print]
- Li, W., Chung, J., & Ho, E. (2011). The effectiveness of therapeutic play, using virtual reality computer games, in promoting the psychological well-being of children hospitalized with cancer. *Journal of Clinical Nursing, 20*(15-16): 2135-2143.
- Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D., Barnholtz-Sloan, J., Velazquez, J., Brat, D., & Cooper, L. (2018). Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences of the United States of America, 115*(13): E2970-E2979.
- Schneider, S., & Hood, L. (2007). Virtual reality: a distraction intervention for chemotherapy. *Oncology Nursing Forum, 34*(1): 39-46.

Figure 5: VR in Cancer Care: Five Selected Papers for Analysis

The five papers selected for analysis (see Figure 5 above) portraying the use of VR in cancer care comprised a publication date range of 2007-2018, and focused on the following topics: chemotherapy, therapeutic play for pediatric cancer patients, a general overview of VR during cancer treatment, radiology, and histology genomics. Two of the papers chosen are recent publications stemming from Proceedings of the National Academy of Sciences of the United States of America, which one of the oncology survey respondents recommended, along with an Epub ahead of print publication (artificial intelligence in radiology)

Following a cumulative approach, where social media elements (twitter feeds, blog posts, news outlets and Facebook pages) were gleamed from all five papers, geographic, demographic, and social media analyses were conducted, with the geographic and demographic totals identified from the twitter feeds.

Geography

The five selected papers (as per figure 5 above) were, as of June 22, 2018 followed by individuals from fifteen different countries, including the United States, Brazil, Venezuela, United Kingdom, Ireland, Spain, France, the Netherlands, Germany, Austria, Czech Republic, Bulgaria, Turkey, China, and Australia.

Demographics

Altmetrics assigns followers of its attention score ratings to one of four categories: members of the public, practitioners (comprising physicians and other healthcare professionals), scientists, and science communicators (including journalists, bloggers, and editors). The five selected VR in cancer care papers were thus followed by 128 members of the public, eighteen practitioners, eighty-five scientists, and seven science communicators.

Grey Literature Document Types

95% of the entire social media output produced by the five VR in cancer care papers derived from tweets, of which there were a total of 248. This was followed by news outlets (3%; 7); Facebook pages (1%; 4), and blog posts (1%; 2). Please see Figure 6 below for a pictorial representation of this data.

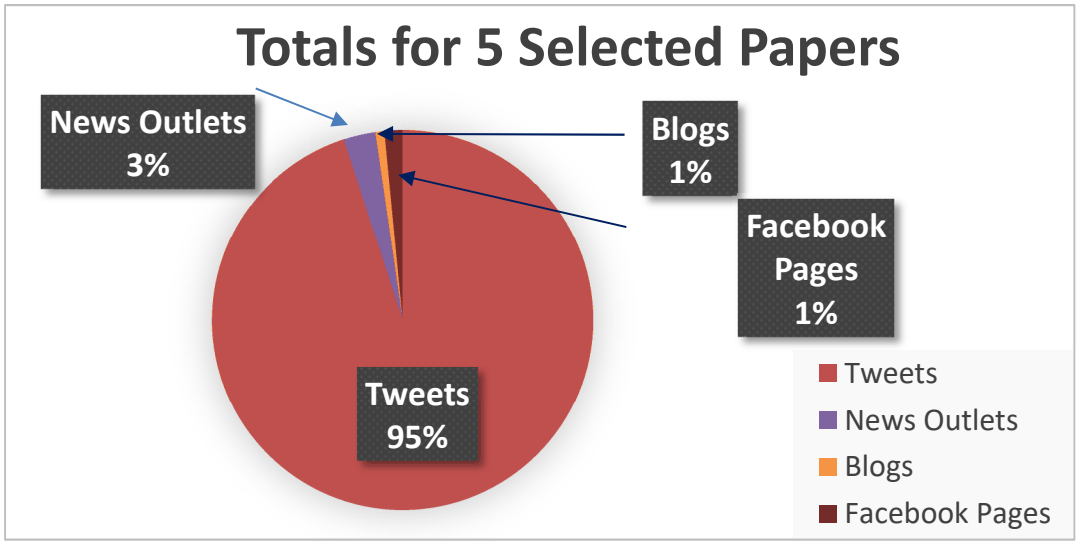


Figure 6: VR in Cancer Care: Grey Literature Document Types

Conclusion

As this paper has hopefully shown, technological marvels, represented in the form of artificial intelligence and virtual reality have had a substantial impact on provision of healthcare, with more changes yet to come as these *disruptive entities* continue to evolve. While some critics argue that AI and VR will wreck havoc in terms of grey literature management, bypassing a set standard of previously identified document types, any means of disseminating information, and using social media to transmit key data in mere seconds cannot be overlooked. While AI and VR are slowly gaining ground as a supplement, so to speak, to a healthcare practitioners own eyes and years of experience, at this point, only the tip of the iceberg has been reached regarding the potential of this technology.

References

1. Rubak, J. (2018). *Introduction to machine learning.* Presented March 1, 2018 at the Tom Baker Cancer Centre [medical physicists session]
2. Virtual Reality Society. (2017). *What is virtual reality?* Retrieved March 3, 2018 from https://www.vrs.org.uk/virtual-reality/what-is-virtual-reality.html
3. Ferry, K., Gelfand, J., Peterman, D., & Tomren, H. (2008). Virtual reality and establishing a presence in Second Life: new forms of grey literature? *The Grey Journal, 4*(3): 159-168.
4. Miller, D., & Brown, E. (2018). Artificial intelligence in medical practice: the question to the answer? *The American Journal of Medicine, 131*: 129-133.
5. Moore, C. (2016). *Artificial intelligence gets an A+ for accuracy diagnosing breast cancer*. Retrieved March 23, 2018 from https://breastcancer-news.com/2016/06/29/artificial-intelligence-gets-accuracy-diagnosing-breast-cancer/
6. Ross, C., & Swetlitz, I. (2017). *IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close.* Retrieved March 23, 2018 from https://www.statnews.com/2017/09/05/watson-ibm-cancer/
7. Schopfel, J., & Prost, H. (2016). Altmetrics and grey literature: perspectives and challenges. *The Grey Journal, 13*(1): 5-22.
8. Peel, N. (2016). *Virtual reality and precision diagnosis: Cancer Research UK's grand challenge shortlist*. Retrieved July 27, 2018 from http://blogs.biomedcentral.com/on-medicine/tag/cruk-grand-challenge/
9. Wanjek, C. (2016). *AI boosts cancer screens to nearly 100 percent accuracy*. Retrieved July 27, 2018 from https://www.livescience.com/55145-ai-boosts-cancer-screen-accuracy.html
10. Hu, J. (2017). *Virtual reality initiative benefits adolescents, young adults with cancer.* Retrieved July 27, 2018 from https://www.healio.com/hematology-oncology/pediatric-oncology/news/online/%7B7e40063f-cc32-4829-98a7-5d2db2ec55de%7D/virtual-reality-initiative-benefits-adolescents-young-adults-with-cancer
11. Clinicaltrials.gov (2016). *Virtual reality intervention in cancer genetics.* Retrieved April 16, 2018 from https://clinicaltrials.gov/ct2/show/NCT01310829
12. Wells, J. (2016). *3-D virtual reality colonoscopy: pursuing a better path to colorectal cancer prevention*. Retrieved April 16, 2018 from https://www.ucsf.edu/news/2016/07/403406/3-d-virtual-reality-colonoscopy-pursuing-better-path-colorectal-cancer
13. Priem, J. (2010). *I like the term #artciclelevelmetrics, but it fails to imply *diversity* of measures. Lately, I'm liking #altmetrics.* [Twitter post]. Retrieved August 9, 2018 from https://twitter.com/jasonpriem/status/25844968813
14. Priem, J., Taraborelli, P., Groth, P., & Neylon, C. (2010). *Altmetrics: a manifesto.* Retrieved August 9, 2018 from http://altmetrics.org/manifesto
15. Konkiel, S. (2016). *Hoe to make better collection management decisions by combining traditional metrics and altmetrics.* Retrieved August 9, 2018 from https://www.altmetric.com/blog/altmetrics-collection-development
16. Altmetric.com (2018). *Who is Altmetric for? Find out how our tools and data can help you.* Retrieved August 9, 2018 from https://www.altmetric.com/audience/
17. Savić, D. (2018). Impact of disruptive technologies on grey literature management. *The Grey Journal, 14*(2)*:* 77-80.

Data Archiving and Networked Services

DANS

Your 7 steps to sustainable data

1. **Prepare your data**
   Select the relevant data files. Check them for privacy aspects and file format against the guidelines issued by DANS.

2. **Go to EASY**
   Log in at https://easy.dans.knaw.nl. If you are new to EASY, you will have to register for an account first.

3. **Start the deposit procedure**
   Go to 'deposit your data', select your discipline and click 'start deposit'.

4. **Documentation and access level**
   Describe the dataset and indicate whether it is open access or whether access restrictions apply.

5. **Upload your data files**
   Select your data files and click 'upload dataset'.

6. **Submit your data files**
   Accept the licence agreement and send your dataset to DANS by clicking the 'submit' button.

7. **Publication by DANS**
   DANS will verify the dataset and publish the description you made. Your data have now been sustainably archived and will be accessible to others on a permanent basis under the conditions you specified.


Newsletter:
DataLink


Twitter:
@DANSKNAW

You Tube:
DANSDataArchiving


E-mail:
info@dans.knaw.nl


Website:
dans.knaw.nl

"The GreyForum is a series of onsite and online courses, seminars, and workshops where grey literature provides common ground for information professionals in the process of knowledge transfer"

Grey Literature and the Circular Economy

GreyForum Series 9.1

Seminar

Grey Literature and the Circular Economy

Date

Thursday, 5 September 2019

Time

13:00 – 17:00 hours

Location

Amsterdam Public Library
OBA Prinsenzaal 6.6
Oosterdokskade 143
1011 DL Amsterdam

Participant Quota

20 Participants

Registration

www.greynet.org/greyforumseries/circulareconomy.html

Contact Information

GreyNet International
Javastraat 194-HS
1095 CP Amsterdam
Netherlands

info@greynet.org
www.greynet.org
Phone +31 (0) 20 331 2420

This GreyForum seminar is organized by GreyNet International together with the Amsterdam based platform gocircularnow.com

Grey Literature and the Circular Economy

GreyNet

Program Outline, Instructors, and Prospective Participants

- **Welcome, Introduction**
- **Why a circular economy? A consumer-focused perspective on achieving a zero-waste economy**
  Lucas Borja Peinado and Michaela Larosse, gocircularnow.com
- **Grey Literature, Vehicle and Driver in a Circular Economy**
  Dominic Farace, GreyNet International
- **Participant's Endgame Challenge**
- **Wrap-up and Take away**

A first of its kind forum on the circular economy introducing the fundaments of this economic model, how the information industry can best adopt to the circular economy, and in particular the importance of grey literature resources in tracking and generating societal awareness.

This program is geared to librarians, archivists, researchers, and other information professionals and practitioners.

Instructor presentations will be delivered in a plenary setting and the participants will be actively engaged by way of examples and discussion throughout.

GreyNet
Grey Literature Network Service
www.greynet.org info@greynet.org

Twenty-First International Conference on Grey Literature

Open Science Encompasses New Forms of Grey Literature

German National Library of Science and Technology

Hannover, Germany • October 22-23, 2019

TIB LEIBNIZ INFORMATION CENTRE FOR SCIENCE AND TECHNOLOGY UNIVERSITY LIBRARY

Day 1 Conference Program

08:30 Registration Desk opens with Coffee and Tea Service

09:00 Opening Session

- **Welcome Address** — *Prof. Dr. Sören Auer, Director of the German National Library of Science and Technology*
- **Keynote Address** — **Unbreaking our Knowledge Sharing Workflows?**
  Barbara Rühling, CEO at Book Sprints Ltd., Berlin, Germany
- **Opening Paper** — **From Digitization and Digitalization to Digital Transformation: A Case for Grey Literature Management**
  Dobrica Savić, Nuclear Information Section, NIS-IAEA, United Nations

Moderator – Margret Plank, German National Library of Science and Technology

11:00 Session 1 – Open Science Principles promote the field of Grey Literature

- **Is Open Science FAIR for Stakeholders?**
  Plato L. Smith, University of Florida; George A. Smathers Libraries, United States
- **Abstracting and Indexing as an enabling interface between open science and grey literature – The approach of the Aquatic Sciences and Fisheries Abstracts service**
  Tamsin Vicary, Food and Agriculture Organization of the United Nations, Italy
  Ian Pettman, Freshwater Biological Association, United Kingdom
- **How Open Science Influences next Developments in Grey Literature**
  Julia M. Gelfand, University of California, Irvine Libraries, United States
  Anthony Lin, Irvine Valley College Library, United States
- **KAUST efforts in Open Access to Open Science**
  J.K. Vijayakumar, King Abdullah University of Science & Technology, Saudi Arabia
- **Hidden Grey Heritage of Science and Research from Pre-Internet Era**
  Tereza Klozová and Jiří Drozda, Research Institute of Geodesy, Topography and Cartography, Czech Republic

14:00 Session 2 – Confronting Obstacles and Challenges to Open Access

- **Open Access – A never-ending transition?**
  Silvia Giannini and Anna Molino, Institute of Information Science and Technologies, ISTI-CNR, Italy
- **AccessGrey: Securing Open Access to Grey Literature for Science and Society**
  Dominic Farace and Jerry Frantzen, GreyNet International, Netherlands
  Stefania Biagioni and Carlo Carlesi, NeMIS Research Laboratory, ISTI-CNR, Italy
- **Going Green – Publishing Academic Grey Literature in Laboratory Collections on HAL**
  Joachim Schöpfel, GERiiCO Laboratory, University of Lille, France
  Hélène Prost, GERiiCO Laboratory, CNRS, France
- **Research Data and Open Science in the Russian University Environment**
  Yuliya B. Balashova, Saint Petersburg State University, Russia
- **Is the Production and Use of Grey Marine Literature a Model for Open Science?**
  Bertrum H. MacDonald, Rachael Cadman, Curtis Martin, Simon Ryder-Burbidge, Suzuette S. Soomai, Ian Stewart, and Peter G. Wells; Dalhousie University, Canada

16:30 Introduction to Conference Posters and Sponsor Showcase

Lightening presentations on each poster in the main conference hall in advance of the Poster Session on Day Two.

Javastraat 194-HS, 1095 CP Amsterdam, Netherlands
www.textrelease.com • conference@textrelease.com
Tel.+31-20-331.2420

Twenty-First International Conference on Grey Literature

Open Science Encompasses New Forms of Grey Literature

German National Library of Science and Technology

Hannover, Germany • October 22-23, 2019

TIB LEIBNIZ INFORMATION CENTRE FOR SCIENCE AND TECHNOLOGY UNIVERSITY LIBRARY

Day 2 Conference Program

09:00 **REGISTRATION DESK OPENS WITH COFFEE AND TEA SERVICE**

09:30 **POSTER SESSION AND SPONSOR SHOWCASE**

The GL21 Call-for-Posters closes **September 20, 2019**. Guidelines for physical posters can be found on the conference site http://www.textrelease.com/gl21callforposters.html. Those presenting conference posters are eligible for the Poster Prize 2019 that will be awarded during the GL21 Closing Session. Posters will be judged on their innovative content, relevance to the conference, graphic design, and presentation.

11:30 13:00 – SPECIAL PANEL SESSION

The Open Science Publishing Flood and Collaborative Authoring

Simon Worthington **Chair**	*German National Library of Science and Technology, TIB Hannover, Germany*
Daniel Speicher	*Bonn-Aachen Int. Center for Information Technology, University of Bonn, Germany*
Denis Roio (Jaromil)	*Dyne.org Foundation, Netherlands*
Thomas Fricke	*Endocode AG, Germany*
Ludwig Hülk	*Reiner Lemoine Institut gGmbH, Germany*

Currently there is a deluge of 'off-piste' collaborative authoring going on in Open Science, the adoption of: software versioning systems like GitHub for writing; simulations and code in platforms like Jupyter; or in 'real time' web authoring 'operational transformation algorithm' based software like Etherpad. Yes the 'digital plumbing' of this publishing flood is just not in place. How do we ID these documents, reuse them for example. These questions have been asked and answered before, but are outside the fast moving, forward looking tech world. As an example, Ted Nelson coined the term 'transclusion' (Nelson 1987) back in the '80s. Transclusion is a step on from the 'hyperlink', another Nelson term, where not just the link of a target is included in a document but instead the whole content (AKA live linked embedding). Currently if you want to link and include another document fragment in your document with persistence, it's unlikely to work. The panelists will explore how this exciting field can better support research.

MODERATOR – PLATO L. SMITH, UNIVERSITY OF FLORIDA; GEORGE A. SMATHERS LIBRARIES, UNITED STATES

14:00 **SESSION 3 – OPEN RESOURCES FOR EDUCATION IN LIBRARY AND INFORMATION SCIENCE**

- **Open Access in the Academy: Developing a Library Program for Campus Engagement**
  Daniel C. Mack; University of Maryland, United States
- **Open Educational Resources and Library & Information Science: towards a common framework for methodological approaches and technical solutions**
  Roberto Puccinelli, Lisa Reggiani, Massimiliano Saccone, and Luciana Trufelli, National Research Council, Italy
- **Strategies for Teaching and Learning about Grey Literature**
  A Multimedia Presentation delivered by GreyNet's LIS Education and Training Committee
  Marcus Vaska, Alberta Health Services, Knowledge Resource Service (AHS-KRS), Canada
  David Baxter and Margo Hilbrecht, Gambling Research Exchange Ontario, Canada
  Sarah Bonato, Centre for Addiction and Mental Health, Canada
  Charles Scott Dorris, Georgetown University Medical Center; Dahlgren Memorial Library, United States

15:45 **CLOSING SESSION –** REPORT CONFERENCE MODERATORS, POSTER PRIZE, FAREWELL

17:00 **POST CONFERENCE TOUR –** GERMAN NATIONAL LIBRARY OF SCIENCE AND TECHNOLOGY
Tour led by Jens Olf, Head of Document Delivery, TIB Library

Twenty-First International Conference on Grey Literature

Open Science Encompasses New Forms of Grey Literature

German National Library of Science and Technology

Hannover, Germany • October 22-23, 2019

TIB LEIBNIZ INFORMATION CENTRE FOR SCIENCE AND TECHNOLOGY UNIVERSITY LIBRARY

Call for Posters

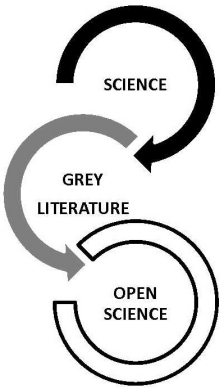
Title of Poster:	**Conference Topic(s):**
Author Name(s):	**Phone:**
Organization(s):	**Email:**
Postal Address:	**URL:**

Guidelines

Persons who seek to present a poster during GL21 are invited to submit an English abstract between 250-300 words. The abstract should describe the project, activity, information product or service. The abstract should likewise include a title, name(s) of the author(s) and their full address information. Abstracts are an important source of information available prior to the conference, which is accessible to both conference delegates and participants.

Related Conference Topics

- ☐ **Research Sharing relies on Open Data**
- ☐ **Open Source Software benefits Grey Literature**
- ☐ **Publishing Grey Literature opens up the Review Process**
- ☐ **Confronting Obstacles and Challenges to Open Access**
- ☐ **Open Resources for Education in Library and Information Science**
- ☐ **Open Science Principles promote the field of Grey Literature**
- ☐ **Other Related Topic:**



Due Date, Format, and Poster Prize 2019

Timely registration is a guarantee for your placement on the conference program. The conference venue is able to accommodate a limited number of physical posters (maximum size *A0 (841 x 1189 mm / 33.1 x 46.8 in). Posters* will be mounted on display panels allowing for optimal viewing. Abstracts in MSWord should be emailed to conference@textrelease.com on or before **September 20th 2019**. Those submitting poster abstracts will receive verification upon receipt accompanied by further guidelines for posters, submission of PowerPoint slides, Conference Registration, etc.

Those presenting conference posters are eligible for the Poster Prize 2019 that will be awarded during the GL21 Closing Session. Posters will be judged on their innovative content, relevance to the conference, graphic design, and presentation.

TextRelease
GL21 Program and Conference Bureau

Javastraat 194-HS, 1095 CP Amsterdam, Netherlands
www.textrelease.com ▪ conference@textrelease.com
Tel.+31-20-331.2420

Author Information

Farace, Dominic
Dominic Farace is Head of GreyNet International and Director of TextRelease, an independent information bureau specializing in grey literature and networked information. He holds degrees in sociology from Creighton University (BA) and the University of New Orleans (MA). His doctoral dissertation in social sciences is from the University of Utrecht, The Netherlands, where he has lived and worked since 1976. After six years heading the Department of Documentary Information at the Royal Netherlands Academy of Arts and Sciences (SWIDOC/KNAW), Farace founded GreyNet, Grey Literature Network Service in 1992. He has since been responsible for the International Conference Series on Grey Literature (1993-2018). In this capacity, he also serves as Program and Conference Director as well as managing editor of the Conference Proceedings. He is editor of The Grey Journal and provides workshops and training in the field of grey literature.
Email: info@greynet.org
ORCID iD https://orcid.org/0000-0003-2561-3631

Frantzen, Jerry
Jerry Frantzen graduated in 1999 from the Amsterdam University of Applied Sciences/Hogeschool van Amsterdam (HvA) in Library and Information Science. Frantzen is the technical editor of The Grey Journal (TGJ). And, since 1996, he is affiliated with GreyNet, Grey Literature Network Service, as a freelance technical consultant.
Email: info@greynet.org
ORCID iD https://orcid.org/0000-0002-3405-7078

Henderson, Kathrine A.
Kathrine Andrews Henderson is research analyst with LAC Group. She is part of a unique team of "virtual" researchers who provide "Library as a Service" to major law firms and corporations. Prior to this Ms. Henderson was the research librarian for the Office of the Auditor General for the State of Arizona. Earlier in her library career, Henderson was as an academic librarian. She was the Instructional Programs Librarian at Thunderbird School of Global Management and served in other roles including time as the business librarian for Arizona State University's Fletcher Library. Kathrine has expertise in business and legal research, intellectual property, and information ethics and has used this expertise to contribute to her field. Recently, she published a chapter on Intellectual Property Ethics in Foundations of Information Ethics, John Burgess and Emily Knox, editors. Other works include co-authoring Case Studies in Library and Information Science Ethics with Elizabeth Buchanan. In January 2018, Henderson was appointed to the Information Outlook Advisory Council for the Special Library Association. In the past, she served as Co-Director of the International Society for Ethics and Information Technology (INSEIT) and as an editor for ACM's Computers & Society. Henderson holds a Masters Degree in Library and Information Science from the University of Wisconsin-Milwaukee and a Bachelor of Science in Management from Arizona State University. In 2017, The School of Information Studies at UWM honored Kathrine as one of 50 Distinguished Alumni as part of the school's 50th Anniversary celebration.
Email: kat.henderson@libsource.com

Lipinski, Tomas A.
Tomas Lipinski is the Dean of the School of Information Studies at the University of Wisconsin Milwaukee. He completed his Juris Doctor (J.D.) from Marquette University Law School, Milwaukee, Wisconsin, received the Master of Laws (LL.M.) from The John Marshall Law School, Chicago, Illinois, and the Ph.D. from the Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. Dr. Lipinski has worked in a variety of legal settings including the private, public and non-profit sectors. In 2006 he was the first named Global Law Fellow, Faculty of Law, Catholic University of Leuven, Belgium where continues to lecture annually at its Centre for IT & IP Law and has been a visiting professor in summers at the University of Pretoria-School of Information Technology (Pretoria, South Africa) and at the Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. He is active in copyright education and policy-making, chairing PLA Legal Issues in Public Libraries Forum, Chair of the ALA Committee on Legislation Copyright Subcommittee, Chair of the ACRL Copyright Discussion Group from 2013 to 2016, Former Chair and at present member of the ALA OITP Copyright Education Subcommittee, a former member and now Expert Advisor to the Copyright and Other Legal Matters Committee of IFLA and serves as head of an NGO delegation with Permanent Observer status to the WIPO, World Intellectual Property Organization and its Standing Committee on Copyright and Other Rights.
Email: tlipinsk@uwm.edu

Author Information *CONTINUED*

Pizzanelli, Miguel
Miguel Pizzanelli, MD, MSc. - I was born in Montevideo in 1962. With Virginia my wife and dear partner, we share raising three children. At this moment we live in Florida, Uruguay. Since 1996 I spent almost all my medical practice time in small rural areas. For 7 years (from 2003 to 2010) we had the experience of living and working in a small rural village of 1500 inhabitants. I have varied interests, reading, I try to play several musical instruments in a self-taught way. I am general practitioner (family and community medicine) from 2003. I was part of the first generation of family and community medicine residents trained in Uruguay. I call this the zero generation (remembering of hard times that passed). I use to disseminate contents in various topics: quaternary prevention, rural medicine, critical thinking development. Quaternary prevention is a concept that defines an attitude ethically center oriented to provide health care focusing on persons trying to share health decisions with them in order to avoid overmedicalization. Since 2012 I began to actively participate in the society of family and community medicine in Uruguay and from that place in CIMF / WONCA. My role leading dissemination and applied of quaternary prevention concept pushed me to lead quaternary prevention working groups, first in my country later in Iberoamerican region and now in WONCA. My interest in classification and systematic terminologies makes me accept the invitation to participate in WONCA International Classification Committee in the quality of associate member from November 2014 up to date. Since 2008 we develop research focus on Barbara Starfield's Primary Care Assessment Tool in Uruguay. I participate actively in national regional and international CIMF WONCA Conferences (Praga 2013, Montevideo 2015, and Rio de Janeiro 2016). I think we need to fight both an individual and collective fight. The Individual fight to set collective interests over personal ones. Only through the collective work of all the family doctors and communities together all over the world, we will achieve "real" Primary Care: comprehensive health care, equity, and people-centered health care, focus on health better than illness, making reality the utopia of health for all in a better world.

Resnick, Melissa P.
Dr. Melissa P. Resnick, M.S., M.L.S., Ph.D.
Email: melissa_resnick@alum.rpi.edu
ORCID iD https://orcid.org/0000-0001-9699-852X

Savić, Dobrica
Dr. Dobrica Savić is Head of the Nuclear Information Section (NIS) of the IAEA. He holds a PhD degree from Middlesex University in London, an MPhil degree in Library and Information Science from Loughborough University, UK, an MA in International Relations from the University of Belgrade, Serbia, as well as a Graduate Diploma in Public Administration, Concordia University, Montreal, Canada. He has extensive experience in the management and operations of web, library, information and knowledge management, as well as records management and archives services across various United Nations Agencies, including UNV, UNESCO, World Bank, ICAO, and the IAEA. His main interests are digital transformation, creativity, innovation and use of information technology in library and information services.
Contact: www.linkedin.com/in/dobricasavic
ORCID ID: https://orcid.org/0000-0003-1123-9693

Schöpfel, Joachim
Joachim Schöpfel is senior lecturer at the Department of Information and Library Sciences at the Charles de Gaulle University of Lille 3 and Researcher at the GERiiCO laboratory. He is interested in scientific information, academic publishing, open access, grey literature and eScience. He is a member of GreyNet and euroCRIS. He is also the Director of the National Digitization Centre for PhD Theses (ANRT) in Lille, France.
Email: joachim.schopfel@univ-lille3.fr
ORCID iD https://orcid.org/0000-0002-4000-807X

Vaska, Marcus
Marcus Vaska is a librarian with the Knowledge Resource Service (KRS), Alberta Health Services (AHS), responsible for providing research and information support to staff at an Alberta Cancer Care Centre. A firm believer in embedded librarianship, Marcus engages himself in numerous activities, including instruction, patient engagement, and research consultation with numerous teams at this facility. An advocate of the Open Access Movement, Marcus' current interests focus on strategies for creating greater awareness of grey literature via various information dissemination and exchange pursuits.
Email: marcus.vaska@albertahealthservices.ca
ORCID iD https://orcid.org/0000-0002-4753-3213

Notes for Contributors

Non-Exclusive Rights Agreement

- I/We (the Author/s) hereby provide TextRelease (the Publisher) non-exclusive rights in print, digital, and electronic formats of the manuscript. In so doing,
- I/We allow TextRelease to act on my/our behalf to publish and distribute said work in whole or part provided all republications bear notice of its initial publication.
- I/We hereby state that this manuscript, including any tables, diagrams, or photographs does not infringe existing copyright agreements; and, thus indemnifies TextRelease against any such breach.
- I/We confer these rights without monetary compensation and with the understanding that TextRelease acts on behalf of the author/s.

Submission Requirements

Manuscripts should not exceed 15 double-spaced typed pages. The size of the page can be either A-4 or 8½x11 inches. Allow 4cm or 1½ inch from the top of each page. Provide the title, author(s) and affiliation(s) followed by your abstract, suggested keywords, and a brief biographical note.
A printout or PDF of the full text of your manuscript should be forwarded to the office of TextRelease. A corresponding MS Word file should either accompany the printed copy or be sent as an attachment by email. Both text and graphics are required in black and white.

REFERENCE GUIDELINES

General

i. All manuscripts should contain references
ii. Standardization should be maintained among the references provided
iii. The more complete and accurate a reference, the more guarantee of an article's content and subsequent review.

Specific

iv. Endnotes are preferred and should be numbered
v. Hyperlinks need the accompanying name of resource; a simple URL is not acceptable
vi. If the citation is to a corporate author, the acronym takes precedence
vii. If the document type is known, it should be stated at the close of a citation.
viii. If a citation is revised and refers to an edited and/or abridged work, the original source should also be mentioned.

Examples

Youngen, G.W. (1998), Citation patterns to traditional and electronic preprints in the published literature. - In: College & Research Libraries, 59 (5) Sep 1998, pp. 448-456. - ISSN 0010-0870

Crowe, J., G. Hodge, and D. Redmond (2010), Grey Literature Repositories: Tools for NGOs involved in public health activities in developing countries. – In: Grey Literature in Library and Information Studies, Chapter 13, pp. 199-214. – ISBN 978-3-598-11793-0

DCMI, Dublin Core Metadata Initiative Home Page http://purl.oclc.org/metadata/dublin_core/

Review Process

The Journal Editor first reviews each manuscript submitted. If the content is suited for publication and the submission requirements and guidelines complete, then the manuscript is sent to one or more Associate Editors for further review and comment. If the manuscript was previously published and there is no copyright infringement, then the Journal Editor could direct the manuscript straight away to the Technical Editor.

Journal Publication and Article Deposit

Once the journal article has completed the review process, it is scheduled for publication in The Grey Journal. If the Author indicated on the signed Rights Agreement that a preprint of the article be made available in GreyNet's Archive, then browsing and document delivery are immediately provided. Otherwise, this functionality is only available after the article's formal publication in the journal.

the grey Journal

15 YEARS

Victorine van Schaick Prize 2008

International Journal on Grey Literature

'Open Access to Grey Research Data'