

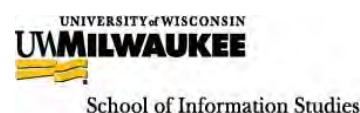
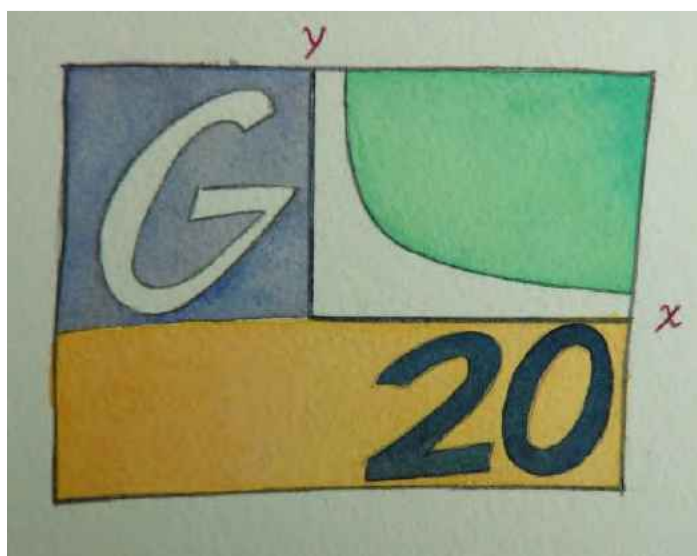
Twentieth International Conference on Grey Literature

Research Data Fuels and Sustains Grey Literature

Loyola University New Orleans, USA • December 3-4, 2018

Conference Proceedings

ISSN 1386-2316





CIP

GL20 Proceedings

Twentieth International Conference on Grey Literature "Research Data Fuels and Sustains Grey Literature". - Loyola University New Orleans, USA December 3-4, 2018 / compiled by D. Farace and J. Frantzen ; GreyNet International, Grey Literature Network Service. – Amsterdam : TextRelease, February 2019. – 161 p. – Author Index. – (GL Conference Series, ISSN 1386-2316 ; No. 20).

ISTI-CNR (IT), TIB Hannover (DE), DANS-KNAW (NL), FEDLINK; Library of Congress (USA), CVTISR (SK), EBSCO (USA), Inist-CNRS (FR), KISTI (KR), NIS-IAEA (AT), NTK (CZ), and the University of Florida; George A. Smathers Libraries (USA) are Corporate Authors and Associate Members of GreyNet International. These proceedings contain full text conference papers presented during the two days of plenary and poster sessions. The papers appear in the same order as in the conference program book. Included is an author index with the names of contributing authors and researchers along with their biographical notes. A list of 60 participating organizations as well as sponsored advertisements are likewise included.

Foreword

RESEARCH DATA FUELS AND SUSTAINS GREY LITERATURE

The definition of research data is as encompassing as the field of grey literature. What should be included and what should be excluded is and remains an issue of concern. Research data can be defined as factual materials collected by diverse communities of practice required to validate findings. While the majority of research data is created in digital format, research data in other formats cannot be excluded. The formats in which research data appear are multiple and the types of research data are diverse. This also holds for the numerous document types in which grey literature appear published.

Today, while emphasis is placed on big data, the fact that the majority of research projects are small to medium size is overlooked. This is but another characteristic that holds true for grey literature. Nonetheless, one should be aware that research publications are not research data, for they are often managed separately from one another. Just as there are a number of stakeholders involved in the production, access, and preservation of grey publications, so too are there stakeholders tasked with the creation and management of research data. Libraries and data management librarians have the responsibility for the curation of the data they collect and preserve. And, it is important to stress the need to maintain appropriate metadata related to research data in order to facilitate their interpretation and further reuse.

Over the past quarter century, grey literature communities have worked diligently to demonstrate how their documents are produced, published, reviewed, indexed, accessed, and further used, applied, and preserved. Today, these communities are now challenged to demonstrate how research data fuels and sustains their grey literature. These communities of dedicated researchers and authors maintain a strong conviction in the uses and applications of grey literature for science and society. Through the years, they have proved willing to share the results of scholarly work well beyond their own institutions. Hence, one can assume they are aware that innovation forfeits with the loss of data as with the loss of information. This 20th International Conference in the GL-Series seeks to address key issues and topics related to grey literature and its underlying research data.

Dominic Farace
GREYNET INTERNATIONAL

Amsterdam,
FEBRUARY 2019

GL20 Conference Sponsors



ISTI, Italy

Institute of Information Science and Technologies
National Research Council of Italy, CNR



CVTISR, Slovak Republic

Slovak Centre of Scientific and Technical Information



KISTI, Korea

Korea Institute of Science and Technology
Information



EBSCO, USA



NIS-IAEA, Austria

Nuclear Information Section;
International Atomic Energy Agency



TIB, Germany

German National Library of Science and Technology –
Leibniz Information Centre for Science and
Technology University Library

GL20 Conference Sponsors (CONTINUED)



DANS, Netherlands
Data Archiving and Networked Services;
Royal Netherlands Academy of Arts and Sciences



NTK, Czech Republic
National Library of Technology



FEDLINK, USA
Federal Library Information Network;
Library of Congress



Inist-CNRS, France
Institut de l'Information Scientifique et Technique; Centre
National de Recherche Scientifique



UW-Milwaukee-SOIS, USA
School of Information Studies (SOIS)
University of Wisconsin, Milwaukee



UF, USA
George A. Smathers Libraries
University of Florida

GL20 Program Committee



Brian Hitson ^{Chair}
Office of Scientific and
Technical Information;
U.S. Department of
Energy



Robert Bell
Loyola University
New Orleans, USA



George Barnum
U.S. Government
Publishing Office



Meg Tulloch
U.S. Government
Accountability Office



Margret Plank
German National Library
of Science and
Technology, Germany



Dobrica Savić
Nuclear Information
Section, International
Atomic Energy Agency,
Austria



Christiane Stock
Institut de l'Information
Scientifique et Technique
CNRS, France



Hana Vyčítalová
National Library of
Technology,
Czech Republic



Silvia Giannini
Institute of Information
Science and Technologies
ISTI-CNR, Italy



Ján Turňa
Slovak Centre of
Scientific and Technical
Information Slovak
Republic



Stefania Biagioni
NeMIS Research
Laboratory
Italy



Joachim Schöpfel
University of Lille
France



Judith C. Russell
University of Florida
Libraries
USA



Plato L. Smith
University of Florida;
George A. Smathers
Libraries, USA



Henk Harmsen
Data Archiving and
Networked Services,
Netherlands



Marcus Vaska
Alberta Health Services
Canada



Dominic Farace
GreyNet International
Netherlands



Tomas A. Lipinski
University of Wisconsin
Milwaukee, USA

Table of Contents

	Foreword.....	3
	Conference Sponsors.....	4
	Program Committee	6
	Program Chair and Conference Moderators	9
	Conference Program.....	11
<i>Program</i>	Opening Session.....	12
	Session One – Research Data and Open Access Compliance.....	19
	Session Two – Data Management and the Role of Librarians.....	51
	Poster Session and Sponsor Showcase.....	99
	Session Three – Current Research Trends in Grey Literature.....	117
<i>Info Adverts</i>	WorldWideScience.org: An International Partnership Supporting Open Science.....	8
	FEDLINK, The Federal Library and Information Network - Library of Congress.....	10
	INIS, The International Nuclear Information System.....	18
	ISTI-CNR, Institute of Information Science and Technologies.....	50
	Ebsco Library, Information Science & Technology Abstracts with Full Text (LISTA).....	98
	The Grey Journal, 15 Years Flagship Journal for Grey Literature.....	114
	TIB, German National Library of Science and Technology.....	115
	KISTI, Korea Institute of Science and Technology Information.....	116
	NTK, National Library of Technology, Czech Republic.....	140
	CVTISR, Slovak Centre of Scientific and Technical Information.....	152
	GL20 Conference Proceedings Order Form.....	160
<i>Appendices</i>	List of Participating Organizations	153
	GL21 Conference Announcement.....	154
	GL21 Call for Papers.....	155
	Author information.....	156
	Index to Authors.....	161

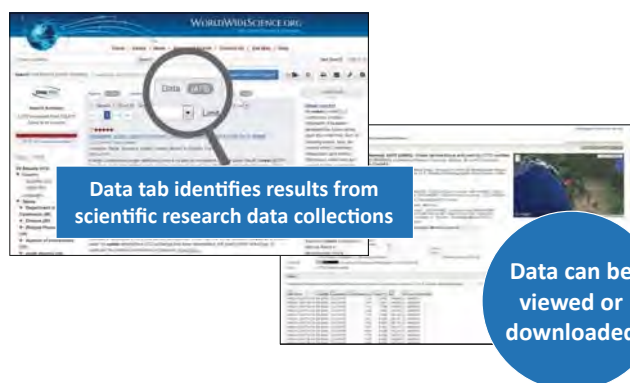
WorldWideScience.org

An International Partnership Supporting Open Science



- ▶ Simultaneously explore over 100 national and international scientific databases and portals from 75 countries
- ▶ Search information in textual, multimedia, software, and scientific data formats
- ▶ Eliminate language barriers through multilingual translations across ten languages

Search and Access Scientific Research Data



Modern Science Demands Reproducibility and Open Access to Publications, Datasets, and Software



Operating Agent:

OSTI.GOV

WorldWideScience.org Operating Agent
Point of Contact: [Lorrie Johnson, JohnsonL@osti.gov](mailto:Lorrie.Johnson@osti.gov)

Program Chair and Conference Moderators



Moderator Day One

Robert Bell
Director of Learning Resources
Loyola University New Orleans

Robert Bell is Director for the Office of Writing and Learning Services at Loyola University. He was born and raised in New Orleans and its confines. He received his education in New Orleans: B.A. from Loyola University New Orleans, M.F.A. from the University of New Orleans. In addition to being the Director for OWLS, he teaches undergraduate writing and literature courses, and directs the Loyola Ireland Summer Abroad program.

He recently published "America's Disaster Culture: The Production of Natural Disasters in Literature and Pop Culture" and edited "Eco Culture: Disaster, Narrative, Discourse," both with Robert M. Ficociello. Additionally, Bell is the Disasters, Apocalypses, and Catastrophes area Co-Chair for the Popular Culture/American Culture Association.

Email: rcbell@loyno.edu



Program Chairman

Brian A. Hitson
Director
U.S. Department of Energy; OSTI

Brian Hitson is Director of the U.S. Department of Energy (DOE) Office of Scientific and Technical Information (OSTI), a DOE corporate function that is managed by the Office of Science. OSTI fulfills agency-wide responsibilities to collect, preserve, and disseminate scientific and technical information emanating from DOE research and development (R&D) activities. Prior to becoming Director, Brian led a range of OSTI's programmatic and administrative activities. He has also managed OSTI's international information exchange programs, as well as the digitization and preservation of a scientific document repository. As Director, he has led strategic efforts to improve discoverability and linkages between diverse, related research objects, including publications, datasets, and scientific software available at OSTI.GOV. He played a key role in the development of WorldWideScience.org and in the establishment of the WorldWideScience Alliance in 2008.

Email: HitsonB@osti.gov



Moderator Day Two

Judith C. Russell
Dean of University Libraries
University of Florida

Judith C. Russell is the Dean of University Libraries at the University of Florida. She was formerly the Managing Director, Information Dissemination and Superintendent of Documents, at the U.S. Government Printing Office (GPO). Russell previously served as Deputy Director of the National Commission on Libraries and Information Science (NCLIS) and as director of the Office of Electronic Information Dissemination Services and Federal Depository Library Program at GPO. She worked for more than 10 years in the information industry in marketing and product development, as well as serving as a government-industry liaison. Her corporate experience includes Information Handling Services (IHS) and its parent company, the Information Technology Group; Disclosure Information Group; Lexis-Nexis (former Mead Data Central), and IDD Digital Alliances, a subsidiary of Investment Dealers Digest. She has an M.L.S. from Catholic University and a B.A. from Dunbarton College of the Holy Cross.

Email: jcrussell@ufl.edu



FEDLINK

an organization
of federal agencies
working together
to achieve optimum use of
resources and facilities
of federal libraries
and information centers
by promoting
common services,
coordinating and sharing
available resources, and
providing continuing
professional education.



101 Independence Ave, SE ~ Washington, DC 20540-4935
FEDLINK Main Number (202) 707-4800
FEDLINK Hotline (202) 707-4900

Strategic Sourcing



Currently, more than 20 federal agencies, both military and civilian, including FEDLINK, participate in the Fed Strategic Sourcing Initiative (FSSI). FSSI was created in 2005 by the Department of the Treasury, the Office of Management and Budget, and the General Services Administration. FSSI provides a variety of products and services that can be purchased more efficiently and at lower costs to the federal government.

strategic sourcing, FSSI agencies also provide centralized acquisition functions for a variety of products to streamline efficiency and reduce costs to the federal government.

Conference Program

OPENING SESSION

- The U.S. Government Publishing Office: Keeping America Informed in the 21st Century and Beyond** 12
Cynthia Etkin, Senior Program Planning Specialist, GPO - U.S. Government Publishing Office, United States

SESSION ONE – RESEARCH DATA AND OPEN ACCESS COMPLIANCE

- When is ‘grey’ too ‘grey’? A case of grey data** 19
Dobrica Savić, Nuclear Information Section, International Atomic Energy Agency, NIS-IAEA, United Nations
- Legal Issues Surrounding the Collection, Use and Access to Grey Data in the University Setting;
How Data Policies Reflect the Political Will of Organizations** 25
Tomas A. Lipinski, School of Information Studies; University of Wisconsin-Milwaukee, United States
Kathrine A. Henderson; LibSource, A LAC-Group Company, United States
- On Open Access to Research Data: Experiences and reflections from DANS** 40
Emilie Kraaikamp, Marjan Grootveld, Hella Hollander, and Dirk Roorda,
Data Archiving and Networked Services, DANS-KNAW, Netherlands

SESSION TWO: DATA MANAGEMENT AND THE ROLE OF LIBRARIANS

- The data librarian: myth, reality or utopia?** 51
Silvia Giannini and Anna Molino, Institute of Information Science and Technologies, ISTI-CNR, Italy
- Research Data Management: What can librarians really help?** 67
Yuan Li, Willow Dressel and Denise Hersey, Princeton University, United States
- Data Management and the Role of Librarians** 75
Plato L. Smith, Jean Bossart, and Sara Gonzalez, George A. Smathers Libraries; University of Florida, United States
- Measuring Reuse of Institutionally-Hosted Grey Literature** 83
Ayla Stein Kenfield, University of Illinois at Urbana–Champaign; Elizabeth Kelly, Loyola University New Orleans;
Caroline Muglia, University of Southern California; Genya O’Gara, Virtual Library of Virginia; Santi Thompson,
University of Houston; Liz Woolcott, Utah State University, United States
- Librarians’ Role in GAO Reports** 91
Meg Tulloch, U.S. Government Accountability Office, GAO, United States

POSTER SESSION AND SPONSOR SHOWCASE

- Published electronic media are becoming Grey** 99
Yui Kumazaki, Satoru Suzuki, Masashi Kanazawa, Katsuhiko Kunii, Minoru Yonezawa, and Keizo Itabashi
Japan Atomic Energy Agency
- Semantic Query Analysis from the Global Science Gateway** 105
Sara Goggi, Gabriella Pardelli, Roberto Bartolini, and Monica Monachini, ILC-CNR, Italy
Stefania Biagioni and Carlo Carlesi, ISTI-CNR, Italy

SESSION THREE – CURRENT RESEARCH TRENDS IN GREY LITERATURE

- Open Data engages Citation and Reuse: A Follow-up Study on Enhanced Publication** 117
Dominic Farace, Jerry Frantzen, GreyNet International, Netherlands; Joachim Schöpfel, University of Lille, France
- The Q-Codes: Metadata, Research data, and Desiderata, Oh My! Improving Access to Grey Literature** 123
Melissa P. Resnick, University of Texas Health Science Center at Houston, United States [et al]
- Analysis of folk literature in grey literature from the National Library of China** 133
Cui Yue, National Library of China, Beijing, China
- When the Virtual Becomes Reality: An Environmental Scan of the Presence of Virtual Reality and
Artificial Intelligence in Health and Cancer Care Environments** 141
Marcus Vaska

Opening Address

The U.S. Government Publishing Office: Keeping America Informed in the 21st Century and Beyond

Cynthia Etkin, Senior Program Planning Specialist, GPO
Office of the Superintendent of Documents
U.S. Government Publishing Office

Abstract

On June 23, 1860, the 36th Congress of the United States approved a joint resolution that created the Government Printing Office (GPO), and directed the Superintendent of Public Printing to have executed the printing and binding of documents approved by the Senate and House of Representatives, and the executive and judicial branch departments. GPO began operation with 350 employees on the day that Abraham Lincoln was inaugurated as the 16th President of the United States, March 4, 1861. The GPO has a long history and it is a remarkable story.

The Printing Act of 1895 overhauled existing printing laws and created a number of new authorities for GPO. Among them was to disseminate Government public documents through three channels: deposit in designated libraries; distribution to Government entities; and sales. Another new authority was to organize the Government's information, with the provision of a cataloging and indexing program. From this landmark legislation came two important public information programs, the Cataloging and Indexing Program (C&I) and the Federal Depository Library Program (FDLP). Both of these programs became integral to GPO's mission of Keeping America Informed.

More recently, with the increase in digital communication and expanding publishing technologies, GPO has streamlined and transformed from a print-only operation to an integrated publishing organization. Recognizing this, Congress re-designated the agency the U.S. Government Publishing Office on December 17, 2014.

The GPO has continually transformed itself throughout its history by adapting to changing technologies. In the ink-on-paper era this meant moving from hand-set to machine type-setting to digital type setting; from slower manual fed presses to high speed presses; and from hand to automated bookbinding.

New strategic priorities for GPO and a National Plan for Access to U.S. Government Information for GPO's public information programs are guiding another transformation for Keeping America Informed in the 21st century and beyond.

Remarks before the Twentieth International Conference on Grey Literature Loyola University, New Orleans, LA December 3, 2018

"Good morning! And let me add my welcome to New Orleans. There are other names you may hear by which the city is known: The Crescent City, The Big Easy, and The Birthplace of Jazz. I call it a special place, and I am happy to be here.

It is an honor to be here this morning representing the U.S. Government Publishing Office, the GPO, and delivering the Opening Address at this the Twentieth International Conference on Grey Literature. I have been at the GPO for more than 20 years now and I suspect, like many of you, like to talk about the place I work. I am proud of the agency, its mission, its history, and its future. This morning I will share a little bit about the history of GPO and then I will describe the path we are taking into the future.

In June 1860, the 36th Congress of the United States approved Joint Resolution 25, *Joint Resolution in Relation to Public Printing*, which created the Government Printing Office (GPO), and directed the Superintendent of Public Printing to execute the printing and binding of documents approved by Congress and by the departments of the executive and judicial branches of the Government. GPO began operation as an agency of the Legislative

Branch of Government nearly a year later with 26 presses, the steam engines to run them, and 350 employees. It was March 4, 1861, the same day Abraham Lincoln was inaugurated as the 16th President of the United States.

The Printing Act of 1895 overhauled existing printing laws and created a number of new authorities for GPO. Among them was to disseminate Government public documents through three channels: deposit in designated libraries; distribution to Government entities; and sales. Another new authority was to organize the Government's information, with the provision of a cataloging and indexing program. From this landmark legislation came two important public information programs, the Cataloging and Indexing Program (C&I) and the Federal Depository Library Program (FDLP). Both of these programs became, and remain, integral to GPO's mission of *Keeping America Informed*. More about these important programs in a minute.

As one might well imagine, GPO's business grew. And around the beginning of the twentieth century — and for several decades through the 1970s, GPO was known as "The Largest Print Shop in the World." The sheer number of presses put GPO in a league of its own. The 1926 inventory shows GPO having close to 200 presses, and it is said that a good sized commercial printer might have half as many presses. But it wasn't just the numbers, the variety of presses also made GPO unique. GPO's inventory contained every type of press found in commerce, allowing an almost unlimited scope of GPO products. "If a Government customer wanted a particular kind of printed document, GPO could produce it."

GPO's growth peaked in the late 1970s, early 1980s. At the time, in addition to the plant in Washington, D.C., there were six regional printing plants and 24 regional bookstores, and almost 9,000 employees. By the end of the 80s the number of presses had declined, but they were replaced with more productive and efficient presses that allowed production of billions of publications. In the 1990s the regional printing plants were closed with the exception of the Denver plant, which closed in 2003. In 2001 the bookstores located throughout the nation began to close.

GPO entered the online world in 1994. Our game-changing moment occurred in 1993 when, with the strong support of the library community, Congress enacted the *GPO Electronic Information Access Enhancement Act* (Public Law 103-40), which directed the GPO to create a publicly accessible system of online access to make available the *Congressional Record*, the *Federal Register*, and other publications as determined by the Superintendent of Documents.¹ GPO was given one year to implement this, and we met the challenge. This landmark legislation gave GPO the necessary authority to assume an essential role in the provision of information services and dissemination of electronic information. Access to this system, which became known as GPO Access, was free for Federal depository libraries and in 1995 it was made free to anyone. As a result of this action, the face of Government information as provided by the GPO began to shift away from tangible print toward a digital presence and in the process improved our operational efficiency and expanded our information dissemination capabilities. Fifteen years later *GPO Access* was replaced with *GPO's Federal Digital System* (FDsys), a significantly reengineered system that is OAIS-compliant (open archival information system). FDsys is a content management system, a preservation repository, and an advanced search engine. There are more than 2.2 million titles available through FDsys and it averages 45 million retrievals per month. Next week on 14 December FDsys will sunset and the third generation of online system of record will be **govinfo**.² **govinfo** was launched in early 2016 in beta.

The impact on the GPO business model and infrastructure has been dramatic. In 1993, before this act, we printed 25,000 copies of the Congressional Record every day. Today, we print 2,500 copies per day. We used to print 35,000 copies of the Federal Register each day. Today, we print fewer than 2,000. Although we print fewer copies of everything, our reach has significantly expanded because of the digital presence.

¹ Public Law 103-40 is codified in Title 44, *United States Code*, Chapter 41— Access To Federal Electronic Information. <https://www.govinfo.gov/content/pkg/USCODE-2017-title44/pdf/USCODE-2017-title44-chap41.pdf>.

² An update since the GL20 International Conference on Grey Literature, **govinfo** is now GPO's official system of record. Find **govinfo** at: <https://www.govinfo.gov/>.

Today headquarters is at the same location on North Capitol St, NW where it first opened its doors in 1861. Now, however, there are 4 buildings that together have 33 acres or more than 13 hectares of floor space with 23 presses; a bookstore; 1,740 employees; and a strong online presence.

FEDERAL DEPOSITORY LIBRARY AND CATALOGING AND INDEXING PROGRAMS

Now I want to touch on the Federal Depository Library Program (FDLP) and the Cataloging and Indexing Program. I mentioned these earlier when describing new authorities that came to GPO with the Printing Act of 1895. Both of these programs are under the purview of the Superintendent of Documents' Public Information Programs.

The Founding Fathers and early legislators recognized the importance of the free flow of information in a democratic society. They thought it essential that the citizenry be informed about its government and its workings so as to allow effective participation in the democratic process. It is this thinking that underlies the establishment of the Federal Depository Library Program (FDLP) as a network of libraries located throughout the country and territories with collections of Government publications of public interest and educational purposes that are freely accessible to the general public.

The library program traces its roots to 1813 when Congress passed a joint resolution for the printing and distribution of the journals of Congress to the executive branches and the legislative branches of every state and territorial legislature, one copy to each university and college in each state, and one copy to the Historical Society incorporated in each state.

Today we have 1,132 libraries — libraries of all types and sizes (we have some of the largest and some of the smallest libraries in the world) — participating in the depository library program. And I am pleased to say that there are ten depository libraries in New Orleans, and I want to recognize that our conference host, Loyola University of New Orleans, has two libraries in our program, the main campus library (since 1942) and the college of law library (since 1978).

GPO provides depository libraries with information products in digital format — and there's still some printed material (and microfiche) going to libraries, in return the libraries must provide free access to the public to these resources, and assistance to those who use them. Depository libraries are a critical link between "We the People" and the Government's information, and GPO's primary information dissemination program.

We now have a number of libraries that have joined the FDLP as "digital only" libraries, though some of our long-standing member libraries are moving in this direction by making available digital content and weeding their tangible collections of Government publications.

The Cataloging and Indexing Program is statutorily mandated to acquire Federal agency information products and develop a comprehensive and authoritative national bibliography of U.S. Government publications. Beginning with 1895, the bibliography was issued as a printed monthly catalog. It was replaced in 2004 with an online public access catalog, the *Catalog of U.S. Government Publications* (CGP).³ The initial record load included cataloging from 1976, which is when GPO first had machine readable records. GPO continues to work on retrospective conversion of records for inclusion in the catalog, while exploring new technologies and methods to enhance discovery of Government information.

In fiscal year 2017 GPO distributed 4,200 tangible titles to depository libraries, and created nearly 11,000 PURLs (persistent uniform resource locator) to digital content accessible through the bibliographic records in the *Catalog of U.S. Government Publications*. There were nearly 33.2 million searches of the catalog from around the world.

NATIONAL ACADEMY OF PUBLIC ADMINISTRATION REPORT

In 2012 at the request of Congress, the National Academy of Public Administration (NAPA) conducted an operational review of GPO, a 21st century readiness study. NAPA's final report, *Rebooting the Government Printing Office: Keeping America Informed in the*

³ *Catalog of U.S. Government Publications* is available from: <https://catalog.gpo.gov>. The statutory authority for the Cataloging and Indexing Program is in Title 44 *United States Code* §§1710-1711. <https://www.govinfo.gov/content/pkg/USCODE-2017-title44/pdf/USCODE-2017-title44-chap17-sec1710.pdf>.

Digital Age,⁴ was released January 2013. Of great importance, NAPA determined that “GPO’s core mission of authenticating, preserving, and distributing federal information remains critically important to American democracy,” while at the same time they made recommendations to strengthen GPO’s business model for the future. GPO is working to implement the recommendations of the NAPA panel.

Among the recommendations of the NAPA Panel that greatly affects the Public Information Programs of the Superintendent of Documents is one related to preservation, safeguarding the historical documents of our democracy for future generations. It has wide-sweeping outcomes: “GPO should work with depository libraries and other library groups to develop a comprehensive plan for preserving the print collection of government documents.” Further they stated that the plan should include cataloging, digitizing, and preserving tangible copies of government publications, as well as a process for ingesting digitized copies into GPO’s system of record for online access. This is a huge undertaking, but the Superintendent of Documents was pleased to see this recommendation, for it gets to the heart of the Public Information Programs — permanent public access to Government information.

MOVING FORWARD

GPO continues to transform and move forward with two documents as guides: *GPO Strategic Plan*, a five-year rolling strategic plan; and *National Plan for Access to U.S. Government Information*, a framework for a user-centric service approach to permanent public access.

GPO STRATEGIC PLAN

The FY 18-22 Strategic Plan⁵ states GPO’s vision as “an informed nation that has convenient and reliable access to their government’s information through GPO’s products and services.” The plan has five strategic goals:

1. Exceed our stakeholders’ expectations
2. Enhance access to Federal Government information
3. Strengthen our position as the government-wide authority on publishing
4. Promote collaboration and innovation within government
5. Engage employees and enhance internal operations

I am not going to talk about all of these, but I do want to say a few things about enhancing access to Federal Government information and promoting collaboration and innovation within government.

Strategies for achieving the second goal, enhanced access to Government information, are to increase the amount of U.S. Government information available for free to the public and enhance access to information to meet evolving user needs; and support access and discoverability through the Federal Depository Library Program and the Cataloging and Indexing Program. This means incorporating missing records and enhancing existing bibliographic records for historical materials. We aim to make the *Catalog of U.S. Government Publications* comprehensive index of every document issued or published by a department, bureau, agency, or office that is not confidential in character. This also means that GPO will continue to add to the collections in **govinfo**, with the goal of offering complete and historic content.

Strategies for achieving the fourth goal, promoting collaboration and innovation within government, include developing partnerships with Federal agencies that will improve identification of and increase access to their information. This also includes the establishment of the Federal Publishing Council, an advisory group made up of Federal employees involved in all facets of the Federal printing and publishing community.

⁴ *Rebooting the Government Printing Office: Keeping America Informed in the Digital Age* can be viewed and downloaded from GPO’s website at: https://www.gpo.gov/docs/default-source/congressional-relations-pdf-files/gpo_napa_report_final.pdf?sfvrsn=2.

⁵ GPO’s FY18-22 Strategic Plan is available for viewing and downloading from: <https://www.gpo.gov/who-we-are/our-agency/mission-vision-and-goals>.

NATIONAL PLAN FOR ACCESS TO U.S. GOVERNMENT INFORMATION

The *National Plan for Access to U.S. Government Information*⁶ sets the strategic direction for the Federal Depository Library and Cataloging and Indexing programs, with the vision “to provide Government information when and where it is needed.” It is guided by GPO’s mission of *Keeping America Informed*, and it was informed by the NAPA study of GPO, external influences, and most particularly by the results of the FDLP Forecast Study. The Forecast Study, already underway when NAPA came to GPO for their study, was a data-gathering effort to obtain information that would allow GPO to better understand the issues and challenges facing depository libraries, document their needs, and their view of the ideal depository library program of the future.

The three strategic priorities set forth in the National Plan are:

1. Establish Library Services and Content Management processes and procedures that apply lifecycle management best practices for all formats and ensure permanent public access to Government information dissemination products in the digital age.
2. Provide a governance process and sustainable network structure that ensures coordination across the Federal Depository Library Program and allows the most flexible and effective management of depository libraries and their resources.
3. Deliver dynamic, innovative, strategic services and mechanisms to support the needs of Federal depository libraries in providing accurate Government information to the public at large in a timely manner.

The first strategic priority looks internally at the organization and process improvement, while the second looks at how GPO can better administer the FDLP to meet the changing needs of the libraries. The third considers what new or enhanced tools and services GPO can provide libraries that will better support them. Some of the outcomes are achievable in the near term, while others can only be achieved over the long-term.

The introduction of online digital content into the FDLP was a major paradigm shift for GPO. GPO has sent Government publications to libraries for more than 200 years. Materials sent to depository libraries remain the property of the U.S. Government; libraries are the stewards of those collections. Collections have always been in the depository libraries and now GPO has a digital collection to curate and a distributed tangible national collection to preserve.

A lot of activities taking place revolve around preservation, because you can’t have permanent public access to content if it isn’t preserved. We are building a sustainable preservation program within our Library Services and Content Management business unit. We’re expanding the number of staff in this area, and we are getting ready to pilot how GPO can provide various preservation services to depository libraries. We have 40 partnerships⁷ with depository libraries and Federal agencies that are providing digital content for ingest into **govinfo**, preserving tangible content in their collections, sharing cataloging records, or providing access to or preserving digital content in their repositories. GPO recognizes that to be successful in preserving Government information, collaboration is necessary. We call our collaboration the Federal Information Preservation Network, or FIPNet.

In the area of digital preservation, GPO is currently undergoing an audit for **govinfo** to be certified as a trustworthy digital repository under ISO 16363: 2012, Space data and information transfer systems -- Audit and certification of trustworthy digital repositories. **govinfo** will be assessed against 109 criteria related to organization infrastructure, digital object management, and infrastructure and security risk management. The latest update on the audit? The auditing team will be making their site visit to GPO later this week.⁸

During the Senate confirmation hearing for a former Public Printer of the United States, the question was asked, “Do we still need the depository library program when people can access Government information from their living room?” The response was, without

⁶ *National Plan for Access to U.S. Government Information* is available for viewing and downloading from: <https://www.fdlp.gov/superintendent-of-documents-public-policies>.

⁷ For more information about partnerships see: <https://www.fdlp.gov/about-the-fdlp/partnerships>.

⁸ I am pleased to provide a further update on the ISO 16363 audit. After the onsite visit, the auditing team recommended to the certification committee that GPO be awarded certification as a trustworthy digital repository under the international standard; the committee concurred. For more information on GPO and the Trustworthy Digital Repository certification process see: <https://www.fdlp.gov/preservation/trusted-digital-repository-iso-16363-2012-audit-and-certification>.

hesitation, a resounding, “YES! But GPO needs to better support them, and help them manage in the digital age.” And that’s what GPO is doing, with the GPO Strategic Plan and the National Plan as guides.

A NEW MIDDLE NAME

On December 16, 2014 Congress redesignated the Government Printing Office to the Government Publishing Office. “This is a historic day for GPO. Publishing defines a broad range of services that includes print, digital, and future technological advancements. The name Government Publishing Office better reflects the services that GPO currently provides and will provide in the future,” said Davita Vance-Cooks, who was then Director of the Government Publishing Office, the agency’s chief executive officer.⁹ Modern publishing operations provide a suite of services to ensure access to the information they disseminate, from conventional printing to eBooks, digital publishing, mobile access, social media, and other strategies. All of these services are provided by GPO today.

CONCLUSION

The importance of an informed public to the effective conduct of self-government prompted James Madison to write in August 1822:

A popular Government without popular information, or the means of acquiring it, is but a Prologue to a Farce or a Tragedy, or perhaps both. Knowledge will forever govern ignorance, and a people who mean to be their own Governors, must arm themselves with the power which knowledge gives¹⁰.

Since its founding, GPO has performed that function by producing and distributing documents created by our government, serving as the “means of acquiring” information for “a people who mean to be their own Governors” through Federal depository libraries and now also through **govinfo** and the *Catalog of U.S. Government Publications*.

GPO is in a good place. Today, GPO is fundamentally different from what it was as recently as a generation ago: smaller, leaner, and equipped with digital production capabilities that are the bedrock of the information systems relied upon daily by Congress, Federal agencies, depository libraries, and the public to ensure open and transparent Government in the digital era.

As it was with its history, GPO will continue to strategize, plan, and transform. And GPO will be *Keeping America Informed* for future generations, beyond the 21st century. Thank you very much.

“

⁹ GPO News Release No. 14-27, *GPO IS NOW THE GOVERNMENT PUBLISHING OFFICE*. December 17, 2014. <https://www.gpo.gov/docs/default-source/news-content-pdf-files/2014/14news27.pdf>.

¹⁰ United States. Government Publishing Office. *Keeping America Informed, the U.S. Government Publishing Office: A Legacy of Service to the Nation, 1861-2016*, Washington, DC: United States Government Publishing Office, 2016. Available from the GPO website at: <https://www.govinfo.gov/content/pkg/GPO-KEEPINGAMERICAINFORMED-2016/pdf/GPO-KEEPINGAMERICAINFORMED-2016.pdf>.



International Nuclear Information System INIS

*organizing the world's information
on nuclear science and technology
and making it universally accessible
for peaceful uses*

over 150 Member States and
international organizations

millions of citations and
abstracts published worldwide

hundreds of thousands of full text
non-conventional 'grey' literature

multilingual thesaurus in Arabic,
Chinese, English, French, German,
Japanese, Russian, Spanish



www.iaea.org/inis

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 data¹ 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).

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

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 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 et 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.

Concept	Situation		
	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, **Trustworthiness**, is missing. **Trustworthiness** 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, widespread 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://onlinelibrary.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. *ERICIM 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 policies 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 *sine 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 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 these 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. WiredData, 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 (<i>Purdue University Research Repository</i>)	https://purr.purdue.edu/	The <i>Purdue University</i> Research Repository (<i>PURR</i>) 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., Kurbanoglu, 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. Burtchael*, 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 2013¹.

DANS supports this target. DANS, short for Data Archiving and Networked Services², 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 output³. 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.

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 Initiative⁵, the Bethesda Statement on Open Access Publishing⁶, and the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities⁷. 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 2017⁸. 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 NARCIS⁹. 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 Unpaywall¹⁰ 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>

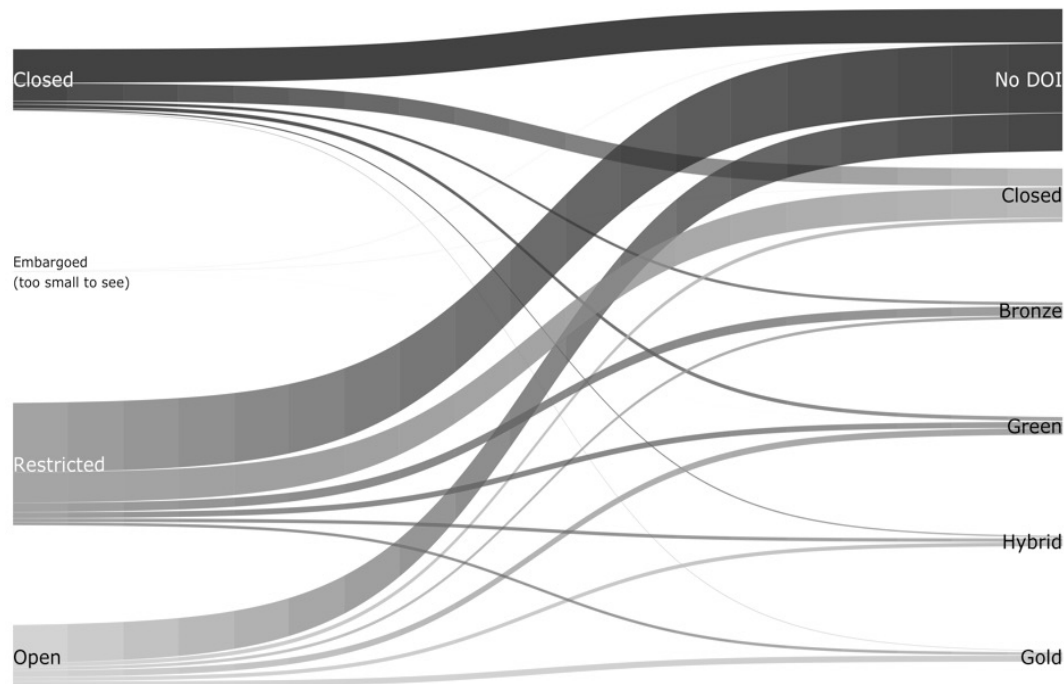


Fig. 1: Access categories for publications. Left axis: situation according to local repositories in the Netherlands (via NARCIS). Right axis: situation according to Unpaywall.

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 decisive¹¹.

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 research¹². 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 Science¹³ is trying to change that.

Ownership

There is a strong trend towards the position that research data should be published under an Open Access regime¹⁴, 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 unproductive¹⁵, 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 file 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 formats¹⁶. 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.

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

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 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 analytics¹⁸. 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 them¹⁹, 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 copyrighted²⁰. 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 lower²¹.

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 certified²²) 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 Commons²³ 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 tool²⁴, or by licensing it using one of six licences²⁵. 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. 2²⁶, 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 two 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.

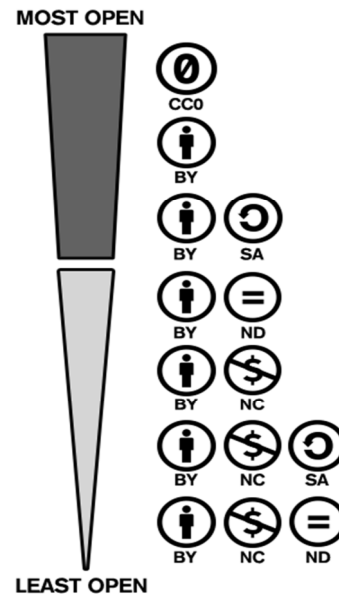


Fig. 2: Creative Commons diagram of CC0 and licences: from least to most open. Source: Creative Commons.

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 ETCBC²⁷. 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 SHEBANQ²⁸ 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 to 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 understandably 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 options place minimal restrictions on

³⁰ <https://www.portable-antiquities.nl/pan/#/public/about>

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 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](https://doi.org/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); Sikkil, 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](https://doi.org/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



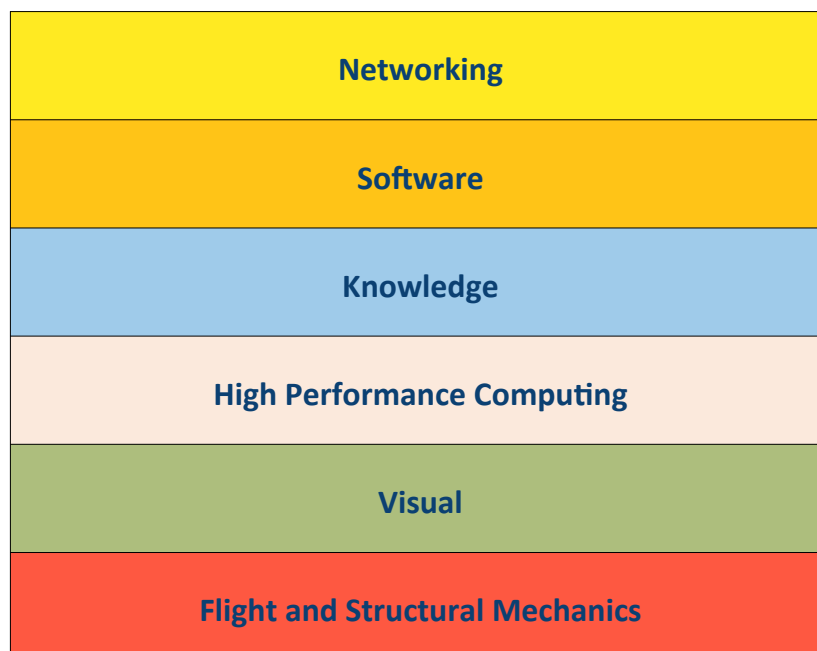
Institute of Information Science and Technologies “A. Faedo”

an Institute of the National Research Council of Italy CNR

ISTI is committed to produce scientific excellence and to play an active role in technology transfer.

The domain of competence covers Computer Science & Technologies and a wide range of applications.

The research and development activity of the Institute can be classified into 6 thematic areas



**CNR-ISTI, Via G. Moruzzi 1
56124 Pisa (PI), Italy
Area della Ricerca del CNR**

**Contact: +39 050 315 2403
segreteria scientifica@isti.cnr.it
<http://www.isti.cnr.it>**

The data librarian: myth, reality or utopia?

Silvia Giannini and Anna Molino,

Institute of Information Science and Technologies, ISTI-CNR, Italy

1. Introduction

It is undoubtedly true that we live in a more and more data-centric world (Cassella, 2016). As citizens and users of the Internet, we produce an enormous amount of data on a daily basis. This may be done consciously (e.g. when we deal with governmental and public organizations, universities and research centers, structured companies, etc.), as well as less intentionally, for instance when we use social networks, and, more generally, as users of the Internet for the widest variety of purposes.

In this scenario, data is becoming of greater importance not only for the average user when it helps making choices and decisions, but mainly for the growing number of companies and entities worldwide founding their activity in collecting and elaborating the exponentially increasing amount of information produced by the citizenry.

Therefore, we are now facing and coping with what has been defined by many commentators a “data deluge”.

In the academic and, more generally, scientific world the need for management, long term preservation, and storage of research data has grown exponentially in the recent past. As the Turing award winner Jim Gray states: “a fourth paradigm [of science] is emerging, consisting of techniques and technologies needed to perform data-intensive science” (Gray, Szalay, 2007). Indeed, in almost all discipline areas “born digital” documents proliferate as files, spreadsheets, databases, digital notebooks, wikis, etc. As a consequence, the management, curation, and archiving of such data are becoming of crucial importance (Bell, Hey and Szalay, 2009).

In this context, the expressions *eScience* and *eResearch*¹ have been identified as umbrella terms describing converging sets of trends and technologies that are radically changing the way science is conducted. Librarians may bring important knowledge and skills complementary to the activities carried on by the *eScience* community, most notably in the management, preservation and archiving of information (Wright et al., 2007). Moreover, another essential component of *eScience* concerns the management of the scholarly communication lifecycle, being this one of the most prominent area of interest in the work performed by libraries.

The Open Science (OS) movement, in turn, is radically changing the perspectives adopted in the scientific production and dissemination, fostering new approaches to research and scholarly communication. The movement is growing considerably in academia and among scientists worldwide. Two fundamental aspects of OS are the Open Access (OA) to scientific publication and the possibility of discovery, sharing and exploit the data used for or produced during the research process. The need for creating Open Data (OD) is profoundly changing the perspectives adopted by researchers during the scientific production, as research data is increasingly recognized as a primary research output.

As far as OD is concerned, firstly it must be pointed out that after the regulation of the Open Access to scientific literature, the legal framework in Europe is recently supporting the approach to OA.

Indeed, last European Recommendations (April 25 2018)² ask Member States to ensure that data management planning becomes a scientific practice and since 2017 the European Commission has made mandatory to open research data for all participants in Horizon 2020 and for any subject area, provided this is allowed from a legal or ethical point of view.

Thus, not surprisingly various funding bodies (e.g. the European Commission, the Wellcome Trust and the RCUK in UK, the Australian Research Council in Australia, or the National Institutes of Health in the U.S.) have been making mandatory the submission of a Data Management Plan (DMP) together with the project proposal. For instance, in the article 29.3

¹ *eScience* is the term preferred in Europe, while in other countries (e.g. Australia) the initiatives aiming at transforming the approach to science are labelled as *eResearch* (cfr. Wright et al., 2007).

² <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32018H0790&from=EN>

of the Annotated Model Grant Agreement³, the EU asks all consortia submitting a proposal in H2020 program to declare: the type of data that would be produced during the project; the strategies for their management in order to guarantee their short- and long-term preservation; how much of the produced data would be openly available.

In this perspective, academic libraries are indeed increasingly involved in the management of research data across the lifecycle (Schmidt and Shearer, 2016), actively participating in tasks such as providing access to data, supporting researchers in managing their data and drafting DMPs, as well as managing data collections.

Given this context, we may ask ourselves: who is currently responsible for the management, curation, and archiving of (research) data? Fearon et al. (2013) observes that:

“the data management space in US in higher education is predominantly owned by the libraries [...], whereas in the UK it is much more dependent on individual institutional cultures and circumstances whether it is the librarians, the academics, or the administrators who take the lead”.

The Research Data Alliance (RDA)⁴ recognizes that: “Many academic libraries are now extending their century-long track record in the professional management of knowledge resources towards the area of research data and therefore seek to maximize research data skills among staff in their organizations”. They identify five main routes to achieve such goal, consisting in: training, expert recruitment, learning-on-the-job, online-courses, and (academic) degrees.

Swan and Brown (2008) in their report commissioned by the UK Joint Information Systems Committee (JISC) recognize a strategic role for libraries in data management, identifying three main potential roles: increasing data awareness; providing archiving and preservation services; developing a new professional strand of practice as data librarianship.

Many commentators have argued that the background knowledge of librarians may be essential in this scenario. For instance, the management of repository’s contents may be seen as a *collection management issue* (Genoni, 2004), while the expertise in classification and description through cataloguing and metadata, as well as the experience in the selection of the information may be crucial for data curation (Witt, 2008).

Starting from this framework, we can consider academic libraries as “aggregator, collector and curator of external scholarship, be it printed or online”. For this reason, in our work we will try to get an idea of the competencies required to a Data Curator, giving an overview of the features of Research Data Management and its conversion into an effective service (RDS), both from a theoretical point of view and through the observation of some concrete examples of data management by librarians. The aim is to understand if librarians are the most accredited candidates to fill the role of Data Curator, giving possible answers and outlining specific qualifications required to those currently operating in academic libraries with the purpose of possibly identifying the figure of *data librarian*.

2. What is (research) data?

2.1. Data: definitions and types

As anticipated in the introduction, the contemporary world may be defined as a data-centric reality, due to the huge amount of data we use and produce in our everyday life.

But what is the meaning of the word *data*?

The broader definition of the term given by the Cambridge Dictionary⁵ is: “information collected for use”; more specifically, data can be seen as “information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer”.

The Oxford English Dictionary⁶ generally speaks about: “facts and statistics collected together for reference or analysis”, providing the more discipline-oriented sub-definition:

³ http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf.

⁴ Research Data Alliance – Libraries for Research Data IG (2015). *How to maximize research data skills in libraries*, <https://www.rd-alliance.org/how-maximize-research-data-skills-libraries.html>.

⁵ <https://dictionary.cambridge.org/dictionary/english/data>

⁶ <https://en.oxforddictionaries.com/definition/data>

“The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media”.

Merriam-Webster⁷ specifies three different connotations of the word, seeing data as: “factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation”; “information in digital form that can be transmitted or processed”; “information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful”.

It is noticeable that all three dictionaries cited take into account the shades of meaning concerning information science and technology, making reference to what is relevant for data-analysis as specific area of interest.

It is possible to identify different types of data, based on their exploitation, purpose, etc., being them identified and categorized in a more or less discipline as well as use oriented.

In our work, we focus on a specific typology of data, namely *research data*. The following subsection is dedicated to the description of this category, trying to outline a possible definition embracing the quantity of information produced in this specific context.

2.2. Research Data

The definition of research data may be quite broad and open to different meanings, which can vary depending on the disciplinary field we consider. Indeed, what a researcher considers to be *research data* depends on the meaning of this data in the research process and this may differ for each scientific discipline. Therefore, different definitions of this concept have been suggested by various actors and entities operating in this field. Here we report those we find more significant, proposed by the “Essentials 4 Data Support” online course of the Research Data Netherlands (RDNL)⁸.

The Queensland University of Technology in its *Manual of Policies and Procedures* defines research data⁹ as: “data in the form of facts, observations, images, computer program results, recordings, measurements or experiences on which an argument, theory, test or hypothesis, or other research output is based. It relates to data generated, collected, or used, during research projects, and in some cases, may include the research output itself. Data may be numerical, descriptive, visual or tactile. It may be raw, cleaned or processed, and may be held in any format or media [...]”.

The UK Engineering and Physical Sciences Research Council (EPSRC)¹⁰ consider research data as: “recorded factual material commonly retained by and accepted in the scientific community as necessary to validate research findings; although the majority of such data is created in digital format, all research data is included irrespective of the format in which it is created”.

Crossley (2004) in her “Introduction to managing research data”¹¹ argues that they are “collected, observed or created for the purpose of analysis to produce and validate original research results”.

Indeed, it must be taken into account that research data may be presented in a variety of formats, both digital or physical, e.g. electronic text documents; spreadsheets; laboratory notebooks, field notebooks, and diaries; audiotapes and videotapes; specimens, samples, and artefacts; methodologies, workflows, standard operating procedures and protocols; metadata, and so on (Scott & Cox, 2016).

Overall, the definition we prefer for its conciseness and effectiveness is the more general one, saying: “Research data is the material underpinning a research assertion”, making reference to all the outcomes produced in the course of the research, from statistics to field observations and answers to questionnaires, in spite of their formats or media.

2.3. Research Data as Open Data

In the context of Open Science, the possibility of making research data as open as possible becomes of crucial importance. Indeed, there are considerable advantages in sharing

⁷ <https://www.merriam-webster.com/dictionary/data>

⁸ <https://datasupport.researchdata.nl/en/start-the-course/i-definitions/research-data/>

⁹ http://www.mopp.qut.edu.au/D/D_02_08.jsp

¹⁰ <https://epsrc.ukri.org/about/standards/researchdata/scope/>

¹¹ <https://www.scribd.com/presentation/138079216/Managing-Research-Data>

materials supporting the research: sharing the research outcomes encourages the cooperation between scientific communities and favors a faster and more efficient research process, as it avoids useless data duplication and stipulates the collaboration between institutions and with the citizenry.

This may be accomplished following a series of practices and principles helping the scientific community in the correct production and reuse of the research results. For instance, the main goal of FORCE 11¹², a community of scientists, librarians, archivists, publishers and funders, is the promotion of the FAIR data principles. The acronym means *Findable, Accessible, Interoperable, Reusable* and it corresponds to a set of guidelines that enables a better realization and sharing of the data.

This brings to light the crucial issue concerning digital and, more specifically, data curation. The affirmation of digital products and services has in fact brought with it a set of strategies, technological approaches and activities that have taken the name of *Digital Curation*.

Digital Curation can be considered a transversal activity to various fields consisting in the creation, the maintenance and the preservation of a digital object throughout its lifecycle. The active management of research data reduces threats to their research value and mitigates the risk of digital obsolescence, enhancing the long-term value of existing data by making it available for further high quality research.

Digital curation and data preservation are ongoing processes, requiring considerable thought and the investment of adequate time and resources. This is the reason why the DCC¹³ in UK has identified some steps to be followed during what has been named *digital curation lifecycle*, as represented in the picture below:

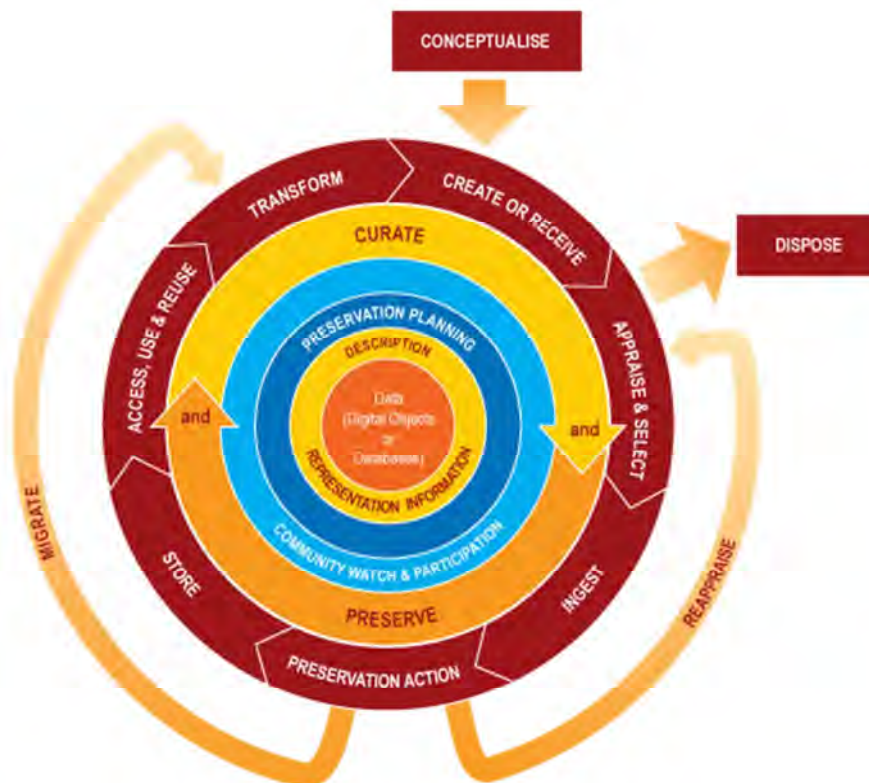


Figure 1: Digital Curation Lifecycle

The need to identify a professional figure who manage and store the growing amount of data in digital format has generated the role of *digital curator*. At the present time, this may not be considered a structured and well-defined character; as a consequence, the breadth of the definition implies that its reference community may include different actors and professional figures.

¹² <https://www.force11.org/group/fairgroup/fairprinciples>

¹³ <http://www.dcc.ac.uk/digital-curation/what-digital-curation>

The concept of Data Curation can be seen as a sort of subset of the Digital Curation. Strictly connected with the academic and research world, it comes from the connection of the digital curation with the development and management of Open Access repositories. At the same time, the data curator becomes a specialization of the digital curator.

The reference community of Data Curation is very often limited to the researchers and the type of data taken into account are the research data associated to the scientific literature.

Thus, the professional figures move towards the librarians or, more exactly, towards the upcoming figure of the *data librarian*, for their wide experience in different disciplinary domains, their skills in the management of metadata sets, in the maintenance of collections and their involvement in the management of research information.

The following paragraph is dedicated to Research Data Management, giving particular attention to its transformation into a specific service granted by some established research entities, namely Research Data Service.

3. Research Data Management and Research Data Service

3.1. What is Research Data Management (RDM)?

Research Data Management (RDM) is a general term to indicate a set of good practices concerning collecting, storing, using, sharing and preserving research data in an effective and productive manner. It involves services, tools and infrastructures that support the management of research data, which may significantly differ across the lifecycle. (Schmidt and Shearer, 2016; Schmidt et al., 2016).

Whyte and Tedds (2011) in their Briefing Paper for the DCC clarify some terminological distinctions between research data management, preservation and curation, arguing that: “Research data management concerns the organization of data, from its entry to the research cycle through to the dissemination and archiving of valuable results. It aims to ensure reliable verification of results, and permits new and innovative research built on existing information. Preservation is about ensuring that what is handed over to a repository or publisher remains fit for secondary use in the longer term (e.g. 10 years post-project). Curation connects first use to secondary use. It is about ensuring that project results are fit to archive, and that valued research assets remain fit for reuse”.

The various aspects of RDM should be seen as research support services distributed across various departments (e.g. Research Offices, IT Services, Libraries), as researchers need support in different areas, such as planning, organizing, documenting and sharing, preparing datasets for deposit and long-term preservation, not forgetting copyright issues (Schmidt and Shearer, 2016).

Therefore, RDM involves a wide range of activities across the data lifecycle, requiring a high level of interaction with both researchers and other support services (e.g. technical services and research officers), such as creating and collecting, processing, analyzing, publishing, archiving and preserving, and re-using data (Schmidt et al., 2016).

Figure 2 illustrates the major steps of the RDM lifecycle, highlighting the common points that might be shared by different scientific communities and realities worldwide:

The Research Data Management Lifecycle

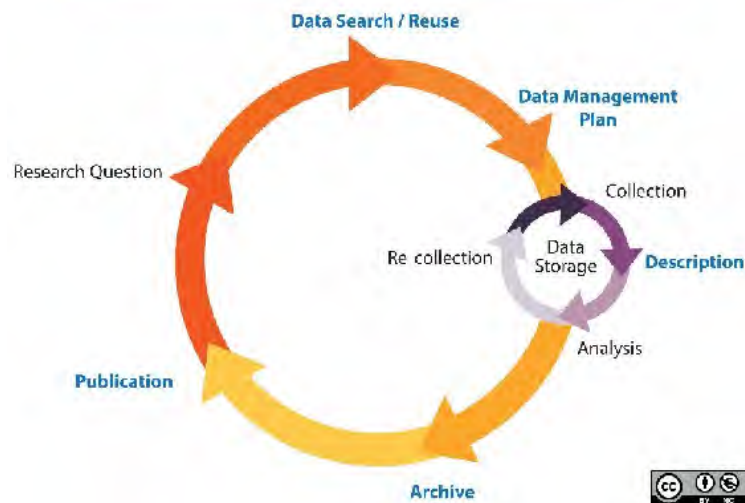


Figure 2: Research Data Management Lifecycle, diagram (University of California)

As for the Research Data Netherlands¹⁴, those operating in RDM should guarantee support within the following influence spheres:

- Legislation and policy;
- Technological infrastructure;
- Culture (i.e. research practices and their exploitation);
- Knowledge of storing, managing, archiving and sharing research data by both researchers and support services;
- Skills (e.g. conversational and influence skills);
- Motivation for collecting and managing research data.

3.2. Libraries' competencies in RDM

As anticipated above, the traditional competencies of librarians may be well-versed in RDM, due to the broad experience in wide-ranging disciplinary domains, the practice in the management of metadata sets, as well as in the creation and preservation of collections. More in general, we can recognize that librarians have always been familiar in dealing with data.

In their Final Report on Research Data Management (2012), the LIBER Working Group on eScience identified *Ten recommendations for libraries to get started with research data management*, underlining the crucial role of librarians in offering support in data management, in the development of metadata and data standards, their participation in the elaboration of institutional policies for data administration, as well as in assisting in the exploitation of interoperable infrastructures granting access and storage to research data (e.g. via the application of persistent identifiers), being possibly involved in subject specific RDM practices.

Schmidt et al. (2016) identifies three main groups into which library services for RDM can be broadly categorized: provide access to data; awareness and support to students and researchers in handling data; managing data collections. Despite the possible overlaps between them, authors also identified distinctive roles for librarians inside each area. Indeed, the first one reflects more traditional library services (e.g. consultation and reference for datasets); the second involves hands on support for researchers across the data lifecycle, e. g. in policy and advocacy on RDM and data sharing, as well as training. The third and final category includes the preparation of data for deposit, the management of metadata, and data preservation activities (Schmidt and Shearer, 2016).

Moreover, libraries have also the opportunity to act as point of contact with the public audience, supporting public engagement with science and acting as a hub to collate links and

¹⁴ <https://datasupport.researchdata.nl/en/start-the-course/vi-data-support/influence-sphere/>

information about citizen science activities, as well as researchers advocating the use of guidelines and templates. In addition to this, librarians are crucially involved in the training of scientists in data management and reuse (Lyon 2012).

This favors the evolvement of RDM practices into Research Data Services (RDS), which find numerous and different applications throughout academic and, more in general, research realities around the world.

3.3. Research Data Management as Research Data Service

The fundamental importance acquired by data collection and reuse in the research process, as well as the establishment of data management mandates by funding bodies have motivated research libraries to develop a set of services that may be generally labelled as Research Data Services (RDS).

Indeed, RDM acquires substance when it becomes a service, i.e. when an infrastructure made up of people and tools is established, providing assistance and advice for RDM practices supporting all the phases constituting a research project.

The results of the studies conducted by Tenopir et al. (2014) on U.S. and Canadian research libraries shows that the provision of RDS would augment institutions' research impact as well as the perception of the library in terms of relevance and prestige, RDS fitting the traditional role of librarians as "stewards of scholarship". At the time of their investigation, the most common services provided by academic libraries in U.S. and Canada could be seen as "extensions" of familiar reference services into the realm of data, dealing mostly with access to and citation of datasets (Tenopir et al., 2014).

We identified some common basic services provided nowadays by research libraries participating in RDS; generally speaking, they are primarily involved in:

- Support in Data Management Plan (DMP) development;
- Digital Curation, i.e. data selection, preservation, maintenance, and archiving;
- Metadata creation and transformation.

Another fundamental aspect that must be taken into account when speaking about established RDS in universities and research bodies worldwide, is the presence of a formal RDM Policy, being this at institutional level (as in the case of many universities in UK, e.g. University of Edinburgh) or presented as codes of conduct at national level, as in the case of the *Australian Code for the Responsible Conduct of Research*, which covers a wide range of topics associated with research including the management of research data and associated materials¹⁵. The availability of such policies endorses the formal establishment of the roles involved in RDS, regulating the smooth functioning of the research support in each institution and contributing to the definition of upcoming figures such as the *data librarian*.

Finally, we can argue that RDS may be reasonably conceived as the conversion of RDM concepts into concrete practices. In order to validate such assertion, we reviewed some literature about RDS experiences worldwide and selected three case studies that we propose in the following paragraph. This helped us to understand how this process turns into reality.

4. Case Studies

4.1. University of Edinburgh Research Data Service (RDS)

The first case study we analyzed is the Research Data Service provided by the University of Edinburgh¹⁶. The aim of the service is to provide tailored support to researchers dealing with the production and/or reuse of research data, offering tools, support and training to university staff and students.

The Research Data Service, led by a *data librarian* whose background is in library services, is part of the Information Services of the University, whose experts contribute in delivering specific tools and software components for the management of the data produced during the research lifecycle. The main goal is to provide tailored services for researchers aiming to achieve good practices in RDM, according to their specific needs. Moreover, at any time of

¹⁵ <https://www.ands.org.au/guides/code-awareness>

¹⁶ <https://www.ed.ac.uk/information-services/research-support/research-data-service/about-the-research-data-service>

such process people working with research data may ask support for training, which will be delivered following specific programs.

It must be pointed out that the University of Edinburgh has a formal Policy for RDM, establishing that data must be “managed to the highest standards as part of the University’s commitment to research excellence”, granting that data will be made available as open as possible, protecting those considered as sensitive and giving the widest outreach to those that may be of public interest. Indeed, the last document’s clause clearly states that: “Exclusive rights to reuse or publish research data should not be handed over to commercial publishers or agents without retaining the rights to make the data openly available for re-use, unless this is a condition of funding”¹⁷.

The support offered may be divided into three major phases: planning, active research project phase, project conclusion. These represent the milestones of the project lifecycle and, subsequently, of the assistance provided by RDS.

Following this pattern, the RDM services delivered at the University of Edinburgh may be schematized as follows:

- 1) *Before*: this consists of the identification of existing datasets; the planning for the data collection and storage; the identification of possible sensitive data, as well as methods and terms for data sharing. We can reasonably argue that the crucial part of this step regards the creation of a Data Management Plan (DMP), required by the majority of funding bodies and universities, which accompanies the research project during its lifetime. The RDS at the University of Edinburgh assist researchers in its development, providing either tools and templates (e.g. DMPonline¹⁸) or personal consultation to discuss the DMP in detail and obtain expert advice.
- 2) *During*: in the active development of the project, consultation is offered about finding existing datasets containing data that might be reused and re-elaborated, some being freely available, others behind paywall. The Data Library here plays a crucial role, giving advices about possibility for exploitation by users, as well as helping researchers in the selection of the data resources based on their type (e.g. surveys, censuses, databases, etc.). Other contributions offered by the RDS regard solutions for data storage during the project lifetime, for the control and safeguard of sensitive data, for the sharing and versioning of data, keeping track of the changes while working with other researchers or research teams.
- 3) *After*: after the conclusion of the research project, RDS grant assistance in recording, sharing and archiving research data for the long-term. This is made via a set of specific tools recording descriptive metadata, providing storage in an open repository for the online discovery and re-use through the association of a persistent identifier (DOI) to researchers’ data resources, and securing long-term archiving in order to keep data safe from accidental deletion or inappropriate access, meeting possible funders’ requirements.

Besides RDM consultancy and support, particular attention is given to training. The RDS at The University of Edinburgh offer a wide range of courses for those unfamiliar with the fundamentals of research data management and sharing, in the form of online courses, classroom-based workshops and seminars.

Indeed, people dealing with research data have the possibility to select a suitable training option among the variety proposed, depending on their specific necessities.

For instance, a free five-week MOOC - created by the Universities of Edinburgh and North Carolina – has been designed to reach learners of various types across disciplines and continents. The subjects covered are: understanding research data; data management planning; working with data; sharing data; archiving data, following the stages of a generic research project.

¹⁷ <https://www.ed.ac.uk/information-services/about/policies-and-regulations/research-data-policy>

¹⁸ <https://dmponline.dcc.ac.uk>

In addition, a free, online course named MANTRA¹⁹ has been realized with the purpose of understanding and reflecting on the management of the data collected throughout the research. This is particularly referred to post-graduate students, early career researchers, and also information professionals. It is composed of a series of interactive online units concentrating on the explanation of the terminology, key concepts, and best practices in RDM.

Inside this training path, a special focus is dedicated to librarians, underlining the central role of such figure in the RDM workflow. Indeed, a *Do-It-Yourself Research Data Management Training Kit for Librarians*²⁰ has been created in order to supply to the needs of academic liaison librarians. It is provided by EDINA and Data Library, University of Edinburgh, in association with the UK Data Archive, Digital Curation Centre (DCC), and Distributed Data Curation Center at the Purdue University Libraries. After an introductory “pre-training”, the course is divided into five main sections, each containing a wide range of materials (e.g. podcasts, presentations, assignments, etc.) and specifically concentrated on: data management planning; organizing and documenting data; data storage and security; ethics and copyright; data sharing. Moreover, materials for post-training study are available, such as the *Data Curation Profiles*, which provide a complete framework for interviewing a researcher in any discipline about their research data and their data management practices, giving practical overviews on what a librarian involved in RDM would face in the daily practice.

Finally, also bespoke group training and face-to-face classes and workshops are planned upon requests or periodically.

4.2. The Research Data Netherlands Front Office – Back Office model

The second case study we have taken into consideration regards the Front Office – Back Office Model of the Research Data Netherlands (RDNL FO-BO Model).

It consists of a federated infrastructure handled by DANS, 3TU.Datacentrum and SURFsara - the three organizations constituting RDNL, a coalition joining three archives in the area of long-term archiving, also open to other third parties - and modelled into a four-layer structure:

- 1) a basic technical infrastructure under the computer centers responsibility;
- 2) back office data services, providing facilities for long-term archiving and accessibility;
- 3) front office services, granting support and training to researchers and students in responsible data management;
- 4) data generators and data users.

The model is graphically represented in figure 3 below:

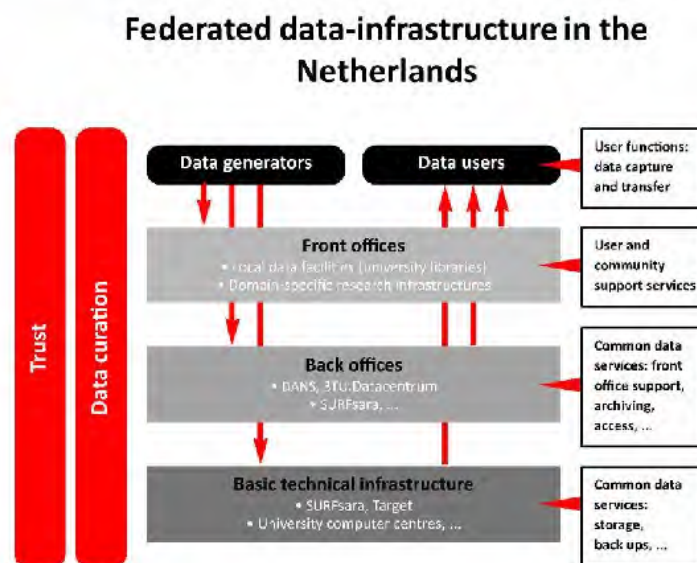


Figure 3: The federated data-infrastructure in the Netherlands

¹⁹ <https://mantra.edina.ac.uk/>

²⁰ <https://mantra.edina.ac.uk/libtraining.html>

RDNL infrastructure moves from the model proposed in the EU *Riding the wave* report (2010), giving a general impression of how the various actors, data types and services should be interconnected in a global e-infrastructure for science. It must be also noticed that Netherlands have recently published the *Netherlands Code of Conduct for Research Integrity* (2018), which delineate specific directions for data management (pp. 20-21).

Globally, the services provided fall into three main categories: information provision, training, and data curation, management and storage.

More in details, the front-offices are situated mainly locally, in most of the cases in research libraries; their focus is on supporting their own research organizations, being primarily responsible for the quality assurance of the data produced.

The front-office is accountable for data collection and acquisition, as well as awareness raising in best practices for data management within its research community of reference, providing information and training to their research personnel. Moreover, the front office hosts so-called Virtual Research Environments or Data Labs, offering research tools and securing mid-term storage facilities for the organization's researchers. In consultation with the back office, the front office also facilitates the transfer of data to a trusted back-office digital repository after the research has been completed. Facilities that are shared by several universities can be hosted and supported by the back offices.

The back-offices are constituted mainly of computer experts, who provide data stewardship, guaranteeing long-term storage and accessibility to the collected data. Their functions are performed mainly by organizations such as DANS, 3TU.Datacentrum, and SURFsara, having a nationwide coverage and expertise on data from various discipline areas (humanities, social studies and sciences).

The back-office also provides consultation, training and support to front-office employees, acting as a center of expertise and innovation. Its fundamental duty is to ensure a sustainable and secure storage and retrieval upon completion of the research project.

It must be pointed out that since the division of labor between front-office and back-office is not always necessarily sharp, especially because organizations may differ in size, staffing capacity, etc. there might be institutes performing front-office tasks only, while outsourcing the back-office ones to a data archive. In this perspective, we can see how the figure of the librarian in this kind of RDM model is always present and plays a relevant role.

4.3. eResearch at Griffith University

The third and final case study we analyzed regards the activities conducted by the eResearch unit at Griffith University (Australia), where the Division of Information Services (INS) integrates what they have named e-research, library, and information and communication technology into a single organization (Brown et al., 2015).

As in the previous case studies, Griffith University responds to the *Australian Code for the Responsible Conduct of Research* (2007) - assigning researchers and their institutions a shared responsibility to manage research data and primary materials well – and to the *Griffith University Code for the Responsible Conduct of Research* (2012). As stated in this policy, researchers are required to manage their data – using methods appropriate to the discipline and to the nature of the data – to the highest standards, while the University is required to provide infrastructures, opportunities to develop professional skills, and access to advice and expertise that enable researchers to meet these standards. Finally, the *Best practice guidelines for researchers: Managing research data and primary materials* (Richardson, 2016) were published, aiming at expanding the Code for the Responsible Conduct of Research in relation to specific aspects of research data management; outlining practical steps that researchers can take; highlighting technology, advisory services and professional enabling development opportunities.

In this context, the first thing to point out is that faculty librarian roles have been modified to address data management. Indeed, while traditional librarian's duties are kept, and their core capabilities (e.g. structured thinking, knowledge of information management theory, etc.) are considered of great value for positioning them in the process of research data management (Brown et al., 2015), the librarians operating in the eResearch team are required to develop additional, more technical and discipline-oriented skills. More in detail, on the one hand librarians within the INS portfolios of Library and Learning Services and

Information Management support researchers in well-established areas such as acquisitions, collection development, copyright advice and information literacy training, and are moving into newer areas such as open access advocacy, publications repositories, research assessment exercises, and bibliometrics. On the other, Griffith's eResearch Services team operates within another portfolio in INS, building and managing technical research infrastructure for supporting researchers. Part of this infrastructure is targeted at specific research needs, while some is associated with university-wide management and discovery. Therefore, librarians in eResearch must explicitly demonstrate how their skills can be combined in productive ways with technical specialties, including software development and business analysis (Simons and Searle, 2014), as they work in close relationship with ICT experts and, for this reason, need to undergo specific and constant training to develop the specific and necessary skills to supply RDM demands.

Librarians working in this unit tend to consider themselves as "generalists", as they are required to have such a broad range of skills, knowledge and expertise that is difficult to acquire a specialization in any of these (Simons and Searle, 2014).

However, as anticipated above, a core set of skills and knowledge for librarians have been identified. As far as skills are concerned, they can be considered a sort of enhancement of the traditional librarian's competences. In fact, they deal with advanced metadata skills, high level communication skills, as in such context the role of the librarian foresees the "translation" of information between research groups. In addition, high level documentation skills are necessary for the production of documentation addressing a wide variety of audiences and purposes.

With respect to core knowledges, there must be a deep knowledge of the broader research environment where the librarian is acting, as well as of the mechanisms and processes of scholarly communication, and of the legal and regulatory framework, concerning mainly contract law and copyright issues, with a specific focus on licensing and data re-use.

Generic technical and managerial skills also play a distinctive role, as research teams usually work on goal-oriented projects, and since in eResearch project teams are comprised largely of software developers.

As far as we understand, the librarian acts as an advisor, even though technical skills make the difference in understanding how eResearch projects are run and in liaising with researchers and research managers.

5. The *data librarian* profile

In the light of the case studies, we tried to identify some basic characteristics of the *data librarian* profile.

Schmidt et al. (2016) briefly describe this upcoming figure in the research scenario as consisting of "Traditional librarian competences and skills into renewed organizational structures". Authors like Sada et al. (2013) highlight necessary technical skills, underlining the fact that many of the competencies of such professional are adopted from the ICT domain. In this perspective, the *data librarian* is seen mainly as responsible for the implementation of collaborative infrastructures for data access and reuse, fostering protocols for data interoperability and dedicating special attention to digital preservation. Indeed, as argued by Cassella (2016) and anticipated previously, just a few steps separate the *digital curator* from the *data librarian*, the latest being a RDM specialist who constantly collaborates with other professionals.

In her presentation for the 2nd DCC/RIN Research Data Management forum, Rice (2008) defines the *data librarians* as "people originating from the library community, trained and specializing in the curation, preservation and archiving of data". As a follow-up of the same event, the diagram reported below has been published in the DCC Data Management Forum²¹:

²¹ <http://data-forum.blogspot.com/2008/12/rdmf2-core-skills-diagram.html>

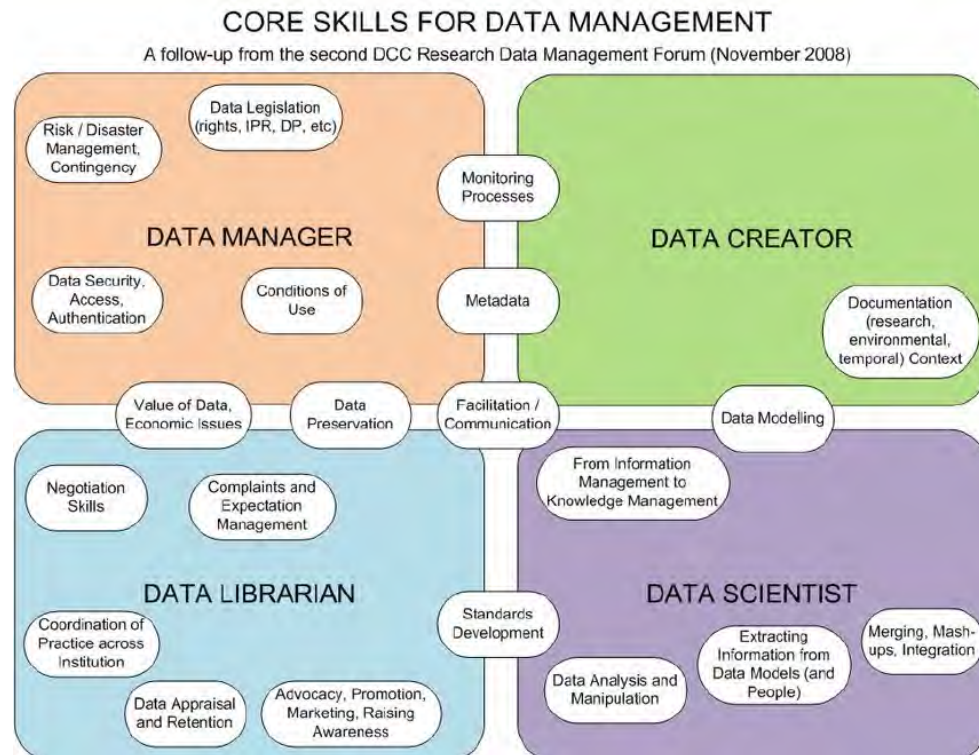


Figure 4: Core skills for data management. Chris Rusbridge and Martin Donnelly, November 2008.

The picture schematically describes the fundamental characteristics and tasks performed by the professional figures operating with data. Even though making a clear distinction between the four, a number of tasks overlap between them, as in the case of the preservation of data or the development of standards. It emerges from this model that at the time the major characteristics of the *data librarian* consisted in:

- Skills in communication and facilitation
- Standards development
- Data selection and evaluation
- Negotiation skills
- Advocacy, promotion and marketing
- Economic issues related to data value
- Preservation
- Complaints and expectations management

Despite its clarity and effectiveness, this model has undergone some critics regarding, for instance, the singular choice of placing the specialization in metadata at the boundary between the roles of data manager and data creator. Another remark concerns the absence of any reference to training for all data professionals and for *data librarians* in particular (Cassella, 2016), which, as we reported in the previous paragraph, is now considered as crucial, thus representing a fundamental aspect of any RDS.

In this perspective, Carlson et al. (2011) identified a list of twelve core educational objectives for a data information literacy program, which might be of interest also for librarians approaching RDM:

- | | |
|--|--------------------------------------|
| • Databases and Data Formats | • Data Curation and Re-use |
| • Discovery and Acquisition of Data | • Cultures of Practice |
| • Data Management and Organization | • Data Preservation |
| • Data Conversion and Interoperability | • Data Analysis |
| • Quality Assurance | • Data Visualization |
| • Metadata | • Ethics, including citation of data |

In the following years, the debate as well as the reality concerning professional roles operating with data has had various developments, as we illustrated in the case studies. Indeed, RDM policies and practices have evolved through time, allowing the growth of RDS in many academic realities.

In such scenario, Schmidt et al. (2016) identified core competencies for *data librarians*. Besides having a basic understanding of the specific disciplinary landscape, as well as being aware of norms and standards, they would be in charge of:

- 1) Provide access to data
- 2) Advocacy and support for managing data (e.g. knowledge of DMP templates and tools, data sharing options, licenses, data citation and reference practices, etc.)
- 3) Manage data collections

More in details, librarians are required to have a good knowledge of existing data centers, repositories and data discovery mechanisms, funders' policies and publication requirements of journals, metadata standards and schemas, data formats, domain ontologies, discovery tools, and so on.

In addition to this, *data librarians* would cooperate in related services such as collections' development and curation, assistance in OA and copyright policies, information literacy, digital curation and preservation, etc.

Cassella (2016) identifies five major areas of reference for the role of the *data librarian*:

- 1) Library science
- 2) Scholarly communication
- 3) Technology
- 4) Disciplinary law, copyright and licenses
- 5) Communication and management

Another fundamental aspect emerged from the analysis of the case studies is the importance of domain-specific competencies. Indeed, the knowledge of the research mechanisms of specific scientific areas would represent an advantage, as the *data librarian* always operates in teams, being part of a staff composed by different researchers and technical figures.

In this respect, Brown et al. (2015) recognize the increasing complexity of the roles of the librarians supporting what they have named eResearch, underlining again that it is in the network of specialists they closely work with that *data librarians* acquire the domain specific competencies differentiating them from other library professionals.

Thus, *data librarians* would present advanced skills in those areas where their colleagues not operating in eResearch units have general or basic knowledges.

More in details:

- Advanced understanding of discipline-based research process, outputs and scholarly communication (e.g. data types and formats)
- Advanced knowledge of ethics, intellectual property, copyright and licensing
- Advanced knowledge of discipline-specific metadata schemas and related standards (item and collection level)
- Knowledge of repository certification schemes and standards
- Knowledge of semantic web standards
- High level communication and documentation skills, project management and business analysis skills

In these regards, we are facing a composite reality, being the role of the *data librarian* one of the most complex to define and identify in such composite scenario, as it emerges firstly from the case studies we observed.

We may reasonably argue that at the present time there is still no overall agreement on the competencies and, most of all, on the tasks that a *data librarian* operating in RDS should perform, this being reflected also in the terminology currently in use to describe and identify such role. As a matter of fact, the *data librarian* may identify a wide range of different context-related professionals. However, most of the commentators agree on stating that *data librarians* usually work in team with other specialists and for this reason having or

acquiring some basic domain knowledge would be an advantage. Moreover, the traditional librarians' ability to count on both existing capabilities and newly acquired skills favors their establishment as core members of a research support team.

6. Conclusions

The emergence of e-science and e-research has opened new paths and trends in scholarly communication and management. In the academic environment, the need for opening research products to a wider audience has become increasingly urgent. For instance, many funder bodies now request Open Access to scientific publications and require the presentation of a Data Management Plan along with the project proposal.

In addition, it has been widely recognized that data sharing would bring major benefits to the scientific community, avoiding useless duplications and saving the researchers' time and resources.

In order to pursue such aim, an efficient management of research data has become essential. This is the reason why in the recent years an increasing number of institutions has been adopting specific policies dedicated to the effective management of the data produced during the research process, leading to the creation and clear definition of dedicated services, namely Research Data Services.

In this scenario, perspectives and concrete realities are quite different, as we observed in the case studies, even though some common aspects may be identified, as the importance of data access, curation, preservation, and, last but not least, the fundamental importance of training either for the professionals operating in this field, or for researchers and students producing and collecting data in their daily work.

In such context, in the recent years the figure of the *data librarian* is acquiring importance, as it may represent one of the possible evolutions of the traditional librarian in the contemporary academic world. Many definitions describing this role are available in the literature, making quite difficult outlining unique skills, knowledge, competences, and tasks. However, it is quite clear that it is a role that would not develop based on the classic skills of the librarian only, although these represent an extreme value and a concrete base for the development of a constantly evolving career. Indeed, it must be pointed out that, due to the recent advancements in science, technology and scholarly communication, almost all professionals working in the research field had to reshape their attitude towards the research processes and scholarly communication. In such prospect, librarians occupy a privileged position either for their conventional background, or their successful adaptation to the transformations and evolutions their profession has undergone over the years.

Finally, we can argue that the traditional competences of the librarian should not be idealized, assuming that they are sufficient for becoming a *data librarian*. On the other hand, the *data librarian* should not be seen as a utopia, since numerous academic experiences show how this role is becoming a concrete reality for many professionals around the world.

As a conclusion, we may say that the *data librarian* is neither myth, nor utopia, but a composite reality.

Bibliography¹

1. AGA – Annotated Model Grant Agreement, version 5.1, 6 December 2018. *Horizon 2020 Programme, European Commission*, http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf.
2. Australian Government (2007). *Australian Code for the Responsible Conduct of Research*, <http://www.nhmrc.gov.au/index.htm>.
3. Australian National Data Service. *Research data policy and the Australian Code for the Responsible Conduct of Research*, <https://www.ands.org.au/guides/code-awareness>.
4. Ball A. (2012). *Review of Data Management Lifecycle Models* (version 1.0). REDm-MED Project Document redm1rep120110ab10. Bath, UK: University of Bath.
5. Bell G., Hey T., Szalay A. (2009). *Beyond the Data Deluge*. *Science*, 323 (5919) 1297-1298; <http://science.sciencemag.org/content/323/5919/1297>.
6. Brown R.A., Wolski M., Richardson J. (2015). *Developing new skills for research support librarians*. *The Australian Library Journal*, 64 (3), 224-234.
7. Carlson J., Fosmire M., Miller C., Sapp Nelson M.R. (2011). *Determining data information literacy needs: a study of students and research faculty*. *Portal: Libraries and the Academy*, 11 (2), 629-657.
8. Cassella M. (2016). *Dal digital curator al data librarian*. *Biblioteche oggi*, 34 (4) 13-21.

9. Commission Recommendation (EU) 2018/790 of 25 April 2018 on access to and preservation of scientific information. Official Journal of the European Union, <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32018H0790&from=EN>.
10. Crossley J. (2014). *An introduction to managing research data*, <https://www.scribd.com/presentation/138079216/Managing-Research-Data>.
11. Digital Curation Centre. *What is digital curation?* <http://www.dcc.ac.uk/digital-curation/what-digital-curation>.
12. Digital Curation Centre. *DMPonline*, <https://dmponline.dcc.ac.uk>.
13. European Union (2010). *Riding the wave. How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data. A submission to the European Commission*, <https://www.fosteropenscience.eu/content/riding-wave-how-europe-can-gain-rising-tide-scientific-data>.
14. Fearon D. Jr., Gunia B., Pralle B.E., Lake S., Sallans A.S. (2013). *Research Data Management Services*, SPEC Kit n. 334.
15. FORCE11, *The FAIR data principles*, <https://www.force11.org/group/fairgroup/fairprinciples>.
16. Genoni P. (2004). *Content in institutional repositories: a collection management issue*. Library Management, 25 (6/7), 300-306.
17. Gray J., Szalay A. (2007). *eScience – A Transformed Scientific Method*. Presentation to the Computer Science and Technology Board of the National Research Council, Mountain View, CA, 11 January 2007; http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt.
18. Griffith University (2012). *Griffith University Code for the Responsible Conduct of Research*, <https://policies.griffith.edu.au/>.
19. KNAW; NFW; NWO; TO2-federatie; Vereniging Hogescholen; VSNU (2018): *Nederlandse gedragscode wetenschappelijke integriteit*. DANS, <https://doi.org/10.17026/dans-2cj-nvwu>.
20. LIBER (2012). *Ten recommendations for libraries to get started with research data management. Final Report of the LIBER Working Group on E-Science/Research Data Management*, <https://www.fosteropenscience.eu/content/ten-recommendations-libraries-get-started-research-data-management>.
21. Lyon L. (2012). *The Informatics Transform: Re-Engineering Libraries for the Data Decade*. The International Journal of Digital Curation, 7 (1), 126-138.
22. McMillan D. (2014). *Data Sharing and Discovery: What Librarians Need to Know*. The Journal of Academic Librarianship, 40, 541-549.
23. OECD (2007). *Principles and Guidelines for Access to Research Data from Public Funding*, <http://www.oecd.org/sti/inno/38500813.pdf>.
24. Osswald A., Strathmann S. (2012). *The Role of Libraries in Curation and Preservation of Research Data in Germany: Findings of a survey*, <https://www.ifla.org/past-wlic/2012/116-osswald-en.pdf>.
25. Perrier L., Blondal E., MacDonald H. (2018). *Exploring the experiences of academic libraries with research data management: A meta-ethnographic analysis of qualitative studies*. Library and Information Science Research, 40 (3-4), 173-183.
26. Queensland University of Technology (2015). *Manual of Policies and Procedures. D/2.8 Management of research data*, http://www.mopp.qut.edu.au/D/D_02_08.jsp.
27. Research Data Alliance – Libraries for Research Data IG (2015). *How to maximize research data skills in libraries*, <https://www.rd-alliance.org/how-maximize-research-data-skills-libraries.html>.
28. Research Data Netherlands. *Essentials 4 Data Support*, <https://datasupport.researchdata.nl/en/start-the-course/i-definitions/research-data/>.
29. Research Data Netherlands (2018). *A federated data infrastructure for the Netherlands: the front-office – back-office model. UK web version*, https://researchdata.nl/fileadmin/content/RDNL_algemeen/Documenten/RDNL_FOBOModel-UK-web.pdf.
30. Rice R. (2008). *Roles and Responsibilities for Data Curation: the Data librarian*. RDMF2: Roles and Responsibilities for Effective Data Management, Manchester, Chancellors Hotel and Conference Centre, 26-27 November 2008. Slides, <http://www.dcc.ac.uk/events/research-data-management-forum/roles-and-responsibilities>.
31. Richardson J. (2016). *Best practice guidelines for researchers: Managing research data and primary materials*. Griffith University, <https://www.griffith.edu.au/library/research-publishing/best-practice-guidelines-for-researchers>.
32. Sada E., Gregori L., Sirito P. (2013). *Un bersaglio mobile: l'evoluzione dei profili degli "information professionals" alla luce dei nuovi scenari accademici*. AIB studi, 53 (1), 92-99.
33. Schmidt B., Calarco P., Kuchma I., Shearer K. (2016). *Time to Adopt: Librarians' New Skills and Competency Profiles. Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Proceedings of the 20th International Conference on Electronic Publishing, Loizides F., Schmidt B. (eds.), IOS Press, 1-8.
34. Schmidt B., Shearer K. (2016). *Librarians' Competencies Profile for Research Data Management*. Joint Task Force on Librarians' Competencies in Support of E-Research and Scholarly Communication, <https://www.coar-repositories.org/activities/support-and-training/task-force-competencies/>.
35. Scott M., Cox S. (eds) (2016). *Introducing Research Data*. University of Southampton, fourth edition, https://eprints.soton.ac.uk/403440/1/introducing_research_data.pdf.
36. Simons N., Searle S. (2014). *Redefining 'The Librarian' in the Context of Emerging eResearch Services*. Paper presented at VALA 2014, <http://www.vala.org.au/vala2014-proceedings/vala2014-session-15-simons>.
37. Swan A., Brown S. (2008). *The Skills, Role and Career Structure of Data Scientists and Curators: An Assessment of Current Practice and Future Needs*. Truro: Key Perspectives. <http://www.jisc.ac.uk/publications/documents/dataskillscareersfinalreport.aspx>.

38. Tammaro A.M. (2016). *Libraries in Digital Age (LIDA): la trasformazione delle biblioteche in era digitale*, <https://annamariatammaro.com/2016/06/21/libraries-in-digital-age-lida-la-trasformazione-delle-biblioteche-in-era-digitale/>.
39. Tammaro A.M. (2018). *IFLA Global Vision Project: Barcelona Kick-Off Workshop*, <https://annamariatammaro.com/category/ifla-global-vision/>.
40. Tammaro A.M. (2018). *L'evoluzione della biblioteca digitale: dall'accesso alle risorse elettroniche alla creazione e alla cura dei dati*, <https://annamariatammaro.com/2018/05/09/evoluzione-della-biblioteca-digitale-dallaccesso-alle-risorse-elettroniche-alla-creazione-e-alla-cura-dei-dati/>.
41. Tenopir C., Sandusky R.J., Allard S., Birch B. (2014). *Research data management services in academic research libraries and perceptions of librarians*. *Library & Information Science Research*, 36 (2), 84-90.
42. UK Engineering and Physical Sciences Research Council. *EPSRC policy framework on research data – Scope and benefits*, <https://epsrc.ukri.org/about/standards/researchdata/scope/>.
43. University of Edinburgh *Research Data Service*: <https://www.ed.ac.uk/information-services/research-support/research-data-service/about-the-research-data-service>.
44. University of Edinburgh. *Research Data Management Policy*, <https://www.ed.ac.uk/information-services/about/policies-and-regulations/research-data-policy>.
45. University of Edinburgh – EDINA. *MANTRA. Research Data Management Training*, <https://mantra.edina.ac.uk>.
46. University of Edinburgh – EDINA. *Do-It-Yourself Research Data Management Training Kit for Librarians*, <https://mantra.edina.ac.uk/libtraining.html>.
47. Whyte, A., Tedds, J. (2011). *Making the Case for Research Data Management*. DCC Briefing Papers, Edinburgh: Digital Curation Centre, <http://www.dcc.ac.uk/sites/default/files/documents/publications/Making%20the%20case.pdf>.
48. Witt M. (2008). *Institutional repositories and research data curation in a distributed environment*. *Library Trends*, 57 (2), 191-201.
49. Wright M., Sumner T., Moore R., Koch T. (2007). *Connecting digital libraries to eScience: the future of scientific scholarship*. *International Journal on Digital Libraries*, 7 (1-2), 1-4.
50. <https://dictionary.cambridge.org/dictionary/english/data>.
51. <https://en.oxforddictionaries.com/definition/data>.
52. <https://www.merriam-webster.com/dictionary/data>.

¹ URL last access: December 2018.

Research Data Management: What can librarians really help?

Yuan Li, Willow Dressel and Denise Hersey

Princeton University Library, United States

Abstract

As national government agencies continue to mandate specific data management requirements and the need for research data management (RDM) grows, many libraries are developing RDM services to help with the research mission of their institution. Research libraries' mission and expertise have always included a variety of research services. What can the library's role be in RDM services? This paper describes the possible roles that libraries and librarians can play throughout the data lifecycle. The Princeton University Library is presented as a case study to demonstrate these roles and the development of RDM services including advocacy, awareness, education, advisory services, data management plan development, and data repository development and promotion. In addition, this paper also discusses the challenges and opportunities associated with RDM services development in libraries and the future plans at Princeton, including the development of a RDM mini course for graduate students and a robust RDM program.

Introduction

In 2013 the United States Office of Science and Technology Policy released a memorandum directing federal agencies with more than \$100 million in research and development funding to develop plans to, among other things, make scientific research data from unclassified research publicly accessible (Holdren). At the same time 74% of libraries surveyed by the Association of Research Libraries indicated they were already engaged in research data management services (Fearon et al.). Since the 2013 memo, both government and journal data management requirements have increased and librarians have continued to use their expertise in information organization, management, and preservation to help meet this growing need.

This paper will present a case study on data management services developed and delivered by librarians at Princeton University, including providing services for the Department of Energy's Princeton Plasma Physics Laboratory. Service development was started by a single science librarian but eventually grew to include the scholarly communications librarian and relevant science subject librarians. Services include data management plan consultation, data management education, and helping connect researchers to the institutional data repository. Services have been developed and delivered using existing staff. It is expected that dedicated staff will be hired in the future to fully develop the program. Staffing costs vary based on the individual institution and location. The paper will highlight outcomes of selected collaborations, discuss feedback on services, and provide a case study for peer institutions.

Literature Review

There are two notable large scale surveys of research data management services in libraries. In 2012 Carol Tenopir et al. wrote a white paper for ACRL titled "Academic Libraries and Research Data Services". The following year David Fearon et al published an Association of Research Libraries (ARL) Spec Kit titled "Research Data Management Services". Tenopir's survey included 221 academic libraries from the United States and Canada. Key findings of the survey showed that at the time less than a third of academic libraries were engaged in research data services, but many more planned to develop services in the future. The most common service presently offered or planned was creation of web guides to help researchers find data sets. Staffing of research data services was most commonly approached by re-assigning existing staff (Tenopir et al.). The ARL Spec Kit surveyed 73 of 125 ARL libraries. Emerging trends identified in this study were research data management support at the grant stage for data management plan requirements and data archiving support. One area the authors regretted as a limitation of their study was the lack of "new, comprehensive case studies" (Fearon et al. 12).

Several case studies from the library perspective were included in *Research Data Management: Practical Strategies for Information Professionals* edited by Joyce M. Ray in 2014. These case studies feature overviews of the development of infrastructure and services at four institutions with fairly well established programs: Cornell University, Purdue University, Rice University, and the University of Oregon. The Cornell University Library (CUL) case study examines their long history of supporting research data needs at Cornell starting in 1982 by collaborating with the Cornell Institute for Social and Economic Research data archive, then providing infrastructure for the other disciplinary repositories in the 1990s, managing the “eCommons” institutional repository launched in 2002, and finally developing the Datastar staging repository for research data. In addition to infrastructure, CUL has also made efforts to develop other research data management services for faculty starting with a Data Working Group from 2006-2008, then participating in the ARL E-science Institute in 2010-2011, and subsequently launching the cross-departmental Research Data Management Service Group. Looking to the future Cornell hopes to improve existing infrastructure and provide more support while refining the scope for data services (Steinhart). Purdue University Library began researching research data management needs of scientists at Purdue in 2004. Their research led them to create the Data Curation Profiles Toolkit to aid in understanding a dataset with the aim of curation, then eventually the Purdue University Research Repository (Brandt). At Rice University, the Library’s Digital Scholarship Services team was seeing a lot of use from the emerging field of digital humanities, so they developed the Rice Digital Scholarship Archive and work closely with researchers to curate locally created digital humanities collections (Henry). To develop its research data management program, the University of Oregon Library performed a needs assessment study of researchers. The needs assessment identified several areas for development including helping researchers assess data management tools such as electronic notebooks, data management plan assistance, institutional repository, and data management training (Westra).

In 2017 the library cooperative OCLC released a four part report series titled *Realities of Research Data Management* aimed at exploring the “context, influences, and choices research universities face in building or acquiring RDM capacity” (Bryant et al. 5). The series starts with a small survey of RDM services before moving on to three separate reports examining scoping, incentives for, and sourcing/scaling of RDM services. The survey identified three categories of service that were practiced at some level in most institutions: education, expertise, and curation (6). To delve into the different phases of RDM service development, the series uses case studies from four different institutions: University of Edinburgh (UK), the University of Illinois at Urbana-Champaign (US), Monash University (Australia), and Wageningen University & Research (the Netherlands).

The case studies showed a distinct variety in approaches to RDM capacity building from these institutions. Despite these differences they identified several key findings. Among other things, the Scoping report exemplified the finding that RDM services at different institutions are shaped by a range of internal and external factors and will thus emphasize different service categories (28-29). A key takeaway from Part Three on Incentives report is that RDM service development is driven by local incentives that usually fall into one of four categories, compliance, evolving scholarly norms, institutional strategies, or researcher demand (18). The variety of internal and external factors as well as local incentives at each institution become clear in Part Four on Sourcing and Scaling as each institution employs a different strategy for developing services in house or utilizing outside resources.

In introducing the case studies produced for the OCLC reports, the authors note that there is no template for research data management service development and adoption will vary across institutions. However, they believe “...the *aggregated* experiences of research universities that have deployed various forms of RDM capacity provide useful markers, lessons and decisions points for other institutional contexts” (10). They point to several notable research data management case studies and large scale surveys (including Tenopir et al. and Fearon et al. mentioned above) produced so far but indicate that they believe more case studies are needed, “But the value of these case studies increases with the number available: in this way, the details of the individual case studies are complemented by a diversity of circumstances and contexts from which to draw them” (10). It is the authors’

hope that this case study of RDM service development at an institution which began by providing support on a case by case basis with plans to grow into a formalized program will add to this diversity.

Librarians' role in Research Data Management

Librarians have a history of collecting, organizing, describing, providing access to and preserving information. Since data is a type of information, it makes sense for librarians to be involved in similar work relevant to the data management lifecycle. While there are a number of models that define the stages of the research data management lifecycle, the UK Data Archive's Research Data Management Lifecycle model is a good place to start.



Pictured above, this data lifecycle includes the stages of planning, collecting, processing and analyzing, publishing and sharing, preserving, and re-using research data (UK Data Archive). Librarians can play a role in each stage of the lifecycle; regardless of the discipline they help support. While the types of data created and curated may vary based on fields of study (i.e. humanities, social sciences, sciences, biosciences), the stages remain constant.

Planning

Currently, librarians that support data management are most active in the planning and archiving stages. Data management planning is often the direct result of funder mandates and data management plans now often accompany grant funded research. Librarians can raise awareness of the benefits of data management planning and can also help guide researchers through the specific requirements of different funder mandates (Bryant et al. 1). Very often, they work to support the creation of formal data management plans (DMP). They do this through educational initiatives such as online tutorials, webinars, in person workshops and personalized consultations. The DMP Tool, created and curated by the University of California Curation Center of the California Digital Library and others (<https://dmptool.org>), has been a popular template amongst librarians who use it to support researchers developing a data management plan of their own. During this process, librarians have the opportunity to educate researchers on good data management practices such as guidance on naming protocols, preferred file formats and storage options.

Collecting

Collecting data is the stage in the lifecycle when researchers gather their data, including performing experiments or locating existing data. Librarians often provide data reference at this point of the process, helping researchers collect, purchase and/or license third party data sets that may be necessary to their research. If researchers are generating their own data, librarians can provide advice on documenting the collection process as well as file organization.

Processing and Analyzing

Processing and analysis of data may seem to be a stage during which librarians would be of little use. However, it is at this phase that librarians can identify and license or purchase data analysis tools for both quantitative (examples: Qlucore, MetaCore, etc.) and qualitative analysis. (examples: REDCap, NVivo) In addition, libraries can offer training on how to most efficiently use vendor provided analyses resources, or provide training in statistical analyses and coding methods that assist in data analysis such as SAS, Python, and R. It is at this point that librarians can provide guidance on data documentation and appropriate metadata to help ensure that the data is discoverable by other researchers (AAU-APLU).

Publishing and Sharing

In addition to the research outputs that will end up in a published paper, researchers may want to share or publish the data underlying the research results. Heather Piwowar showed that open data led to higher citation rates for the published article (Piwowar et al.), and many funders are requiring that data be made publicly available, assuming it does not contain sensitive information. Librarians can recommend options for publishing and sharing data including relevant repositories. Many institutions now offer an open access repository, often created and maintained by their library. Some repositories can be used to host data as well as other file formats. In addition to providing storage for the data, librarians can also provide tools for synchronizing/sharing data with collaborators. Open access repositories also promote the dissemination of research by making materials more easily discoverable and accessible to the public and other researchers. It is also useful at this stage for librarians to instruct researchers on data rights management.

Preserving

Preservation strategies include maintaining a long-term archival space and migration of data when needed. This is not always the same as publishing data, especially if the data is not able to be made publicly available in an open repository. Librarians who support data curation can ensure that file formats are usable or migrate data to new file formats as technologies evolve. They can also help generate permanent identifiers for data so it remains continually available.

Re-Using

The phase of re-using research data brings the research data life cycle around full circle. All steps prior to this make re-use possible. Librarians can instruct on the need for proper data citation when researchers' use others' data, particularly since it will affect citation counts and impact calculations which are meaningful metrics to many researchers.

Case study at Princeton University

Similar to many other research libraries, Princeton University Library (PUL) started Research Data Management (RDM) services as a response to funders' data management plan requirements. The library had already been providing a few long-standing services in economics and the social sciences that support data collection, analysis, sharing, and reuse. However, when the NSF announced its plans to begin requiring data management plans with grant applications a need for RDM services in the sciences was recognized. Initially science librarians worked together to create a template for NSF data management plans. Then as more and more funders started requiring data management plans a science librarian was tasked with creating research data management services for the sciences as a half-time eScience librarian. These services were based on services being developed by peers and included data management plan assistance, data management education, and assistance depositing in repositories. When the eScience librarian was promoted to lead the Engineering Library in 2015, library administration did not replace her but worked to develop a more robust response to campus needs. In absence of a dedicated librarian for the role, the Scholarly Communications Librarian and Engineering Librarian have continued the services initially developed on an ad-hoc and responsive basis. This has enabled the library to continue to engage in the campus response to research data management needs. It is our goal to develop our RDM services to a more robust program along the data lifecycle. The following are services that we provide at each stage of the research data lifecycle.

Planning

Raising awareness. Since the White House directive on public access to federally funded research in 2013 (Holdren), many funders have developed distinct and varying public access policies. In 2014, a few projects were developed to raise awareness on campus about the new policy and the importance of RDM. A website and a Libguide were created to focus on general information and resources on RDM. The RDM website and Libguide¹ serve as very useful resource on campus. A handout with information about the public access policies on publications and data from the top 5 funders² at Princeton University was also developed with a focus on informing faculty about the new requirement from funders. The library worked with the Office of the Dean for Research and distributed the handout to faculty. Our Population Research Librarian also created a guide specifically on NIH public access policy³ and detailed in steps how to comply with the mandate.

Educating the campus on research data management. In addition, as an important part of our RDM services, we educate the campus about the importance of data management and data management best practices by providing individual consultations and group workshops. Whenever we receive a question or request regarding data management, the Research Data Management (RDM) team, which consists of two librarians and one OIT (Office of Information Technology) staff, will schedule a meeting to talk about the details of the data management needs and offer customize solutions. For group workshops, we offer both tailored workshops for individual departments, such as a workshop specifically on best practices in data management; and a general session for the campus. A recent example is the Data Management Plan workshop offered for in March 2018. The workshop was a collaborative effort with colleagues across campus, including Library, OIT, the Office of Dean for Research, Research Computing, and Princeton Institute for Computational Science and Engineering (PICSciE). It had a great turnout with over ninety attendees, including faculty, graduate students, and grant managers.

Assisting in creating data management plan. Many funders with a public access policy also require a data management plan as part of the grant application. This requirement presents challenges to some faculty who have never written a data management plan before despite guidelines provided by funders. Therefore, there are rising needs for assistance in creating such a plan on campus. To address this emerging need, the Library formed a RDM Team (consisting of two librarians and one OIT staff), to develop services that could assist campus in creating data management plans. The first project that the team engaged was with OIT to adopt the DMPTool at Princeton. Following the DMPTool adoption, the team created promotional materials, such as postcard and a services brochure, to promote the DMPTool and RDM services. It didn't take long for the team to get requests to review other data management plans for grant applications.

In 2014, the Department of Energy released their public access plan in response to the 2013 White House directive. As a Department of Energy (DOE) national laboratory, the Princeton Plasma Physics Laboratory (PPPL) was tasked with writing a data management plan for data produced at PPPL under DOE funding as well as making the data underlying publications publicly available. The RDM Team consulted on the data management plan PPPL wrote in response to DOE's public access requirements.⁴

Customized support in data management plan creation. It's not often but we have gotten requests for guidance developing data management plans for a specific purpose other than for a grant requirement. In 2017, The Molecular Biology department requested help developing data management guidelines for use of shared equipment such as new

¹ Research Data Management at Princeton <https://libguides.princeton.edu/rdm>

² Research Funder Open Access Mandates <http://library.princeton.edu/sites/default/files/FunderRequirements-April14-2015.pdf>

³ NIH Public Access Policy Guide <https://libguides.princeton.edu/c.php?g=84087>

⁴ For further details see Dressel, Willow. *Helping a Department of Energy Laboratory Respond to Public Access Requirements*. https://escholarship.umassmed.edu/science_symposium/2016/posters/1. University of Massachusetts and New England Area Librarian e-Science Symposium.

microscopes that generate large amounts of data. The RDM Team and the liaison librarians worked together to develop recommendations for the department.

Collecting

Data Reference. Social Science Librarians at Princeton have been assisting researchers in determining their data needs and locating data sources such as using ICPSR and financial data sources like Bloomberg over decades. There has been an increase in research and teaching involving text mining across the social sciences, humanities and sciences at Princeton and liaison librarians of all disciplines are providing assistance finding appropriate corpora ranging from using the Scopus API or understanding licensing limitations of other databases to free corpora such as Project Gutenberg.

Purchasing and licensing datasets and data analysis tools. PUL has a long history of assisting researchers in licensing existing 3rd party data, such as statistical dataset- Data-Planet, ICPSR datasets, iPOLL databank, social science electronic data library-Sociometrics, integrated public use microdata series-IPUMS, and Wharton research data services-WRDS. Other databases recently acquired or licensed outside of social science include Biocyc a collection of genome database and LexisNexis Web Services Kit for text mining. As to the data analysis tools, the statistical packages are R/R Studio, Stata, and SPSS.

Processing and Analyzing

Offering Training on data analysis and visualization. The library's Data and Statistical Services⁵ (DSS) has a long history providing data and statistical consulting. Experts are available to advise Princeton University students, faculty, and staff on choosing appropriate data, application of quantitative research methods, and the interpretation of statistical analyses, data conversion, and data visualization. DSS provides statistical and software assistance in quantitative analysis of electronic data as part of independent research projects, such as junior papers, senior theses, term papers, dissertations, and scholarly articles. Students can schedule individual consultation through their website or just walk in the DSS lab to get help during the walk-in hours. DSS also offers online tutorials and group workshop on a variety of topics related to data analysis.

The Center for Digital Humanities⁶ (CDH) is an interdisciplinary research center and academic unit within the Library. CDH offers many services regarding data, including training on data visualization, grant on data curation (for humanities data sets, and consultation on issues related to humanities data curation), digital humanities software tools, and events (Year of Data 2018-2019).

Publishing and Sharing

Data sharing, publishing, and preservation. Princeton's Office of Information Technology (OIT) provides a data repository called DataSpace. When the service was launched, OIT consulted with cataloging librarians at Princeton on metadata for discovery. Librarians also work with OIT to promote DataSpace as a local solution for data sharing, publishing, and preservation. When Princeton Librarians consulted on the Princeton Plasma Physics Library's Data Management Plan, they also assisted OIT with setting up the DataSpace collection for public access compliance, including advising on providing study level documentation as well as linking data to the published article. Librarians also direct researchers to appropriate and available repositories for their disciplines to maintain grant compliance (i.e. PubMed Central for NIH Open Access Policy).

GIS data and services. The Map and Geospatial Information Center is located in the Lewis Science Library. The Center provides access to paper maps, geospatial data, digital maps and geographic information systems (GIS) services. In addition, the GIS center provides ongoing reference, research consultation and instruction to users with all levels of experience. The Center has eight workstations with 30-inch monitors loaded with GIS and satellite image processing software package, and commonly used geographic data. Faculty, students, staff and others in the Princeton University community are welcome to contact or visit the Center for additional information. At the center, users can access over 300,000 paper maps, charts,

⁵ Princeton Data and Statistical Services <https://dss-princeton-edu.ezproxy.princeton.edu/>

⁶ Princeton Center for Digital Humanities <https://cdh.princeton.edu/>

aerial and satellite photographs; create, modify, and print custom maps by using GIS software packages and third party data; analyze data graphically by conducting “what-if” scenarios⁷.

Re-using

The library is in full support of data reuse through licensing/purchasing data, providing workshops on legal and ethical considerations when reusing the data, providing guidance on citing data, and supporting campus workshops on generating reproducible data.

Conclusion and Next Steps

Challenges and Opportunities

Developing a comprehensive RDM program is not an easy job for a research library, even a large academic one. It is not only costly in terms of staffing and technical infrastructure (e.g. a data repository), but it also takes time to build relationships with faculty, researchers, students, service and resources providers (e.g. grant administrators), and to educate and engage the campus through the data management lifecycle. However the opportunities that RDM services bring to the library are too compelling to ignore. RDM services increase the library’s relevance to research, strengthen the relationship with campus primary stakeholders, confirm the library’s expertise in information management, demonstrate the library’s central role in supporting the institution’s mission, proves library’s ability in innovation and staying abreast of change, and affirms the leading role in open access.

Future plans

PUL was able to start small (for a particular purpose) and build up from there. From one science librarian, working half time, developing a service to help faculty comply with the funder’s DMP requirement, to an RDM Team providing multiple services, including DMPTool, DMP assistance, education and training, data storage, data publishing and sharing, and data preservation.

The development of a week long course for graduate students. In summer 2018, the library began collaborating with Research Computing, OIT, and PICSciE to develop a week long mini course on RDM for graduate students, which is to be offered in January 2019. The associate CIO of OIT initiated and led this effort based on his own experience as a doctoral student at Princeton. The mini course will be the first program to comprehensively teach skills and knowledge on RDM through the data lifecycle at Princeton. The mini course involves the presenters from several partners on campus, including Research Integrity and Assurance, Office of Research and Project Administration, OIT, and Research Computing. The long term goal is to develop it as a transcribed course at the Graduate School, which will be required for every graduate student.

Building towards a more robust RDM program. Over the last few years, Princeton’s University Librarian has been collaborating with the Dean for Research and the CIO of OIT to propose to the university administration an RDM program to meet the increasing needs on campus. The proposed program includes multiple dedicated staff with specific expertise in data management/curation and an infrastructure for data storage, discovery, publishing, and preservation. The proposal was approved last fall. The work on the staff recruitment and the development of an infrastructure for data storage and curation are currently underway. With all the good faith and work, we look forward to a more robust RDM program offered on campus in the near future.

⁷ GIS Center <http://library.princeton.edu/collections/pumagic>

References

- AAU-APLU. *AAU-APLU Public Access Working Group Report and Recommendations*. Association of American Universities and Association of Public and Land-grant Universities, 2017, <https://www.aau.edu/key-issues/aau-aplu-public-access-working-group-report-and-recommendations>.
- Brandt, D. Scott. "Purdue University Research Repository: Collaborations in Data Management." *Research Data Management : Practical Strategies for Information Professionals*, edited by Joyce M. Ray, West Lafayette, Indiana: Purdue University Press, 2014. *catalog.princeton.edu*, <http://www.jstor.org/stable/10.2307/j.ctt6wq34t>.
- Bryant, Rebecca, et al. "The Realities of Research Data Management." *OCLC Research*, <https://www.oclc.org/research/publications/2017/oclcresearch-research-data-management.html>. Accessed 28 Dec. 2018.
- Fearon, David Jr, et al. *Research Data Management Services, SPEC Kit 334 (July 2013)*. Association of Research Libraries, 2013. *publications.arl.org*, <http://publications.arl.org/Research-Data-Management-Services-SPEC-Kit-334/>.
- Henry, Geneva. "Data Curation for the Humanities: Perspectives from Rice University." *Research Data Management : Practical Strategies for Information Professionals*, edited by Joyce M. Ray, West Lafayette, Indiana: Purdue University Press, 2014. *catalog.princeton.edu*, <http://www.jstor.org/stable/10.2307/j.ctt6wq34t>.
- Holdren, John. *Increasing Access to the Results of Federally Funded Scientific Research*. *Memo, Executive Office of the President Office of Science and Technology Policy, 22 Feb. 2013*, https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.
- Steinhart, Gail. "An Institutional Perspective on Data Curation Services: A View from Cornell University." *Research Data Management: Practical Strategies for Information Professionals*, edited by Joyce M. Ray, Purdue University Press, 2014.
- Tenopir, Carol, et al. *Academic Libraries and Research Data Services Current Practices and Plans for the Future*.
- UK Data Archive. "Research Data Lifecycle." *UK Data Service*, 2019 2012, <https://www.ukdataservice.ac.uk/manage-data/lifecycle.aspx>.
- Westra, Brian. "Developing Data Management Services for Researchers at the University of Oregon." *Research Data Management : Practical Strategies for Information Professionals*, edited by Joyce M. Ray, West Lafayette, Indiana: Purdue University Press, 2014. *catalog.princeton.edu*, <http://www.jstor.org/stable/10.2307/j.ctt6wq34t>.

Data Management and the Role of Librarians

Plato L. Smith, Sara Gonzalez, and Jean Bossart

George A. Smathers Libraries; University of Florida, United States

Abstract

'Research Data Science' is defined by Committee on Data of the International Council for Science Research Data Alliance ([CODATA-RDA](#)) as an ensemble of (a) Open Science principles and practices ([FAIR](#)) and research data management and curation skills, (b) the use of a range of data platforms and infrastructures, (c) large scale analysis, (d) statistics, (e) visualization and modeling techniques, (f) software development and annotation, and (g) more. Data management and the role of librarians must now include developing expertise and training with faculty, students, and staff on "research data science" directly and/or indirectly through collaborative library/faculty partnerships. To meet this need, librarians at the University of Florida have developed a new research support service called Academic Research Consulting & Services ([ARCS](#)) to assist faculty, students, and staff with their data management and research needs. The library-centered, campus-wide focused UF Data Management and Curation Working ([DMCWG](#)) and ARCS work in collaborative partnerships with the campus units such as the UF Informatics Institute, UF Data Carpentry Club (<https://github.com/UF-Carpentry>), and UF Data Science & Informatics ([DSI](#)) undergraduate student organization to provide support to pre- and post-grant research and teaching. This new role of librarians is to facilitate library/faculty collaborations and broker resources that contribute to the facilitation of promulgating 'research data science' skills at scale for their respective institutions. This paper will discuss the developing outreach activities, inter-departmental collaborations, some initial outcomes, and future goals of leveraging capacity, infrastructure, and resources to develop data management efforts across communities of practice within an institution's current organizational culture. One aim of this paper is to highlight the importance and significance of developing good library and faculty partnerships built on character, integrity, and humility as the cornerstones for the roles of librarians as collaborators in promoting socio-technical data management programs.

Introduction

All data is not Findable, Accessible, Interoperable, and Reusable (FAIR). Researchers face many data management challenges resulting from the need to comply with funding agencies' data management and sharing requirements. While major funding agencies such as the American Heart Association (AHA), National Institutes of Health (NIH), NSF, and United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA) have data management and sharing requirements guidelines and/or statements, researchers seeking funding struggle with complying, fulfilling, and negotiating data lifecycle management due to existing capacity, infrastructure, resources, and support. Data lifecycle requirements include:

- a) Key components of a data management plan [6] and data lifecycle processes [7]
- b) Data management life cycle [1]
- c) Research data management services for scalability, support, & sustainability [8].

The coordination, collaboration, and connection of diverse stakeholders (See Fig 3) across multiple departments, organizations, and units in support of good data management requires consistent organization, technology, and resources to maintain stability [2] over time.

In January 2011, the National Science Foundation began requiring researchers to include a research data management (RDM) plan in all of their funding proposals prompting academic libraries to develop data management tools and roles [9]. The mandate was further expanded through executive orders and legislation in 2013, when many grant applications were required to include a RDM plan for the purpose of making the research data more openly available [10]. Up until this mandate, most academic libraries were offering only limited RDM services. This mandate provided the catalyst for new services and challenges for academic libraries [11, 12]. Universities and academic libraries responded with developing

their own institutional framework to comply with the mandates [14]. Even as funding agencies and policy makers supported making research data available to others, data sharing is still often met with resistance and the data sharing process is complex [14].

What are some of the barriers to data sharing? Tenopir *et al* examined differences in data sharing practices across researchers' age groups, geographic regions, and subject disciplines [15]. Variations in data management practices are often seen across academic disciplines [16]. Deposition and sharing practices vary across journal publishers, repositories, and universities [17]. The data itself seems to be a barrier to data sharing. Data tends to be stored in local domains and have minimal structures and documentation making shared data difficult to fully be utilize by other researchers [18]. Data newly generated as part of a grant is often easier to organize and store than legacy datasets which offer additional challenges [19]. A 2016 article in PLoS One by VanTuyl and Whitmire evaluated how effective data sharing has been and the usability of the data that was shared [20].

One way to increase data sharing is to reward those who share data and those who utilize shared data. In peer reviewed literature, rewarding researchers who create well-documented data sets is one method of encouraging good data curation practices. The scholarly communication community is encouraging data publication [21]. Studies have shown that scholarly publications which used a publicly available data set had a higher number of citations than similar studies without publicly available data [22].

While there are many tools and recommendations for creating a good RDM plan [23], there are arguments for and against research libraries taking on the role as research data managers as there are many competing agencies in the data curation field [24]. Pinfield *et al* examined the roles and relationships involved in RDM and the major drivers and components of developing an RDM plan [25]. As libraries take on a more active role in RDM, a shift in the libraries' role from support to partnership with researchers is emerging [26]. Librarians who are becoming more involved with research data are being asked to assist with preparing RDM plans and insertion of data into data repositories [27]. Since data management is a new service many academic libraries are asked to provide, there are challenges both intellectual and economic related to offering these new services [28]. Although more researchers are aware of the need for management of data sets, RDM is still primarily conducted at the institutional level using a locally developed framework [29] in need of a socio-technical systems thinking [3] approach.

Socio-technical data management collaborations

This paper introduces the concept of socio-technical data management collaborations as one approach in developing the data management and curation skills required for librarians to understand and contribute to research data science as collaborators and consultants. The socio-technical systems thinking approach framework (See Fig. 1) was selected for this paper as the preferred model to begin investigating the interrelated changing dynamics of a socio-technical system in the context of data management collaborations. "Social-technical systems theory has around 60 years of development and application internationally by both researchers and practitioner [3]". Understanding and analyzing data management and the required collaborations from a socio-technical systems thinking perspective at an academic research institution is beyond the scope of this paper. However, the application and use of *steps in analyzing and understanding an existing socio-technical system* (See Table 1) of data management can aid in the development of socio-technical data management collaborations and a Business Model Canvas (See Fig 2) on which to build benchmarks, metrics, and success.

Table 1.
The steps involved in analyzing and understanding an existing socio-technical system [3].

Step	Task description
1	Gather relevant data from appropriate sources, including key actors, stakeholders, subject matter experts, and internal and external documents.
2	Analyze and classify data, using techniques such as template analysis (King, 2004). Initial template consists of the socio-technical framework.
3	Identify and group key system factors. Visually represent the groups of factors on each node of the framework.
4	Consider the implication of the external environment in which the system is embedded within the node which it relates.
5	Systematically consider relationships between each set of factors, and identify contingencies and direction of relationships.
6	Visually inspect the hexagon framework and assess underexplored or related areas, and reappraise evidence or seek input from colleagues and subject matter experts (e.g., with expertise in buildings and infrastructure).
7	Add any additional relevant factors that emerge from the data during analysis or following previous step.
8	If appropriate: Generate a timeline of key factors leading up to the event or scenario, grouped by six factors. Classify as: (long-standing issues (3+ months); issues immediately preceding the event (0-3 months); and factors involved on the day.
9	Test analysis on key stakeholders for accuracy, omissions and interpretations, and modify as necessary after discussion.
10	Generate key inferences regarding the system and how it works.

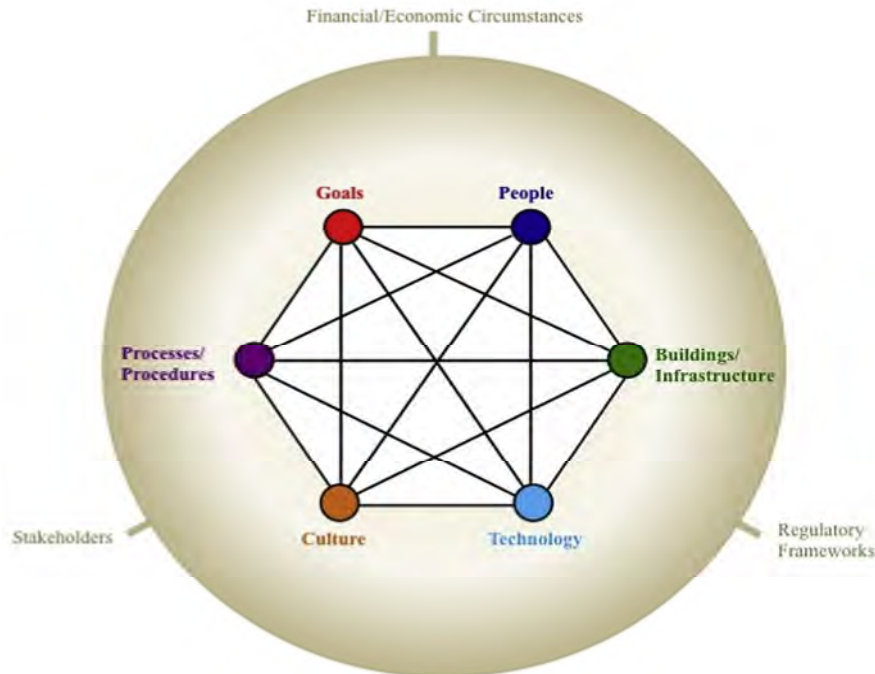


Figure 1. Socio-technical system, illustrating the interrelated nature of an organizational system, embedded within an external environment [3].

Figure 1 illustrates the conceptual framework of key interrelated nodes that figure prominently within an existing organizational system. Analyzing and understanding each of the nodes and their impact within an organization system using data management and cyberinfrastructure as variable for analysis can provide useful data in support of capacity, infrastructure, and resources necessary for socio-technical data management collaborations within and across departments, funded projects, institutions, labs, organizations, and units.

The socio-technical framework was used a frame of reference for this paper. Below are a few developing socio-technical data management collaborations via the DMCWG at the University of Florida.

- **UF Marston Science Library (MSL)** – provides DMCWG monthly meetings venue, STEM research, resources, support, training and partnership in 1st Data Symposium event
- **UF Office of Research** – provides access to senior stakeholders' guidance and support
- **UF Informatics Institute (UFII)** – provides venue, promotion, and lunch for the DMCWG training workshops and membership to the UFII Faculty Affiliates
- **UF Research Computing** – provides university-wide capacity, infrastructure, resources, support, and membership to the Research Computing Advisory Committee (RCAC)
- **UF College of Engineering** – provides data management plan referrals and partnership
- **UF Clinical and Translational Institute (UF/CTSI)** – provides data management and analysis and informatics consultations, expertise, and support for funded projects
- **UF Institute of Food and Agricultural Sciences (UF/IFAS)** – provides DMP projects
- **UF Biomedical Informatics (UF BMI)** – provided invitation to 5 yrs. strategic plan retreat

The Archive Development Canvas (Detailed-Level Version)

This is a brain-storming tool when starting up new data archives or services or extending/developing existing ones. The User Guide and other component tools in the cost/benefit advocacy tool kit can help complete it. Prompts are in grey text. As you complete each section you should begin to see connections to the others. The value proposition (benefits) is central.

Key Partners Host institution? Funders? Data creators/depositors? Data users? Project /service partners? Supporters/volunteers/interns (testing, user champions, etc.)?	Key Activities Acquisition and creation of Products (datasets, tools, etc.)? Services (platform, helpdesk, training, promotion etc.)?	Benefits What are the benefits? (Use the Benefits for a Data Archive worksheet and the KRDS Framework to develop this) Can you measure benefits? (see Key Metrics)	Beneficiaries Who benefits? (Use the Benefits for a Data Archive worksheet and the KRDS Framework to develop this)	Beneficiary Relationships Personal/Automated? Grant/contract/non-regulated relationship?
Key Resistances Competitors? Beneficiaries of status quo? Potential roadblocks (legal, existing policies, culture and practices, etc.)?	Key Resources Data and metadata? Staff knowledge and skills? Technical and organisational Infrastructure (tools, ontologies, depositor/user agreements, etc.)? Professional networks?			Channels To raise awareness? To evaluate service benefits? To provide access /delivery /support? To improve integration?
Cost Structure Existing institutional cost structure (salaries, equipment, utilities, etc.)? Fixed costs/variable costs? Direct/indirect costs? Non-costed activities (volunteers, etc.)? Activity based costing (if known)? Dataset based costing (if known)?		Key Metrics Deposit metrics? User metrics? Service metrics? Impact metrics? Costs of inaction?	Funding Streams "Core" public funding? Project funding? In-kind (infrastructure, accommodation, etc.)? Deposit/access charges? Other (consultancy, training, donations, volunteers, etc.)?	

Developed from Business Model Canvas www.businessmodelgeneration.com for the CESSDA SaW Project by Charles Beagrie Ltd ©2017. This work is licensed under the Creative Commons Attribution-Share Alike 4.0 Unported License. <https://creativecommons.org/licenses/by-sa/4.0/> Requested attribution: Archive Development Canvas (Detailed Version), Charles Beagrie Ltd and CESSDA 2017, <http://dx.doi.org/10.18448/16.0009> Project funded by the EU Horizon 2020 Research and Innovation Programme under the agreement No 674939

cessda saw

Figure 2: The Archive Development Canvas (Detailed-Level Version)¹

Figure 2 is an illustration of the Consortium of European Social Science Data Archives (cessda) Strengthening and widening (saw) the European infrastructure for social science data archives brain-storming tool when starting up a new data archives or services or extending/developing existing ones as part of a cost/benefit advocacy tool kit. Fig 2 is introduced as a complimentary tool for use in conjunction with Fig 1 to explore ways of developing sustainability for socio-technical data management collaborations. Oftentimes researchers are required to provide data management plans without the funding resources to sustain data lifecycle management. Developing a data management plan that includes the economic and financial considerations for long-term preservation require investments from multiple stakeholders (See Fig 3). Researchers must now consult, collaborate, and work with campus groups addressing data management in order to effectively and successfully comply with funding agencies' mandates.

¹ Charles Beagrie Ltd. (2017). CESSDA SaW Archive Development Canvas (Detailed Version). <http://tinyurl.com/ybpy94hu>.

Data Management and Curation Working Group (DMCWG)

The Data Management and Curation Task Force (DMCTF) (2012 – 2016) was formed to develop a culture of data management at UF. The task forces developed data management initiatives, projects, presentations, and the Data Management Librarian faculty position. The Data Management Librarian hired in 2016 developed a DMCWG Charge², Year End Reports³, and organized for the 1st Annual Data Symposium. The DMCWG has 53 subscribers, comprised of chairs, directors, faculty, IT, and staff, and meets every 4th Monday in Marston Science Library.

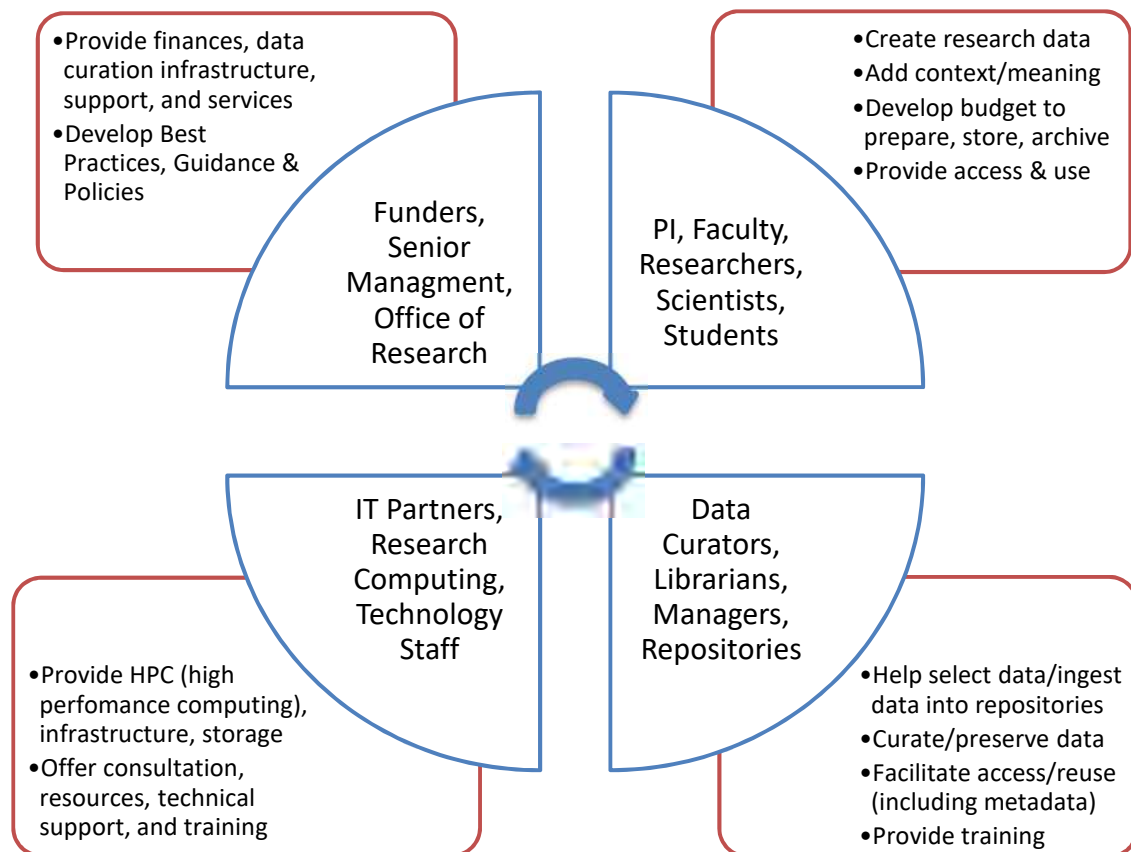


Figure 3: Stakeholders' roles aligned with Data Curation Lifecycle Responsibilities⁴

Figure 3 was developed from the DCC Curation Lifecycle Model mapped to key stakeholders involved in data curation lifecycle process as illustrated by the DAF implementation guide. Figure 3 is an example of some key stakeholders and their overlapping roles in the curation, management, and planning of research data during its lifecycle. In order to ensure compliance with evolving funding agencies data sharing requirements, stakeholders must collaborate, communicate, and aggregate best practices, guidelines, and recommendations across domains.

The DMCWG developed and distributed a small pilot data management survey to researchers at UF in spring 2017. The survey recorded 159 starts and 139 completes. The survey participants included faculty 52.94% (81), staff 16.99% (26), graduate/professional student 22.22% (34), postdoctoral fellow 3.92% (6), resident 1.31% (2), undergraduate 1.31% (2), and other 1.31% (2). Of the 159 respondents, 156 indicated that they would participate in the survey, although not all participants answered every question. Due to limited scope of this paper, only one question from the survey is included to support socio-technical data management collaboration.

² DMCWG/DMCTF. (2016). DMCWG. <http://ufdc.ufl.edu/AA00014835/00076>.

³ DMCWG Year End Reports. (2019). DMCWG Year End Report (2018) <http://ufdc.ufl.edu/AA00014835/00135>; DMCWG Year End Report (2017) <http://ufdc.ufl.edu/AA00014835/00117>; DMCWG Year End Report (2016) <http://ufdc.ufl.edu/AA00014835/00085>.

⁴ JISC, University of Glasgow – HATII, & DCC. (2009). [DAF Implementation Guide](#), p. 3. (Adapted)

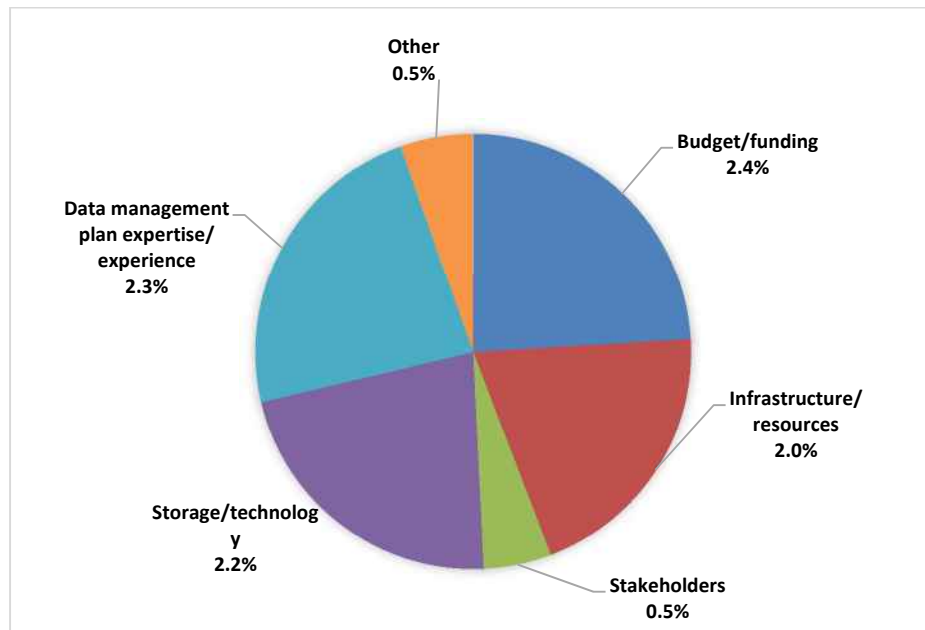


Figure 4. Data management and storage barriers (Q15)

Figure 4 identifies some barriers to data management and storage that are also included in the nodes in the socio-technical framework in Figure 1. When asked to identify barriers regarding management and storage of data, the researchers were given six possible choices and could choose more than one response, if it applied. Of the 258 responses by the 108 researchers that answered this question, 78 researchers chose more than one answer. As shown in Figure Q15, the largest barrier was budget/funding at 24.0% followed closely by data management expertise/experience at 23.3%, storage/technology at 22.1%, and infrastructure and resources at 20.2%. Thus, a socio-technical framework with an embedded business model can help librarians assist researchers in addressing and solving some data management issues.

Some DMCWG accomplishments include but not limited to: several general data management training workshops with an average attendance of 12 participants; several discipline-specific training workshops; several DMP support projects resulting in successful awards totaling \$6.2 million dollars (US); 1st Annual Data Symposium⁵; two sub-awards to the UF Libraries; lead on a UF Data Management and Analysis Core (DMAC) for a National Institutes of Health (NIH) National Institute of Environmental Health Sciences (NIEHS) grant proposal.

The Carpentries @ UF

The [Carpentries @ UF](#) was established in 2016 to provide data and software carpentries workshops. There were 18 workshops in 2018 (up to 10/24/18) that included R for Geospatial Data; One Day R workshop; Research Bazaar; Data Carpentry (R); Software Carpentry (Python); Data Carpentry Workshops for the 1st Annual Data Symposium; Instructor Training; and Multi-Week Data Carpentry. The Carpentry @ UF premium membership is \$15k annually and includes extended collaborations with UFII, UF Biodiversity Institute (UFBI), UF Marston Science Library, iDigBio, UF Florida Museum of Natural History (FLMNH), Wildlife Ecology, and the White Lab.

UF Data Science and Informatics (DSI)

The [UF DSI](#) is a recognized UF undergraduate student organization that is creating a data science community through training workshops. Training workshops include use of Jupyter notebooks and the anaconda distribution, use of R Studio for R workshops, use of MySQL for

⁵ 1st Annual Data Symposium. (2018). Enabling Data Reproducibility and Sustainability. March 19, 2018. <http://cms.uflib.ufl.edu/envisioning-data-symposium/Index.aspx>.

SQL workshops, and developing learning pathways for students. UFII fellows and students attend of help at workshops. DSI is funded by UFII and collaborates with UF Libraries/MSL.

Academic Research Consulting & Services (ARCS)

Academic Research Consulting and Services ([ARCS](#)) offers a wide range of research support services to the University of Florida community. Hosted by the George A. Smathers Libraries, ARCS has been developed to provide expert services throughout the research process – from data collection through publication. The team also has expertise in the several aspects of data science, including: Statistical analysis (SAS, SPSS); Writing of computer code in using R and Python for analysis of data; Bioinformatics and genomic data analysis using Galaxy, HiPerGator, and R; Geospatial analysis (ArcGIS); Data management and archiving; 3D printing/visualizations.

Testimonials

"Participating in DMCWG gave me chance to work with so many different experts in developing and promoting Data Management service, librarians, research administration and IT managers. Also, I have learnt a lot from the DM training and workshops. At first so many topics (DM planning, Data Sharing and Storage, data security and backup and so on) sound overwhelming, but fortunately, there are more practical tools than ever before to help to achieve the goal. Thanks to the workshop, I know where to start and how to manage the data and made them more accessible." – UF Hough Graduate School of Business, Information Systems and Operations Management graduate (former UF Libraries' student employee), February 4, 2018

Concluding thoughts

"Long-term curation and preservation represent a complex set of challenges, which are exceptionally difficult for data centres and institutions to address individually. They will require a step change in current investment and approaches, and concerted effort on fundamental research, development of shared services, expertise and tools to assist organisations in this work" (JISC Circular 6/03 (Revised), 2003) [5].

References

1. University of New South Wales (UNSW). (2017). Data Governance Policy, Appendix 1 - Data Management Life Cycle. p. 7. Available from: <https://www.gs.unsw.edu.au/policy/documents/datagovernancepolicy.pdf>.
2. Joint Information Systems Committee (JISC), University of Glasgow Humanities Advanced Technology & Information Institute (HATII), and Digital Curation Center, 2009. Data Asset Framework [formerly Data Audit Framework] Implementation Guide. Retrieved May 30, 2013 from http://www.data-audit.eu/docs/DAF_Implementation_Guide.pdf.
3. Davis, M. C., Challenger, R., Jayewardene, D. N. W., & Clegg, C. W. (2014). Advancing socio-technical systems thinking: A call for bravery. *Applied Ergonomics*, 45(2 Part A), 171–180. <https://doi.org/10.1016/j.apergo.2013.02.009>.
4. Charles Beagrie Ltd. (2017). CESSDA SaW Archive Development Canvas (Detailed Version). Available from <http://tinyurl.com/ybpy94hu>.
5. JISC. (2003). JISC Circular 6/03 (Revised). An invitation for expressions of interest to establish a new Digital Curation Centre for research into and support of the curation and preservation of digital data and publications. Accessed January 19, 2019 online from <http://tinyurl.com/y95rq5wp>.
6. Digital Curation Centre. Checklist for a Data Management Plan [Internet]. Version 4.0. Edinburgh; 2013 [cited 2017 Jan 23]. Available from: <http://www.dcc.ac.uk/resources/data-management-plans>
7. United States Geological Survey. USGS Data Management [Internet]. Data Lifecycle Overview. 2013 [cited 2017 Jan 20]. Available from: <https://www2.usgs.gov/datamanagement/why-dm/lifecycleoverview.php>
8. Jones S, Pryor G, Whyte A. How to Develop Research DataManagement Services: A Guide for HEIs'. DCC How-to Guides [Internet]. Edinburgh; 2013. Available from: <http://tinyurl.com/hpf6nod>
9. Antell K, Foote JB, Turner J, Shults B. Dealing with Data: Science Librarians' Participation in Data Management at Association of Research Libraries Institutions. *Coll Res Libr*. 2014;75(4).

10. Diekema AR, Wesolek A, Walters CD. The NSF/NIH Effect: Surveying the Effect of Data Management Requirements on Faculty, Sponsored Programs, and Institutional Repositories. *J Acad Librariansh.* 2014;40(3–4).
11. Cox AM, Pinfield S. Research data management and libraries: Current activities and future priorities. *J Librariansh Inf Sci.* 2014;46(4).
12. Saunders L. Academic Libraries' Strategic Plans: Top Trends and Under-Recognized Areas. *J Acad Librariansh.* 2015;41(3).
13. Hickson S, Poulton KA, Connor M, Richardson J, Wolski M. Modifying researchers data management practices: A behavioural framework for library practitioners. *IFLA J.* 2016;42(4).
14. Fecher B, Friesike S, Hebing M. What drives academic data sharing? *PLoS One.* 2015;10(2).
15. Tenopir C, Dalton ED, Allard S, Frame M, Pjesivac I, Birch B, et al. Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLoS One.* 2015;10(8).
16. Weller T, Monroe-Gulick A. Understanding methodological and disciplinary differences in the data practices of academic researchers. *Libr Hi Tech.* 2014;32(3).
17. MacMillan D. Data sharing and discovery: What librarians need to know. Vol. 40, *Journal of Academic Librarianship.* 2014.
18. Wallis JC, Rolando E, Borgman CL. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PLoS One.* 2013;8(7).
19. Marshall B, O'Bryan K, Qin N, Vernon R. Organizing, contextualizing, and storing legacy research data: A case study of data management for librarians. *Issues Sci Technol Librariansh.* 2013;74.
20. Van Tuyl S, Whitmire AL. Water, water, everywhere: Defining and assessing data sharing in Academia. *PLoS One.* 2016;11(2).
21. Kratz JE, Strasser C. Researcher perspectives on publication and peer review of data. *PLoS One.* 2015;10(2).
22. Piwowar HA, Vision TJ. Data reuse and the open data citation advantage. *PeerJ.* 2013;1.
23. Michener WK. Ten Simple Rules for Creating a Good Data Management Plan. *PLoS Comput Biol.* 2015;11(10).
24. Jørn Nielsen H, Hjørland B. Curating research data: the potential roles of libraries and information professionals. *J Doc.* 2014;70(2).
25. Pinfield S, Cox AM, Smith J. Research data management and libraries: Relationships, activities, drivers and influences. *PLoS One.* 2014;9(12):1–28.
26. Cox AM, Verbaan E. How academic librarians, IT staff, and research administrators perceive and relate to research. *Libr Inf Sci Res [Internet].* 2016;38(4):319–26. Available from: <http://dx.doi.org/10.1016/j.lisr.2016.11.004>
27. Tenopir C, Sandusky RJ, Allard S, Birch B. Research data management services in academic research libraries and perceptions of librarians. *Libr Inf Sci Res [Internet].* 2014 Apr [cited 2017 Aug 7];36(2):84–90. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0740818814000255>
28. Kennan MA, Corral S, Afzal W. "Making space" in practice and education: research support services in academic libraries. *Libr Manag.* 2014;35(8/9).
29. Patel D. Research data management: a conceptual framework. *Libr Rev.* 2016;65(4/5).
30. Greene JC, Caracelli VJ, Graham WF. Toward a conceptual framework for mixed-method evaluation designs. *Educational evaluation and policy analysis.* 1989 Sep;11(3):255–74.
31. Creswell JW, Klassen AC, Plano Clark VL, Smith KC. Best practices for mixed methods research in the health sciences. Bethesda (Maryland): National Institutes of Health. 2011 Aug 1;2013:541–5.

Measuring Reuse of Institutionally-Hosted Grey Literature

Ayla Stein Kenfield, University of Illinois at Urbana–Champaign;

Elizabeth Kelly, Loyola University New Orleans;

Caroline Muglia, University of Southern California;

Genya O’Gara, Virtual Library of Virginia;

Santi Thompson, University of Houston;

Liz Woolcott, Utah State University, United States

URL: <https://reuse.diglib.org>

Abstract

Grey literature is often hosted in digital libraries such as institutional repositories (IRs) and managed by digital librarians who are tasked with proving the value of these collections. Newly developed multimedia COUNTER standards for IRs help guide collection and analysis of standardized use data but lack qualitative or storytelling measures that can add much needed nuance to assessment. Content reuse, or how often and in what ways digital repository materials are utilized and repurposed, can bridge this gap for IRs. Reuse of digital repository collections and materials is an important assessment measurement because, unlike use, reuse shows engagement with collections and impact of digital repository resources in a more meaningful fashion.

The research team for Developing a Framework for Measuring Reuse of Digital Objects received a \$70,850 grant from the Institute of Museum and Library Services (LG-73-17-0002-17) to address this problem. The grant’s primary deliverables focused on the creation and application of an in-depth needs assessment of the digital repository community to determine desired functionality of a reuse assessment toolkit. The research team’s outputs include the development of well-defined requirements and use cases which serve as the building blocks for an assessment toolkit that goes beyond use and focuses on transformation.

The researchers employed surveys, focus groups, and data tagging and analysis to assess digital repository needs for measuring reuse. Specific data points regarding reuse of IR materials were extracted and analyzed for this conference presentation. Measuring reuse of IR materials has emerged as a complex issue. Potential use cases thus far include: collecting data on what content was not reusable in order to develop deselection criteria; mapping dataset reuse to data management plans; assessing use and reuse of IR materials by stakeholders such as grant funding agencies; and faulty or missing citations in electronic theses and dissertations. A broader concern emerged regarding weighting types of reuse differently depending on the type of digital collection as well as the mission and priority of the hosting institution. Participants identified lack of best practices, documented workflows, assessment training, and staffing as the greatest barriers to assessing reuse.

As the grant ended in June 2018, the research team turned its focus to synthesizing the results of the needs assessment to address the challenges of measuring reuse. This presentation explores the findings of this research, specifically its impacts on Grey Literature hosted in IRs.

Introduction

Reuse of digital repository works and collections is an important assessment measurement because, unlike use, reuse shows engagement with collections and impact of digital repository resources in a more meaningful fashion. This paper focuses on the findings of a needs assessment from the “Developing a Framework for Measuring Reuse of Digital Objects” (Measuring Reuse) grant project that are directly pertinent to grey literature and institutional repositories (IRs).

The Measuring Reuse grant focused on carrying out an in-depth needs assessment of the digital repository community, spanning the Cultural Heritage and Knowledge Organization (CHKO) spectrum, to determine desired functionality of a reuse assessment toolkit. Via surveys and focus groups, the project team elucidated from this needs assessment the features, functionalities, and use cases the digital repository community prioritized in a reuse assessment toolkit.

During this year-long needs assessment, the team gathered data using two surveys, five in-person focus groups and five virtual focus groups. The researchers derived use cases from the coded focus group data, which were, along with features of a future toolkit for assessing reuse, prioritized by the digital repository community in the second survey. Measuring reuse of grey literature and IR collections has emerged as a complex issue, with all the complications of assessing reuse of published material with additional challenges unique to grey literature and its myriad forms.

The paper will conclude with avenues of future research from the Measuring Reuse project centering assessment of IRs and grey literature, and the research team's next steps. The remaining sections of this paper are as follows: Grant and Project Background, Literature Review, Focus Group Data, Discussion, and Conclusion.

Grant and Project Background

This work is a result of the 2017-2018 Institute of Museum and Library Services (IMLS) National Leadership/National Forum (NL/NF) grant project, Developing a Framework for Measuring Reuse of Digital Objects¹ (hereafter referred to as "Measuring Reuse"), which began on July 01, 2017 and ended June 30, 2018. The grant itself was written based on recommendations from the 2015 white paper, "Surveying the Landscape: Use and Usability Assessment of Digital Libraries", which identified multiple areas in the digital library assessment literature where more research and work would benefit the digital repository community. The Measuring Reuse project team were original members of the User Studies Working Group (USWG) of the Digital Library Federation (DLF) Assessment Interest Group (AIG), which wrote the "Surveying the Landscape" white paper. After the release of the white paper and in order to best tackle its recommendations, the USWG further divided itself into several subgroups. The members of that Content Reuse Subgroup became the project team².

In the following section, the literature review will provide an overview of the professional scholarship on assessment of grey literature reuse overall with a special interest on institutionally-hosted grey literature.

Literature Review

For the purposes of this paper, grey literature will be defined to include all of the document types identified in the GL Survey 2004, the results of which are listed on the GreyNet website ("Grey Literature - GreySource"). Previous research by the project team was conducted in conjunction with other members of the DLF AIG and culminated in the previously mentioned 2015 white paper, "Surveying the Landscape: Use and Usability Assessment of Digital Libraries," which identified three areas of focus within digital library assessment: use and usability studies, return on investment, and content reuse (Chapman et al. 2015). The white paper outlined the paucity of research surrounding digital content reuse, particularly with non-imaged-based objects and identified challenges in assessing content reuse. Further research by the Measuring Reuse project team in 2017-2018 explored more recent reuse assessment literature, some of which expanded the boundaries of the original research project's definitions of digital repository content and which is particularly relevant to assessing the reuse of grey literature. Dataset reuse, in particular, is the subject of new research, as is the reclassification of digital collections themselves as datasets (O'Gara et al. 2018).

Reuse of digital repository collections and items is an important assessment measurement because, unlike use, reuse shows engagement with collections and impact of digital repository resources in ways that are more valuable to stakeholders. Reuse data can be looked at to show the success of digitization and born-digital hosting efforts as well as to distinguish positive reuse, such as text-mining digitized books for a digital humanities (DH) project, from negative reuse, such as plagiarism. Buchanan and McKay discuss the use of plagiarism-checking software and manual examination to find instances of plagiarism from items in digital library collections to analyze connections between open access publications,

¹ Developing a Framework for Measuring Reuse project website: <https://reuse.diglib.org/>

² Content Reuse Subgroup OSF Repository: <https://osf.io/36npw/>

prestige, and plagiarism (2017). Similarly, Biagioni et al. assessed reuse of grey literature through citations to find evidence that open access publishing, such as that of conference proceedings, may increase the likelihood of scholarly publications citing grey literature (2017). Ferreras-Fernández et al. also found that open access theses in IRs received significant numbers of citations in academic research (2016). Finally, reuse assessment of technical and markup language documentation may also point researchers in the digital humanities to examples of other researchers producing similar projects using theirs as templates (Warwick et al. 2009).

There are challenges to assessing reuse, as identified previously by the research team and then further examined in literature regarding grey literature reuse. A lack of persistent identifiers in grey literature prevents capitalizing on the potential of altmetrics tools for more accurate reuse assessment (Cancedda and De Biagi 2017; Schöpfel and Prost 2016). Big data tools could allow grey literature repositories to better analyze (or help others analyze) their collection materials, but a dearth of sufficient resources and skilled personnel prevent widespread use among the grey literature community (Crowe and Candlish 2013). Existing infrastructure, such as the COUNTER standards for assessing use of electronic resources, focus primarily on subscription resources and do not filter out web robots, making them difficult to be modified for grey literature reuse assessment (Greene 2017). And open data can be much more difficult to track reuse of than data that requires permission for reuse (Parsons and Summers 2016).

Some possible solutions in the literature provide further areas of exploration for the project team. The use of linked data and publication codes such as ISBN, ISNI, ISIL, and ISSN could combat inconsistency in persistent identifiers for grey literature (Cancedda and De Biagi 2017). The COUNTER Robots Working Group is striving to recommend spider detection techniques for measuring the use of open digital content, including grey literature such as research data and institutional repository materials (Greene 2017). IRUS-UK, the Institutional Repository Usage Statistics United Kingdom service, is engaging in similar attempts by adding a short “Tracker Protocol” code to its repository that strips robots from usage data and converts the remainder to COUNTER-compliant statistics (MacIntyre and Jones 2016). The UK Data Service Communications team tracks the impact of research that uses their data in part by requiring researchers to register with the UK Data Service in order to use some datasets, and by requesting that researchers contact them with notifications of any publications that cite their collections (Parsons and Summers 2016).

Several new and recently proposed systems in the Open Science community in particular may provide needed infrastructure for reuse assessment of grey literature. The OpenAIRE-Connect project has exciting implications for reuse assessment. The “one-stop-shop” lets researchers “publish” portions of their research that are not published on their own traditionally, such as research methods, alongside their data and articles, and links all the related objects together easily. It also provides a notification service so researchers know when their research materials have been used and reused (Manghi et al. 2017). The Socionet researcher information system, too, proposes a semantic setup that facilitates researcher publication of the entire grey literature life-cycle. One function, Socionet Statistics, provides statistics on research artifact reuse (Parinov, Kogalovsky, and Lyapunov 2014).

Project Overview

In-depth explanations of the Measuring Reuse data collection and analysis methods are published elsewhere and will not be reiterated in full here (O’Gara et al., 2018; Kelly et al. 2018a). For clarity’s sake, the data collection methods will be covered briefly in the remainder of this section.

The research team employed several methods of data collection throughout the course of the Measuring Reuse project, including two surveys; one at the beginning of the project to gauge the digital repository community’s interest in certain topics (O’Gara et al. 2018); and the second at the end of the project that asked the community to indicate support and prioritize potential use cases, features, and functionality of a future toolkit for assessing reuse of digital repository materials (Kelly et al. 2018b).

The second method of data collection generated the bulk of the data from the project. The research team held three rounds of both in-person and virtual focus groups. The first

two rounds consisted of two group sessions for each focus group venue, and the last round consisted of only one session for both the in-person and virtual focus groups, for a total of 10 focus group sessions over the course of the year-long project.

From the focus group data, the research team identified specific use cases from the focus group session notes, compiled them, then synthesized the individual use cases into generalized thematic use cases (Kelly et al., 2018b). Additionally, for the purpose of provoking discussion, the research team developed working definitions that differentiated between use and reuse. Use was defined as, “Discovering and browsing objects in a digital library, often described as “clicks” or “downloads,” without knowing the specific context for this use”. Reuse was characterized as “how often and in what ways digital library materials are utilized and repurposed. In this definition, we do know the context of the use”. The research team fully expected these definitions to evolve, informed by the findings of the needs assessment, and actively solicited feedback from project participants and members of the digital repository community (Stein et al. 2018). It should be noted that as of the time of writing, the research team is still in the process of solidifying what to consider use and reuse or the differences between them.

In the next section, the data analysis process and findings for this paper are stated.

Focus Group Data

Data from the three rounds of in-person and virtual focus groups was re-analyzed in order to pull out specific examples and issues relating to content reuse of grey literature discussed by participants.

Reuse: What and Why

Throughout the focus groups, the project team sought to refine its definition of reuse. Of particular interest was the goal of expanding the definition beyond scholarly types of reuse (such as citations) to also include non-scholarly types of reuse that could be identified by altmetrics, such as web-based genealogy or news. Participants differentiated between scholarly and non-scholarly reuse of their assets, noting that both should be measured but would be weighed differently depending on institutional priorities and mission. Similarly, the impact measures of IRs, data repositories, and digitized special and archival collections might differ. In conjunction with each other, compelling reuse cases for digitized special and archival collections could potentially be used to convince administration that scholarly repositories should also be invested in at institutions that do not currently have or prioritize them.

The focus groups also served to provide additional examples of what the participants considered reuse that the project team had not previously identified in its foundational survey results. These included the following, which the research team has since extended with illustrations applicable to grey literature:

1. Repurposing metadata in aggregated consortial or membership repositories
 - o Inclusion of data in aggregated data repositories (like The UK Data Service) would serve as reuse
2. Indexing records with abstract and record location information, and selling it on portable storage media
 - o Health data could potentially serve as an example of this
3. Building a database from opened, digitized vital records and connecting that to living people
 - o Seen frequently with genealogical data, such a database, e.g., Ancestry.com, would be a type of reuse

Collecting reuse data for grey literature could serve not only to demonstrate the value of these documents, but also to help understand where gaps exist in these collections. Some potential use cases for grey literature administrators include the ability to connect reuse of grey literature collections to community or industry values; and to generate reports of collection item reuse to share internally, making all staff aware of the types of materials being reused. Gathering data on what content was not reusable in order to develop deaccessioning criteria was voiced as a potential benefit of assessing reuse of data collections. Repositories with datasets wished that they could see when Data Management

Plans referred to their institutional data repositories so they could assess deposit and eventual reuse of these datasets. One participant noted that IRs were increasingly being looked to as official repositories for what they called “data bundles including images, hyperlinks, pdf files, and databases.” Grant agencies were then using the IR to verify grant applicant claims, and focus group participants did not know how to measure the value of providing such a service.

Reuse Data Collection - Focus Group Use Cases and Examples

Methods for collecting grey literature reuse data include the following:

- citations (including Google Scholar) to grey literature in scholarly works and non-academic works (like Wikipedia)
- web analytics to trace reuse via URLs
- reports within a department and then sharing across institutional units
- mentions of grey literature on external blog posts
- historic data that's been transformed and extracted into a digital dataset
- consuming/sharing of IR objects and differences in how the same objects are consumed/shared from other repositories

Limitations to collecting reuse data, especially for grey literature, range from:

- lack of sharing of use/reuse data from aggregators and outside partners; scarcity of multi-genre grey literature aggregators
- deficit of standardized citation formats, particularly for vital and genealogical records, and incorrectly formatted citations of genealogical records
- absence of available reuse data as a result of poor search engine optimization and, therefore, diminished discovery
- separation of digitized data formats from their analog originals and any corresponding codebooks or documentation
- Outdated and colonial vocabulary in collection materials, metadata, and technical documentation

Several participants identified what they called the “problem” of electronic theses and dissertations (ETDs). Inadequate oversight or culture of attribution were indicated as problems leading to IR-hosted theses and dissertations with faulty citations, or no citations at all. These pose a challenge to digital repository staff hoping to find citations to their collections in IR-hosted ETDs.

Reuse data analysis

Once reuse data has been collected, further challenges may prevent stakeholders from analyzing the data in ways that can be beneficial to their institutions. Transforming quantitative data into meaningful stories that are suitable for consumption by a variety of audiences is difficult. Aggregated data in particular create complications for capturing individual-specific data, because aggregated data can obscure the important contextual information that reuse data emphasizes. Some aggregators, for example genealogical research databases, may have differing metadata priorities than others. Finally, legal limitations on data protection and privacy, such as the European Union General Data Protection Regulation (GDPR), may (and reasonably so) limit the ability of digital repository stakeholders to analyze reuse data without exposing individuals’ personal information.

Reuse data reporting

Institutions collecting reuse data vary in how that data is shared. For public or federal grant-receiving institutions it may be shared via periodic reports to government officials. For other organizations, reports on new services and programs developed from collected data may be sent to their administration. Lack of best practices on what to include in these reports further complicates reuse assessment, however. There are no guidelines on how to measure quantity of reuse. For example, is repeated reuse of an object indicative of more meaningful or valuable objects than those with a single reuse? Are some types of reuse “better” or more “important” than others?

Facilitating reuse assessment

Focus group participants noted a number of common or recommended practices that should be followed in collecting reuse data. The ensuing examples are of particular interest to grey literature repositories.

First, as citations (or lack thereof) have been identified as barriers to assessing reuse of grey literature, bibliometrics should, and traditionally do, include quantitative measures like the number of times an object has been formally cited. In addition, qualitative measures, such as whether the reuse crosses disciplines, should also be noted, as interdisciplinary reuse may have a greater impact overall.

Several documentation features have also been identified as recommended practices. Documentation on how to measure information in referatory services (for example, a data repository aggregator that directs users to your collection like OpenDOAR or re3data) that also enables tracing a users' journey to or from the referencing service is needed. Guidelines on reuse of specific content types with special properties, like theses and dissertations, is needed. And finally, codes of practice for meeting the needs of indigenous and/or marginalized groups that may have specific reuse requirements is essential. While there were not any specific examples of grey literature from indigenous communities discussed in the focus groups, cultural appropriation and the need for general codes of ethics for use and reuse of indigenous and marginalized cultural heritage materials emerged as an important theme that could be relevant to grey literature as well; for example, data sets of location data of sacred sites. Participants suggested some system features that could accommodate reuse measurement and, in particular, these best practices, consist of rights statements included with all digital materials; and auto-generated citations for digital library objects.

Discussion

The scope of the Measuring Reuse project was not limited to grey literature; the purview encompasses any and all materials in digital repositories hosted and curated by CHKO, inherently including digital grey literature collections. However, because the boundaries of the project are so broad, many of the findings can be related to assessing reuse of grey literature; therefore, this paper only covers use cases and findings explicitly mentioning a grey literature document type ("Grey Literature - GreySource"). The majority of focus group participants did hail from academic libraries, despite the project team's best efforts, and the majority of data we collected on measuring reuse of grey literature materials involved ETDs and research data.

Akin to the breadth of materials in digital repositories, the focus group feedback regarding features and functionality for a future toolkit supporting measuring reuse appears overwhelming at first. Ultimately, the community needs can be distilled to a handful of high level use cases. During the data analysis phase of the project, the research team reduced the many specific and sundry user stories and examples into 18 generalized use cases which were divided into three categories: Data Collection, Analysis, and Reporting; Collection Development; and Privacy, Rights Management, and Ethics (Kelly et al. 2018). The overwhelming majority of participants indicated the need for best practices on all aspects of reuse assessment: data collection, analysis, reporting, and, albeit more implied than stated, how to *enable* reuse of their digital grey literature materials.

Considering the latter, traditional publications in the Information Age typically receive some sort of unique and persistent identifier, with Digital Object Identifiers (DOIs) being the most popular. Standardized identifiers such as DOIs, Archival Resource Keys (ARKs), and Handle System identifiers (Handles) facilitate assessment work, especially for reuse. The opportunity for the grey literature field to vastly improve data gathering practices on the impact of their collections is tangible, but unless the community is able to seize it through ubiquitous, concerted assignment of persistent unique identifiers, the chance will become completely unattainable Schöpfel and Prost argue in their 2016 article. The results of the Measuring Reuse project reinforce their conclusions and further add that this is the opportune moment for tracking and evaluating transformative uses, rather than passive consumption, of grey literature works as reuse data.

The increasing pattern of making grey literature available via IRs not only aids in discovery of the materials, but can also facilitate reuse and citation, since institutional

repository software may be set up to automatically generate Handles for each deposited item.

Depositing into IRs also helps address another discovery challenge for grey literature: the lack of a major aggregator, meaning a potential source of reuse data that is available to digital repository collections aggregated via services (such as Google Scholar or SHARE³) is not available to grey literature collections. Obstacles to the creation of a unified grey literature aggregator are exacerbated by the sheer breadth of grey literature content types. Certain types of grey literature, such as research data, have their own difficulties with discovery, aggregation, and citation, even with the proliferation of standards such as DataCite. Making collections of grey literature available via institutional scholarly repositories, which are often set up with OAI-PMH servers, makes it possible for aggregators to harvest their metadata, thus feeding grey literature records to genre or thematically focused harvesting services.

Discovery systems powered by metadata aggregation are often limited in scope to certain content types, such as DataCite for research data and the Global ETD Search by the Networked Digital Library of Theses and Dissertations (NDLTD) for ETDs. They have their own challenges, typically relying on institutional membership dues, which automatically precludes participation from financially-vulnerable organizations. A full exploration of the pros and cons of metadata aggregation discovery systems is out of scope for this paper.

Findings such as these are important because the research team can further synthesize the results to develop use cases, features and functionalities for the future toolkit that are designed and included expressly for the needs of the grey literature and institutional repository communities.

Conclusion

The Measuring Reuse grant was a year-long project wherein the research team conducted a needs assessment of the digital repository, spanning across CHKO domains. The intent of gathering and prioritizing the use cases, features, and functionalities is to build a toolkit that will guide digital library professionals in assessing the reuse of the collection materials. Grey literature collections pose unique challenges in assessing reuse, some of which may be ameliorated by using IRs to collect, preserve, and provide access to grey literature.

Perhaps the biggest limitation of the project was the overrepresentation of academic libraries among the focus group and survey participants despite strident efforts to recruit from a wide variety of institutions and CHKO domains. The special difficulties facing administrators and maintainers of grey literature repositories wishing to collect reuse data point to a new area of research and development.

The next logical stage of the project is to begin development of the reuse assessment toolkit, which the project team has applied for funding to pursue. Regardless of whether or not the grant request is successful, the research team intends to move forward with creating a reuse assessment toolkit. The goal of the toolkit will be to enable any digital repository professional to gather, analyze, and report on assessment data that show how their digital materials are innovatively and transformatively reused.

³ <http://www.share-research.org/>

References

- Biagioni, Silvia Giannini Stefania, Sara Goggi, and Gabriella Pardelli. 2017. "Grey Literature Citations in the Age of Digital Repositories and Open Access." *Grey Journal (TGJ)* 13 (1): 23–31.
- Buchanan, George, and Dana McKay. 2017. "The Lowest Form of Flattery: Characterising Text Re-Use and Plagiarism Patterns in a Digital Library Corpus." In 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 1–10. <https://doi.org/10.1109/JCDL.2017.7991570>.
- Cancedda, Flavia, and Luisa De Biagi. 2017. "International Identification and 'White and Grey Literature': Identities, Retrieval, Reuse and the Certainty of Knowledge While Sharing and Connecting Information." *Grey Journal (TGJ)* 13 (1): 32–36.
- Chapman, Joyce, Jody DeRidder, Megan Hurst, Elizabeth Joan Kelly, Martha Kyrrilidou, Caroline Muglia, Genya O'Gara, Ayla Stein, Santi Thompson, Rachel Trent, Liz Woolcott, and Tao Zhang. (2015), "Surveying the landscape: use and usability assessment of digital libraries," working paper, Digital Library Federation Assessment Interest Group, User Studies Working Group, December. <https://doi.org/10.17605/OSF.IO/9NBQG>.
- Crowe, June, and J. R. Candlish. 2013. "Data Analytics: The next Big Thing in Information." *Grey Journal (TGJ)* 9 (3): 157–59.
- Ferreras-Fernández, Tránsito, Francisco García-Peñalvo, José A. Merlo-Vega, and Helena Martín-Rodero. 2016. "Providing Open Access to PhD Theses: Visibility and Citation Benefits." *Program* 50 (4): 399–416. <https://doi.org/10.1108/PROG-04-2016-0039>.
- Greene, Joseph. 2017. "Developing COUNTER Standards to Measure the Use of Open Access Resources." In *9th International Conference on Qualitative and Quantitative Methods in Libraries (QQML2017)*, Limerick, Ireland, 23-26 May 2017. <https://researchrepository.ucd.ie/handle/10197/8464>.
- "Grey Literature - GreySource, A Selection of Web-based Resources in Grey Literature." (n.d.). Retrieved October 7, 2018, from <http://www.greynet.org/greysourceindex/documenttypes.html>.
- Kelly, Elizabeth Joan, Caroline Muglia, Genya O'Gara, Ayla Stein, Santi Thompson, and Liz Woolcott. 2018a. "Measuring Reuse of Digital Objects: Preliminary Findings from the IMLS-Funded Project." In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries - JCDL '18*, 351–52. Fort Worth, Texas, USA: ACM Press. <https://doi.org/10.1145/3197026.3203871>.
- Kelly, Elizabeth Joan, Ayla Stein Kenfield, Genya O'Gara, Caroline Muglia, Santi Thompson, and Liz Woolcott. 2018b. "Setting a Foundation for Assessing Content Reuse: A White Paper From the Developing a Framework for Measuring Reuse of Digital Objects Project." <https://doi.org/10.17605/osf.io/bqjvr>.
- MacIntyre, Ross, and Hilary Jones. 2016. "IRUS-UK: Improving Understanding of the Value and Impact of Institutional Repositories." *The Serials Librarian* 70 (1–4): 100–105. <https://doi.org/10.1080/0361526X.2016.1148423>.
- Manghi, Paolo, Michele Artini, Claudio Atzori, Miriam Baglioni, Alessia1 Bardi, Sandro La Bruzzo, Michele De Bonis, et al. 2017. "OpenAIRE: Advancing Open Science." *Conference Papers: International Conference on Grey Literature* 19 (January): 107–12.
- O'Gara, Genya, Liz Woolcott, Elizabeth Joan Kelly, Caroline Muglia, Ayla Stein Kenfield, Santi Thompson. (2018), "Barriers and solutions to assessing digital library reuse: preliminary findings", *Performance Measurement and Metrics*, Vol. 19 Issue: 3, pp.130-141. <https://doi.org/10.1108/PMM-03-2018-0012>.
- Parinov, Sergey, Mikhail Kogalovsky, and Victor Lyapunov. 2014. "A Challenge of Research Outputs in GL Circuit: From Open Access to Open Use." *Grey Journal (TGJ)* 10 (2): 87–94.
- Parsons, Rebecca, and Scott Summers. 2016. "The Role of Case Studies in Effective Data Sharing, Reuse and Impact." *IASSIST Quarterly* 40 (3): 14–19.
- Schöpfel, Joachim, and Hélène Prost. 2016. "Altmetrics and Grey Literature: Perspectives and Challenges." In *Eighteenth International Conference on Grey Literature: Leveraging Diversity in Grey Literature*. GL Conference Series. New York, NY, USA: TextRelease. <https://hal.univ-lille3.fr/hal-01405443/document>.
- Stein, Ayla, Santi Thompson, Elizabeth J. Kelly, Caroline Muglia, Genya O'Gara, and Liz Woolcott. 2018. "Developing a Framework for Assessing Use and Reuse: ALA Midwinter 2018." OSF. February 28. <https://doi.org/10.17605/OSF.IO/YWH6K>.
- Warwick, Claire, Isabel Galina, Jon Rimmer, Melissa Terras, Ann Blandford, Jeremy Gow, and George Buchanan. 2009. "Documentation and the Users of Digital Resources in the Humanities." *Journal of Documentation* 65 (1): 33–57.

Librarians' Role in GAO Reports

Meg Tulloch

U.S. Government Accountability Office, GAO, United States

Problem: In a time when information is freely available, what role is there for librarians in the research process, most specifically the U.S. Government Accountability's research process that leads to GAO Reports?

Method: Librarians play several roles in the research process that involve finding, collecting, and analyzing large sets of information and/or data. The librarians at the GAO play a formal role in the research that supports the findings in GAO reports. As determined by the engagement teams in conjunction with methodologists and librarians, literature reviews with specific criteria can be used to provide evidence for findings. Literature and background searches also support the research behind the reports. This presentation/paper will discuss three types of searching—literature review, literature search, and background searching—and their role in the GAO process. It will also discuss our method for assessing Web-based sources of information.

Result: Through literature reviews, librarians can produce evidence for GAO reports. Literature and background searches also support the research behind the reports.

Abstract:

Working at the U.S. Government Accountability Office (GAO), librarians know that their research and documentation are part of an important process that can provide background information, help refine a researchable question that the agency is answering for the U.S. Congress, or support a finding in a GAO report. Carefully documented literature searches form the backbone of this process. The librarians with a team of analysts and other specialists, such as methodologists, economists, and scientists, define the criteria for the searches and the evaluation of the results. The librarian performs the searches and participates in the evaluation process often with the methodologist. The librarian is responsible for documenting searches and their returns in the Record of Research. As well, she captures the results for evaluation in a way that can be linked back to the searches. At the end of the process, the librarian has produced a Record of Research and a results document or Data Collection Instrument. Sometimes, the librarian also writes analysis about the results to support a finding.

Introduction

The United States Government Accountability Office (GAO) is an independent, nonpartisan agency within the legislative branch of government. "Our mission is to support the Congress in meeting its constitutional responsibilities and to help improve the performance and ensure the accountability of the federal government for the benefit of the American people. We provide Congress with timely information that is objective, fact-based, nonpartisan, non-ideological, fair, and balanced."¹

GAO's work aids congressional oversight in several ways—auditing agency operations, including government programs and policies; investigating allegations of governmental wrongdoing; outlining policy options for Congress to consider; and issuing legal decisions and opinions. As the supreme audit institution for the United States, one of GAO's most important roles is to investigate and understand how the federal government spends taxpayer dollars. For this reason, GAO has been nicknamed the "congressional watchdog."

GAO is headquartered in Washington, D.C. and has eleven field offices around the country to support its work. GAO has over three thousand employees; the majority of whom work for the mission teams who create GAO reports. Thirteen of the fourteen teams audit specific subject areas such as defense, contracting, health care, education, homeland security, physical infrastructure, natural resources, financial management, information technology,

¹ "What GAO Is." GAO Web site, 3 Jan. 2019, <https://www.gao.gov/about/what-gao-is/>.

international affairs, and many other topics.² In fiscal year (FY) 2018³, GAO produced six hundred thirty-three reports as part of its auditing function.⁴

GAO librarians work within the fourteenth mission team, Applied Research and Methods (ARM). ARM mainly supports the thirteen other teams in creating GAO reports.

ARM offers expertise in many areas including cost analysis, engagement design, economics, data analysis, evaluation, library research and literature review, science, statistics, surveys, technology, engineering, and IT security. In addition, five technical chiefs provide expertise in statistics, economics, technology, actuarial science, accounting, and cutting-edge science.⁵

Library Services provides GAO with research products and services, with access to accurate and reliable information, and with proficiency in organizing and preserving information in all formats. Librarians provide direct support to the mission teams responsible for GAO reports, including as stakeholders. Librarians work with the teams to develop research and search strategies and the results of those strategies. They also purchase access to accurate and reliable information and data sources and provide access to information through interlibrary loan and document delivery. Librarians work with other ARM Specialists to develop guidance on topics such as literature searches and reviews and information reliability.

In FY2018 GAO librarians conducted two hundred nineteen literature searches and fifty-four background searches in support of GAO reports.⁶ This paper will discuss the librarians' process and role in creating GAO reports.

Literature Searches and Reviews Guidance

GAO has several types of internal guidance that applies to literature searches and reviews. For the engagement process GAO staff follows the *Electronic Assistance Guide for Leading Engagements* (EAGLE II). "It contains GAO guidance, consistent with GAO policies and protocols and GAGAS [GAO's Government Auditing Standards] requirements, and is a definitive source for implementing GAO's Engagement Management Process."⁷ Within EAGLE II under "Plan & Proposed Design," GAO staff can refer to a section on obtaining background information (2.3.2). This section says that, "an engagement team conducts and documents a sufficient review of existing literature or other appropriate research to assure themselves that they understand the nature and background of the program or agency under review."

The section goes on to describe the literature searches more specifically. It says that "the engagement team conducts literature searches, or has ARM Research Librarians conduct the searches for them, as appropriate. Literature searches can provide contextual sophistication or can be the basis of a GAO product, such as an evaluation synthesis."⁸ ARM also has three internal guidance papers-- *Using Studies Conducted by Outside Researchers for GAO Engagements*, *Guidelines for Assessing Web-based Sources of Information*, and *Sample Objectives, Scope, and Methodology (OSM) Language for Literature Searches and Reviews*. The first guidance paper describes the literature search process as follows:

The literature search identifies studies that are relevant to the researchable questions and that meet specified criteria. ARM specialists and research librarians can help determine what the criteria should be in order to conduct a comprehensive yet specific literature search.⁹

² "Our Teams." *GAO Web site*, 3 Jan. 2019, <https://www.gao.gov/about/careers/our-teams/>.

³ The fiscal year for the U.S. Government is October 1st through September 30th. FY2018 is from Oct. 1, 2017 through September 30th, 2018.

⁴ "About GAO, Performance." *GAO Web site*, 4 Jan. 2019, <https://www.gao.gov/about/what-gao-is/performance>.

⁵ "ARM Overview." *GAO Intranet*, 3 Jan. 2019, ARM Home page (internal web site).

⁶ "ITSM Library Subject Statistics Between FY2015 and FY2018." *ITSM Reports*, Generated 1 Nov. 2018, (internal document).

⁷ "EAGLE II Overview." *GAO Intranet*, 3 Jan. 2019, EAGLE II Home page (internal web site).

⁸ "EAGLE II "Planning & Proposal." *GAO Intranet*, 3 Jan. 2019, EAGLE II Home page (internal web site).

⁹ "Using Studies Conducted by Outside Researchers for GAO Engagements." *ARM Guidance*, p. 4, Accessed 3 Jan. 2019, ARM Guidance Web site (internal web site).

Guidelines for Assessing Web-based Sources of Information helps analysts determine the sources that might be used as part of a background or literature search are authentic, authoritative, and reliable (see Appendix 4 for definitions). For information to be used as evidence, the information must be authentic and authoritative, and almost always it must also meet the test for reliability. As the guidance says,

Reliability Assessment

Information from U.S. government and military Web sites should not be assumed to be reliable even if the source is authentic and authoritative. The information could be inaccurate, outdated, or incomplete. Depending on the nature of the information and the auditors' use of the information in relation to the audit's objectives, findings, and conclusions, the auditors may need to corroborate the information. The auditor should always use professional judgment when assessing the value of information for use as GAO evidence.¹⁰

The following are examples of web-based content that is considered low risk and generally usable as evidence: Congressional Budget Office Reports, Code of Federal Regulation or U.S. Code from the Government Publishing Office or an approved third party such as Westlaw or LexisNexis, peer-reviewed/scholarly articles, professional and technical standards from the authoring web site, U.S. federal, state, and local government web sites.

Sample Objectives, Scope, and Methodology (OSM) Language for Literature Searches and Reviews provides examples and recommendations on how a literature search and review are addressed in the formal written part of a GAO report. It suggests that five key elements should be included to adequately address how the literature review was conducted and how the results were reviewed. The five elements are 1) how the articles or reports were identified 2) a list of the databases searched 3) criteria used to select articles for the literature review 4) number of articles identified 5) description of the methodical review.¹¹ For the second element, the librarians provide detailed search strings for the databases searched, which are entered in the Record of Research as is shown in Appendix 2. When a literature review provides evidence for an objective a librarian may be asked to write the OSM for the literature review or review what an analyst or methodologist has written (see Appendix 5).

Literature Searches and Reviews Tools and Process

In order to follow the EAGLE II and ARM guidance, GAO librarians developed different tools to document the research process. The Record of Research is filled out by the mission team asking for the literature search. The team outlines their information needs, whether the information or studies will be used as background information, evidence or to answer a researchable question, and documents their own searching for information. The team also lists limiting factors, such as date range and desired document types (see Appendix 1). The librarian then records her searches and the number of results into the Record of Research (see Appendix 2). For a background search, the librarian will use a select set of databases for searching. For a literature review, the librarian will be as comprehensive as possible in her searching. In either case, the librarian will often send an initial set of results to make sure that she has understood the research request. Once she has feedback, the librarian may continue searching. For a literature review or a complex background search, the searching process can be iterative, with several exchanges between the librarian and the analysts. The results from the searching are exported to RefWorks, de-duped, organized, and then exported as citations with abstracts to a results document or Data Collection Instrument (see Appendix 3) for analysis by the librarian, ARM Specialist, or team analyst. The analysis is

¹⁰ "Guidelines for Assessing Web-based Sources of Information." *ARM Guidance*, April 2017, p. 7, Accessed 4 Jan. 2019, ARM Guidance Web site (internal web site).

¹¹ "Sample Objectives, Scope, and Methodology (OSM) Language for Literature Searches and Reviews." *ARM Guidance*, March 2011, p. 1-2. Accessed 3 Jan. 2019, ARM Guidance Web site (internal web site).

recorded in the Data Collection Instrument (DCI) to include evaluation of the materials using pre-determined criteria.

The Library Services members will also help the team find full-text of any or all of the results for the background or literature search. This process can involve finding full-text in existing library materials, borrowing or requesting full-text from outside organizations, or purchasing the materials or data. The librarians are constantly evaluating materials to make sure that the right mix of library materials are readily available, easily accessible, and licensed so that the materials can be saved with the engagement workpapers as evidence in the audit process.

As noted earlier, librarians can be involved in writing or reviewing the OSM. The sample in Appendix 5 was written by a librarian.

Conclusion

Working at the U.S. Government Accountability Office (GAO), librarians know that their research and documentation are part of an important process that can provide background information, help refine a researchable question that the agency is answering for the U.S. Congress, or support a finding in a GAO report. As members of the ARM team, they can be involved with any report that the mission teams are working on. Librarians were listed as contributors in twenty GAO reports in FY2018.

Appendix 1 : Sample Record of Research, Searching Details

2

2. Purpose**2.1 Intended Use of Literature Search (check all that apply)**

Note: There is ARM Guidance on [Literature Searches and Studies Conducted by Others](#)

If the intended use of the literature search changes over the course of the job, you and your librarian may make changes to this form documenting updated searches.

	Initial background search to help understand engagement topic and/or to provide context for background section of report
	Provide a source of evidence for one or more objectives/researchable questions (examples include open source projects and literature searches conducted to identify articles for a review of studies) • Please provide details on how the search results will be used _____ Results will be systematically presented as a section in the report _____ Results will be referenced as contextual information in the report _____ Results will be used for identifying and/or developing criteria or identifying examples or case studies for further examination Note: Your methodologist can help you identify how the literature search can fit into your methodology.
	Identify subject matter experts or other stakeholders for interviews
	Other (please explain):

3. Search Preferences**3.1 Types of materials desired (check all that apply):**

☐	Scholarly/Peer Reviewed Material	☐	Government Reports
☐	Conference Papers	☐	General News
☐	Dissertations	☐	Trade and/or Industry Articles
☐	Working papers	☐	Association/Nonprofit/Think tank Publications
☐	Books	☐	Legislative materials (Congressional transcripts, bills, etc.)

NOTE: Contact your General Counsel stakeholders regarding legal material.

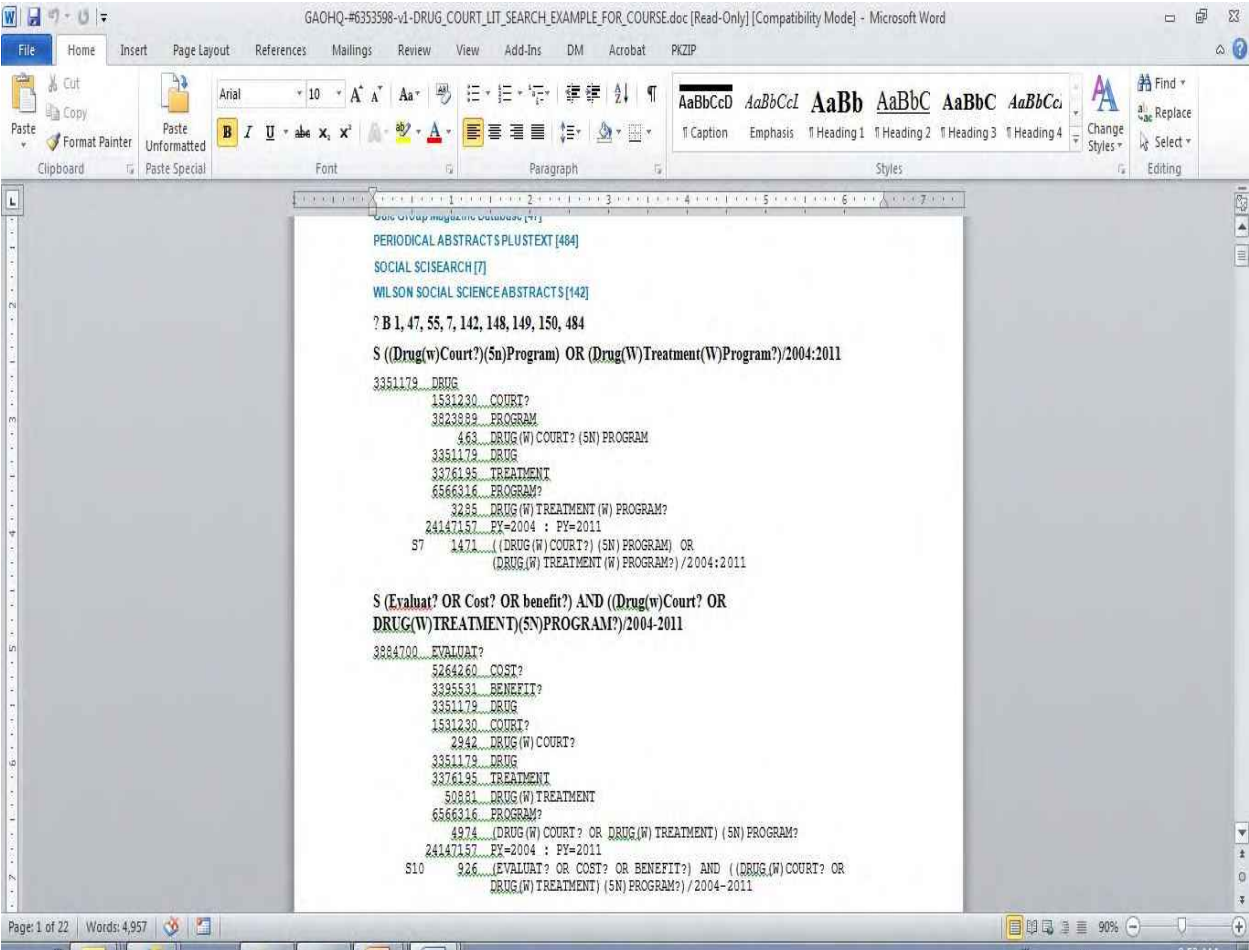
3.2 Publication date range:

Note: Discuss with your librarian if you're not sure how far to go back.

☐	Previous 5 years
☐	Previous 10 years
☐	Other (please specify):

Researcher's Name:

Appendix 2 : Sample Search Strings



Appendix 3 : Sample Data Collection Instrument (DCI)

ALL_STAFF-#1779998-v3-101120_ROA_-_ABSTRACT_REVIEW_-_FOR_FINDINGS.XLSX [Read-Only] - Microsoft Excel															
BN3															
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Reference Type	Authors, Primary	Title Primary	Periodical Full	Pub Year	Keywords	Abstract	First reviewer Determination (Y/N) (ACL)	ACL Notes	second reviewer Determination (Y/N) (GN or CFK)	GN/CFK Notes	Final Determination (Y/N)	Final Notes	Date requested from librarian	DM link to article	
Journal Article	Allogoskou K, van d	O-0907 The Impact Of Breastfeeding On The Incidence And Severity Of Neonatal Abstinence Syndrome	Archives of Diseases in Childhood	2014	Immunology Abstracts; Epidemics; Pediatrics; Opioids; Breastfeeding; Neonates; Child; us; Abstracts; Hospital; Effects; F 06.905; Vaccines	Background and single light of the current epidemic in the abuse of opioids, a major increase in neonates with neonatal abstinence syndrome (NAS) is likely. Incorporation of breastfeeding as a first pillar of treatment of NAS seems appropriate. We aimed to quantify the impact of breastfeeding on the incidence and severity of NAS. Methods: Pooling of published NAS cohorts, with specific emphasis on the impact of breastfeeding on the incidence (per 1000 opioid administration) and duration (duration of opioid treatment, duration of hospitalization) of NAS. Results: Three studies [1-3] were retrieved and resulted in a pooled dataset of 400 neonates (218 breastfed, 54.5%). There is a significant reduction in NAS (54 vs 17%, number needed to treat 5-6). The same trends are observed when the duration of opioid treatment (difference -18 to -23 days) or the length of hospital stay (difference -4 to -10 days) are considered. Conclusions: Breastfeeding is associated with a clinical significant reduction on both the incidence and the duration of NAS in opioid exposure newborns. Incorporation of breastfeeding as a first pillar of treatment for relieving the NAS symptoms seems to be a very natural, and effective way of addressing this problem. Abdulkadir ME et al. Pediatrics 2006;vol:118, no: 4, pp: 811-816. Valle-Strand GK et al. Acta Paediatr 2013	Y		Y		Y				
23	Journal Article	Allogoskou K, van d	O-0907 The Impact Of Breastfeeding On The Incidence And Severity Of Neonatal Abstinence Syndrome	Archives of Diseases in Childhood	2014	uterine opioid were delivered preterm. There is currently no neonatal abstinence syndrome (NAS) scoring tool known to accurately evaluate preterm opioid-exposed infants. This can lead to difficulties in titrating pharmacotherapy in this population. Purpose: To describe NAS symptoms in preterm opioid-exposed infants in comparison with matched full-term controls. Methods: This was a retrospective cohort study from a single tertiary care center of methadone-exposed infants born between 2006 and 2010. Using modified Finnegan scale scores recorded every 3 to 4 hours beginning at 6 hours of life until 24 to 48 hours after medication discontinuation, NAS symptoms were compared between 45 preterm infants and 49 full-term matched controls. Concurrent neonatal medical diagnoses were also compared.	N		Maybe?		Y				
B.RefList of Abstracts Round2															
C.Articles (from abstracts)															
D.Articles (from others)															
E.Lit review stats															

Appendix 4 : Definitions of Authentic, Authoritative, and Reliable

From GAO ARM Guidance Paper: *Guidelines for Assessing Web-based Sources of Information*

Authentic: The accessed Internet source is, in fact, what it declares to be or the Internet source's provenance can be verified.

Authoritative: The Internet source is the official site for an organization (public or private sector) or is recognized as an legitimate source for information on the subject matter.

Reliable: The information presented in the internet source is sufficiently unbiased and reasonably complete, timely, and accurate with regards to the analyst's intended purposes; as a result, it is associated with a low risk of the analyst making incorrect or improper conclusions based on that information (GAGAS 6.56-6.72).¹²

GAGAS = generally accepted government auditing standards (GAGAS) or GAO's Government Auditing Standards (Yellow Book)

Appendix 5 : Sample OSM¹³

Review of Literature on Child Care Subsidies

To determine what is known about the impact of child care subsidies on employment, we conducted a literature search for studies that analyzed relationships between child care subsidies or changes in child care costs and employment outcomes. To identify existing studies from peer-reviewed journals, we conducted searches of various databases, such as EconLit, ProQuest, PolicyFile, and Social SciSearch. We also asked all of the external researchers that we interviewed to recommend additional studies. From these sources, we identified 31 studies that appeared in peer-reviewed journals between 1995 and August 2009 and were relevant to our research objective on the effect of child care subsidies on employment outcomes. We performed these searches and identified articles from June 2009 to October 2009.

To assess the methodological quality of the selected studies, we obtained information about each study being evaluated and about the features of the evaluation methodology. We based our data collection and assessments on generally accepted social science standards. We conducted an extensive literature review, examined summary level information about each piece of literature, and then from this review, identified articles that were germane to our report. We then evaluated the methods used in the research, eliminated some research if we felt the methods were not appropriate or rigorous, and then summarized the research findings. In addition, for articles directly cited in the report, we performed an initial in-depth review of the findings and methods, and then a GAO economist performed a secondary review and confirmed our reported analysis of the finding. As a result, the 31 studies that we selected for our review met our criteria for methodological quality. We supplemented our synthesis by interviewing four of these studies' authors. We also conducted an interview with an official at the Office of Planning, Research and Evaluation within the Administration for Children and Families.

¹² United States, Government Accountability Office. *Government Auditing Standards*, Government Accountability Office, 2011, p. 150-156. Accessed 4 Jan 2019. <https://www.gao.gov/products/GAO-12-331G>.

¹³ United States, Government Accountability Office. *Child Care: Multiple Factors Could Have Contributed to the Recent Decline in the Number of Children Whose Families Receive Subsidies*, Government Accountability Office, May 5, 2010, p. 37-38. Accessed 4 Jan 2019. <https://www.gao.gov/products/GAO-10-344>.

Library, Information Science & Technology AbstractsTM with Full Text

Available via EBSCOhost[®]

The definitive professional information
resource designed for librarians and
information specialists...

*Library, Information Science & Technology AbstractsTM
with Full Text* is an indispensable tool for librarians
looking to stay current in this rapidly evolving field.

Comprehensive content includes:

- Full text for more than 270 journals and nearly 20 monographs
- Indexing for more than 550 core journals, 50 priority journals and nearly 125 selective journals
- Includes books, research reports, proceedings and author profiles
- Access to 6,800 terms from reference thesauri
- Coverage extends back as far as the mid-1960s

Subject coverage includes:

- Bibliometrics
- Cataloging
- Classification
- Information Management
- Librarianship
- Online Information Retrieval
- And much more...

Contact EBSCO Publishing to learn more about
Library, Information Science & Technology AbstractsTM with Full Text,
or to request a free trial.

Phone: 800.653.2726

Email: request@ebscohost.com

www.ebscohost.com





Published electronic media are becoming Grey

Yui Kumazaki, Satoru Suzuki, Masashi Kanazawa, Katsuhiko Kunii,
Minoru Yonezawa, and Keizo Itabashi
Japan Atomic Energy Agency, JAEA, Japan

Abstract

The library of the Japan Atomic Energy Agency (JAEA) focuses on collecting conference proceedings and technical reports which are very important and referred to as materials for our users in the field of nuclear science and technology. Increasing numbers of cases are published in electronic media such as CD/DVD-ROM and flash memory. Meanwhile, problems exist as to how to manage such electronic media in a library such as long-term preservation and permanent access, because such electronic media are updated or changed in terms of specifications. For this reason, it is well known that respective electronic media would be unavailable. Therefore it's necessary to have appropriate playback equipment and software corresponding to the electronic media. The lifetime of electronic media is much shorter compared with paper and microform. The authors would like to present the JAEA Library's current activities on long-term preservation and use of electronic media. They conclude the fact that even published electronic media are becoming grey literature in certain environment.

1. Introduction

Japan Atomic Energy Agency (hereinafter, this is called "JAEA") (ref 1) is a comprehensive R&D institute dedicated to nuclear energy in Japan. The JAEA Library is one of the largest nuclear information centers in Japan. It provides technical information services to researchers and engineers in the nuclear science and technology.

The primary roles of the Library (ref 2) are:

- To collect and make academic information in nuclear field available for library users
- To repose and disseminate documented results of research and development activities

The following practical work of the Library is detailed:

- General work of library (acquisition and technical service, network system, photocopy service, interlibrary loan and so on)
- Dissemination of the R&D results/achievements via its own institutional repository, JAEA Originated Papers Searching System (JOPSS)
- Publishing the JAEA Reports (technical reports of JAEA) and the "JAEA R&D Review" (annual publication that introduces JAEA's R&D results/achievements in a form of digest)
- Collection and dissemination of Fukushima Accident information via Fukushima Nuclear Accident Archive (FNAA)
- Taking part in the International Nuclear Information System (INIS) program, academic information sharing engaged in International Nuclear Library Network (INLN)

In this article, the electronic media are the ones ranged and limited to materials electronically published, recorded and passed into electronic media tangible, in accordance with the definition found in the survey report issued by the National Diet Library, Japan (NDL) (ref 3).

2. Facts of materials of electronic media at JAEA Library

2.1. Collection of electronic media

As of today, the JAEA Library includes some 50,000 books, 2,000 journal titles, 800,000 research and/or technical reports and about 365,000 US nuclear licensing documents of US Nuclear Regulatory Commission (DOCKET). The breakdown of the Library collections are as follows (Table 1, Fig. 1).

Table 1: JAEA Library Collections

	Print	Microforms	Electronic media	Online	Total
Books	46,667	0	1,201	2,285	50,153
Technical reports	74,783	705,035	4,469	0	784,287
DOCKET	0	365,502	0	0	365,502
	121,450	1,070,537	5,670	2,285	1,199,942

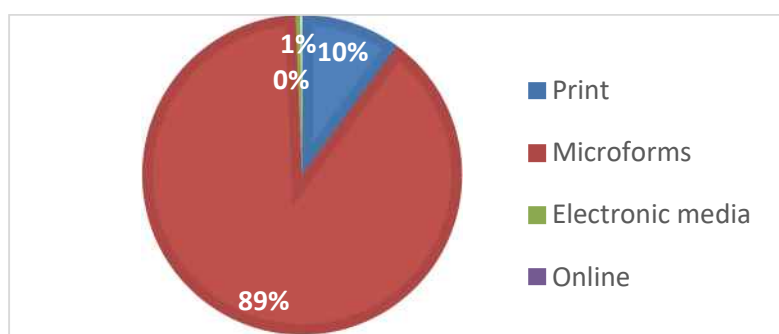


Fig. 1: JAEA Library Collections

The proceedings specific to nuclear science take priority in the library's collections. Notable is the collection of the grey literature having strived over decades since the establishment of the library. We have investigated the type of JAEA achievement announcements over the past three years. Of JAEA's achievement announcements, 76% were proceedings of papers with presentations. From this point, it can be said that the proceedings are important for the field of nuclear science and technology (Fig. 2).

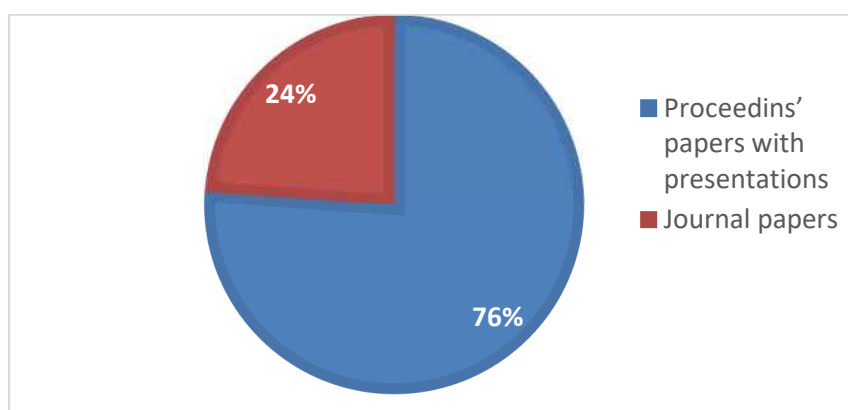


Fig. 2: Submission of JAEA R&D results over the latest three years (Apr 2015 to Mar 2018)

Furthermore, it is important for the Library to stay and collect proceedings as many proceedings as possible. Occasionally, for example, the library looks for JAEA personnel who took part in conferences and requests to grant the proceedings to the library to enrich the collection, courtesy of them. That is why the Library can often satisfy users' requests, for example, such requests as those via Document Delivery Service from other libraries in Japan. Proceedings are increasingly delivered and distributed at conferences in the form of electronic media such as CD/DVD-ROM, flash memory (Fig. 3). And they can even be downloaded by participants from websites controlled by ID/PW.



Fig. 3: Changing publication form of conference proceedings from paper to electronic media

The collection status of books and conference proceedings by the Library by such media are, shown in Table 2. The books in Table 1 are separated with books and conference records.

Table 2(a): Breakdown and purchase time of electronic media in books of JAEA Library

	CD/ DVD	Flash Memory	Floppy Disk	Video Cassette	Magneto -Optical Disc	Total
1990s	2	0	0	38	0	40
2000s	132	0	0	18	1	151
2010s	45	2	5	0	0	462
Total	589	2	5	56	1	653

Table 2(b): Breakdown and purchase time of electronic media in proceedings of JAEA Library

	CD/ DVD	Flash Memory	Floppy Disk	Video Cassette	EBook reader*	Total
1990s	16	0	0	0	0	16
2000s	194	0	0	0	0	194
2010s	303	33	0	0	1	337
Total	513	33	0	0	1	547

*EBook reader: Proceeding stored in TRECSTORE ebook reader 3.0

3. Current situation and issues

3.1. Current situation of the JAEA Library

One of the serious and often urgent concerns for libraries would be coping with the unavailability of documents held in the form of electronic media. It would be neither simple nor cost effective for libraries to make and keep the hardware and software corresponding with such media available for reading at any time. These items are not only of case-by-case but of year-by year following updates to procedures or routine maintenance. Meanwhile, alongside such inconveniences, there is the other concern that the Library cannot take an enough time to maintain electronic media and take measures such as emulation. Therefore, lifetime of electronic media is much shorter than paper and microform.

Below are some challenges the JAEA Library has faced so far:

- most of books and technical reports, electronic medium-based, collected and held since 1990's cannot be read out at present without corresponding and suitable software
- whereas the Library has held old personal computers (PCs) with old versions of Operating Systems (OSs) like Windows 95 to read out the respective electronic media. Such PCs with OSs are mostly limited to be used to applications. They cannot be connected to any electronic devices from of the current PCs related under the in-house security rules

- some of articles of proceedings are displayed on suitable monitors and fortunately might have been printed out if properly connected to usable printers. However, some printers are too old, even worse without toners, and do not have the corresponding electronic format needed for use (Fig. 4)¹



Fig. 4: Windows 95 machines and their printers kept in JAEA Library, purchased before 2000

In this regard, the Library has experienced difficulty managing electronic media, to collect, hold, and provide accessibility in the long term. Improving the collection of electronic medium documents is necessary at the Library in near future.

The Library attempts to share important and useful information regarding how to deal with electronic media stemming from its own practical experiences with other libraries worldwide, as likely measures we could take even if not definite. This article guides the following two topics of discussion:

1. associated issues having arisen for years at the Library in practice,
2. some measures having been taken and sought to be taken afterward against such issues. It is believed such concerns would be not only of ours but yours as well and would be shared, as a matter of “grey literature being coming, anytime.”

3.2. Investigating on long-term preservation and use of electronic media by NDL, Japan

National Diet Library of Japan (NDL, Japan) has implemented surveys on how to effectively preserve electronic media in the long term since several years ago (ref 3), which they mention in their 2018 report. It also focuses on how to deal with media packaged.

In the 2018 report, NDL, Japan has investigated the interoperability by migration from such media to the other media or something like databases. The experiences and findings remain as follows:

- such migration is usable and comparatively easier to be conducted
- however, for electronic media like CD/DVD-ROM, it is necessary to be applied with applications to read out. Electronic configurations applying virtual machine like VMware (ref 4), should be installed and used. In addition, further investigation is necessary for how to apply the methods of migration.

4. Measures for using electronic media materials at JAEA Library

4.1. Policies toward necessary measures

In order to ensure preservation and permanent access to all the electronic media of the collection of the JAEA Library, the following surveys to check such media and seek for measures for those:

- Could the media be read out or not?
- What are the conditions regarding hardware and software to have the media read out?
- Are effective methods of migration of the media for long-term preservation foreseen?

On the other hand, resources like manpower, budgets, and documents/data management skills are insufficient at the Library. Taking account of this practical and uneasy situation, procedures prioritized as the measures are:

¹ In Japan by the domestic law of copyright and its guideline, libraries are neither allowed to electronically duplicate materials nor to electronically provide them to users. Therefore, the user who wants to copy the proceedings of the electronic medium needs to print the relevant part and it is necessary for my library engaged in Japan to use the printer.

1. to list proceedings before they are requested by library users or through photocopy service and to check if they can be read out or not,
2. to seek for dealing with proceedings of electronic media which are readable out only by Windows 95/98 as of today,
3. to deal with proceedings of electronic media which are not able to be read out as of today.

In view of the above, we decided to take countermeasures against this situation. Our final goal should not be to migrate the electronic media, but should be to figure out the ways and measures for long term preservation and accessibility. Preserving the media is not the challenging but rather making them available for library users, i.e., to make them readable out by any measures, which we have not explored.

4.2. Measures via OSs

The Library has possessed a PC compatible with Windows 95 and made it available as of today to make PDF files available to be read out, for most of electronic files in proceedings published in the form of CD/DVD-ROM and of PDF files. Doing so highly depends on the versions of Adobe Acrobat or Adobe Acrobat Reader.

However, we have problems with this as well:

- The PC with Windows 95 was purchased before 2000, and the state is unstable and incompatible. For example, the touch pad is temporarily unusable.
- It cannot print articles because it does not have toner for the printer connected to the Windows 95 PC is produced anymore.

In this regard, the solution to succeed in reading out some electronic files has been identified. To avoid challenges arising from deeply limited use of PC compatible with Windows 95, its counter-measure is to provide a PC compatible with Windows 10 for 32 bit with NTVDM (NT Virtual DOS Machine)(Fig. 5).

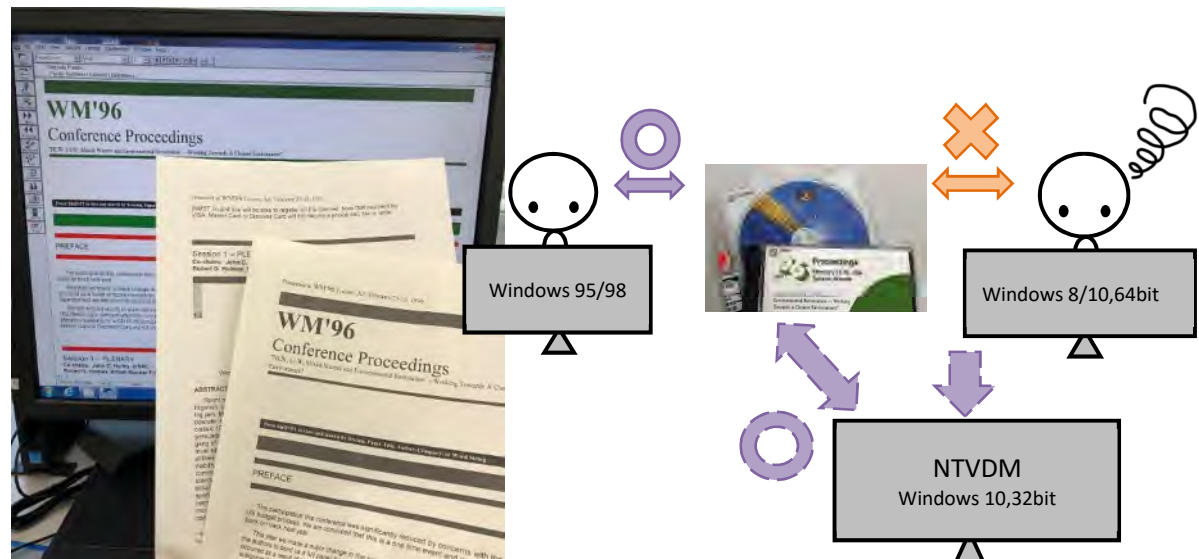


Fig. 5: Read and print proceedings using Windows 10 for 32 bit

The following, as references, are trials with the consideration, sought for and worked out:

- the fault to let the respective programs run via the present PCs occurs due to its incompatibility with the 16 bit based ordering commands for the respective programs. The programs were previously run via PCs with Windows 95/98
- NTVDM can make such programs in days of Windows 95 usable by activating the virtual 86 mode. This mode enables the orders of old CPU to run via the present CPU. This program is available on Windows 10 for 32 bit but not on Windows 10 for 64 bit.

Considering this, we decided to purchase Windows 10 for 32 bit newly and enable the playback of the old version file. However, not all Windows 10 for 32 bit can play back all the past conference proceedings. We are investigating now.

4.3. Example of migration - migration simply and easily -

Here is one example for easily and easily migrating electronic media stored in an electronic book reader made in Germany, "TRECSTORE (ref 5) ebook reader 3.0" at the Library. It was either difficult for library users to find out some documents of them by the reader due to the proceeding with more than 1000 pages, nor was it impossible for them to connect to printers and print out due to the reader like "stand-alone". A sort of simple migration was done for it, by reading out the electronic file via the reader and saved in a DVD-ROM for use.

5. Closing remarks

We believe that the most important and necessary is not only to distribute/share proceedings to/with participants but also to publish them to make them accessible for the potential users through libraries. Proceedings distributed to participants in the form of electronic media are one area of concern of grey literature. Those are unlikely to be available for potential users for a long time and will disappear, to some extent.

The JAEA Library has a serious concern as entitled "Published materials are becoming Grey". We have a wish "Published materials should not become Grey." and would like to ask conference organizers or academic associations who possess the electronic media of proceedings to recognize the concern. To avoid such situation loomed up, becoming Grey, it is necessary to take action requesting the academic associations to do as follows:

- Not only distributing/sharing the proceedings in the form of electronic media between participants but publishing those in print and/or online accessible for users in general
- In addition, having agreement on the publishing with authors regarding copyright
- Furthermore possibly being with DOI in the case of online publishing

It is hoped, acknowledging such request, little by little, the academic associations would change their standpoints on the concern with "becoming Grey".

Proceedings in the form of CD/DVD-ROM and flash memory are easy to carry around, easy to search, and very convenient for users. Whether to publish or distribute those only depends on the conference's scope, in which any users cannot be involved, we, libraries neither. Therefore it's true those convenience and scope of the disclosure are not matters we deny adversely.

The Library persists to improve and enhance the proceedings' collection in particular the skills to read out as measures against the issues and difficulty associated with, mentioned above. The hope is to be able to share important and useful information stemming from practical experiences of the Library with other libraries worldwide, believing such concerns would be not only of ours but shared as well.

References

- 1) Japan Atomic Energy Agency: <https://www.jaea.go.jp/english/>, (accessed Dec. 27th 2018).
- 2) Library of Japan Atomic Energy Agency: <https://tenkai.jaea.go.jp/english/library/> (accessed Dec. 27th 2018).
- 3) National Diet Library, Japan : <http://www.ndl.go.jp/jp/preservation/dlib/index.html> (in Japanese), (accessed Dec. 27th 2018).
- 4) VMware: <https://www.vmware.com/jp.html> (in Japanese), (accessed Dec. 12th 2018).
- 5) TRECSTORE: <http://www.trekstor.de/home-en.html>, (accessed Dec. 27th 2018).



Semantic Query Analysis from the Global Science Gateway

Sara Goggi, Gabriella Pardelli, Roberto Bartolini, and Monica Monachini, ILC-CNR, Italy
Stefania Biagioni and Carlo Carlesi, ISTI-CNR, Italy

1. Introduction

Nowadays web portals play an essential role in searching and retrieving information in the several fields of knowledge: they are ever more technologically advanced and designed for supporting the storage of a huge amount of information in natural language originating from the queries launched by users worldwide.

Given this scenario, we focused on building a corpus constituted by the query logs registered by the GreyGuide: *Repository and Portal to Good Practices and Resources in Grey Literature*¹ and received by the *WorldWideScience.org*² (*The Global Science Gateway*) portal: the aim is to retrieve information related to social media which as of today represent a considerable source of data more and more widely used for research ends.

The following quotation by Bronson gives a good description of the *WorldWideScience* search engine:

The database is available at <<http://worldwidescience.org/>>. It is based on a similar gateway, Science.gov, which is the major path to U.S. government science information, as it pulls together Web-based resources from various agencies. The information in the database is intended to be of high quality and authority, as well as the most current available from the participating countries in the Alliance, so users will find that the results will be more refined than those from a general search of Google. It covers the fields of medicine, agriculture, the environment, and energy, as well as basic sciences. Most of the information may be obtained free of charge (the database itself may be used free of charge) and is considered "open domain." As of this writing, there are about 60 countries participating in WorldWideScience.org, providing access to 50+databases and information portals. Not all content is in English. (Bronson, 2009)

While the World Wide Web keeps on growing, the development of ever more sophisticated search tools within the universe of public and private infrastructures allows to optimize the users' approach to technology: a new generation of web users – such as the so-called "millennials" – is exponentially connected for getting access and share information on social networks.

In 2005 Bettelle states: "Most searchers –and linguists may be no exception – are instead incredibly lazy, generally typing in a few words and expecting the engine to bring back perfect results, ignoring that it is only the act of offering more data in the query that often dramatically improves the results" (Battelle 2005: 23-25). Nowadays it is possible to retrieve knowledge from the web also by means of query logs, linguistic elements which are useful for monitoring a wide range of information.

This Corpus – called GSGCorpus (Global Science Gateway Corpus) – has been processed with Natural Language Processing (NLP) tools: it talks the *web language*, made up of terms originating from the most various domains and styles. The analysis mainly concentrates on the semantics of the queries received from the portal clients: it is a process of information retrieval from a rich digital catalogue whose language is dynamic, is evolving and follows – as well as reflects – the cultural changes of our modern society.

¹<http://greyguide.isti.cnr.it/>

GreyGuide is the online forum and repository of good practices and resources in Grey Literature. It was created - and is now edited - by GreyNet International (content provider) and ISTI-CNR, Pisa Italy (service provider): its launch was in December 2013 and since then GreyGuide provides a unique resource in the field of grey literature, which was long awaited and responds to the information needs of a diverse, international grey literature community. GreyNet International is one of the WorldWideScience Associate Members <https://worldwidescience.org/alliancemembers.html>.

² <https://worldwidescience.org/>

It is a global science gateway comprised of national and international scientific databases and portals. WorldWideScience.org accelerates scientific discovery and progress by providing one-stop searching of databases from around the world. WorldWideScience.org is maintained by the U.S. Department of Energy's Office of Scientific and Technical Information as the Operating Agent for the WorldWideScience Alliance.

2. Methods and Tools

This project includes eight months of query logs³ registered between July 2017 and February 2018 for a total of 445,827 queries.

The preliminary phase has essentially dealt with the huge amount of non-relevant information, the so-called *noise* which had to be filtered and eliminated.

Therefore, in order to analyze the available information a considerable pre-processing on four levels has been carried out:

- at the first level, the set of queries has been cleaned: duplicates, alphanumeric strings, strange graphical forms, IP addresses, etc. have been eliminated;
- at the second level, filters have been added and alphabetical order inserted for having a first picture of the contents of these queries;
- the third step consisted of several trials for choosing the focus;
- lastly, natural language processing (NLP) tools have been applied for processing the information and building the sample.

Since the corpus is made up of queries collected in only eight months and the cleaning process reduced them consistently, as a result the final is relatively small. In addition, only the queries in English have been registered while those in other languages have been eliminated (there are a few in French, Spanish, Italian, Portuguese, Polish, Albanese, Galician, Corsican, and so on).

Coming to the NLP analysis, the software team has decided to follow these two steps:

1. free information extraction: it measured the frequency of all the words contained in the corpus. This preliminary investigation provided us with the whole scenario of the lexical variety of the queries and allowed us to focus on a set of terms from which we built a micro-ontology with meaningful terms relating to the queries launched on the portal;
2. ontology-based extraction: the extraction has been performed again using this micro-ontology which has been essentially used for enriching the domain. In this way, the search engine retrieved each single occurrence of those terms (monograms, bigrams, trigrams) which can be found starting from the ontology.

In the following paragraph, it is described the process of information retrieval from a rich digital catalogue of queries.

2.1 NLP Analysis

The free information extraction from the GSGCorpus measures the frequency of the words contained in the corpus; examines the lexical variety of the queries and finally focuses on a set of terms to build a micro-ontology. This extraction was preceded by the cleaning process already described above and of which some examples are listed here:

- Graphical variants:
 - (*micro-fluidic* "micro fluidic*" microfluidic*)
 - (*trypodes basapyrazilique pdf, trypodes Base pyrazilique PDF*)
 - (*Adam Smith, Adams, SM.*)
- Queries in languages other than English:
 - (*alimentos proteicos, Alteracion proteica, Entrainement isometrique, 1. Peste-des-petits-ruminants virus fusion protein F) gene, complete cds, Trypanosoma cruzi: contribution Ã l'identification de substances chimiques et naturelles ayant une activitÃ© trypanocide, alfa1 fetoproteina, hrabanus maurus lehrer, abt und bischof, anabolismo proteina, AND Tanztheater Pina Bausch: Spiegel Gesellschaft, poliù tico, Espanña*)
- Duplicates:
 - (*spinal injury) electrical nerve electrical stimulation*)
 - (*spinal injury) electrical nerve electrical stimulation*)

³ The General Query Log is the record of each SQL statement received from clients, in addition to their connection and disconnection time.

➤ Alphanumeric strings and IP addresses:

- *AND colegio MÃ©xico hazaÃ±a*
- *TÃ©cnicas*
- *http://www.repositoriodgb.buap.mx:2095/*
- *http://www.scielo.org.mx/pdf/tca/v6n3/v6n3a2.pdf*
- *http://www.sciencedirect.com/science*
- *http://www.sciencedirect.com/science/article/pii/S0019850199001133*
- *http://www.sciencedirect.com/science/article/pii/S0048712002732908*
- *http://www.sciencedirect.com/science/article/pii/S0176161712001848*
- *http://www.sciencedirect.com/science/article/pii/S0211563811001684*
- *http://www.sciencedirect.com/science/article/pii/S0308814614003628*
- *http://www.sciencedirect.com/science/article/pii/S1632347500719722*
- *http://www.xvideos.com/video27283249/jynx_maze_anal_banged_on_bangbros_chong as_in_1080p_ch13211_*
- *https://AsPredicted.org/wsxx7.pdf*
- *https://cirworld.com/index.php/jssr/article/view/3380*
- *https://doaj.org/article/c4a86ed60b7a4961baf52e2b45951d65*

➤ Disambiguation:

- *AND terapeuta errores Jaquelin cortez*
- *AND terapeuta errores jaquelin Fortes*
- *sandra massoini*
- *sandra massoni*
- *bio pelicula [spanish]*
- *bio pelicula [termine inesistente]*
- *social media reslut [termine inesistente]*
- *facebook adverstis [termine inesistente]*
- *social media marketing [termine inesistente]*
- “Empty” words like articles and prepositions.

As for the most frequent words, the extraction from the GSGCorpus has provided:

- I. the decreasing frequency of monograms like nouns, adjectives and adverbs, that is the number of occurrences of each word;
- II. the decreasing frequency of bigrams and trigrams and the number of occurrences of each of them in the corpus.

On the basis of their frequency, monograms have been divided in three areas depending on their frequency: high, medium and low. In the highest there are a very few words, while in the lowest there are many but with an irrelevant number of occurrences and the presence of *hapax legomena*⁴. Table 1 and Figure 1 show the twelve most recurrent monograms and how the adjective <social> ranks first with 2412 occurrences.

TERMS	# OCCURRENCES
SOCIAL	2412
MARKETING	1714
MANAGEMENT	1682
EFFECT	1578
MEXICO	1160
SCIENCE	1135
EDUCATION	1122
BUSINESS	1054
CHILD	994
PSYCHOLOGY	916
HEALTH	910
RESEARCH	906

Table 1– Twelve most recurrent monograms

⁴ In [corpus linguistics](#), a *hapax legomenon* (*/ˈhæpəks lɪˈɡomɪnɒn/* also */ˈhæpəks/* or */ˈheɪpəks/*,^{[1][2]} pl. *hapax legomena*; sometimes abbreviated to *hapax*) is a [word](#) that occurs only once within a context, either in the written record of an entire language, in the works of an author, or in a single text.

As shown in Figure 1, the most relevant nouns in the corpus are: <business>, <child>, <education>, <effect>, <health>, <management>, <marketing>, <Mexico>, <psychology>, <research>, <science>.

The only adjective retrieved is <social>.

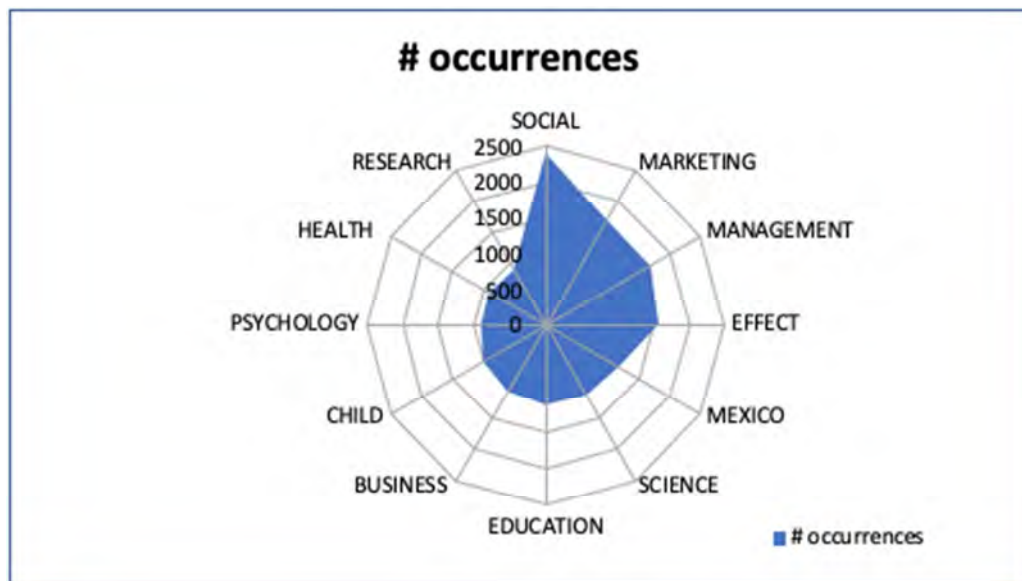


Figure 1 – Most frequent words

Figure 2 shows the bigrams with the higher frequency in the Corpus.

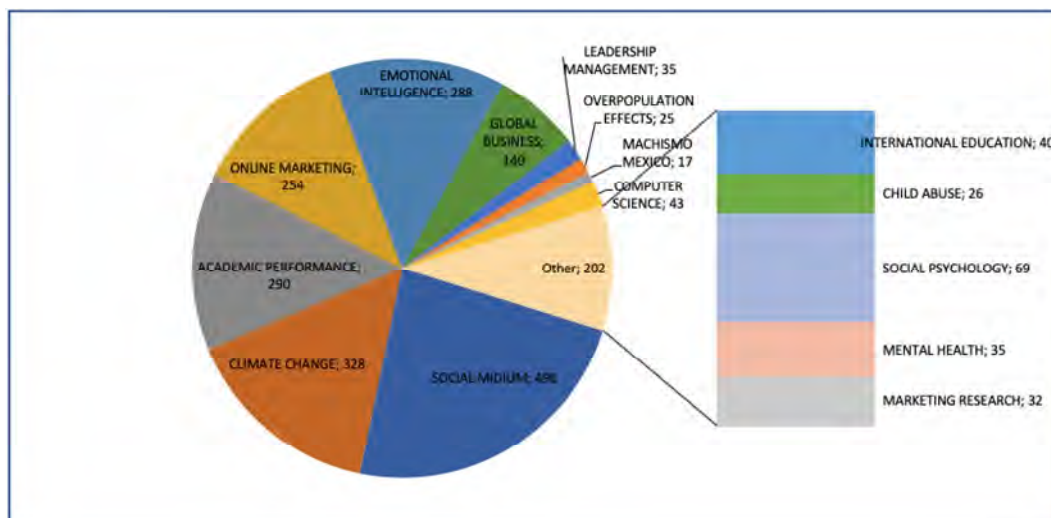


Figure 2 – Most frequent bigrams

3. A case study

At the end of this pre-processing phase, we chose to focus on a flow of queries launched on the *WorldWideScience* platform concerning only the bigram *social media*. A case study has been carried out involving medicine, psychiatry and 'social media'.

Why *social media*?

- ✓ nowadays social media are obviously a very effective means of communication but can even vehiculate knowledge as their various types (eg.: blogs, YouTube, Facebook, Twitter, etc.) are by now often quoted in bibliographical references amongst the more traditional categories (books, journals and so on);
- ✓ the subject involves document types pertaining to Grey Literature.

NLP analysis allowed to browse the corpus through the most and less queried terms: once *social* has been identified as the most frequent one, the analysis was channeled into '*social media*' and the pertinent contexts.

Ontology-based extraction: enrich the domain; retrieve each occurrence of those terms contained in the ontology by using a search engine.

In particular:

- ✓ Some low-frequency terms (hapax) carry a negative connotation⁴, similarity and diversity. in relation to the use of 'social media';
- ✓ An analysis of negative connotations in connection with child/children, is further investigated.

3.1 Results

The topics concerning social media selected from the corpus of queries (Figure 3):

- ❖ The terms of the queries which reveal who are the subjects involved:
 - *The celebrities* [The term “celebrities” connote people according to the fame and public attention accorded by the mass media]
 - *The influencers* [The “influencers” can be defined as internet users who have established a relevant number of virtual relationships]
 - *The Millennials* [The Millennials - or generation Y - are those born in full digital revolution]
 - *The students*
 - *The teenagers*
- ❖ The terms of the queries which follow the expression <social media **cause**> with a negative polarity:
 - *anxiety*
 - *depression*
 - *dietary diseases*
 - *disadvantages*
 - *distraction*
 - *insecurity*
 - *sleep deprivation*
 - *teenager becomes antisocial*
 - *teenager neglect real world interaction*
- ❖ The terms of the queries which follow the expression <social media **impact on**>:
 - *anxiety*
 - *democracy*
 - *families*
 - *hospitality industry*
 - *maintaining relationships with others*
 - *sales*
 - *society*
 - *teens' lives*
- ❖ The terms of the queries which follow the expression <social media **Social media influence**>:
 - *brand loyalty*
 - *consumers purchase decision*
 - *criminality*
 - *fashion trends*
 - *teenagers' body image*
- ❖ The terms of the queries which follow the expression <social media **help**> with a positive polarity:
 - *business growth*
 - *maintain relationships with people*
 - *young people stay connected to distant people*
- ❖ The terms of the queries involving the words<child/children>: bigrams and trigrams with a negative polarity, see Figure 3:

⁴ “In [linguistics](#), a **polarity** is a [lexical item](#) that can appear only in environments associated with a particular [grammatical polarity](#) – affirmative or negative. A polarity item that appears in [affirmative](#) (positive) contexts is called a **positive polarity item** (PPI), and one that appears in negative contexts is a **negative polarity item** (NPI)”, Wikipedia.

- child abuse
- child labor Indonesia
- Child Marriage
- Child psychiatry
- depression child
- domestic violence children
- domestic violence children behavior
- internet safety children
- Obesity child
- obesity children
- Psychopatic assassin children
- punishment children
- The Evil Child Marriage
- violent video games child

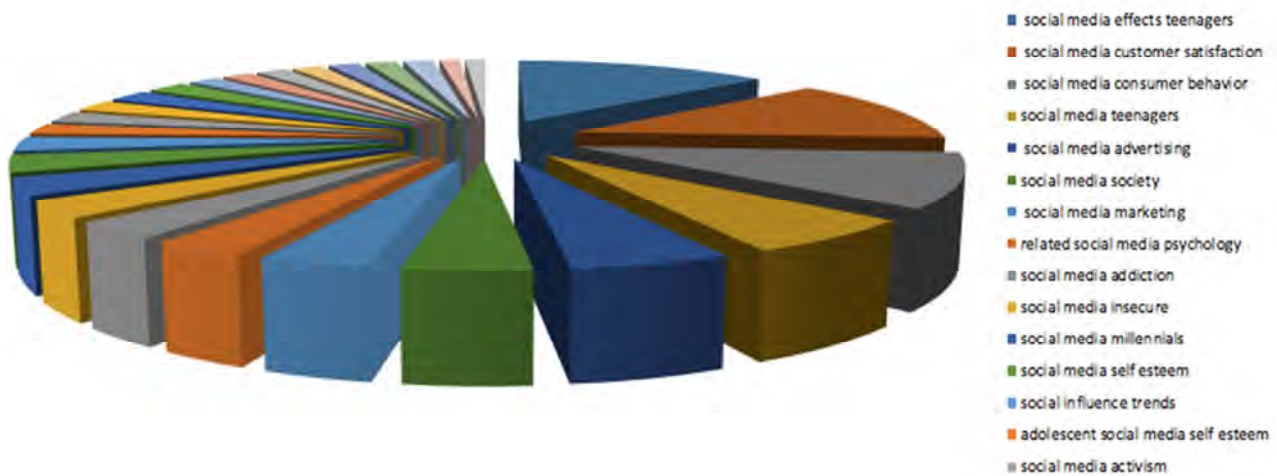


Figure 3 'Social media' occurrences

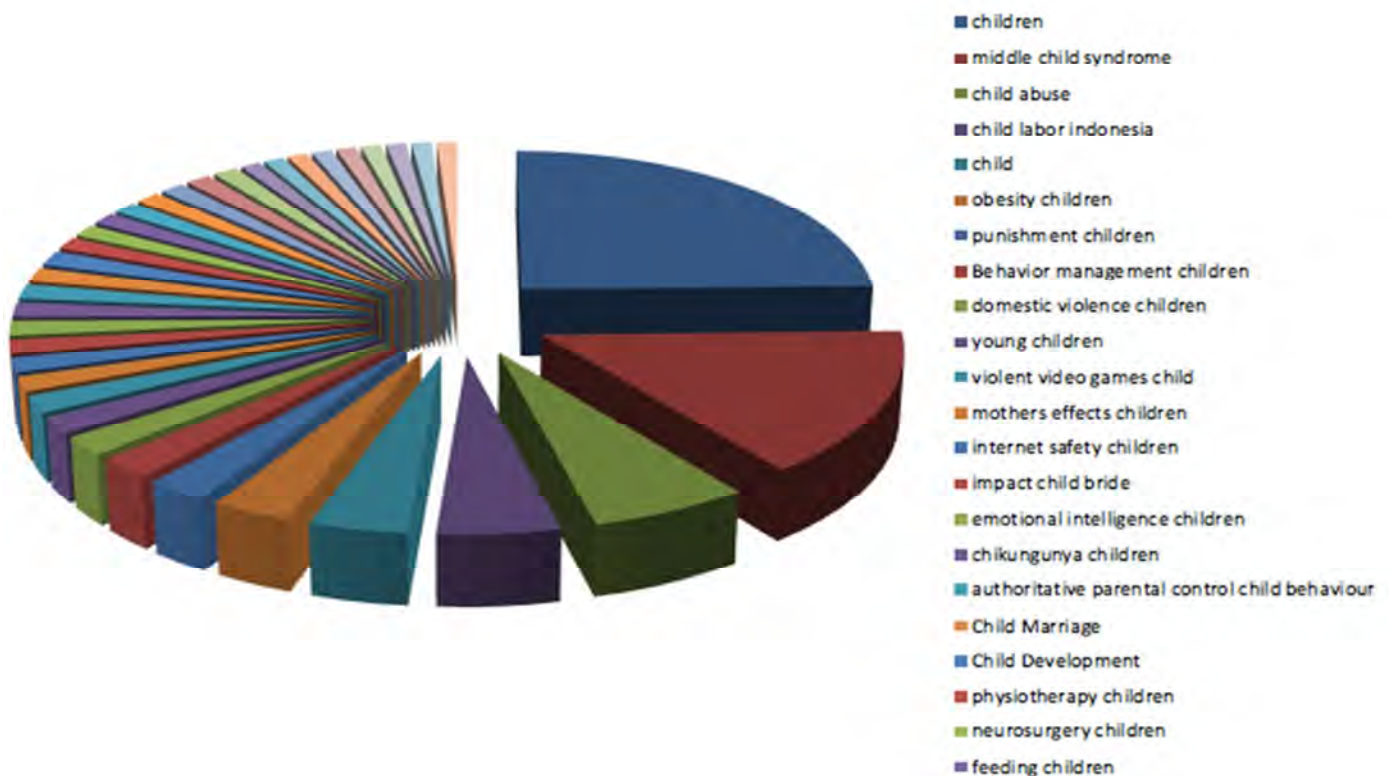


Figure 4 – Child/children context queries

4. Conclusion and final remarks

This work, which does not have any sociological purpose, provides two possible keys for interpreting the sample data of the GSGCorpus.

- I. use of terms resulting from the recently exploded digital technology, in particular the contexts of the bigram <social media>⁵ with either positive, negative or neutral polarity (see Figures 3, 5).
- II. use of terms with a mostly negative polarity which have been retrieved from the contexts of the words <child/children>: examples are given by this string of words which have been extracted from the corpus <prevalence child marriage Burkina Faso>⁶ as well as by the search of the book “The Evil Child Marriage” by Radhika Kapu⁷.

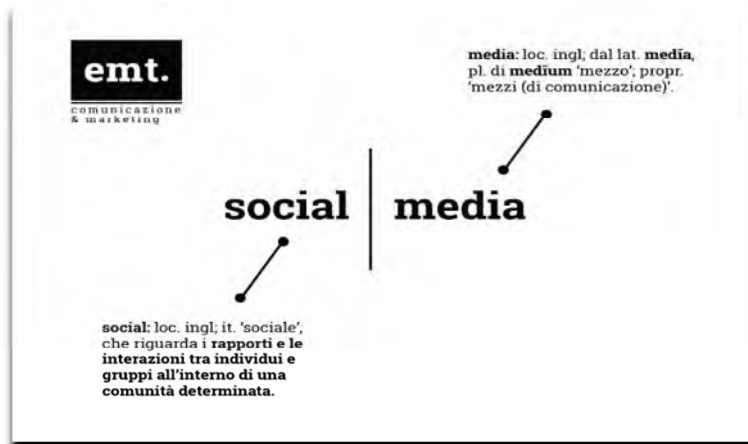


Figure 5. Meaning of the two terms

a) From the sample of queries where the bigram <social media> appears, polar verbs (affirmative and negative) and bipolar verbs (without a clear-cut tendency) have been extracted. In agreement with Manfred Klenner and Stefanos Petrakis, “we observed verbs with a relatively clear positive or negative polarity preference, as well as cases of verbs where positive and negative polarity preference is balanced (we call these bipolar-preference verbs)”.

Just to make an example, a verb like ‘to cause’ has a negative polarity while ‘to help’ has a positive polarity; ‘to impact’ and ‘to influence’ have a neutral polarity (please see par. 3.1 for a detailed description of the terms with a negative polarity used in conjunction with <social media>).

The main results show that the use of social media

- can cause physical harm and widespread diseases;
- can help strengthening virtual relationships;
- can sustain e-commerce and therefore business grown.

b) From the sample of queries associated to the words <child/children>, only a negative polarity of the lexical forms emerges. The use of the following verbs, nouns and adjectives clearly certifies this assumption: ‘to abuse’, ‘to labor’, ‘marriage’, ‘psychiatry’, ‘depression’, ‘(domestic) violence’, ‘obesity’, ‘assassin’, ‘punishment’, ‘domestic’, ‘psychopathic’, ‘violent’.

Terms extracted from the corpus of queries are largely referring to topics pertaining to the major problems of today’s society, eg. *alcoholism*, *depression*, *obesity*, *pornography*, *drugs*, *violence*, etc.

Some critical issues can be identified in the following points: a diachronic analysis of the terms was not possible given the short temporal window taken into account; queries in different languages and many spelling/grammatical errors made our task more complicated by weighing the cleaning process down.

⁵ <https://enricomtomassi.com/cosa-social-media/>

⁶ Burkina Faso has a child marriage prevalence rate of 52%. On average, almost one out of two girls in Burkina will be married before the age of 18. The rates of child marriage vary from one region to another, and are as high as 86% in the Sahel region and 76% in the East region.

< <https://www.girlsnotbrides.org/child-marriage/burkina-faso/> >

⁷ https://www.researchgate.net/publication/323771530_Child_Marriage_The_Social_Evil

Essential Bibliography

Battelle, J. 2005. *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*. Nicholas Brealey Publishing.

Bronson Fitzpatrick R., (2009) WorldWideScience.org: The Global Science Gateway, *Medical Reference Services Quarterly*, 28:4. {363-370, DOI:10.1080/02763860903256121}.

Chiari, I. 2007-2008. *Liste di frequenza. Analisi del testo letterario 1 -*, 2007-2008

Hendry David G., Jenkins J. R., McCarthy Joseph F. (2006). Collaborative Bibliography, *Inf. Process. Manage.*, May 2006, volume 42,3. Pergamon Press Inc. Pages 805-825. {<http://dx.doi.org/10.1016/j.ipm.2005.05.007>}.

Kalgarrieff A., Grefenstette G. *Introduction to the Special Issue on the Web as Corpus*. ACL {<https://www.mitpressjournals.org/doi/pdf/10.1162/089120103322711569>}.

Klenner M., Petrakis S. (2012). Polarity Preference of Verbs: What Could Verbs Reveal about the Polarity of Their Objects? In G. Bouma et al. (Eds.): *NLDB 2012*, LNCS 7337, pp. 35–46, 2012. Springer-Verlag Berlin Heidelberg 2012

Smith A., Anderson M. (2018). *Social Media Use in 2018*, Report, March 1, 2018. Pew Research Center Internet & Technology {<http://www.pewinternet.org/2018/03/01/social-media-use-in-2018/>}.

Trevisan M., Dini L. Barbu E., Barsanti I., Lagos Ni., Segond, F., Rhulmann M., Vald, Ed (2012). Query Log Analysis with GALATEAS LangLog, in *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, ACL, Stroudsburg. Pages 87-91. {<http://dl.acm.org/citation.cfm?id=2380921.2380939>}.

Web Search Queries as a Corpus

<https://www.celi.it/blog/2016/06/web-corpora/>

<<http://www.fondazionemilano.eu/blogpress/weaver/2014/03/22/181/>>

<https://www.uniba.it/docenti/gatto-maristella/attivita-didattica/materiale-didattico/Gatto_light.pdf>

<<https://worldwidescience.org/>>

<<http://greyguide.isti.cnr.it/>>

https://www.researchgate.net/publication/323771530_Child_Marriage_The_Social_Evil rarity item

Appendix

SEMANTIC EXTRACTION SCHEME

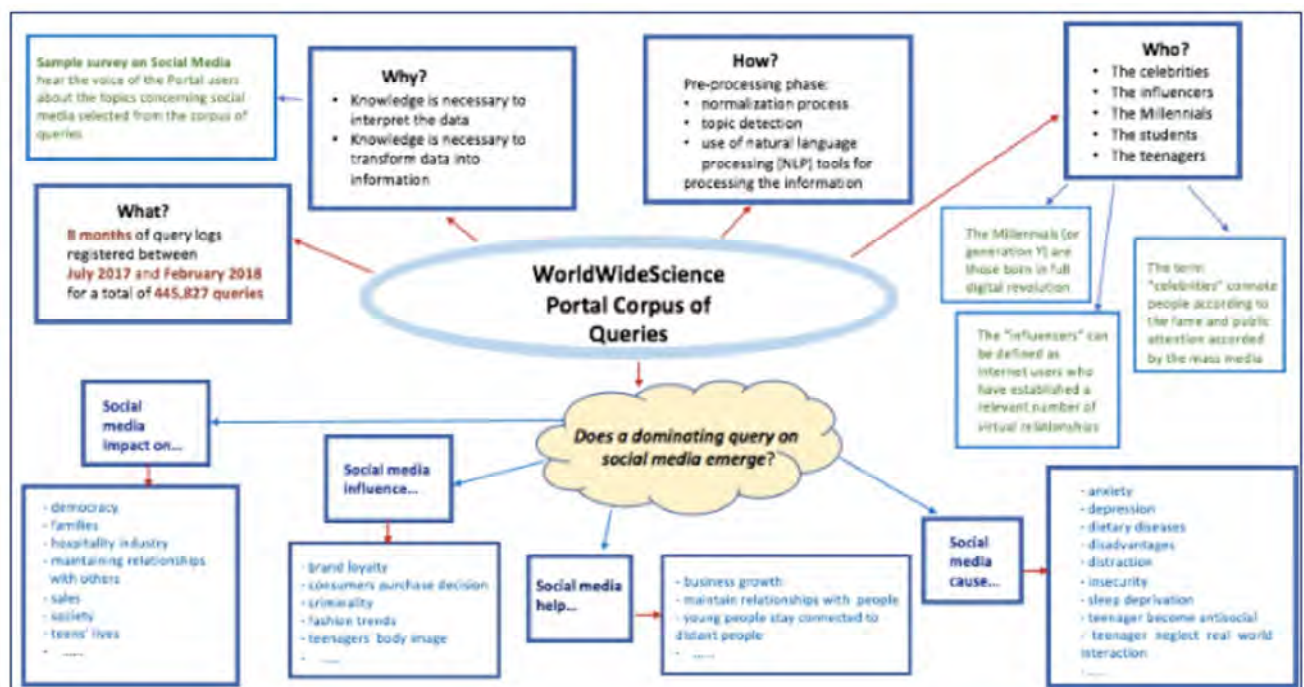


Fig. A1

THE GLOBAL SCIENCE GATEWAY

<<https://worldwidescience.org/index.html>>



Fig. A2

REPOSITORY AND PORTAL TO GOOD PRACTICES AND RESOURCES IN GREY LITERATURE

<<http://greyguide.isti.cnr.it/>>

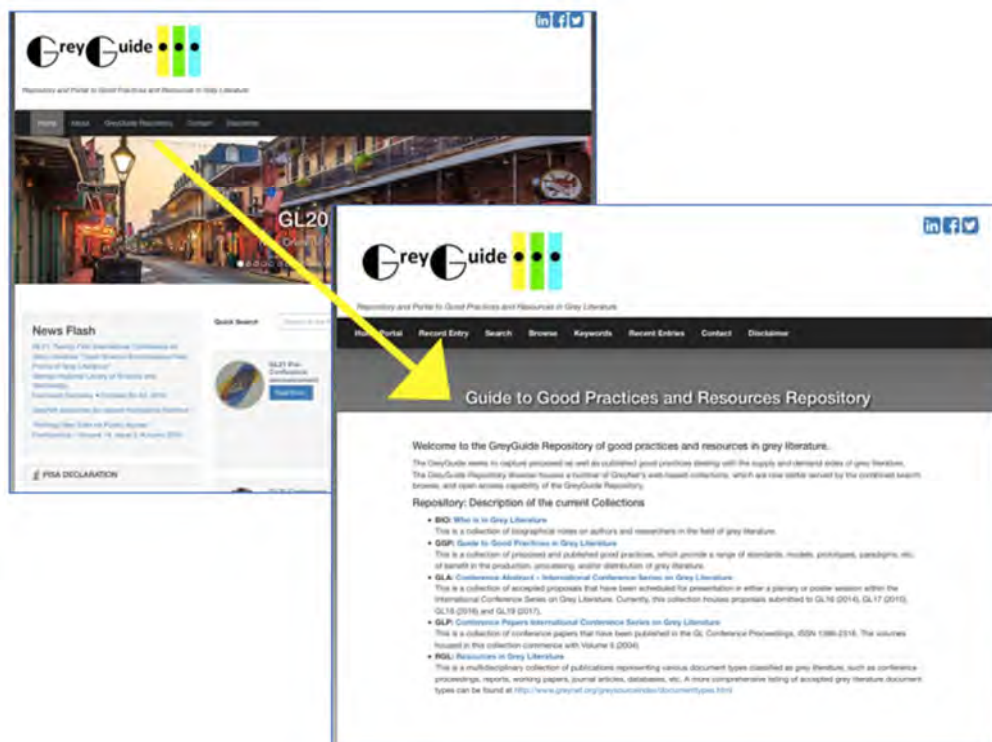


Fig. A3



International Journal on Grey Literature

Subscription Order Form

For *GreyNet* Members free or reduced rates, see <http://www.greynet.org/membership.html>

TGJ Volume 15, 2019	Type of Subscription:	Amount in Euros	Total
THE GREY JOURNAL - Printed ISSN 1574-1796 Annual Subscription, including Postage and Handling	☐ Institutional	€ 240	€
THE GREY JOURNAL - PDF/ CD-Rom ISSN 1574-9320 Annual Subscription, including Postage and Handling	☐ Institutional	€ 240	€
THE GREY JOURNAL - PDF/ Email / PWP ISSN 1574-180X Annual Subscription, including Electronic Handling	☐ Institutional	€ 240	€

Customer information

Name:	
Organisation:	
Postal Address:	
City/Code/Country:	
E-mail Address:	

Check one of the boxes below for your Method of Payment:

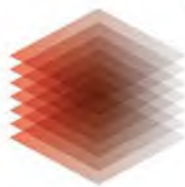
- ☐ Direct transfer to TextRelease, Rabobank Amsterdam, Netherlands
BIC: RABONL2U IBAN: NL70 RABO 0313 5853 42, with reference to 'TGJ/The Grey Journal'
- ☐ MasterCard/Eurocard ☐ Visa card ☐ American Express
- Card No. _____ Expiration Date: _____
- Print the name that appears on the credit card, here _____
- Signature: _____ CVC II code: _____ (Last 3 digits on signature side of card)
- Place: _____ Date: _____

Note: Credit Card transactions will be authorized via Ingenico|Ogone, a designated payment service for Visa Card and MasterCard.

Correspondence Address:

TextRelease

Javastraat 194-HS, 1095 CP Amsterdam, Netherlands
Tel +31-(0) 20-331.2420 • info@textrelease.com • www.textrelease.com



TIB LEIBNIZ INFORMATION CENTRE
FOR SCIENCE AND TECHNOLOGY
UNIVERSITY LIBRARY



**“AS AN INFORMATION CENTRE FOR
THE DIGITISATION OF SCIENCE AND
TECHNOLOGY, OUR OBJECTIVE IS TO
SUPPORT RESEARCHERS AT ALL STAGES
OF THEIR WORK BY PROVIDING THEM
WITH OUR SERVICES.”**

Professor Dr. Sören Auer

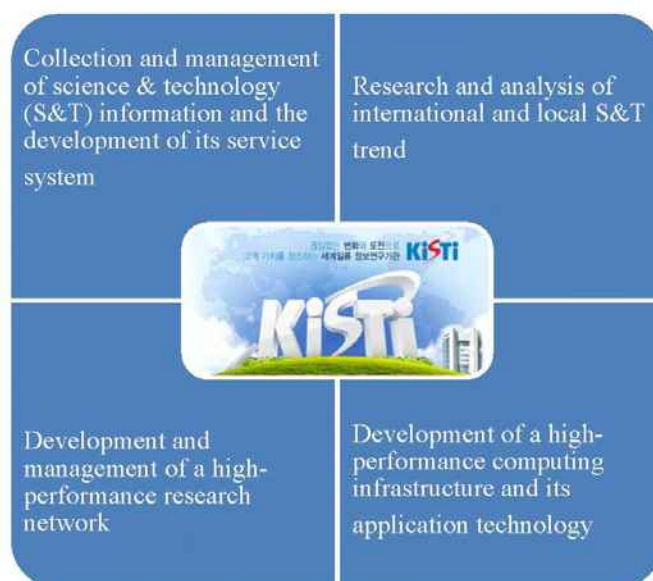
Korea Institute of Science and Technology Information (KISTI)

English version - <http://en.kisti.re.kr/>

* Vision

World-class information research institute creating values for customers

* Main functions



* Management and service of Korean R&D reports

KISTI exclusively manages, preserves, and serves Korean R&D reports for citizens and government officials. It provides Korean R&D reports and their information with National science & Technology Information Service (NTIS) and National Discovery for Science Leaders (NDSL).

* Contact information

KISTI email address: hcpark@kisti.re.kr

Headquarters: Tel : +82-42-869-1004, 1234 Fax: +82-42-869-0969

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 Literature¹. 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 EASY² data archive.

In 2017, GreyNet's Enhanced Publications Project was further broadened to include a Data Papers Project³, 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 Principles⁴ 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 sharing⁵. It is by way of a data paper that the FAIR Data principles are implemented. The FAIR Data Principles formulated in 2014 are related to the data and/or datasets deposited in archives. Prior to FAIR was the Data Seal of Approval⁶, 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 tail⁷ 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 format⁸ 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

<i>Question 6</i> 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%	<i>Question 7</i> When doing research, have you or a colleague ever reused data? Answered: 40 Skipped: 4 Yes 50% No 50% Other -
<i>Question 8</i> 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%	<i>Question 9</i> 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 paper⁹ 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 workshops¹⁰ 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 in Family Medicine

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. 3CGP is placed under Attribution-Non-Commercial-Share-Alike 4.0 International (CC BY-NC-SA 4.0) license.

Introduction

In recent years, grey literature has become more important, especially in research areas, such as General Practice/Family Medicine (GP/FM) (Jamouille, 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 (Jamouille 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 (Jamouille 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 (Jamouille 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 (Jamouille, Grosjean, et al., 2017). In building this taxonomy, Jamouille and colleagues (Jamouille 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, Jamouille and colleagues (Jamouille 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 Bommel, 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 (Jamouille, Grosjean, et al., 2017). Each one of these 182 terms was given only one definition (Jamouille & 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. *Methods 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 Bommel, J. H. (1994). A model for structured data entry based on explicit descriptonal 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.

Analysis of folk literature in grey literature from the National Library of China

Cui Yue, National Library of China, Beijing, China

Abstract:

As a nationwide library, national bibliographic center and document information center, the National Library of China (NLC) is not only comprehensively collecting official Chinese publications but as also, long before, attached importance to the acquisition of grey literature. The NLC's acquisition of grey literature has changed from a basic, scattered collection to a comprehensive collection and finally to a key collection. The institution was established through, in succession, the academic dissertation library, local chronicles and the genealogical documents collection.

At the same time, the NLC was exploring the construction of characteristic Chinese grey literature. Through the analysis of the present grey literature collections in NLC, the amount of folk literature, conference proceedings, research reports and document assembly has already taken shape.

Therefore, a new grey literature system has been formed, among which the most significant increase is in folk literature with distinctive regional characteristics and historical value.

Broadly speaking, folk literature refers to all literature bearing historical and cultural information and kept in folkloric. It contains the literature produced by the people and the official literature lost in folkloric.

This literature covers historical, cultural, artistic and other fields, such as folk poetry, intangible cultural heritage, memoirs, revolutionary historical materials, etc.

At present, literature collections and sorting institutions, such as libraries are not collecting and researching of folk literature to a great enough extent. As a result, researchers cannot easily find comprehensive literature. As a library that bears a mission of preserving cultural heritage, the collection and utilization of folk literature should be emphasized.

Taking folk literature in the grey literature collected by NLC over the past 10 years as a sample, this paper analyzes the content, composition, quantity, type, geographical distribution, value and significance of this kind of literature. To solve the problem of the construction of folk literature in grey literature, this paper innovates acquisition methods and the collection scope of folk literature within grey literature. Some ideas and suggestions are put forward on how to establish folk literature collection institutions and rationally utilize and develop existing resources.

1 Background

As early as January 1912, the Library had begun to collect books. In June 1921, the graduation thesis and dissertation submitted by foreign graduates were accepted in collection[1]. In April 1956, the National Library of China publicly collected scientific research materials through newspapers. In July 1964, the National Library of China, systemically collect nationwide academic conference materials[2].

By 80s, the process of the collection in the National Library was basically in a scattered and chaotic state. Fortunately, it was not until 1985 that the National Library had established a special force, "Domestic Data Collecting Department (DDCD)" which specialized in the collection of GL, and began to investigate, collect and arrange all the grey literature out there. In late 80s, the whole number of copies was up to 100,000[3]. It was unexpected such mass of materials were gathered in a short time. However, due to limited human resources, there were some staggering difficulties in terms of document processing. Therefore, according to the situation mentioned above, in early 90s, NLC adjusted its strategy, primarily focused on thesis, local history and genealogy instead of comprehensive collection. Meanwhile, certain institutions such as dissertation library, local chronicles and genealogy literature center were established. Afterward, on the other hand, the DDCCD was abolished. At the beginning of 2003, the newly revised "Regulations on the Selection of National Library Documents" re-listed the collection of grey literature, formulated the reference range for collection[4]. In general, the National Library's collection of grey literature has roughly gone

through 4 stages, "the fragmentation of the 1980s" - "the full entry, 1980s to the 1990s" - "the selective entry, early 1990s - mid-1990s" - and the "Special Collections, late 1990s - Present".

2 Status of grey literature

2.1 Overview

The "Chinese Information Group" established in 2008 by National Library of China, for providing valuable and professional services for some major national-level projects such as legislative decision-making and scientific research projects, have collected various and numerous GLs from a wide range of channels, like government, college, institute, etc. And as of now, it really has done an excellent job. According to the statistics from 2008 to 2017, a total of up to 150,000 books were included in the Chinese Literature Group (CLG) in the past 10 years. With the experience of the previous literature collection, major collection scope set by CLG included: academic conference literature, research reports, folk literature, data compilation, and other types of literature.

In a decade of collecting, it's observed that the folk literature accounted for around one-third of the total number of grey literature, ranking first. Followed by data compilation and research report literature, just as what you can see from the Fig. 1. It is not difficult to realize that the folk literature in the grey literature database from the National Library currently occupied a large proportion, which has prominent features in terms of both document characteristics and document value. As shown in Fig. 1:

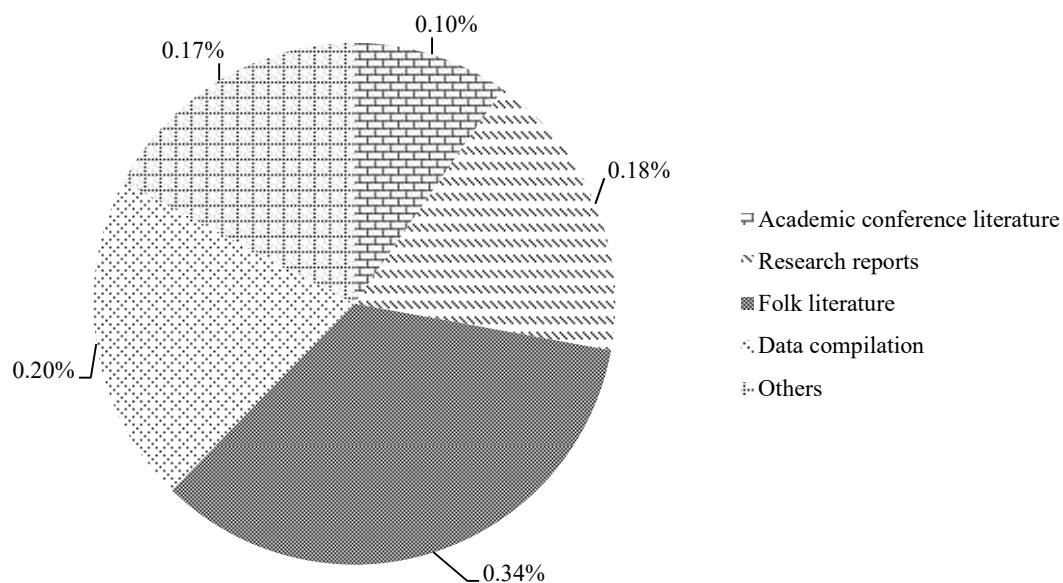


Fig. 1 The proportion of each type of grey literature during 2008-2017

2.2 Source

At present, the grey literature in our library, for the most part, is purchased from booksellers or accepted through donation projects. The purchase should be preceded by lists of book title for our selection supplied by booksellers. Acceptance of donations can be through individuals or groups. The first one is more dispersed but varied, while donations from group organizations are more fixed and systematic, and more continuous. According to Fig. 2, the donation frequency was higher in east China and north China, and their willingness to donate was more positive. On the other hand, it also reflects the uneven distribution of literature collection, should be strengthened to collect literature from other regions. As shown in Fig. 2:

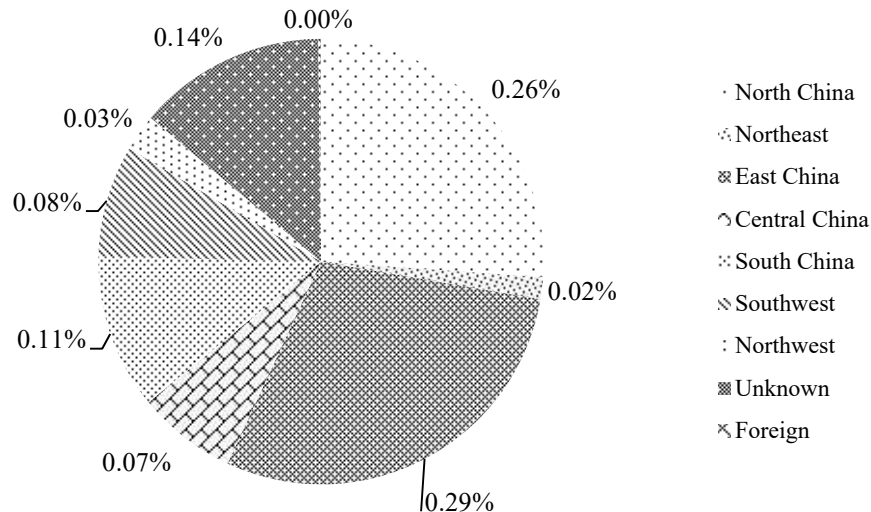


Fig.2 Donation distribution of grey literature in China

After ten years collecting of grey literature, the national library has a mature collection method and literature system of grey folk literature, the collection of literature is concentrated and the source of collection is stable, and has formed the relatively outstanding folk literature specialized collection.

3 Folk literature

3.1 Definition

Academically, it is clearly defined by Zhengzhenman: "Literatures formed and used in daily life of the people". The folk literature is produced by and spreads among people. So, to some degree, it could embody all aspects of people's life in the long historical process. Considering its historical, national and artistic nature, it could demonstrate kaleidoscope of society and history from multi-dimensional perspective for future generations and plays a key role of cultural understanding.

3.2 Classification

The grey folk literature in the National Library collection could generally be divided into three categories: folk customs, folk literary work, folk history materials. There are more than 20,000 kinds and 50,000 volumes of grey folk literature are collected in NLC. Among which folk literary work accounts for 67%, ranking first. Followed by folk history materials accounts for 28%. Specifically, the folk customs include folklores, parables, art and crafts, and folk culture materials. The most distinguished type among them could be ratified as intangible cultural heritage literature. Folk literary work can be represented by poetry and anthology. The poetry kind dominates folk literary work, accounting for 52%. The folk history materials mainly consist of revolutionary historical materials, college chronicle, memoir historical materials, etc. While the number of revolutionary historical materials has prominent position in the folk history materials collection, accounting for 79%.

3.3 Collection

To sum up, there are mainly two ways of obtaining grey literature in national library: institution and individual. Among them, the collection of personal literature is relatively scattered, irregular and discontinuous, which is difficult to trace this part of literature in the literature source. However, donations from group organizations are relatively fixed and systematic. According to statistics, the collection agencies of folk literature are mainly distributed in poetry groups, societies, associations, universities, education institutions, literary and artistic institutions, research institutions and so on.

4 Folk literature collection construction

The types of folk literature are varied and characteristic, so the resources construction of folk literature cannot be generalized. It should be collected, developed and utilized according to their own characteristics. The following is an elaboration on the construction of the typical folk literature.

4.1 The construction of intangible cultural heritage materials

The intangible cultural heritage documents record and preserve the original appearance of the intangible cultural heritage, providing effective help for the protection of the endangered intangible cultural heritage and research. Intangible cultural heritage documents are of historical, cultural and artistic value. Collecting such literature can inject new vitality into the library, and it is also an important supplement of literature resources.

According to statistics, there are 547 kinds of grey intangible cultural heritage documents in the national library. It can be divided into: Materials of intangible cultural heritage census, records of intangible cultural heritage, intangible cultural heritage research, laws and regulations related to intangible cultural heritage protection, intangible cultural heritage activities[5]. Materials of intangible cultural heritage census refer to a series of intangible cultural heritage census information generated in the work of intangible cultural heritage census; Materials of records of intangible cultural heritage refer to record the origin, development and inheritance of intangible cultural heritage; Materials of research on intangible cultural heritage refer to general or special research on intangible cultural heritage; Materials of intangible cultural heritage activities refers to the materials of lectures, exhibitions and publicity activities organized around intangible cultural heritage and protection. According to the statistics, the materials of intangible cultural heritage census information accounts for 62%[6], ranking first. On the one hand, it shows that the intangible cultural heritage census has achieved remarkable results in recent years; On the other hand, it also shows that the intangible cultural heritage census is the most basic and most direct way to protect the intangible cultural heritage. The documentation of records of intangible cultural heritage accounts for 23%, ranks the second. Indicating that under the influence of social development and environmental change, the endangered intangible cultural heritage can be recorded and preserved in the form of pictures and texts, which is also an important part of the protection of intangible cultural heritage.

Systematic solicitation. Literature of intangible cultural heritage should be collected in a targeted and systematic way according to its characteristics. Based on the above analysis, the literature of intangible cultural heritage census has the characteristics of large amount of systematic literature. Based on this, collection of intangible cultural heritage census materials according to literature sources, release time, release batches, etc. Secondly, according to the regional characteristics of intangible cultural heritage, systematic solicitation can be conducted in different regions. Thirdly, systematic solicitation can be carried out according to the grade of intangible cultural heritage, such as the collection of intangible cultural heritage system divided into national、provincial、municipal and county levels. Finally, a systematic collection of intangible cultural heritage batches can be conducted according to the intangible cultural heritage list.

Make special catalogue of intangible cultural heritage. Making a special catalogue can better understand the literature collection, timely find missing documents to supplement, and ensure the integrity of the literature. At the same time, it is conducive to literature collection and collation. The catalogue can be classified according to literature type, literature source, literature content and literature grade.

4.2 The construction of folk poetry materials

Chinese poetry originated in the pre-Qin Dynasty and flourished in the Tang Dynasty. The Chinese iambic verse originated in the Sui and Tang Dynasties and gain its popularity in the Song Dynasty. The poetry comes from the folk and belongs to the real grassroots literature. The poets of the Tang Dynasty, such as Li Bai, Du Fu, BaiJuyi are all well-known in the oriental culture. At present, the National Library's poetry and literature documents have a more comprehensive interview system, the interview direction is clear, the interviewees are

very active, and we can acquire continuous relevant literature. which is the main reason for the significant growth in poetry literature in NLC database.

According to statistics, there are more than 5,000 kinds of grey folk poetry literature in the collection, among which contemporary poetry accounts for 98%. Showing poetry in China is greatly favored by the public and have a potential to develop further. On the other hand, it also shows that contemporary poetry is a kind of closer acquisition, easier to collect.

A total of 166 poetry associations, societies and other poetry-related groups in the country regularly donate poetry documents to the National Library. Donation groups are divided by region, as shown in Fig. 3. it can be easily seen that there are more poetry groups in Central South part of China, and their willingness to donate is more positive. As shown in Fig. 3 :

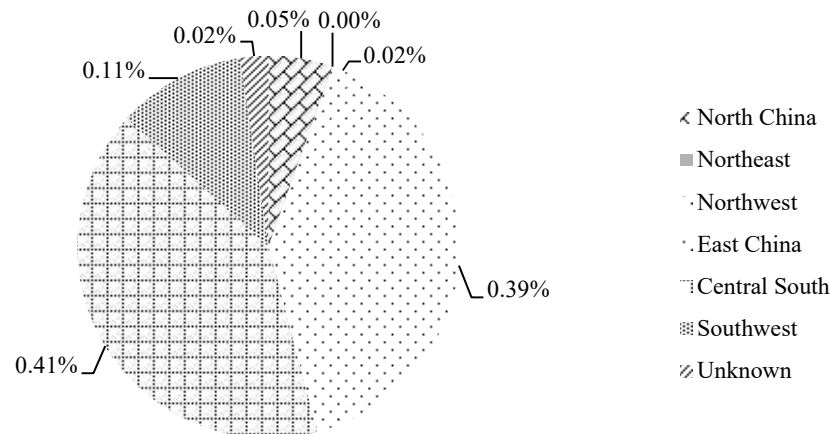


Fig. 3 Poetry donation distribution in China

According to the statistics, the folk poetry donation has some fixed poetry association and social groups. First, maintain the donation relationship with the institute, college, and certain poetry group, through which we could obtain stable and continuous poetry documents. And the Internet can be helpful too, like Chinese poetry net, and any other sites involving Chinese poetry, more selection of poetry and literature can be actively selected. Second, through the classification of poetry depending on the content of literature, we could set special catalog for better research and development.

4.4 The construction of revolutionary historical materials

The historical materials of revolution not only record the history of revolution, but also provide materials for the research of the history of revolution at a certain stage and provide help for the propaganda of traditional revolution education. At the same time, the historical data of revolution in grey literature recorded the history of revolution from a macro perspective, and vividly and succinctly described the history of revolution from the individual perspective. It complements the revolutionary history in many ways.

At present, the revolutionary history materials in the National Library Collection is expected to be more than 2,000 copies, including: History of revolutionary war; Memoir in revolutionary; Record of heroes and other historical materials. Among which History of revolutionary accounts for 38%, ranking first. Followed by Memoir in revolutionary and Record of heroes. As shown in Fig. 4 :

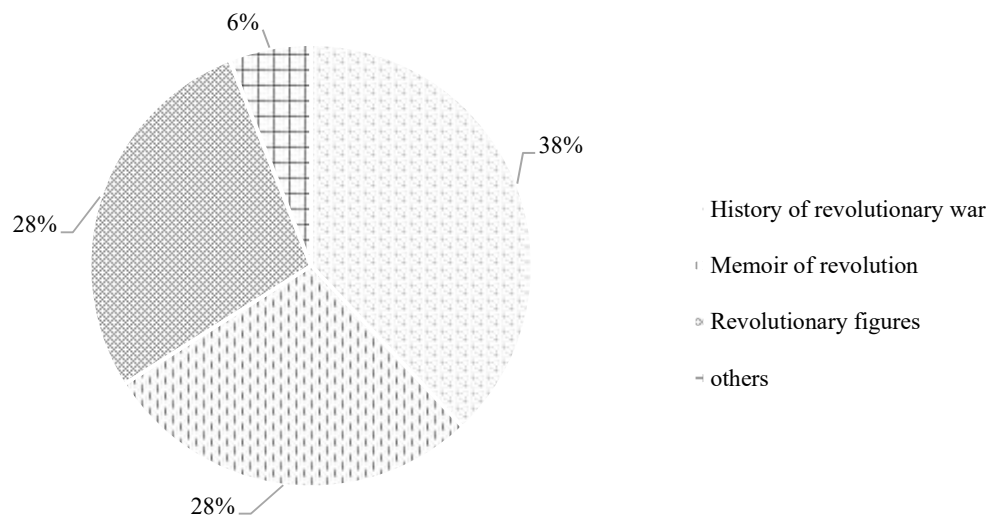


Fig. 4 Classification of revolutionary historical materials

The history of revolutionary war materials mainly records relevant materials about the war in the warfare process. The memoir in revolutionary age materials mainly refers to oral and written account of revolutionary events. The record of heroes materials mainly refers to the life story, biography and revolutionary of hero and heroine, etc.

Tracking the literature according to the source of the literature. it could ensure the validity, continuity and stability. As it statistic, revolutionary historical materials mainly come from individuals and institutions, and personal publishing part possess a large proportion, taking up 73.2%. while the rest materials come from institution such as Civil affairs bureau, Seminar, Retirement center for veteran cadres, Archives bureau, etc.

Secondly, Classifying on the basis of the types of revolutionary historical materials, for further development and utilization. In fact, our library has begun to compile the "Revolutionary History" bibliography as early as the early 1980s. By compiling certain bibliography, we could better understand the contents of the collection and conduct more efficient and reasonable classification management. We could according to the different publishing institutions to classify revolutionary historical materials, such as the History Research Office catalog, the Revolutionary Research Society institution catalog, the Archives bureau catalog etc. From the perspective of the literature form, it can be divided into historical materials catalog, memory catalog, research catalog, publicity and education catalog, etc.

5 Conclusion

At present, the collection of grey literature in China is still in the stage of continuous exploration. There are still many improvements in the collection and utilization of the folk literature. The folk literature is diversified, the literature sources are relatively scattered, and it is not easy to track and collect. In addition, the public's cognition of grey literature donation is simple, and the folk groups know little about the rules of grey literature entry. So, the publicity and promotion of grey literature collection should be strengthened to enhance the awareness of folk groups on grey literature and enhance their awareness of donation. Without active solicitation, the grey folk literature may be lost. At the same time, the collection of grey literature also needs the support of the government to actively promote the introduction of the grey literature legal deposit system. It provides a better guarantee for the collection of folk literature. Then, build resource sharing platform and open access resource. By building a resource sharing platform, the folk literature of different units and different regions are comprehensively integrated.(1)The collection information of folk literature is provided by culture-related units, libraries of various types, professional non-governmental organizations and individuals, etc. so as to achieve the largest collection of the substantive resources of folk literature.(2) The most extensive retrieval and utilization

of the data is realized through the folk literature bibliographic data Shared by the libraries. (3) The public will gain a deeper understanding of the grey folk literature by publishing the solicitation information, collection process and collection system on the Shared platform. Finally, the digitization of folk grey literature. At present, social development has achieved digitalization, networking and is now entering the direction of artificial intelligence. There are some limitations in the acquisition of physical books and the quantity of literatures. Some folk literatures only have one volume, so it is difficult to collect literatures. The valuable grey literature can be digitized and saved permanently, and the digital literature can be organized and analyzed. And then by using advanced technologies such as holographic images, valuable folk history and endangered intangible cultural heritage techniques will be reproduced and restored through three-dimensional images to provide more service space for the cultural field and the technical field.

References

- [1] Beijing library business research committee. collection of historical materials of Beijing library (1909-1949)[M]. Bibliography and Document press,1992
- [2] Compilation of historical materials of Beijing library (2) editorial committee. Compilation of historical materials of Beijing library (ii):1949-1966[M]. Beijing library press,1997
- [3] Yao rong. Development and utilization of Chinese materials[J]. journals of Library Science,2004(4)
- [4] Wang dongbo, Zhao xiaohong. Improve the policy of document collection and selection and build the national general library -- an overview of the amendment of the regulations on document collection and selection of national library[J]. Journal of the National Library of China, 2004(1)
- [5] Cui yue. Some Observations of Improving the Document Catalog of Intangible Cultural Heritage—From the Perspective of Full Disclosure of the Literature Content[J].Library Development,2018(1)
- [6] Cui yue. Research on Intangible Cultural Heritage Literature Collection Development[J].Journal of the National Library of China.2018(3)

National Repository of Grey Literature (NRGL)



NRGL is
digital
repository
for grey
literature

**Free
online
access**

Features

Provider:

National Library of Technology
Prague, Czech Republic

Records:

over 400,000 records

Collection provenance:

Czech Republic

Partners:

over 130 organizations (Academy of Science,
Public Research Institutions, Universities, State
Offices, Libraries, NGOs etc.)

International Cooperation:

OpenGrey, OpenAire, ROAR, OpenDOAR, BASE

Goals

- Central access to grey literature and the results of research and development in the Czech Republic
- Support of science, research and education
- Systematic collection of metadata and digital documents
- Long-term archiving and preservation
- Cooperation with foreign repositories

What else?

Conference on Grey Literature
and Repositories

<http://nrgl.techlib.cz/conference/>

Informative Web pages

<http://nrgl.techlib.cz>

www.nusl.cz

NTK
2,5x
4x
Národní technická knihovna
National Library of Technology

NUŠL
národní
úložisko
šedé
literatury

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; Knowledge Resource Service, 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 patterns¹ 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 world². Introduced as a form of grey literature via Second Life³, 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 breakthroughs in artificial intelligence⁴. 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 history⁴. 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 electrocardiograms⁴, the invention of cloud networking in 2006⁴ is considered the first proven use of AI and VR in the modern era focusing on healthcare. Although the arguments for AI and VR in clinical settings are plentiful, ranging from enhancing imaging and increased processing speed in electronic medical record (EMR) applications⁴, 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 cells⁵.” However, a column authored in 2017 disputed a claim by IBM that Watson was the new revolution to cancer care⁶.

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 Schopf 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 brain⁴. 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 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 intimidating 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-May 2016.¹¹ 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 #articlelevelmetrics, 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, it 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. Schopf 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)

AI & VR in Cancer Care OR Health Care: Literature Search

Resources Consulted:

- MEDLINE (Ovid)
- PubMed
- CINAHL
- MEDLINE (Ebsco)
- Google Scholar

Limits Applied:

- English language
- 2007 – present

Terms Brainstormed:

artificial intelligence
virtual reality
machine learning
healthcare
public health
cancer
oncology
technology
2-D; 3-D

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 norm⁷. 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 gleaned 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 gleaned 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.

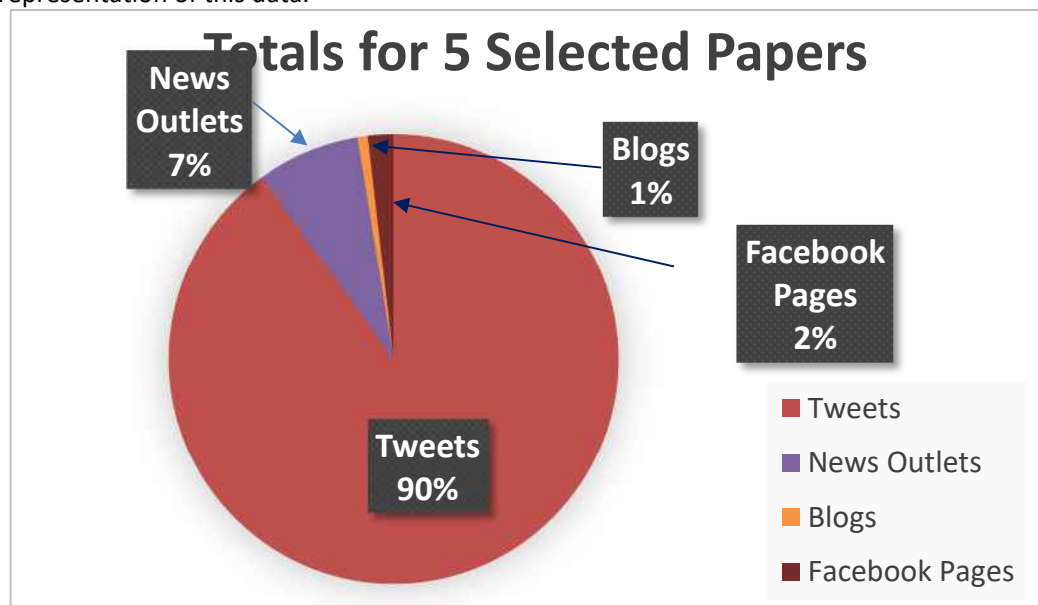


Figure 4: AI in Healthcare: Grey Literature Document Types

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.

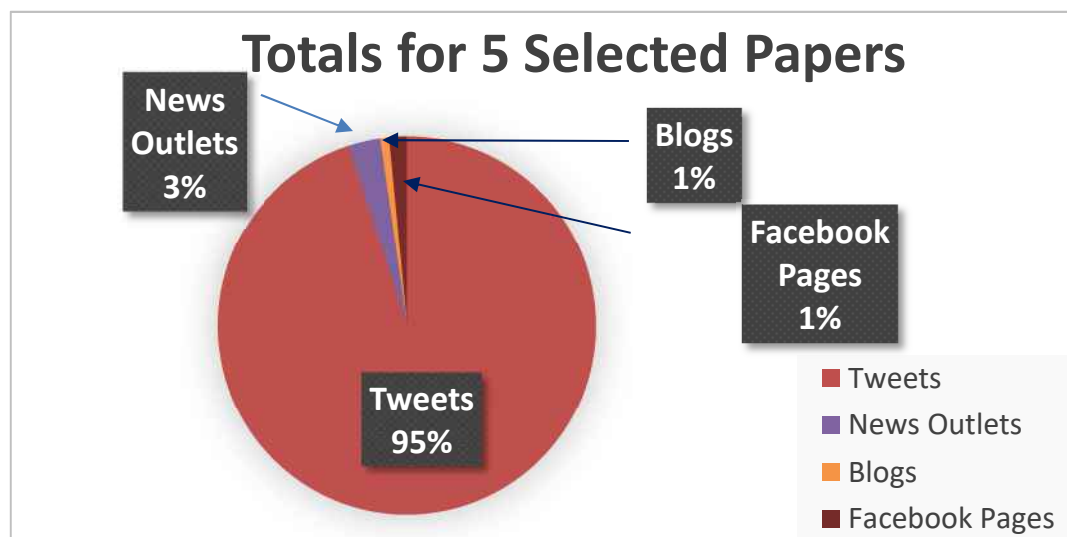


Figure 6: VR in Cancer Care: Grey Literature Document Types

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. Schopf, 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 #articlelevelmetrics, 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. Konkil, 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.

Slovak Centre of Scientific and Technical Information **SCSTI**

Achieve
your goals
with us



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

List of Participating Organizations

Access Innovations, Inc.	United States
Air Force Civil Engineer Center AGEISS Inc.	United States
Alberta Health Services	Canada
Data Archiving and Networked Services, DANS-KNAW	Netherlands
Deep Web Technologies	United States
Duke University Medical Center Library	United States
EBSCO	United States
Federal Library Information Network; Library of Congress, FEDLINK-LOC	United States
Gambling Research Exchange Ontario, GREO	Canada
George Mason University, GMU	United States
Georgetown University Medical Center; Dahlgren Memorial Library	United States
German National Library of Science and Technology, TIB	Germany
GreyNet International, Grey Literature Network Service	Netherlands
IBM Federal	United States
Indiana University Libraries	United States
Information Today Inc.	United States
Institut de l'Information Scientifique et Technique, Inist-CNRS	France
Institute of Computational Linguistics, ILC-CNR	Italy
Institute of Informatics and Telematics, National Research Council	Italy
Institute of Information Science and Technologies, ISTI-CNR	Italy
Irvine Valley College	United States
Japan Atomic Energy Agency, JAEA	Japan
Korea Institute of Science & Technology Information, KISTI	Korea
LibSource, LAC-Group Company	United States
Louisiana State University Libraries, LSU	United States
Loyola University New Orleans	United States
National CK University; College of Management	Taiwan
National Institute of Standards and Technology, NIST	United States
National Library of China, NLC	China
National Library of Technology, NTK	Czech Republic
National Research Council of Italy, CNR Central Library	Italy
NeMIS Research Laboratory, ISTI-CNR	Italy
Network and Informative Systems Office, CNR	Italy
Nuclear Information Section; International Atomic Energy Agency, NIS-IAEA	United Nations
Office of Scientific and Technical Information; OSTI-DOE	United States
PricewaterhouseCoopers, PwC	Netherlands
Princeton University	United States
Sandia National Laboratories	United States
Slovak Centre of Scientific and Technical Information, CVTISR	Slovakia
Texas A&M University Libraries	United States
TextRelease, Program and Conference Bureau	Netherlands
University of California, Irvine Libraries	United States
University of Florida; George A. Smathers Libraries	United States
University of Ghent, Heymans Institute of Pharmacology	Belgium
University of Houston, University Libraries	United States
University of Illinois at Urbana-Champaign	United States
University of Liège, Department of General Practice	Belgium
University of Liège, HEC School of Management	Belgium
University of Lille	France
University of Maryland; University Libraries	United States
University of Minnesota Libraries	United States
University of Rouen, Department of Information and Medical Informatics	France
University of Southern California, USC	United States
University of Texas Health Science Center at Houston, School of Biomedical Informatics	United States
University of the Republic of Uruguay	Uruguay
University of Wisconsin, Milwaukee	United States
U.S. Geological Survey, USGS	United States
U.S. Government Accountability Office, GAO	United States
U.S. Government Publishing Office, GPO	United States
Utah State University, USU	United States
WorldWide Science Alliance	United States



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



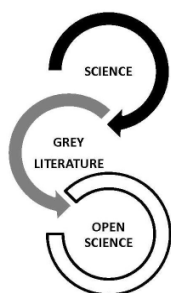
Conference Announcement

For more than a quarter century, grey Literature communities have explored ways to open science to other methods of reviewing, publishing, and making valuable information resources publicly accessible. This Twenty-First International Conference on Grey Literature seeks to demonstrate how the principles of science and advancements in information technology have impacted the field of grey literature and in turn how grey literature by implementing these has contributed to the open science movement.

Open science is defined as the movement to make scientific research, data and dissemination accessible to all levels of an inquiring society¹. Grey literature by definition seeks to make publications produced on all levels of government, academics, and business openly accessible different from those controlled by commercial publishing. As such the open science movement incorporates the work carried out by grey literature communities and renders an even broader framework encompassing newer forms of grey in both textual and non-textual formats.

Open science encompasses the life and physical sciences as well the social sciences and humanities as does grey literature. Open science recognizes the value of grey literature in the process of knowledge generation and as such acknowledges contributions made by researchers, authors, and their communities of practice. Open science changes the way research is done and allows for convergence with the field of grey literature. It is within the open science movement that grey literature and its wealth of information resources are valued and properly exploited for society as a whole.

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

Dateline 2019

• March 31	• April 12	• April 18	• April 25	• Sep. 15	• Sep. 20	• Oct. 15	• Oct. 22-23
Close, Call for Papers	Program Committee Meeting	Authors Notified	Open, Call for Posters	Close, Early Conference Registration	Close, Call for Posters	Submission Conference Papers	GL21 Conference Hannover

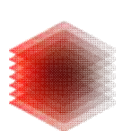
¹<https://www.fosteropenscience.eu/taxonomy/term/7>



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 Papers

Title of Paper:

Author Name(s):

Organization(s):

Postal Address:

Postal Code – City – Country:

Conference Topic(s):

Phone:

Email:

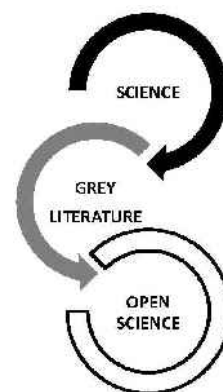
URL:

Guidelines for Abstracts

Participants who seek to present a paper dealing with grey literature are invited to submit an English language abstract between 300-400 words. The abstract should address the problem/goal, the research method/procedure, an indication of costs related to the project, as well as the anticipated results of the research. The abstract should likewise include the title of the proposed paper, conference topic(s) most suited to the paper, name(s) of the author(s), and full address information. Abstracts are the only tangible source that allows the Program Committee to guarantee the content and balance in the conference program. Every effort should be made to reflect the content of your work in the abstract submitted. Abstracts not in compliance with the guidelines will be returned to the author for revision.

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 and Format for Submission

Abstracts in MS Word must be emailed to conference@textrelease.com on or before **March 31, 2019**. The author will receive verification upon its receipt. By mid-April, shortly after the Program Committee meets, authors will be notified of their place on the conference program. This notice will be accompanied by further guidelines for submission of full text papers, biographical notes, accompanying research data, PowerPoint slides, and required Author Registration.

Author Information

Bartolini, Roberto 105

Roberto Bartolini - Expertise on design and development of compilers of finite state grammars for functional analysis (macro-textual and syntactic) of Italian texts. Expertise on design and implementation of compilers of finite state grammars for analysis of natural language texts producing not recursive syntactic constituents (chunking) with specialization for Italian and English languages. Skills on acquiring and extracting domain terminology from unstructured text. Skills on semi-automatic acquisition of ontologies from texts to support advanced document management for the dynamic creation of ontologies starting from the linguistic analysis of documents.

Email: roberto.bartolini@ilc.cnr.it

Biagioni, Stefania

105

Stefania Biagioni graduated in Italian Language and Literature at the University of Pisa and specialized in Data Processing and DBMS. She is currently an associate member of the research staff at the Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" (ISTI), an institute of the Italian National Research Council (CNR) located in Pisa. She is currently involved in the activities of the ISTI Networked Multimedia Information Systems Laboratory (NMIS). She has been head librarian of the Multidisciplinary Library of the CNR Campus in Pisa till August 2017. She was the responsible of ERCIM Technical Reference Digital Library (ETRD) Project and currently is the coordinator of the PUMA (Publication Management) & MetaPub, a service oriented and user focused infrastructure for institutional and thematic Open Access repositories looking at the DRIVER/OpenAIRE vision, <http://puma.isti.cnr.it>. She has coauthored a number of publications dealing with digital libraries and grey literature. Her research interest are focused on digital libraries, knowledge sharing and transfer in scientific area, scholarly communication infrastructures, Open Access and Open Science. She has been dealing with grey literature since 90's. Since 2013 she is involved on the GreyGuide Project.

Email: stefania.biagioni@isti.cnr.it

ORCID iD <https://orcid.org/0000-0001-9518-0267>

Carlesi, Carlo

105

Carlo Carlesi, graduated in Computer Science, worked since 1970 at the IEI (now ISTI) of the CNR in Pisa. He is currently a Research Associate of the Institute ISTI and he is involved in the following projects: PUMA - Publication Management. The Digital Library service allows public access (when permitted) through Internet to the published documents produced by CNR Organizations. And GreyGuide, portal and repository of good practice and resources in the field of grey literature.

Email: carlo.carlesi@isti.cnr.it

ORCID iD <https://orcid.org/0000-0001-9808-6268>

Etkin, Cynthia

12

Cynthia Etkin is the Senior Program Planning Specialist in the Office of the Superintendent of Documents (SOD), U.S. Government Publishing Office (GPO). She brings to this position 21 years of experience at GPO and almost 15 years of experience managing Federal depository and law library operations in academic libraries. Cynthia's job focuses on policy and planning for the public information programs of the SOD directly supporting GPO's mission of "Keeping

America Informed" and the public's right to freely access its government's information. Recent efforts include preparing for GPO's trustworthy digital repository audit, developing the National Plan for Access to U.S. Government Information, and implementing an eLearning platform. She continues to work on implementing the National Plan while anxiously following Congressional efforts to modernize the statutory authority of the Depository Library Program. Email: etkin@gpo.gov

Farace, Dominic

117

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

117

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>

Giannini, Silvia

51

Silvia Giannini graduated and specialized in library sciences. Since 1987 she has been working in Pisa at the Institute for the Science and Technologies of Information "A. Faedo" of the Italian National Council of Research (ISTI-CNR) as a librarian. She is a member of the ISTI Networked Multimedia Information Systems Laboratory (NMIS). She is responsible of the library automation software "Libero" in use at the CNR Research Area in Pisa and coordinates the bibliographic and managing activities of the ISTI library team. She cooperates in the design and development of the PUMA (Publication Management) & MetaPub, an infrastructure software for institutional and thematic Open Access repositories of published and grey literature produced by CNR.

Email: silvia.giannini@isti.cnr.it

ORCID iD <https://orcid.org/0000-0001-7323-3786>

Author Information (CONTINUED)

Goggi, Sara

105

Sara Goggi is a technologist at the Institute of Computational Linguistics "Antonio Zampolli" of the Italian National Research Council (CNR-ILC) in Pisa. She started working at ILC in 1996 working on the EC project LE-PAROLE for creating the Italian reference corpus; afterwards she began dealing with the management of several European projects and nowadays she is involved with organisational and managerial activities mainly concerning international relationships and dissemination as well as organization of events (e.g. LREC conference series). Currently one of her preminent activities is the editorial work for the international ISI Journal Language Resources and Evaluation, being its Assistant Editor. Since many years (from 2004) she also carries on research on terminology and since 2011 - her first publication at GL13 - she is working on topics related with Grey Literature. Email: sara.goggi@ilc.cnr.it

Henderson, Kathrine A.

25

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.

Hersey, Denise

67

Denise Hersey holds an MA in American History from University of Massachusetts, a BA from University of Pennsylvania, and an MLS from Southern Connecticut State University. She is currently the Head of the Science Libraries at Princeton University, having previously held the position of Assistant Director of Clinical Information Services at the Cushing/Whitney Medical Library at Yale University. Denise has also worked as a corporate librarian and at liberal arts institutions. Her current professional interests are in the communication of science to the public, and using user experience methods to help inform changes within libraries.

Kelly, Elizabeth

83

Elizabeth Joan Kelly, Digital Programs Coordinator at Loyola University New Orleans, manages digitization activities for Special Collections & Archives and is also responsible for collecting, maintaining, and assessing usage data for the library's digitized collections. Kelly publishes and presents on archives, digital library assessment, and library pedagogy, and co-founded the Digital Library Federation Digital Library Pedagogy group.

ORCID iD <https://orcid.org/0000-0002-7306-3331>

Li, Yuan

67

Yuan Li - As a Scholarly Communications Librarian at Princeton University, Yuan Li manages the Princeton University Library's efforts to support scholarly publication innovations and reforms, and supervises and coordinates activities related to the Princeton Open Access policy and Repository. Previous roles include Scholarly Communication Librarian at Syracuse University, Digital Initiatives Librarian at University of Rhode Island, and Digital Repository Resident Librarian at University of Massachusetts Amherst. She holds an MLS from the University of Rhode Island, a MS in Applied Computer Science from the National Computer System Engineering Research Institute of China, and a BS in Computer Science from Yanshan University (China). As a Scholarly Communications Librarian at Princeton University, Yuan Li manages the Princeton University Library's efforts to support scholarly publication innovations and reforms, and supervises and coordinates services related to open access, copyright, and data management. Previous roles include Scholarly Communication Librarian at Syracuse University, Digital Initiatives Librarian at University of Rhode Island, and Digital Repository Resident Librarian at University of Massachusetts Amherst. She holds an MLS from the University of Rhode Island, a MS in Applied Computer Science from the National Computer System Engineering Research Institute of China, and a BS in Computer Science from Yanshan University (China).

Lipinski, Tomas A.,

25

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

Author Information (CONTINUED)

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. tlipinsk@uwm.edu

Molino, Anna

51

Anna Molino graduated in Linguistics at the University of Pisa in 2010. Since 2013, she works at CNR - ISTI ("Istituto di Scienza e Tecnologie dell'Informazione - A. Faedo") as member of the Networked Multimedia Information Systems Lab. (NeMIS). She has worked as project assistant and financial manager in various EU funded and national research projects for the Digital Libraries group of the NeMIS lab. She contributes in the language revision and translation of scientific papers. Email: anna.molino@isti.cnr.it

Muglia, Caroline

83

Caroline Muglia, Head of Resource Sharing and Collection Assessment Librarian at the University of Southern California (USC), manages the InterLibrary Loan and Document Delivery department and leads the collection assessment efforts for the Library system. In this capacity, she is responsible for qualitative and quantitative assessment and evaluation of all resources, the return on investment, and ways in which the library resources support research at the institution. Her current research interests include collection assessment, open education resources (OER), and streaming media opportunities in libraries. Email: muglia@usc.edu

Monachini, Monica

105

Monica Monachini is a Senior Researcher at CNR-ILC. Field of expertise: computational linguistics, computational lexicography, semantics, lexical semantics, language resources, ontologies, lexicon, terminologies, metadata, validation, methods for retrieving information in different areas (biology, environment, civil protection, oceanography, social media, humanities and social sciences, ...), infrastructural issues related to language resources. Active in many standardisation activities for harmonising lexical information. Involved and responsible of the Pisa team in many international projects for language engineering. Over the last years, she has published articles in the field of lexical resources and information extraction in different areas. Currently, she focused her activities on digital humanities. Member of various Scientific Committees; UNI delegate for ISO/TC37/SC4. Email: Monica.Monachini@ilc.cnr.it

O'Gara, Genya

83

Genya O'Gara is the Associate Director of the Virtual Library of Virginia (VIVA), a consortium of 72 academic libraries. In this position she implements consortial projects, coordinates assessment, develops collection management workflows, negotiates on behalf of members, supports committees and working groups, and assists in the preparation and management of consortial grants. She publishes and presents on emerging models of content development and assessment, with a focus on digital collections, scholarly publishing, and collaborative collection development.

Pardelli, Gabriella

105

Gabriella Pardelli was born at Pisa, graduated in Arts in 1980 at the Pisa University, submitting a thesis on the History of Science. Since 1984, researcher at the National Research Council, Institute of Computational Linguistics "Antonio Zampolli" ILC, in Pisa. Head of the Library of the ILC Institute

since 1990. Her interests and activity range from studies in grey literature and terminology, with particular regard to the Computational Linguistics and its related disciplines, to the creation of documentary resources for digital libraries in the humanities. She has participated in many national projects. Member of board at Institute for Computational Linguistics. She is author and co-author a number of publications dealing with Computational Linguistics, Computational Terminology and Grey Literature. Email: gabriella.pardelli@ilc.cnr.it

Pizzanelli, Miguel

123

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.

Savić, Dobrica

19

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

Author Information (CONTINUED)

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: orcid.org/0000-0003-1123-9693

Schöpfel, Joachim

117

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 GERiCO 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>

Smith, Plato L.

75

Plato Smith is the Data Management Librarian at the University of Florida with experience in academic research libraries, digital libraries, and data management. He received his doctorate in the field of Information Science from the School of Information within the College of Communication and Information at Florida State University, Florida's iSchool, Summer 2014. From 2005 to 2012, he was Department Head for the FSU Libraries' Digital Library where he developed, populated, and managed digital collections in the FSU Libraries' digital content management system, DigiNole Repository, and electronic theses and dissertations (ETDs) institutional repository.

Email: plato.smith@ufl.edu

ORCID iD <https://orcid.org/0000-0003-1814-0151>

Stein Kenfield, Ayla

83

Ayla Stein Kenfield, Metadata Librarian at University of Illinois at Urbana-Champaign (UIUC). She supports the metadata needs for scholarly communication, data curation, and preservation in the Library. She has published and presented on digital repository evaluation, metadata development for data repositories, and digital library system migration. Her research interests include digital repositories; metadata and linked data; and the place of metadata in critical librarianship.

ORCID iD <https://orcid.org/0000-0002-6829-221X>

Thompson, Santi

83

Santi Thompson, Head of Digital Research Services at the University of Houston (UH), serves as Primary Investigator for the grant. At UH Santi develops policies and workflows for the digital components of scholarly communications, including digital research support and digital repositories. He publishes on the assessment of digital repository metadata, software, and content reuse. Santi is currently the co-facilitator of the DLF AIG.

ORCID iD <https://orcid.org/0000-0002-0337-6439>

Tulloch, Meg

91

Meg Tulloch is Executive Director of the Federal Library and Information Network (FEDLINK), Library of Congress. She is the former Library Director of the National Defense University Libraries in Washington, D.C. and Norfolk, Virginia. Previously, she was the Europe Region Librarian for the U.S. Army and oversaw 26 libraries in four different countries. She has also worked as a librarian at Vanderbilt University's Walker Management Library and Kutztown University of Pennsylvania's Rohrbach Library. Much of her career has focused on how technology can assist the researcher through digital library tools, using digital materials. Additionally, Meg taught "Introduction to Poetry Writing" at the University of Virginia while a graduate student there. She holds a Masters in Library and Information Science, a Masters in Fine Arts in poetry writing, and a Bachelors in American Literature. She is currently pursuing a Doctorate of Liberal Studies from Georgetown University. Her dissertation will explore fragmented twenty-first century literature.

Email: mtulloch@loc.gov

Vaska, Marcus

141

Marcus Vaska is a librarian with the Knowledge Resource Service, Alberta Health Services, 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: mmvaska@ucalgary.ca

ORCID iD <https://orcid.org/0000-0002-4753-3213>

Woolcott, Liz

83

Liz Woolcott, Head of Cataloging and Metadata Services at Utah State University, manages the MARC and non-MARC metadata creation of the University Libraries and is the co-founder of the Library Workflow Exchange. She publishes and presents on workflow and assessment strategies for library technical services, innovative collaboration models, the impact of organizational structures on library work, creating strategic partnerships for libraries, and building consortial consensus for metadata standards.

'Research Data Fuels and Sustains Grey Literature'**Loyola University New Orleans, USA • December 3-4, 2018****Publication Order Form****TWENTIETH INTERNATIONAL CONFERENCE ON GREY LITERATURE**

Publication(s):	No. of Copies	x	Amount in Euros	Subtotal
GL20 CONFERENCE PROCEEDINGS - Printed Edition ISBN 978-90-77484-33-3 ISSN 1386-2316 <i>Postage and Handling excluded^{*)}</i>		x	109.00 = €	
GL20 CONFERENCE PROCEEDINGS - PDF Edition ISBN 978-90-77484-33-3 ISSN 1386-2316 <i>Forwarded via email</i>		x	109.00 = €	
GL20 Conference Proceedings - Online Edition ISBN 978-90-77484-33-3 ISSN 2211-7199 <i>Password Protected Access</i>		x	109.00 = €	



POSTAGE AND HANDLING PER PRINTED COPY^{)}*

Holland		x	5.00	€
Other		x	15.00	€
TOTAL EURO =				€

Customer Name:

Organisation:

Postal Address:

City/Code/Country:

E-mail Address:

☐ Direct transfer to TextRelease, Rabobank Amsterdam
BIC: RABONL2U IBAN: NL70 RABO 0313 5853 42, with reference to "GL20 Publication Order"

☐ MasterCard/Eurocard ☐ Visa Card ☐ American Express

Card No. _____ Expiration Date: _____

Print the name that appears on the credit card, here _____

Signature: _____ CVC II code: _____ (Last 3 digits on signature side of card)

Place: _____ Date: _____

NOTE: CREDIT CARD TRANSACTIONS WILL BE AUTHORIZED VIA OGONE/INGENICO DESIGNATED PAYMENT SERVICES

TextRelease
www.textrelease.com

GL20 Program and Conference Bureau
Javastraat 194-HS, 1095 CP Amsterdam, Netherlands
Tel. +31-(0) 20-331.2420 Email: info@textrelease.com

Index to Authors

A-B-C

Bartolini, Roberto	105
Biagioni, Stefania	105
Bossart, Jean	75
Cardillo, Elena	123
Carlesi, Carlo	105

D-E-F

Darmoni, Stefan	123
Dressel, Willow	67
Etkin, Cynthia	12
Farace, Dominic	117
Frantzen, Jerry	117

G

Giannini, Silvia	51
Goggi, Sara	105
Gonzalez, Sara	75
Grootveld, Marjan	40
Grosjean, Julien	123

H-I-J

Henderson, Kathrine A.	25
Hersey, Denise	67
Hollander, Hella	40
Hsu, Chiehwen Ed	123
Itabashi, Keizo	99
Ittoo, Ashwin	123
Jamoulle, Marc	123

K

Kanazawa, Masashi	99
Kelly, Elizabeth	83
Kraaikamp, Emilie	40
Kumazaki, Yui	99
Kunii, Katsuhiko	99

L-M-N

Li, Yuan	67
Lipinski, Tomas A.,	25
Molino, Anna	51
Monachini, Monica	105
Muglia, Caroline	83

O-P-Q

O'Gara, Genya	83
Pardelli, Gabriella	105
Pizzanelli, Miguel	123

R-S

Resnick, Melissa P.	123
Roorda, Dirk	40
Savić, Dobrica	19
Schöpfel, Joachim	117
Shamenek, Frank S.	123
Smith, Plato L.	75
Stein Kenfield, Ayla	83
Suzuki, Satoru	99

T-U-V

Thompson, Santi	83
Tulloch, Meg	91
Vander Stichele, Robert	123
Vanmeerbeek, Marc	123
Vaska, Marcus	141

W-X-Y-Z

Woolcott, Liz	83
Yonezawa, Minoru	99
Yue, Cui	133