Data: Is it Grey, Maligned or Malignant?*

Julia M. Gelfand and Daniel C. Tsang (United States)

Abstract:

Cancers, growths, past events, social issues, conditions, and trends are each proverbially described as on a spectrum from maligned to malignant and scientists, physicians, journalists, commentators, politicians and other specialists offer opinions and commentary on what frames the answer to this question of the title. This paper explores not just the color and tone of data, but attempts to resolve what characterizes whether data is maligned or malignant. Hues of greyness distinguish the perils of failing to share, publish nor make accessible research data and the contemporary consequences to scholarship and open access are critical objectives in today's information arcade. Access to data is determined by those who can afford it, discover and know about it, and can thus manipulate it. Grey literature can take the offensive approach to further the role of data, and promote it to advance the common good, contribute to social responsibility and human actions. Data, while increasingly ubiquitous and abundant is the driver of evidence-based foundations, and the link to academic credibility, communication, discourse, dialogue, and the platform for greater open access. Grey data, possessing some of the attributes of grey literature, difficult to identify, acquire and access, when endangered or threatened, not archived or preserved, requiring methods to organize, sort and stratify, forces nontraditional publishing to pursue data publication to enhance perpetual access and new interpretations for its utility in future learning and research applications. We know that there is a somewhat elevated likelihood that open data policies lead to more widespread knowledge and information sharing, greater self-confidence among information providers and scholars alike, but we know less of whether these patterns have any short or long term benefits or disadvantages for individuals or society and of the factors that moderate and mediate these effects. In the meantime, the new reality is that data is central to the work of science, social sciences and basic human conditions of health and wellbeing and data policies mostly proceed from a grey containment to this new reality. The argument that as libraries become active publishers by digitizing content, creating new content, supporting researchers by addressing new domains and formats, that other interpretations of grey data and data more generally are increasingly plausible and that further research on the factors moderating and mediating the effects of data management is needed. This paper explores the continuum for data from maligned to malignant and anticipates data approaching the benign stage emphasizing new hues of grey and open access.

Definitions of Grey Literature & Grey Data

There is much to reconcile – the analogies versus disconnects between Grey Literature and grey data and the relationships between maligned and malignant and how those terms can apply or describe the expanding sphere of grey data today. The grey literature movement with origins now nearly four decades old documents how parallels between original grey sourcing impacts new components of scholarship such as data in the contemporary research setting. From the 1970s when technical and government report literature was the domain of grey literature championed by Charles Auger in his early compilations[1] or a more expansive definition, coined in 1997 that became known as the "Luxembourg definition," which included, "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."[2] An additional clause was added in 2004 noting that grey literature "is not controlled by commercial publishers, or "where publishing is to the primary activity of the producing body."[3]

Fast forward to 2010 when this conference met in Prague and there was a call for some revisions to the definition that reflected more online, electronic or born digital publishing. It was proposed then that four attributes should now qualify the basic definition:

- Documents character of grey literature – more multidisciplinary in content and form
- Includes legal nature of works of the mind – protecting intellectual property
- Reinforces quality level for peer review, validation
- Links to intermediation bridging collection status to readers or users

* First published in the GL16 Conference Proceedings, March 2015.

And the new "Prague Definition" for grey literature was defined: "Grey literature stands for 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, ie, where publishing is not the primary activity of the producing body."

This definition has been challenged as still being too narrow considering new forms and practices of scholarship and research underway and the methods of publishing now widely available. The emphasis on different forms of publishing and the added dissemination streams of content reaching people in new ways; the constant influx of new technologies; incorporation of new media; containing raw research including data, interviews, previews; new forms of single and combined authorship; and the added concerns for preservation, stewardship and perpetual access regardless of future revisions; are all concerns for a current definition of what constitutes being grey or called "grey literature."

Data for the purposes of this paper is defined as research data and we have embraced the standard OMB definition, "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings."[4] Data can be as effusive as grey literature, reflect disciplinary intersections as well as stand on its own and be as varied as spreadsheets, laboratory or field notebooks, codebooks, protein or genetic sequencing, spectra, test plotting, samples, specimens, digital objects, database elements, models, algorithms, formulas, simulations, methodologies, workflows, operating procedures, protocols, and the like. What research data does not include are trade secrets, commercial information, information protected under law, and personnel or medical information that would clearly constitute a clearly unwarranted invasion of privacy. Balancing privacy and data usage will remain one of the biggest challenges in working with data and will contribute to the ongoing "greyness" associated with data.

Data Science as a discipline is also becoming a programmatic area that crosses mathematics, statistics and economics and spans across all disciplines. The definition at our own institution includes how it "may encompass the full spectrum of theories and methods that use data to understand and make predictions about the world around us." This includes "fundamental research on statistical methods, prediction algorithms, data management techniques and policy issue; as well as a broad range of domain-specific data-driven research problems in the sciences, engineering, humanities, education, medicine and business/management."[5] It is a broad range of "theories, algorithms, methodologies and tools that allow us to use data to better understand and make predictions about the world around us."[6]

Some examples of common elements that bridge grey literature with grey data include:
correspondence, project files, grant and ethics applications, technical reports, research reports, signed consent forms.[7]

Maligned or Malignant?

Why, you may inquire do we pose the spectrum of maligned or malignant to define our thesis? Maligned is used in the sense of speaking ill of someone or something, being injurious, slanderous or defaming by causing one to conclude very differently about given circumstances. An example may be when shown to be evil in disposition, nature or intent in order to conclude different outcomes. Malignant, known for its pathological or cellular definition of tending to produce a bad outcome such as death from cancer and can be highly invasive, dangerous, cause harmful influence or effect suggests something out of a researcher's control but will have to be explained due to some other force or context. Thus our imagery shows the extremes of confusion to an example of the recent anthrax episodes that generated fear and uncertainty about the potential exposure and outcomes.

We were influenced by the psychologist, Professor K. Alison Clarke-Stewart who framed some important questions about social issues such as infant day care as "maligned or malignant?"[8] She assessed what kinds of available data perhaps best suggest there could be separation anxiety among infants from their mothers who returned to work during their early infancy and what risks these infants would have for future emotional insecurity and social maladjustment when placed in daycare. She also asks what the empirical evidence is concerning the effects of infant day care – "it is truly bad for babies or it is undeservedly maligned,"[9] and goes on to examine what evidence there is to support either supposition.

The conclusions of social science and other forms of scientific research reveal that there is much to learn about the effects of something being studied or the outcomes. We can conclude that perhaps there are elevated or declining likelihoods of something happening, we can establish patterns, review longevity data, and understand that different forces contribute to maturation, change and outcomes. We know about situational evidence or trends in work environments, lifestyles that can influence outcomes without observing or studying those contributing factors, such as climatic or environmental occurrences, catastrophic or unexpected events, political interpretations or activities, market surges or declines, and thus the issue today is not whether data is sufficient for the research being undertaken but what else can we learn from the entire research experience and landscape that influences what we observe and learn. That begins with the research question, how the research is conducted and whether sufficient data is collected to offer any conclusions, definitive or not.

Perceptions of data by different communities

Data used to be the stepchild of the research publication process – considered as not worthy of including with a publication, too complex to describe adequately, or too bulky to attach and beyond the needs of most readers. Funding agencies until recently never gave much thought about how to manage research data, even data generated with taxpayer funds, so data were left to rot literally (as computer tapes or disks deteriorate over time), without any thought of preservation and retention of the data underlying scientific research. Like grey literature, research data can be termed as largely fitting the concept of grey data – unfit to see the light of day. They tended to be left on floppy disks or other removable storage devices, often unintelligible to anyone but the original data compiler and over time were unable to be deciphered. Unless the data are collected as part of a well-funded large-scale study, individual datasets are often forgotten after the research is finished and a paper produced or publication issued. Although the scientific ethos mandated reproducibility, research data was more often than not left in filing cabinets or on computer drives inaccessible to others.

That has begun to change. With the advent of concepts like Data Paper[10] and the development of mandates from US federal agencies including the National Institutes of Health (NIH) and the National Science Foundation (NSF) to share and deposit research data for public access, one suspects that data has suddenly been elevated and given a more prestigious status. The concept of data paper is to create a searchable metadata document where instead of focusing on a journal article where hypotheses and conclusions are central, the data becomes the primary emphasis. Increasing examples of this are known. This has been done by the environmental sciences community that studies biodiversity, and these scholars achieve recognition like in other forms of scholarly and creative outputs, by increasing the visibility, usability and credibility of the data resources now cited and published.

When Megan Smith, Chief Technology Officer of the United States responded in an interview when asked how her office is working to make large sets of government data public, "Scientists and universities and the general public can do extraordinary things with it. It could be weather or climate data; it might be data from the Department of the Interior or NASA or water data. Whole industries are being built from things that taxpayers have helped the government know."[11]

Still, efforts to develop new ways of managing research data while encouraging researchers to share their data have been stymied at the individual level. Researchers may be unwilling to share their data, or see little utility in making it available to a public likely to be uninformed about how to interpret the data. Efforts to teach individual researchers about research data management are time consuming and in the main not supported by funding agencies. Funding for research data management is more likely to become more available to institutions than to individual researchers or research projects. Also, the mandates encourage repurposing data may become a more central practice than constantly funding variations of the same research question. Data has demonstrated that it is more than "nice to have," or being relegated to second class citizenry, for when data is misunderstood it changes the tenor or disposition in fundamental and profound ways.

Risks & Conditions of Data: Unevenness, authoritativeness, misinterpretation

Data is often easily misinterpreted, especially if research results do not contain enough information to place the research in context. How the data was collected, the sample size versus the universe, whether the data was adequately described, and who funded the research are issues any outsider might want to have answered.

Metadata is often inadequate and inconsistent, filenames are subject to private whim, and documentation of steps in the research process are often lacking. Researchers thus can make bold and extravagant statements about research findings – unfortunately not backed up by their data upon closer perusal. A study may not be in fact generalizable but that may not deter the mass media to tack on to the latest research and proclaim a scientific "breakthrough." Traditionally research subjects in some fields were limited to who were readily available, say, college undergraduates of a certain ethnicity and demographic. Whether such findings are in fact generalizable to a larger population, another setting, or another time period is dubious.

Repurposing the research frontier

With the last few decades demonstrating that an increasing amount of research is done collaboratively with different specialists sharing how to examine and study a problem, bringing their specialized skills and objectiveness to the frontlines, and having the desire to share research findings with other likeminded citizenry, several conclusions are evident. They include the need to share research data widely. This can be from the vantage point of international or global sharing with the focus on making sure colleagues and inhabitants in developing countries who are already challenged by not having as dependable or consistently good access to information, get it.[12] It also serves the ongoing development of ICTs, statistical analysis and the intensity of interdisciplinarity across many subject areas and the entire academic landscape.

The directive of working or conducting research in a more open and networked environment, now increasingly commonplace suggests that the cyberinfrastructure that we come to rely upon is more complex than we perhaps anticipated as it was launched. With the unknowns came opportunities and the role of the Internet over the last two decades has promoted connectivity, exchange, collaboration, and reduced barriers and thus redefined new indicators for competitive intelligence and for intellectual competitiveness, a sense of evidence-based models and the ongoing need for best-practices. Supercomputing now available in a handheld device offers new metrics and ease of computation that can be achieved more economically and efficiently. Being committed to the tenets of sustainability, whether it be economic, social or environmental[13] reinforces how the research process ensures integrity, accessibility and stewardship of research data in the digital age.

Data is often delegated to the cloud. Out of sight but not out of mind, data has found a parking spot in the cloud where costs are less and options more plentiful in terms of what to park.

Role of Libraries: Finding & supporting data

Library catalogs, finding aids and most recently the discovery systems that consolidate access to all content via this new breed of discovery tools support data in new and novel ways. By engaging with social networks, media, and other innovative technologies such as QR Codes, discovery systems from a user, contextual and content owner or publisher perspective will impact a greater reliance on data.

The Internet Manifesto 2014[14] widely adopted by the library community and its publishing partners ensures equitable access to the Internet and its services in support of freedom of access to information and freedom of expression and data related content is no exception to this. As the current chair of IFLA's Committee on Freedom of Access to Information and Freedom of Expression, (FAIFE) writes about this update,

"The world has changed significantly since 2002 both physically and digitally, and we now have a greater experience and understanding of the role of the Internet and digital resources in our services, and in developing connected societies where individuals have the skills that they need to exploit the opportunities that technologies can bring. We also have a greater understanding of the threats that can be posed through the Internet including the impact on human rights of inappropriate monitoring and surveillance, and from criminal activity....reflects this experience and reinforces the vital role of library and information services in ensuring equitable access to the Internet and its services in support of freedom of access to information and freedom of expression."[15]

Libraries, even experiencing these trends remain insufficiently staffed and ill-equipped to handle research data management, but they welcome the opportunity to manage data much like they have taken on different elements of library holdings. Managing research data is not the same as oversight for digitizing printed materials or archiving and preserving digital objects. Data needs to be curated, backed up but also restructured as formats and platforms change. In other words data curation is not just dumping data in a repository, providing access, telling users where it is and considering that sufficient. The level of data curation will vary depending on the resources, skill-sets and time a library or institution can devote to it.

Typically a library or data repository receives a dataset at the end of a research life cycle, often decades after – thus posing often intractable challenges to the data archivist.
Efforts now are being devoted in some institutions to involve researchers at the beginning or during the research life cycle in best practices in research data management. But such efforts are labor intensive, varies by discipline, and often unfunded.

Data Professionalization

With the emphasis of this paper positioned on the role of libraries and librarians handling research data, we should not discount how the commercial world is also responding to the need to create more reliable data management services. The advertising industry is partnering with the computer industry to explore how it "turns Big Data into new ideas." Ogilvy & Mather, among the largest agencies with global revenue, has a new idea about stepping up a longstanding commitment to data-driven decision-making by forming a unit named OgilvyAmp where "Amp" is short for "amplify" and is being directed by a leader with the new job title of global chief data officer. The previous emphasis on information evangelists and such gurus has led to this latest series of tasks that include data strategy and planning, analytics services and data management continuing the way innovative companies such as Ogilvy Mather has amplified the creative process when advertising is developed by paying attention to data with sayings like "Advertising people who ignore research are as dangerous as generals who ignore decodes of enemy signals." The goal of companies such as Ogilvy is to "remove data as a distraction and position it as a tool in the creative palette." Descriptions note, "As data becomes more complex, we want to move data up in the process I support of the creative...You need someone to translate stories out of the data instead of just giving the creative a dashboard." "Creative departments should embrace data as part of their raw material, instead of seeing it as taking their power away." "Data doesn't replace creativity...but great creativity has lasting power that data may not have predicted."[16]

In other innovative business and management applications, data has value if it contributes to efficient and cost-effective savings. Data gains tractions if it has positive outcomes for:

- Reduces building & investment costs
- Allows for better and improved cash flow during projects
- Encourages new opportunity costs – using new apps to interpret results of search or data analysis
- Promotes a more green and sustainable environmental landscape[17]

Data mastery is highly associated with a digital vision. As companies and organizations create their digital vision, they are experiencing a digital transformation, although a compelling vision of a digital future still remains unclear. Technology can remove obstacles and extend capabilities but to enhance users or customers' experiences, streamline operations and transform operational methods and business models is a never ending process.[18]

Back to the library

Today we see positions posted in nearly every research library around the world trying to recruit for professional librarians who can offer appropriate and relevant services to be delivered by a team of research data curation librarians. Sometimes called eScience or eResearch Librarians, Data Librarians or Digital Initiatives Librarians, the work that these professionals now conduct reflects the scholarly communication environment we find very commonplace, where we want to engage users as authors, and develop collections and services that directly support teaching, learning and research missions of the institutions that are trying to re-imagine the contemporary research library.[19] These positions are commanding a lot of attention from library users seeking directions in where to publish, how to publish, how to provide accompanying research data and how to share and repurpose data. Not easy to fill, there remains insufficient competition for all the slots that are being created. Related is the province of statisticians and indicators of professional forecasting does not suggest that there will be an adequate supply of them either for the needs in the marketplace.

In addition, new data intensive environments rely upon the work of semantics, digital humanities, web technologies, human computer interaction, data-mining, information retrieval, communication, strong web programming skills to forecast for tomorrow's new discoveries. As Steve Lohr writes, we still must be mindful that this popular new trade contains far too much handcrafted work or "data wrangling," "data munging" or "data janitorial work." It has been estimated that between "50-80% of data scientists' time" is devoted to "collecting and preparing unruly digital data, before it can be explored for useful nuggets."[20]

Repositories

Many libraries increasingly have institutional repositories and digital commons to contain the scholarship from their affiliates. These repositories are in addition to the thematic or subject repositories that have evolved where one stop deposit and consultation is increasingly the norm to stay informed about the latest contributions to a field. Examples of this include ArXiv, with nearly a million ePrints in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance and Statistics (http://arxiv.org/), the Social Sciences Research Network (http://www.ssrn.com/en/) that supports a score of subject areas. There are also major incentives to not disenfranchise non-recognized or non-distinguished members of the academy but to encourage undergraduates and the student population to also deposit their creative outputs. Examples[21] of journals containing student work, independent projects and group work are increasingly gaining recognition as seen by the detailed work at Illinois Wesleyan University, a small independent liberal arts institution. Such efforts are examined in the increasingly common efforts to promote open data, open education and just reduce the barriers for access. A recent emphasis that addresses these issues is the intersections between scholarly communication and information literacy that librarians have been embracing for the past few years where new library services have been developed to offer supplemental instruction about the value of each of these significant roles.[22]

Just as we continue to explore grey literature, grey data, open data, open science, open education we celebrate ten years as plans are underway for the Tenth International Conference on Open Repositories to be hosted cooperatively by the Coalition of Networked Information (CNI) and several institutions in June 2015. That proposed agenda is a call for many topics to be explored including how repositories can support the needs of research data via data registries, storage, curation lifecycle management, management and digital preservation tools. In addition, the CNI is also tracking the establishment of digital scholarship centers or labs in universities and colleges and hosted a program in April 2014 that offered participants a way to examine the intersections of Scholarly Communication, Visualization and Digital Scholarship.[23] The data element was very visible in this work environment.

How data is structured

At a recent forum on our campus highlighting the new Data Science Initiative[24] we learned how distinctions between statistics and computing have merged and blurred. Related to this may be the parallels of machine learning to the operations of brain functioning. IBM has Watson and it has been shown that learning has similarities between the human and a programmable robot. The same deep learning problems can be applied across the sciences to physics, chemistry, biology and related disciplines. Depending on computer science, programming, machine learning, statistical analysis, varied technologies, one can integrate applications with understandings so that they can be ready for analysis.

Impacts of Open Access

The Open Access movement has contributed significantly to perceptions, expectations and directions of open data. Publishing opportunities and scholarly communication directives illustrate many examples of how government and the private sector have independently and collaboratively produced new information products, resources and methods of handling data. The pricing of free content does not always come without other associated costs and as soon after the launch of the Educational Resources Information Center (ERIC) by the US Department of Education in 1966 we have seen a very broad range of products and methods of information production.

Highlights include the following launches as noted on Peter Suber's Open Access Timeline:

- Agricola database in 1970 from the National Agriculture Library
- USENET created in 1979
- Text Encoding Initiative (TEI) in 1987
- Psycoloquy among first free online journal that became peer-reviewed in 1989
- First web page was debuted by Tim Berners-Lee in 1990
- Preprint Archives, Mathematical Physics and arXiv launched in 1991
- GenBank launched by the National Center for Biotechnology Information (NCBI) in 1992
- CERN launched its preprint server in 1993
- Human Genome Project, Social Science Research Network (SSRN) and NASA Technical Report Server went live in 1994
- Scientific Electronic Library Online (SciELO) went online in 1997
- Wikipedia was born in 2001
- Budapest Open Access Initiative (BOAI) becomes reality in 2002
- Several scholarly communication initiatives launched in 2002
- Public Library of Science (PLoS) obtains funding to start tow open access journals in 2002
- Google began to digitize and index millions of public domain and copyrighted books from five major libraries in 2004
- From 2004 on, aggressive activity on a global scale took place exploring how to produce open access content and share worldwide – from governments, academic institutions, non-profits, philanthropic foundations, and hosts of individuals with journal declarations being among the most common examples of new information resources[25]

Examples of Data

One flavor of data does not apply across everything. Some will be greyer than others but making generalizations remain problematic. One should be careful and cautious when defining data and its properties. Some examples and current experiences follow:

Survey Data

Common elements include but are not restricted to consumer behavior, trends and, demographics. In the social sciences, survey data is typically structured as a rectangular alphanumeric file, arranged by respondent responses to multiple-choice questions. Areas where such data are relied upon include demographic, public opinion, political participation and consumer behavior research. Given the current importance of social media in social movements etc., researchers are likely to become less interested in structured survey data and will want ways to analyze data that are structured differently, e.g. as tweets. In other words data mining will become a necessary core skill.

Big Data

Getting the basics down for understanding Big Data is easily done by reviewing Cathy O'Neil and Rachel Schutt's book, *Doing Data Science.*[26]

The report, *Big Data: Seizing Opportunities, Preserving Values* issued by the White House this year and currently under review[27] suggests the importance placed on big data due to the increasingly available sensors, cameras and geospatial technologies that can track global movements. The purpose of this report was to determine how big data will transform the way we live and work and alter the relationships between government, citizens, businesses and consumers. The perception is that big data technologies will be transformative in every sphere of life. Big data according to some experts is "fundamentally reshaping how Americans and people around the world live, work and communicate. It is enabling important discoveries and innovations in public safety, health care, medicine, education, energy use, agriculture and a host of other areas. But big data technologies also raise challenging questions about how best to protect privacy and other values in a world where data collection will be increasingly ubiquitous, multidimensional and permanent."[28] As exciting as knowledge discovery is, and how intensely the Internet has changed how we live our lives, big data as it is defined does not come without challenges and concerns.

The definition of big data is "The capability to manage a huge volume of disparate data, at the right speed and within the right time frame, to allow real-time analysis and reaction." Big data is typically broken down by three characteristics, including volume (how much data), velocity (how fast that data is processed), and variety (the various types of data).[29] Data virtualization and data warehousing is critical for businesses as well as for research practices.

What the US Federal Government has learned so far is that there is concern with data practices.

- Big data tools can alter the balance of power between government and citizen
- Big data tools can reveal intimate personal details
- Big data tools could lead to discriminating outcomes[30]

The policy recommendations made so far include:

- Advance the Consumer Privacy Bill of Rights
- Pass National Data Breach Legislation
- Extend Privacy Protections to non-US Persons
- Ensure Data Collected on Students in School is used for Educational Purposes
- Expand Technical Expertise
- Amend the Electronic Communications Privacy Act[31]

It is easy to agree with Emma Uprichard that we are all consumed by the "challenges of big data." We are led to believe that big data "brings new hope to big social problems and social policy that big data will help us deal with crime and terrorism, intervene with social problems and social policy and may be cheaper to use than organizing large-scale official surveys."[32]

From this we can conclude that technologies are driving the potential for current practices that may accurately or inaccurately represent what we value most. In the spheres of science, technology, and medicine, some ongoing concerns include, data formats, the ambiguity of human language, the need to repeat applications in data projects.

Examples of tools that make data less grey include:

- Spreadsheets – Excel has been around a long while and offers methods to create charts and calculations to allow for models
- ClearStory Data[33] – software that recognizes many data sources, pulls them together and presents the results visually
- Statista[34] – commercial sourcing of data that gives public & private data for eSearchers information in multiple options of delivery offering infographics
- Trifacta[35] – uses machine learning technology to find, present, and suggest types of data that might be useful to see and explore
- Paxata[36] – automates data preparation for analysis
- MapReduce[37] – a software framework that allows developers to write programs that process massive amounts of unstructured data in parallel across a distributed cluster of processors or stand-alone computers
- Data Conversion Laboratory (DCL)[38] – been in the marketplace since 1981; uses optical character recognition (OCR) to integrate data from multiple sources such as scanning, MS documents,

Each of these can be applied to social science data, medical, environmental, science, and interdisciplinary contexts that blend and create new understandings and lend to new knowledge generation. Data and digital conversion, storage, archiving, description and analysis are all the more possible due to "automation and technology that lets us transform things without needing to build things anew every time."[39]

Credibility of Data

"Big data may mean more information, but it also means more false information," and even when the information may not be false the problem is "that the needle comes in an increasingly larger haystack."[40] That is the problem most researchers and scholars face.

Contemporary studies of mindfulness and how it contributes to creativity and innovation have been in both the scholarly and popular media. "There are some who believe the increasing power of Big Data (using powerful computers to sift through and find patterns in massive amounts of information) is going to rival the human consciousness at some point. But there's also growing skepticism about how effective Big Data is at solving problems."[41]

All data are not equal – "experts" may "trust" certain data over others. Source credibility is one big factor – who is compiling the data and is that institution's work reliable? Are its methods transparent? And is there enough documentation to allow a subsequent user to understand how the original data was collected, arranged, described, massaged, and cleaned. This has led to the concept of apply a data seal of approval to data archives – for example, ICPSR, the major social science data repository at the University of Michigan, has obtained a "Data Seal of Approval" for its work.

Researchers are concerned about relationships and conclusions that their findings and data suggests. Spurious correlation, defined by Pearson in 1897 describes the correlation between ratios of absolute measurements that arises as a consequence of using ratios rather than because of any actual correlations between measurements[42] and is one of the main motives for the field of compositional data analysis which deals with the analysis of variable that carry only relative information, such as proportions, percentages and parts-per-million.[43] Spurious data, relationships and correlations may offer respite and humor but the seriousness of when they enter mainstream media or are introduced erroneously is problematic. These examples of relationships of total revenue generated by computer arcades correlating with the number of computer science doctorates awarded provides mixed up interpretations as the computer gaming industry matures. Another example is the number of oil imports correlating with consumption of high fructose corn syrup.[44] Pearson's definition is obviously not to be confused with misconceptions about correlation and causality.

Data Publishing & Publications

New information products issued by a variety of sources globally support individual researchers and functions of work conducted in different sectors. Specific resources developed by the academic community with which we are most familiar include the Data Management Plan created by the California Digital Library (CDL), and other widely respected efforts generated at Purdue, in the UK and at many institutions.[45]

The University of California has also created DASH, a data sharing platform for researchers at the 10 different campuses.[46] These resources are open source and can be used by institutional adoption at other institutions. Each of them contributes in unique ways to data sharing and ultimately to new knowledge generation.

Funding agencies requirements vary regarding the retention periods for data generated from sponsored research, depending on discipline, and form of data. Data longevity thus may vary by different indicators. This implies that data can be "weeded" – a heresy to some data archivists whose archives had missions to preserve and retain data in perpetuity. Before deciding to weed data, attention needs to be given to whether the data can be repurposed to meet current research needs, and whether format migration is feasible. The retention vs de-selection debate remains controversial and problematic, with tools and economic consequences allowing for a longer lifespan and less expensive storage of data with ease of providing better metadata describing the data.

Looking Forward: The next horizon for grey data

Recent research and scholarship has shown a large wave of interest in data management in this accelerating digital age. We must continue to be vigilant about the "validity of research data, standards that do not keep pace with the high rate of innovation; restrictions on data sharing that reduce the ability of research to verify results and build on previous research; and huge increases in the amount of data being generated, creating severe challenges in preserving that data for long term use."[47]

Reflecting on the Internet Manifesto, the current trajectory and the landscape we think data, especially grey data instills, we can't forget the diverse nature of library work and the opportunities we now have with diverse technologies and new skills individuals must develop and realize that it too, will continue to be part of an evolution of services and resources in libraries. As Steven Bell writes, "But what if, in addition to other core values instilled in LIS programs such as protecting patron privacy and defending intellectual freedom, LIS students learned and practiced methods for tackling tough problems and developing thoughtful solutions – in any situation. Academic libraries need LIS graduates who can assess situations and resources, identify the source and nature of a problem, and then craft an appropriate solution. In other words, educate [the next generation] to identify, frame, think through and solve problems the way designers do."[48] Perhaps we will have strategists, technologists, designers and curators in the role of librarians to manage data.

Conclusions: Maligned or Malignant?

If scholars think that data is maligned it may be because of this "liberal notion" we possess regarding how evidence-based decision-making is superior to one without evidence. The specialty field of logistics and decision sciences points to this. Let's be clear that not all evidence contains data and not all data constitutes evidence concerning the issue at hand. Data may be relevant or it may not be. Hence the "wrong" or misappropriate date that is selected should be worthy of being maligned. Data is often not presented in ways that lets us judge if it is worthwhile or contains relevant data. Hence the dilemma – is it maligned or malignant? How much malaise or data fatigue is there in the work in which we engage? How do we plan accordingly? Strategically, for policy-makers, data is seen through preconceived notions, if it matches one's preconceptions, it is considered "good" data; otherwise it may be maligned.

Marcus Banks writes in 2010 how his interest in grey literature has shifted over the previous five years.[49] He speculates that grey data is compatible with grey literature due to the Web 2.0 directives and new opportunities to disseminate via many new avenues such as blogs, social media, tweeting. The end product or tweets has become "grey data" but we urge our audience and readers to consider the research by-products of both formal science and the social content or new methods to communicate it as worthy of retention and preservation. Libraries are currently trying to influence governments and large enterprises to save, authenticate and preserve older runs or versions of data for future generations to compare, study and align with recent or current datasets, but that continues to be a challenge based on available resources, financial and human.

Malignant data can be construed as being misunderstood, clumsy or insufficiently analyzed – a sense of false positives, incomplete data, or a set of data that cannot be replicated using the same techniques and instruments for whatever reasons. This is increasingly problematic in the data sharing world. Funding sources may not want to consider supporting multiple efforts to study or engage in similar projects, however, it is the challenging outcome that may be providing clarity and a definitive analysis, not the original source. Building on previous work is the hallmark of research and taking it to the next level with more learned resourcefulness, an open and innovative mind makes it ultimately successful. Malignant data may still be vigorous but it is sick, perhaps suffering from malaise and fatigue and needs a new push to find its path.

Just two weeks ago, Pope Francis addressed the European Parliament with a grimly somber diagnosis about the state of Europe, that he described as a "continent's malaise," referring to how he perceives it to have lost its way, its energies sapped by economic crisis and a remote technocratic bureaucracy. It is increasingly a bystander in a world that has become "less and less Eurocentric," and that frequently looks at the Continent "with aloofness, mistrust and even at times, suspicion."[50] He went on to describe the aging product as "grandmotherly, no longer fertile and vibrant...and we cannot allow the Mediterranean to become a vast cemetery" and concluded another speech to the Council of Europe in Strasbourg on this same trip, by expressing more optimism, "It is my profound hope that the foundations will be laid for a new social and economic cooperation."[51]

We concur with O'Neil and Schutt that we would like to "encourage the "next-gen" data scientists to become problem solvers and question askers, to think deeply about appropriate design and process, and to use data responsibly and make the world better, not worse."[52] And to echo Pope Francis with his optimistic pledge for cooperation and not despair. That also makes data less maligned, malignant and grey.

References

[1] Auger, C.P., ed (1989). *Information Sources in Grey Literature, 2d ed.* London: Bowker-Saur.

[2] "Luxembourg Definition" of Grey Literature (http://opensigle.inist.fr/handle/10068/697932)

[3] Schopfel, J, Farace, D.J. (2010), "Grey Literature." In Bates, M.J., Maack, M.N. eds., *Encyclopedia of Library and Information Sciences, 3rd ed.*. Boca Raton, FL: CRC Press, 2029.

[4] OMB Circular 110 (http://www.whitehouse.gov/omb/circulars_a110#36)

[5] Data Science Initiative, University of California, Irvine (http://datascience.uci.edu)

[6] See http://datascience.uci.edu/about/

[7] Adapted from Defining Research Data (http://library.uoregon.edu/datamanagement/datdefinned.html) by the University of Oregon Libraries.

[8] Clarke-Stewart, K.A. (1989). "Infant Day Care: Maligned or Malignant?" *American Psychologist* 44, #2, 266-273.

[9] *Ibid*, 266.

[10] Data Paper (http://www.gbif.org/publishingdata/datapapers)

[11] Dominus, S. (2014). "You Can Affect Billions of People," *New York Times Magazine*, November 2: 14.

[12] Fienberg, Sharing Research Data (1985)

[13] Committee on Intellectual Property Rights and the Emerging Information Infrastructure, Computer Science and Telecommunication's Board Commission on Physical Sciences, Mathematics and Applications, and the National Research Council *(2000). Digital Dilemma: Intellectual property in the information age.* Washington, DC: National Academies Press. http://uclibs.org/PID/31331

[14] Internet Manifesto 2014. See http://www.ifla.org/publications/node/224

[15] eMail announcement from Martyn Wade, November 25, 2014.

[16] Elliott, S. (2014). "At Ogilvy, New Unit Will Mine Data," *New York Times*, October 21, 2014: B6.

[17] Lanser, J. (2014). "A Whole Lot of Data," *Orange County Register,* October 12. Real Estate Section: 6. 14.

[18] Westerman, G., Bonnet, D., and McAffee, A., (2014). *Leading Digital: Turning Technology into Business Transformation*. Boston, MA: Harvard Business Review Press, 113.

[19] See current job descriptions for Data Management positions such as at the University of Michigan (http://umjobs.org/job_detail/103342/research_data_curation_librarian or at the Pennsylvania State University Libraries (http://www.libraries.psu.edu/psul/jobs/facjobs/sdl.html) or at McGill University (http://www.mcgill.ca/library/files/library/14-al9904-01-coordinator-dat-curation-scholarly-communications.pdf) or at the University of Texas at Arlington (https://uta.service-now.com/jobs/recruiting_job_posting_detail.do?sysparm_document_key=u_recruiting_job_posting,a8f9b7706f6c75009f15cccc5d3ee4c3)

[20] Lohr,S. (2014). "For Big-Data Scientists, "Janitor Work' is Key Hurdle to Insights," *New York Times*, August 18: B4.

[21] See http://digitalcommons.iwu.edu/

[22] Davis-Kahl, S and Helmsley, M., eds. (2013). Common Ground at the Nexus of Information Literacy and Scholarly Communication. Chicago, IL: ACRL. http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/booksanddigitalresources/digital/commonground_oa.pdf

23 Coalition for Networked Information, Spring 2014 meeting, St. Louis, Mo. (http://www.cni.org/events/membership-meetings/past-meetings/spring-2014/)

[24] Data Science Initiative, University of California, Irvine (2014). (http://datascience.uci.edu/). Forum held October 24.

[25] Suber, P. (2008). Open-Access Timeline (http://legacy.earlham.edu/~peters/fos/timeline.htm

[26] O'Neil, C. and Schutt, R., (2013). Doing Data Science. Sebastopol, CA: O'Reilly Media.

[27] White House (2014) Big Data: Seizing Opportunities, Preserving Value (http://www.whitehouse.gov/sites/default/files/docs/big data privacy report 5.1.14 final print.pdf)

[28] *Ibid.* Fact Sheet, http://www.whitehouse.gov/sites/default/files/docs/big data privacy report 5.1.14 final print.pdf

[29] Hurwitz, J., Nugent, A., Halper, F., Kaufman, M. (2013). *Big Data for Dummies*. Hoboken, NJ: Wiley: 280.

[30] *Ibid*. Fact Sheet.

[31] *Ibid*. Fact Sheet.

[32] Uprichard, E. (2014), "Big Data Doubts," *Chronicle of Higher Education*, October 17: B14-15.

[33] http://www.clearstorydata.com/

[34] http://www.statista.com/

[35] http://www.trifacta.com/platform/

[36] http://www.paxata.com/

[37] http://hortonworks.com/hadoop/mapreduce/

[38] http://www.dclab.com/

[39] Gross, M. (2014). "Data Conversion in the 21st Century," *Information Today*, October: 26.

[40] Taleb, N. (2013). "Beware the Big Errors of 'Big Data.'" *Wired*, February 8. (http://www.wired.com/2013/02/big-data-means-big-errors-people/)

[41] Huffington, A., (2014). *Thrive: The Third Metric to Redefining Success and Creating a Happier Life*. New York: Random House: 140.

[42] Pearson, K (1897). "Mathematical Contributions to the Theory of Evolution – On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs," Proceedings of the Royal Society of London, 60:489-498. DOI: 10.1098/rspl.1896.0076

Proceedings of the Royal Society of London **60**: 489–498. doi: 10.1098/rspl.1896.0076.

[43] Pawlowsky-Glahn, V. and Buccianti, A., eds. (2011). Compositional Data Analysis: Theory and Applications. Hoboken, NJ: Wiley.

[44] Spurious Correlations. http://www.tylervigen.com/

[45] https://www.lib.purdue.edu/researchdata

[46] http://cdluc3.github.io/dash/

[47] Ensuring the Integrity, Accessibility and Stewardship of Research Data in the Digital Age (2009). Washington, DC: National Academies Press.(http://www.nap.edu/penbook.php?record id=12615)

[48] Bell, S. (2014). "MLD: Masters in Library Design, Not Science: From the Bell Tower." Library Journal, November 20, 2014. http://lj.libraryjournal.com/2014/11/opinion/steven-bell/mld-masters-in-library-design-not-science-from-the-bell-tower/

[49] Banks, M. (2010). "Blog Posts and Tweets: The Next Frontier for Grey Literature." In Farace, D., and Schopfel, J., eds., *Grey Literature in Library and Information Studies*. Berlin: DeGruyter Saur: 222-223

[50] Higgins, A. (2014). "At European Parliament, Pope Bluntly Critiques a Continent's Malaise," *NY Times,* November 26: A4, A10.

[51] *Ibid*.

[52] O'Neil and Schutt, chapter. 16.

Additional Resources:

Bessis, N. and Dobre, C., eds., (2014) Big Data and Internet of Things: A Roadmap for Smart Environments. Studies in Computational Intelligence, v.546. Cham: Springer.

Big Data Now: Current Perspectives from O'Reilly Media (2011). Sebastopol, CA: O'Reilly.

Chen, M., Mao, S., Zhang, Y., Leung, V, (2014). Big Data: Related Technologies, Challenges and Future Prospects. Cham: Springer.

Gauntner Witte, G. (2014). "Content Generation and Social Network Interaction within Academic Library Facebook Pages," Journal of Electronic Resources Librarianship 26 (2): 89-100.

Krier, L. and Strasser, CA (2014). Data Management for Libraries: A LITA Guide. Chicago, IL ALA TechSource.

Manning, P. (2013). Big Data in History. New York: Palgrave Macmillan.

Marcum, D, and George, G. (2010). The Data Deluge: Can Libraries Cope with eScience? Littleton, CO: Libraries Unlimited.

Pryor, G., Jones, S., and Whyte, A., eds. (2014). Delivering Research Data Management Services: Fundamentals of Good Practice. London: Facet.