Open Scholarship Initiative Proceedings, Volume 1, 2016
doi: http://dx.doi.org/10.13021/G8R30G

Report from the Information Overload and Underload Workgroup

Bryan Alexander, Kim Barrett, Sioux Cumming, Patrick Herron, Claudia Holland, Kathleen Keane, Joyce Ogburn, Jacob Orlowitz, Mary Augusta Thomas, Jeffrey Tsao

Abstract

The duality of information overload and underload is a defining issue of our age. Scholarly information is abundant but not universally accessible to all scholars and learners, thereby hindering or prohibiting equitable engagement in ongoing scholarly conversations. Access is a core aspect of the issue of overload and underload—both access to research materials and access to venues where one can contribute to the scholarly corpus—but it is not the only aspect. Our group agreed that the problem of overload is preferable to that of underload; however, the dual nature of the issue makes that conclusion more nuanced, dynamic, and situational. In this report we explore the many factors and causes of information overload and underload and also develop ideas for solutions. A summary of the issues is provided.

OSI2016 Workgroup Question

Information underload occurs when we don't have access to the information we need (for a variety of reasons, including cost)—researchers based at smaller institutions and in the global periphery, policymakers, and the general public, particularly with regard to medical research. Overload occurs when we can access everything but are simply overwhelmed by the torrent of information available (not all of which is equally valuable). Are these issues two sides of the same coin? In both cases, how can we work together to figure out how to get people the information they need? Can we? How widespread are these issues? What are the economic and research consequences of information underload and overload?

A dynamic dialectic

Information underload and overload are connected. The information-overloaded world ironically suffers from under-loading: its inhabitants are incompletely informed, being given too many irrelevant pieces of data that obscure the ones they need. In contrast, information underload is rooted in settings where information either does not exist or is not being supplied; at its core underload is caused by a lack of access and/or an inability to discover information resources even if they are available. As we remedy problems of underload, we create more problems of overload; and as more information is created, supplied, and accessed, the more people without that information are at an underload disadvantage.

Information overload and underload both lead to underutilization of knowledge and anxiety. The paradox of choice suggests

that more information can lead to its own problems, but *this is a problem of great privilege.*[1] As more information is created and becomes widely accessible, overload challenges inevitably arise. Still, problems of overload are qualitatively preferable than those of underload. It may be difficult to complete a puzzle with many pieces, but it is impossible to do so if some pieces are missing entirely.

Colloquially, information consumption has been described in dietary terms: there is too much ("infobestity," "infoxication") and too little (info starvation).[2] Pragmatically there is rarely an exact quantity of data that one needs, but it's helpful to think of information sufficiency as a happy medium between the two extremes—information satiety, if you will.

1. Overload

Information overload is not qualitatively new, as attested in Ann Blair's book *Too Much to Know: Managing Scholarly Information before the Modern Age* (Yale University Press: 2010). The volume of societal knowledge, both scholarly and non-scholarly, long ago surpassed the cognitive limits of the individual human's mind. Nonetheless, information over-load is quantitatively different now from ever before.

First, information is produced at a much greater rate than ever before. We have entered an era of data deluge. Modern computing systems now produce over 2.5 quintillion bytes of data every day;[3] over 200 billion emails are sent per day;[4] and the total amount of global scientific scholarly output is doubling every 9 years.[5] Once relegated to the scholarly class, the information processing-intensive task of knowledge production and assimilation now dominates modern life.

Second, advances in information storage and especially communication technology are enabling individual human beings to access an increasingly large fraction of the increasingly large amount of information produced.

Third, modern societies are growing more dependent on information and its computational processing, from information-intensive service industries such as finance, communications, and entertainment, to information-oriented public sectors such as education and health care. For example, intangible goods—information-based rather than material-based goods—now comprise an ever-greater share of the gross domestic product (GDP) of every G7 nation.

In other words, today's information overload is now characterized by a growing dependency on relevant, accurate information to survive and thrive and to a degree that is dramatic by volume and breadth alike. These conditions are catalyzed by the appearance of the digital, the information explosion, and the number of people with a need to interact with and produce scholarship. Taken together, all of these changes constitute an historic break from the past. We live in an increasingly infocentric world that generates overload (and underload) problems. While information needs have almost always exceeded our information processing abilities, as described by Blair, the desires for information probably will always exceed our abilities to satisfy them. Both the gap and its impact are massive.

Filters

Blair posits (p. 3) that information management basically requires four crucial operations: storing, sorting, selecting, and summarizing (the 4 S's). Regardless of the technological tools available to us today to perform these operations (e.g., search engines and text mining), the need to perform these operations has not changed much over the past several centuries—arguably, since the early modern era. We suggest that information overload occurs when the latter three of these operations have not kept up with the first. That is, overload occurs when information is abundant and accessible but not necessarily organized, filtered, or presented in useful and appropriate ways to maximize access to and use of this information.

Broadly speaking, we may call these latter three operations "filtering," and they are crucial to dealing with information overload. Filters take a variety of forms and paths. In sociopolitical contexts, filters may be considered a form of censorship over which a user may have no control, such as Internet blocking. In basic searching, filters allow users to include or restrict certain document details, such as format or date. In the context of information overload, technological filters are designed to assist users with winnowing out the least applicable information to fit their research purpose more precisely. Simply put, these filters are critical to and pervasive throughout the research process.

Meta-information

Hypothetically, the most accurate filtering would be done by an omniscient, superintelligent being whose conceptual understanding of the raw information and user needs would enable users to be presented with the exact information which they are either seeking or in some cases not looking for at all (i.e., information they need but don't know they need). In the absence of such a being, however, filters all rely on meta-information: information about the information, such as information about other people who have used and/or cited the information, how this information connects to other information, the reputations of the authors of the information, how the authors are connected to other, and the reputations of the journals or distribution channels of the information. These are just a few examples of meta-information that filters rely on.

Perhaps most important, in a way analogous to the fact that there are more mutual funds than stocks, there is at least as much meta-information as there is raw information itself. There is even, recursively, meta-information about meta-information, much like Google's PageRank algorithm, information associated with the "reputation" of another piece of information might depend on a nested layer of even more information associated with the reputation of the person who may have first given the piece of information its reputation.[6]

It is thus critical to the effective and continued development of filters that meta-information be at least as "openly accessible" as the raw information itself.

We should also bear in mind the widespread reliance on content rather than meta-information or metadata for searching, as software indexes main body texts. Indeed, as David Weinberger argues (*Everything Is Miscellaneous*, Times Books: 2007), we may see content serve the majority of indexing and search functions in the near

future, at least for automated search and discovery.

Filter success

Filter success may be achieved through robust online discovery processes, as provided by various online products developed for libraries (e.g., Primo, Summon, and WorldCat).[7] Discovery software is continually under development and costly to libraries. However, discovery success is dependent on the open availability and dependable curation of content. If the needed items aren't digitized, or the metadata for them is incomplete or nonexistent, it makes no difference how good or how expensive your discovery tools are or where in the world you are located. This is the yin-yang of digital content.

Search engines have been touted as easier, faster, and/or cheaper means to find content, but success often depends *a priori* on having detailed knowledge about what is being sought (i.e., relevant keywords, a citation, item DOI, ISBN, author ORCID identifier, etc.) to guarantee targeted hits. Typically, research conducted via search engines using only keywords and Boolean logic is inefficient and incomplete; hence filter failure occurs.

Filter failure

The term "filter failure" was coined by Clay Shirky to subvert the notion that too much information can be a bad thing;[8] instead, it's merely a sign we haven't put in place the proper sieve. However "filter failure" extends not only to the filters we *don't have*, but the ones we have that simply *don't work*.

There are filters that block too little, presenting us with an ever-expanding glut of information of questionable quality and value, and additionally the presence of duplicative hits. Both of these results can produce personal psychological responses to overload. Lack of productivity, anxiety, fatigue, and fear of missing out (aka FOMO) are examples of reactions and emotions experienced when coping with overload.[9] There are also filters that block too much (leading to underload), filters that warp information based on a biased protocol or design (delivering misleading information), and filters that simply dysfunction.

Even when filters are discerning, neutral, and competent, the increasing personalization of search and answer results threatens to envelop each of us in our own filter bubble, as Eli Pariser warned in a TED talk and book (*The Filter Bubble*, Penguin: 2012).[10] Search providers can sort results to match their models of our interests, reducing the chance of encountering sources and documents that are novel or unexpected. Worse, these filter bubbles can contribute to the growth of echo chambers, when users move within a universe of uncontested and likeminded opinions.

Despite the existence of abundant and useful filters that strike a balance between universality and targeted relevance, people must be aware that they are available, know where to find and how to use them, and also teach others to use them. Potential users must also have the wherewithal to access or pay for these services. This suggests, as many librarians know all too well, that information overload is often caused or exacerbated by a lack of information literacy. Modern digital citizens need regular, deep, and broad education about the skills and attitudes that accompany information literacy.[11]

Discovery and curation as alternatives to filtering

The more traditional concepts of information discovery and curation offer complementary approaches to under-standing information overload.

Curation, as it relates to the creation process, delegates the responsibility for filtering content to "experts" charged with selecting information relevant to any given problem. Selective, peer-reviewed journals entrust this task to a body of reviewers and editors who are assumed to have achieved credibility and expertise (with most peer-review still conducted in a single-blind fashion, however, the veracity of this assumption cannot be tested). Thus, an investigator in a given field can address overload by following only the content of a subset of scholarly journals. With the increasingly interdisciplinary nature of most scholarship, however, this is a perilous strategy.

Curation has also been accomplished by the long-standing practice of developing scholarly review articles. We also recognize that the authors of such reviews are themselves subject to overload and potentially vulnerable to biases; the curation embodied in a scholarly review can never be perfect. Furthermore, our academic reward systems tend to undervalue the scholarly contribution represented by a well-researched and well-written review article, while the journals publishers are chasing an increasing number of reviews to boost their citation statistics. These issues result in difficulties in persuading the most-qualified commentators to author review articles, as well as issues of self-plagiarism whereby very similar analyses are repackaged over a series of articles by a single author or team.

Curation at its heart is meant to facilitate the long term availability and viability of scholarly information for many audiences and for many purposes. It often operates outside of formal distribution channels and agents but often within institutional contexts. Curation promotes discovery and prevents underload over time, stabilizes content, and assures continuity of access. Good creative practices and discovery cannot occur without curation in mind from the earliest stages of scholarly discourse.

Solutions to overload

Social and technological solutions exist to address information overload, some of which build on Blair's 4 S's mentioned above. Collaborative approaches among and between stakeholders that cross domains and institutions offer the most efficient means of addressing and managing overload.

Librarians embedded in educational institutions (K-12 and higher education) have been on the frontline of teaching information literacy for decades, but the pedagogy has changed dramatically to incorporate skills critical to today's students and their overcrowded information environment. Increased collaboration among teachers and librarians is critical to academic success and in the workplace. With greater competency comes an increased ability to filter, manage, and organize information for the purposes of producing new scholarship, finding pertinent research on topics affecting personal decisions, and evaluating information relevancy and sufficiency throughout a lifetime.

Consequently, teaching information literacy is more important than ever. With increased knowledge of how information systems and economies behave and how they can be influenced, students and scholars will be better prepared to make these systems and economies more open and understandable to others. Those with much at stake include researchers, learners of all kinds and ages, scholarly societies, librarians, and publishers. Inventors and tool creators in the community of scholars have a stake as well, and can grow through more open information. All stakeholders can benefit from collaborations across domains and institutions.

Social solutions

The rise of social media and the user as producer has led to a growing movement of democratic curation. Sites such as Pinterest, YouTube, and Diigo let users select, present, arrange, and comment on materials, most of which are publicly available. Users can also query their social networks for information or documents, as when we ping Facebook or Flickr for what one's friends are sharing. Google searches often return Google+ posts, sometimes linking results to the searcher's social network. In short, our online associations are increasingly serving an information-filtering role.

Technology solutions

One of Blair's achievements is describing how editors, authors, and publishers created innovative new textual and physical features to cope with information overload. We can recognize familiar forms of these, such as marginal annotations, structured indexes, and encyclopedias; others, like the florilegium, are used less often now but were relied upon for centuries.

Similarly, we argue that our present era is experimenting with new forms of information and discovery aids. Our team generated a quick list, including:

- overlay journals (a push approach, driving content towards readers)
- personalized search (a pull approach, using search, discovery, mining to draw content towards consumers)
- increased use of open metadata
- enhanced metadata and standards
- more machine-readable data
- vendor-built intermediary tools, such as Primo and Summon
- academic networks, like Academia.edu and ResearchGate
- publisher-built tools (Scopus or Web of Science)
- open platforms (Wikipedia).

Specifically, the digital environment of modern scholarly communication offers some new tools for discovery and curation to aid the chronically-overloaded scholar, which also may rely on the collective networked expertise of other scholars. For example, overlay/virtual journals present content addressing a specific topic from a single publication or a range of publications over time, perhaps with associated commentary. Similarly, efforts that compile and promote the "best" publications over a specified period, whether selected by an expert jury or on the basis of post-publication use and citation metrics, alert investigators to impactful work in venues they might otherwise have overlooked.

Two additional technological methods that provide ready manipulation and deeper understanding of information are data visualization and text mining. Data visualization simplifies and improves hu-

man interpretation of data by re-representing the data in graphical form.[12] Visualizations can be as straightforward as drawings and or as complex as simulations and immersive environments. Text mining, often referred to as text data mining (TDM), is a hybrid of the more mature information overload-reducing technologies of information retrieval, information extraction, and data mining.[13] Text mining fosters an ever faster and more extensive creation of knowledge, facilitating the inclusion of far more text information in knowledge formation than purely manual efforts. Scholarly archives are wellsprings for text mining, enabling countless medical, economic, social, and environmental discoveries, from speculating on the impact of Moliere's plays on modern television comedies, to predicting the health risks of new cancer drugs.

2. Underload

Information underload is the condition of the *under* delivery of meaningful information caused by barriers of both access to and entrance into scholarly dialogue. These barriers arise from a complex set of factors, including but not limited to sociopolitical, financial, technological, and geographical challenges. Underload can occur between scholars, as well as between academia and the general public.

Among the barriers to access are sociopolitical ideologies and regimes that might block or censor materials. Many libraries outside of the U.S. limit access to their physical collections without a letter of reference or other documentation of need. Religious factors might influence access, readership or interpretation of scholarship.

Technological challenges include connectivity, bandwidth, access to computers, the available format(s) of content, or other hardware limitations. Significantly, access to mobile devices is more prevalent worldwide than larger computers and screens. Published scholarly material is rarely optimized for mobile devices, and especially the cell phones that are the primary portal to the Internet for those in developing countries. Technology and formats of information are elements of accessibility (physical impairments), both as impediments and potential solutions.

We should also remember that mere access to technology is not sufficient. The lack of access to and knowledge of sophisticated software hinders the production and use of scholarship wherever it occurs. At the same time, many inventive scholars have little opportunity or means to develop new software and technology innovations. Some information may be held in proprietary systems or held by commercial interests and therefore not shared. Legal issues and regimes govern ownership, and the ready use, of information as well.

Simple geographic location makes a difference. Distance to a local library (if one exists), distance to other libraries or archives where physical materials reside, and lack of local technology or Internet access all present formidable barriers. The geography of underload is not equal. The majority of developing, emerging, Global South countries are significantly more impaired and impacted by underload.[14] As well, many communities within developed, "Global North" countries are similarly suffering in contrast to their wealthier in-country peers due to income and infrastructure inequality.

Education is another significant factor. Both basic literacy and information literacy are essential to engaging fully with scholarly work. A lack of computer skills may impede ready access to information and the creation of new knowledge. The extent of language skills and the quality of writing influence whether scholarly work is easily read and understood, and poor skills may prevent new scholarship from being accepted and published. Language and writing styles can also present barriers to reading scholarly work or to having one's work accepted.

In short, fewer information options are available to those who face this multitude of barriers.

The lack of access to technology and content also causes the inability not only to consume, but to *contribute* fully to the scholarly conversation and record. Funds may be unavailable for research and publication. The grant system that supports science research in North America and Europe is not typical in most other countries and regions. Underrepresented communities cannot produce new scholarship at the rates that they need or want, due in large part by the lack of funding for conducting research and gaining access to published research. Scholars in any country might not be able to conduct, use and share research because of the cost of performing the research, acquiring publications, and fees (however small) that may be associated with publishing. There is often a lack of funding to digitize local content in order to make it available to researchers elsewhere to access and use.

3. Challenges

Changes to systems of scholarly communication and measures of impact face the challenge of overcoming existing promotion and tenure structures and criteria. Although these systems are evolving, innovation is not readily recognized as valid unless it is expressed in familiar terms and acceptable within traditional rubrics. Disciplinary self-interest might perpetuate the status quo and block, or at least ignore, the open access outlets and Global South voices that are clamoring to be heard. Institutional self-interest and competition are influenced by perceptions of prestige, rankings, and other measures of quality. Industry self-interest, driven by revenue projections and shareholder interests, dominate many markets and methods for sharing. Adequate resource bases are, of course, essential to the health of scholarly communication that thrives on vigorous and equitable participation in the conversations.

Solutions

Collaborations can empower solutions. Such library organizations as the American Library Association and the International Federation of Library Associations and Institutions have robust agendas that address information literacy, social justice, and freedom of inquiry internationally. In 2014 the Association of European Research Libraries (LIBER) authored the Hague Declaration in an effort to foster ethical and legislative reforms to encourage the adoption of open access policies and infrastructure. The Hague Declaration addresses revisions to copyright in order to allow unfettered access to all knowledge stores for the purposes of content mining. [15] The Declaration cites the importance of machine reading and subsequent processing to solving the grand challenges of the modern era, including global health and climate change. As of May 2016, the

Hague Declaration has attracted more than 500 individual signatories as well as over 230 institutional signatories. UNESCO has a vested interest in collaboration to achieve its ambitious goals for sustainable development that include education.

The International Network for the Availability of Scientific Publications (INASP) is "an international development charity working with a global network of partners to improve access, production and use of research information and knowledge, so that countries are equipped to solve their development challenges." As a nongovernmental organization (NGO), INASP is devoted to addressing and solving issues of "the availability, access and use of international research information by researchers in developing countries and the production, quality, dissemination and access of research outputs from researchers in those same countries." INASP claims that "by building capacity at individual, institutional, national and international levels, we have seen significant improvement in the research and knowledge sector in many of the countries with which we work." This organization is an important player in mitigating underload.

The Wikimedia Foundation, the nonprofit that hosts Wikipedia and its sister projects, boasts ambitious aims to "share the sum of all human knowledge with every person on the planet" (freely licensed for reuse, and in the language of one's choice). In pursuit of its mission it sees 8000 views per second and nearly half a billion readers per month across nearly 300 languages. While full-text scholarly articles are only available for some readers, their content is freely summarized for all. Wikipedia is, according to Geoffrey Bilder at Crossref, the 5th highest referrer to all DOIs online.[16]

Not all solutions rely on institutions and organizations. As individuals, people might seek education (formal and informal) as well as professional development that allows them to increase their information literacy skills in all their many aspects, as well as their success in scholarly information use and exchange, and their ability to advocate for solutions and changes in their organizations, communities, professions, or other venues. To some extent information over/underload solutions rest on individual action and responsibility.

To address publishing challenges directly, partnerships can develop to make Global South journal content more present. Tools such as altmetrics can expose and assign value to more kinds of scholarly contributions, both informal and formal. Of course, an increase in open access provides more avenues for sharing and communicating. Discovery can be advanced through metadata solutions—more metadata, more open metadata, and more effective application to aggregated collections. Judicious use of social technologies can contribute toward bridging gaps in access and sharing. The use and acceptance of blogs, RSS feeds, academic networks, and other social media tools for crowd sourced curation will help, although these tools can also fracture the environment and leave one unsure as to which forums, if any, capture the totality of relevant and reliable information (a condition that overlay journals, described above, might address.).

Economic implications

The digital divide runs deep and wide.[17] Fully 4.6 billion people—approximately three-fifths of the world's population—remain without access to the Internet.[18] For many of the over three billion who do enjoy Internet connectivity, numerous social and technical barriers to retrieving information remain. While the digital divide excludes more than half of the world's population, a relative deficit in knowledge provision exacts economic costs from the overwhelming majority of the world's inhabitants. The economic effects of underload on an individual level may involve a lack of access to crucial and timely information about competitive product prices, an inability to reach marketplaces for trading goods, missed opportunities for employment and skill development, prolonged illnesses and lost wages, and even an underutilization of know-how and resources in times of environmental, health, and political crisis. Moreover, access to information and the optimal utilization of knowledge therein promises not just skill development or new jobs but employment advancement, even empowering individuals with the opportunity to join the ever-expanding and highly complex digital economy.

A lack of Internet access is one way to suffer lost economic opportunities, but even populations with connectivity can watch their access to knowledge slip away through a number of threats, such as closed governance, substandard digital infrastructure, and a lack of access to high quality information-processing tools and the means to acquire requisite skills. Such underloading for individuals with access can begin to replicate conditions otherwise seen only on the other side of the digital divide.

On a societal level, the implications of improved access to knowledge seem to multiply. The Internet already directly accounts for a large share of the global GDP, is necessary for much of the remainder of global GDP, and is increasing its share. From 2007-2011 alone, 21 percent of economic growth came from the Internet.[19]

Economist Ricardo Hausmann and network physicist César Hidalgo have recently demonstrated knowledge productivity is the leading indicator of a nation's GDP, capturing such knowledge productivity in a measure known as economic complexity. As Richard Hausmann writes (*The Atlas of Economic Complexity*, 2014),

> .... [T]he wealth of nations is driven by productive knowledge…. The secret to modernity is that we collectively use large volumes of knowledge, while each one of us holds only a few bits of it. Society functions because its members form webs that allow them to specialize and share their knowledge with others.

For a complex society to exist and sustain itself, people with knowledge of design, marketing, finance, technology, human resource management, operations and trade law must be able to interact and combine their knowledge to make products. These same products cannot be made in societies that are missing parts of this capability set.

To state it bluntly, the capacity of a person or nation to possess and combine knowledge increasingly drives that person's or nation's prosperity. Underload poses a fundamental threat to economic

opportunities for individuals and nations alike, and threatens geopolitical stability.

4. Other factors

Digitization and data-sharing

A fundamental limitation to sufficient information access lies with content not yet made available on the open web. This includes archival assets that have not yet been digitized or published online. Targeted efforts to convert more such content into a machine-shareable form are still very much needed, especially for scholars in emerging countries who want their resources to be available for research and teaching. This limitation also applies to the data underlying research and scholarship, much of which is digitized but never shared beyond one author's computer or institution. The continuing push for a model of open science (as for other fields), which includes the publication or at least the accessible archiving of raw data, will not only improve access to information, but, because we can interrogate it, will also make the information we consume more robust.

Evolving legal, business and copyright models

Obviously, open access publishing, public access, preprint servers, institutional repositories, open commenting and review, and other approaches are alleviating underload, and even overload to some extent, as witness the rise of sources like arXiv and bioRxiv. Sci-Hub, a service that distributes articles freely but without permission of copyright holders, represents a more radical approach to addressing the problem of underload. There are continual challenges to and push back from prevailing legal and copyright models. Advocacy is making a difference here. Acknowledging that open access is here to stay, business models are changing to make money from open access, some by reducing costs and facilitating new modes of access, and others by misleading potential contributors or double-dipping with subscription charges and optional open access fees per article.

Advocacy and development

Increasingly, access to scientific information is being seen as a right for taxpayers (e.g., in the United States, United Kingdom, Ireland, Sweden, and France) who support science research through government agencies. Not-for-profit institutions and NGOs are requiring access to the results of research in which they invest. Globally, access to information is being considered a human right and a critical component of social justice.[20] Advocacy and policy are working hand-in-hand toward more open and transparent systems of scholarly exchange.

Automation, machines, and computing

We can't ignore the fact that one of the acute ironies of the imbalance of entrance into the flow of scholarly information is that machines have far greater access to data and information than humans. They are the primary consumers and accumulators of information today and will increasingly be making decisions without human intervention, oversight, or even our knowledge of the computations and algorithms governing these decisions. Machines are keeping our memories and writing our narratives, and answering our questions perhaps without validation or authentication. We can look to the computer-created news stories that already

exist. Current ideas and practices of information literacy may not go far enough in exposing, understanding, and explaining these facts.

Increasingly, computers are taking on human functions. To do these well, machine-reading requires good metadata and ontologies, along with restriction-free access to large bodies of information to enable processing. Take, for example, text mining. As mentioned above, this is the process of identifying, synthesizing and creating knowledge via machine reading of digital texts. Text mining addresses information overload in two ways. First, it radically reduces the time to read materials, making even the largest disciplines readable in a reasonable amount of time. Second, text mining accelerates the ability to turn what has been read into actionable knowledge. It might involve the identification of useful pieces of information, the juxtaposition of otherwise disparate facts leading to new insights, or wholly imaginative speculations stimulated by more creative generative mining techniques. We can imagine a future text mining application serving as a research assistant, either to create new work, to save a researcher's time, or both.

Following this line of thought we can offer an axiom: *people, through their machines, must be allowed to analyze anything they have a legal right to access*. Hindering computational analysis is nothing short of preventing analysis. Text mining engages in the extraction of small digital objects, such as facts, data, and ideas. By definition, such entities are not copyrightable (17 U.S.C. §102 (b)). U.S. Copyright law provides an exception via the Fair Use Doctrine (17 U.S.C. §107) that would permit text mining; however, licenses often prohibit such uses, thereby undermining Fair Use and inhibiting innovative research, both non-commercial and commercial.

A mandatory copyright exception for text mining should be made, and publishers and vendors should be encouraged to remove obstructions to third-party text and data mining. Rather than request free and unfettered access of copyrighted materials, we are asking that machine means of access, such as crawling and scraping, be permitted at a minimum, and that derivative works from those analyses be unencumbered by legal threats. A lack of clarity around copyright and ownership of derived works and orphan works, as well as the complexities of licensing, all stand in the way of facilitating more creative and optimal computational use of scholarly content. Librarians and their principles stand at the center of the opportunity to provide access to text mining for reduction of overload and underload alike.

Automation challenges

We do not wish to appear as Pollyannas or utopians, however, as many challenges await an automation-based solution to information overload and underload. For one, not all information is machine-readable in its present forms. Information containers (books, articles, videos, manuscripts and the like) vary widely and are not inherently machine-readable. For full machine processing, these forms must be converted to a machine-friendly format—a formidable task.

Automation of filtering for successful discovery is very complex. It requires intensive development of well-organized and reliable repositories as well as of codebases. The skills to develop intelligent automation are in high demand and require advanced education. There are also

sociopolitical challenges in forming agreements and alliances that lead to comprehensive automation solutions. All of this is expensive, in terms of time and intellect, and often money. Open source solutions mediate some of these challenges but nevertheless require significant investment in efforts of coordination and community development.

Information literacy about automation is largely untrodden territory. Even some librarians or researchers who use automated tools on a regular basis may not have a deep understanding of the underlying mechanisms and the consequences of algorithmic and information architecture decisions. Some of this deficit of knowledge arises from the fact that this is a still a new frontier in information science. This also occurs because these systems are opaque, proprietary, "black boxes": There is simply no way to know how they work because their mechanisms are secret, and thus evaluating their outputs amounts to guesswork.

The opacity of what machines are doing in the background of our lives creates illusions of fairness, equity, privacy, or full access to everything. None of these is assured or necessarily true for all seekers and users of information.

The growing issue of automated control over scholarship requires that we become better informed about the nature of computation and its control over what is accessible from and delivered to our devices. We should push for, if not demand, transparency, education, and dialogue about the implications of the choices the machines make (through human programming choices) and their potential risks, principles and standards, and mitigation strategies and options—plus far more than we can elucidate at this time.

5. Summary statement

Both information overload and underload will be with us for the foreseeable future. In a world where ready access to high quality information is increasingly essential to quality of life, people with less access and fewer choices will continue to be disadvantaged in participating in the marketplace of ideas and a competitive information economy. The simplest solution to information underload is removing the current barriers, including notably the sheer cost. Open access provides a framework for making that transition, and although it is not a simple transition to make, it will undoubtedly improve the lives of billions. Information access is considered by many to be a human right that parallels other basic economic or socio-cultural necessities that benefit the human condition and maximize its potential. We must advocate forcefully for this access.

At the same time, open access will create new challenges of information overload. Well-vetted scholarly information is competing for our attention along with other information that purports to have credibility without following peer review or other assessment processes. Better exposure and discovery options for scholarly products are needed, as well as the means to understand and apply them. There is no single solution to the glut of information, which can be overwhelming, counterproductive, or potentially dangerous. Even with a surfeit of information there is inevitably loss, barriers, missed opportunities, and discovery challenges. Like overload, there is no single solution to the problem of information underload, which can be dis-

empowering, disheartening, constraining, and dangerous. Solving both challenges will require all stakeholders to be both deliberate and inventive, ideally within a framework of open collaboration that is built upon common values, shared metadata, sound standards, and a commitment to a far more open system of scholarly endeavors.

OSI Information Overload and Underload Workgroup

Bryan Alexander, higher education publishing consultant and futurist

Kim Barrett, Dean of the Graduate Division, University of California San Diego (UCSD)

Sioux Cumming, Program Manager, Online Journals, International Network for the Availability of Scientific Publications (INASP)

Patrick Herron, Senior Research Scientist, Information Science + Studies, Duke University

Claudia Holland, Head of Scholarly Communication and Copyright, George Mason University

Kathleen Keane, Director, Johns Hopkins University Press

Joyce Ogburn, Dean of Libraries, Appalachian State University

Jake Orlowitz, Head of The Wikipedia Library

Mary Augusta Thomas, Deputy Director, Smithsonian Libraries

Jeff Tsao, Distinguished Member of the Technical Staff, Sandia National Laboratories

Notes:

[1] See https://en.wikipedia.org/wiki/The_Paradox_of_Choice

[2] See https://en.wiktionary.org/wiki/infobesity and https://en.wiktionary.org/wiki/infoxication

[3] Nawsher Khan, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, et al., "Big Data: Survey, Technologies, Opportunities, and Challenges," The Scientific World Journal, vol. 2014, Article ID 712826, 18 pages, 2014. doi:10.1155/2014/712826

[4] "Email Statistics Report, 2015-2019," The Radicati Group, March 2015, as of May 27, 2016: http://www.radicati.com/wp/wp-content/uploads/2015/02/Email-Statistics-Report-2015-2019-Executive-Summary.pdf

[5] Richard Van Noorden, "Global scientific output doubles every nine years" (blog post), May 7, 2014, Nature newsblog, as of May 27, 2016: http://blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html
[6] See https://en.wikipedia.org/wiki/PageRank
[7] A good definition of discovery is provided by librarytechnology.org.
[8] Matt Asay, "Shirky: Problem is filter failure, not info overload" (blog post), C|Net website, January 14, 2009, as of May 27, 2016: http://www.cnet.com/news/shirky-problem-is-filter-failure-not-info-overload
[9] For a definition of FOMO, see https://en.wikipedia.org/wiki/Fear_of_missing_out
[10] Eli Pariser, "Beware online 'filter bubbles" (TED talk video), March 2011, as of May 27, 2016: https://www.ted.com/talks/eli_pariser_beware_online_filter_bubbles?language=en
[11] See the Association of College and Research Libraries document, "Framework for Information Literacy for Higher Education", for a current take on just how much work this entails.
[12] For a definition of data visualization, see https://en.wikipedia.org/wiki/Data_visualization
[13] For a definition of text mining, see https://en.wikipedia.org/wiki/Text_mining
[14] For a definition of the global North-South divide, see https://en.wikipedia.org/wiki/North%E2%80%93South_divide
[15] http://thehaguedeclaration.com/
[16] https://meta.wikimedia.org/wiki/Wikipedia_as_the_front_matter_to_all_research
[17] For a definition of the digital divide, see https://en.wikipedia.org/wiki/Digital_divide
[18] http://ahumanright.org/
[19] Mathieu Pélissié du Rausas, James Manyika, Eric Hazan, Jacques Bughin, Michael Chui and Rémi Said, "Internet Matters: The Net's sweeping impact on growth, jobs, and prosperity," McKinsey Global Institute, May 2011, as of May 27, 2016: http://www.mckinsey.com/industries/high-tech/our-insights/internet-matters
[20] See the Lyon Declaration at http://www.lyondeclaration.org