Information Services & Use 42 (2022) 285–299
DOI 10.3233/ISU-220160
IOS Press

The scholarly record of the future: A technologist's perspective

Todd A. Carpenter*
Executive Director, NISO (National Information Standards Organisation), Washington, DC, USA
ORCID: https://orcid.org/0000-0002-8320-0491

Abstract. This article is adapted from Todd Carpenter's keynote presentation at the APE Annual Conference on January 11, 2022. It focuses on the scholarly record of the future from a technological perspective. It describes how the transition to data-driven science and an increased focus on discovery and interoperability of content, will shape the article of the future. For a "digitally native" scholarly record, developing and maintaining a digital infrastructure with high-quality metadata and identifiers is crucial. It is also emphasized that technology does not determine the direction or eventual destination of the scholarly record of the future.

Keywords: Scholarly record, digital transformation, identifiers, metadata

1. Introduction

There are many approaches to looking at the future. In this talk, I will take a particularly technical approach to thinking about the Future of the Scholarly Record. Without going into a full history of the technologies in use in scholarly communications, I'll start with a brief review of some of the key transitions that have brought us to the present day, before I delve into the future trends in technology, and their applications. As Steve Jobs said, during a 2005 graduation address at Stanford University, "You can't connect the dots looking forward; you can only connect them looking backwards. So, you have to trust that the dots will somehow connect in your future [1]". In this way, we can predict the future by examining the past, but we also recognize that predicting the future is as much a leap of faith.

Scholars have been sharing the results and data associated with their findings for centuries. An illuminated manuscript from the 12th century, a diagram of the terrestrial climate zones in the Riphaean mountains from the collection of the Walters Art Museum [2] is just one of many examples. Over time, the form of incunabula grew into monographs. A separate practice of sharing letters among scientists describing their research and results became formalized as articles with the launch of journals in the 17th century. The recording, compilation and sharing of specific observations was also a practice that became systematized in the 19th century, most notably with weather records. Despite their limitations, these systems served the scholarly communications community reasonably well for centuries.

2. Digital developments and the scholarly record

Skipping forward to the late 20th century with the development of computer technology and the development of early computer networks. Efforts to connect computer systems and creating resource

*E-mail: tcarpenter@niso.org.

sharing networks was an early goal of those involved in computer networking in the 1960s and 1970s. The Internet was designed in part for defense, but also for science, with sharing data and papers in mind. A 1985 map of the NASA SPAN network, a space-data focused network built around the same time as the NSFnet, highlights the many research institutions involved in astrophysics research at the time.

These visions of an integrated, multi-media ecosystem of linked digital content were features of Vannevar Bush's ideas on information organization [3] and Ted Nelson concept of HyperText and later HyperMedia [4] in the 1970s. They were finally implemented in the 1980s in a variety of technology systems such as ZOG, HyperTies, and HyperCard, and were the precursors to today's World Wide Web. In the early 1990s, Tim Berners-Lee at CERN proposed [5] to create a platform for sharing scholarly content when he developed the basic elements of our current web technology stack, Uniform Resource Locators (URLs), Hypertext Transfer Protocol (HTTP) and the HyperText Markup Language (HTML). While not particularly groundbreaking technologies at the time, the two strongest benefits of Berners-Lee's approach were its basis on open standards, and that the approach was free from license fees. Interestingly, in 1991 at the third ACM HyperMedia conference in San Antonio, Tim Berners-Lee's proposed paper on his World Wide Web idea was rejected [6]. He was forced to demo his WWW technology in the exhibit hall as a poster session because the organizers thought his application was a weak implementation of far more robust systems at that time.

We are only now taking full advantage of the visions of HyperText from the 1980s because of a confluence of distributed technology, connectivity, processing power, storage, and standards. Part of the reason the process of implementing robust digitally native solutions has been slowed over the years has been people's natural tendency to transfer the old methods to solving a problem to the new technology. For centuries people were comfortable with print, so the publishing industry and technologists provided users a page replica, PDF, what they expected. This is a classic skeuomorph, an item with design elements that still preserve the look and feel of the previous iteration but no longer are required in the new design. It is similar to the apocryphal quote often attributed to Henry Ford [7], "if you asked people what they wanted from cars in the 1900s, they would say faster horses". Rather than implementing the amazingly transformative things that could be achieved via digital distribution, the publishing industry began by recreating the printed page.

Instead of thinking about "faster horses", let's think about very different digital things. Over the past 15 years, there has been a compulsive use of the term "digital native", particularly by marketing executives, to describe a set of people born into using technology. But really, what is it that makes someone a digital native? Or more importantly for this conversation, what constitutes a digitally native scholarly record? Was it born out of digital data collection or data sharing? Was it analyzed in the cloud? Is it machine generated content, code, or images? What will a truly digital native scholarly record look like? Realistically, it should be something more significant than simply porting a representation of content from a printed page and transferring it onto a screen, be a desktop, a tablet, or a mobile phone.

It seems clear that digitally native documents need to be:

Multi-format and multimedia when appropriate
Interoperable with other resources
Machine-readable
Adaptive in design for different displays
Accessible for people with different capabilities
Transformable to other forms or environments as needed
Atomize-able, so users can reuse components
Described with high-quality metadata

Preservable
Persistently Linkable
Trackable

And to be these things, many of them need to adhere to standards that facilitate these functionalities. My own organization, the National Information Standards Organization (NISO) has been helping publishers, libraries, and the various intermediaries in scholarly communications support this transition. Many of the elements in this list are areas around which NISO has helped to develop and manage standards that support these functionalities. Without standards applied across the community, the interoperability that is required to bring this hyper-connected, hypermedia world to life.

3. Catalyzing growth in research data sharing

There are several areas where the scholarly record is changing radically to take advantage of being digitally native. The first change I'll cover is focused on sharing of research data sets. In a print environment, data was difficult to publish, hard to consume, and difficult to share. Researchers weren't rewarded for sharing it and there was little incentive to prepare it for distribution. Therefore, the scholarly record didn't contain much of it. Certainly, there were ledgers of observations that were recorded in logbooks, as I noted previously regarding weather or environmental observations, and very occasionally these observational tables might be translated into a published form, but this was seldom undertaken.

This started to change as technology advanced making data sharing easier, and more interesting as a tool to advance science. In the early 2000s, there was a growing awareness of the potential for data, data analysis and data sharing in research. One leading book of that time entitled The Fourth Paradigm: Data Inclusive Science Discovery [8], edited by Tony Hey, Stewart Tansley, and Kristin Toole, outlined how data-intensive science will transform how researcher will be done using massive data sets. It predicted that, "increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets."

Also, around this same time, there began a massive investment in the data management infrastructure around the globe. In the United States, a group of DataNet projects [9] funded by US National Science Foundation (NSF), were launched in 2007. This project was a significant investment in support of NSF's Cyberinfrastructure Vision for 21st Century Discovery [10]. The Project eventually encompassed a $100 million investment in five data networks over 10 years, including, DataONE [11], Data Conservancy [12], SEAD Sustainable Environment - Actionable Data [13], DataNet Federation Consortium [14], and Terra Populus [15]. Similarly, other data repository investments were made outside of the US around this time including the Australian National Data Service (now ARDC) [16], the EU Open Data Portal (now Data Europa) [17], UK JISC Repositories Support Project (now closed) [18] and DRIVER (Digital Repository Infrastructure Vision for European Research) [19].

Not long after these investments began, there was a simultaneous effort to get researchers to use data, as well as to motivate them to share their data. One way to motivate them, it was thought, was through citations, which is one motivator for getting articles published. So, a team organized by International Council for Science (ICSU): Committee on Data for Science and Technology (CODATA) and International Council for Scientific and Technical Information (ICSTI), set out to explore ways to establish an ecosystem of citing data, making it a 'first-class' research object in scholarly communications. The joint CODATA-ICSTI Data Citation Standards and Practices working group was launched during the 27th General Assembly in Cape Town in October 2010.

The group was tasked with several goals: survey existing literature and existing data citation initiatives; obtain input from stakeholders in library, academic, publishing, and research communities; host workshops to establish the state of the art; and to work with the ISO and major regional and national standards organizations to develop formal data citation standards and good practices. The group eventually hosted a dozen meetings worldwide, produced two reports *Out of Cite Out Of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data* [20] and a National Academies Report, *For Attribution: Developing Data Attribution and Citation Practices and Standards* [21].

As this work progressed, those who attended the newly organized Beyond the PDF meeting—which eventually would become Future of Resource Communications and E-Scholarship (FORCE11)—also began discussing ways to incentivize scholars to use communication tools, such as data, as the basis for sharing their research. They sought to galvanize interest and awareness of data sharing, particularly through the use of data citation to drive credit toward those who share their datasets. The Amsterdam Manifesto on Data Citation [22] was drafted during a reception at the Future of Research Communications conference in Amsterdam in March 2013. It received an award at the end of the conference and eventually led to the formation of a working group within FORCE11 to develop the Joint Declaration of Data Citation Principles, which were later published in 2014. Those Principles eventually were endorsed by 125 organizations and nearly 300 individuals.

In January 2014, representatives of a range of the stakeholders interested in data use and reuse came together at the Lorentz Center in Leiden, to think about how to further enhance this nascent data ecosystem. From these discussions, the notion emerged that, through the definition and widespread support of a minimal set of community-agreed guiding principles, data providers and data consumers - both machine and human - could more easily engage with the vast quantities of information being generated by contemporary data-intensive science. The group found a home within FORCE11, and the FAIR Data Publishing Group was organized in September 2014 [23]. The group's "main aim is to create and put up for community endorsement a document that is a general 'guide to FAIRness of data', not a 'specification'." Published in Nature Scientific Data in 2016, the FAIR Guiding Principles made a significant splash on the scientific stage. It has subsequently led to the launch off a variety of related initiatives focused on the implementation of these ideas. Many of these are focused on the necessary standards, the 'specifications' that the original FAIR principles avoided.

It has taken about 15 years for the initial investments to catalyze growth in data sharing. This in turn led people to focus on the use and application of these systems and built it into their workflows. Then it's taken time to develop a culture around data sharing. Realistically, 15 years is significant progress in transformation of one feature of scholarly output. Culture changes much more slowly than the technology that supports it can change.

4. Discovery of content through identifiers and metadata

Another of the elements of this transformed ecosystem is not just about the content, but it's also about the discovery of that content. This is one of the elements of FAIR, but it applies to all forms of the scholarly record, and it is another important element of the transformed landscape of scholarly communications, which I'll address.

Although they are not the foundation of discovery, the venerable card catalog card is an important launching point for a discussion of computer-aided discovery. Card catalogs were among the first application of computer technology to facilitate the discovery of content in libraries. Most institutions didn't have access to multi-million-dollar computers, such as those at the Library of Congress, to manage

their collections so cataloging cards were the distributed output of those first computer systems and then shipped to libraries around the globe. This was how discovery worked in libraries from the 1960s to the 1990s, at which point individual libraries could afford their own computers with digital MARC records.

Of course, those catalog cards were useful things when wandering around a traditional library filled with stacks of books arranged neatly on shelves. However, this is not how many libraries are organized any longer, nor is it how most users search for content. For example, many libraries have instituted automated shelving systems, such as the one at the James B. Hunt library at North Carolina State University. The library's bookBot robotic book delivery system is a central architectural feature of the library, and it is the "stacks" of the library. It is a fully automated retrieval system, which is visible through a glass pane in the lobby of the library and contains roughly 2 million volumes. It is accessible through an online browsing system, which will have a robot deliver the items to the patron automatically. Through a touch screen interface on the adjacent wall, users can "browse" the books in the BookBot Library. But this is not browsing in the traditional sense. It is a fully computer-mediated experience in which a user browses the metadata of the objects in the collection. In libraries like the Hunt library, one cannot browse the stacks of a digital library, one can only browse the metadata.

For humans, **a key problem in experiencing the digital world is that one can't walk the stacks in a digital library or browse the shelves in an online bookstore**. Digitally, all you can browse or search is metadata. This has profound implications for how we discover content and what publishers must do to facilitate that discovery. Publishers should ask themselves this: In your organization, is it creating content the machines can use, can navigate, and then provide to the human readers? How much are you investing in the quality of your metadata? Metadata is truly your product's front door. And as someone who spends a lot of time talking about metadata and with people who use it. It needs to be a lot better.

Another implication of this is that a user's discovery experience is no longer driven by simply matching her words with the item she wants to find—even if she knows exactly what the term was, which is often not the case. Her discovery experience is driven by the complex algorithms that are constantly fine-tuned both to the content available, but also her experience and her behaviors. There is a lot of algorithmic machine-learning technology getting users to the results they are searching for embedded in the machines that consume this content and their algorithms determine what we find. Publishers should carefully consider whether their content is sufficiently organized to support this environment.

This environment is significantly impacted by the roles of the computational tools that navigate this ecosystem on our behalf. To illustrate one specific, but impactful example, let us consider the role of the Google bots that crawl the internet.

Figure 1. shows an approximation of the number of websites that Google crawled per day in 2021 according to the website WorldWideWebSize.com [24]. It fluctuates between about 15 billion and 50 billion sites per day. Let's presume it's a lite day for Google and its crawler only indexes 24 billion sites, which is roughly three sites per person in the world. Again presuming each of those sites contains an average of ten individual pages, each with 600 words (a good SEO average [25]). This would mean Google's indexing robots alone are 'reading' roughly 18,000 words per person per day. This is roughly the equivalent of reading seventy-two pages of a book per person per day, or a novel every three and a half days [26]. Imagine all the publishing industry executives who would pop champaign corks if we lived in a world where every person on the planet read through one hundred books per year.

The publishing industry has for too long been focused on one type of user of the content we create: Humans. There are other audiences consuming the content we create, and they are increasingly important in our space. They are machines and they 'read' far more content than people do. Increasingly, they are becoming quite good at it as well, with advances in machine learning and natural language processing.

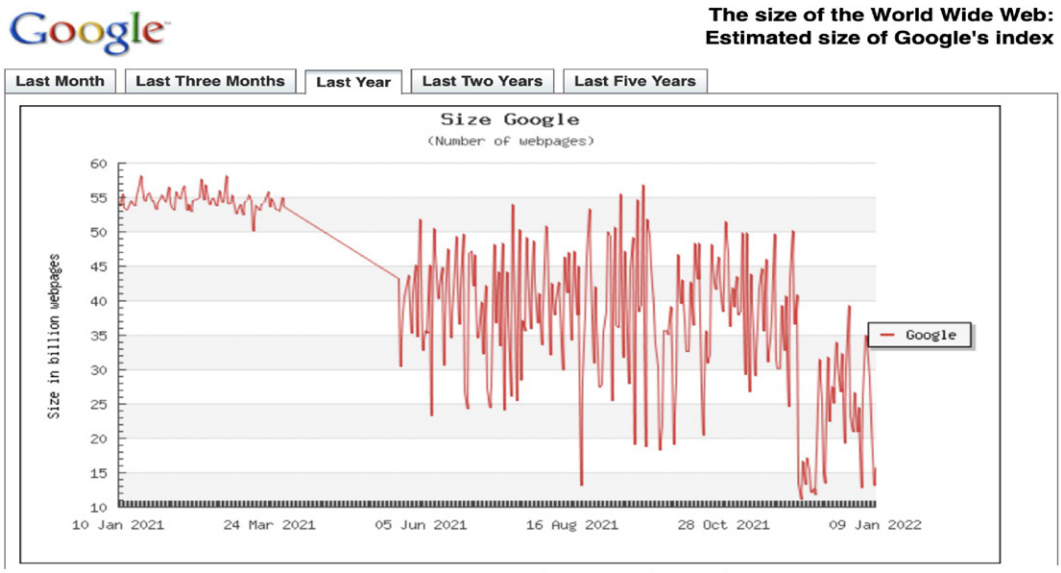

Fig. 1. Estimated number of websites included in Google's index in 2021.

Publishers should start creating and distributing content as much for the human readers as well as for their machine counterparts. Publishers need to provide content structured in a way machines can read the content we are distributing. The way we do that is two-fold–well-structured content and the second is identifiers and metadata. There have been advances in machine consumption of content and natural language processing. When we get to the future, we will dig into those in more detail, but first let us focus on the changes in the ecosystem that are supporting this transformation.

5. Developing an interoperable infrastructure

The International Science Council issued a report in February of 2021 entitled Opening the Record of Science [27]. It proclaims: "All disciplines, whether or not data-intensive, operate in a digital world where all the elements of the research process are connected or connectable in ways that permit them to be linked together as parts of a research workstream, with the possibility of digital interoperability across the 'research cycle'". It also talks a great deal about openness and how that is a lynchpin of the future of the record of science. Understanding how that digital world functions and the infrastructure that supports it is critical to leveraging its power and the analytics that can be derived from it. Several elements of this infrastructure have developed over the past dozen years that are positioning scholars to leverage the power of our connected scholarly record.

The introduction of the Digital Object Identifier (DOI) and its various manifestations, including CrossRef and DataCite, are the first of these elements. CrossRef was launched in 2001 and has seen tremendous growth over its twenty years, becoming a significant feature in scholarly communications. DataCite is a more recent application of the DOI for permanently linking to data sets launched in 2009. It has since grown to include 879 repositories with over 29 million DOIs registered for data sets. In September 2021 alone, DataCite served 13.9 million successful resolutions of DOIs to 3,769,169 objects [28].

ORCID has become another critical identifier in scholarly communications, identifying and describing the authors and contributors in scholarly communications. Launched in 2012 ORCID has grown to include more than 11 million id assignments and continues to grow at a rapid pace. In 2020, 2,639,296 researchers registered for new ORCID identifiers [29].

The Funder Registry managed by CrossRef is a registry of identifiers for funding organizations that provide resources to research projects in the scholarly community. The initial registry data was donated to Crossref by Elsevier after funding information was added to the CrossRef Content Registration schema in 2012. Initially the database included about 4,000 funders and has grown to more than 21,536 funders [30]. The registry connects more than 2.5 million published works with funding data from these organizations.

Beyond funders of research, there has been a recognized need for a cross-community, publicly available identifier for research organizations for some years. To address this, the Research Organizations Registry (ROR) was launched in 2019. Launched with data from Digital Science's GRID database, the ROR system has assigned identifiers to more than 97,000 institutions.

Connecting all the various research outputs that might come from a research endeavor is the next step in navigating the people, institutions and outputs that are generating new knowledge. A newly developing ISO international standard launched by Australian Research Data Commons (ARDC) is the Research Activity Identifier (RAiD). This new identifier will connect the various elements of a research project, from funding, to researchers, to protocols, to papers, data sets and other outputs. It is expected to be published and available later in 2022.

Based on these existing and developing identifier and metadata systems now have a connected network of resources that connects organizations, funders, researchers, their publications, and their data sets. Soon we'll be able to connect these across research projects. We can then use this knowledge graph to discover and navigate the digital-native record of science. In doing so, we can benefit from tremendous economies of scale, because things are done in a standardized way, as well as leverage the power of computational navigation and digitally connected resources.

A perfect example of this is the newly launched OpenAlex Project [31]. In 2021, Microsoft announced it would be closing the Microsoft Academic Graph resource [32], as part of its dialing back the work it has done in this space of scholarly communications, such as the Microsoft Academic Search service, which has also been retired. What's interesting in the team at Our Research, a very small non-profit, has taken up the work of maintaining and continuing to develop this resource. What originally was undertaken by a many-billion-dollar company with tens of thousands of staff, and untold millions to direct toward research projects like this, can now be managed by a team of only a couple of people with less than $1 million per year.

The many investments in the scholarly infrastructure allow for tremendous downstream applications such as this. There are a variety of small start-up projects that are taking the investments in this open infrastructure and are creating real value of scholars and institutions. What are the amazing things that existing organizations or the next startup can do based on this infrastructure, that doesn't need the resources of a trillion-dollar company behind getting those ideas off the ground?

Yet, it's important to note that funding for infrastructure is a real problem. No one likes to talk about infrastructure. Fewer people want to pay for it. If one think about one's own home, generally people don't really think about the plumbing until it breaks. Funding maintenance of scholarly infrastructure is an even bigger problem. Traditional funding mechanisms for resourcing research projects are generally short-term and term limited. Yet there are few funding models that provide ongoing operational support.

Format Transformations in the Future

Style	1992	2002	2012	2022	2032
Articles	Print	PDF	PDF/HTML	HTML5	Distributed
Monographs	Print	Print	Print/EPUB?	EPUB	EPUB
Research Data	Astronomy, maybe	Astronomy, Chemistry	Explosion of Data	Focus on Data Management	FAIR-REST ready
Preprints	ArXiv	DSpace	Ubiquitous Repositories	Subject repositories	Distributed Web
Annotation	Not implemented	Coming soon	Prototypes (Annotea)	Hypothes.is	Native web annotation
Discovery	A/I Services	Metasearch	Google	Google!	AI-driven
Authoring	Word/Word Perfect	Word	Word	G Docs	Cloud/AI supported
Video	TV	Adobe Flash Video Player	YouTube	Video Platforms	Embedded everywhere
Presentations	In-person	In-person	In-person	Zoom	AR/VR
Distribution	Agents	Online one-off	Big Deal	Open Access	Open Science
Identifiers	ISBN/ISSN	DOI	ORCID	RoR	RAiD

Fig. 2. Format transitions from 1992–2032.

6. Transformations in the scholarly record format

Building on this infrastructure, we are now well positioned to explore the future of scholarly communications. As noted at the outset, with the quote from Steve Jobs, we'll take one last look at where we've been before moving onto my thoughts of what will be. I'll do this by examining the last thirty years of technology as a grounding for what areas will change in the next twenty. I do this by examining the various formats and outputs of scholarly communications, such as articles, books, datasets, annotation, discovery, authoring, presentations, distribution and the newly available supporting identifiers.

In the start of 1992, the World Wide Web was only 5 months old. There were hardly any online journals, and almost no online books, save Project Gutenberg, which hosted fewer than 100 titles on the internet. Research data was an issue that NASA and the astrophysics community had started to address at scale, with something like 200 GB of data stored and managed. The first preprint server, ArXiv has just started in 1992. Annotation, first environed by the developers of hypertext was dropped from the implementation of the WWW. The only video available was generally via television/cable. Meetings and presentation were almost always held in-person. The primary delivery method for the majority of content was physical. There were basically two information supply chain identifiers for content, ISSN, and ISBN.

Moving forward a decade, by 2002, several online journal aggregators began to take off, and articles began being distributed online, primarily in PDF format. Monographs were primarily distributed in print, although glimmers of online book publishing were beginning to take shape. DSpace, a joint development project led by Hewlett Packard and MIT launched in 2002. Data publication remained the domain primarily of astronomy, but a few other domains had begun exploring it and some other services were gaining traction. Annotation was still an unrealized vision. Discovery began to take shape as metasearch

across multiple sites, and while Google had been launched some three-plus years prior to 2002, it hadn't achieved market domination in search. Online video, primarily as Flash video, started to become more available, but it was limited by most people's bandwidth. In the identifier space, the DOI system was launched for services and standardized in 2000, taking off and finding considerable adoption by 2002. Online content distribution was still mainly delivered via print distributed by agents and jobbers, but online sales to individual institutions was picking up with some consortia starting to pick up group buying.

Another decade passes and by 2012, the majority of articles were online, and open access as a model was starting to gain traction 10 years after the first Budapest Declaration. Monographs were still primarily distributed in print, although online book publishing were beginning to take shape. Many institutions have deployed repositories, but their use was modest. Early annotation systems were being introduced and work had begun on addressing some of the challenging technical problems of a robust ecosystem for annotations. Discovery began to take shape as indexed discovery, with the first applications of knowledge graphs and linked data beginning to take off. Google introduced its knowledge graph and began to solidify its hold on the discovery marketplace, although many vendors remained. Presentations remained primarily in-person, but the tools for reasonable online sharing were emerging. Online video began to expand with the launch of YouTube. Research Data started to gain speed because of the significant investments in infrastructure as I discussed earlier. The identifiers infrastructure grew to include both DataCite (as an application of DOI) and ORCIDs.

Bringing forward this discussion of format transitions to the present day, we find much of the infrastructure in place. Most articles today are created for the web and are distributed in a HTML format, though PDF still exists. Books are increasingly available in EPUB, and if the primary market for scholarly monographs is via the library, that will push more and more into a digital format, probably EPUB. Research data has become an accepted model of distributing content, but now the focus is more on the data management of that content rather than simply posting the files. Google has become the entrenched search default for most users. Authoring, which for years has been based on the author's device has moved to the cloud and easily supports more collaborative writing. Video is now ubiquitous and there are many forms of video sharing applications. Distribution has moved from big deals to Open Access becoming a bigger element of how content is shared. The identifier landscape is filling out as I discussed earlier. Conferences are increasingly virtual, this is because of the pandemic, but the trend started before the pandemic.

7. Trends for the future

What do these trends in the development of the scholarly record imply for the future? Based on the past, we can identify some future trends about this developing ecosystem and some implications of those trends. Before digging too deeply, it is important to stress that technology needs to work with culture and our social environment. Applications of technology are almost always driven by social factors, not technical ones, as described in the use of skeuomorphs. Just because a technology is capable of a certain function, doesn't mean that is how users will apply the technology in their day-to-day application. Technology needs to work with culture. Technology may facilitate change, but it doesn't determine the direction or eventual destination of those changes.

7.1. High quality metadata is the key to discovery and interoperability online. But whose responsibility is the metadata creation necessary for a digitally native scholarly record?

As discussed earlier, metadata is critical to discovery, interoperability, and reuse. It is a key element of many of the FAIR principles and is crucial to allowing users to discovery and use the content that they find. Metadata really is the key to interoperability and discovery. Institutions are recognizing this as a core component of marketing in the 21st century, with resources traditionally invested in advertising and outreach being redirected toward metadata management and improvement. Tools will be configured to highlight the need for improved metadata at an object's source.

Unfortunately, there are three core problems with metadata: First not everyone knows what it is or why it's important. Second, it's complicated to do correctly. Third, is that metadata change over time, so it needs to be maintained. Yet because most people don't understand why it is important or how to manage its complexity, the value to maintaining it is undervalued. This leads to metadata that degrades over time and becomes increasingly more problematic; clogging the high-quality metadata that does exist. Of course, researchers already have too much work to do and are not always well placed to describe their work or connect it meaningfully to the rest of the ecosystem. As the pace of publication has increased and the speed of science accelerates, who is going to create all the metadata that will enable this digitally native scholarly record? Each project may require a data specialist to manage the data and connect its many resources with identifiers and metadata. Beyond educating the world about the importance of metadata, we are going to need to train and hire more librarians who know and can manage metadata. They need to be more integrated in the scholarly workflow of content creation, distribution, and reuse.

7.2. The open-science paradigm allows for content to be replicated, moved, analyzed, and republished in different forms

One of the key concepts behind the open science movement is the ability to share content. The International Science Council report discussed above focusses a great deal of attention on open sharing and the value that it can bring to the scientific process. Today every major publisher has an open access option. Preprint repositories are growing rapidly. Over the decade the amount of open access content has grown from less than 1% in 1996 [33] to 27.9% in 2018 [34], or an estimated 18.6 million objects, and is most certainly far higher today. Now have an extensive network of items that can be—is encouraged to be—reposted anywhere, by anyone. But this capacity for objects to move around the network and be replicated, repurposed, and reused also has its downsides. For example, there is no easy way to assess the impact or reach of an object that has been replicated, since collecting usage from around the network and analyzing it becomes impossible at scale, when you don't even know where copies might reside. There are other downsides, such as the proliferation of versions, and authority control.

7.3. In a distributed data ecosystem, how do we build a notification system to connect the disparate pieces?

One particularly important challenge with a distributed network is that a certain copy of a thing might not reflect a change in the status of the canonical version or another related thing at a different point in the network. Say, a published article has been withdrawn, how does someone viewing the preprint server know since there is currently no way to reflect that change in status? Or that a preprint has been published and is now "officially vetted"? Or say the data set for this paper is now available or updated, and that

users can now go look at it? The scholarly record is constantly changing, growing, and adapting, and we need to develop ways to make the network more "aware". NISO has launched two related projects, one on retractions [35] and another on publisher and repository interoperability [36]. Other work on this is also ongoing elsewhere. The EU funded Confederation of Open Access Repositories (COAR) [37] project has launched a new initiative called Notify [38], which aims to connect the ecosystem of repositories with a notification system, for recognizing when the status of things in the system changes. It's based on the same technology that you might see on a website—such as "This has been liked on Facebook by 200 people" or "300 people are tweeting about this". In a scholarly context it might be that this object has been peer reviewed, or was it retracted, or the related data set is now available. The network of science is growing, and it will need to be robustly connected to ensure these distributed elements are contextually aware of the status of related things in the network.

7.4. The economics of creation of the scholarly record and its distribution are being up ended because of openness and the shift away from subscriptions. Can publishers effectively manage supporting dual systems infrastructure (i.e., both subscription and open) without more effective workflows?

The shift away from subscriptions is upending not only the business models, but also the support infrastructure that publishers have in place to manage those models. Publishers who are not purely open access in their business model, which is most publishers, are having to create multiple workflows and support systems to manage these two income models and having to adjust them as each new change in the business model appears. It took the publishing industry decades to develop effective methods of receiving payments, processing and fulfilling orders for print content to customers. In a world where a publisher has not tens of thousands of library customers around the world, but they have millions of authors, each with their own mandates, their own institutional policies, and their own subscribe-to-open plans, the complexity of that order workflow expands exponentially. Answering the questions, who pays for what, how, and how much will become customized to the author, which will create a nightmare situation for managing the sales and order process. It is also going to require connections between historically unconnected systems. In a traditional publishing house, there was no need for the author submission system to be tied to the accounting system since the two business functions were segregated. Now they aren't and there is a business requirement to connect those systems. It will take years of effort to build an effective and robust infrastructure that supports a model built around something other than the subscriber pays, be that author payments, or institutional read and publish arrangement, or funder supported models.

7.5. Scholarly publishers are investing heavily in automated processes to support the increased pace of content creation and to maintain profitability. Can machine learning and automated metadata creation support the process of creating a new scholarly record?

Publishers have invested heavily in technology to support the increased workflow. For example, machine reading systems are in production today that can do pattern matching on text to discern plagiarism. Keyword and entity extraction tools also exist, which can connect items for discovery on themes within a field. Unfortunately, within the technology world, there is an article of faith that we can just have the machines do all the work. And if the computational power doesn't exist today, it will very soon. I'm more skeptical. Not because various domains of artificial intelligence are not making significant progress; they are. It is that some of the challenges that we undertake in scholarly communications are hard by nature. Take cataloging for example, there has been debate for decades about vocabulary use in

some domains. Frequently, the most interesting discoveries take place at the boundary points between domains, which fit into neither and new terminology needs to be developed. Humans have a hard time describing things accurately. Or plagiarism detection as another example, which may have many forms. A work published in Portuguese and then translated verbatim into English is still plagiarism, but existing automated tools are not sufficiently robust to capture that kind of deception. We will be astounded at the pace that machine language processing tools will develop and play an increasingly valuable role in supporting a variety of publishing functions. However, they are and will remain supporting tools, not the primary methods of producing the highest-quality outputs that scholars rely on and expect.

7.6. The pandemic has transformed the nature of scholarly (all?) interactions and virtuality will remain a core component of work moving forward. Can we create an access control system that is truly viable and adoptable, world-wide across all institutions?

The pandemic has allowed a large portion of our community the flexibility to work remotely, and it has shown institutions that their staff can be effective remotely. Despite the move toward open content, readers will continue to exist in a world where not all content will be free, nor will all services be free. We will have to maintain access control systems for some things. Those things might be the content of some part of the scholarly record, or it might be access to scholarly research tools or infrastructure. It might be a method for validating your affiliation so that you can get access to discounted APCs. The robustness and security of those systems will be a key feature moving forward. An important question will be how to make a viable system that applies across all institutions, across all services, and across users. How do we ensure it is secure, but also user friendly? I point here to the work of NISO, STM, Géant and Internet2 on the Seamless Access initiative. In some ways, though the scholarly publishing community can only be a partner in these conversations, since the technology stack needs to be implemented at the publisher or service provider, the subscribing institution, and be adopted by the end-users to be effective. This delicate partnership will take a long time to grow as trust is difficult to build, easy to break, and relies on functional technologies that are regularly changing.

7.7. People will only change the systems they use if they are motivated to do so, either by better benefit to them, or because they are forced to

Finally, we need to talk about motivations. And let's be clear, users of technology generally don't just move, they are driven to move. Sure, there are some of us who try new things, because we enjoy it. Most people don't. They just want it to work to serve their needs. WordPerfect functions as a word processing program. Why don't people use it anymore? In some ways it is because the feature sets of modern word processors are more extensive. In others it is because the previous version one used to work with is no longer compatible with your new hardware of operating system. Motivations don't need to be external, but they need to be sufficient to overcome the inertia of existing approaches. New tools might have new capabilities, might be sufficiently cheaper, or they might work with a new infrastructure when the old ones didn't. As we consider the changes that new technologies will bring to an ecosystem, it is important to consider what motivates the users and drives them to change their behavior.

7.8. Is the recognition system adapting to incorporate these new scholarly record creation tools such that it will motivate its use in practice?

I'll end here by focusing on what motivates a scholar to share her results. It may be requirements from her employer, determination to progress in her career, or simply the motivation to have an influence on the broader world or the trends in her domain. In many ways it comes back to impact, and this should be considered in the light of these changes. Revisiting data sharing as a practice, which I described earlier, the early adopters understood that the real driver of data sharing at scale would come with development of a recognition system and the tools necessary to drive that. When scholars began to see real meaningful recognition for their data sharing, either through beneficial reuse of data, data citation, or a funding requirement or some combination, then data sharing will really take off. The same could well happen for the sharing of code, or videos, augmented reality experiences, or assessment tools. It's hard to know which technology will rise to the level of widespread use, but it will need infrastructure to support it, to make it easy and consistent, before everyone adopts it.

8. The scholarly record in 2032

Taking these trends in mind, let's extend our view of tools and methods a bit further, looking out to 2032. Many interesting things will happen in the expansion of the scholarly record. Articles and monographs will continue to exist, but print will be an outlier for most but a small group of mass market items. Articles and preprints will become interchangeable as the nature of the fixity of the scholarly record adapts, but with connections that support contextual awareness. Data will continue to grow and new forms of content forms will become supported, such as visualizations, audio visual, and code, which will become increasingly important formats in the new scholarly record. Data will be shared not simply as a file, but it will be plugged into the entire ecosystem, most likely in a domain-specific repository with other related data, tools, and analysis tools. Presentations will no longer just be on video calls like this conference; Virtual reality will become a regular feature of engagement at distance. Authoring will take place in the cloud, although very likely it will not take place through typing, but by text to speech. Authoring will also be AI supported, connecting knowledge the author might not be aware of, and facilitating easy citation and document linking as part of the authoring process.

We're in the very early days of this transformation. Many of us working in the scholarly communications industry can recall a time before ubiquitous computing and natural language processing of huge text corpus. It is unrealistic to expect the world to transform magically because of the next new device or system release. These changes will take time. For reference, it has been estimated that in the 1450s, some two decades after the invention of the Gutenberg printing press in Europe, fewer than 10% of books included page numbers. It wasn't until the first decade of the 16th Century that scholars started to use page numbers [39] to mark the location of their work. That was a period of rapid change in how content was created and distributed, which led to deep societal change. Using this one guide for how technology changes as a guide, we're right on track with some of these new developments.

Technology changes far faster than the cultural changes that are required to drive widespread adoption. While this talk has primarily been about technology, in the end it will be the culture that will drive the changes in how the scholarly record will change.

References

[1] Stanford Report, June 14, 2005. *You've got to find what you love, Jobs says*.https://news.stanford.edu/news/2005/june15/jobs-061505.html.
[2] Illuminated Manuscript, Compendium of computistical texts. Website: https://www.flickr.com/photos/medmss/5613680005/.
[3] V. Bush, As we may think, *The Atlantic Monthly* **176**(1) (1945), 104.
[4] T. Nelson, *Literary Machines*. Ted Nelson, Swarthmore, 1982.
[5] B.-L. Tim and C. Robert, (12 November 1990). WorldWideWeb: Proposal for a HyperText Project. Archived from the original on 2 May 2015: https://web.archive.org/web/20150502080527/http:/www.w3.org/Proposal.html.
[6] C. Mike, That Time Berners-Lee Got Knocked Down to a Poster Session. April 21, 2015. https://hapgood.us/2015/04/21/that-time-berners-lee-got-knocked-down-to-a-poster-session/ (accessed on 12/18/2021).
[7] Quote Investigator (Website). My Customers Would Have Asked For a Faster Horse https://quoteinvestigator.com/2011/07/28/ford-faster-horse/ (Accessed on 12/18/2021).
[8] H. Tony, T. Stewart and T. Kristin, *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Seattle, WA, 2009, ISBN: 978-0-9825442-0-4.
[9] National Science Foundation, Office of Advanced Cyberinfrastructure. *Sustainable Digital Data Preservation and Access Network Partners (DataNet)*Program Solicitation. Website. https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141 (Accessed on 12/18/2021).
[10] National Science Foundation, Cyberinfrastructure Council. NSF 07-28, Cyberinfrastructure Vision for 21st Century Discovery. March 2007. https://www.nsf.gov/pubs/2007/nsf0728/index.jsp (Accessed on 12/18/2021).
[11] DataONE. https://en.wikipedia.org/wiki/DataONE (Accessed on 12/18/2021).
[12] Data Conservancy. https://dataconservancy.org/ (Accessed on 12/18/2021).
[13] SEAD Sustainable Environment — Actionable Data. http://sead-data.net/ (Accessed on 12/18/2021).
[14] DataNet Federation Consortium. http://datafed.org/ (Accessed on 12/18/2021).
[15] Terra Populus. http://www.terrapop.org/ (Accessed on 12/18/2021).
[16] Australian National Data Service. https://ardc.edu.au/ (Accessed 12/18/2021). Note: ARDC was formed on July 1, 2018, by combining the Australian National Data Service, Nectar, and Research Data Services.
[17] Data Europa. EU Publications Office. http://data.europa.eu (Accessed 12/18/2021).
[18] Closing RSP. *Repositories Support Project*. https://rspproject.wordpress.com (Accessed 12/18/2021).
[19] Digital Repository Infrastructure Vision for European Research (DRIVER). https://cordis.europa.eu/project/id/212147 (Accessed 12/18/2021).
[20] Task Group on Data Citation Standards and Practices. Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. Data Science Journal. 12, pp.CIDCR1–CIDCR7. DOI: 10.2481/dsj.OSOM13-043.
[21] National Research Council. 2012. *For Attribution: Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop*. Washington, DC: The National Academies Press. doi:10.17226/13564.
[22] Amsterdam Manifesto, FORCE11.org website. https://force11.org/info/amsterdam-manifesto/ (Accessed March 6, 2022).
[23] FAIR Data Publishing Group, FORCE11.org website. https://force11.org/group/fair-data-publishing-group/ (Accessed March 6, 2022).
[24] The size of the World Wide Web: Estimated size of Google's index. WorldWideWebSize.com website. https://www.worldwidewebsize.com/ (Data from January 9, 2022).
[25] How Much Content Is Good for SEO Rankings? Whiteboard Marketingwebsite. Tue. Feb 13, 2018 https://whiteboard-mktg.com/blog/how-much-content-is-good-for-seo-rankings/ (Accessed March 7, 2022).
[26] This is part hypothetical estimation, part analogy. Not every crawl of every website is indexed. In fact, there are matching algorithms that only return the changes between an indexed version and the latest version. What characterizes "reading" by machines and how it differs from human reading is also an entirely different philosophical conversation, well beyond the scope of this paper.
[27] International Science Council. 2021. *Opening the record of science: making scholarly publishing work for science in the digital era*. Paris, France. International Science Council. DOI:10.24948/2021.01.
[28] DataCite Resolution Statistics, September 2021, https://stats.datacite.org/resolutions.html (accessed on 12/18/2021).
[29] ORCID Annual Report 2020: Connecting Research and Researchers. March 11, 2021, https://info.orcid.org/now-available-orcids-2020-annual-report/ (accessed on 12/18/2021).
[30] CrossRef. About the Funder Registry. https://www.crossref.org/pdfs/about-funder-registry.pdf.
[31] OpenAlex: About. https://openalex.org/about (Accessed on March 3, 2022).

[32] Microsoft Academic Blog. *Next Steps for Microsoft Academic – Expanding into New Horizons*. May 4, 2021. https://www.microsoft.com/en-us/research/project/academic/articles/microsoft-academic-to-expand-horizons-with-community-driven-approach/ (Accessed on March 3, 2022).
[33] É. Archambault, D. Amyot, P. Deschamps, A.F. Nicol, F. Provencher, L. Rebout et al., European Commission *Proportion of open access papers published in peer-reviewed journals at the European and world levels–1996–2013. 2014.* http://science-metrix.com/sites/default/files/science-metrix/publications/d_1.8_sm_ec_dg-rtd_proportion_oa_1996-2013_v11p.pdf (Accessed on March 6, 2022).
[34] P. Heather et al., The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles, *PeerJ* **6** (2018), e4375. doi:10.7717/peerj.4375.
[35] NISO Voting Members Approve Work on Recommended Practice for Retracted Research (Press Release). September 22, 2022. https://www.niso.org/press-releases/2021/09/niso-voting-members-approve-work-recommended-practice-retracted-research (Accessed on March 6, 2022).
[36] NISO Announces New Project to Integrate Publisher and Repository Workflows (Press Release). October 27, 2021. https://www.niso.org/press-releases/2021/10/niso-announces-new-project-integrate-publisher-and-repository-workflows (Accessed on March 6, 2022).
[37] Confederation of Open Access Repositories https://www.coar-repositories.org/about-coar/ (Accessed on March 6, 2022).
[38] EC–Confederation of Open Access Repositories (EC-COAR). The Notify Project website. https://www.coar-repositories.org/notify/ (Accessed on March 6, 2022).
[39] B. Naomi, Words Onscreen: The Fate of Reading in a Digital World. Oxford University Press; Reprint Kindle edition. September 1, 2016. ISBN-13: 978-0199315765.