In this perspective I outline my thoughts on why creating interconnected data and metadata systems is beneficial for the development of a robust knowledge ecosystem. My viewpoint is grounded in my experience working in the technical and product development areas of scholarly publishing, seeing first hand the code, systems, and costs, involved in keeping the lights on, so to speak. I have worked on building some of the earliest social web infrastructures for science (Connotea, Nature Network), contributed to the establishment of new standards for data citation, been involved in the creation of open source publishing platforms (https://github.com/elifesciences), and managed product and software development teams in this arena for about twenty years. This is very much a personal perspective informed by these experiences.

I am also going to frame this perspective in relation to the work that https://democratizingdata.ai/ is doing. Other papers in this collection will look more closely at the emergence and evolution of US Government requirements about evidence based policy making, and the historical development of directives, initiatives, and infrastructure to support that, but my own view is trans-governmental, as publishers occupy a space in the ecosystems apart from funders, researchers, governments and institutions, but in support of all of those.

I acknowledge that what I present here doesn't provide a complete picture, nor is it referenced, but I hope that it can help to foster a broader appreciation of some of the opportunities we have today and open some perspectives on how we can collaboratively make the most of them.

The key points I make are:

- Our knowledge systems have never been static, we can, and should, be reexamining how we operate. We do not need to be locked in to working today, as we have in the past.
- The volume of knowledge production has grown in scale to the extent that we now need new forms of infrastructure and investment to make the most of the knowledge that we are creating.
- There is a significant shift now form the journal to individual impact, and Publishers are changing to support that.
- The work of https://democratizingdata.ai/ is directly aligned with these changes, and as a publisher I welcome them.
- We have a long way to go ahead of us.

We are now working within a knowledge ecosystem that has scaled up remarkably over the last few decades. OpenAlex's data (https://openalex.org/works?sort=cited_by_count%3Adesc&column=display_name,publication_year,type,open_access.is_oa,cited_by_count&group_by=publication_year), an open-source/open data alternative to Google Scholar, showed only a handful of publications per year in the mid-19th Century. That number increased to about half a million papers per year by the mid-1960s, exceeded a million in 1975, and an astonishing 10 million by the end of 2020. This is

a remarkable increase in the scale of scholarly output since the first issue of the Philosophical Transactions of the Royal Society about 350 years ago.

Continued growth in investment in research has been a driver of this (https://ourworldindata.org/grapher/public-health-expenditure-share-gdp). R&D expenditure in the US alone has grown at a steady 5% CAGR from 2010 through to 2021, with health expenditure growing at about 3% CAGR over the same timeframe. The move to online publishing has also contributed in the increase in scale, and again this is relatively recent, with the first online journals approving in the early 1990s (the BMJ whom I work for went fully online in 1995, a year ahead of the New York Times).

We have this great increase in scale. I'm going to state that scale is good when it comes to our ability to understand the world, yet operating in a coordinated way at scale is challenging. This is a broad statement, so I'll take a moment to unpack it. In terms of why scale is good, my perspective is informed by a concept from cybernetics about the degrees of freedom our systems need to exhibit to maintain stability in a complex environment. This is Ashby's law of requisite variety and is often phrased as: "To adequately deal with the diversity of problems the world presents, you need a repertoire of responses that are (at least) as nuanced as the problems you face." Science aims to understand the world, which is essentially open-ended in scale, so we need systems that support as diverse a set of participants as we can. But why is it difficult to operate at scale? I want to present three thought experiments about organising knowledge in the world.

In one scenario, governments have a controlling role in managing the systems that accredit knowledge. They lay out a central agency with its rules and regulations on what constitutes knowledge. This allows for strong coordination, and the government agency can modify its own rules. If different rules are developed by different governments, then they can come together under a treaty process to create alignment. Conceptually, this is how the modern patent system operates, but it struggles with scalability. As the number of patent applications grows, the agency face a bottlenecks in processing those applications.

The second model might involve a single centralised commercial entity accrediting knowledge. If this entity can generate a profit, it has the potential to scale the required resources for operation. It can coordinate internal policy and practice but comes with the problem of not being independent and only able to allocate resources where profits can be made.

The third model is a distributed one, where communities of experts gather once a minimum threshold of interest has been generated. They create their community of practice and validate knowledge claims through a shared medium - a journal. The advantage of this model is the flexibility of scalability, but it presents a significant coordination problem, as each journal operates independently. By its distributed nature, it can be challenging to get an accurate count, but in 2020, STM estimated the number of journals to have exceeded 45,000, up from 25,000 in 2001. In this environment, if we want to introduce a new practice (e.g., data citation), we have to

wait until that practice permeates a critical volume of this set of journals before the new practice becomes embedded, making changing practices at scale difficult.

And yet as our knowledge, systems and practices continue to evolve. How we cite, how we use measures such as the impact factor, and even peer review, which only assumed its current form in the early 1980s, have all evolved.

The business model supporting scholarly publishing is also evolving, transitioning from library-subscription-based to a transactional model where payment is made for services involved in publishing an individual paper (Open Access). With this shift, the payer often transfers from the library to the individual researcher or granting body.

Most of our current digital publishing systems in use were built around the subscription model, and a model where the journal is the key container. This maps through to submissions, systems, and the key types of metadata that we attache to articles and research objects. As a publisher it used to be sufficient to produce key usage indicators at the journal level. In an Open Access model, wherein the payer purchases services around individual articles, it becomes critical to provide data at the individual article level. Consequently, we are now trying to build systems that shift our perspective from journal-centric to article-centric.

Being able to give an individual information about how they interact with a service in the round is common in most modern web applications that we use, but remains somewhat lagging in the scholarly publishing space. We should be able to provide clear information about the performance of submitted articles, the status of articles they have reviewed, or data they have used in their papers reused by others. This could hep with the story around impact, and furthermore, having a solid understanding of a researcher's interest can help us build more robust and resilient systems.

The point I want to make here is that how we generate knowledge, and the processes that we put in place around those activities are constantly evolving. We can choose to be proactive in how we decide we want those systems to change in the future.

As publishers we are seeing a clear market demand for a move to being able to support an understanding of impact. The BMJ has always tried to innovate in this area, having been the first journal to publish a centrally randomised condoled trial (1948), and the first to introduce patients as peer reviewers for papers impacting on specific disease areas, but the need to evidence impact is only increasing. Impact can often be ambiguous and excellence can often be fetishized, so it's important to connect data to impact narratives where possible. For example, BMJ has invested in the development of BMJ Impact Analytics, which uses Natural Language Processing (NLP) to extract references to academic papers from policy documents, helping to understand the research-to-impact connection.

PLOS has developed a set of Open Science Indicators which "identify and quantify instances of specific Open Science practices"

(https://theplosblog.plos.org/2023/10/open-science-indicators-q2-2023/). The Chan Zuckerberg Initiative has invested in the creation of software mentions in academic papers (https://medium.com/czi-technology/new-data-reveals-the-hidden-impact-of-open-source-in-science-11cc4a16fea2).

Although it's easy to imagine the abstract graph of research objects' connections and their interactions with different steps in the research process, each of these initiatives has required time, investment, and some level of sustainability if these signals are to be generated over a significant duration.

To illustrate why this kind of work is beneficial, let's think a bit about how it relates, and is different to citations. Citations are acknowledgments from one paper to another, admitting to prior work, but critically the influence of the work is not delimited only by the citation. For example, many people who read the paper do not necessarily cite it. Perhaps the work changes their perspective, leading to a shift in behaviour. There may be uses of the paper's data that do not result in another paper but are still valuable. There may be instances of the paper appearing in policy documents without formal citation. The full life of the research object casts a digital shadow, with citations being just one projection of influence. One of the enduring reasons that citations are so powerful is that we can measure them, and we have invested in significant infrastructures to make that possible, from creating handles and DOIs for research objects, to publishers depositing that information with a central independent agency (CrossRef), which was created and funded by publishers for this purposes.

Several systems have been established over the past few years to offer a better glimpse into the more complex life of research objects beyond citations - Altmetric.com tracks social media mentions of research papers, and more recently, Overton.io has been developed to text mine policy documents and to map the connections from those documents to the research literature. For a few years, many initiatives built on top of the Microsoft Academic Graph, but that resource was retired. Subsequently, the Allen Institute absorbed some of that open data into Semantic Scholar but didn't keep the graph updated. OpenAlex has filled the void to create an ongoing resource, which is fantastic for now but the numbers of organizations working on making this information available remain few in number. Having an increase in the number and variety of organisations generating this information increases the likelihood that we may find a sustainable and more resilient way of generating this information, through diversification of effort. The layer of connection between researchers, datasets, publications, and government agencies that https://democratizingdata.ai/ is creating is the kind of information we need to access to, in order to create these kinds of connections effectively.

I want to come back to now to why I think these efforts are important. I've referenced impact, and that is probably the key driver, but there are many other benefits available to us if we can make our systems, and the quality and relevance of the information that our systems generate, scale along with the scaling that we see in the production of knowledge.

There are many efficiency gains to be had, mostly encapsulated by Clay Shirky's idea of filter failure vs information overload. Better data could help us establish enhanced connections across communities or disciplines of practice. The "holy grail" in medical research is the transition from bench to bed, which involves applying fundamental research in medical practice. But bridging that gap requires the ability to connect across multiple boundaries of practice. If we can scale insights about who is doing what and with which type of data, we can make it easier to connect researchers who are solving common problems. More excitingly, we can find ways to connect disparate people, where the cross-pollination of ideas can stimulate real innovation.

One way to frame this is to consider is whether datasets could be regarded as social objects, akin to a perspective we took when contemplating the architecture of Web 2.0 sites. In this framing, popular sites involving social sharing could be thought of as having a core social object. For Instagram, that object is the photograph; for Spotify, it's the song; and for Facebook, it's the friendship connection. Could initiatives like democratizingdatabase.ai transform governmental datasets into social objects, and by aggregating information about their use, create interest communities that transcend disciplinary silos? To this end, I would love to see the https://democratizingdata.ai/ project explicitly expose identifiers like ORCID records and FundRef records. They've already begun bridging the gap between governmental datasets and the publications that cite those datasets. These publication identifiers could further bridge to other identifiers that could create a richer tapestry of connections.

These kinds of connections should enable us to allocate resources more effectively. From a publishing perspective, one of the hardest things to do is finding the right reviewer for the appropriate paper. Needing to cycle through multiple reviewers is an inefficiency in the system. Additionally, when we apply selection criteria to a journal, such as the scope or community interest, there will be papers that get rejected from one journal only to be published elsewhere. This also represents a waste which better data could help us reduce.

There are also benefits to be derived around maintaining the integrity of the scholarly record. Fake papers are an increasing problem (currently about 2% of papers submitted to journals are fake, and the number is growing). An indicator of potential fraud is when a researcher's co-authorship network varies wildly between sequential papers they publish. To analyze this requires systems that develop individual profiles of researchers' interests and activities. The STM integrity hub is another example of this work (https://www.stm-assoc.org/stm-integrity-hub/) where publishers are collaborating on detecting signals at scale of fake papers.

Of course we are talking here about stepping up the granularity and specificity of connections between people. Research objects, and connected entities in the research graph. This also opens up the door to tying this information to behavioural data. BMJ had over 770 million access requests to the content that we published in 2023, most of them anonymous. We believe that if we can understand our audiences better, we can better server those audiences, but we need to acknowledge the potential pitfalls a program like this can present and identify what steps we must take to operate in a way that is respectful and valuable to the overall research ecosystem. I've seen the phrase "surveillance publishing" which I think captures that unease

well. By building upon a vast network of available data, we can increase our ability to automate aspects like marketing messages, review requests, and journal submission requests. Each one of these activities represents a call for the researcher's attention. We have to determine if we can provide recommendations to the researcher that are ultimately more valuable to them than not, and that drive efficiencies in the system, avoiding the risk is that these systems become net consumers of attention without adding any efficiency for researchers. I like to think about this not in terms of increased messaging, but increased quality of the messaging that currently happens.

The opportunity to use data to focus attention and build new connections is clear. While the impact factor is mostly criticized, it represents a proxy for attention, but it is a poor proxy that does not scale well in our new environment. By using better information we should be able to direct attention in a more equitable way than the limited dimensions in use today. Another essential aspect of how systems like these operate safely and respectfully is that we must be transparent about what we're doing and work through appropriate consent frameworks. Having researchers at the heart of our journals is a necessary condition for us to create these kinds of systems in a way that works for the community.

I want to conclude with the following note. The scientific creation of knowledge may seem deeply entrenched in our societies, but it has only been around for a few hundred years—a blink of an eye. Our use of the internet and internet scale technologies may seem fixed, but they are even newer. Devices and techniques that can function at scale are only now becoming more commonplace within our research infrastructures. The future lies ahead, and we can shape it for our mutual benefit. Creating good ways of scaling collaboration, decoupling systems while finding common ground through identifiers, and ensuring trustworthy provenance of data can all help us create a more resilient, scalable, and better research future—something we can collectively help to build. I've been working on many of these types of systems over the past twenty years, and I'm surprised to still find myself engaged in this work. I thought it might be easily solved, but making the data available remains a key issue, and I am excited for what the next twenty years brings.