OPEN KNOWLEDGE INFRASTRUCTURE[1]

13 MAY 2025
Keynote at the international Workshop "Open Metadata: Libraries and Open Science", organized by Stanford University Libraries, honoring the 80th anniversary of the National Library of Korea, Seoul.

SUMMARY

Information infrastructures for exchanging and producing knowledge in a digital form can be seen as the most important cultural achievement in the 21st century. Artificial Intelligence now boosts the further development of information infrastructures in new ways by providing ready-made, interpretative knowledge resources, resulting in an unprecedented abundance of information and a societal challenge to find an appropriate control of the balance between human and artificial knowledge resources. Science and libraries, as societal institutions with a global scope, traditionally fulfil the role of custodians and curators of knowledge. Yet, an open knowledge infrastructure provided by science and libraries has not emerged in the form of a unified digital system that is used on a daily basis in families, schools, industry and politics. The role of libraries in further building such a unified open knowledge infrastructure should therefore focus on the global collaboration and embrace artificial intelligence while addressing the challenge how humans maintain and control the responsible use of information.

Introduction

Infrastructure for information became an essential, if not the most decisive development of human society in the 21st century. Like the commodification of electricity in the 19th century or the invention of cars or nuclear power in the 20th century, information is shaping the wellbeing and the fate of humanity today and will continue to be a dominant issue for the future. More and more information resources become available and more and more digital tools, including Artificial Intelligence (AI), penetrate areas of human life that were formerly working with little or no digital information: eating with calory counters, recipes and cooking devices, sleeping with monitoring tools, motion with heart rate control and performance measurement, playing with gaming environments. And, obviously, social behavior is almost disappearing as a pure, non-digital form of interaction. Even love is made dependent on digital helpers. Today, it would not come to a surprise if parents educated their children that they must not date any person that is not vetted through a digital social platform.

However, the most relevant human need with respect to information is curiosity, the need for knowledge. Sleeping, eating, moving, playing, communicating and loving – every single way

1 *This paper is created by a human body, without using search engines, literature databases or AI – only with a digital word processor and vocabulary look-up. As such, the paper is drafty, subjective, speculative, and entirely non-scientific.*

that humans make use of digital information involves knowledge. Knowledge is the common denominator of digital information use. But make no mistake: as opposed to digital information, knowledge is not digitally encoded. Our senses, our brain and our behavior have to convert digital information into analogue signals and generate knowledge. Knowledge is analogue. We use digital information to be better prepared, react faster or to behave in a way, we did not think of without digital help – but all has to go through our sensory filters and has to arrive in our brains to be further processed. For the sake of the argument I am developing in this paper – for understanding knowledge infrastructure, libraries and AI – it is essential to make this point painstakingly clear: digital information and tools are resources that produce sensory inputs for our brains. And our brains are not digital. A human is a biological system. The difference between the analogue, non-digital nature of knowledge and the possibly-digital nature of information is particularly important for understanding AI. AI is a form of information processing that feeds our sensory inputs, triggering biological information processing in our brains. We should not, and I even would say, we must not, treat AI as sentient or even conscious thing, like humans tend to do with cats or dogs. Neither AI nor cats nor dogs have a culture of knowledge the way humans have a culture of knowledge. When we attribute to AI that it has, for example "hallucinations" and thereby imply that AI can have conscious states without hallucinations, we are making a mistake. If we do so, we as humans create a world in which AI can be intelligent because we make the world so that AI is quasi-intelligent. It is only a simple logical mistake to attribute intelligence and control over knowledge to AI. But making that simple logical mistake can make it happen. If we attribute AI to get control over knowledge, it will get it because we make it so. It is all in our hands.

Knowledge Infrastructure

I define knowledge as enacted or embodied information in humans. In this sense, "knowledge infrastructure" is an optimized means for humans to generate knowledge. A human interacts with a knowledge infrastructure to get input for the brain that generates knowledge that is used to optimize behavior. The most obvious and historically rooted examples of knowledge infrastructures are Libraries. Libraries make collected information resources available. In the context of paper materials, libraries are a local societal institution. In the context of digital information, libraries are a powerful network and, probably, the most important societal institution with respect to the future of digital information in humanity, globally. We will come to that promise of libraries, later.

Some other examples of knowledge infrastructures: News, that is newspapers, journals, magazines and, now, digital news channels provide current and interpreted information resources that are designed to support the knowledge generation process in humans. News is not knowledge – as explained above. But news is a typical knowledge infrastructure. News has a higher degree of interpretation and, thereby, fabrication potential than typical library resources. Social Media, as another form of knowledge infrastructure, recently took over parts of the roles of news platforms. Today, it is nothing unusual that humans retrieve government election results on Instagram. Whether or not this amalgamation of knowledge, news, social life and fabricated information was a positive development remains to be seen.

Science and knowledge infrastructures

What is the role of scientific knowledge infrastructures as opposed to other infrastructures such as those for information, news or entertainment? Scientific infrastructures certainly fulfil their roles in society, for bringing innovation and fueling economy. But they also seem to fail to rise to the expectations. Consider the role of science during the COVID-19 pandemic. Did science really perform well in consulting politics about the management of the pandemic and did science successfully explain to society how to deal with COVID-19? Luckily, the vaccination

came to a rescue for large parts of society and science played a role in that. Science got out of the pandemic fine.

But look at what also happened. The massive rise of conspiracy theories in several countries indicates that large parts of society had an intuition that something is wrong. To be clear: I am not saying that conspiracy theories are justified. They are not. What I am saying is that the observable desire to question the "truths" that were communicated by the major knowledge infrastructures clearly shows that many people did not believe anymore in what was said. This is to be taken very seriously. Also, the success of prominent political figures and their news outlets to bend knowledge and use this bended knowledge for their power indicates that there is a fundamental problem in the way humans currently deal with information and knowledge. The confusion is ubiquitous and dangerous.

Who or what, if not scientists and librarians can limit the dangers of the current confusion about information and knowledge? Scientists know that truth is relative. But does society know that? Science is the generator of new knowledge and responsible for many of the innovations that make this new world of information and knowledge possible. But the question now is whether the world of information and knowledge has changed in a radical way and science did not sufficiently change with it? We now live in a world where information is abundant. Universal access and mostly open access to scientific information and, even more important, the mass of digital information that humans are currently processing might change the nature of truth. It is obvious that a combination of science and libraries would be the most powerful and adequate knowledge infrastructure that can tackle this challenge of explaining the changed nature of truth.

AI and knowledge infrastructures

Similar to the challenge of explaining "truth" in an overwhelmingly noisy world of digital information, AI is another challenge to be tackled in knowledge infrastructures. AI was developed in science and is now part of everyday life. Whether AI is really about intelligence is still an open research question. Scientists know that. But did science explain that to society? And even if AI was about intelligence, the intelligence would not be artificial but human. Currently, AI is simply an information processing tool that can be prompted by humans for responses that humans are most likely to expect. A better name for things we now call AI would be Artificial Information Production Systems. AI might have more information and better capacities to produce sensory inputs required for knowledge generation than humans have. But it nonetheless stays a simulation.

AI is also currently the most prominent example to explain the need for open knowledge infrastructure. AI, in its most renowned form of ChatGPT, has been brought to society by commercial players who based their initial language models on mostly non-scientific information, sometimes called "garbage". Yet, it was a success. Today, newer and other versions of Large Language Models can generate scientific papers based on scientific information. The mass applications of AI, however, are in commodity contexts. And the mass applications are in the hands of commercial providers driven by commercial interests.

Where are the AI applications provided by open knowledge infrastructures in science or libraries that explain knowledge and the relativity of truth and AI to society? Have they been built yet? Are they used in schools and everyday life by everybody? No. Instead, the inventions of science and library facilities are applied in commercial or political contexts for making money, manipulating society, gaining control and increasing influence and power. Why are science and libraries, as guardians and curators of knowledge, not successful to produce an AI application that is widely used to deal with scientific information or knowledge as a broader cultural resource? Here is where open knowledge infrastructure is needed. Science and libraries

have the power over the biggest reliable corpus of information in the world. And science has huge computing facilities, is capable and has the mission to produce such a system. Science and libraries are the most potent actors in open knowledge infrastructure.

Obstacles

So, why do Science and libraries fail to provide a powerful, widely usable open knowledge infrastructure, based on AI or other established or even not yet existing technology? I will only briefly mention some of the potentially underlying obstacles. First, of course, there is limited funding. Another obstacle is the political dependence of science and libraries, specifically but not only, with respect to public funding. Yet another obstacle is self-interest and the science assessment system. The competition for limited funding and the intrinsic self-interest in science result in a fragmented information system. Not only geographic boundaries that are influenced by the political dependence of science but also disciplinary competition within science lead to different cultures and ever differentiating scientific communities. In knowledge infrastructures, this fragmentation is massively visible as can be seen by the current trend to develop sovereign digital infrastructures for each region, each nation, each discipline, each continent.

Of course, digital sovereignty is needed, specifically in these times of peculiar geopolitical developments. National libraries, for example, are needed as culturally localized pillars in a global system. In fact, the global library system might even be the only system that can overcome some of the obstacles that hinder the emergence of a global, open knowledge infrastructure. Competition in Science, as mentioned, is limiting the development of unified, simple systems. Say, medicine or physics all have their own storage, compute, curation and preservation. In this sense, they compete about limited funding instead of developing joint aspirations for building a unified system. Of course, redundancy must be in the system in order to prevent single points of failure. But the tendency to always build an individual infrastructure for each discipline can also be questioned. The knowledge infrastructure community could be more self-critical with respect to answering the question how many parallel systems are really necessary. Consider the fashionable term “interoperability”, for example. It seems to be used not anymore as a serious attempt to build unavoidable connections between infrastructures but rather to excuse the lack of interest to take risks in managing dependencies, building alliances and sharing resources. In discussions about science politics that are about budget allocation it can be observed that interoperability is misused in this way. And, if somebody dares to mention terms such as “convergence”, “complementarity” or “integration”, the discussion gets defensive and heated.

Transcending boundaries through library metadata

Libraries have developed over millennia, independently on different continents and in diverse cultural contexts. Aren‘t libraries an example of convergent evolution of human culture, where the same societal functions of making knowledge resources available and accessible for the Many emerged independently in different cultural settings? It is worth taking time, sitting back and thinking about the question whether there are any other comparable societal institutions on every continent, in every cultural or political context? Libraries are a very rare, if not unique, example of a societal institution that exists across regions, cultures and political regimes.

And libraries have collaboration and networking in their DNA. Even in the era of physical materials such as books, in which libraries where predominantly local institutions, confined by a limited mobility of humans or books, libraries found ways to exchange materials between libraries in order to connect humans and knowledge resources. A regional, national and international “inter-library-exchange-network” was the result. The emergence of this library network was based on bibliographic data and, later, card

catalogues that helped to identify what knowledge resource was available in one library and missing in another library. As soon as it was possible, the card catalogues for bibliographic data were converted into a digital resource in the last century, which marks the birth date of metadata. And libraries were among the very early adopters of large scale IT-systems. In many universities and other institutions, the first enterprise-scale IT systems and applications of mainframes in the 1970s were established by libraries to process metadata that could connect humans to knowledge resources. And the digital metadata system soon became a local, national, global or disciplinary exchange network.

Libraries have built large aggregations of metadata for knowledge resources – be it in the western, eastern, northern or southern world. These aggregated resources are connected but still do not fulfil the requirement of a truly global system of knowledge that is available to all mankind – an open knowledge infrastructure. Current technology allows that libraries can overcome, can transcend residual boundaries. It is not necessary anymore to build aggregations that are monopolizing certain disciplines or regions. It is now possible to expose metadata locally and to build interoperable systems that generate knowledge resource discovery on the fly, in runtime. This has been demonstrated in a limited fashion (with somewhat arcane technology) through OAI-PMH, the Open Archives Initiative Protocol for Metadata Harvesting or, for images, through IIIF, the International Image Interoperability Framework. But libraries can do even better. It is this challenge that, I hope, Open Metadata initiatives, such as this conference" can tackle. What is needed is to fulfil the mission of libraries as collaborative societal institutions that can transcend competition and geopolitical boundaries to build a truly global and open knowledge infrastructure.

Metadata in the age of AI

Before the conclusion, I will briefly address the question of the role of metadata in the age of AI. It can be stated that metadata as artifacts resulting from intellectual, human curation is no longer needed. AI, as machine learning based on artificial neural networks and natural language processing, does not necessarily require metadata because it generates means for the discovery of knowledge resources that are derived "sub-symbolically" by statistical methods applied to unstructured data. Thereby, one of the core qualifications in the library profession for intellectual knowledge organization, often called cataloguing, is threatened to be rendered obsolete by AI. However, it remains to be seen whether statistical, sub-symbolic approaches to process unstructured data will be the only solution for building open knowledge infrastructure. There are good reasons to assume that a combination of sub-symbolic AI and more traditional, symbolic AI, involving metadata, semantic networks, ontologies, knowledge graphs and the like will deliver a better solution for open knowledge infrastructure. Academically, this question might only be answered by research. But societally, the answer is clear: we need libraries as collaborative institutions and librarians as curators and custodians of knowledge to control AI. Therefore, libraries should focus on a hybrid approach to open knowledge infrastructure, involving both AI based on unstructured data and AI based on structured metadata. As mentioned in the beginning, humans should not incautiously or even deliberately hand over control to AI when it is possible to maintain human control over AI.

A "paradoxical intervention"

I conclude with a rather unusual thought to further develop open knowledge infrastructure.

There are many good examples of what Science and libraries already do to progress open knowledge infrastructure. But, long story short, I am not convinced that it will be fast enough. There is a considerable risk that human society can be severely damaged in this new world of information and knowledge. Maybe, it all goes right and humanity adapts fast enough. Maybe,

concerned voices will look pessimistic and dystopian in a bright future that has solved all problems of disinformation, misuse of power over information, the lack of digital sovereignty and ethical breaches of AI-technology. But maybe not, and today's optimism will look naïve in a brittle future society.

In those situations of uncertainty, professional risk management would recommend to have a contingency plan. What could be a contingency plan? What would it take for science and libraries to master the challenge of fulfilling the role as a guardian and a curator of knowledge in a radically changed world?

Many attempts look at controlling the quality and amount of information available. However, there is no way to limit the increase of the abundance of information. The amount of information is ever increasing because humanity finds ever better ways to record, store, process and communicate information. This process is as irreversible as the increase in entropy as it is defined in physics. With the increase of the amount of information, there is more noise and less information in a given signal and knowledge becomes more relative. Attempts at limiting the production of information and the capacity of humanity to produce knowledge, individually and culturally, are noble but futile. Yes, in scholarly communication, for example, we can implement policies about producing less papers or using less research outputs in research performance measurement as many funders recommend at the moment. Yes, we can further develop "slow science", decelerate information production, read less but with more depth. But it will not reverse, only decelerate the process. Science and libraries have to find ways to deal with the abundance of information.

The current political environment is giving rise to more concerns. Scientific information resources in the US are acutely at risk and data rescue activities are underway, globally. More than ever, the need for a robust and resilient knowledge infrastructure is paramount. Many current political considerations move into a direction of national sovereignty. And the call for a massively parallel and redundant architecture is well justified. However, it will be crucial, especially in this current situation of "looking inward" that science and libraries do not lose sight of the overall purpose of knowledge infrastructure in society, namely to provide a unified access to a diverse culture of responsible use of information, foster curiosity and innovation for the betterment of humanity. Against this backdrop of not forgetting the need for global collaboration and common vision during times when self-referential sovereignty becomes a major driver, the proposal for a single, unified knowledge infrastructure may seem like a "paradoxical intervention". Unified, non-redundant systems introduce single points of failure and are prone to attacks. However, a unified system can still be based on an architecture that is robust and resilient as the development of the World Wide Web or Wikipedia has proven.

Eventually, there might be no alternative to embracing these risks and the dissipation of information and knowledge. So, I would like to ask the question, whether we need an international moonshot project, a global program for open knowledge infrastructure?

It is certainly worth to do some initial blue-sky-thinking.

What could be the 5 steps needed for a global program for open knowledge infrastructure?

1. Define an aspiration

The aspiration could be a unified global system that is the epitome of reliable knowledge in the world, a new form of global encyclopedia, jointly carried by libraries and science, used in schools, everyday life, industry and science. It should provide curated, simple language – in all languages – and built-in information literacy and AI literacy. There must be initial discussions about "Greenfield" vs. "Brownfield" designs, i.e. building from scratch or re-using the existing infrastructures. Maybe, a two-pronged approach is needed, one totally new and another built on existing elements such as Wikipedia,

national as well as disciplinary infrastructures.

2. Establish an initial international legal framework or entity and a governance that provides the strongest possible firewall between science and libraries on the one hand and political or economic interests on the other hand.

This will certainly require lawyers but also an initial group of determined persons. This step should not be the greatest obstacle and should only require moderate funds.

3. Establish a code of conduct that forces participants of the program to transcend self-interest.

This could be a second version of a legal framework or entity that will involve social and political scientists but also philosophers and practitioners. It will require the definition and might even involve the establishment of autonomous executive power or jurisdiction. This task is harder and more expensive than building a first version of an international legal framework or entity but also doable.

4. Gather the finest minds and most constructive persons for a work program.

Transparency, inclusion and all other aspects of balance must be included from the beginning, also in the first three steps. But the main challenge is to bring together the right mixture of people from information science, data science, computer science, behavioral and social science as well as law and philosophy to define a work-program in an iterative, agile design. (I am thinking about something like the Macy Conferences on Cybernetics and Systems Theory that laid the foundation for many of the theoretical constructs that are now the basis for AI. However, differently to the Macy Conferences, practitioners should be included from the beginning.) People operating large information infrastructures will be needed to start building functional real-world systems immediately alongside theoretical considerations.

5. Fund raising

Having enough funds available will probably be the most difficult task. The current budget situation is restricted by the rather weak economic situation overall and the competition with other demands of science and society such as security, national sovereignty, climate, ageing and health. Also, the indispensable strong firewall between science as well as libraries and politics as well as economy will limit the motivation to give public or private funding. This budget challenge can be attenuated and managed, though, by an iterative and incremental design of the process.

End

As you might expect, I see Open Metadata initiatives as a good starting point for such a moonshot project. We need a vision and protagonists stepping forward for a truly global open knowledge infrastructure.