Fachbeitrag

Peter Kraker, Maxi Schramm and Christopher Kittel

Discoverability in (a) Crisis

Auffindbarkeit in der Krise

<http://doi.org/10.1515/abitech-2021-0003>

Abstract: The coronavirus pandemic has exposed a host of issues with the current scholarly communication system, one aspect being the discoverability of scientific knowledge. Observing the many shortcomings of discovery workflows in the course of COVID-19 confirms that discoverability itself is in crisis. In this article, we analyze the discoverability crisis and its root causes. We introduce open discovery infrastructure as a promising approach for the development of innovative discovery solutions. We present several efforts based on open discovery infrastructure that have emerged to provide better discovery of coronavirus research and discuss what is still needed to overcome the discoverability crisis.

Keywords: discovery, coronavirus, open science

Zusammenfassung: Die Coronavirus-Pandemie hat eine Reihe von Problemen des aktuellen wissenschaftlichen Kommunikationssystems aufgedeckt, auch im Hinblick auf die Auffindbarkeit wissenschaftlicher Erkenntnisse. Die Beobachtung der vielen Discovery-Probleme in Bezug auf die COVID-19-Forschung bestätigt, dass sich die Auffindbarkeit selbst in einer Krise befindet. In diesem Artikel analysieren wir die Auffindbarkeitskrise und ihre Ursachen. Wir führen die Open Discovery-Infrastruktur als vielversprechenden Ansatz für die Entwicklung innovativer Discovery-Lösungen ein. Wir stellen verschiedene Bemühungen vor, die auf der offenen Open Discovery-Infrastruktur basieren, die zur besseren Auffindbarkeit der Coronavirus-Forschung entwickelt wurde, und diskutieren, was noch zur Überwindung der Auffindbarkeitskrise erforderlich ist.

Schlüsselwörter: Auffindbarkeit, Coronavirus, Open Science

1 Introduction

In the current pandemic, we have seen an explosion of scientific knowledge on the coronavirus. Depending on who is counting, more than 100 000 papers on COVID-19 and Sars-CoV-2 have been published to date.¹

At the same time, many research groups have pivoted to coronavirus research without prior experience or adequate preparation. They were immediately confronted with two discovery challenges: (1) having to identify relevant knowledge from unfamiliar (sub-)disciplines with their own terminology and publication culture, and (2) having to keep up with the rapid growth of data and publications and being able to filter out relevant findings.²

Both challenges pose significant problems for researchers, leading to delays, unnecessarily duplicated work, and findings that are based on questionable prior results.^{3,4} Observing the many shortcomings of discovery workflows in the course of COVID-19 confirms that discoverability itself is in crisis. We currently do not have the tools to get a quick overview of research results and the context information to be able to evaluate them correctly.

1 See e. g., Schmidt, Martin, Benedikt Fecher, Christian Kobsda, et al. "Open Access in Near Time.", April 1, 2020. <https://elephantinelab.org/open-access-near-time/> (17.12.2020).

2 Johnson, Carolyn Y. "Inside a Lab Where Scientists Are Working Urgently to Fight the Coronavirus Outbreak." *Washington Post*. <https://www.washingtonpost.com/science/2020/02/14/inside-lab-where-scientists-are-working-urgently-fight-coronavirus-outbreak/> (12.10.2020).

3 OECD. "Why Open Science Is Critical to Combatting COVID-19." OECD Policy Responses to Coronavirus (COVID-19), 12 May 2020, 3–4. <http://www.oecd.org/coronavirus/policy-responses/why-open-science-is-critical-to-combatting-covid-19-cd6ab2f9/> (17.12.2020).

4 Ioannidis, John P. A. "A Fiasco in the Making? As the Coronavirus Pandemic Takes Hold, We Are Making Decisions without Reliable Data." *STAT* (blog), March 17, 2020. <https://www.statnews.com/2020/03/17/a-fiasco-in-the-making-as-the-coronavirus-pandemic-takes-hold-we-are-making-decisions-without-reliable-data/> (17.12.2020).

2 The Discoverability Crisis

The discoverability crisis, while highlighted by the current pandemic, is by no means a novel situation. Modern science has been growing exponentially since its inception more than 450 years ago.⁵ As such, information overload has always played a role in science and research. But today, with three million publications per year, conservatively estimated,⁶ as well as many new output formats, such as datasets, preprints, and source code, discoverability has become a question of managing not only the magnitude of the output, but also of a plethora resource types.

The challenges with respect to this information overload are reflected in a lack of reuse of scientific knowledge: depending on the discipline, between 7 % and 38 % of research papers are never cited, rising to 63 % of those without a disciplinary classification.⁷ In the case of data sets, the uncitedness even increases to up to 85 %.⁸ We can also see effects for reuse in practice; even in application-oriented disciplines like medicine, only a minority of research results are ever applied in clinical practice, and if so, then with a long delay.⁹

The phenomenon that public knowledge remains hidden has recently been given a name: “Dark Knowledge”.¹⁰ The term, coined by a group around the Berlin ecologist Jonathan Jeschke, describes knowledge that cannot be found and reused. In short: you can’t see the forest for the trees.

Open science is often seen as an antidote to dark knowledge. To a certain extent, this is true: Open Access has dramatically increased the accessibility of scientific articles. But little has changed in terms of discoverability. All of this shows that we are in a veritable discoverability crisis. This crisis has a negative impact on the efficiency,

effectiveness, and quality of science, as it impedes communication within the scientific community and its transfer to practice.

3 Problems with the classic literature search

One of the reasons for the discoverability crisis is a lack of innovation in closed and proprietary search engines. These search engines currently do not have the means to provide a quick overview of current research results and the context information to be able to evaluate them correctly.

Google Scholar, which is probably the most used academic literature search engine,¹¹ is a prime example of this. Below is an image of the top results of Google Scholar for the search term “covid-19”.¹² It can be seen that the results are scarcely contextualized. There is very little information apart from basic metadata and a short snippet showing the search term in context. This means that users have very little information as to whether the articles apply to their research interest. Results from different disciplines are intermixed, but they are not annotated as such. In this case, the only way to get an overview is to sift through the articles by hand.

The unstructured result lists with ten results per page, which are offered by Google Scholar and many of its commercial competitors such as Scopus and Web of Science, work well when the users’ information need is already clearly defined. However, if one wants to get an overview of a research topic, it can take weeks, if not months, before one has identified the most important topics, publication venues and authors. This is too long in most situations, but especially when a public health emergency occurs.

When Google Scholar launched about fifteen years ago, it was a groundbreaking literature search engine. Since then, however, the amount of scientific literature has almost quadrupled.¹³ During this time, the function-

5 Price, Derick J De Solla. *Little Science, Big Science*. Columbia Univ. Press, 1963, 17–42.

6 Johnson, Rob, Anthony Watkinson, and Michael Mabe. “The STM Report: An Overview of Scientific and Scholarly Publishing.” https://www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf, 8.

7 Nicolaisen, Jeppe, and Tove Faber Frandsen. “Zero Impact: A Large-Scale Study of Uncitedness.” *Scientometrics* 119,2 (2019): 1338. doi:10.1007/s11192-019-03064-5.

8 Peters, Isabella, Peter Kraker, Elisabeth Lex, et al. “Research Data Explored : An Extended Analysis of Citations and Altmetrics.” *Scientometrics* 107,2 (2016): 723. doi:10.1007/s11192-016-1887-4.

9 Brownson, Ross C., Matthew W. Kreuter, Barbara A. Arrington, et al. “From the Schools of Public Health.” *Public Health Reports* 121,1 (January 2006): 97. doi:10.1177/003335490612100118.

10 Jeschke, Jonathan M., Sophie Lokatis, Isabelle Bartram, et al. “Knowledge in the Dark: Scientific Challenges and Ways Forward.” *FACETS* 4,1 (June 1, 2019): 423–441. doi:10.1139/facets-2019-0007.

11 Note that to date – to the best of our knowledge – no comprehensive survey of academic search engine usage has been published. Therefore, while the prevalence of Google Scholar is unquestionable, more research is necessary.

12 It is noteworthy that Google Scholar reports over 1.3 million results, which is way beyond any upper bound of any estimate. It becomes a bit more realistic when one selects articles since 2020, which yields about 97 000 results. However, when one then selects articles since 2019, Google Scholar suddenly reports fewer articles (78 500).

13 Bornmann, Lutz, and Rüdiger Mutz. “Growth Rates of Modern Science: A Bibliometric Analysis Based on the Number of Publica-

Google Scholar

covid-19

Articles

About 1.330.000 results (0,07 sec)

Any time

Since 2020

Since 2019

Since 2016

Custom range...

Sort by relevance

Sort by date

☒ include patents

☒ include citations

☐ Create alert

[HTML] Covid-19—navigating the uncharted [HTML] nejm.org

AS Fauci, HC Lane, RR Redfield - 2020 - Mass Medical Soc

Covid-19—Navigating the Uncharted Fauci, Lane, and Redfield comment on the early clinical features and epidemiology of cases reported in Wuhan, China, along with current mortality data, noting that the outbreak is a stark reminder of the need for constant ...

☆ Cited by 766 Related articles All 35 versions

[HTML] The COVID-19 epidemic [HTML] nih.gov

TP Velavan, CG Meyer - Tropical medicine & international health, 2020 - ncbi.nlm.nih.gov

The current outbreak of the novel coronavirus SARS-CoV-2 (coronavirus disease 2019; previously 2019-nCoV), epi-centred in Hubei Province of the People's Republic of China, has spread to many other countries. On 30. January 2020, the WHO Emergency Committee ...

☆ Cited by 808 Related articles All 12 versions

[HTML] COVID-19: immunopathology and its implications for therapy [HTML] nature.com

X Cao - Nature reviews immunology, 2020 - nature.com

Severe coronavirus disease 2019 (COVID-19) is characterized by pneumonia, lymphopenia, exhausted lymphocytes and a cytokine storm. Significant antibody production is observed; however, whether this is protective or pathogenic remains to be determined. Defining the ...

☆ Cited by 459 Related articles All 10 versions

[HTML] Remdesivir for the treatment of Covid-19—preliminary report [HTML] nih.gov

JH Beigel, KM Tomashek, LE Dodd... - The New England ..., 2020 - ncbi.nlm.nih.gov

Background Although several therapeutic agents have been evaluated for the treatment of coronavirus disease 2019 (Covid-19), none have yet been shown to be efficacious. Methods We conducted a double-blind, randomized, placebo-controlled trial of intravenous ...

☆ Cited by 1312 Related articles All 6 versions

Fig. 1: Top search results for the search term “covid-19” on 11 December 2020

ality and interface of Google Scholar have only changed slightly. Between 2010 and now, search and filter functions as well as the presentation of the results stayed almost unchanged.¹⁴ In contrast, Google web search results have been contextualized more and more in recent years with elements such as images, maps, related searches and entries from the knowledge graph. Google has not invested enough in Google Scholar to keep up with the growth of scientific knowledge. As a result, Google Scholar is of limited use these days for discovery use cases that go beyond known item search.

This lack of innovation would not be a problem if other tools could build on the Google Scholar Index. However, this is not possible as the index may not be reused, following the motto: “Use permitted, re-use prohibited”. Innovators in this market first have to build up their own index – a tedious undertaking that is not made easier by Google’s numerous special agreements with publishers. These agreements give Google the ability to index full texts

of articles that are otherwise behind a paywall. Other commercial providers such as Elsevier (Scopus) and Clarivate (Web of Science) follow the same strategy when it comes to reuse of their indices. Their business model may differ – Google monetizes user data for personalized advertising, whereas Elsevier and Clarivate mainly collect license fees – but they are still built around exclusive ownership of the index.

4 Open discovery infrastructure

As a result of the huge barriers in the form of proprietary indices raised by commercial services, the market for academic discovery systems has been moving very slowly. For a long time, it was divided between a handful of proprietary systems such as Web of Science, Scopus, Google Scholar, and Microsoft Academic Search. However, in the shadows of these giants, an alternative, open discovery infrastructure has been created. It builds on a network of tens of thousands of libraries, archives, repositories, and aggregators that offer their (meta-)data via an open data interface such as OAI-PMH.¹⁵

tions and Cited References.” *Journal of the Association for Information Science and Technology* 66, no. 11 (November 2015): 2217–2221. doi:10.1002/asi.23329.

¹⁴ Kraker, Peter, Maxi Schramm, and Christopher Kittel. “Open Knowledge Maps: Visuelle Literatursuche basierend auf den Prinzipien von Open Science.” *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare* 72,2 (October 15, 2019): 464–467. doi:10.31263/voebm.v72i2.3202.

¹⁵ Kraker/Schramm/Kittel 2019: 473.

Fig. 2: Open discovery infrastructure

These metadata are then harvested by meta-aggregators, which in turn also offer open data interfaces, including Open APIs, OAI-PMH and full data dumps. These data interfaces are used by value-added services, such as visual discovery systems (e. g., Open Knowledge Maps, Scholia), text and data mining applications (such as ContentMine), services that enrich this content further (e. g., Unpaywall), and open researcher profiles (e. g., ORCID).

Innovation thrives in an open discovery infrastructure. As can be seen in Figure 2, the entities in the infrastructure can build on each other and thus set in motion a positive, self-reinforcing cycle. Open infrastructure is therefore the strongest driver for innovation in discovery. At last, many of the technologies that are common to users in other areas (e. g., in the discovery of general knowledge), such as semantic search, visual discovery and (social) recommender systems, are also becoming available for the stakeholders of scientific knowledge. It should be noted that many of the technologies underpinning these systems such as knowledge domain visualization, bibliometrics and text mining were already around as early as the 1960s and 1970s.¹⁶ But

it was not until the arrival of open discovery infrastructure that these systems became more widely available.

In addition, open infrastructures have many advantages beyond innovation that are relevant to libraries. Open infrastructures follow open standards, and through the use of open licenses for software, data and content, they facilitate migration between systems and thus avoid the lock-in effects of closed offerings. Open infrastructures are therefore community-owned, and often also community-driven, meaning that they allow for community participation in their governance, enabling direct involvement in the decision-making process around the infrastructures.¹⁷

5 Open discovery systems for COVID-19

In the coronavirus pandemic, it quickly became apparent that tailored approaches are needed to stay on top of the

¹⁶ Börner, Katy, Chaomei Chen, and Kevin Boyack. "Visualizing Knowledge Domains." *Annual Review of Information Science & Technology* 37 (2003): 182–195; Hobbs, Jerry R., Donald E. Walker, and Robert A. Amsler. "Natural Language Access to Structured Text." In *Proceedings of the 9th Conference on Computational Linguis-*

tics, Volume 1, 127–132. COLING '82. CZE: Academia Praha, 1982. doi:10.3115/991813.991833.

¹⁷ Mayer, Katja, Katharina Rieck, Stefan Reichmann, et al. "Empfehlungen für eine nationale Open Science Strategie in Österreich / Recommendations for a National Open Science Strategy in Austria," October 21, 2020, 8. doi:10.5281/zenodo.4109242.

Fig. 3: Interface of LitCovid

literature, both due to the unprecedented growth of knowledge and the immediacy of the issue at hand. In the case of COVID-19, the discoverability of outputs has a direct impact on the lives of people around the world. Below, we will discuss several efforts of the open discovery infrastructure that specifically address coronavirus research. Note that we will focus on human discovery rather than machine discovery. We will center on the features that go beyond a simple list-based search that traditional search engines such as Google Scholar provide.

5.1 LitCovid

LitCovid¹⁸ is a service that builds on PubMed and – at the time of writing – offers discovery covering almost 60,000 articles on COVID-19. With a specific query, articles are retrieved from PubMed on a daily basis. In a next step, articles are reviewed, annotated and categorized, using both automated machine learning/text classification and human review.¹⁹

LitCovid offers a frontend to this corpus, which includes two visualizations: a timeline of the number of

¹⁸ Chen, Qingyu, Alexis Allot, and Zhiyong Lu. “Keep up with the Latest Coronavirus Research.” *Nature* 579, no. 7798 (March 10, 2020): 193–193. doi:10.1038/d41586-020-00694-1.

¹⁹ “LitCovid. Frequently Asked Questions.” <https://www.ncbi.nlm.nih.gov/research/coronavirus/faq> (16.10.2020).

articles published and a geographic visualization of the countries mentioned in abstracts. Users can navigate the corpus using these visualizations, they can browse by category, or they can search the corpus using a subset of the PubMed query capabilities. Search results can again be filtered by date and country, but also by chemical, and journal. In addition, metadata and articles can be downloaded in bulk.

LitCovid is a solid discovery application that offers better filtering, better annotation and a better overview of geographic distribution and time. However, it only allows for sorting articles by date; it should therefore work best for researchers who already have an overview of the literature in a specific topic and/or country and would like to stay on top of the latest articles. The reliance on PubMed is both an advantage and a disadvantage: on the one hand, LitCovid benefits from the high overall metadata quality in PubMed. On the other hand, articles that are not indexed in PubMed will also not be displayed in LitCovid. This includes articles that are not from biomedicine, including epidemiology, and it also relates to document types that are only very selectively indexed in PubMed, such as pre-prints and non-English language publications.

5.2 SemanticScholar

SemanticScholar offers a popular open research dataset called CORD-19,²⁰ which – at the time of writing – includes more than 280 000 articles about COVID-19 and related coronavirus research in a wider sense (the dataset also includes literature on SARS and MERS). CORD-19 is more suited to machine learning applications than human consumption, but SemanticScholar also offers several user interfaces for discovery on top of this data. These include:

- Recent Research: a list-based interface to all research on COVID-19 in SemanticScholar (around 100 000 publications), which can be filtered according to discipline, publication date, open access status, publication type, author and publication venue. The list can be sorted by recency, but also by relevance, citation count, and most influential papers. Influential papers are those that have been influential on other research. A paper's ranking is determined by the number of papers it has been highly influential on. "Influential citations are determined utilizing a machine-learning model analyzing a number of factors including

the number of citations to a publication, and the surrounding context for each."²¹

- Adaptive Research Feed: a personalized feed of research articles, which lets you evaluate articles with thumbs up ("more like this") and thumbs down ("fewer like this"). Based on an AI model,²² SemanticScholar then creates recommendations for articles, which you can in turn evaluate again to provide additional information as to which articles you would like to see.
- SciSight: a set of visual tools to explore CORD-19, including bar charts showing the distribution of relevant metadata items, a network visualization for authors and a choropleth showing the relations between proteins, genes, and cells as well as diseases and chemicals.

With this offering SemanticScholar goes in many ways beyond the scope and capabilities of LitCovid. SemanticScholar not only covers biomedical research, but also allows for cross-domain discovery. With the Adaptive Research Feed, SemanticScholar also offers a personalized application. However, SemanticScholar lacks LitCovid's more fine-grained categorization into themes as well as the geographical information. Furthermore, SemanticScholar seems to be a fully automated approach, whereas LitCovid already has a level of human review. In addition, some of the more advanced SemanticScholar tools are still stamped with beta, and for some, such as the Adaptive Research Feed, not a lot of information is available. SemanticScholar therefore seems to be useful for users looking for a cross-disciplinary perspective and/or research outside of biomedical research. SemanticScholar also caters to those who want to get more information on the relevance of research outputs, and who are looking to try advanced discovery tools, even when they are still experimental.

5.3 OpenAIRE COVID-19 Open Research Gateway

The **OpenAIRE COVID-19 Open Research Gateway** offers discovery covering – at the time of writing – more than

²⁰ Wang, Lucy Lu, Kyle Lo, Yoganand Chandrasekhar, et al. "CORD-19: The COVID-19 Open Research Dataset." ArXiv:2004.10706 [Cs], July 10, 2020. <http://arxiv.org/abs/2004.10706> (17.12.2020).

²¹ "Frequently Asked Questions | Semantic Scholar." <https://www.semanticscholar.org/faq#influential-citations> (11.12.2020). For more information, see Valenzuela, M., Vu A. Ha, and O. Etzioni. "Identifying Meaningful Citations." In *AAAI Workshop: Scholarly Big Data*, 2015.

²² "Frequently Asked Questions | Semantic Scholar." <https://www.semanticscholar.org/faq#research-feeds-signaling> (16.10.2020).

168 000 coronavirus-related research outputs. Similar to CORD-19, the data covers not only COVID-19, but also SARS and MERS. The gateway is based on the OpenAIRE Research Graph.²³ Research outputs in the graph are included based on an elaborate set of criteria that differ depending on the original source of the data (see “Sources and methodology”²⁴): for dedicated COVID-19 sources, all outputs are taken; for all other sources, a specific query is used to determine coronavirus papers. In contrast to LitCovid, OpenAIRE also mines the full text of papers to find relevant candidates. Additionally, OpenAIRE has created a Coronavirus Disease Research Community on Zenodo.²⁵ Here, researchers can suggest coronavirus-related research to be included in the community and by extension also in the Open Research Gateway. The Zenodo community and the gateway are both curated by a dedicated team.

The OpenAIRE research graph can be downloaded and accessed in several ways;²⁶ the subset used for the COVID-19 Open Research Gateway is available via data dumps from Zenodo.²⁷

The OpenAIRE gateway, while having a similar scope to SemanticScholar in terms of topical coverage, is especially interesting as it offers search across other output types than publications including research data and research software, and it also enables researchers to find related projects. The OpenAIRE gateway does not provide distribution in disciplines or topics, but offers geographical filters, and filters for funder, Zenodo community and language. Sorting is restricted to relevance and recency. With the research communities, OpenAIRE also adds a collaborative element to the discovery of research in this area. The OpenAIRE research gateway is therefore best suited for users who would like to do a broad search across disciplines, research outputs and related entities. It is also useful for researchers who would like to highlight relevant research outputs that are not included in the corpus via the related Zenodo community.

5.4 CoVis

CoVis²⁸ is a visual discovery tool for COVID-19 research. It has been developed in cooperation between two research infrastructures that focus on visualization and knowledge synthesis, ReFigure and Open Knowledge Maps. ReFigure enables users to compile figures from different scientific papers into a single visual dashboard. These ReFigures might be a collection of results that are tied together by a single element or may be created to address a specific question, such as whether a certain drug is effective. Open Knowledge Maps is the world’s largest visual search engine for research, enabling users to create knowledge maps in any discipline based on more than 250 million research outputs. Knowledge maps provide an overview of a topic by showing the main areas at a glance, and resources related to each area annotated with keywords, comments and tags.

With CoVis, we address the challenge of identifying seminal coronavirus research among the rapidly increasing body of knowledge around coronavirus research. In the EU-funded project, a team of experts compiles the most reliable resources (articles, preprints, reviews, and datasets) into an open database. Resources are included depending on their impact, or potential impact, for moving the field forward. When key findings are addressed and substantiated by multiple research sources, the curation team creates a dedicated ReFigure. The resulting data is fed into a knowledge map, providing a visual overview of the collected research output. The collection is not meant to be exhaustive, but to offer a single reference point for definitive research in key areas of biomedical research.

CoVis is an open infrastructure following the principles of open science. In contrast to the aforementioned tools, not only the data, but the whole infrastructure can be reused. Content on CoVis is licensed under CC BY 4.0. The CoVis database is made available under CC0. Our software is open source and hosted on GitHub under the MIT license.

CoVis compiles significantly fewer resources than the other tools (at the time of writing, it included 108 resources and collections), but these resources are vetted by experts, further contextualized with keywords and comments, and can be accessed via a unified knowledge map. CoVis is therefore especially useful for biomedical researchers who

²³ Manghi, Paolo, Claudio Atzori, Alessia Bardi, et al. “OpenAIRE Research Graph: Dumps for research communities and initiatives.” 2020. <https://pub.uni-bielefeld.de/record/2945732> (17.12.2020).

²⁴ “Sources and methodology.” <https://covid-19.openaire.eu/content> (11.12.2020)

²⁵ Zenodo is the catch-all repository of the EU: <https://zenodo.org>.

²⁶ See <http://api.openaire.eu/>.

²⁷ Bardi, Alessia, Iryna Kuchma, Gina Pavone, et al. “OpenAIRE Covid-19 Publications, Datasets, Software and Projects Metadata.” *Zenodo*, August 12, 2020. doi:10.5281/zenodo.3980491.

²⁸ Kraker, Peter, Girija Goyal, Maxi Schramm, et al. “CoVis: A New Tool to Discover Reliable COVID-19 Research.” *eLife* (blog). eLife Sciences Publications Limited, July 23, 2020. <https://elifesciences.org/labs/6c623581/covis-a-new-tool-to-discover-reliable-covid-19-research> (17.12.2020).

Fig. 4: Screenshot of the CoVis knowledge map

want to quickly get into coronavirus research or want to stay on top of seminal outputs in the area. It is also interesting for stakeholders outside of academia who want to identify seminal research in a contextualized format. Furthermore, CoVis is a relevant tool for researchers who are looking to nominate resources for inclusion in the collection or the opportunity to become part of the curation team.

6 What is needed to overcome the discoverability crisis

As we have seen above, open discovery infrastructure provides many new and unique tools, which go far beyond the functionality of traditional search engines. This is an important first step to overcoming the discoverability crisis. However, to provide adequate discovery for research and all its stakeholders in societies around the world, further effort is needed. Below, we have listed four different measures which we believe are necessary to overcome the discoverability crisis, complete with examples of initiatives that work in that direction.

6.1 Better alignment of tools with use cases

One of the problems of the current scholarly infrastructure is that it is not always created in a user-driven manner. Many tools are designed from the systems' rather than the users' perspective. While experimentation from a purely technical perspective can lead to new and unexpected solutions, more often than not it produces systems that do not cater to the needs of the research community.

One example of this are the tools for dataset discovery, which have often been adapted from literature discovery without addressing the specificities of datasets in comparison to publications. This is why we have founded a GO FAIR implementation network together with many infrastructures relevant to open and FAIR research data, including CESSDA, DataCite, EUDAT and OpenAIRE. In this network, which has now 27 members, both the users' and the systems' perspective are taken into account. Based on a broad stocktaking of more than 100 use cases and a detailed structuring of the open infrastructure for data discovery, we will then implement user interfaces and user-facing services for cross-disciplinary data discovery.

6.2 Support underserved communities

Historically, discovery tools have not been equitable and inclusive. Both Scopus and Web of Science have a strong focus on English-speaking articles from Western journals.²⁹ As a result, many countries, languages and disciplines have been neglected and are now suffering from larger discoverability problems. Traditionally, these communities are also not involved in the development of discovery tools and infrastructures.

One example of this are the social sciences and humanities (SSH). The SSH are characterized by a high bibliodiversity, in types of outputs, number of languages and wealth of publishers and publication venues. This diversity is not catered to by many of the current discovery solutions. As a result, the use and reuse of SSH research is suboptimal.³⁰ In the H2020 project TRIPLE led by Hum-Num, we attempt to address this challenge by offering a multilingual discovery solution that caters to the specific needs of these disciplines. The core of the platform is a search engine based on ISIDORE, which is extended by innovative services for visual discovery, annotation, recommendation and trust building.³¹

6.3 Combination of AI and collaborative approaches (augmented intelligence)

Currently, innovations in discovery are overwhelmingly driven by automated approaches based on machine learning and artificial intelligence (AI). Discovery is a process that most users are currently tackling on their own – and as a result the same process is repeated over and over again. We therefore see huge potential for collaborative approaches, where users share the results of their discoveries and build on each other's knowledge.

In our opinion, systems that combine AI with human intervention, known as augmented intelligence systems, would be best suited. While AI is needed to deal with the sheer size of the output, the ultimate call needs to be with humans. After all, when structuring scientific research, researchers from different disciplines or even the same

discipline may have different opinions as to how the field is shaped.

At Open Knowledge Maps, we want to create a large-scale system for collective knowledge mapping where different individuals and communities come together to map out their fields aided by computational approaches. The maps and the underlying knowledge structures will subsequently be openly shared via user interfaces and data interfaces so that they can be reused by others.

Here we see libraries and librarians as important partners. Such a collaborative system cannot function without the experts in knowledge curating, structuring and management. Together with other stakeholders from science and society, we want to create a system that enables discoverability of scientific knowledge and makes it (re-) usable for everyone.

6.4 Sustainability of open infrastructure

Recent studies have shown that open infrastructure is not sustainably funded, and many systems could not survive more than six months without continued grant funding.³² Overall, the funding options for non-profit organizations and open-source projects are severely limited. There is little money available, especially for front end services.

This effectively hands control over the way in which researchers and other stakeholder groups interact with science to commercial companies. Commercial companies, however, are primarily tasked with maximizing shareholder value and therefore do not always act in the interests of the scientific community and society. Usage data also remain with these for-profits, which can lead to problems if this data is passed on or sold. In addition, proprietary solutions often lead to lock-in effects, meaning that switching to a different system would be associated with considerable costs. As a result, annual license fees can often be increased well above inflation. This is a situation that should be very familiar to those who advocate for switching from closed access to open access for publications.

Leaving academic discovery to proprietary and closed solutions also restricts innovation, as there are fewer in-

²⁹ Vera-Baceta, Miguel-Angel, Michael Thelwall, et al. "Web of Science and Scopus Language Coverage." *Scientometrics* 121,3 (December 1, 2019): 1806–1810. doi:10.1007/s11192-019-03264-z; Mongeon, Philippe, and Adèle Paul-Hus. "The Journal Coverage of Web of Science and Scopus: A Comparative Analysis." *Scientometrics* 106, no. 1 (October 19, 2015): 218–224. doi:10.1007/s11192-015-1765-5.

³⁰ See e.g., Nicolaisen and Frandsen, "Zero impact", 1231.

³¹ See <https://gotriple.eu> (28.1.2021).

³² Ficarra, Victoria, Mattia Fosci, Andrea Chiarelli, et al. "Scoping the Open Science Infrastructure Landscape in Europe." *Zenodo*, October 30, 2020, 47–49. doi:10.5281/zenodo.4153809; Thaney, Kaitlin. "The Open Scholarship Ecosystem Faces Collapse; It's Also Our Best Hope for a More Resilient Future." *Impact of Social Sciences* (blog), June 19, 2020. <https://blogs.lse.ac.uk/impactofsocialsciences/2020/06/19/the-open-scholarship-ecosystem-faces-collapse-its-also-our-best-hope-for-a-more-resilient-future/> (17.12.2020).

centives to improve one's own offer in a market with only a handful of providers and high entry barriers. Therefore, it is important to sustain open infrastructure, not only as a way to avoid lock-in effects and exploding costs for research organizations, but also as a driver of innovation. Consortial funding models have emerged as a particularly viable funding model in this context. Infrastructures such as the Open Library of Humanities (OLH) and the Directory of Open Access Journals (DOAJ) have successfully employed such models. In this crowdfunding-type approach, organizations become supporting members and contribute an annual membership fee. In return, supporting members are usually involved in the infrastructure's governance.

Open Knowledge Maps also employs this model: in our case, we invite our supporting members to co-create the platform with us. The Board of Supporters provides input on the technical roadmap and has one third of the vote on which features and sources are implemented on Open Knowledge Maps. This deepens the strong collaboration that we have had with libraries from the start: librarians are amongst our most prominent advisors and we have developed innovative open-source projects together with libraries. With BASE as the main data source, Open Knowledge Maps builds on library infrastructures and increases the visibility of the content contained therein, e.g., from university repositories. This includes many sources and document types that are not indexed by commercial products, making the service ideal for specialized collections that libraries maintain in large numbers.

7 Summary and conclusion

The coronavirus pandemic has put a spotlight on the discoverability crisis once and for all. It has become abundantly clear that in order to address the many challenges of our globalized world, including, among others, infectious diseases, climate change, and global inequality, a multitude of stakeholders has to work together. For that to happen, we must be able to discover relevant research and to build on each others' knowledge. This is not possible with closed and proprietary discovery infrastructures that have not kept up with the enormous growth of scientific knowledge.

Instead, an alternative open discovery infrastructure has emerged, which has reuse of content, data, and software as its core principle. This infrastructure enables continuous innovation, as it removes many of the barriers towards innovation. The open infrastructure has enabled a host of new discovery tools, many of which have been adapted to address the specific challenges of the coronavirus pandemic.

As a next step, it is important to work towards making this infrastructure more inclusive, participatory and collaborative. And above all, we need to guarantee sustainability of the open discovery infrastructure so that we can rely on it for both today's and tomorrow's challenges.

Acknowledgments: The research presented in this paper is funded in part by the EOSC Secretariat as part of the COVID-19 Fast-Track Funding and in part by the European Commission as part of the H2020 project TRIPLE (grant agreement no. 863 420). The authors want to thank Dr. Girija Goyal and Dr. James Akin for inspiring discussions around many aspects of the discoverability crisis and for providing insights into discovery issues related to coronavirus research.

Authors

Peter Kraker

Open Knowledge Maps
Gumpendorfer Straße 16/1/8
1060 Vienna
pkraaker@openknowledgemaps.org
orcid.org/0000-0002-5238-4195

Maxi Schramm

Open Knowledge Maps
Gumpendorfer Straße 16/1/8
1060 Vienna
maxi@openknowledgemaps.org
orcid.org/0000-0003-1408-5394

Christopher Kittel

Open Knowledge Maps
Gumpendorfer Straße 16/1/8
1060 Vienna
christopher.kittel@openknowledgemaps.org
orcid.org/0000-0002-1894-5040