Growing New Scholarly Communication Infrastructures for Sharing, Reusing, and Synthesizing Knowledge

Joel Chan
Wayne Lutters
University of Maryland College Park, College of Information Studies
USA

Jodi Schneider
University of Illinois Urbana-Champaign, School of Information Sciences
USA

Karola Kirsanow
Sílvia Bessa
Protocol Labs Research
USA

Jonny Saunders
University of Oregon, Institute of Neuroscience
USA

ABSTRACT

Sharing, reuse, and synthesis of knowledge is central to the research process. These core functions are in theory served by the system of monographs, abstracts, and papers in journals and proceedings, with citation indices and search databases that comprise the core of our formal scholarly communication infrastructure; yet, converging lines of empirical and anecdotal evidence suggest that this system does not adequately act as infrastructure for synthesis. Emerging developments in new institutions for science, along with new technical infrastructures and tooling for decentralized knowledge work, offer new opportunities to prototype new technical infrastructures on top of a different installed base than the publish or perish, neoliberal academy. This workshop aims to integrate these developments and communities with CSCW's deep roots in knowledge infrastructures and collaborative and distributed sensemaking, with new developments in science institutions and tooling, to stimulate and accelerate progress towards prototyping new scholarly communication infrastructures that are actually optimized for sharing, reusing, and synthesizing knowledge.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing systems and tools**; **Empirical studies in collaborative and social computing**; • **Applied computing** → **Digital libraries and archives**; • **Information systems** → **Document representation**.

KEYWORDS

scholarly communication, synthesis, knowledge organization, infrastructure, sensemaking

CSCW'22 Companion, November 8–22, 2022, Virtual Event, Taiwan

ACM ISBN 978-1-4503-9190-0/22/11.
https://doi.org/10.1145/3500868.3559398

ACM Reference Format:
Joel Chan, Wayne Lutters, Jodi Schneider, Karola Kirsanow, Sílvia Bessa, and Jonny Saunders. 2022. Growing New Scholarly Communication Infrastructures for Sharing, Reusing, and Synthesizing Knowledge. In *Computer Supported Cooperative Work and Social Computing (CSCW'22 Companion), November 8–22, 2022, Virtual Event, Taiwan.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3500868.3559398

1 INTRODUCTION

How do researchers, scholars, and scientists share, reuse, and synthesize knowledge? Alongside informal channels of communication such as personal communications and interactions, scholars today rely heavily on a central infrastructure of scholarly publishing: a system of monographs, abstracts, and papers in journals and proceedings, overlaid with citation indices and search databases. Together, this comprises the formal scholarly record. Through this scholarly communication infrastructure, scholars have access to a vast and rapidly growing literature, across disciplines — on the order of hundreds of millions of documents accessible through just a few clicks and keystrokes from a single search engine [9] — from which they can draw to construct new knowledge.

How well does this infrastructure serve scholarly needs of sharing, reusing, and synthesizing knowledge? Consider a researcher who wants to understand what interventions might be most promising for mitigating online harassment. To synthesize an understanding of this complex interdisciplinary problem to advance the state of the art, she needs to work through detailed answers to a range of questions: which theories have the most empirical support in this particular setting? Are there conflicting theoretical predictions that might signal fruitful areas of inquiry? What key phenomena should be kept in mind when designing an intervention (e.g., perceptions of human vs. automated action, procedural considerations, noise in judgments of wrongdoing, scale considerations for spread of harm)? What intervention patterns are both theoretically predicted to be effective in this setting, *and* lack direct evidence of efficacy? The answers to these questions lie at granular levels of theoretical and empirical claims and their interrelationships, not publications. For example, *"viewers in a Twitch chat engaged in less bad behaviors after a user was banned by a moderator for bad behavior"*, and *"banning bad actors from a subreddit in 2015 was somewhat effective at mitigating spread of hate speech on other subreddits"* are claims that

interrelate in complex ways, both supporting other claims/theories that are in tension with each other. This level of granularity is crucial not just for finding relevant claims to inform the synthesis, but also for constructing more complex arguments and theories, by connecting statements in logical and discursive relationships. Our researcher also needs to work through a range of contextual details. For example, to judge which studies, findings, or theories are most applicable to her setting, she needs to know key methodological details including the comparability of different studies' interventions, settings, populations, and outcome measures. She might need to reason over the fact that two studies that concluded limited efficacy of bans had ban interventions that were quite short, on a forum with no identity verification. To judge what findings have been established with sufficient certainty and where the frontier might be, she would need to know, which findings came from which measures (e.g., self-report, behavioral measures), and the extent to which findings have been replicated across authors from different labs, and across a variety of settings (e.g., year, platform, scale). Where appropriate, she would need integrated access to methods, protocols, and data.

How might this researcher retrieve, reuse, and synthesize this information from our current scholarly communication infrastructure? Search engines operate over documents and their metadata (authorship, date, publication outlet), not claims and their context and relationships; scholarly documents are primarily archived as unstructured text, often in PDFs. If she is lucky, she might come across a published synthesis that is both on topic, with sufficient coverage, and up to date; that would likely be difficult for an interdisciplinary topic such as hers! Perhaps she might also be able to piece together leads on papers and authors and verbal statements of key ideas from knowledgeable colleagues. Failing all this, she would need to do the laborious work of citation tracing, manually checking references because citation databases typically do not surface information about why and in what way scholarly works cite each other, and collecting documents through keyword searches and then screening first the titles and abstracts and then their full text, and then constructing her own database of claims and data, often in a bespoke system of notes, annotations, and spreadsheets.

This example illustrates how **the current scholarly communication infrastructure is not acting as infrastructure for synthesis**. The term *infrastructure* is meant to evoke reliability, sustainability, and "smooth functioning" that enables people to focus on the task at hand instead of fussing over preparatory overhead [5, 20]. Instead of this smooth functioning, as the preceding example illustrates, researchers make do with an assemblage of workarounds and hacks to find and synthesize relevant literature. While these hacks often work well enough for the task at hand, they are rarely transferred in systematic ways across projects and people, violating the dimensions of "reach or scope" and "embodiment of standards" of infrastructure [20]. These are also not "one-time" costs: scientific problems require many such queries, and likely spawn new queries as projects evolve and more is learned. It is unsurprising, then, that synthesis of prior literature is often subpar [2, 8, 13, 14].Similarly, systematic reviews are increasingly struggling to keep up with the pace of knowledge, with many becoming outdated soon after they are published [17], but are rarely updated [6].

In this workshop, we ask: **how might we design scholarly communication infrastructures that are actually optimized for sharing, reusing, and synthesizing knowledge?** And where might these issues intersect with CSCW, to stimulate fresh sociotechnical progress and theoretical development? This set of questions has deep roots within CSCW, connecting with long-standing threads of work on of work has continued to the present, studying the collaborative, practical, and social work of constructing and maintaining scientific knowledge infrastructures [15, 16, 21, 23], distributed cognition and organizational memory [1], systems for peer production of knowledge, such as Wikipedia [12], complex sensemaking in collaborative and distributed settings [3, 7, 22], and social computing and human-machine systems for synthesis [10, 19]. In the cognate field of library and information science, too, there are decades of standards and platform-level work on new scholarly communication infrastructures that incorporate more sophistical models of semantics and scientific argumentation to support synthesis [4, 11, 18].

We aim to stimulate fresh progress by bringing these deep roots into conversation with emerging trends in science and infrastructure reform and innovation, such as new institutions for science outside the academy that are deliberately structured to pursue different incentive structures, such as Invisible College, the Arcadia Institute, and LabDAO; new coordination structures and technologies, including open-source community publishing platforms like PubPub, and patterns from emerging technology around Decentralized Science and new decentralized Web infrastructures, such as IPFS; and new experiments in science funding, from crowdfunding to larger investments in creating new kinds of scientific institutes.

The potential intersection of these communities and technological developments offers new opportunities to grow new technical infrastructures on top of a different installed base than the publish or perish, neoliberal academy. We aim for these new infrastructural patterns to synergize and combine with long-running advances and innovations in open science (sharing of code, data, and protocols) and open access and preprints. We hope this new wave of sociotechnical innovation will catalyze bottom-up evolution and growth towards new infrastructures for sharing, reusing, and synthesizing knowledge.

Some examples of concrete topics include:

- "living" syntheses that respond appropriately to retracted, outdated, or inconsistent findings as knowledge evolves;
- integrating formality and machine readability — e.g., ontologies, semantic markup, or argumentation patterns of relationships between lines of work — while respecting the limits of machine intelligence, maintaining diversity of perspectives and epistemologies, and balancing the need for tacit and contextual knowledge required for synthesis;
- integrating novel forms of scholarly authorship and knowledge sharing into everyday individual and collaborative scholarly practices and workflows, and novel assemblages, bricolages, and repurposing of existing tooling to enable these integrations;
- the role of AI and machine learning systems in assisting — vs. automating — the construction of new synthesis-optimized knowledge sharing infrastructures;

- the role of decentralized peer-to-peer technologies, and new developments in hypertext and personal knowledge graphs, in lowering barriers to community-owned and maintained syntheses of scientific knowledge;
- and applications and design patterns from crowdsourcing and social computing for new forms of scholarly communication infrastructure.

2 WORKSHOP GOALS, STRUCTURE, AND PARTICIPANTS

The purpose of this proposed two-day workshop is to catalyze new collaborations, prototypes, and research ideas at the intersection of research, tooling, and practice communities. Thus, the workshop will be structured around **synchronous and asynchronous work sessions in working groups**. This will include scheduled time to form and refine proposals, and short lightning talks about concrete proposals for artifacts to create, such as new syntheses/aggregations of resources and research theories and findings, pilot user studies and prototypes with novel tooling, and new designs and lo-fi prototypes. We will also schedule for an extended synchronous session at the end of the workshop to share progress, with ample time for discussion, feedback, and planning of next steps. All sessions will also be recorded and shared with participants to allow for continued participation if time zones conflict. We also plan to organize **a venue for invited publication of the mature results of the collaborations**, approximately 3-6 months after the conclusion of the workshop. We envision the venue being a combination of community-owned open access publications, such as on PubPub, and a co-authored overview article in a more traditional publication such as CSCW or CACM.

We estimate that the overall available time for synchronous work, accounting for time zone differences, will be on the order of 5-6 hours across the two days of the workshop. To maximize the likelihood of this focused time leading to concrete outputs, we will also design **asynchronous structured activities in the lead-up to the workshop**. Prospective participants will submit materials for consideration, appropriate for their background and interests: *researchers* will submit short summaries of their past relevant research, and/or detailed annotated bibliographies; *tool-builders* will submit video demos, and/or links to learn and try out their tools; *practitioners* will submit case studies of their attempts to shift practice (along with problems and opportunities/solutions they have discovered), and/or training materials for practice innovations, and example datasets. These submissions will be reviewed by the organizers according to criteria of relevance for the core workshop themes, as well as balance across communities of research, tooling, and practice. Accepted submissions (announced in early-October) will then be curated into a shared github repository of materials. In the subsequent weeks leading to the workshop, participants will comment on and iterate on the ideas and resources in the repository. The materials in this repository will serve as important context and resources to fuel impactful collaborative work during the workshop. We will also host a **chat platform**, such as Discord or gitter, to facilitate interactions between participants prior to the workshop. Prior to the workshop, the organizing team will propose topical and time-zone-based groupings of participants. Participants will be invited to use these proposals as starting points for forming working groups.

Given the time constraints and goals of producing and sharing work outputs, we expect that the maximum number of participants will be 20-30 (to aim for no more than 5-6 working groups). We aim to recruit participants from personal connections (to organizations like Arcadia and Invisible College), Twitter (where many from the new infrastructural reforms are being discussed), and Discords (where many DeSci efforts are being coordinated).

3 PROPOSED SCHEDULE

Prospective participants will submit proposed materials by late September, and receive notifications by early October. **Pre-workshop activities**, including commenting and iterating on materials, as well as initial formation of working groups, in the github repository and over our hosted chat platform, will begin shortly thereafter, and continue up to the workshop dates.

We plan to limit **synchronous joint sessions** to 3 hours in the morning of US East timezone, to maximize overlap in schedules across Asia, Europe, and the US. This schedule will be adjusted as needed based on the final set of participants. To maximize informal and spontaneous interactions, we plan to host the workshop on gather.town. All sessions will also be recorded and shared with participants to allow for continued participation if time zones conflict. The following is a proposed rough schedule (in US Eastern time).

Day 1:

- 09:30 - 09:40 - Welcome and kickoff
- 09:40 - 10:10 - Working groups finalize proposals
- 10:10 - 11:40 - Work sprint 1
- 11:40 - 12:25 - Working group proposal lightning talks
- Various times (depending on time zone): Work sprint 2

Day 2:

- 09:30 - 12:30 - Working group proposal progress reports and discussion
- 12:40 - 12:50 - Closing call to action, next steps

4 BACKGROUND OF THE ORGANIZERS

Joel Chan is an Assistant Professor in the University of Maryland's College of Information Studies. His research investigates systems that support creative knowledge work, such as scientific discovery and innovative design. His recent work focuses on studies of scientific thinking (including their synthesis practices), and tools for searching and synthesizing scientific literature. His research has received funding from the National Science Foundation, the Office of Naval Research, the Institute for Museum and Library Sciences, Adobe Research, and Protocol Labs.

Wayne Lutters is a professor in the University of Maryland's College of Information Studies. Wayne's research interests are at the nexus of CSCW, social computing, and social informatics. He specializes in field studies of IT-mediated work, from a socio-technical perspective, to better inform the design and evaluation of collaborative systems. Recent projects have focused on the human-side of information infrastructure for distributed science. He has served as a Program Director for Human-Centered Computing at the National Science Foundation. He earned his M.S. and Ph.D. in Information and Computer Science from the University of California, Irvine.

Jodi Schneider is an Associate Professor at the School of Information Sciences, University of Illinois at Urbana-Champaign where she runs the Information Quality Lab. She studies the science of science through the lens of arguments, evidence, and persuasion with a special interest in controversies in science. Her recent work has focused on systematic review automation, semantic publication, and the citation of retracted papers. She has held research positions across the U.S. as well as in Ireland, England, France, and Chile. Her work has been funded by the Alfred P. Sloan Foundation, the European Commission, IMLS, NIH, Science Foundation Ireland, and an NSF CAREER award.

Karola Kirsanow is a Research Program Manager at Protocol Labs, an open-source research, development, and deployment lab creating new internet technologies. There she leads a team that builds research public goods, identifying and supporting high-impact research projects in the distributed systems space and designing experiments to align researchers and research funders. Her previous research background is in human evolutionary biology and palaeogenetics, including work funded by the Leakey Foundation and the European FP7 framework programme.

Sílvia Bessa is a Research Program Manager in the Network Research team at Protocol Labs, where she designs new mechanisms to incentivise and accelerate research to build public goods. She's a strong believer that community-driven research is the best-known way to protect humanity's knowledge from individual interests. Her past research applied computer vision and machine learning to breast cancer imaging, funded by national and European Programs, and in close collaboration with the Portuguese National League Against Cancer and Champalimaud Foundation.

Jonny Saunders is a PhD candidate at the University of Oregon's Institute for Neuroscience. They are a transdisciplinary research worker studying ill-defined categories of complex sounds in a mouse model of phonetics, embedding distributed systems of knowledge sharing in experimental tooling, and applied strategy for information liberation from the history of digital social movements. They search between the seams of technology, labor, and politics for points of leverage to pry apart the systems of hierarchy, extraction, and privatization that structure knowledge work. Their hope is that by organizing with researchers across disciplines that we might be able to contribute our diverse skills towards building liberatory digital infrastructures of communication and collaboration — and realize the role we might play in building a better world beyond the broader digital enclosure movement.

REFERENCES

[1] Mark S. Ackerman, Juri Dachtera, Volkmar Pipek, and Volker Wulf. 2013. Sharing Knowledge and Expertise: The CSCW View of Knowledge Management. *Computer Supported Cooperative Work (CSCW)* 22, 4-6 (Aug. 2013), 531–573. https://doi.org/10.1007/s10606-013-9192-8

[2] Adrienne Alton-Lee. 1998. A Troubleshooter's Checklist for Prospective Authors Derived from Reviewers' Critical Feedback. *Teaching and Teacher Education* 14, 8 (1998), 887–90.

[3] Marcela Borge, Craig H. Ganoe, Shin-I Shih, and John M. Carroll. 2012. Patterns of team processes and breakdowns in information analysis tasks. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (CSCW '12)*. Association for Computing Machinery, New York, NY, USA, 1105–1114. https://doi.org/10.1145/2145204.2145369

[4] Anita de Waard. 2010. From Proteins to Fairytales: Directions in Semantic Publishing. *IEEE Intelligent Systems* (2010). https://ieeexplore.ieee.org/document/5456415

[5] Paul N Edwards, Steven J Jackson, Geoffrey C Bowker, and Cory P Knobel. 2007. *Understanding infrastructure: Dynamics, tensions, and design.* Technical Report.

[6] Ann-Margret Ervin. 2008. Motivating authors to update systematic reviews: practical strategies from a behavioural science perspective. *Paediatric and perinatal epidemiology* 22, 0 1 (Jan. 2008), 33–37. https://doi.org/10.1111/j.1365-3016.2007.00910.x

[7] Kristie Fisher, Scott Counts, and Aniket Kittur. 2012. Distributed Sensemaking: Improving Sensemaking by Leveraging the Efforts of Previous Users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 247–256. https://doi.org/10.1145/2207676.2207711

[8] Padhraig S. Fleming, Jadbinder Seehra, Argy Polychronopoulou, Zbys Fedorowicz, and Nikolaos Pandis. 2013. Cochrane and non-Cochrane systematic reviews in leading orthodontic journals: a quality paradigm? *European Journal of Orthodontics* 35, 2 (April 2013), 244–248. https://doi.org/10.1093/ejo/cjs016 Publisher: Oxford Academic.

[9] Michael Gusenbauer. 2019. Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics* 118, 1 (Jan. 2019), 177–214. https://doi.org/10.1007/s11192-018-2958-5

[10] Nathan Hahn, Joseph Chang, Ji Eun Kim, and Aniket Kittur. 2016. The Knowledge Accelerator: Big Picture Thinking in Small Pieces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 2258–2270. https://doi.org/10.1145/2858036.2858364

[11] Tobias Kuhn and Michel Dumontier. 2017. Genuine semantic publishing. *Data Science* 1, 1-2 (Jan. 2017), 139–154. https://doi.org/10.3233/DS-170010

[12] Guo Li, Haiyi Zhu, Tun Lu, Xianghua Ding, and Ning Gu. 2015. Is It Good to Be Like Wikipedia?: Exploring the Trade-offs of Introducing Collaborative Editing Model to Q&A Sites. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 1080–1091. https://doi.org/10.1145/2675133.2675155

[13] Barbara E. Lovitts. 2007. *Making the Implicit Explicit: Creating Performance Expectations for the Dissertation.* Stylus Publishing, Sterling, Va.

[14] Jonathon McPhetres, Nihan Albayrak-Aydemir, Ana Barbosa Mendes, Elvina C. Chow, Patricio Gonzalez-Marquez, Erin Loukras, Annika Maus, Aoife O'Mahony, Christina Pomareda, Maximilian Primbs, Shalaine Sackman, Conor Smithson, and Kirill Volodko. 2020. *A decade of theory as reflected in Psychological Science (2009-2019).* Technical Report. PsyArXiv. https://doi.org/10.31234/osf.io/hs5nx 00000 type: article.

[15] Dave Randall, Rob Procter, Yuwei Lin, Meik Poschen, Wes Sharrock, and Robert Stevens. 2011. Distributed ontology building as practical work. *International Journal of Human-Computer Studies* 69, 4 (April 2011), 220–233. https://doi.org/10.1016/j.ijhcs.2010.12.011

[16] David Ribes and Charlotte P. Lee. 2010. Sociotechnical Studies of Cyberinfrastructure and e-Research: Current Themes and Future Trajectories. *Computer Supported Cooperative Work (CSCW)* 19, 3 (Aug. 2010), 231–244. https://doi.org/10.1007/s10606-010-9120-0

[17] Kaveh G. Shojania, Margaret Sampson, Mohammed T. Ansari, Jun Ji, Steve Doucette, and David Moher. 2007. How Quickly Do Systematic Reviews Go Out of Date? A Survival Analysis. *Annals of Internal Medicine* 147, 4 (Aug. 2007), 224. https://doi.org/10.7326/0003-4819-147-4-200708210-00179

[18] Simon Buckingham Shum, Enrico Motta, and John Domingue. 2000. ScholOnto: an ontology-based digital library server for research documents and discourse. *International Journal on Digital Libraries* 3, 3 (Oct. 2000), 237–248. https://doi.org/10.1007/s007990000034

[19] Pao Siangliulue, Joel Chan, Steven P. Dow, and Krzysztof Z. Gajos. 2016. IdeaHound: Improving Large-scale Collaborative Ideation with Crowd-Powered Real-time Semantic Modeling. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, Tokyo Japan, 609–624. https://doi.org/10.1145/2984511.2984578

[20] Susan Leigh Star and Karen Ruhleder. 1996. Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information systems research* 7, 1 (1996), 111–134. https://pubsonline.informs.org/doi/abs/10.1287/isre.7.1.111 03070 tex.publisher: INFORMS.

[21] Andrea K. Thomer, Michael Bernard Twidale, and Matthew J. Yoder. 2018. Transforming Taxonomic Interfaces: "Arm?s Length" Cooperative Work and the Maintenance of a Long-lived Classification System. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 173:1–173:23. https://doi.org/10.1145/3274442 00000.

[22] Sijia Xiao, Coye Cheshire, and Amy Bruckman. 2021. Sensemaking and the Chemtrail Conspiracy on the Internet: Insights from Believers and Ex-believers. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–28. https://doi.org/10.1145/3479598

[23] Alyson L. Young and Wayne G. Lutters. 2017. Infrastructuring for Cross-Disciplinary Synthetic Science: Meta-Study Research in Land System Science. *Computer Supported Cooperative Work (CSCW)* 26, 1 (April 2017), 165–203. https://doi.org/10.1007/s10606-017-9267-z