2nd Workshop on Digital Infrastructures for Scholarly Content Objects (DISCO'22)

Wolf-Tilo Balke
Hermann Kroll
balke@ifis.cs.tu-bs.de
kroll@ifis.cs.tu-bs.de
Institute for Information Systems,
TU Braunschweig
Braunschweig, Germany

Yuanxi Fu
Jodi Schneider
fu5@illinois.edu
jodi@illinois.edu
School of Information Sciences,
University of Illinois at
Urbana-Champaign
Champaign, Illinois, USA

Anita de Waard
Research Collaboration Unit, Elsevier
Jericho, Vermont, USA
a.dewaard@elsevier.com

CCS CONCEPTS

• **Information systems → Information retrieval**; • **Applied computing → Digital libraries and archives**; **Publishing**.

KEYWORDS

semantic publishing, robustness, reproducibility, argumentation, narrative, fact checking, knowledge graphs, scholarly publishing

ACM Reference Format:
Wolf-Tilo Balke, Hermann Kroll, Yuanxi Fu, Jodi Schneider, and Anita de Waard. 2022. 2nd Workshop on Digital Infrastructures for Scholarly Content Objects (DISCO'22). In *The ACM/IEEE Joint Conference on Digital Libraries in 2022(JCDL '22), June 20–24, 2022, Cologne,Germany.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3529372.3530943

1 DISCO GOALS, GENESIS, AND EXPECTED AUDIENCE

The goal of the Digital Infrastructures for Scholarly Content Objects (DISCO) workshop is to raise awareness of quality issues, improved discovery, and re-use challenges in digital infrastructures for scholarly content, and to collect potential solutions among an audience of diverse expertise.

The first workshop on Digital Infrastructures for Scholarly Content Objects (DISCO'21)[1] was held in conjunction with the 2021 ACM/IEEE Joint Conference on Digital Libraries as a one-day workshop on September 30th, 2021, [16], online due to the COVID-19 pandemic. The DISCO'21 proceedings[2] were published as volume 2916 within the open access CEUR-WS proceedings platform include 2 keynotes, 3 long papers and 3 short papers.

This year the second DISCO workshop is dedicated to propelling an ongoing dialogue between the computer science, information science, and library science communities necessary for building innovative, value-adding, and sustainable digital infrastructures in digital libraries. We invite academic researchers, librarians, and industrial practitioners to participate and share their knowledge in this forum.

As digital libraries make the dissemination of research publications easier, they also create an information flood severely challenging findability and enable the propagation of invalid or unreliable knowledge. Relevant problems include: retraction and inadvertent citation and reuse of retracted papers [2, 17]; propagation of errors in literature and scientific databases [6, 7]; non-reproducible papers; known domain-specific issues such as cell line contamination [3]; bias in research datasets and publications [4, 10, 19]; systematic reviews that arrive at different conclusions about the same question at the same time [8, 20]. The digital environment facilitates broad interdisciplinary reuse beyond the originating scientific community; thus, marking known problems and tracing the impact on dependent and follow-on works is particularly important (but still under-addressed). Further, context-specific information inside a paper may not be immediately reusable when extracted by automated processes, leading to apparent contradictions [15]. Current mitigating approaches use the underlying reasoning for information retrieval [1, 13], develop new infrastructures analyzing the reasoning [5, 11, 21] or certainty [14] of statements, or use visualization to highlight possible discrepancies [8, 11]. Moreover, new retrieval models based on narrative intelligence try to foster coherence and plausibility of scientific argumentation [9, 12, 18].

2 TOPICS AND OUTCOMES

Topics include:

- Fact checking and knowledge updates for scholarly publishing, scholarly databases, and expert knowledge
- "Living" documents and innovation in publishing
- Semantic publishing, metadata, ontologies
- Scholarly database curation, scholarly knowledge graphs
- Argumentation, identifying and tracing dependencies between papers
- Mining, representing, and exploiting narrative structures in and across papers
- Infrastructure for robustness and reproducibility (e.g., multiverse analyses, data storage and citation, etc.)
- Infrastructure for knowledge and evidence synthesis, systematic review, question answering on expert knowledge
- Annotation and integration of scholarly content

[1] https://infoqualitylab.org/events/disco2021/
[2] http://ceur-ws.org/Vol-2976/

JCDL '22, June 20–24, 2022, Cologne, Germany and online
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9345-4/22/06.
https://doi.org/10.1145/3529372.3530943

- Quality assurance and quality assessment of automatic knowledge mining processes, recovering from retracted, outdated, or inconsistent findings

The lessons learned in these workshops will serve as foundation for a roadmap on digital infrastructure development in digital libraries.

3 ORGANIZING COMMITTEE

Jodi Schneider is Assistant Professor at the School of Information Sciences, University of Illinois at Urbana-Champaign where she runs the Information Quality Lab.

Anita de Waard is VP of Research Data Collaborations at Elsevier, and developing cross-disciplinary frameworks for sharing data and tools to store, share and search experimental outputs.

Wolf-Tilo Balke heads the Institute for Information Systems as a full professor at Technische Universität Braunschweig, and serves as a director of L3S Research Center at Leibniz University Hannover, Germany.

Hermann Kroll is a PhD student at the Institute for Information Systems at Technische Universität Braunschweig, focusing on narrative intelligence.

Yuanxi Fu is a PhD student in Information Sciences at the University of Illinois at Urbana-Champaign focusing on argumentation in science.

REFERENCES

[1] Kevin D Ashley. 2014. Applying argument extraction to improve legal information retrieval. In *Proc. of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*. 1–9. http://ceur-ws.org/Vol-1341/paper3.pdf

[2] Sorana D. Bolboacă, Diana-Victoria Buhai, Maria Aluaș, and Adriana E. Bulboacă. 2019. Post retraction citations among manuscripts reporting a radiology-imaging diagnostic method. *PLOS ONE* 14, 6 (06 2019), 1–14. https://doi.org/10.1371/journal.pone.0217918

[3] Amanda Capes-Davis, George Theodosopoulos, Isobel Atkin, Hans G. Drexler, Arihiro Kohara, Roderick A.F. MacLeod, John R. Masters, Yukio Nakamura, Yvonne A. Reid, Roger R. Reddel, and R. Ian Freshney. 2010. Check your cultures! A list of cross-contaminated or misidentified cell lines. *International Journal of Cancer* 127, 1 (2010), 1–8. https://doi.org/10.1002/ijc.25242

[4] Aled M. Edwards, Ruth Isserlin, Gary D. Bader, Stephen V. Frye, Timothy M. Willson, and Frank H. Yu. 2011. Too many roads not taken. *Nature* 470, 7333 (01 Feb 2011), 163–165. https://doi.org/10.1038/470163a

[5] Yuanxi Fu and Jodi Schneider. 2020. Towards knowledge maintenance in scientific digital libraries with the keystone framework. In *Proc. of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. ACM, 217–226. https://doi.org/10.1145/3383583.3398514

[6] Walter R. Gilks, Benjamin Audit, Daniela De Angelis, Sophia Tsoka, and Christos A. Ouzounis. 2002. Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics* 18, 12 (12 2002), 1641–1649. https://doi.org/10.1093/bioinformatics/18.12.1641

[7] Steven A Greenberg. 2009. How citation distortions create unfounded authority: analysis of a citation network. *BMJ* 339 (2009). https://doi.org/10.1136/bmj.b2680

[8] Tzu-Kun Hsiao, Yuanxi Fu, and Jodi Schneider. 2020. Visualizing evidence-based disagreement over time: the landscape of a public health controversy 2002-2014. In *Proceedings of the Association for Information Science and Technology*, Vol. 57. e315. https://doi.org/10.1002/pra2.315

[9] Hermann Kroll, Denis Nagel, and Wolf-Tilo Balke. 2020. Modeling Narrative Structures in Logical Overlays on Top of Knowledge Repositories. In *Conceptual Modeling - 39th International Conference, ER 2020, Vienna, Austria, November 3-6, 2020, Proceedings (Lecture Notes in Computer Science, Vol. 12400)*. Springer, 250–260. https://doi.org/10.1007/978-3-030-62522-1_18

[10] Latrice G. Landry, Nadya Ali, David R. Williams, Heidi L. Rehm, and Vence L. Bonham. 2018. Lack Of Diversity In Genomic Databases Is A Barrier To Translating Precision Medicine Research Into Practice. *Health Affairs* 37, 5 (2018), 780–785. https://doi.org/10.1377/hlthaff.2017.1595 arXiv:https://doi.org/10.1377/hlthaff.2017.1595 PMID: 29733732.

[11] Yang Liu, Tim Althoff, and Jeffrey Heer. 2020. Paths Explored, Paths Omitted, Paths Obscured: Decision Points & Selective Reporting in End-to-End Data Analysis. In *Proc. of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, 1–14. https://doi.org/10.1145/3313831.3376533

[12] Moran Mizrahi and Dafna Shahaf. 2021. 50 Ways to Bake a Cookie: Mapping the Landscape of Procedural Texts. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*. ACM, 1304–1314. https://doi.org/10.1145/3459637.3482405

[13] José María González Pinto and Wolf-Tilo Balke. 2017. Result set diversification in digital libraries through the use of paper's claims. In *Digital Libraries: Data, Information, and Knowledge for Digital Lives (Lecture Notes in Computer Science)*, Songphan Choemprayong, Fabio Crestani, and Sally Jo Cunningham (Eds.). Springer International Publishing, 225–236. https://doi.org/10.1007/978-3-319-70232-2_19

[14] Mario Prieto, Helena Deus, Anita de Waard, Erik Schultes, Beatriz García-Jiménez, and Mark D. Wilkinson. 2020. Data-driven classification of the certainty of scholarly assertions. *PeerJ* 8 (Apr 2020), e8871. https://doi.org/10.7717/peerj.8871

[15] Graciela Rosemblat, Marcelo Fiszman, Dongwook Shin, and Halil Kilicoglu. 2019. Towards a characterization of apparent contradictions in the biomedical literature using context analysis. *Journal of Biomedical Informatics* 98 (Oct 2019), 103275. https://doi.org/10.1016/j.jbi.2019.103275

[16] Jodi Schneider, Anita de Waard, Wolf-Tilo Balke, Xiaoguang Wang, Ningyuan Song, Bolin Hua, and Yuanxi Fu. 2021. Digital Infrastructures for Scholarly Content Objects. In *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2021, Champaign, IL, USA, September 27-30, 2021*, J. Stephen Downie, Dana McKay, Hussein Suleman, David M. Nichols, and Faryaneh Poursardar (Eds.). IEEE, 346–347. https://doi.org/10.1109/JCDL52503.2021.00069

[17] Jodi Schneider, Di Ye, Alison M. Hill, and Ashley S. Whitehorn. 2020. Continued post-retraction citation of a fraudulent clinical trial report, 11 years after it was retracted for falsifying data. *Scientometrics* 125, 3 (01 Dec 2020), 2877–2913. https://doi.org/10.1007/s11192-020-03631-1

[18] Dafna Shahaf, Carlos Guestrin, and Eric Horvitz. 2012. Metro maps of science. In *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*. ACM, 1122–1130. https://doi.org/10.1145/2339530.2339706

[19] Thomas Stoeger, Martin Gerlach, Richard I. Morimoto, and Luís A. Nunes Amaral. 2018. Large-scale investigation of the reasons why potentially important genes are ignored. *PLOS Biology* 16, 9 (09 2018), 1–25. https://doi.org/10.1371/journal.pbio.2006643

[20] Ludovic Trinquart, David Merritt Johns, and Sandro Galea. 2016. Why do we think we know what we know? A metaknowledge analysis of the salt controversy. *International Journal of Epidemiology* 45, 1 (02 2016), 251–260. https://doi.org/10.1093/ije/dyv184

[21] Huimin Zhou, Ningyuan Song, Wanli Chang, and Xiaoguang Wang. 2019. Linking the thoughts within scientific papers: Construction and visualization of argumentation graph. *Proc. of the Association for Information Science and Technology* 56, 1 (2019), 757–759. https://doi.org/10.1002/pra2.205