# Final report for G-2016-7209

SoftCite: Raising visibility and improving incentives for software work as a contribution in the scientific literature.

PI: James Howison, University of Texas at Austin

## Summary

Software contributions to scholarship should be more visible to the scholarly reputation system. With greater visibility, we hope, comes greater rewards, and with greater rewards, we hope for greater quality and collaboration in the production and maintenance of software in science and scholarship. To this end we have produced three contributions: softcite-dataset, the softcite-kb toolchain, and CiteAs.org. Below we briefly describe each, and other grant-funded activities. We then provide screenshots and illustrations before discussing goals, changes, challenges, listing outputs, highlighting possible “Grantees in the News” and platforms to list.

**Softcite-dataset** is a gold standard dataset of software mentions in biomedicine and economics, manually annotating 4,971 academic PDFs finding 4,093 software mentions, together with 2,541 annotations of details such as URL, publisher, and version. This dataset has been made available through Github[[1]](#footnote-1), deposited with Zenodo[[2]](#footnote-2), and we have published, in the JASIST journal, a full description of our work to create this dataset, including reporting learnings about dataset creation and interdisciplinary collaboration (Du et al., 2021). The dataset is suitable to enable the Named Entity Recognition of software in the research literature, whether those mentions are formally cited or simply informally mentioned in free-text. We are excited that the dataset has begun to be used by the science of science community, including a 2020 hackathon by the UK Software Sustainability Institute, as well as in-house researchers at the Chan Zuckerberg Meta project. We have presented this work at academic and industry events including SciPy (2020), ODSC (2021) and workshops at ACM CSCW and ACM CHI.

**Softcite-kb** is a prototype software mention discovery tool, available on Github[[3]](#footnote-3), that implements a machine learning model trained on the softcite-dataset, together with a full pipeline of PDF conversion, entity recognition, disambiguation, and knowledge base presentation. The tool is showing good performance using a BiLSTM-CRF+Elmo model with f-scores over 0.80 and scalable performance. We have obtained continued funding for the further development of this tool, as discussed below, and we have a draft paper under preparation for the CKIM conference, titled, “Mining Software Entities in Scientific Literature: Document-level NER for an Extremely Imbalanced and Large-scale Task”. We obtained additional computing resources from the Texas Advanced Computing Center to run this pipeline on the CORD-19 collection of academic PDFs, and in newly funded work we are expanding this to ~16 million PDFs available through the Unpaywall system (from ImpactStory, an organization previously supported by the Sloan Foundation).

**CiteAs.org** is a specialized search engine that provides recommended citations for software tools, as well as educating and encouraging software producers to make clear requests for citations. The tool was published in 2017 and has been maintained since. CiteAs has been promoted through participation in scientific software workshops and relevant grant programs, including an invited presentation at the CZI “Essential Open Source Software” PIs meeting (Howison, 2020), a hackathon at the UK RSE conference in 2018, the widely read blog of the US Research Sustainability Institute (URSSI), and at the Scientific Software Registry Collaboration Workshop, a group funded by the Sloan Foundation. We have conducted interviews with users, and are working towards a publication describing the tool and reactions to it.

Additionally, we have implemented a crowdsourced extension to the softcite-dataset, working with TagWorks to re-annotate our training data set and to annotate 2,743 new sections of articles from snippets, work completed in December 2020. We set this up to provide a comparison of the effectiveness and cost of crowdsourcing in this space, and we are working towards a publication of that effort. TagWorks is a company founded by Nick Adams, and we understand that it builds on a tagging interface previously supported by the Sloan Foundation.

We have undertaken educational efforts about software citation, including developing draft materials for the Sloan supported Carpentries (the Make Code Citable[[4]](#footnote-4) lesson), participation in the RDA Software Citation Working Group, and presentations at the follow-up RDA Software Citation Implementation Working Group.

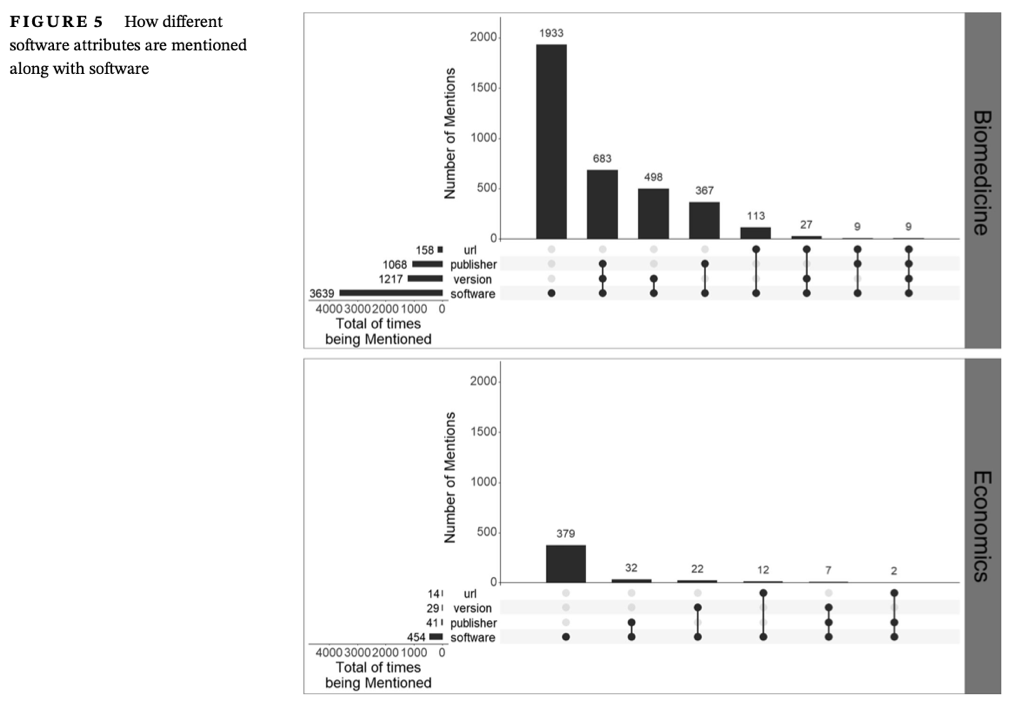
We involved over 30 undergraduate and masters students in the annotation of the softcite-dataset, training them in annotation and github collaboration. The annotation group involved participants from University of Texas at Austin, as well as Huston Tillotson, a local historically black serving institution. Finally, we trained two doctoral students in research in this area, including managing research teams (Caifan Du and Hannah Cohoon).

## Screenshots

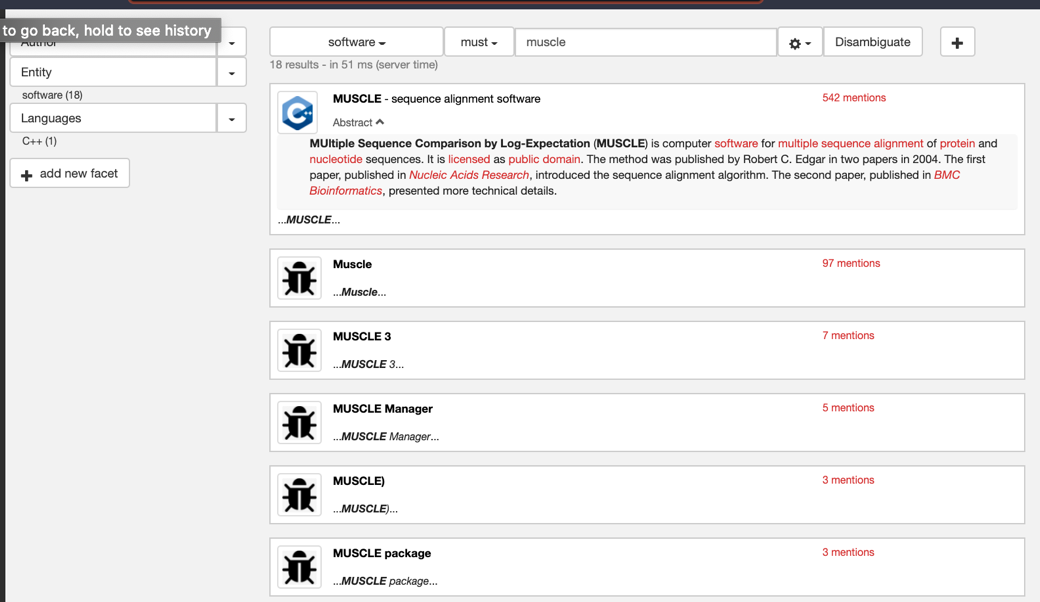
Softcite-dataset was a four year effort involving over 30 masters and undergraduate students, managed by the PI and two doctoral students. We developed a coding scheme and together annotated just under 5,000 articles. The figure below, taken from our 2021 JASIST paper, shows our dataset (and the ability to “paint” a PDF with the software mentions).



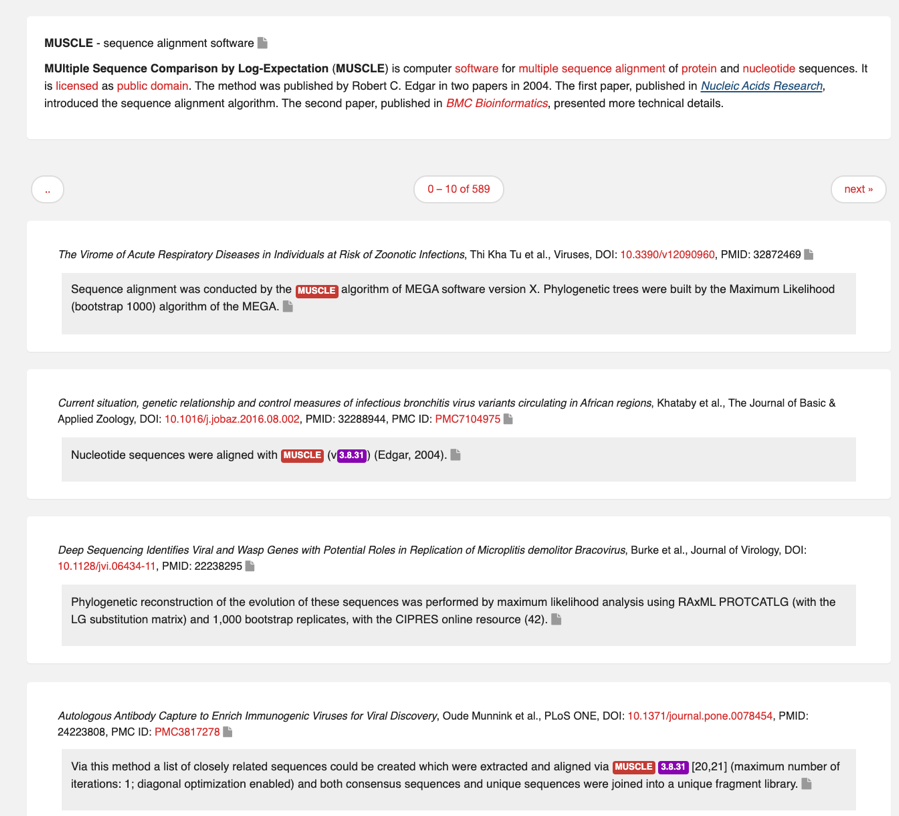
We annotated by biomedicine and economics articles, finding distinct patterns. While there were substantially more mentions in the biomedicine literature, in both mentions were extremely sparse, generating challenges for machine learning. We are addressing these challenges in our softcite-kb toolchain.



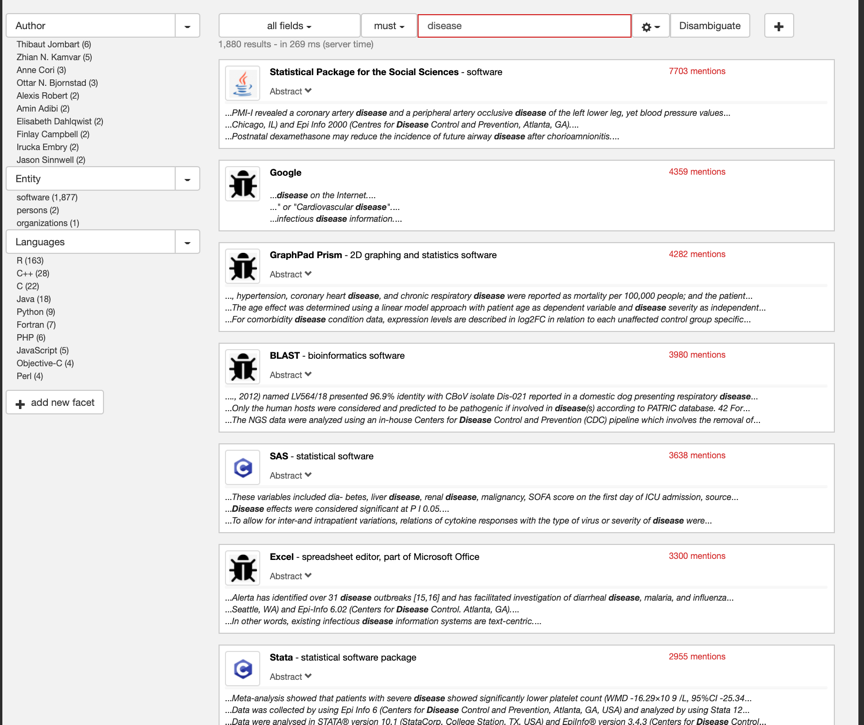
The softcite-kb toolchain is under continued development. Nonetheless, we have a prototype browser that shows identified software mentions extracted from the over 100,000 PDF articles in the CORD-19 dataset on scholarship relevant to COVID-19. We are building out an API for open access to these data. The first screenshot shows a search (done live while creating this report) for the gene software MUSCLE. The results show both success (disambiguated mentions of MUSCLE) and challenges (other entries show other instances of MUSCLE which should be incorporated in the top entry, still others show non-software occurrences of Muscle.)



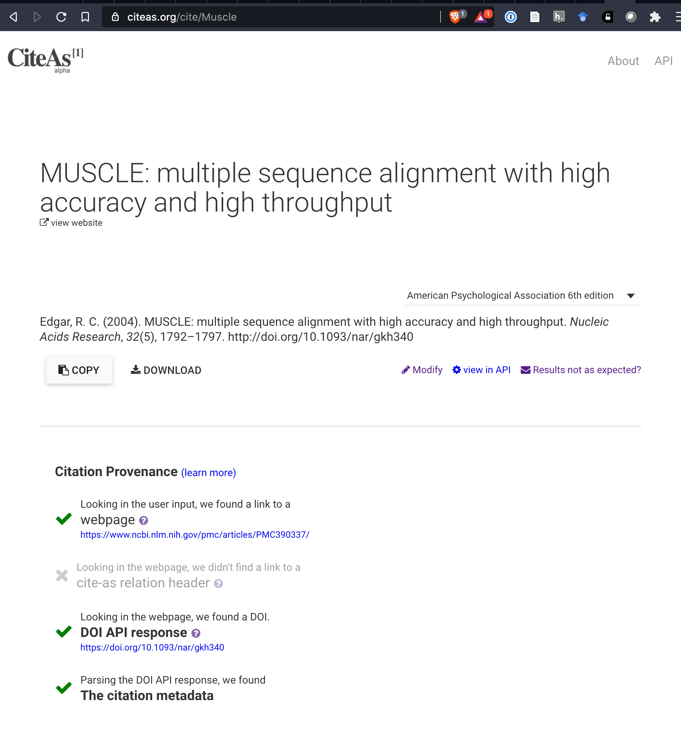
Clicking on the top entry enable us to see details of each mention; these are linked to the DOI for the article (and thus to journal, authors etc). The formal citations are also identified. Each piece of metadata will be available via the API.



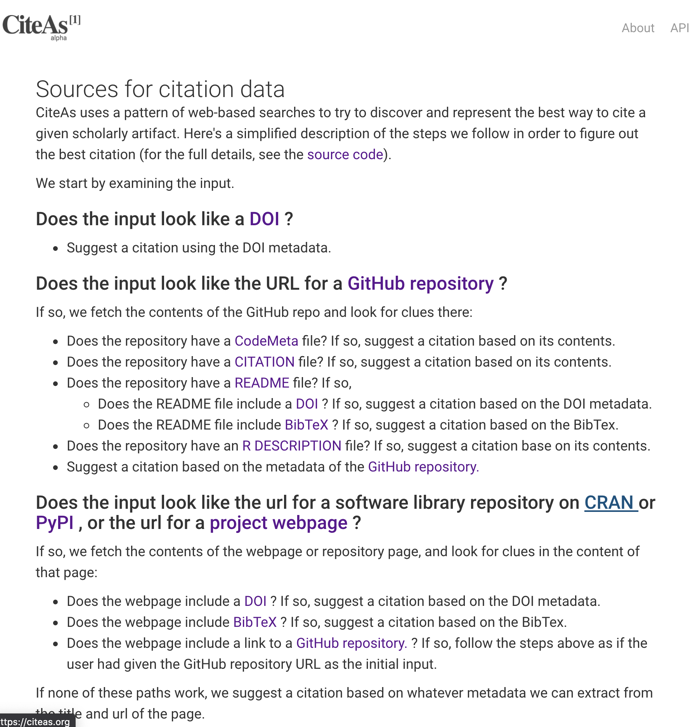
Finally, we show a search for a common word in the CORD-19 dataset “disease”, showing the ranked software packages mentioned in this collection of articles.



CiteAs.org enables scientists and scholars to enter the name, URL, or other identifier for a piece of software and receive a formatted citation. These recommendations come from searching requests for citations from the creators of the software. Below we search for the string “muscle”.



In addition, CiteAs seeks to educate software producers to make clearer requests for citation, by demonstrating the places it looks for requests for citation and thus gently suggesting that producers put their requests in these spots.



## Original Goals and Metrics

1. Development and publication of gold standard dataset of software mentions
2. An open source machine learning system for recognizing software mentions
3. Launch and dissemination of open source systems to:
   1. Suggest software citations based on article text (CiteSuggest)
   2. Offer preferred citations for a piece of software (CiteMeAs)
   3. Illustrate the impact of a piece of software (Software ImpactStory)
4. Published research papers evaluating social and technical opportunities and barriers for each prototype

## Key Changes

During the course of the grant period, as discussed with our program officer and in our interim reports, we decided to concentrate on expanding the gold standard dataset of software mentions and ensuring its quality. To concentrate on that goal, we scaled back efforts towards realizing the CiteSuggest and Software ImpactStory prototypes (each of which required a working software mention detection system). Happily, now that the software mention recognizer is complete, we have obtained additional grant support from the Mellon Foundation (via a sub-contract from Karthik Ram and rOpenScience managed through UCBerkeley) and are now able to build towards tools like CiteSuggest and Software ImpactStory.

In addition, during the grant the team shifted from Howison, Piwowar, and Priem to Howison and Patrice Lopez. Priem and Piwowar (from ImpactStory) worked to implement CiteAs and to provide sampled access to the literature for the annotation team creating the gold standard dataset. As work began towards the Software Mention recognizer, Priem and Piwowar identified Patrice Lopez as an ideal collaborator and employed him. Lopez is the creator of the GROBID PDF conversion tool, which provide to be vital to the project. Howison and Lopez continue to work together towards extensions of the softcite set of data and tools.

Finally, we had hoped to present sociotechnical studies of the barriers to adoption for each of our prototypes. We were able to do this with CiteAs, through interviews with users and presentation to publishers, describing these in our posters and presentations. Without functioning prototypes for the other tools we were unable to pursue this. The doctoral student, rather than undertaking that work, worked to organizing the creation and quality assurance on the softcite-dataset.

## Challenges

The key challenges faced were:

1. The extent of effort needed to build the gold standard dataset was significantly greater than we anticipated.
2. The difficulty of converting PDF articles to useable text was anticipated, but we thought we could deal with that after our content analysis work. In fact, usable conversions were key to calculating agreement statistics for the content analysis work and therefore we had to spend time aligning the XML versions of articles to assess when training was sufficient to move to individual coding.
3. Even with well-trained annotators, when Patrice Lopez joined as our machine learning expert he identified issues with the gold standard set, and we decided to conduct extensive expert review.
4. It was unclear how much provenance to maintain about our annotation effort
5. We, as many do, experienced interdisciplinary tensions through the project, taking more time to synchronize our expectations and understandings.

We address these issues in detail, hoping that other scholars can learn from them, in our recently published JASIST paper (Du et al., 2021), documenting our annotation process in far more detail than usually done, meeting recent calls to better document gold standard datasets (e.g., Gebru et al., 2020)

## List of outputs

1. CiteAs.org
2. The CiteAs API and source code: <https://github.com/ourresearch/citeas-api>
3. The softcite-dataset at Zenodo: <https://zenodo.org/record/4445202#.YANCG-j0k2w>
4. The softcite recognizer and knowledge base: <https://github.com/softcite/softcite_kb>
5. Du, C., Cohoon, J., Lopez, P., & Howison, J. (2021). Softcite dataset: A dataset of software mentions in biomedical and economic research publications. *Journal of the Association for Information Science and Technology*, *forthcoming*. <https://doi.org/10.1002/asi.24454>
6. Caifan Du, James Howison, Patrice Lopez, Norman Gilmore, Johanna Cohoon, Nick Adams, & Karthik Ram. (2021). *Softcite: Data-driven Software Visibility for Science*. Open Data Science Conference (ODSC East).
7. Invited Keynote presentation (September 2018) “Software makes science better, but is it research? Arguments for a research agenda in scientific software work” Presented at (dot)Astronomy X. [https://figshare.com/articles/Software\_makes\_science\_better\_but\_is\_it\_research \_DotAstonomy\_X\_Presentation/7127723](https://figshare.com/articles/Software_makes_science_better_but_is_it_research%20_DotAstonomy_X_Presentation/7127723)
8. Cohoon, J., Howison, J. (2019). Routes to Sustainable Software: Transitioning to

Peer Production. Poster session presented at the 2019 Collegeville Workshop on Sustainable Scientific Software, Collegeville, MN.

1. Du, Caifan, Cohoon, J, and Howison, J (2020). CiteAs: Bridging the gap in software citation. SciPy 2020 Poster. <https://github.com/howisonlab/softcite-dataset/blob/master/docs/papers/SciPy2020-CiteAs-Poster.pptx.pdf>
2. Du, C. F., Cohoon, J., Howison, J., Priem, J., & Piwowar, H. (2019). Studying Processes in Software Citation Towards Improved Collaboration Among Scientists. Process Theory Workshop (ACM CSCW). <https://github.com/howisonlab/softcite-dataset/blob/master/docs/papers/SoftCite_CSCW_Process_Workshop_2019.pdf>
3. Du, C., Howison, J., & Lopez, P. (2020). Softcite: Automatic Extraction of Software Mentions from Research Literature. SciNLP workshop at AKBC2020. <https://github.com/howisonlab/softcite-dataset/blob/master/docs/papers/Softcite-SciNLP-pre.pdf> <https://github.com/howisonlab/softcite-dataset/blob/master/docs/papers/Softcite-SciNLP-AKBC2020-Abstract.pdf>
4. URRSI blog post on CiteAs: <http://urssi.us/blog/2018/10/01/citeas.org-discovering-and-improving-software-requests-for-citation/>
5. Du, C. F., Cohoon, J., Howison, J., Priem, J., & Piwowar, H. (2019). Studying Processes in Software Citation Towards Improved Collaboration Among Scientists. CSCW Workshop “Mapping the How of Collective Action” (ACM CSCW).
6. Caifan Du, Piwowar, H., Howison, J., & Priem, J. (2020). CiteAs: Bridging the Gaps in Software Citation. Scientific Software Registry Collaboration Workshop, University of Maryland. <https://github.com/howisonlab/softcite-dataset/blob/master/docs/papers/CiteAs_SciTS2020_May2020.pdf>
7. Caifan Du & James Howison. (2020). How to Study and Support Labor of a Nascent Category? Reflection on the Investigation and Intervention of Scientific Software Work. Presentation at ACM CHI Workshop on Worker Centered Design. <https://github.com/howisonlab/softcite-dataset/blob/master/docs/papers/CHI20-WorkerCenteredDesignWorkshop-Du.pdf>

The project supported two doctoral students and over 30 undergraduate and masters students. For many of the undergraduate and masters students this was their first involvement in research. In addition to the research experience, we also taught them about content analysis, RDF, git and github based collaboration.

We think that a story in the “Grantees in the News” could be arranged around the softcite-dataset paper, together with the detailed JASIST paper outlining its creation. In addition we would be excited to promote the CiteAs.org site (drawing on the URSSI blog post linked above.)

For the Platforms supported under Data and Computational Research, we would be excited to list:

1. CiteAs.org
2. Softcite-kb

For the Platforms supported under Scholarly Communications we would be excited to list:

1. Softcite-dataset
2. CiteAs.org

## References

Du, C., Cohoon, J., Lopez, P., & Howison, J. (2021). Softcite dataset: A dataset of software mentions in biomedical and economic research publications. *Journal of the Association for Information Science and Technology*, *forthcoming*. https://doi.org/10.1002/asi.24454

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2020). Datasheets for Datasets. *ArXiv:1803.09010 [Cs]*. http://arxiv.org/abs/1803.09010

Howison, J. (2018). *Challenges and Pathways to Sustainability in Scientific Software Ecosystems*. https://doi.org/10.6084/m9.figshare.7043093.v1

Howison, J. (2020, February 24). *Sustainability in Scientific Software: Ecosystem complexity and Software Visibility*. https://doi.org/10.5281/zenodo.3885803

1. https://github.com/howisonlab/softcite-dataset/ [↑](#footnote-ref-1)
2. https://zenodo.org/record/4445202#.YANCG-j0k2w [↑](#footnote-ref-2)
3. https://github.com/softcite/softcite\_kb [↑](#footnote-ref-3)
4. https://github.com/carpentries-incubator/citable-software [↑](#footnote-ref-4)