Extracting body-text citations at scale

Matt Marx & Aaron Fuegi (Boston University)

Bryan, Sampat, and Ozcan (2019) extracted non-patent citations from the body text of patents for 250 journals over a 30-year period. As part of the Innovation Information Initiative, the Sloan Foundation has provided us with initial funding to create a comprehensive inventory of body-text citations in U.S. patents since 1836. We provide a progress report, including a taxonomy of citation types, methodologies for extracting and matching them to scientific papers at scale, and initial results including the trend of front-page vs. body-text citations over time.