itx®

REDUCED COST AND IMPROVED PERFORMANCE VIA DATA SCIENCE

SUCCESS STORY

SSRN's recent growth ignited rapid expansion in its vast collection of academic research papers. To keep pace with the new submissions, the research repository asked ITX for help modernizing its legacy citation system. ITX's data science team applied machine learning algorithms to confidently identify, accurately allocate, and securely store millions of citations automatically – at a fraction of the cost – allowing SSRN to release the power of its rich but untapped data set.
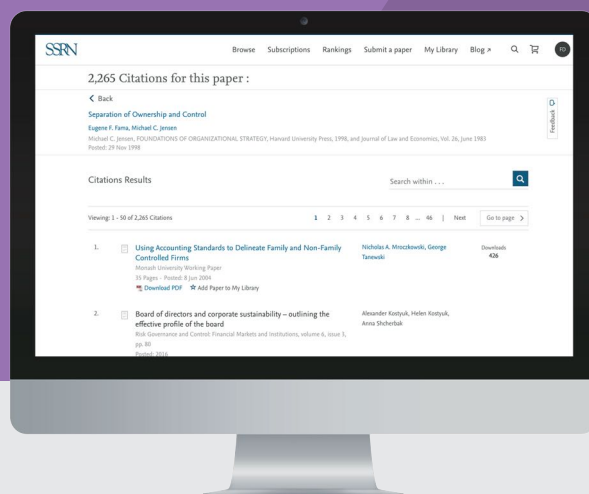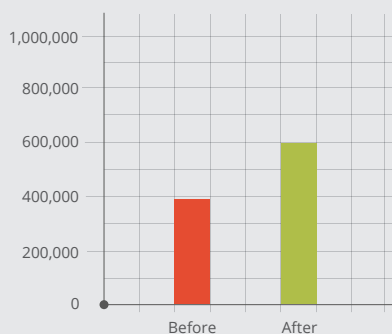
PERFORMANCE: 16 MONTHS POST-LAUNCH

↑ 60%

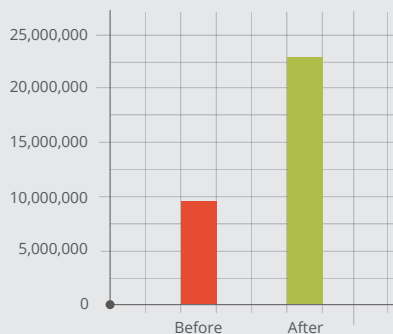**INCREASE IN RESEARCH PAPER SUBMISSIONS**

demonstrates authors' confidence in SSRN's new processing capabilities

↑ 150%

SPIKE IN CITATIONS PROCESSED

shows algorithm's ability to process large data sets quickly and accurately

↓ 100%

ELIMINATED BACKLOG OF RESEARCH PAPERS AWAITING PROCESSING

cut lead time required to process a paper's citations from 6 months to 9 seconds

TESTIMONIAL

ITX worked with SSRN to replace an antiquated, manual process with a completely automatic one that has removed all manual processing costs associated with the previous solution. Using the GROBID algorithm, ITX did a brilliant job prototyping a solution, working out models to prove the algorithm's efficacy, and delivering the system in a timely fashion. We've been delighted with the results.

— *Michael Parsons, Director of Product Management, SSRN*

MEET THE CLIENT

SSRN
Tomorrow's Research Today

SSRN, formerly known as Social Science Research Network, is devoted to the rapid worldwide dissemination of research. Through continuous innovation, they focus on creating tools that enhance researcher workflow and productivity, building bridges to close the divide between the previously separate worlds and workflows of working papers and published papers.

GOAL	Through data science, leverage a machine learning algorithm to extract, assess, and display citations within academic research papers. Create an automated process to eliminate a sizable backlog of work. Construct a scalable, extensible database architecture, within the AWS hosting environment, to process and store millions of citation records.

STRATEGY	ITX data scientists conducted comprehensive experimentation and testing to devise multiple proofs of concept. We tested citation extraction algorithms against an array of document complexities and key performance measures. We compared data storage solutions to identify the one best able to accommodate a large and expanding collection of citation content.

## UNDERSTAND THE CHALLENGE ❯ IMPLEMENT TARGETED SOLUTIONS

- The existing CiteReader extraction system is a manual data entry process. It is slow, inefficient, and expensive.
- Citation counts increase too slowly, frustrating authors that their scholarly submissions were not timely distributed to the research community.
- Delayed processing left researchers without actionable awareness of the research indicators regarding which SSRN papers they should pay attention to.
- The existing technology had grown obsolete; the citation extraction process lacked necessary scalability and storage capacity to keep pace with the influx of new content.

- ITX data scientists studied, tested, and selected the algorithm best suited to identify and extract citations, match them against SSRN content, and display them in the article Abstract.
- The team added a new CrossRef look-up service that reveals the interconnectedness of scholarly research across the web.
- After comprehensive review, SSRN chose the DynamoDB database, hosted on AWS, based on its extensibility and scalability to accommodate a large and rapidly growing data set.
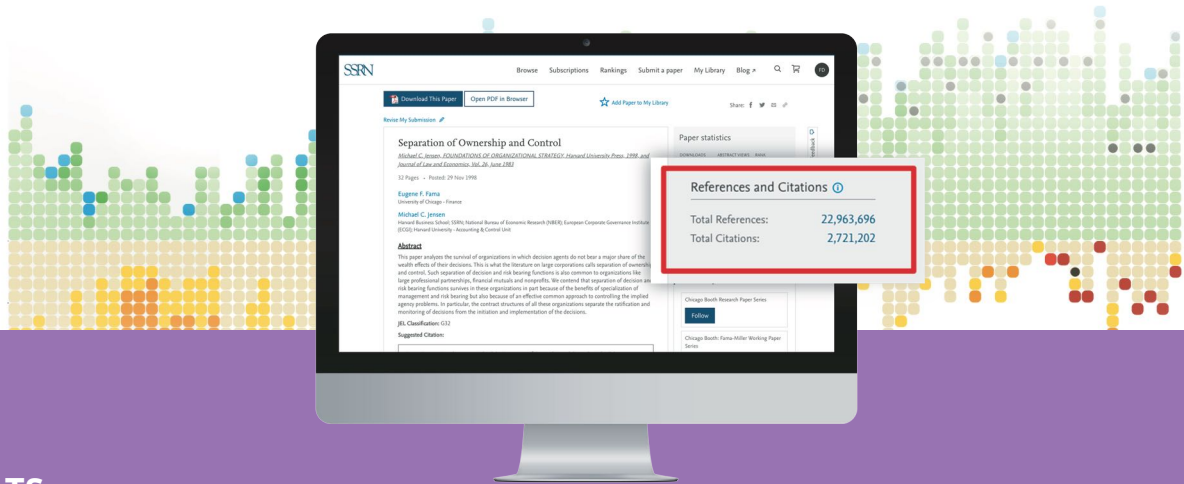
RESULTS

❯ SSRN utilized machine learning algorithms to eliminate their backlog of research papers and accelerate the practice of scholarly research.

❯ By unleashing the power of the GROBID algorithm, ITX data scientists drove revenue and reduced costs through process automation; and accurately measured, monitored, and improved the research experience for end users.

❯ The ITX Data Science response continues to power SSRN's citation extraction and classification system, providing an evergreen solution that continues to deliver value.