Impactstory
PIs: Jason Priem and Heather Piwowar
Title: Tear down this wall: building the Open Research Profile

Researchers love sharing their work with others, and love seeing metrics on who has read and used their research.

ResearchGate (RG) and Academia.edu have capitalized on this interest, building platforms that offer easy ways for scholars to upload papers, see who reads them, and discuss. As a result, millions of scholars have made profiles on these commercial sites, and tens of millions of papers have been uploaded to their servers. More than 10% of all open fulltext on Google Scholar points to ResearchGate.

If only this were good news! But because these for-profit companies are so hungry to build closed walled gardens, each well intentioned author action is actually bad news: each paper uploaded to ResearchGate and Academia.edu is a paper that is not made Open Access, openly accessible, discoverable, and reusable by everyone.

Ultimately, the problem is that ResearchGate and Academia.edu want to build self-sufficient publishing, identity, and review platforms rather than open repositories. Uploaded publications and reviews aren't available via OAI-PMH endpoints or exposed via public APIs. Uploaded papers don't have license information, so reuse terms are not assured. Worse, reading the "open" uploaded PDFs is restricted to people who become members of the walled gardens (something many researchers don't realize, partly because the platforms cut backroom deals with a few indexers as they have done with Google). Even if the intensions of ResearchGate and Academia.edu are honourable, make no mistake: the scholarly resources they hold will become more closed over time. They are millions of dollars in debt to venture capitalists and need a viable business model.

This colonization of online scholarly identity threatens the long-term vision of open science.

The solution is a profile system that is better than ResearchGate and Academia.edu for scholars AND for scholarship, a compelling alternative to their walled gardens. It will be built on several new, game-changing data sources:

- Microsoft Academic Graph (MAG), a comprehensive linked graph of researchers, institutions, publications, and citations,
- oaDOI, a new system we've built that aggregates all open-access in a single database,
- CrossRef Event Data, a new free source of altmetrics.

Setup will be as easy as typing your name and selecting a few of your publications. Fulltext papers will be linked automatically, and adding new papers unites a one-click user experience with best-practice repository management: behind the scenes we automatically upload it to a full

open repository (figShare, Zenodo, OSF). Publications will be shown as part of a “research package” that also includes preprints, datasets, software, and other diverse papers.

Users will view meaningful metrics including citations, predicted citations (based on machine-learning using altmetrics, downloads, and article attributes), downloads (from publishers when available, estimated when necessary), fulltext mentions of datasets and software (from text mining), altmetrics (from altmetric.com and crossref event data). There will be an institution-level dashboard that aggregates this all in an real-time report. The system will be two-way synced to ORCID profiles, as well as post-publication peer review frameworks like PubPeer and workflow system like the Open Science Framework. Like all our previous projects, code, data, and APIs will be free and open.

How will this result in increased access to research results?
This project will increase access to research results in three ways:

1. Millions of preprints will move from the walled gardens of RG and Academia.edu to institutional repositories and disciplinary repositories like BioArXiv. In the longer term, we avert the disaster of scholarly identity being colonized by these for-profit closed systems, and instead create a vibrant open profile ecosystem that promotes third-party apps to filter, recommend, text-mine, recombine, and remix these preprints.
2. More researchers will share alternative research products like code, data, and workflows openly. The new system will feature alternative products prominently, along with metrics demonstrating their impact. This will help encourage users to see them as first-class results of research that are worth investing in and sharing, building a culture of value around diverse research products.
3. Over the long term, this system will allow the replacement of the traditional, review-then-publish journal system with a publish-openly-then-review system--a model envisioned years ago with PLOS ONE and altmetrics, but not yet realized. The key will be the application of state-of-the-art machine-learning techniques to the enormous dataset of scholarly publications, their authors, their citations, and altmetrics data like online reviews and discussion. The proposed system is the first to gather all this in one place. Eventually, it will power credible post-publication peer review in the form of customized, impact-aware article recommendation streams presented to individual users. This is the future of open science. We must make it open for all.

Risks and mitigations

- *Chicken-and-egg problem impedes growth*. It’s a well-known problem that users aren’t interested in new social networks until there are already many users there. Mitigations:
    - The system is useful even for the first user: altmetrics, downloads, citations, and openness statistics are interesting and valuable to users even if none of their friends has a profile yet. We don’t have to be limited to bogus metrics on activity within a walled garden.
    - Profiles exist for everyone, right away. Because profiles are built from public, open data, we can pre-populate them immediately, allowing users to search,

follow, and view their colleagues right away. There's no ghost town.
    - Interest from institutions will help power early growth. ResearchGate and Academia.edu are notorious for spammy and aggressive marketing; partly as a result, trusted institional representitives like librarians are generally reluctant to recommend these systems. Our proposed system, on the other hand, will have the express support of researchers' institutions, since we will be working closely with those institutions in building dashboards and reporting systems, as well as leveraging established credibility as an open, free, and mission-driven nonprofit.
- *Researcher name and institution name information will be too difficult to aggregate and disambiguate*. This is a notoriously difficult problem, with a large research literature dedicated to it. Mitigations:
    - The recent availability of the Microsoft Academic Graph as open data is a game-changer, since it allows a relatively small team like ours to leverage their years of work and investment in solving this problem using state-of-the-art machine-learning approaches. Moreover, since the data is only available for purposes that don't generate revenue, our status as a grant-funded nonprofit makes us particularly suited to use this data effectively.
- *This proposal has much in common with the profile system Impactstory, which has struggled to attract a significant user base*. Mitigations:
    - A profile will be quick and easy to make in the proposed system, which was never true for Impactstory (an Impactstory profile is quick for researchers with up-to-date ORCID profiles, but this is a very small number of scholars)
    - The newly-available MAG data means the proposed system includes citations, which researchers value much more than Impactstory's altmetrics.
    - The proposed system has a (much-requested) Institutional component. This is possible thanks to the newly-available open Crossref Event Data altmetrics stream, which may be aggregated at an insitutional level, unlike Altmetrics.com data (due to API licence contract restrictions).

Timeline

- Year 1: We will have a Minimum Viable Product version in the first three months. From there, we will iterate in sprints over the next nine months, focusing on listening to feedback from researchers and institutional stakeholders. Significant development is complete at end of year. We will finish the year with 10k monthly active users, (2k in the last month) and 10 institutional users.
- Year 2: We continue fine-tuning the application, particularly recommendation, search, and prediction algorithms as more data accumulates from users. We do a lot of marketing. We improve the UX in response to feedback from users and data from usage logs. Over 100k users at the end of year based on Paul Graham's recommended target of 10% week-over-week growth. 100 institutional users.
- Year 3: We transition into supporting and marketing the application, emphasizing growth of a project that is now well-tuned to user needs. Over 1M users, 500 institutional users.

Appendix

High-level budget (total: $1,185,000 over three years)

Year 1 ($560k):

- front-end developer ($120k including fringe)
- back-end developer ($120k including fringe)
- data scientist ($120k including fringe)
- project management and communication (50% FTE including fringe from each of the two PIs: $120k)
- scalable servers and data storage ($30k)
- indirect costs (10%)

Year 2 ($375k)

- front-end development (25% FTE from each of the PIs, $60k)
- back-end development (25% FTE from each of the PIs, $60k)
- data scientist ($120k)
- project management and communication (25% FTE from each of the PIs $60k)
- scalable servers and data storage ($40k)
- indirect costs (10%)

Year 3 ($250k)

- front-end development (25% FTE from each of the PIs, $60k)
- back-end development (25% FTE from each of the PIs, $60k)
- project management and communication (25% FTE from each of the PIs $60k)
- scalable servers and data storage ($50k)
- indirect costs (10%)

Project Team and Organizational Background

Impactstory is incorporated as a 501(c)(3) nonprofit corporation in North Carolina. Impactstory began life as total-impact, a hackathon project at the Beyond Impact workshop in 2011. As the hackathon ended, a few of us migrated into a hotel hallway to continue working, eventually completing a 24-hour coding marathon to finish a prototype. Months of spare-time development followed, then funding. We've got the same excitement for Impactstory today.

Impactstory has successfully designed, implemented, and disseminated production-level, web-scale applications including Impactstory.org, Depsy, and oaDOI. These applications have supported tens of thousands of users, and handle many millions of rows of data daily.

Dr. Heather Piwowar (co-founder). Heather has been a passionate advocate and investigator of open science since 2007, publishing what Peter Suber called the "first study to document a [..] correlation between OA data and citation impact". Her Bachelor's and Master's degrees from MIT in Digital Signal Processing have given her a strong mathematics, statistics, and modeling background. In addition, Heather has 15 years of software development. Heather will take the lead on the clustering algorithms and OA Boost Algorithm development, and co-lead the backend development and dissemination efforts.

Jason Priem (co-founder). Jason has been a passionate supporter of open science since his PhD studies began in 2009 and he helped create the field of "altmetrics" to help measure the impact of diverse, open scholarly products. Since then altmetrics has become an important subdiscipline of scientometrics, and has been called one of the "five schools of open science". Jason's pre-academia background is in art, education, educational technology, and interface design, and he has leveraged this experience to help build acclaimed user-focused interfaces for several successful software applications including Impactstory, Depsy, oaDOI, and FeedVis. Jason will lead the user interface and design aspects of the project, and co-lead backend development (particularly altmetrics-related areas) and dissemination efforts.