Lessons Learned from Developing RIALTO: A Research Intelligence System

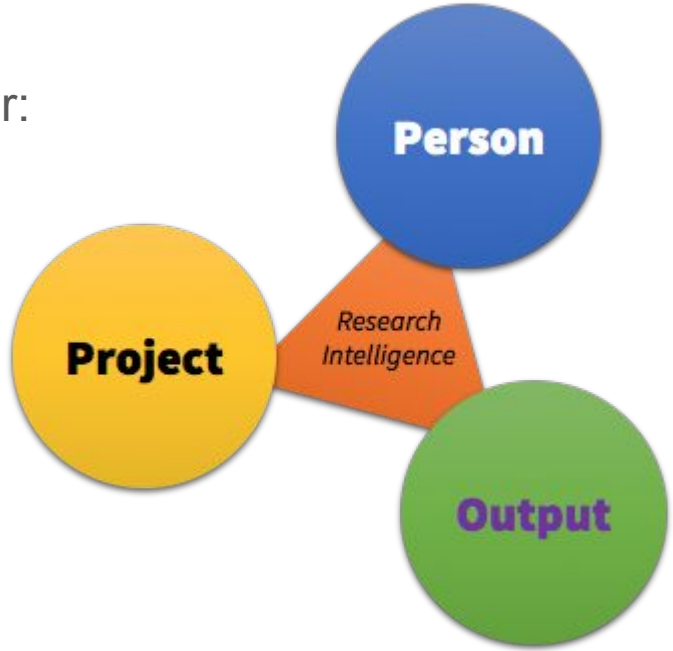
November 18 2020
OPERA Conference

Stanford | LIBRARIES

What is “Research Intelligence”?

Measuring, linking and analyzing research output for:

- Compliance Reporting
- Return on “investment” (or impact)
- Collaboration analysis
- Strategic planning
- Trend analysis



What we know

WHO

HOW

WHAT

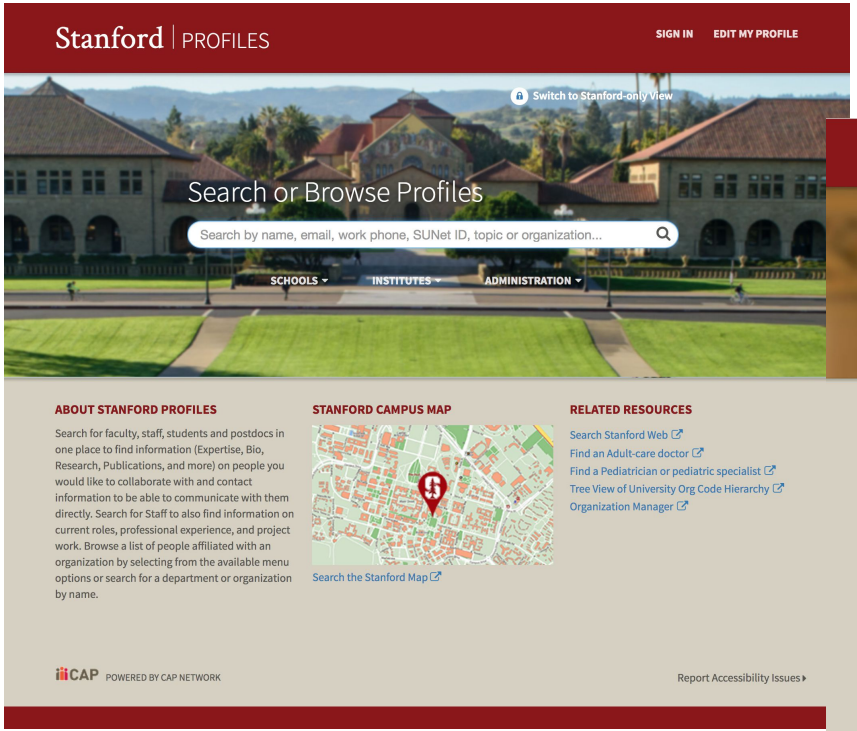
Sort of…

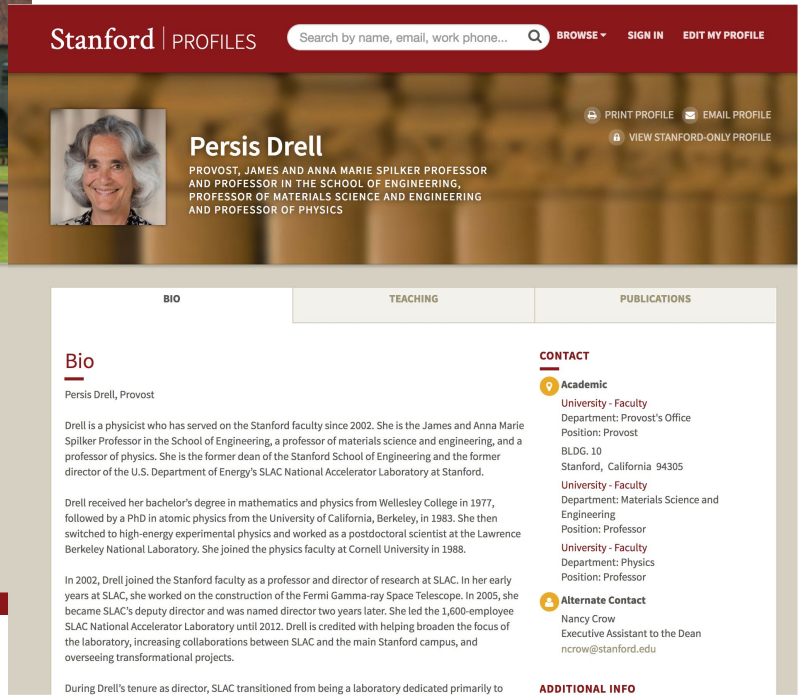
The What

Different systems track different kinds of outputs to varying degrees of completeness, but most are not integrated:

- Books published
- Publications published
- Students graduated
- Courses taught
- Patents filed
- Others

Current public facing Profiles systems





Profiles are not complete and the data that backs it are not well positioned to provide analytics or aggregation.

The current situation is a bit like a company that tracks income and records outputs, but doesn’t measure the organization’s decisions or strategy impact either.

Research intelligence is understanding how a research organization’s decision making and strategy affects its inputs and outputs, as well as the impact the research outputs have on the subject specific domains.

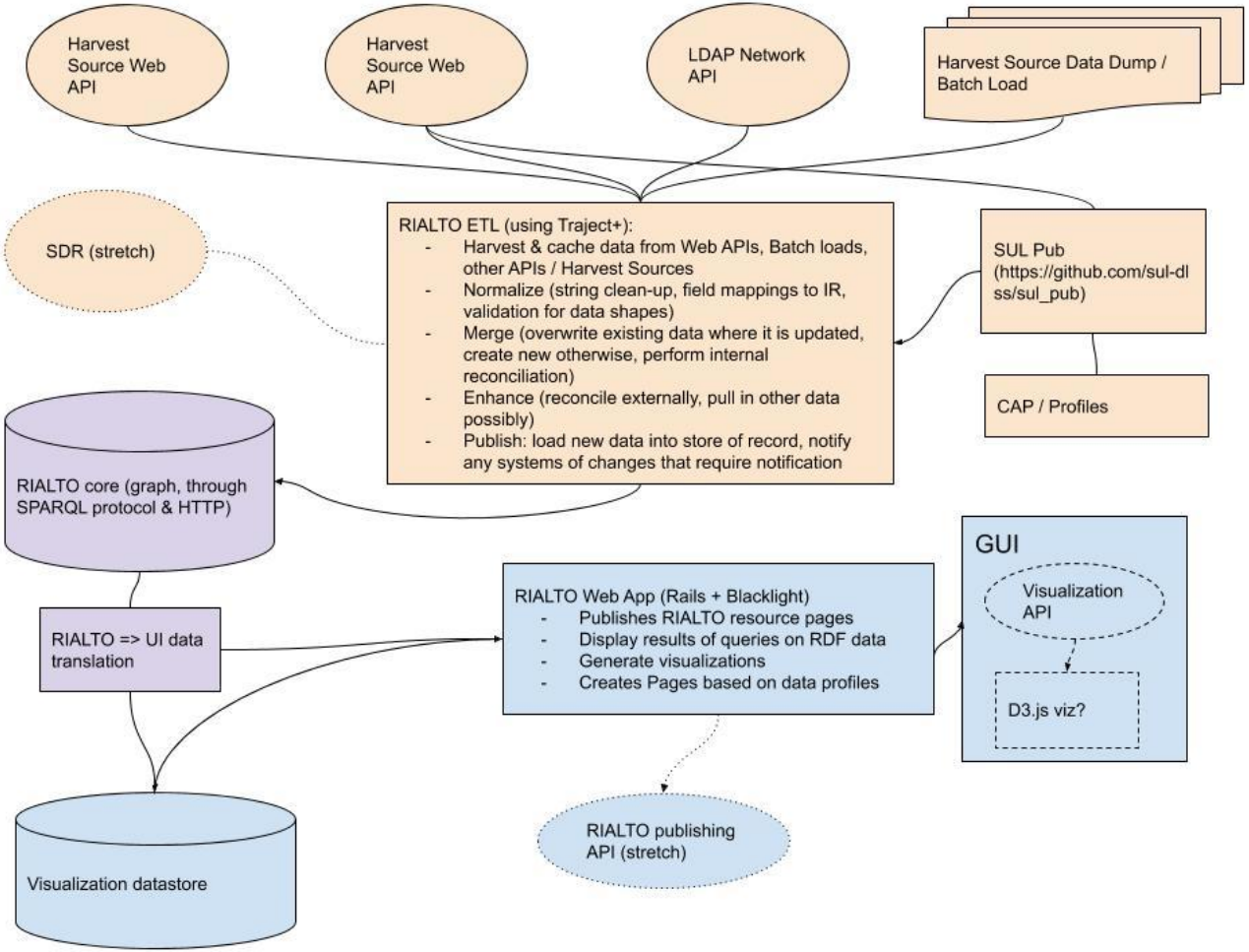
Research intelligence is an expanding domain

- Clarivate (InCites, Converis, Web of Science)
- Digital Science (Dimensions, Symplectic Elements)
- Elsevier (Pure, Scopus, SciVal)
- Open stack (VIVO)
- Custom solutions (what we did!)

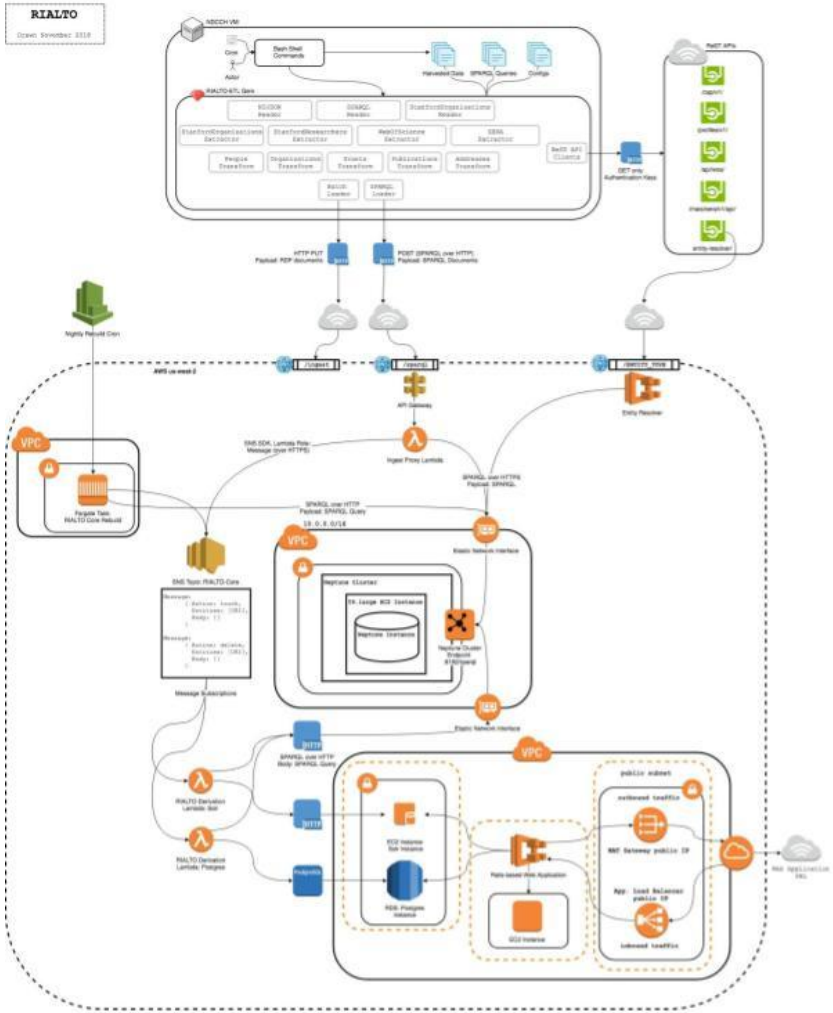
Top Use Cases We Worked On

- As the Director of the Office of International Affairs, I need to be able to map Stanford faculty members' international co-authors by institution so that I can make strategic decisions about potential university-wide partnerships.
- As a university administrator, I want to know about trends in the areas of research being produced by Stanford faculty so I can better understand where faculty interests area
- As a university administrator, I need to understand the impact of cross-disciplinary institutions on research impact so I can assess the ROI
- As a university administrator, I need to know which publications resulted from which grants so that I can fulfill reporting requirements to funding agencies

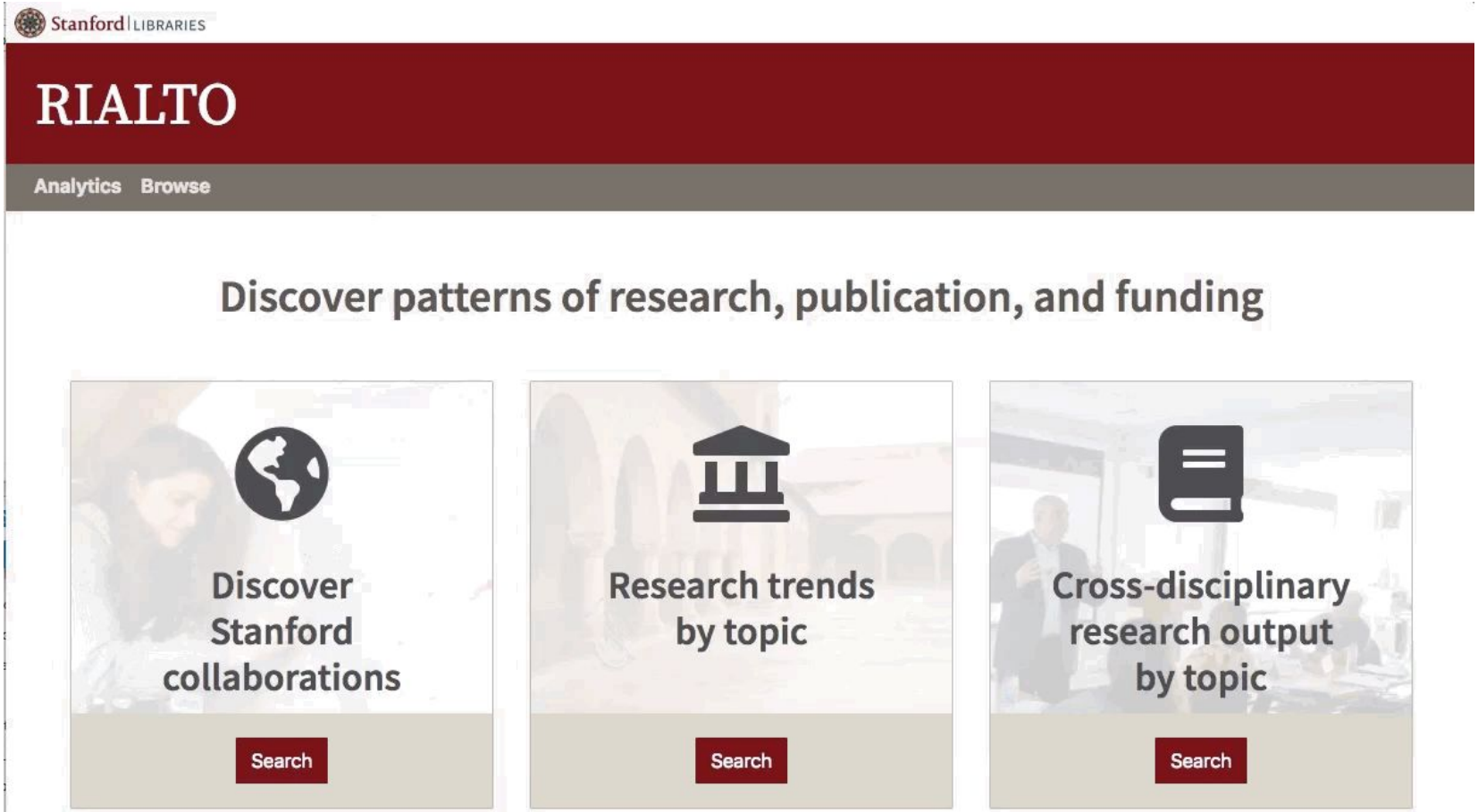
High Level Architecture



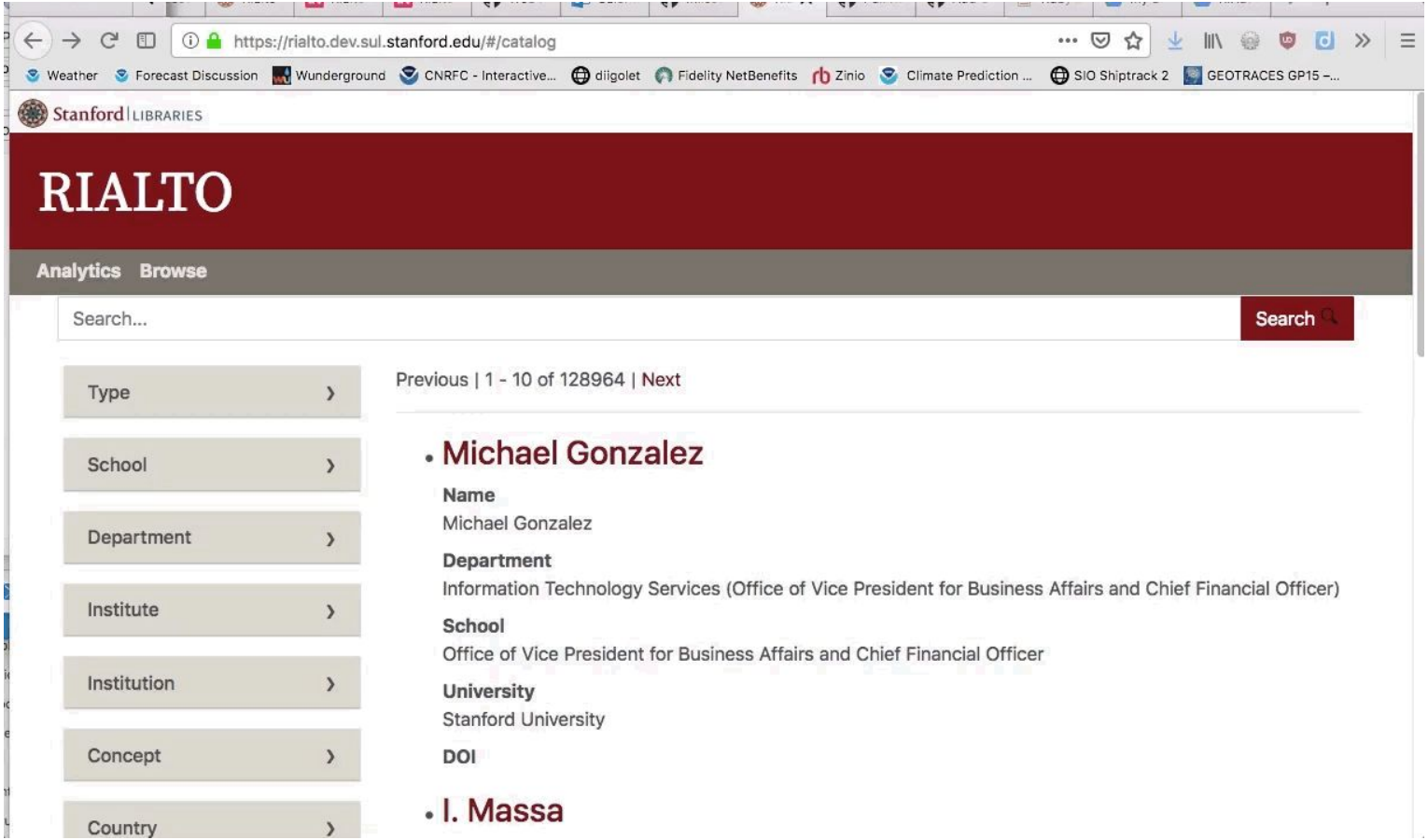
Technical Diagram



Analytics


Stanford | LIBRARIES
RIALTO
Analytics
Browse
Discover patterns of research, publication, and funding
Discover Stanford collaborations
Search
Research trends by topic
Search
Cross-disciplinary research output by topic
Search

Browse


https://rialto.dev.sul.stanford.edu/#/catalog
Weather
Forecast Discussion
Wunderground
CNRFC - Interactive...
diigolet
Fidelity NetBenefits
Zinio
Climate Prediction ...
SIO Shiptrack 2
GEOTRACES GP15 –...
Stanford | LIBRARIES
RIALTO
Analytics
Browse
Search...
Search
Type
School
Department
Institute
Institution
Concept
Country
Previous | 1 - 10 of 128964 | Next
Michael Gonzalez
Name
Michael Gonzalez
Department
Information Technology Services (Office of Vice President for Business Affairs and Chief Financial Officer)
School
Office of Vice President for Business Affairs and Chief Financial Officer
University
Stanford University
DOI
I. Massa

Benefits for stakeholders

- Reports built to specifically identified use cases
- Aggregation by internal only Stanford data
  (e.g. department and school affiliations)
- Stanford only data available that does not exist in similar commercial systems
  (e.g. non public sponsored projects)

Note, most of the data are already available to most of the stakeholders we've talked to, but getting it all, iterating, collating and then filtering by field is a lot of work!

Challenges

- Coverage gaps and incomplete data (coverage varies by subject area)
- No historical data (only current Stanford researchers represented)
- Name ambiguity leads to many false positive and false negative connections between authors, publications and grants
- Research areas are determined at the level of a journal (so all publications in a given journal will have the same research area)
- Out of date unless you constantly harvest new data
- Hard to get performance at web speed with an RDF based data store
- Architectural complexity
- System has many built in external dependencies (e.g. APIs for data)
- Data inconsistencies (usual stuff: misspellings, inconsistent use of fields, duplicate data, bad data, etc.)

Future Possibilities

- APIs, APIs, APIs
- ORCID capture and sync with ORCID profiles
- Additional data sources
- Algorithmic and AI intelligence around the linking of data

Many use cases we considered involved questions around “return on investment” based on one factor or another.

So the question is, how do you measure impact, or “return” on academic research?

Number of publications produced?

Number of grants received?

Number of citations generated?

Number of collaborators?

Altmetrics score?

Students mentored or graduated?

Courses taught?

And over what time periods?

If it takes 5 years for one important paper to become seminal in the field and start generating many citations, should this count as higher impact than a large number of more incremental papers that happen to generate a higher total number of citations over the same timeframe? If so, how do you do this? How does one algorithmically know a work is seminal?

If we build systems that get the answers to these questions wrong, we may end up making strategic decisions or incentivizing work based on flawed metrics.

Facebook measures engagement with numbers of clicks and likes, and success as ad revenue generated. Which leads to misleading and bad content, simply because it optimizes for this.

Let’s avoid the same mistakes in academia by being cautious about what we optimize for when building research intelligence systems.

Having lots of data is not sufficient.

Having lots of linked data is not sufficient.

Having lots of correctly linked data is not sufficient.

Having lots of correctly linked data that is reasonably complete in scope in not sufficient.

Having lots of correctly linked data that is reasonably complete in scope with fancy reports is not sufficient.

What is first required is a good understanding of what the actual metrics are.

In conclusion…

Research intelligence requires more than aggregated data and graphs. It requires careful thought on how we interpret the data and how this interpretation is used for strategic decision making.

So what’s next? Let’s talk!

Peter Mangiafico

pmangiafico@stanford.edu