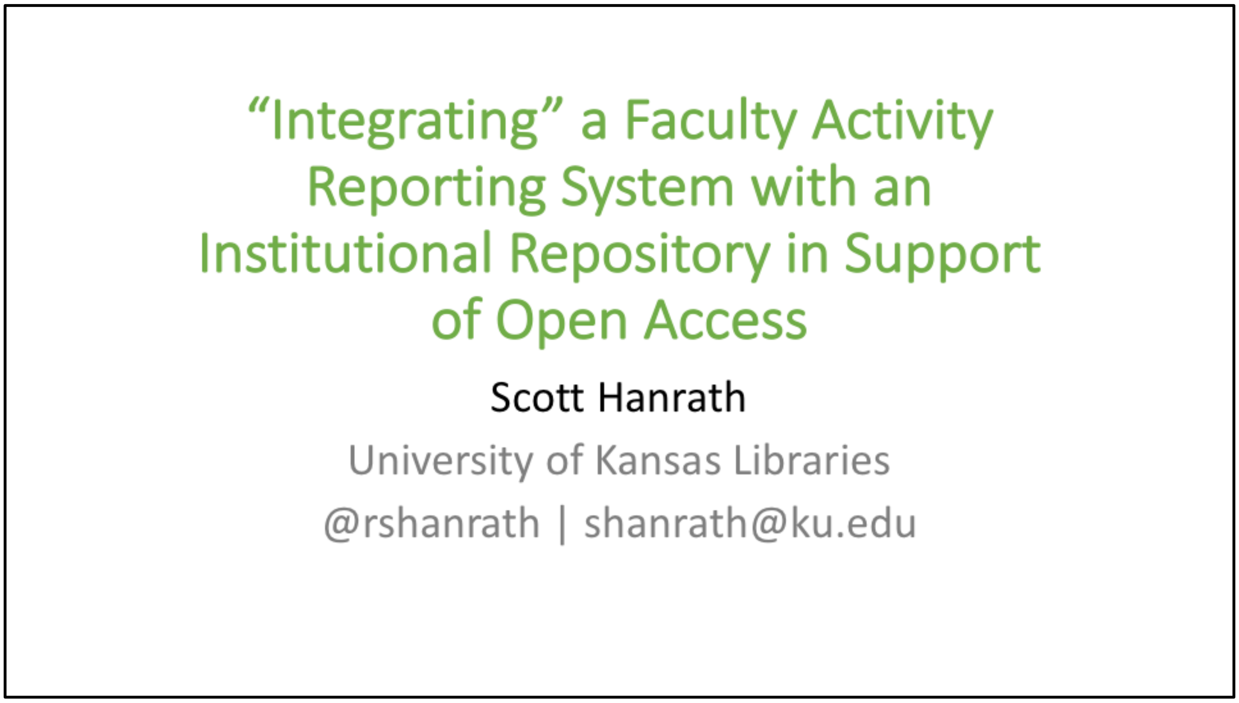
Integrating a Faculty Activity Reporting System with an Institutional Repository in Support of Open Access

Scott Hanrath

University of Kansas Libraries

@rshanrath | shanrath@ku.edu

Need to re-title…



“Integrating” a Faculty Activity Reporting System with an Institutional Repository in Support of Open Access

Scott Hanrath

University of Kansas Libraries

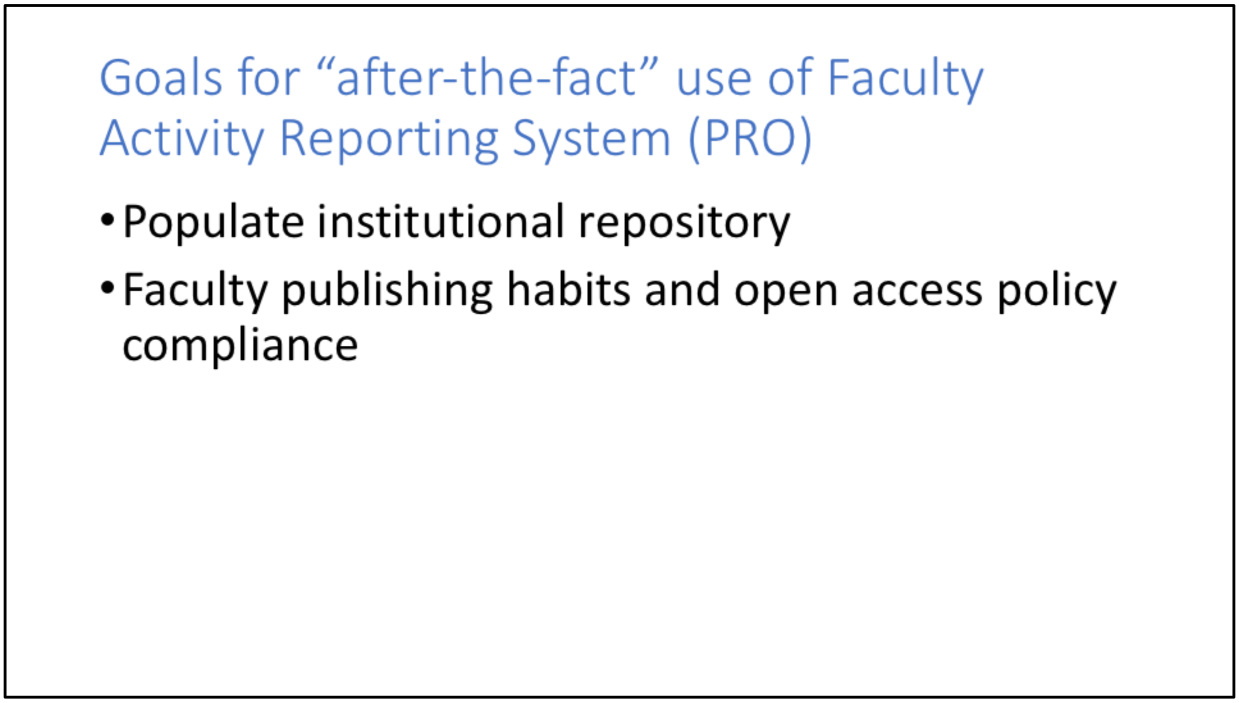
@rshanrath | shanrath@ku.edu

In many ways I'll describe how we've tried to do the best we can with what we've got.

What could we in the Libraries do with a data source provided to us, given that we didn’t have much input into the system and processes around it? Is is “good enough” that we could we avoid setting up parallel processes and systems to meet our needs?

Not exactly a "making lemons out of lemonade" or "blood from a turnip" situation. But neither has our experience been exactly turn-key or an example of a meticulously planned start-to-finish process that worked from goal to solution.

I hope that sharing our experiences will be of value to others making use of institutional data sources, esp. data sources which may already exist and over which you may not have great influence.



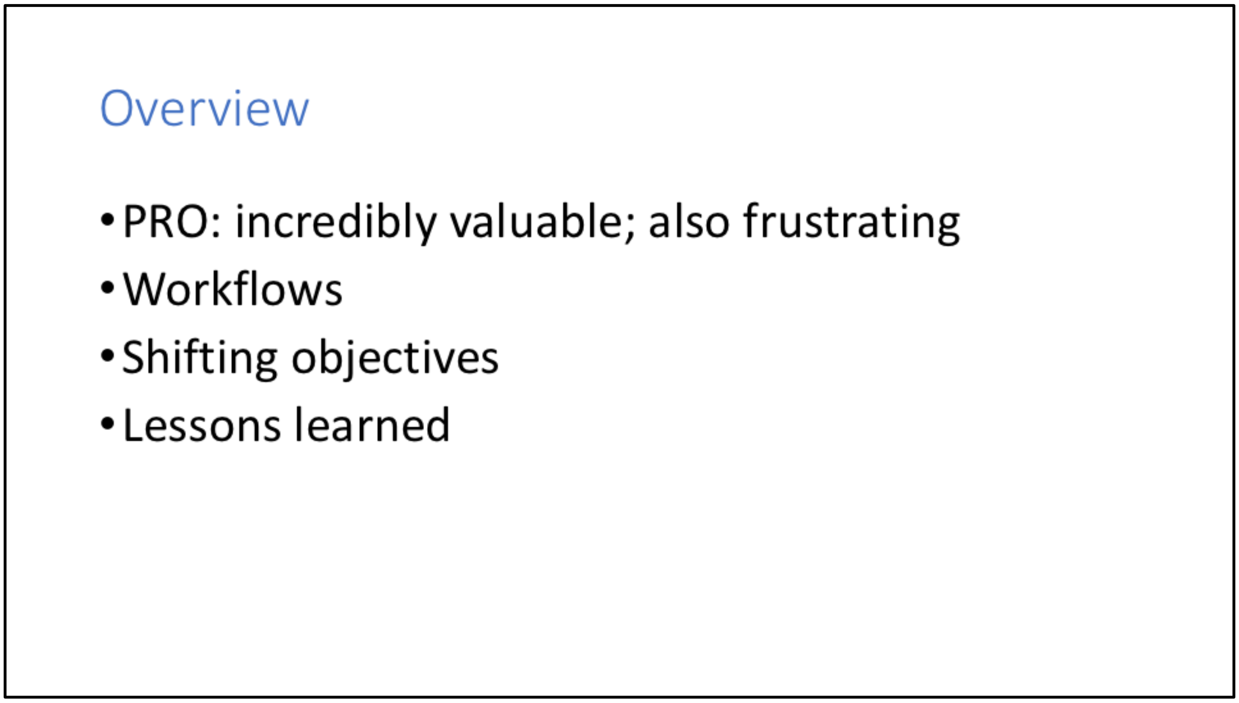
Goals for “after-the-fact” use of Faculty Activity Reporting System (PRO)

- Populate institutional repository
- Faculty publishing habits and open access policy compliance

Specifically, I’ll talk about our “after the fact” use of what at KU we call PRO (Professional Record Online). I say after the fact, because PRO was not created with our goals in mind.

- Populate the IR
    - Specifically with open versions of publications, not metadata-only records or restricted publications

- Gain a sense of faculty publishing habits, at least with regard to the sharing policies of their publishers, and compliance with an open access policy.



Overview

- PRO: incredibly valuable; also frustrating
- Workflows
- Shifting objectives
- Lessons learned

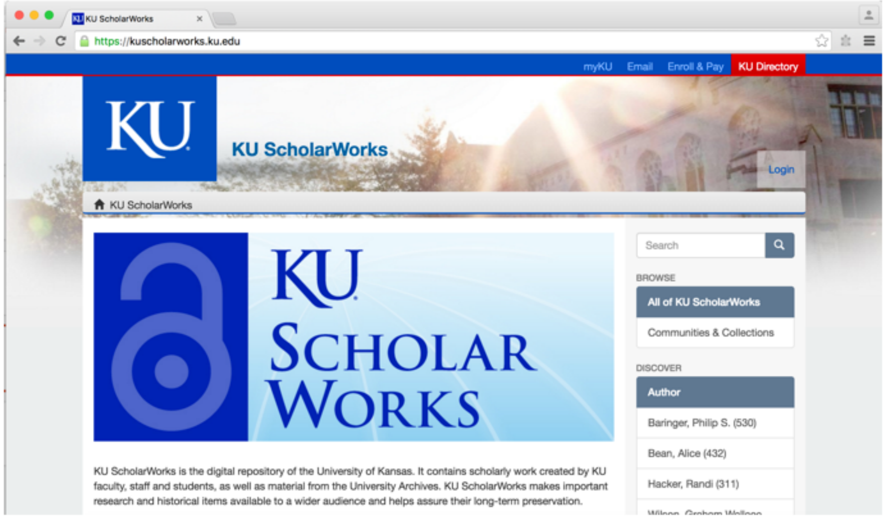
Touch on

- where the Faculty Activity Reporting System has been incredibly valuable and where it has, from our perspective in the Libraries, been occasionally frustrating
- go through our workflows and how they have changed
    - Esp as we have transitioned from the initial rollout to regular, on-going use
- how we shifted our thinking about our primary objective in using the data
- Summarize some lessons learned

Obviously I'm speaking about a team effort here which has draw upon the skills of my talented colleagues, and my take is no doubt colored by my particular perspective as a administrator working with our repository services.

First, a little background.

KU ScholarWorks

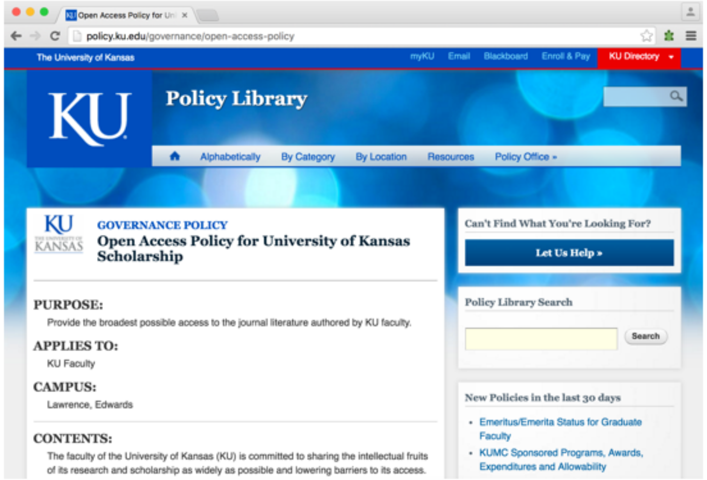


KU ScholarWorks is the digital repository of the University of Kansas. It contains scholarly work created by KU faculty, staff and students, as well as material from the University Archives. KU ScholarWorks makes important research and historical items available to a wider audience and helps assure their long-term preservation.

We've run our institutional repository KU ScholarWorks since 2007. (It's Dspace)

- broad repository of scholarship from KU, journal articles are primary target,
- but also includes Theses and Dissertations,
- materials from the University Archives,
- some (small) data and gray literature

Faculty Open Access Policy



KU KANSAS

GOVERNANCE POLICY

Open Access Policy for University of Kansas Scholarship

PURPOSE:

Provide the broadest possible access to the journal literature authored by KU faculty.

APPLIES TO:

KU Faculty

CAMPUS:

Lawrence, Edwards

CONTENTS:

The faculty of the University of Kansas (KU) is committed to sharing the intellectual fruits of its research and scholarship as widely as possible and lowering barriers to its access.

Can't Find What You're Looking For?

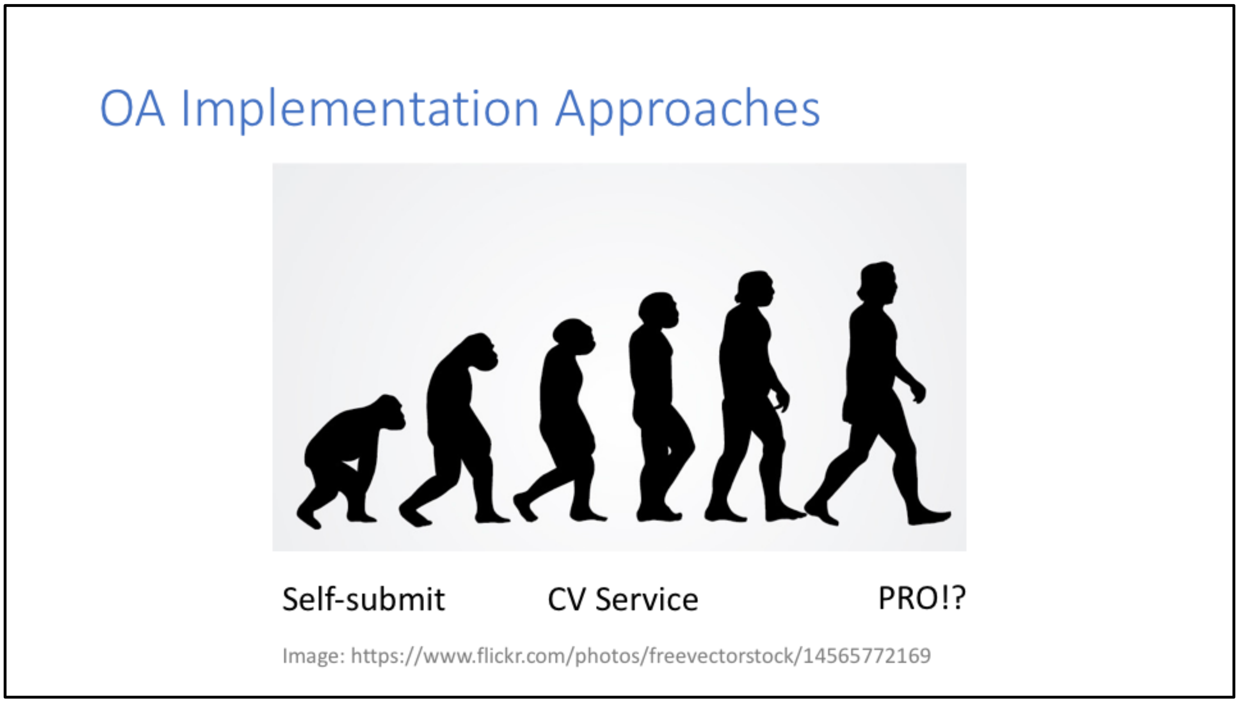
Policy Library Search

New Policies in the last 30 days

- Emeritus/Emerita Status for Graduate Faculty
- KUMC Sponsored Programs, Awards, Expenditures and Allowability

KU Faculty passed an Open Access policy in 2009

- Applies to published journal articles authored by KU Faculty from 2009 forward,
- goal is to make articles openly available.
- Libraries serve as the Provost's designate in implementing the policy, which means working with faculty to get a copy into the repository
- carrot, not stick approach. No "enforcement”.



OA Implementation Approaches

Self-submit CV Service PRO!?

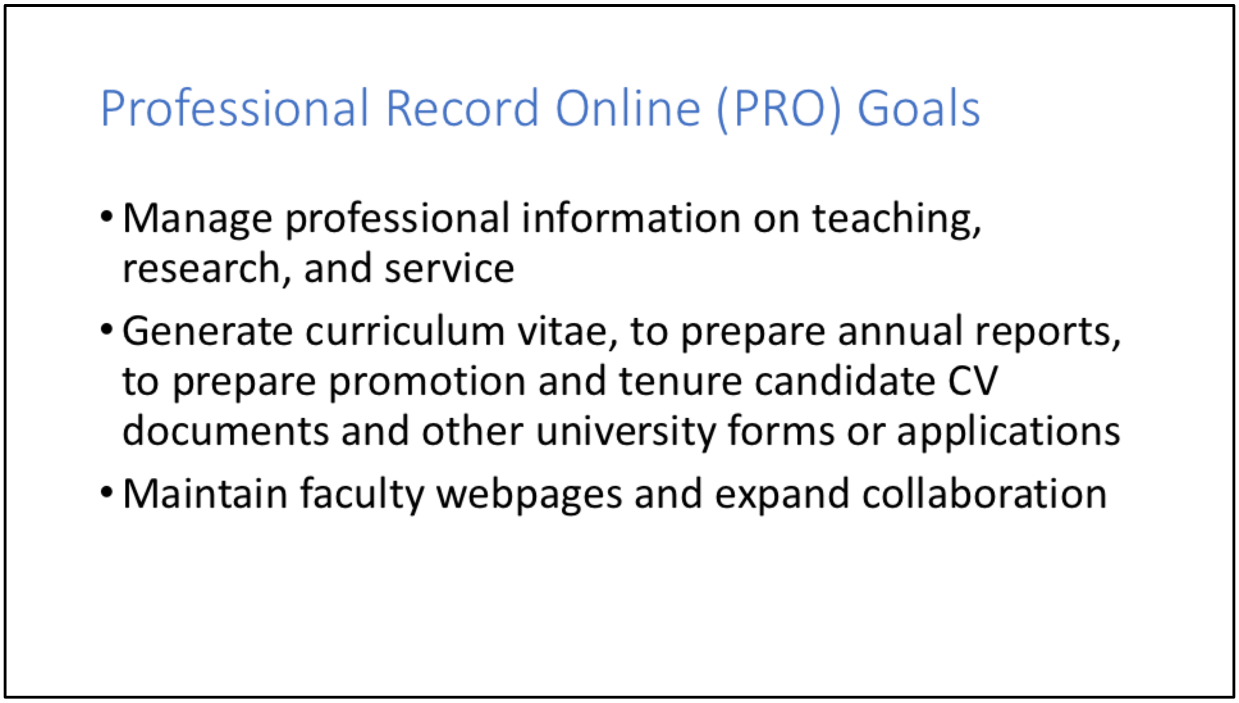
Image: https://www.flickr.com/photos/freevectorstock/14565772169

- self-submit
  - like others, we discovered that faculty members frequently lack the zeal to alter their normal publishing workflows to include depositing a copy of their article in the repository, along with excellent metadata

- * CV Service:
  - outreach based, esp. targeting high profile researchers.
  - Successful up to a point
    - time-consuming
    - snapshot of time

Also some canned search notification attempts

So the Libraries were very interested when the Provost’s office in 2011 announced a Faculty Activity Reporting System initiative as launched as PRO -- Professional Record Online – which would involve, among other things, creating a record all of our faculty publications.



Professional Record Online (PRO) Goals

- Manage professional information on teaching, research, and service
- Generate curriculum vitae, to prepare annual reports, to prepare promotion and tenure candidate CV documents and other university forms or applications
- Maintain faculty webpages and expand collaboration

- Comprehensive “faculty activity”, not just research and publication
- administrative reporting (dept/school/college aggregate reports, annual evaluations), supports the Promotion and Tenure process
- Faculty "webpages" or "profiles"   - delivered via the campus Drupal CMS

"Faculty member" is the primary dimension or unit of analysis. That has implications for us that I'll come back to.

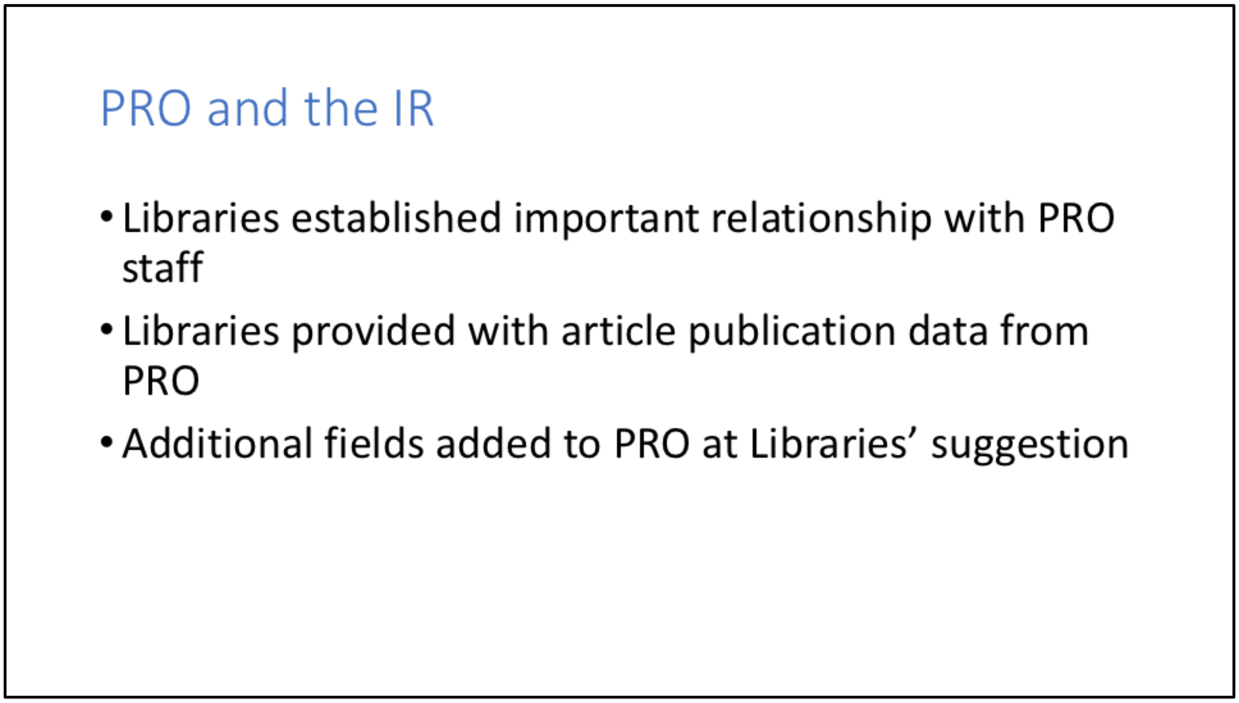
Notice that it does not say "support the open access policy" or "populate the institutional repository”

Centrally-driven by Provost office, the Digital Measures Activity Insight product was selected. There is a campus steering committee (currently has no Libraries representation).

- Implemented by school, discipline area over a couple of years
- CVs were collected and manually entered by PRO staff, then reviewed by faculty
- little to no enhancement of the data
- After the initial CV entry, faculty are responsible for keeping PRO current

    - There is an import option that faculty choose to import references from a BibTeX file
- New faculty will have CVs manually keyed by PRO staff

Some auto-population for courses taught and grants from other systems, but nothing (so far we know) for publications



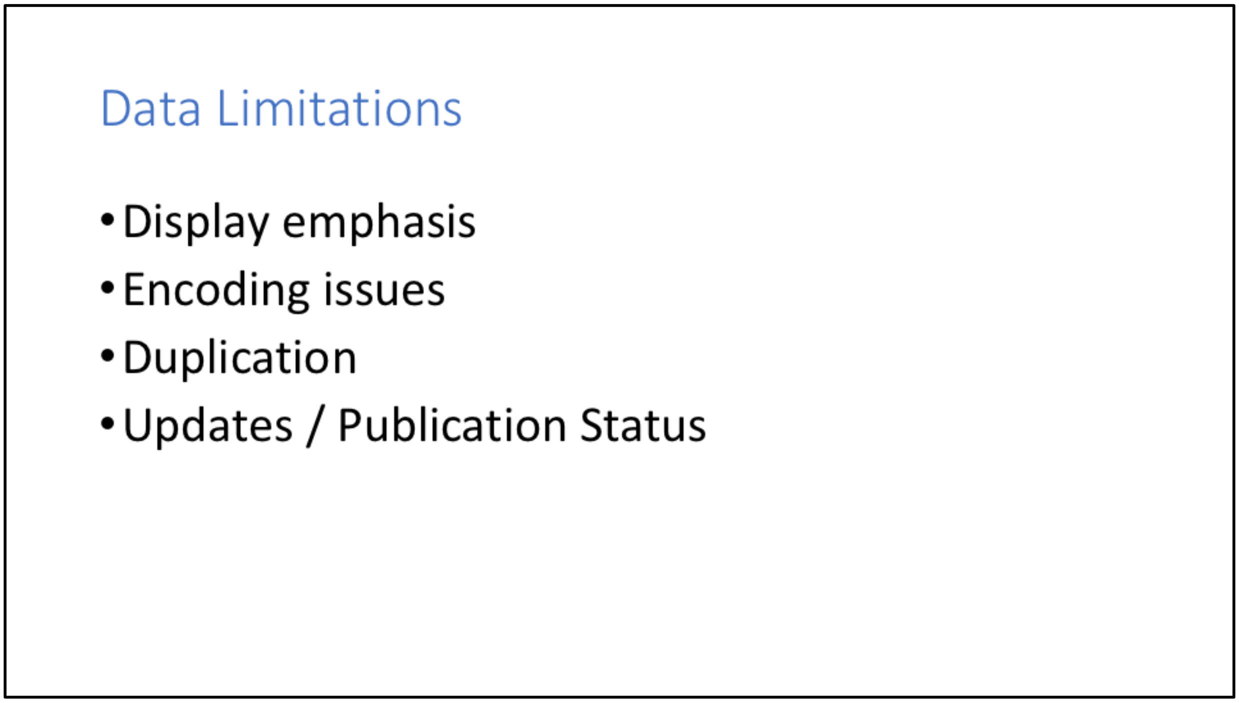
PRO and the IR

- Libraries established important relationship with PRO staff
- Libraries provided with article publication data from PRO
- Additional fields added to PRO at Libraries' suggestion

So from the Libraries perspective, PRO was very much a "pre-existing condition".

We did not have much, if any, input into the product selection or initial roll-out.

- Libraries established important relationship with PRO staff
  - They were receptive and have been good to work with
- Agreed to provide Libraries with article publication data from PRO
  - Some concern about faculty privacy, limited access to the to just articles
  - Rolling export of initial implementation (Excel files) with annual exports going forward (Excel files)
  - The product has an API, but we do not have access
- Agreed to add additional (optional) fields to PRO
  - URL/Handle of articles already in the IR
  - ORCID
  - Upload a document



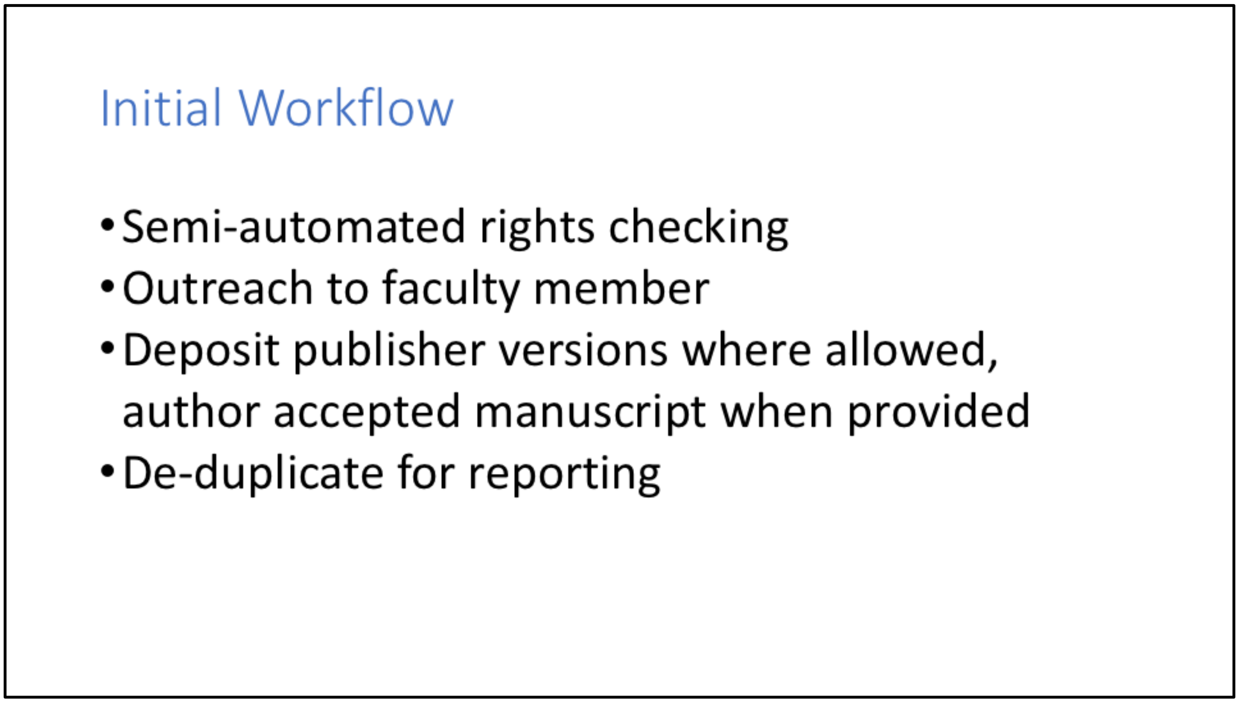
Data Limitations

- Display emphasis
- Encoding issues
- Duplication
- Updates / Publication Status

- Display emphasis
    - what looks good on web page or CV?
    - disciplinary conventions/preferences can obscure data
    - HTML

- Encoding
    - unicode (or lack thereof)
    - HTML, often appearing to be from cut-and-paste
    - Export and transfer process happens through PRO staff using Excel
- Duplication
    - PRO unit is a faculty member, not an article
    - co-authored articles repeated, sometimes with different metadata
    - different export sets in the initial roll-out contained the same faculty members (joint or overlapping appointments)
    - sometimes get DOIs or other helpful identifiers, but not required or intentionally added
- Updates / Publication status
    - PRO allows for in-process or under review articles to be entered (to support annual evaluations, P&T processes)

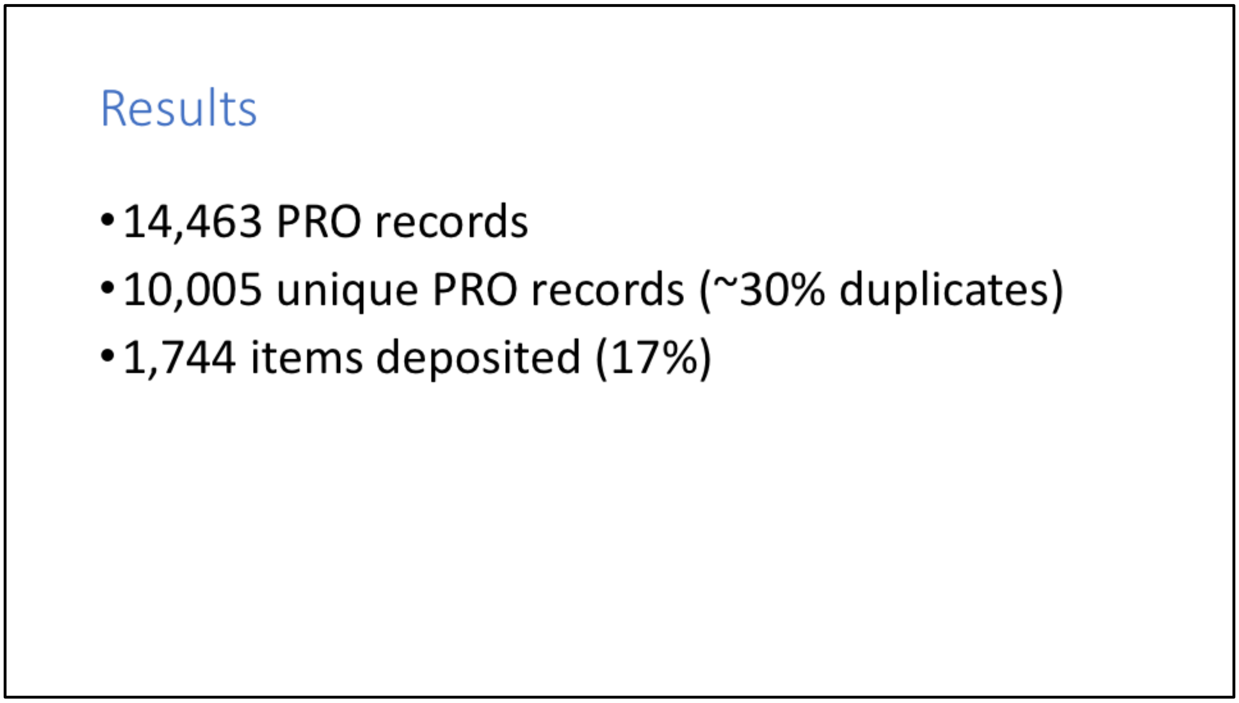
- metadata can change
- no created or last-modified dates available from PRO
- Some faculty add data in something like real-time, but many others do so only as they are required to do so for annual evaluation or P&T processes
- The end result is that knowing when PRO contains a “new” article is tricky



Initial Workflow

- Semi-automated rights checking
- Outreach to faculty member
- Deposit publisher versions where allowed, author accepted manuscript when provided
- De-duplicate for reporting

- Semi-automated rights checking, via SherpaRomeo API lookup based on Journal Title
  - SR API found entries for around 54% of articles
  - varies by discipline but looking at a set of Pharmacy, Engineering, Humanities, and Education articles, found hits in **48% to 55% of cases**
  - Pretty good: but finding half, means not finding half
  - manual look ups where no matches where found
- Email summary to faculty members
  - all articles, publisher policies
  - request for author manuscripts, offer of assistance
  - very low response rate to these emails
  - abandoned and shifted to summary reports to department chairs with offer of follow-ups
- Deposit publisher versions where possible, deposit author-accepted manuscript when provided
- De-duplicate articles for reporting
  - used Open Refine's clustering features to identify duplicate titles
  - manual process, but still time-saving

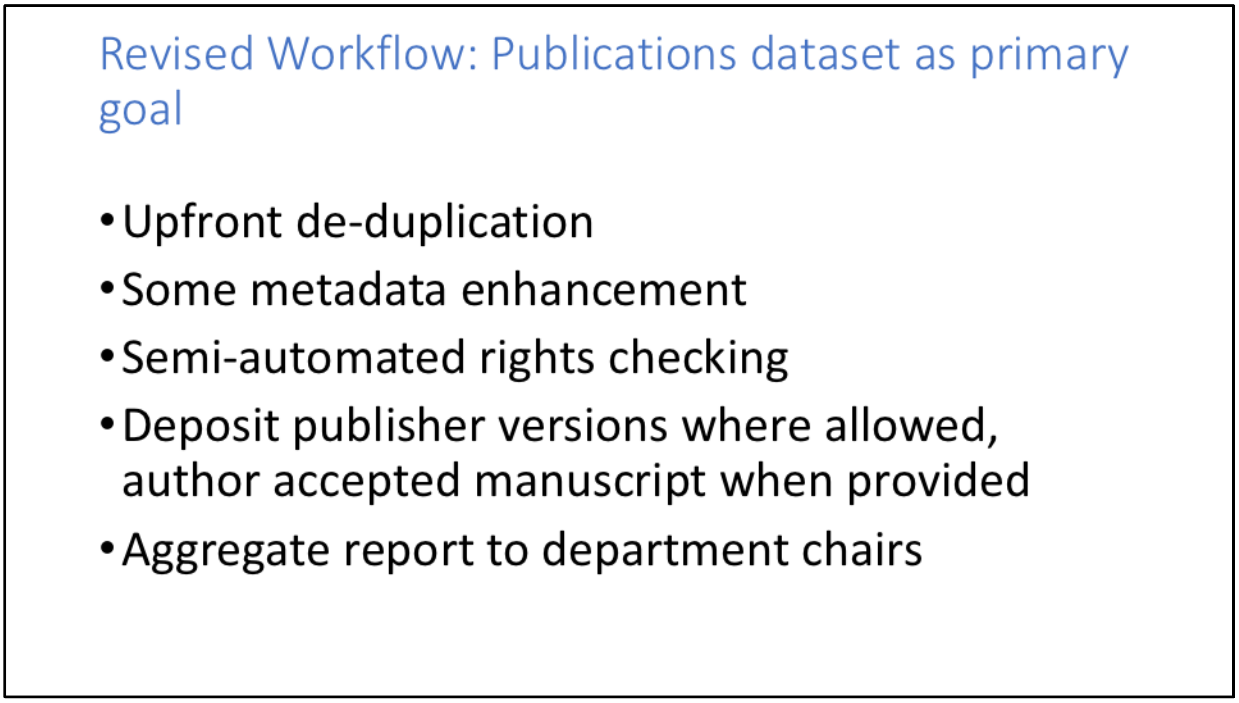


Results

- 14,463 PRO records
- 10,005 unique PRO records (~30% duplicates)
- 1,744 items deposited (17%)

Looking 2009-2014 period -- OA Policy timeframe (do have article data going back much further)

- 14,463 PRO records
- 10,005 unique PRO records (~30% duplicates)
  - likely says as much about the process as the data-as-entered
  - should improve
- 1,744 items deposited (17%)
  - the 17% figure could be considered an open access compliance rate
  - but on-going discussions about what the best number is
    - should it be based on unique articles, like this number?
    - what about the notion of faculty member compliance?
    - Should unit be author-article?
    - what about departments/school/college level?
    - Different levels of reporting will require slicing the data differently



Revised Workflow: Publications dataset as primary goal

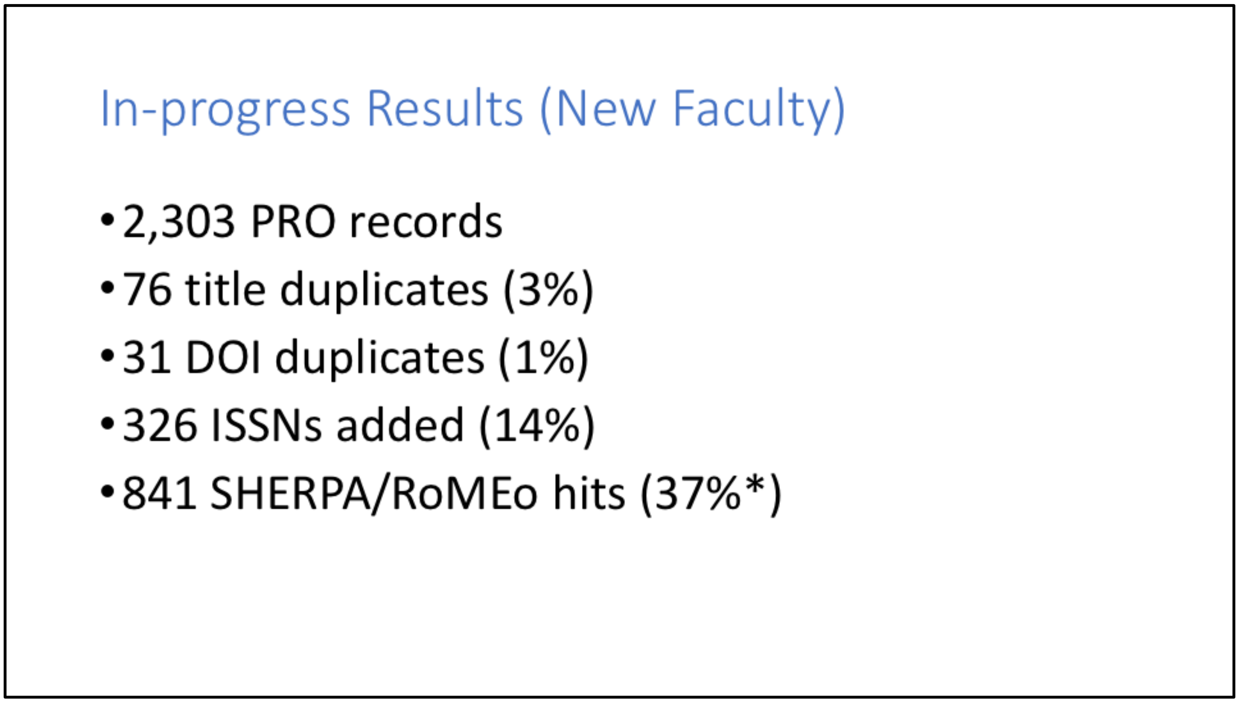
- Upfront de-duplication
- Some metadata enhancement
- Semi-automated rights checking
- Deposit publisher versions where allowed, author accepted manuscript when provided
- Aggregate report to department chairs

New Goal: treat a cleaned dataset of publications as a primary, first-class asset

- not just a means of getting content into the repository
- account for reporting needs upfront
- We are not interested in having records for non-open articles in the repository
- But we do want to know what we don't have
- Variations on this distinction seem to be becoming important as recent discussions about vendor with publishers and aggregators has show

- Do de-duplication first
  - smaller problem because we should only be getting co-authored papers, not faulty export duplicates
  - Use DOI and normalized titles (lowercased, trimmed, stripped of punctuation) as keys
    - look at using Levenstien distance matching next.
  - Done via Python scripts rather than Refine to more easily replicate
- Metadata enhancement
  - for items with DOIs, retrieve titles, journal names, authors, ORCID, and ISSNs from CrossRef API
  - Rights checking

    - use ISSNs in addition to Journal Titles for SherpaRomeo checks
- Instead of lots of email correspondence with lots of faculty members, provide an aggregate annual report to department chairs, offering further information and assistance if they are interested



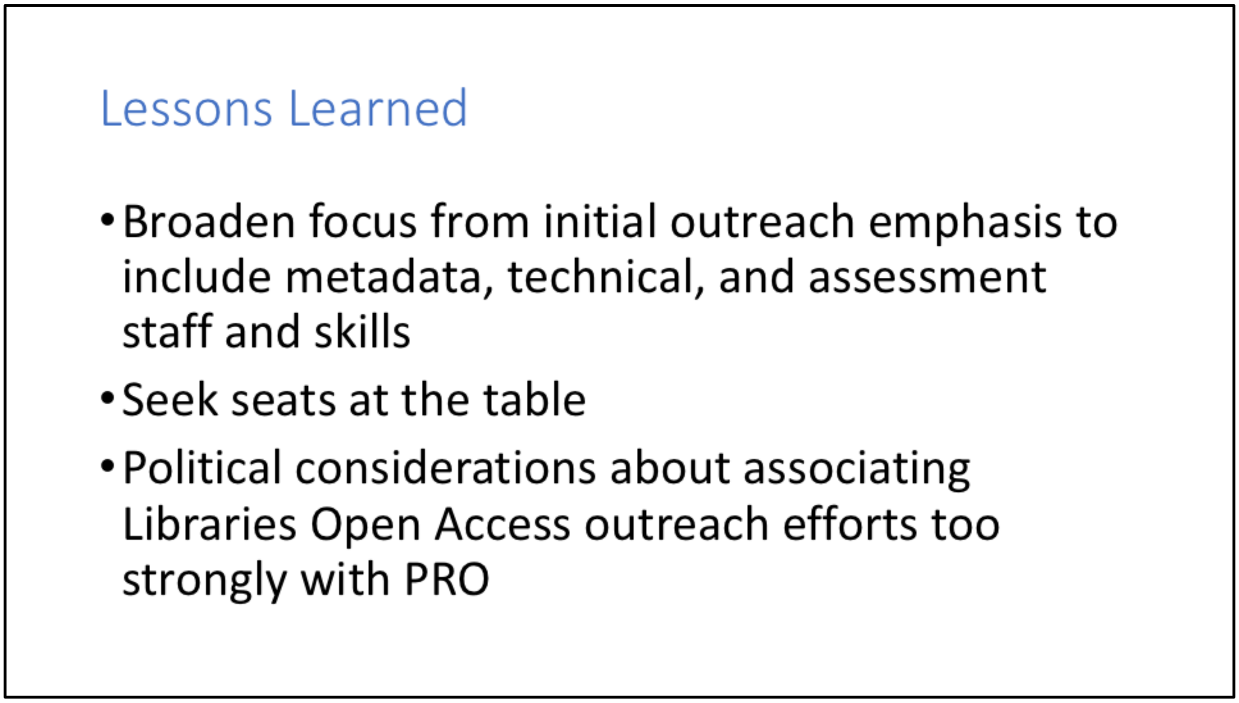
In-progress Results (New Faculty)

- 2,303 PRO records
- 76 title duplicates (3%)
- 31 DOI duplicates (1%)
- 326 ISSNs added (14%)
- 841 SHERPA/RoMEo hits (37%*)

New workflows and scripts are still in development, but here's what came out of a recent test run of the new faculty publication export we received.

- 2,303 PRO records
- 76 title duplicates (3%)
- 31 DOI duplicates (1%)
- 326 ISSNs added (14%)
- 841 SHERPA/RoMEo hits (37%) – this will get better as our process is refined. Kinks in the process.

Cleaner data with ongoing export processes that with the original "from the beginning of time" process.



Lessons Learned

- Broaden focus from initial outreach emphasis to include metadata, technical, and assessment staff and skills
- Seek seats at the table
- Political considerations about associating Libraries Open Access outreach efforts too strongly with PRO

Staffing shift from focus on outreach. The larger team contributing to the effort includes metadata, technical, and assessment staff and skills

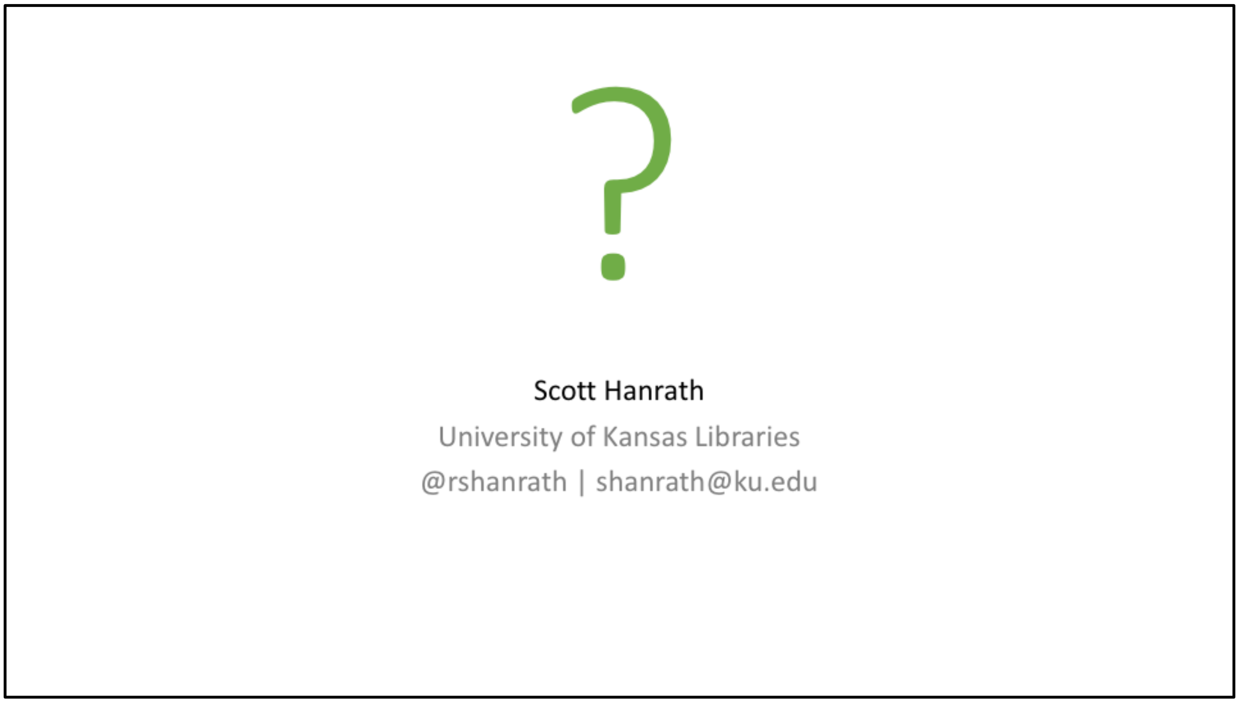
- seeking efficiency and greater automation, requires different tool sets
- looking to future efforts (e.g, like harvesting and/or pushing to ORCID), metrics
- some tensions here, but healthy discussions
  - for faculty who do upload documents to PRO (which are few, but important to us) need to weight responding quickly vs waiting until we have all the data for the reporting period to deal with the data more efficiently and avoid re-processing the data multiple times
  - So far the volume of uploaded documents is low

Seek seats at the table

- through engagement with PRO staff, were able to include ORCID, add upload feature
- continue to refine the access we have to data to better meet our needs
- Systems don’t speak to each other and that is limiting
- Libraries have an opportunity show our value to greater campus efforts as faculty reporting system evolves

Politics: PRO isn't universally embraced and loved by faculty.

- Some faculty consider it intrusive; some consider the sharing of data a breach of rights and privacy.
- Not a natural part of publishing workflow for all faculty and tends to be updated as required for university processes
- Convenient for us, esp. opportunity to upload article manuscripts, but there is a desire to keep OA outreach efforts distinct from faculty activity reporting to avoid inappropriate associations between the two activities



?

Scott Hanrath
University of Kansas Libraries
@rshanrath | shanrath@ku.edu