arXiv IT Update

Simeon Warner
(Cornell University Library)

Prepared for arXiv SAB and MAB meetings
NYC, 2015-09-21/22

arXiv.org

arxiv.org/stats/monthly_downloads

Cornell University Library

We gratefully acknowledge support from the Simons Foundation and member institutions

arXiv.org > stats

Login | Logout

arXiv monthly download rates [CSV]

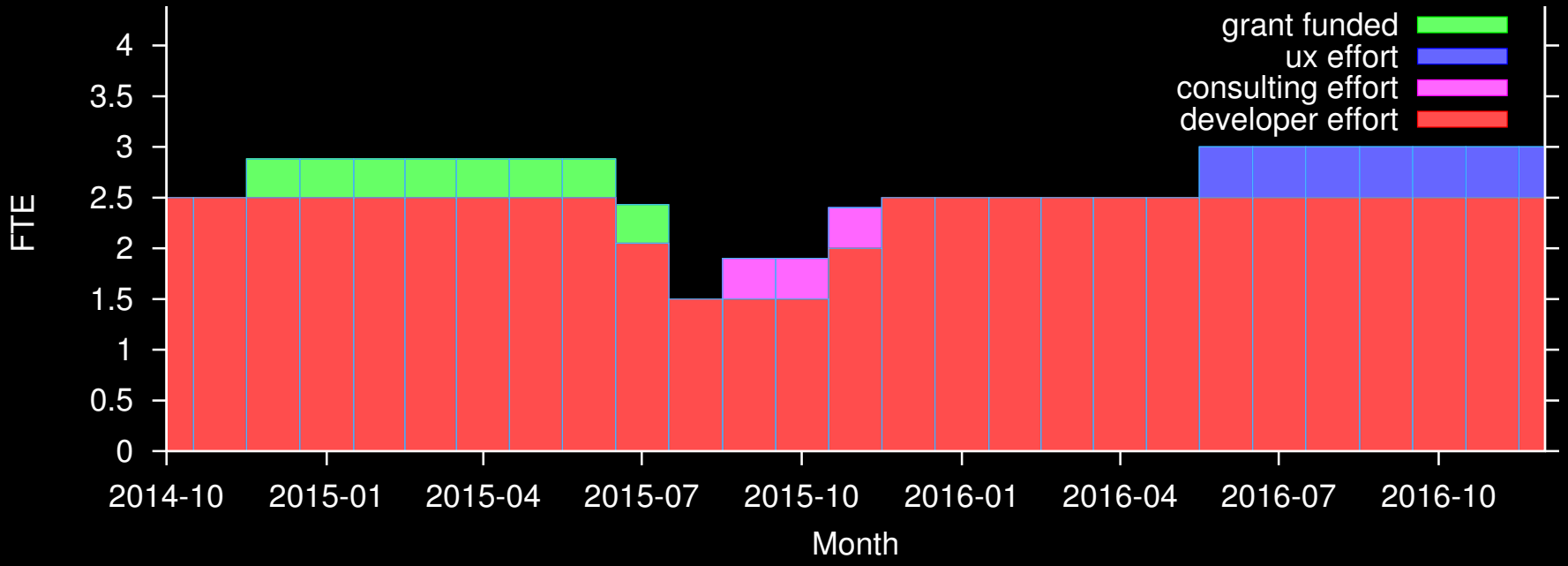
Continued successful operation

arXiv continues to work well, serve a large and growing number of users, and deal with an ever higher submission rate.

Previous estimate of 1-1.5FTE IT effort to “keep the lights on” still seems reasonable.

IT staffing

- 2.5 FTE development effort budget
- 0.25 FTE-year Sloan-Hypothesis grant effort
- Joe Skovira left in July, job posted August, consultant back-fill
- Hope to have 0.5 FTE UX sometime in 2016
- Martin transitioning to 100% as arXiv Lead Developer



2014 arXiv Roadmap – CU

https://confluence.cornell.edu/display/culpublic/2014+arXiv+Roadmap

Technical

Items are listed in approximate priority order.

Implement secure communication (SSL/https) for all login and account interactions on arXiv - All interactions with the submission, modera
when users login on unsecured wifi. Completed

Improve tools and interfaces to support mo ow moder
work of moderators as quick and convenient as e SAB. Co

Add automatic classification checks to subr classifier
submission system for new articles, including p m classifie
further work being done to refine notifications.

Provide summary documentation of the arX gies and a

Do category aliasing for cs/math/stat and new c tegory re
primary<->id correspondence. In the past aliase and docu
in cs/math/stat not started.

Add automatic overlap detection comparing ors and n
existing corpus and staged submissions. Devel these wa
detection not started.

Investigate and expand the id range to `yymm` => 8135),
submissions will get yymm.nnnnn identifiers, ex

Add support for ORCID and other author ide er interope
implementing storage of affiliation in the profile

Ingest data from discontinued Data Conserv ta Conser
DOIs from EZID to ancillary files thus making them citeable. Work in progress, data has been copied from Data Conservancy.

Submitter email addresses should not be harvestable - Currently submitter email addresses are stored in the metadata (.abs) files and in the
complaints from users where collaborating services have made the emails available. We should instead keep the submitter email addresses only

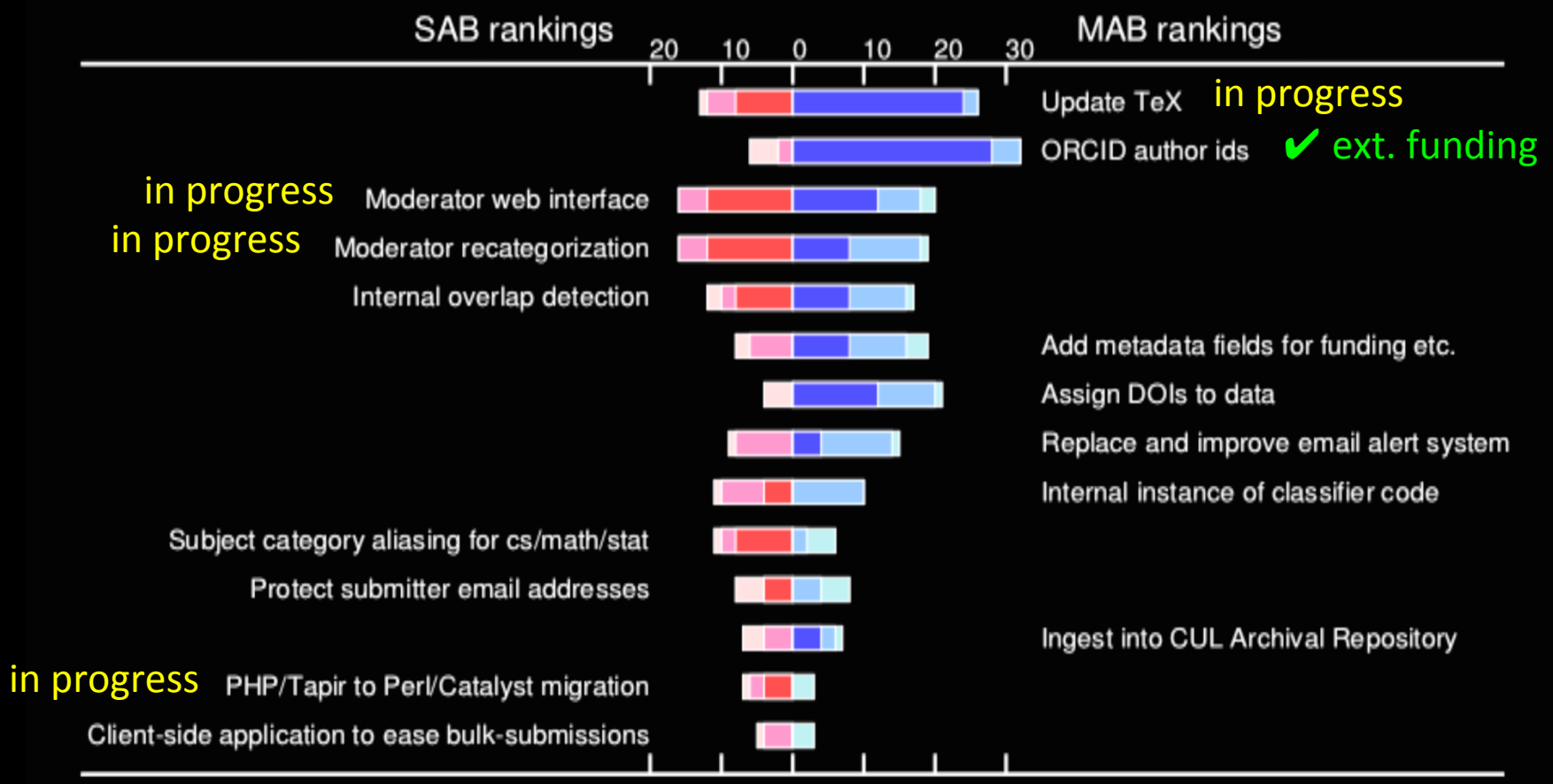
Replace email alert system for better maintainability, to allow easy subscribe/unsubscribe, and for flexible options - Replace email alert

2014, September-December

- Added title similarity checks to reduce numbers of duplicate submissions
- Changed identifiers to 5-digits in month starting arXiv:1501.00001

Also work on moderator tools, basic ORCID support, and migration of Data Conservancy data.

2015-01 SAB & MAB prioritization



2015 arXiv Roadmap – CU

https://confluence.cornell.edu/display/culpublic/2015+arXiv+Roadmap

Technical

Items are listed in approximate priority order and may be adjusted based on ongoing discussions with the Scientific and Member Advisory Boards.

Add automatic hold-on-submit based on flags from classifier - There are some flags from the classifier (such as line numbers, two copies) that should result the hold status for admin action when submitted. Moderators should not get any notification of such submissions until admins have had a chance to resolve obvi

Add ORCID author identifier support - We would providing institutional statistics for member organiza accounts via the ORCID OAuth process. Completed

Improve moderator web interface, add personal significantly extended and improved in 2014. Work v

Ingest data from discontinued Data Conservancy repository. In order to preserve access to datasets u and http://blogs.cornell.edu/dsps/2013/06/14/arxiv-d

Allow moderators to recategorize articles via the admins as intermediaries. Moderators should be abl

Develop and integrate internal automatic overlap warnings for administrators and moderators based c

Subject category aliasing for cs/math/stat - There these require extra work because there are pre-070 primary archive to identifier prefix correspondence w should instead work out and document procedures f

Update, reorganize and better document the TeX need to put effort into updating our TeX installation, within the current development team. We need to up 2011), and also update our ghostscript installation.

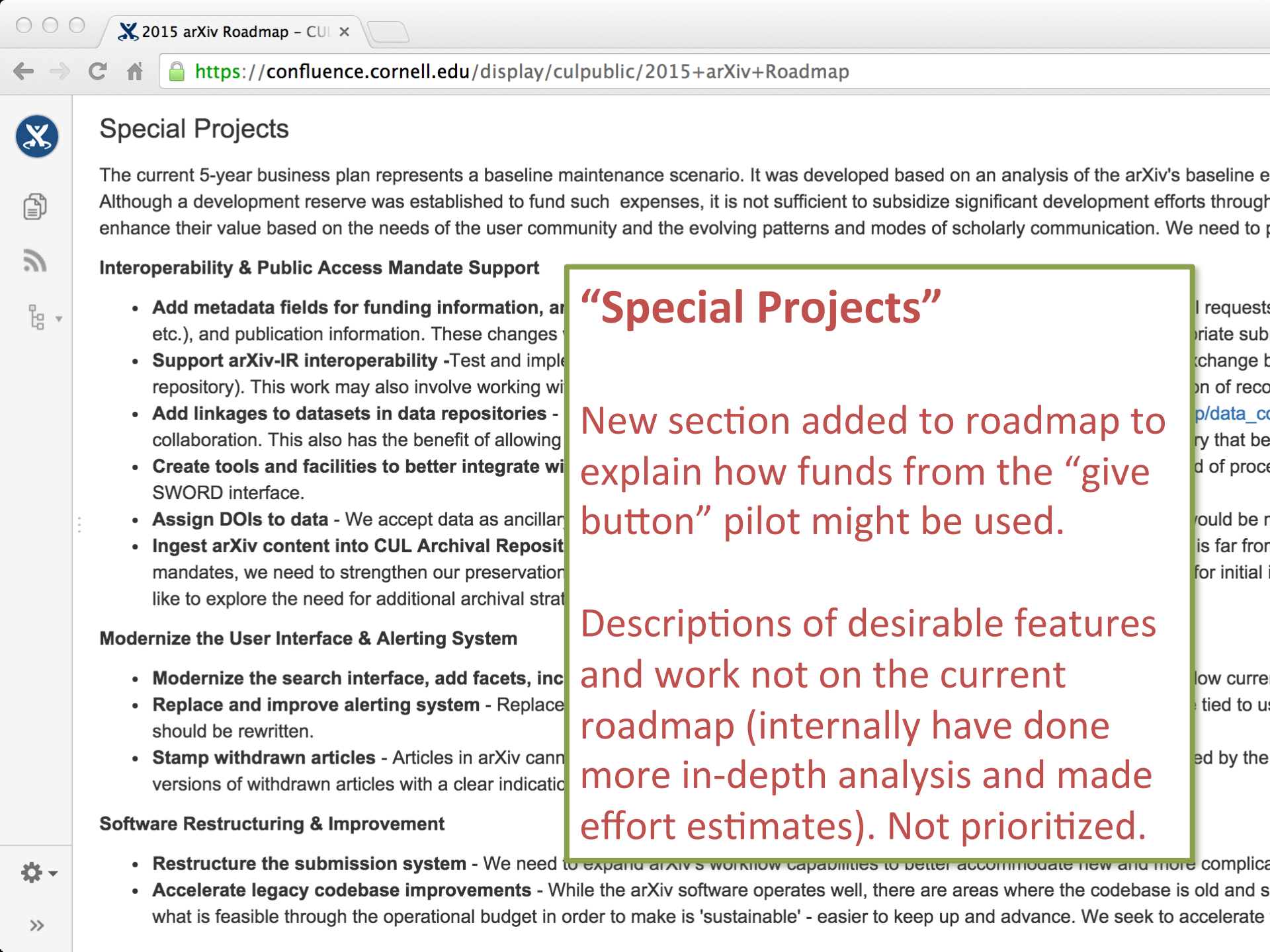
Migrate functions away from old PHP/Tapir codebase and into Perl/Catalyst - We have been gradually replacing old PHP/Tapir code with more maintainabl code.

Develop and integrate internal instance of classifier code - We should integrate the classifier code into the arXiv production system rather than using API to research machine. This was agreed by the SAB on 2013-09. Work was postponed in summer 2014 to allow quick initial deployment and to allow Paul Ginsparg uncertainties here because we haven't seen Paul's code and perhaps when we do we will want to rewrite some of the client-side code to reflect that understandi

2015, January-September

- Completed migration of Data Conservancy data
- Completed basic ORCID integration
- Added auto-holds based on classifier flags

Ongoing work on moderator tools, TeX system, and old code migration.

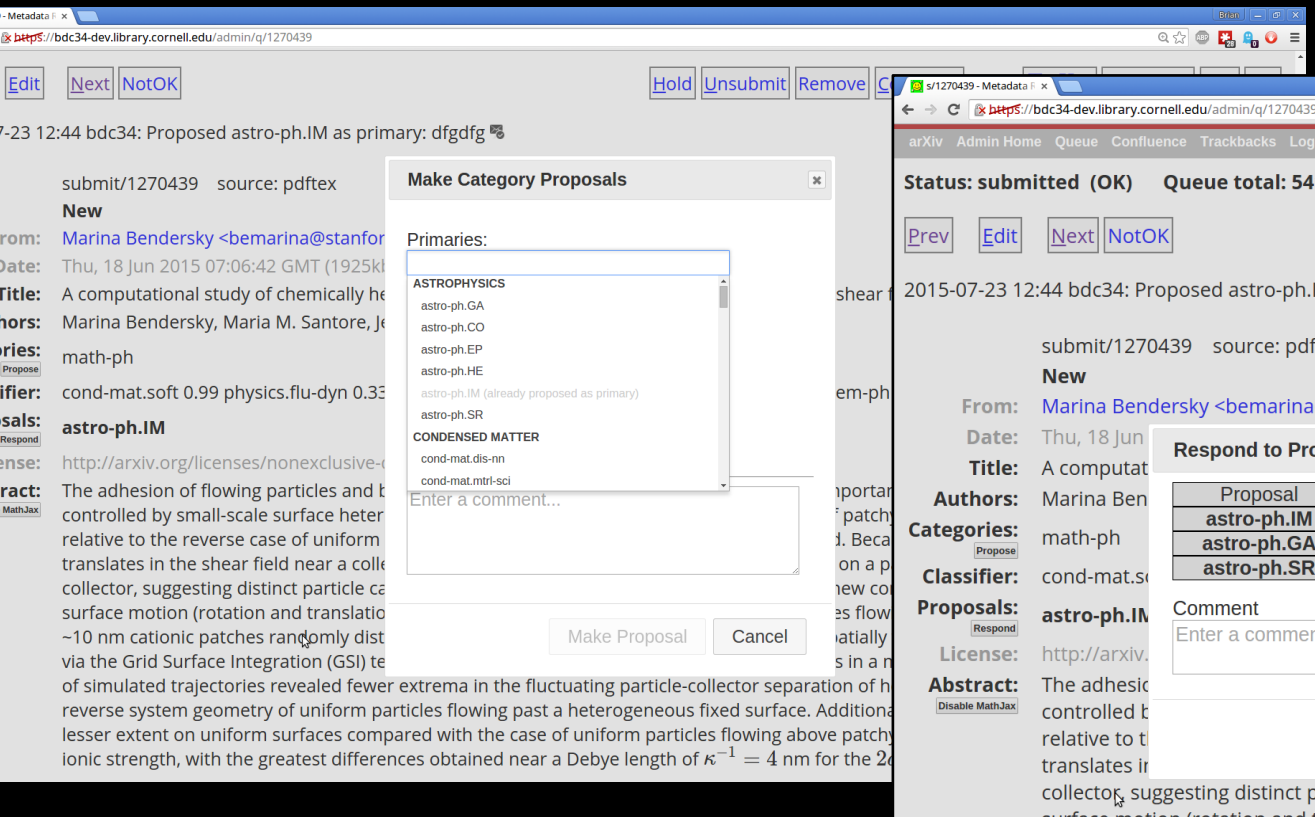

2015 arXiv Roadmap – CUL
https://confluence.cornell.edu/display/culpublic/2015+arXiv+Roadmap
Special Projects
The current 5-year business plan represents a baseline maintenance scenario. It was developed based on an analysis of the arXiv's baseline e
Although a development reserve was established to fund such expenses, it is not sufficient to subsidize significant development efforts through
enhance their value based on the needs of the user community and the evolving patterns and modes of scholarly communication. We need to p
Interoperability & Public Access Mandate Support
Add metadata fields for funding information, ar
etc.), and publication information. These changes
Support arXiv-IR interoperability -Test and imple
repository). This work may also involve working wi
Add linkages to datasets in data repositories -
collaboration. This also has the benefit of allowing
Create tools and facilities to better integrate wi
SWORD interface.
Assign DOIs to data - We accept data as ancillar
Ingest arXiv content into CUL Archival Reposit
mandates, we need to strengthen our preservation
like to explore the need for additional archival strat
Modernize the User Interface & Alerting System
Modernize the search interface, add facets, inc
Replace and improve alerting system - Replace
should be rewritten.
Stamp withdrawn articles - Articles in arXiv cann
versions of withdrawn articles with a clear indicatic
Software Restructuring & Improvement
Restructure the submission system - We need to expand arXiv's workflow capabilities to better accommodate new and more complica
Accelerate legacy codebase improvements - While the arXiv software operates well, there are areas where the codebase is old and s
what is feasible through the operational budget in order to make is 'sustainable' - easier to keep up and advance. We seek to accelerate
"Special Projects"
New section added to roadmap to explain how funds from the "give button" pilot might be used.
Descriptions of desirable features and work not on the current roadmap (internally have done more in-depth analysis and made effort estimates). Not prioritized.

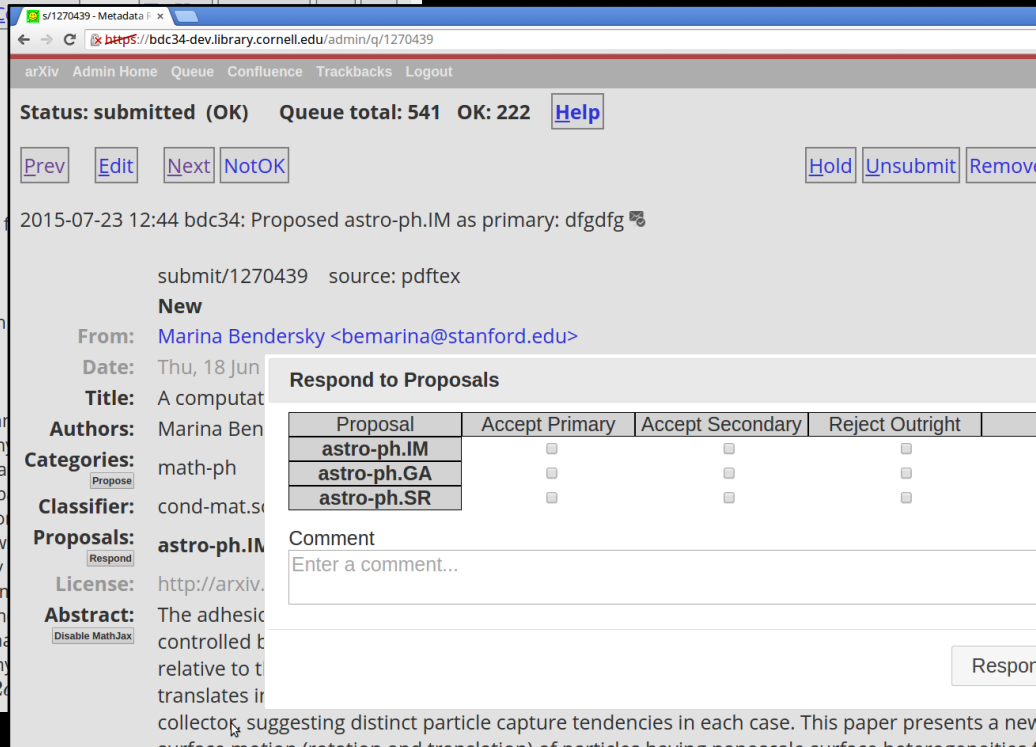
Process and prioritization

- Two-week work chunks (“sprints”)
- For each sprint Team+Chris+Simeon meet to review and plan work
  - Transitioning to Martin leading these meetings
  - Chris advocates for moderation and admin related issues
  - Simeon advocates for infrastructure and maintenance issues
  - Follow roadmap for major work (thus incorporating SAB and MAB input)
- Martin sends regular updates to interested parties (Greg, Joe, Paul, Oya, etc.)
- Process working well, even if development work takes longer than we would like

Reclassification

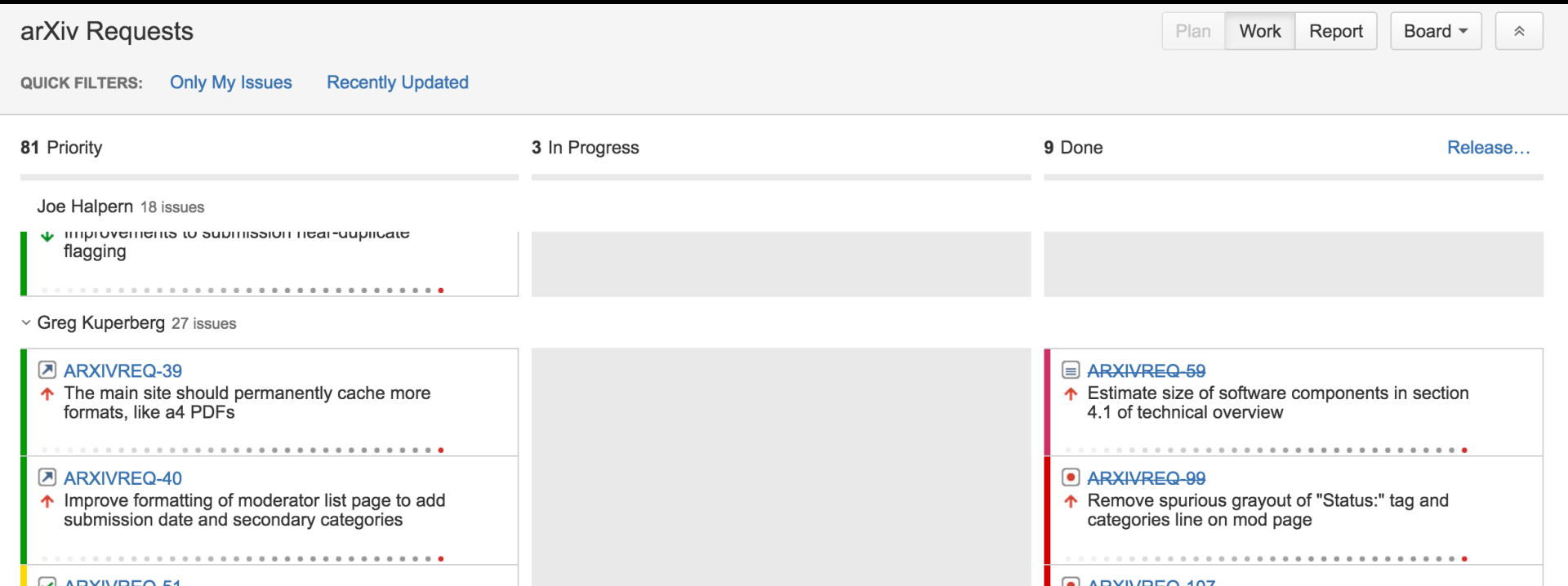
- Focus of current development work
- Phase 1: mods propose, admins deal with responses
- Phase 2: mods accept/reject on web





Feature requests - ARXIVREQ

- Have implemented email alias and issue tracker for IT requests and questions (primarily for Paul, Greg, Joe)
- Helps us address issues of follow-up and transparency
- ... but still require developer effort to implement



Proposals

1. Shut down mirrors
   - mirrors are a significant development impediment, little use
   - need to keep LANL presence so that xxx.lanl.gov URIs resolve (+independent backups)
   - support A4 paper on main site (also requested by Greg)
2. http -> https everywhere
   - web very rapidly moving to all SSL
3. Search collaboration with Cornell CS
   - build on partnership for full-text search
   - lessen reliance on CS grad student for production but increase flexibility – have API switch between local copy and expt. copy
   - possibly expand collaboration community (funding?)
   - tie to proposed user study re. search and browse

Major rewrite?

- Code base old, Perl not ideal and contributes to hiring difficulties
- Need to resolve authority and direction issues
- Need to understand/accept stable system during work
- Need to break into tractable chunks
- While resulting system will be easier to maintain and update, that effort will not be vastly reduced. More about ability to move forward
- Major undertaking, perhaps 8-12 FTE-years? (developers, UX, project management) on top of staff to keep current arXiv going
- For grant we'd need to show benefit beyond arXiv:
  - open code (do everything on github)
  - generalized and reusable components
  - make arXiv data more available/reusable/interoperable
  - perhaps couple with expansion to non-TeX disciplines?