

arXiv IT Update

Simeon Warner
(Cornell University Library)

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

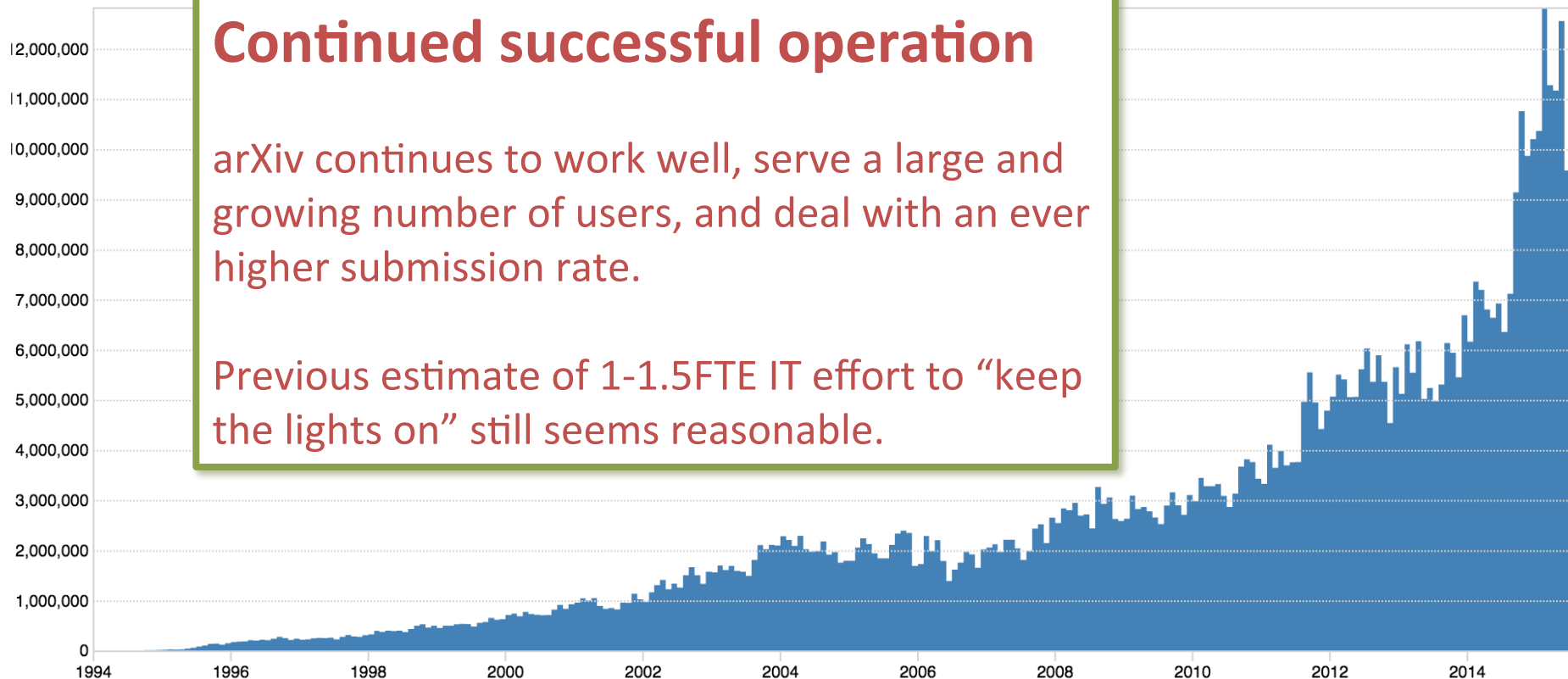


arXiv monthly download rates [CSV]

Continued successful operation

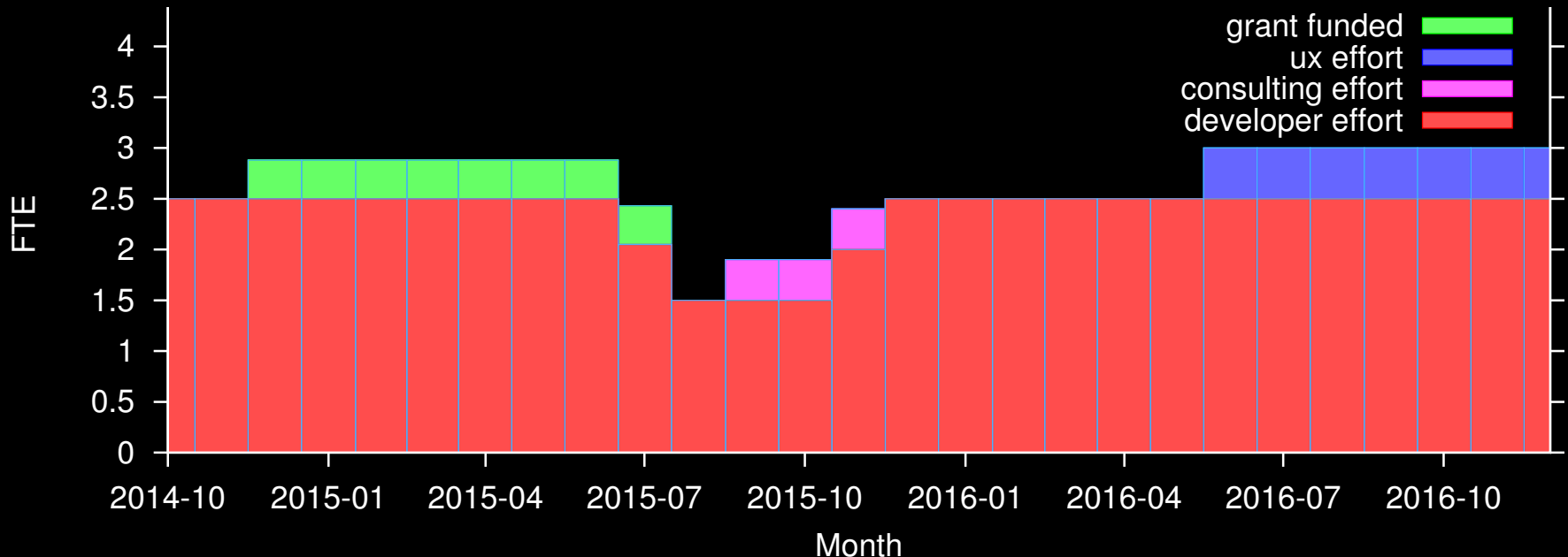
arXiv continues to work well, serve a large and growing number of users, and deal with an ever higher submission rate.

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



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, moderation, and account management systems should be secured. All interactions with the submission, moderation, and account management systems should be secured when users login on unsecured wifi. **Completed**

Improve tools and interfaces to support moderation - Make the work of moderators as quick and convenient as possible.

Add automatic classification checks to submission system for new articles, including post-publication - **Further work being done to refine notifications.**

Provide summary documentation of the arXiv

Do category aliasing for cs/math/stat and new categories. Do primary<->id correspondence. In the past aliases were used in cs/math/stat **not started.**

Add automatic overlap detection comparing existing corpus and staged submissions. Develop a system for **detection not started.**

Investigate and expand the id range to yymm - **submissions will get yymm.nnnnn identifiers, experimental**

Add support for ORCID and other author identifiers - Implementing storage of affiliation in the profile

Ingest data from discontinued Data Conservancy - 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 metadata files. We should instead keep the submitter email addresses only in the metadata files. **Complaints from users where collaborating services have made the emails available.**

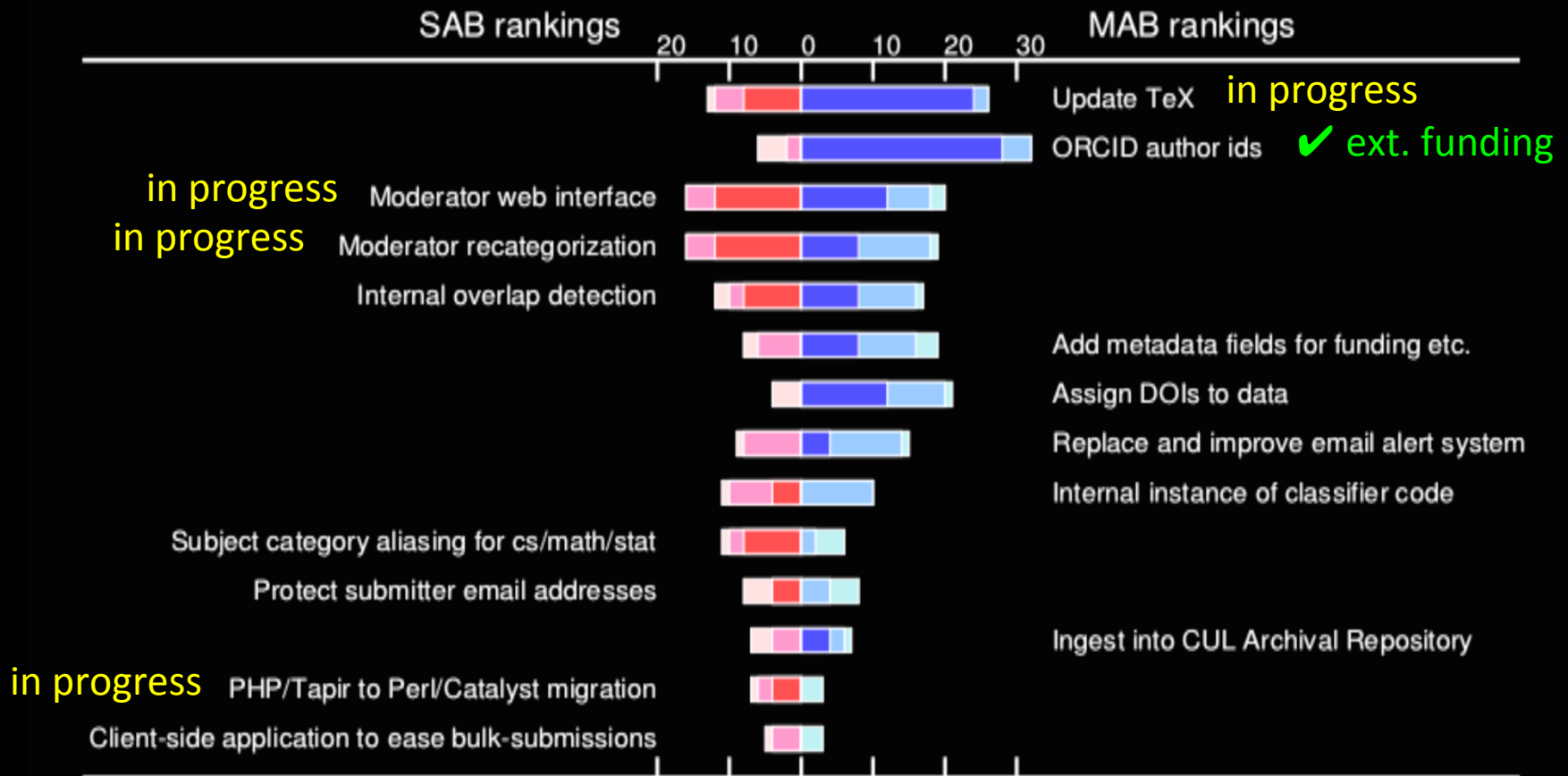
Replace email alert system for better maintainability, to allow easy subscribe/unsubscribe, and for flexible options - Replace email alert system with a more maintainable system.

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



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 in 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 obvious issues.

Add ORCID author identifier support - We would like to be providing institutional statistics for member organizations and accounts via the ORCID OAuth process. [Completed](#)

Improve moderator web interface, add personalization - Work was significantly extended and improved in 2014. Work will continue in 2015.

Ingest data from discontinued Data Conservancy repository - In order to preserve access to datasets used in the repository, we are migrating data to a new location. See <http://blogs.cornell.edu/dsps/2013/06/14/arxiv-dsps/>

Allow moderators to recategorize articles via the admin interface - Admins as intermediaries. Moderators should be able to request recategorization of articles.

Develop and integrate internal automatic overlap warnings for administrators and moderators based on classifier flags - Work was done in 2014 to integrate the classifier flags into the moderation workflow.

Subject category aliasing for cs/math/stat - There are many subject categories that span multiple years and these require extra work because there are pre-0700 subject categories. We need to update the primary archive to identifier prefix correspondence work and should instead work out and document procedures for handling these.

Update, reorganize and better document the TeX system - We need to put effort into updating our TeX installation, including the current development team. We need to update our documentation (last updated in 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 maintainable Perl/Catalyst 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 a separate 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 to address 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 understanding.

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.

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 v
- **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 indicatio

Software Restructuring & Improvement

- **Restructure the submission system** - We need to expand arXiv's workflow capabilities to better accommodate new and more complicated
- **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

The screenshot displays the arXiv admin interface. On the left, a 'Make Category Proposals' dialog box is open, showing a list of primary categories: ASTROPHYSICS (with sub-categories astro-ph.GA, astro-ph.CO, astro-ph.EP, astro-ph.HE, astro-ph.IM (already proposed as primary), and astro-ph.SR), and CONDENSED MATTER (with sub-categories cond-mat.dis-nn and cond-mat.mtrl-sci). Below the list is a text area for 'Enter a comment...' and buttons for 'Make Proposal' and 'Cancel'.

The main interface shows a submission for 'astro-ph.IM' with the following details:

- Status:** submitted (OK) Queue total: 541 OK: 222
- Buttons:** [Prev] [Edit] [Next] [NotOK] [Hold] [Unsubmit] [Remove]
- Submission Info:** 2015-07-23 12:44 bdc34: Proposed astro-ph.IM as primary: dfgdfg
- Metadata:** submit/1270439 source: pdftex
- From:** Marina Bendersky <bemarina@stanford.edu>
- Date:** Thu, 18 Jun 2015 07:06:42 GMT (1925k)
- Title:** A computational study of chemically heterogeneous surfaces
- Authors:** Marina Bendersky, Maria M. Santore, Joseph A. Scheraga
- Classifier:** cond-mat.soft 0.99 physics.flu-dyn 0.33
- Proposals:** astro-ph.IM
- License:** http://arxiv.org/licenses/nonexclusive-terms/v1.0/
- Abstract:** The adhesion of flowing particles and their capture are controlled by small-scale surface heterogeneity. This heterogeneity translates in the shear field near a collector, suggesting distinct particle capture tendencies in each case. This paper presents a new surface geometry of uniform particles flowing past a heterogeneous fixed surface. Additional simulations revealed fewer extrema in the fluctuating particle-collector separation of the reverse system geometry of uniform particles flowing past a heterogeneous fixed surface. Additionally, the capture of particles on a patchy surface is controlled by the relative to the patch size. This paper presents a new surface geometry of uniform particles flowing past a heterogeneous fixed surface. Additionally, the capture of particles on a patchy surface is controlled by the relative to the patch size. This paper presents a new surface geometry of uniform particles flowing past a heterogeneous fixed surface. Additionally, the capture of particles on a patchy surface is controlled by the relative to the patch size.

On the right, the 'Respond to Proposals' table is visible:

Proposal	Accept Primary	Accept Secondary	Reject Outright
astro-ph.IM	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
astro-ph.GA	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
astro-ph.SR	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Below the table is a 'Comment' field with the text 'Enter a comment...' and a 'Respond' button.

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

arXiv Requests

Plan Work Report Board ▾ ⌵

QUICK FILTERS: Only My Issues Recently Updated

81 Priority 3 In Progress 9 Done Release...

Joe Halpern 18 issues

- ↓ improvements to submission near-duplicate flagging

Greg Kuperberg 27 issues

- ARXIVREQ-39
↑ The main site should permanently cache more formats, like a4 PDFs
- ARXIVREQ-40
↑ Improve formatting of moderator list page to add submission date and secondary categories
- ARXIVREQ-51

- ARXIVREQ-69
↑ Estimate size of software components in section 4.1 of technical overview
- ARXIVREQ-99
↑ Remove spurious grayout of "Status:" tag and categories line on mod page
- ARXIVREQ-107

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?