2016 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. The focus of work in 2016 is improvement of facilities for administrators and moderators in order to streamline their workflows, and to improve clarity and transparency of arXiv communications. A number of other tasks address legacy software issues.

Allow moderators to recategorize articles via the web interface - We want to encourage moderator use of the web interface to streamline their workflow and to avoid unnecessary reliance on admins as intermediaries. Moderators should be able to make specific category change recommendations via the web interface that result in alerts to other appropriate moderators.

> *In late 2015 and 2016 a new set of features was added to the admin interfaces that allow the admins to work with category proposals on submissions. The goal of this was to explicitly codify decisions about categories as data and to use it to efficiently deal with redundant category proposals or crosses.*
>
> *These features were originally rolled out to admins only so feedback and refinements could be made before rolling it out to moderators. The functionality was made available to the moderators in late August 2016.*
>
> - *ARXIVDEV-2503*
> - *status: complete*

Update, reorganize and better document the TeX system - TeX is currently a central component of our article processing, approximately 85% of submissions are TeX or PDFTeX source. We need to update our TeX installation, improving our packaging so that it can more easily be deployed and updated, better documenting our installation, and increasing experience within the current development team. We need to update the TeX binaries to the current version of TeX Live (currently TeX Live 2011, should use 2014), update our set of style files (last update was 2011), and also update our ghostscript installation.

> *The updated TeX system will skip TeX Live 2014 favor of TeX Live 2016. The integration with TeX Live 2016 is currently undergoing "burn-in" testing against recent submissions in our development environment. Testing is going well, and the latest patches from TeX Live 2016 have been applied. The anticipated release date of this update is January 2017.*
>
> - *ARXIVDEV-1561*
> - *status: deployed February 2017*

Migrate functions away from old PHP/Tapir codebase and into Perl/Catalyst - Complete replacement of old PHP/Tapir code with more maintainable and better integrated Perl/Catalyst code.

> *Work to migrate the user administration functionality–the bulk of the tapir codebase--is nearly complete. Several smaller features will still need to be migrated following the migration of the user administration functionality.*
>
> - *ARXIVDEV-2525*
> - *status: work in progress*

Improve ORCID author identifier support - Deprecate the local arXiv author ids in favor of ORCID iDs throughout the arXiv user interface, add ORCID iDs to arXiv API.

> *As part of work to deprecate the local author ids and support ORCID iDs as the primary form of author identification, the ORCID form now simply displays the page (or provide atom / json) without requiring that an arXiv author id exist and without a redirect to an arXiv author id page. ORCID iDs still need to be added to the arXiv API.*
>
> - *ARXIVDEV-3133*
> - *status: deferred to 2017 and/or arXiv NG*

Use HTTPS for all of web interface - The web is rapidly moving toward secure communication (using HTTPS, HTTP over SSL) and all arXiv accesses should occur over HTTPS. All existing HTTP URLs should still resolve but they should do this via a redirect to the equivalent HTTPS URL.

> *HTTPS was enabled site-wide in March 2016 and enforcement began at the end of September.*
>
> - *ARXIVDEV-3133*
> - *status: complete*

Refine process for addition of overlap notices - Following the September 2015 recommendation of the Scientific Advisory Board, we will implement a process to put submissions on hold when overlap with other articles is detected by Paul Ginsparg's system. This will provide authors an opportunity to respond to warnings before overlap notices are added.

> *In recent meeting between Paul Ginsparg, Jim Entwood and Martin Lessmeister, agreement was reached that an acceptable interim solution would be for Paul Ginsparg to directly put submissions on hold with desired overlap notices which would be subject to review by the mod-admins.*

Consolidate web presence at the main site - The arXiv mirror network was very useful in the early days of arXiv and web. However, geographic locality of servers is much less useful now than it was when the mirror network was established, the mirrors are little used, and maintaining support for mirrors impedes development of new features on arXiv. We will discontinue the remaining six arXiv mirrors and improve capacity of the main site network. A presence at Los Alamos National Laboratory will be retained to support http://xxx.lanl.gov/ URLs and we will investigate support for TeX processing for A4 paper.

> *The French mirror was decommissioned earlier this year. The following mirrors (excluding LANL) remain: CN, DE, IN, ES.*

- *ARXIVDEV-1066*
- *status: postponed following discussions in annual SAB meeting*

Subject category aliasing for cs/math/stat - There are three subject category merges (aliases) requested in order to better represent subject areas that span major discipline boundaries. Some of these require extra work because there are pre-0704 (old identifiers, see http://arxiv.org/help/arxiv_identifier) submissions where the primary category is becoming an alias and thus the historical primary archive to identifier prefix correspondence will be broken. In the past aliases have been made on an ad-hoc basis and without the need to change existing primary archive designations. We should instead work out and document procedures for such changes. Includes work to create tools for the bulk re-categorization of submissions affected by this and later merges.

- *ARXIVDEV-1179*
- *status: deferred to 2017 and/or arXiv NG*

Streamline moderator and admin interactions - Work with arXiv moderators to streamline their interactions with the system and with arXiv administrators. Will include better handling of warnings from classification and overlap detection systems, better presentation of discussions of issues and more facilities directly in the moderator web interface.

> *This has been partially achieved by ARXIVDEV-2503 but more work is needed, especially in the area of adding more facilities directly in the moderator web interface. It has been difficult to prioritize these features given the diverse set of moderator needs and preferences. With the addition of the UI/UX developer to our team (Manolo Bevia; .5 FTE), we hope to leverage his expertise to take some of the guesswork out of prioritization.*

Tidy arXiv code and version control systems, improve test coverage, remove unused code - arXiv's codebase has evolved over more than 20 years and while there have been a number of reorganizations over this time, and work to add tests, focused work on improving structure and test coverage will improve maintainability and development productivity. We also need to remove unused code and re-evaluate legacy bug reports.

> *Work supporting the consolidation of arXiv's many git repositories is complete, but will require further testing before deployment on production machines.*

- *ARXIVDEV-2752*
- *status: work in progress*

User Support and Moderation

Define and implement new tools and interfaces for moderators - Continue ongoing work with moderators and arXiv IT to define and implement new tools and interfaces to support the work of moderators.

- *Admins continue to meet with arXiv IT team to clarify workflow and specifications for the proposal system and moderator interface. Admins have also relayed feedback from moderators to developers related to tools and interface. Also in this scope of work the admins lead the work on the Moderator Survey conducted in fall 2016.*
- *Status: 2016 work complete*

Implement new administrative staffing configuration - Hire, train, and integrate Operations Manager into arXiv administration unit

- *Jim Entwood was brought on in February in the new position of Operations Manager. Additional staffing changes include hiring Rebecca Rich Goldweber to the new position of Associate Administrator and promoting Andrea Salguero and Jake Weiskoff to the position of Senior Administrator. This staffing configuration allows the admin team to handle increasing workload, improve quality of service to users and moderators, and adapt to emerging needs.*
- *Status: complete*

Complete and publish physics category descriptions - Work with physics subject committee and its chair to get final approval for physics subject category descriptions.

- *Work in progress.*
- *Status: deferred to complete in 2017*

Publish current policy statements where needed - Work with Scientific Director to compose and publish appropriate policy statements regarding endorsements and appeals.

- *Status: complete*

Systematize arXiv's individual archive pages - The "home" pages for individual archives within arXiv have grown organically over time and are inconsistent with each other and in some cases outdated.

- *Status: deferred to 2017*

Business Model & Governance

Continue testing the arXiv Scientific Director position – In 2014, we created a new position to provide intellectual leadership for arXiv's operation and appointed Chris Myers as the interim Scientific Director. We'll continue to test and refine the job description and also consider the effectiveness of the current arXiv team model. Myers' appointment extended to the end of 2016. Position needs to be assessed, revised (if necessary), and posted in 2016.

After Chris Myers's departure in April 2016, we've started to assess the position and are in the process of revising the job description. Dr. Myers was instrumental in evaluating the position and proving useful feedback to help us refine the set of responsibilities and expectations. We aim to fill the position in early 2017.

Continue the membership drive & identify new funding sources - We continue to be encouraged with the five-year pledges and increasing number of arXiv member institutions. Creating a broad and international network of supporters requires ongoing efforts. We are entering the fourth year of our 5-year business plan. One of the goals this year is to start planning for the next 5-years. Based on last year's online fund raising pilot (Give button), we aim to assess and repeat the strategy to create an additional revenue stream.

We repeated the online fund raising campaign and raised an additional $32,000. We also secured additional members, bringing the total to 200 institutions. This year, the Simons Foundation increased its annual contribution to arXiv from $350,000 to $400,000. In addition to raising funds to support the current arXiv operation, we successfully secured two grants in the amount of $650,000 to initiate the next-generation arXiv initiative, which is envisioned to be a 3-year project with a total budget of $3+ million.

Continue assessing and refining the operation of the new organizational and governance model -The arXiv principles aim to clarify the authority, responsibilities, and constraints of CUL, MAB, and SAB. Ironing out problems and developing a working system will require some time to test and observe the inner operation of the governance model. We will continue our engagements with the advisory boards and experiment with different communication strategies to share our vision, priorities, and challenges and to seek their input.

This is an ongoing process as we review our goals, strategies, and performance annually, especially during the MAB and SAB meetings. One of the key outcomes of this process is modifications to the arXiv team structure based on the input gathered. In order to strengthen the daily oversight of arXiv, during 2016 we transitioned to a team model that has two full-time managers, one for IT and the other one for user support, and appointed a new CTO. Here is the current team model: arXiv Sustainability Initiative

New Partnerships & Communication

arXiv's role in scholarly communication ecology - We will continue following the new developments in regard to public access mandates and related compliance issues. Also of interest to the arXiv team are plans for integration of standardized metadata by use of IDs like ORCID, Grant-IDs, or Institutional IDs; SHARE & CHORUS; and depositing and linking research data associated with papers. Based on last year's investigation of

interoperability issues, we will continue exploring how arXiv should enable communication/exchange between arXiv and institutional repositories, (for instance, pushing copies of papers published by a scientist to his/her home institution's repository). This work is added to the Special Projects category below as we need additional funds to accomplish our goals in this domain.

> *We completed a preliminary assessment of how we can enable one-stop deposit and compliance reporting in support of new public access requirements such as the one from the UK (HEFCE/REF). One of the goals of the arXiv-NG initiative is to create an open architecture to allow other service providers to develop APIs as overlay services, for instance to simplify scientists' publishing process by enabling submission of an article to a journal at the same time as it is uploaded to arXiv.*

Special Projects

The current 5-year business plan represents a baseline maintenance scenario. It was developed based on an analysis of arXiv's baseline expenses during 2010-2012. It does not factor in any new functionality requirements or other unforeseen resource needs. Although a development reserve was established to fund such expenses, it is not sufficient to subsidize significant development efforts through surplus funds. Stewardship of resources such as arXiv involves not only covering the operational costs but also continuing to enhance their value based on the needs of the user community and the evolving patterns and modes of scholarly communication. From users' perspective, arXiv continues to be a successful, prominent subject repository system serving the needs of many scientists around the world. However, under the hood, the service is facing significant pressures. The conclusion of the 2015 SAB and MAB annual meetings was that the arXiv team needs to embark on a significant fund raising effort, pursuing grants and collaborations. Based on the ideas and recommendations gathered last year, we've concluded that we need to first create a compelling and coherent vision to be able to persuasively articulate our fund raising goals. We'd like to use the approaching 25th anniversary of arXiv as an important milestone to engage us in a series of vision-setting exercises. Please see the arXiv Review Strategy for more details.

> *We successfully completed the tasks identified in the **arXiv Review Strategy** and secured funds from the Simons Foundation and the Allen Institute for Artificial Intelligence to launch the arXiv-NG initiative. As we develop functional requirements and create use cases, the arXiv team will be informed by the special projects identified below.*

Interoperability & Public Access Mandate Support

- **Add metadata fields for funding information, article status and migration of old content** - arXiv team has received several requests for support for additional metadata such as funding information, version information (author manuscript, publisher version, etc.), and publication information. These changes will require extensions of our internal metadata format and handling in appropriate submission interfaces, admin interfaces, moderator screens, search systems, and data export facilities.
- **Support arXiv-IR interoperability -**Test and implement the interoperability requirements identified to enable communication/exchange between arXiv and institutional repositories (e.g., pushing copies of papers published by a scientist to his/her home institution's repository). This work may also involve working with publishers/societies represented in arXiv to explore issues such as version of record, linking pre-print to formal published version, etc.
- **Add linkages to datasets in data repositories** - Based on our experience with the Data Conservancy pilot (http://arxiv.org/help/data_conservancy), a loose coupling to existing external data repositories seems more likely to be sustainable than close collaboration. This also has the benefit of allowing arXiv to work with many repositories, so that users can use the data repository that best matches their need, their community expectations, etc.
- **Create tools and facilities to better integrate with Computer Science conferences** - Scope out a project to ease the upload of proceeding (or other collections) by reducing the amount of custom programming required for the submission of proceedings via the SWORD interface.
- **Assign DOIs to data** - We accept data as ancillary files (http://arxiv.org/help/ancillary_files) but offer relatively little support. It would be more helpful to assign DataCite DOIs from EZID to ancillary files thus making them citeable.
- **Ingest arXiv content into CUL Archival Repository** - While arXiv adopts good practices for data backup and management, it is far from being an archival collection. As we increase our collaboration with other repositories and consider supporting public access mandates, we need to strengthen our preservation strategies. Work is require to script creation of submission packages (SIPs) for initial ingest (and regular incremental updates) of arXiv content to CULAR (Cornell University Library Archival Repository). Also, we'd like to explore the need for additional archival strategies (e.g., working with Portico or Lockss).

Modernize the User Interface & Alerting System

- **Modernize the search interface, add facets, include author identifiers** - The arXiv search interface could be improved to follow current best practices using facets and better result ordering.
- **Replace and improve alerting system** - Replace the email alert system to allow easy subscribe/unsubscribe via web interface tied to user accounts, ensure scalability and allow customization. The current code is very old and hard to maintain, the bulk of it should be rewritten.
- **Stamp withdrawn articles** - Articles in arXiv cannot be entirely removed once announced, but a withdrawal notice may be added by the submitter or by arXiv administrators (http://arxiv.org/help/withdraw). We would like to develop a system to stamp previous versions of withdrawn articles with a clear indication of withdrawal while retaining the original content and version history.

Software Restructuring & Improvement

- **Restructure the submission system** - We need to expand arXiv's workflow capabilities to better accommodate new and more complicated workflows associated with document analysis, overlap detection, and improved moderator interactions.
- **Accelerate legacy codebase improvements** - While the arXiv software operates well, there are areas where the codebase is old and should be migrated or rewritten to make it more efficient to maintain and further develop. So we need to invest in arXiv beyond what is feasible through the operational budget in order to make it "sustainable" - easier to keep up and advance. We seek to accelerate the modest progress possible within the baseline maintenance scenario.