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 in a submission going straight to 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 technical issues. Completed.

Add ORCID author identifier support - We would like to support ORCID identifiers for better interoperability with other repositories implementing authority control and also as a route toward providing institutional statistics for member organizations (because ORCID has implemented storage of affiliation identifiers in the profile data). ORCID identifiers will be associated with arXiv user accounts via the ORCID OAuth process. Completed.

Improve moderator web interface, add personal checkbox - We want to encourage moderator use of the web interface to streamline their workflow. The moderator web interface was significantly extended and improved in 2014. Work will improve clarity based on the feedback we have received and provide each moderator with the ability to mark submissions as checked. Postponed pending experience with facilities to allow moderators to recategorize articles via the web interface.

Ingest data from discontinued Data Conservancy pilot - The Data Conservancy pilot that ran from 2010 through 2013 has been discontinued and Johns Hopkins are going to shut down the pilot repository. In order to preserve access to datasets uploaded with over 600 articles we need to pull the Data Conservancy data into arXiv as ancillary files (see http://arxiv.org/help/data_conservancy and http://blogs.cornell.edu/dsps/2013/06/14/arxiv-data-conservancy-pilot/). Completed.

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. Work in progress.

Develop and integrate internal automatic overlap detection for new submissions - Develop pipeline for checking of new submissions against existing corpus and staged submissions. Develop warnings for administrators and moderators based on overlap check results. Make these warnings available for administrators and moderators. Completed. Title-based checks implemented internally, in parallel with calls to Paul Ginsparg's overlap detection system.

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. Not started.

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 put effort into updating 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. Work in progress.

Migrate functions away from old PHP/Tapir codebase and into Perl/Catalyst - We have been gradually replacing old PHP/Tapir code with more maintainable and better integrated Perl/Catalyst code. Work in progress.

Develop and integrate internal instance of classifier code - We should integrate the classifier code into the arXiv production system rather than using API to code running on Paul Ginsparg's 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 time to tidy his code. There are 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. Postponed pending agreement on sharing code or decision to rewrite.

User Support and Moderation

Define and implement new tools and interfaces for moderators – Continue working with moderators and arXiv IT to define and implement new tools and interfaces to support the work of moderators. See "Improve tools and interfaces to support moderators" in Technical section above. Work in progress.

Improve arXiv administrative processes – Work with Scientific Director and others to evaluate arXiv administration processes, and to define and implement an optimal administrative staffing configuration, in light of evolving moderation tools and staffing needs. New admin staffing configuration defined. Job position for arXiv Operations Manager created and posted. Interview process begun.

Publish arXiv category definitions – Complete the development of public subject category descriptions for existing physics categories. Only a small number of physics categories currently have public descriptions. Defining the scope and boundaries of the categories will help users, moderators, and administrators. Work in progress.

Review arXiv endorsement policies – Review current arXiv endorsement procedures and policies across all subject categories, seeking greater uniformity and transparency. Work with IT to implement any polices that can be programmatically enabled. New endorsement policy approved by SAB during July 2015 meeting. Public documentation and development work to do.

Systematize the arXiv moderation appeal processes – Work toward a uniform arXiv moderation appeal process across all subject categories. Provide public documentation of the process. New appeal process proposed, modified, and accepted by SAB in first quarter 2015.

Review arXiv user communication – Begin to review the many "stock" messages used by arXiv administrators when communicating with submitters and other arXiv users. Some of these messages are outdated, cryptic, or unnecessarily brusque. Work toward identifying these and improving their usefulness. Modification of messages ongoing, as those needing change are encountered.

Develop arXiv moderator assessment metrics – Define, develop, and implement metrics for evaluating moderator performance, to share with subject committee chairs. A tool has been developed to calculate and report on a moderator's level of activity using moderator web interface. Report generated on demand.

Develop criteria for accepting new subject domains – arXiv is occasionally asked to add a new subject domain. While there is considerable experience within the organization about what to consider regarding new domains, this has not been documented. Gather and document a set of criteria to be used when considering the feasibility of adding new subject domains to arXiv. A set of questions for assessing new domains is now documented on SAB wiki.

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.

Setting development priorities - arXiv operates on limited resources therefore it is critical for us to identify and set priorities. In support of this goal, we'll experiment with an online survey to get input from both SAB and MAB in ranking and commenting on this year's technical agenda. Completed.

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 third year of our 5-year business plan. One of the goals this year is to start planning for the next 5-years. The idea of adding a Give button was provisionally approved by SAB & MAB, contingent on a pilot proposal that will lay out the details. Also, we want to explore other funding opportunities from federal and private agencies. Work in progress, during the last 3 years, each year we add 5-10 new members.

Continue assessing and refining the operation of the new 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. Ongoing process as we review our goals, strategies, and performance annually, especially during the MAB and SAB meetings.

New Partnerships & Communication

arXiv's role in scholarly communication ecology - We continue to get questions and requests from libraries, publishers, societies, and funding agencies in regard to arXiv's role in supporting emerging OA mandates and providing features in support of compliance requirements. We will continue following the new developments in regard to open access mandates from funders 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. We will continue to explore issues related to depositing and linking research data associated with papers. Also, we are partnering with Hypothes.is on an Alfred P. Sloan foundation grant to explore open annotations for scholarly communication. Work to be continued in 2016 - see Special Projects section below.

Interoperability of arXiv with other institutional and subject repositories. One of the important factors in our sustainability efforts is enabling interoperability and creating efficiencies among repositories with related and complementary content to reduce duplicate efforts and increase efficiencies. We will investigate interoperability requirements to 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). We formed a MAB subcommittee to identify needs and assess if and how arXiv can provide such functionality. Also, we'll continue to exchange information with publishers/societies represented in arXiv, especially in exploring issues such as version of record, linking pre-print to formal published version, etc. Draft interoperability needs assessment and requirement document completed. 2015_arXiv_IR_interop_plan_draft.pdf - project added in the Special Projects category as we need additional funds to accomplish our goals in this domain.

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. We need to pursue grants and engage in collaborations to secure funds to support the following goals:

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 exploring 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 required 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.

2015 Roadmap was revised and re-posted in August 2015 (additional special projects)