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, moderation and admin system should be over SSL. The login in particular exposes arXiv to sniffing attacks for user, moderator or admin credentials, and is especially and issue when users login on unsecured wifi. Completed.

Improve tools and interfaces to support moderators - Work with the SAB and moderators to define and develop better tools that allow moderators to interact more directly and efficiently with the arXiv system and administrators. The overarching aim is to make best use of available moderator effort by making the work of moderators as quick and convenient as possible, consistent with achieving the quality goals and following policies set out by the SAB. Completed several improvements, further work in progress.

Add automatic classification checks to submission system and integrate with moderator alerting - Need to set Paul Ginsparg's classifier software and model data up on a production server and understand how we will deploy software and model updates from him. Will need to work out how to call this from the submission system for new articles, including pipelining the text extraction needed for the classification to work. Based on response from classifier add information to moderator email and also make classification information available on /mod/ pages. Consider thresholds for alert for reclassification. Completed, further work being done to refine notifications.

Provide summary documentation of the arXiv code to SAB - Provide documentation of the arXiv codebase arrangement, technologies and areas of personnel expertise. Do this as an exercise of updating, fleshing-out, and organizing information on the arXiv wiki. Provide dumped snapshot to SAB. Completed.

Do category aliasing for cs/math/stat and new category for q-fin - There are three category merges (aliases) requested and one new category requested. Some of these are problematic because there are pre-0704 (old id) submissions that exists with primary category that is becoming an alias and so will "break" the primary<->id correspondence. In the past aliases have been made on an ad-hoc basis by Simeon Warner. We should instead work out and document procedures for such changes. We will need to create some tools for the bulk re-categorization of submissions affected by merges. New q-fin category created, aliases in cs/math/stat not started.

Add automatic overlap detection comparing new submissions with existing corpus and generate notifications for administrators and moderators - Work with Paul Ginsparg to obtain bulk overlap check software and install on production servers. Develop pipeline for checking of new submissions against existing corpus and staged submissions. Developer warnings for administrators and moderators based on overlap check results. Make these warnings available in moderator emails and on /mod/ and /admin/ pages. Completed duplicate submissions checks based on title similarity, implementation of full overlap detection not started.

Investigate and expand the id range to $yymm.nnnnn$ - We have already had months with > 8000 submissions (1210=> 8452, 1304 => 8135), and must soon expect to have > 9999 submissions in a month. This requires migration to $yymm.nnnnn$ ids instead of the current $yymm.nnnn$. From January 2015 submissions will get yymm.nnnnn identifiers, existing submissions will retain yymm.nnnn identifiers.

Add support for ORCID and other author identifiers associated with authors - 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 is implementing storage of affiliation in the profile data). Work in progress.

Ingest data from discontinued Data Conservancy pilot and assign DOIs to data as ancillary files - We have discontinued the Data Conservancy pilot but not yet pulled the data into arXiv. We continue of accept ancillary files but offer relatively little support. Plan to pull the DC data in and to assign DataCite 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 listings files. While the web user interface does hide this information somewhat, the email addresses to too easily harvestable and potentially misused. We have had complaints from users where collaborating services have made the emails available. We should instead keep the submitter email addresses only in the database, and not in the metadata or listings files. Implemented protections in view-email pages, not yet removed from metadata files.

Replace email alert system for better maintainability, to allow easy subscribe/unsubscribe, and for flexible options - Replace email alert system: allow changes via web interface tied to user accounts, make maintainable code, ensure scalability and customization. Work not started.

User Support and Moderation

Define and implement new tools and interfaces for moderators - Work with existing moderators and 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. Completed several improvements, further work in progress.

arXiv moderator coverage and evaluation - 1) Work with subject committees to attain a full complement of moderators. Over 12 new moderators added in physics during 2014; 2 empty physics mod positions as of Dec 2014. 2) Develop and implement metrics for evaluating moderator performance, to share with subject committee chairs. Work not completed: defining and developing countable metrics for moderator performance.

General Physics moderation - Work with Physics Advisory Committee to implement an effective moderation process for the General Physics category. Gen-ph moderator in place as of November 2014.

Scientific Director - Work with various stakeholders to define the role of the Scientific Director with respect to daily moderation processes and workflows. Work in progress.

Evaluate administrative processes - Work with Scientific Director to evaluate arXiv administration processes and staffing, especially in light of evolving moderation tools. Work in progress.

arXiv category definitions - Work with Physics Advisory Committee to develop public subject category descriptions for existing physics categories. Only a small number of physics categories currently have descriptions. Defining the scope and boundaries of the categories will help users, moderators, and administrators. Unedited descriptions have been gathered from moderators for most categories; yet to do: work with the Scientific Director and Physics Advisory Committee for editing and approval.

Business Model & Governance

Recruit an Interim arXiv Scientific Director -We created a new position to provide intellectual leadership for arXiv's operation. First we would like to test this position and refine the job description. In support of this goal, we'll recruit a temporary Scientific Director at 0.50 FTE capacity for an 18-month term to explore how such a position can contribute to arXiv in formulating overall scientific direction of the service and its policies. Chris Myers as SD recruited.

Test and refine 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. We continue to evaluate our arXiv's governance model and the team structure. We formed several MAB and SAB subgroups this year in order to focus on specific issues such as the IT prioritization, recruitment of a scientific director, evaluation of membership & revenue model, etc. Another important milestone was appointing two new SAB members based on a nomination and election process specified by the SAB bylaws.

Further refine the benefits of being a member for participating institutions - During the governance planning meetings, several ideas emerged as potential free services for members; however, implementing these features may require significant staff time. Therefore these services need to be considered in the context of arXiv's current maintenance and development priorities. We will engage the MAB and SAB in discussing possible ideas and assess their potential value and requirements from the arXiv team. Work in progress - discussion item for SAB/MAB meetings.

Continue the membership drive - We are very encouraged with the five-year pledges received so far. We want to increase the number of arXiv member institutions to create a large and international network of supporters. We recently formed two SAB subcommittees. The first one aims to assess and revise our tier model to consider other classes of member categories such as scientific societies (to be implemented during the next 5-year plan). The second group will focus on fund raising policies & strategies such as the potential addition of a 'Give' button to arXiv homepage to encourage donations. This is an ongoing activity and will continue. With advice from MAB, a decision was made not to change the tier model and membership class categories for the current 5-year business period. The idea of adding a Give button was provisionally approved by SAB & MAB, contingent on a pilot proposal that will lay out the details.

New Partnerships & Communication

Continue the dialogue with publishers/societies - Continue information exchange with publishers/societies represented in arXiv, especially in exploring the evolving public policies and open access mandates. One of the ideas we are considering is implementing a pilot with the IOP for facilitating deposit of author version of articles to arXiv after they are published. Work in progress (IOP pilot delayed due to the publisher's current focus on CHORUS). The current OA landscape is changing rapidly due to emerging OA mandates. We will continue to follow the changes and evaluate arXiv's role within the new scholarly communication models.

Define and communicate measures of success - Over the next several months, we want to create an assessment model to help CUL continue to fine-tune the sustainability model. Working with SAB and MAB, we will draft desired outcomes and success measures (for instance, dynamics of the governance model, level of financial support, enhancements to arXiv, improvements to moderation system, etc.). Work in progress. One of the new strategies we tried this year was engaging the MAB members in an informal survey of arXiv user experiences from their home institutions.