arXiv.org の次世代システムの公開と戦略

arXiv.org Next Generation -Its Opening and Strategy-

引原隆士（京都大学図書館機構長/
arXiv.org会員コンソーシアム代表）
Takashi Hikihara
Professor, Director of Kyoto Univ. Library Network
Representative of the NII Japan Consortia for arXiv.org

Data is on arXiv.org site

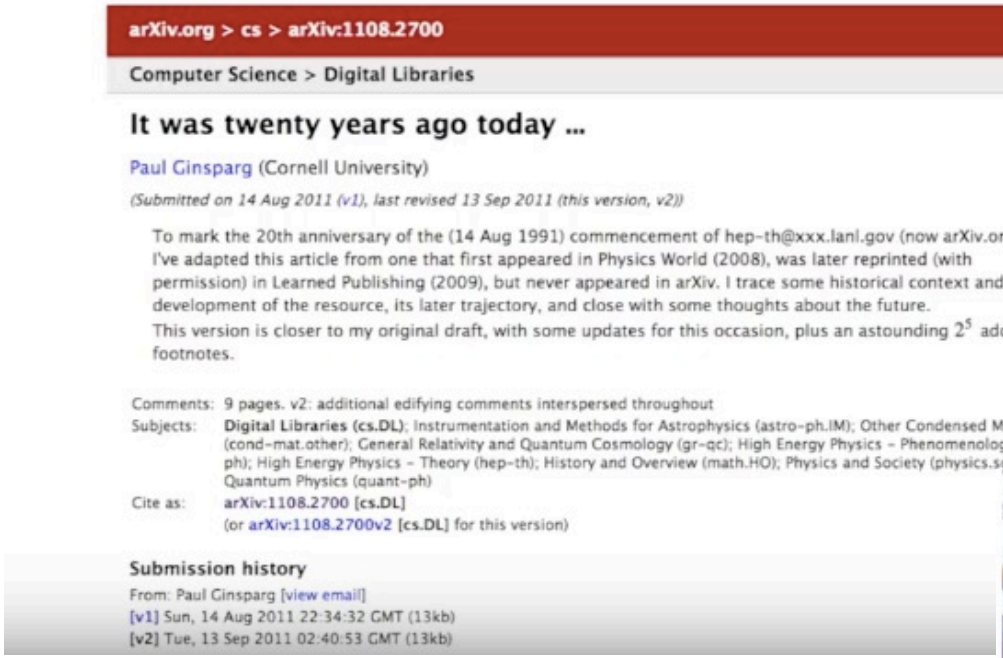
https://arxiv.org

Video

https://www.youtube.com/watch?v=ntoxZzh0ha8

Origin of arXiv.org

arXiv:1108.2700



arXiv.org > cs > arXiv:1108.2700 Search

Computer Science > Digital Libraries

It was twenty years ago today ...

Paul Ginsparg (Cornell University)

(Submitted on 14 Aug 2011 (v1), last revised 13 Sep 2011 (this version, v2))

To mark the 20th anniversary of the (14 Aug 1991) commencement of hep-th@xxx.lanl.gov (now arXiv.org), I've adapted this article from one that first appeared in Physics World (2008), was later reprinted (with permission) in Learned Publishing (2009), but never appeared in arXiv. I trace some historical context and early development of the resource, its later trajectory, and close with some thoughts about the future. This version is closer to my original draft, with some updates for this occasion, plus an astounding 2^5 added footnotes.

Comments: 9 pages. v2: additional edifying comments interspersed throughout

Subjects: **Digital Libraries (cs.DL)**; Instrumentation and Methods for Astrophysics (astro-ph.IM); Other Condensed Matter (cond-mat.other); General Relativity and Quantum Cosmology (gr-qc); High Energy Physics - Phenomenology (hep-ph); High Energy Physics - Theory (hep-th); History and Overview (math.HO); Physics and Society (physics.soc-ph); Quantum Physics (quant-ph)

Cite as: arXiv:1108.2700 [cs.DL]
(or arXiv:1108.2700v2 [cs.DL] for this version)

Submission history

From: Paul Ginsparg [view email]
[v1] Sun, 14 Aug 2011 22:34:32 GMT (13kb)
[v2] Tue, 13 Sep 2011 02:40:53 GMT (13kb)

https://www.youtube.com/watch?v=ntoxZzh0ha8

HP735@Los Alamos

Short history of arXiv.org

- **1991 GINSPARG, Paul, Repository Alert System**
  hep-th@xxx.lanl.gov (before internet)
  High Energy Physics
- **2011 → Cornell University Library**
- **Categories Expansion:**

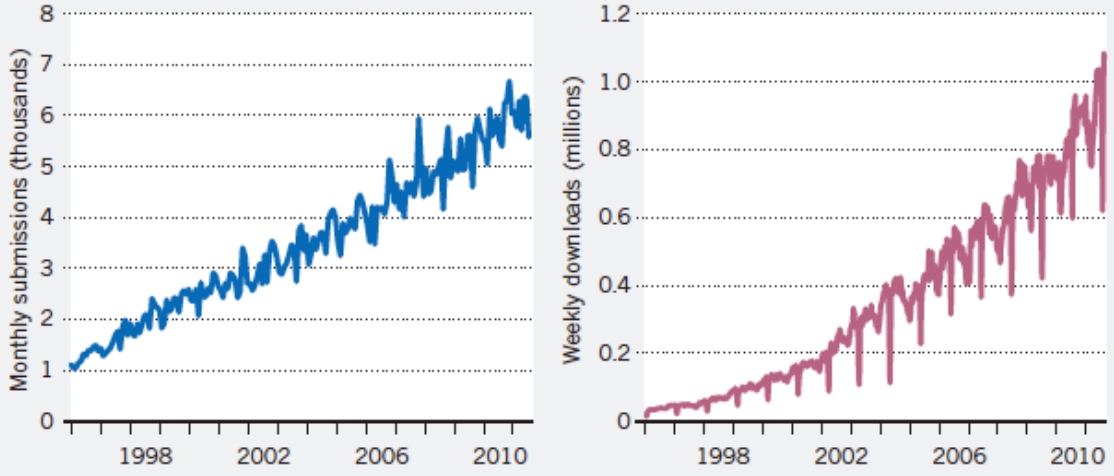
  Physics (1991),
  Mathematics (1997),
  Computer Science (1998),
  Quantitative Biology (2003),
  Statistics (2007),
  Quantitative Finance (2008),

  EESS (Sep. 18, 2017), Econ (Sep. 26, 2017)

GINSPARG, Paul. "ArXiv at 20". nature. vol. 476, p. 145-147, 2011

DIGITAL PIONEERS LEAD THE WAY TO SHARING RESEARCH ONLINE

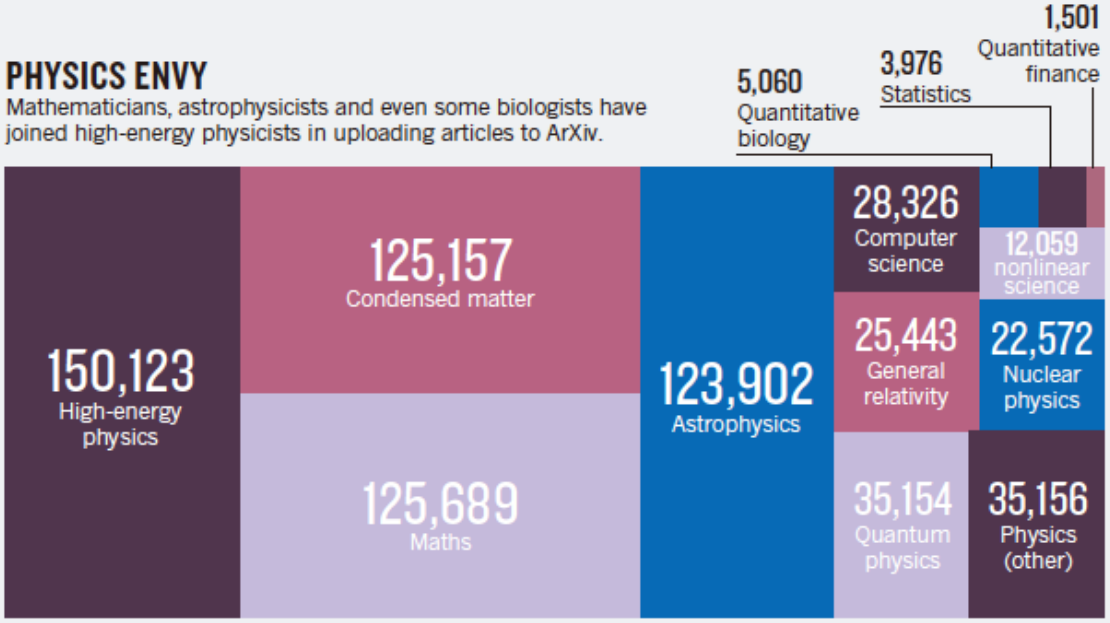
The popularity of the arXiv preprint server has grown inexorably since its launch in the early 1990s. Academics enjoy the universal access, low cost and speed of online distribution.



SOURCE: ARXIV

PHYSICS ENVY

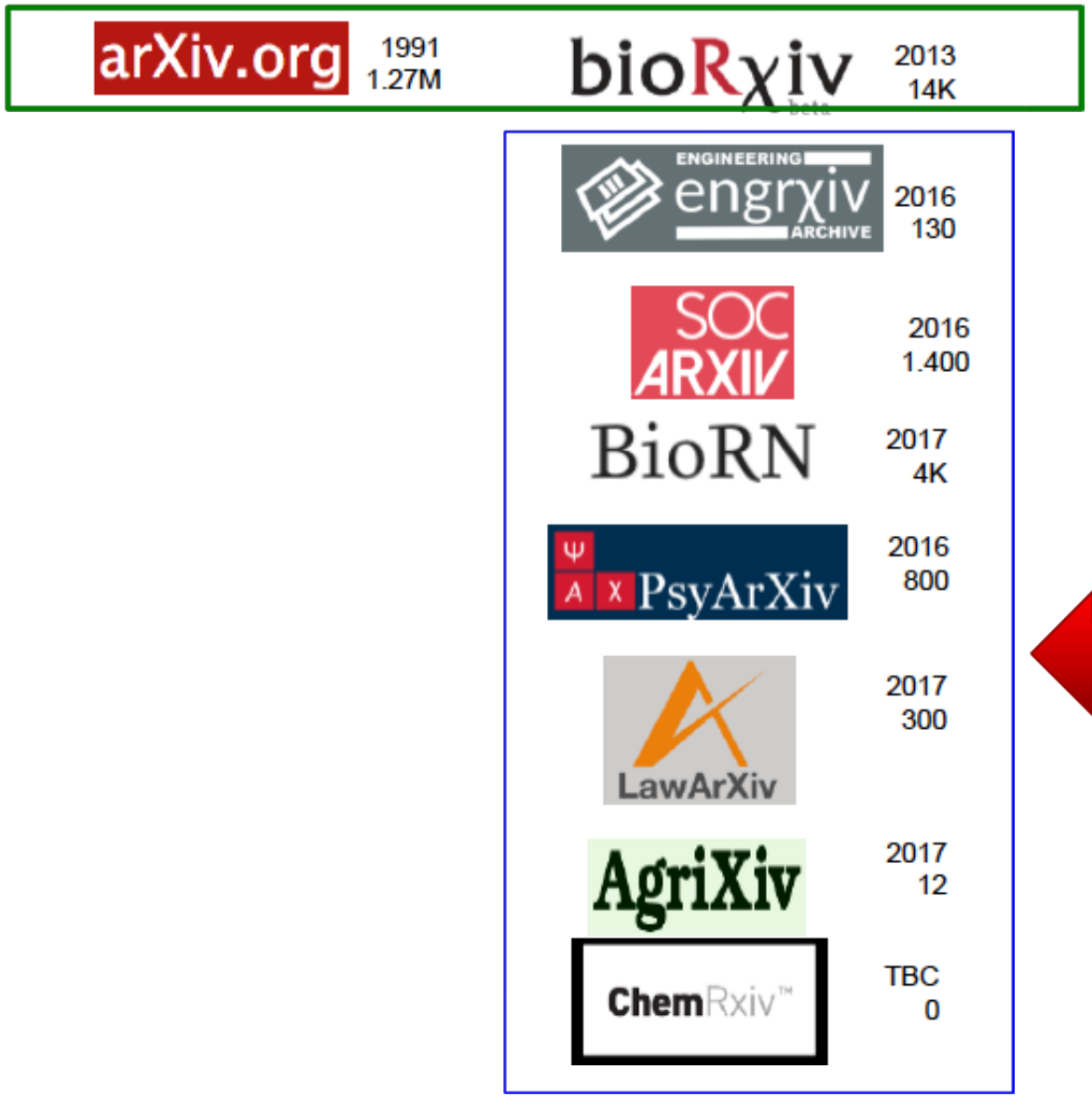
Mathematicians, astrophysicists and even some biologists have joined high-energy physicists in uploading articles to ArXiv.



694,000
Total uploads to ArXiv
(as of 3 August 2011)

GINSPARG, Paul. “ArXiv at 20”. nature. vol. 476, p. 145-147, 2011

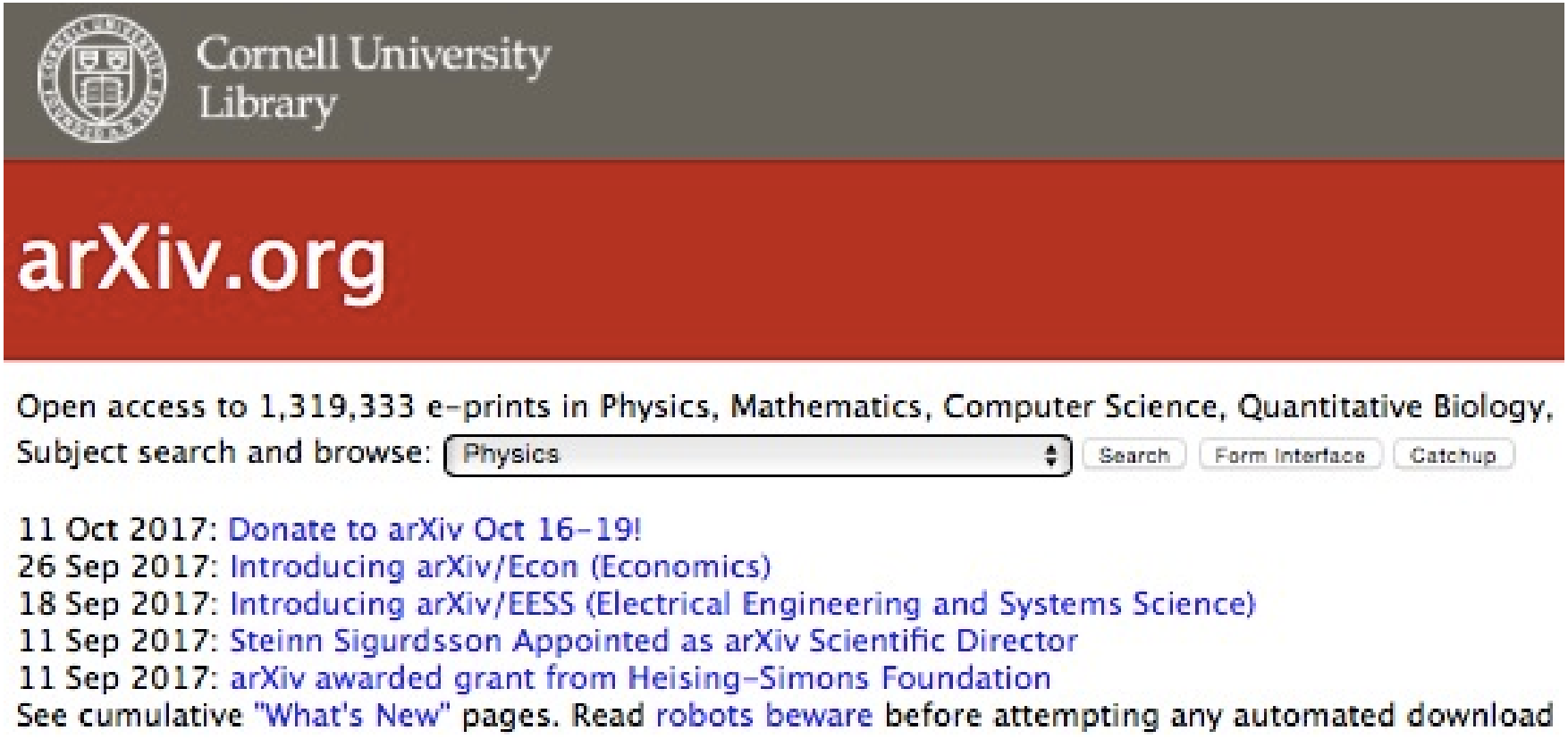
Preprint archive families



Publishers' archive

Their archives are business.

News


Cornell University
Library
arXiv.org
Open access to 1,319,333 e-prints in Physics, Mathematics, Computer Science, Quantitative Biology,
Subject search and browse: Physics
Search
Form Interface
Catchup
11 Oct 2017: Donate to arXiv Oct 16-19!
26 Sep 2017: Introducing arXiv/Econ (Economics)
18 Sep 2017: Introducing arXiv/EESS (Electrical Engineering and Systems Science)
11 Sep 2017: Steinn Sigurdsson Appointed as arXiv Scientific Director
11 Sep 2017: arXiv awarded grant from Heising-Simons Foundation
See cumulative "What's New" pages. Read robots beware before attempting any automated download

arXiv.org Now

The warm gas atmosphere of the HD 100546 disk seen by Herschel*

Evidence of a gas-rich, carbon-poor atmosphere?

Simon Bruderer[1], Ewine F. van Dishoeck[1], Steven D. Doty[2], and Gregory J. Herczeg[3]

ABSTRACT

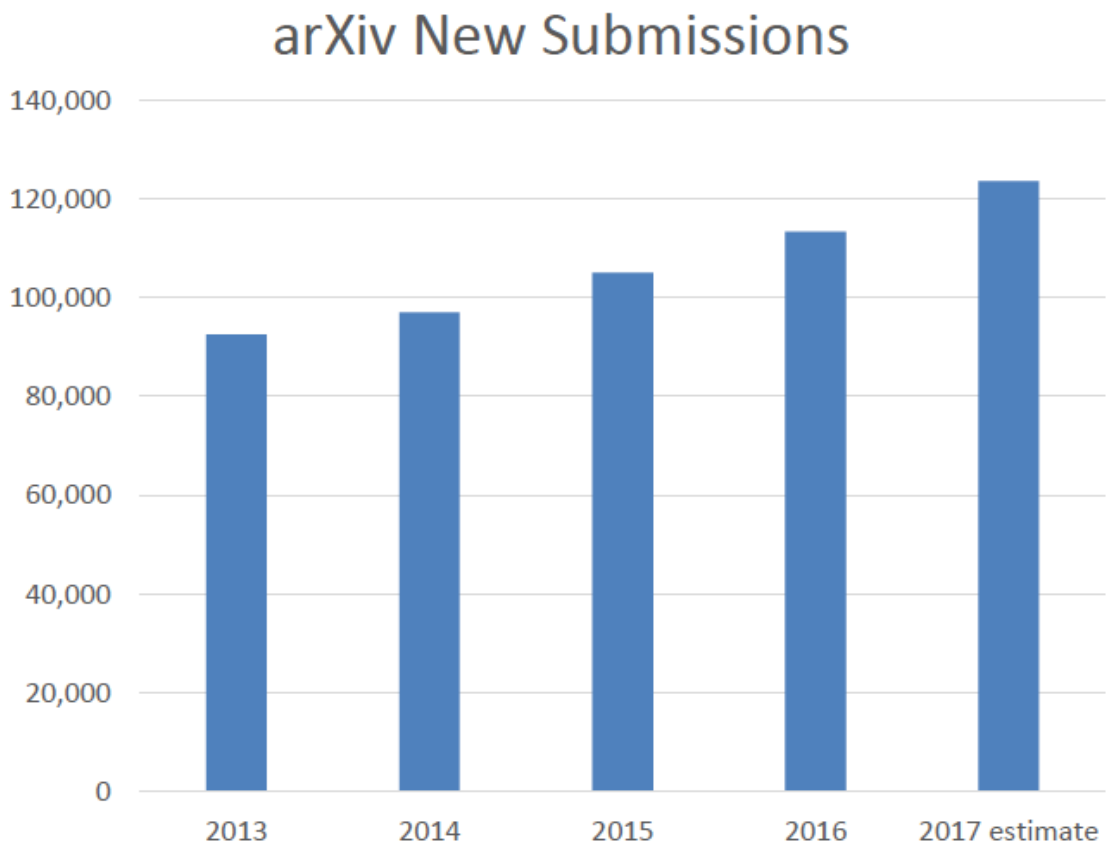
- 2012
  - 84,000 new submissions
  - 64 million downloads
- 2013
  - 92,500 new submissions
  - 67 million downloads
- 2014
  - 97,000 new submissions
  - 90 million downloads*
- 2015
  - 105,000 new submissions
  - 139 million downloads*
- 2016
  - 113,380
  - 139 million downloads*

Cornell University Library

arXiv.org

* The numbers are sensitive to robot downloads and it is hard to remove all from our numbers so potential significant over-counting – we put less effort in cleaning up this data 2014 on.

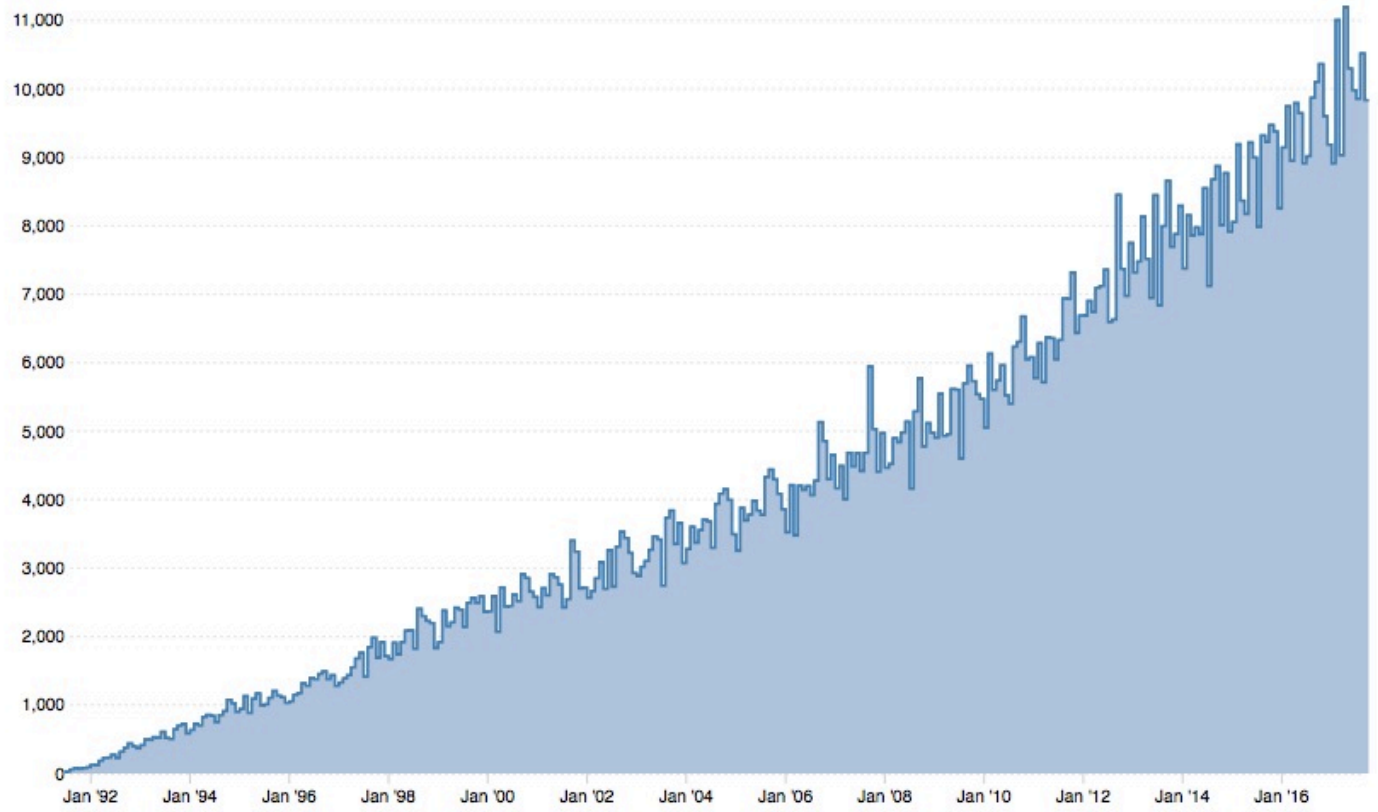
Data: Submission



500-600/day

Statistics

arXiv monthly submission rates [CSV]



Blue: Number of new submissions received during each month since August 1991.
Hover over the graph to see the exact count for a given month.

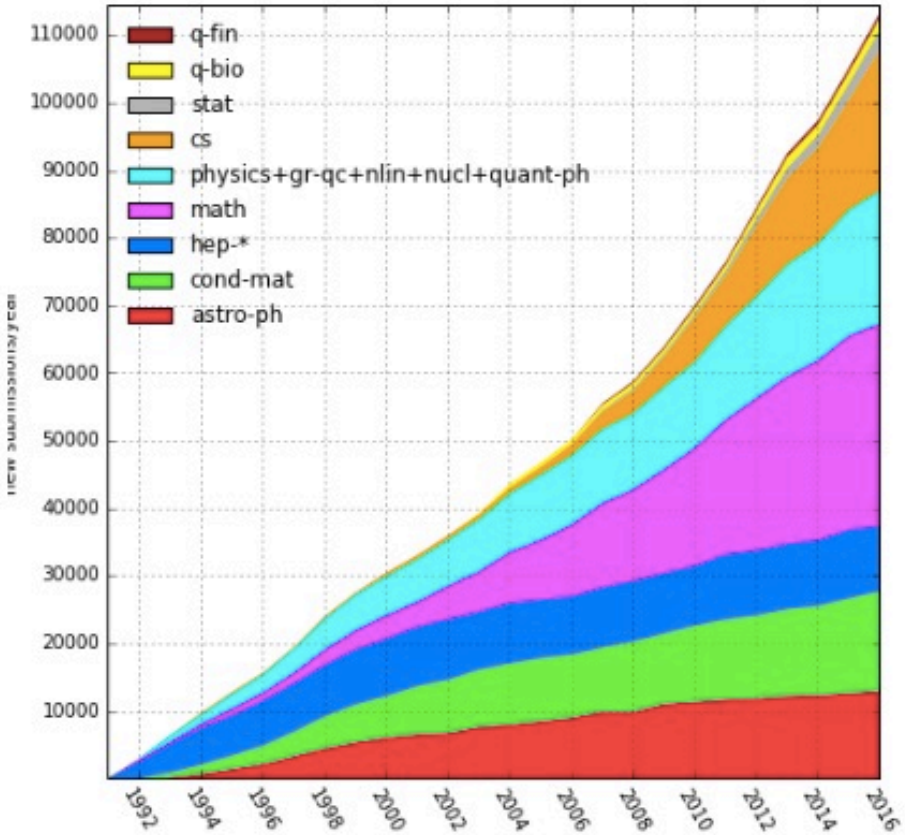
Total number of submissions shown in graph as of October 28th, 2017 (after 26.2 years) = 1,317,057

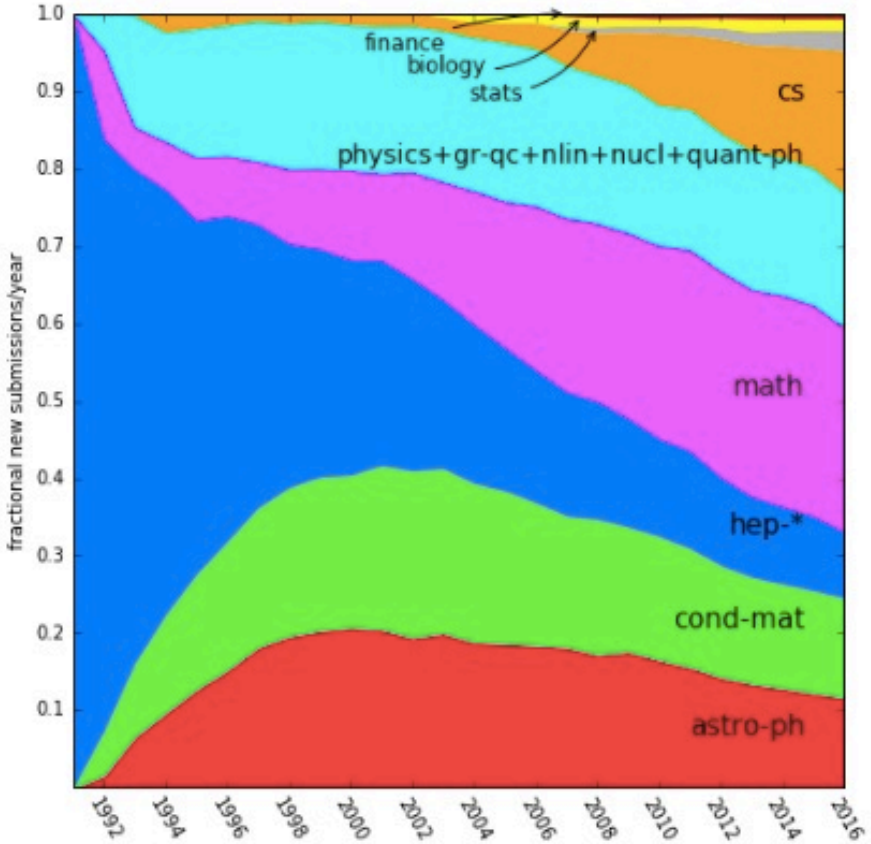
The total number of submissions excludes 2,431 articles that were migrated to arXiv rather than being submitted directly, and includes 155 articles that have been deleted. The total number of articles available is 1,319,333.

See also other arXiv usage statistics.

Statistics 1991-2016

Data for 1991 through 2016, updated 31 December 2016.





Left: number of new submissions/year as a function of calendar year for "hep" = High Energy Physics (hep-th+hep-ph+hep-lat+hep-ex), "cond-mat" = Condensed Matter Physics, "astro-ph" = Astrophysics, "math" = Mathematics (math+math-ph), "other physics" = physics+nucl+gr-qc+quant-ph+nlin, cs, stats, biology = q-bio, finance = q-fin.

The graph on the right shows the same data as at left, but with the submission rates divided by the total for each year, giving the fractional submission rates for each of the domains, and highlighting the growth in submission rates from new domains.

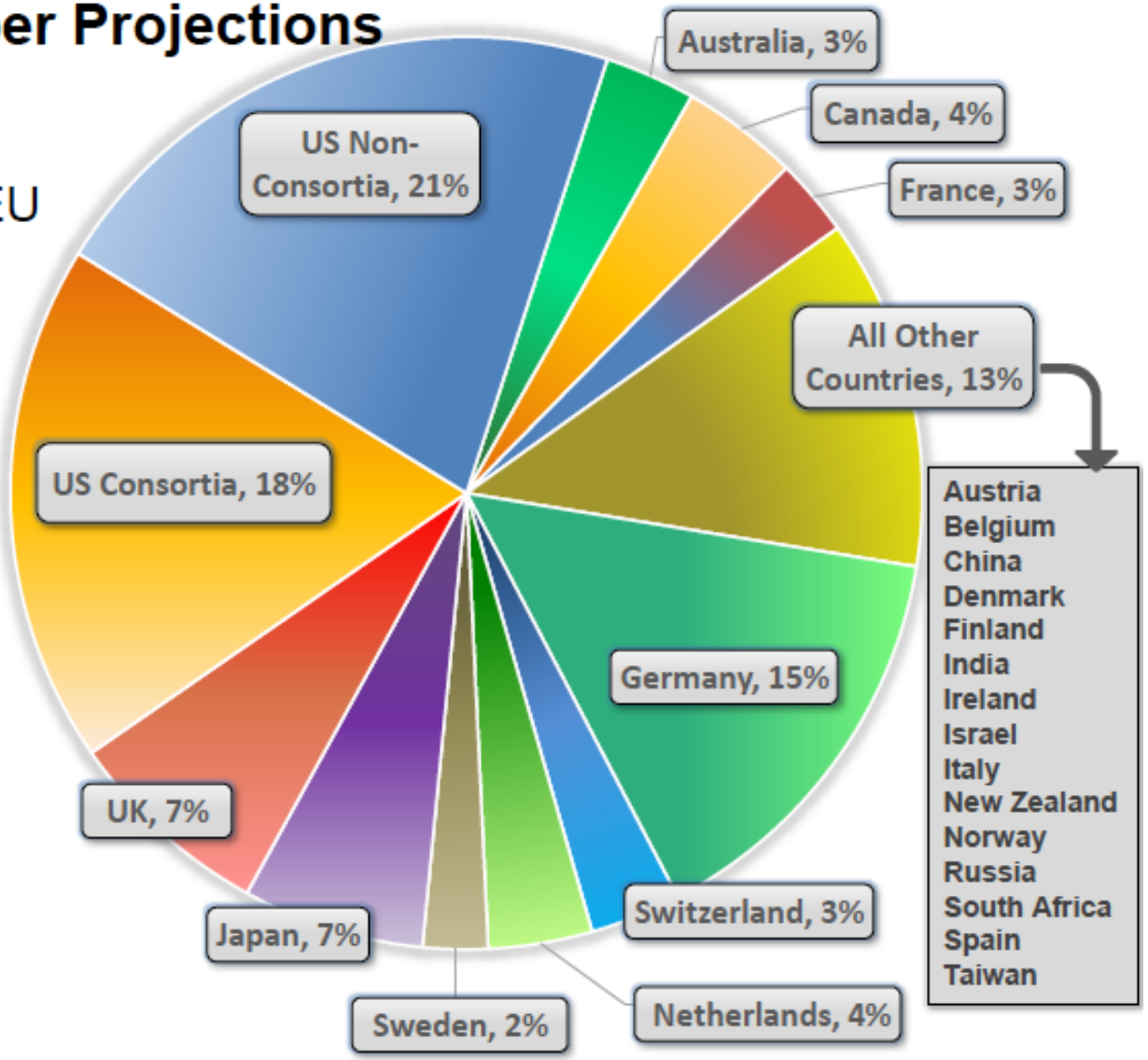
Data : Statistics

2017 Member Projections

25
Countries + EU

217
Members

$425K
Contributions



Infrastructure

- 14 locally-hosted servers (prod, dev VMs), 5 mirrors, shared file system

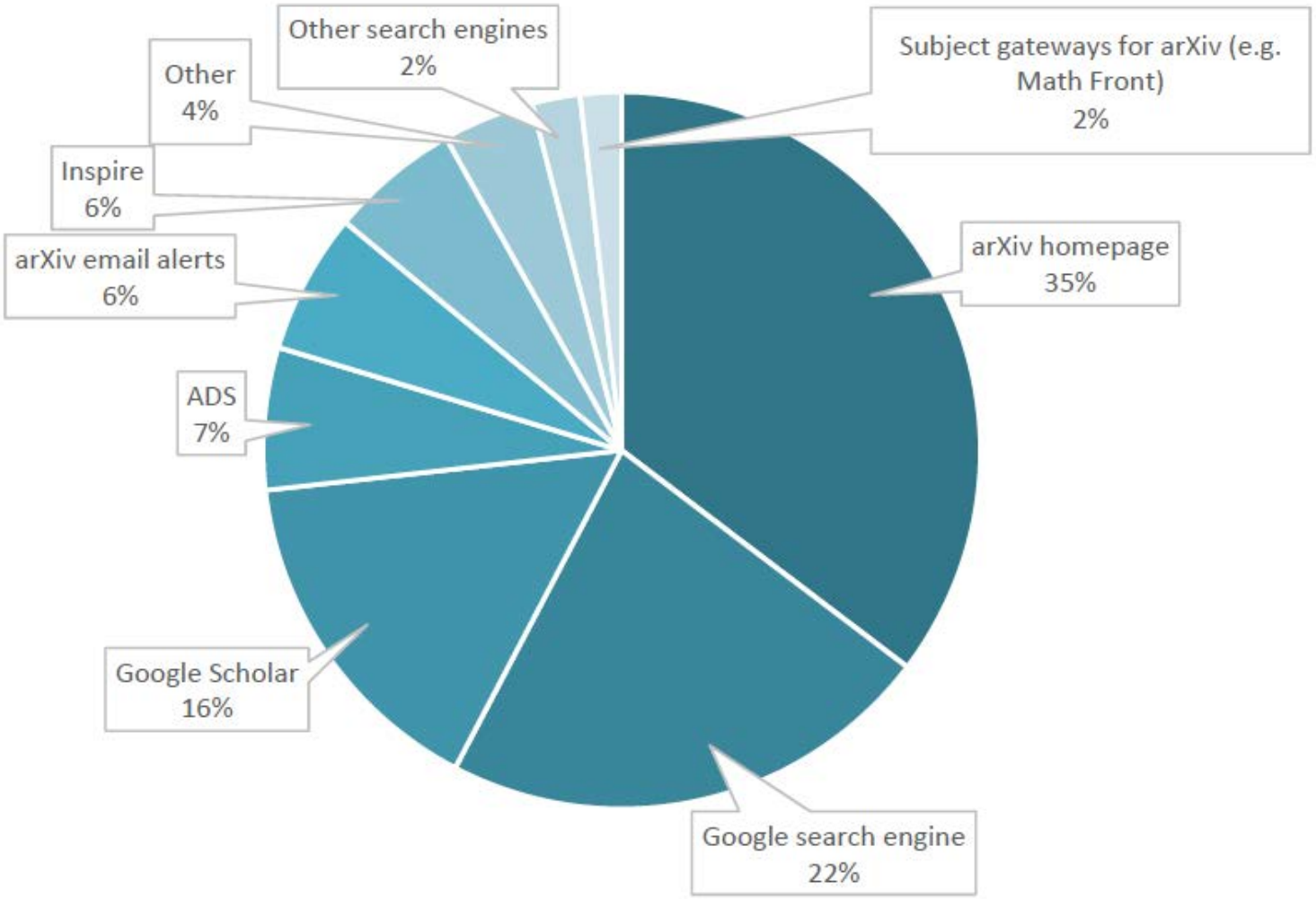
arXiv mirror sites

In addition to the main site at Cornell University Library, there are several mirror sites for arXiv content. These are updated daily but have fewer features than the main site:

- cn.arXiv.org (China)
- de.arXiv.org (Germany)
- in.arXiv.org (India)
- es.arXiv.org (Spain)
- lanl.arXiv.org (née xxx.lanl.gov, U.S. mirror at Los Alamos)
- arXiv.org (U.S. primary site at Cornell University)

- Shifting to Amazon Web Services
- Pressure point: database upgrade

Where do you go to find arXiv paper



arXiv@25: Key findings of a user survey

arXiv.org operations update (Stats from past 12 months)

164 Moderators

- total submissions 235,444
  - 3,032 | cross reference
  - 24,588 | journal reference
  - 128,958 | new submission
  - 75,693 | repeated submission
  - 3,173 | withdraw
- touched by Moderators/Administrators 36,380

- removed 6,886
- bounced to fix 4,984
- proxy submissions 4,631
- auto-Hold 3,790
- iThenticate checks 650
- overlap notes added 787
- % papers with DOI 28%

Standpoint of arXiv.org

Encouragement of research:
- ✓ Do research
- ✓ Write the paper
- ✓ Submit the paper

Research papers
- ✓ Reference
- ✓ Archive

Journals
- ✓ Pay subscription on page charge
- ✓ Take weeks to years

No more Journals
- ✓ Most journals are no longer printed
- ✓ Fully searchable online
- ✓ Cost is to refreeing
- ✓ Stable unique identifier
- ✓ Publishers has no more cost on **Baumol's cost disease**

https://ja.wikipedia.org/wiki/ボーモルのコスト病

- ✓ Receive email and/or check web
- ✓ Clear simple interface authors vs readers **Community**
- ✓ Sources and/or printable

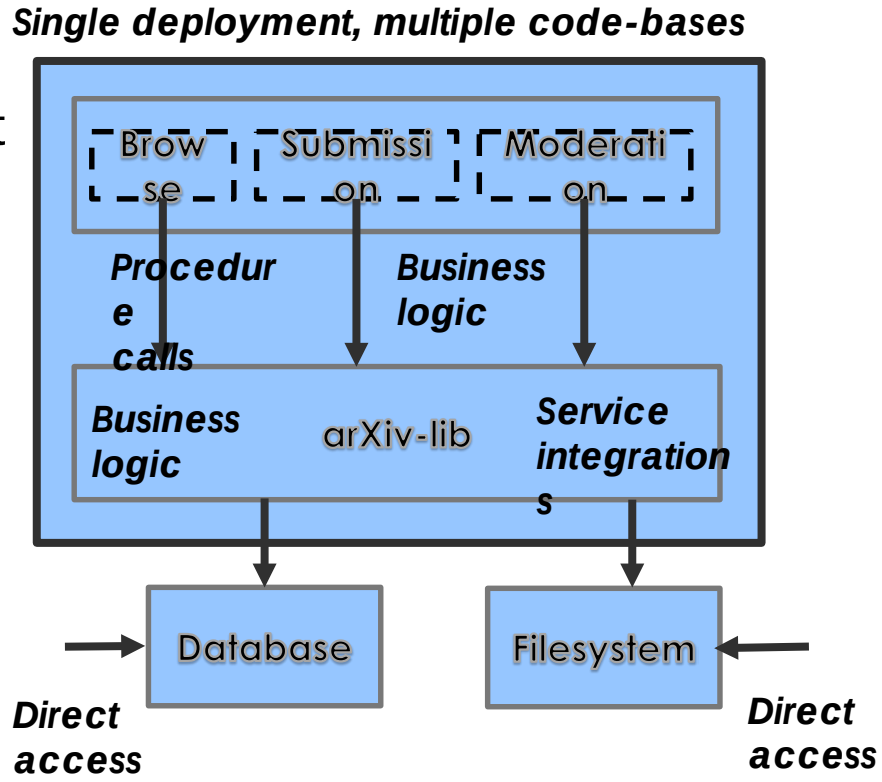
You do not have to include everything in all collections...

What is the arXiv.org, what is it for?

- **Description at multiple levels:** system context, subsystems, components.
- **Audience:** stakeholders and developers.
- **Not a requirements analysis.** Describes decisions and their rationale, and the most important requirements of the system as a whole, but allows for agility and changing requirements throughout the project.
- **Both prescriptive and descriptive:** commemorates technical decisions in context, provides guidance for implementation, but also evolves as new decisions are made throughout the development process.

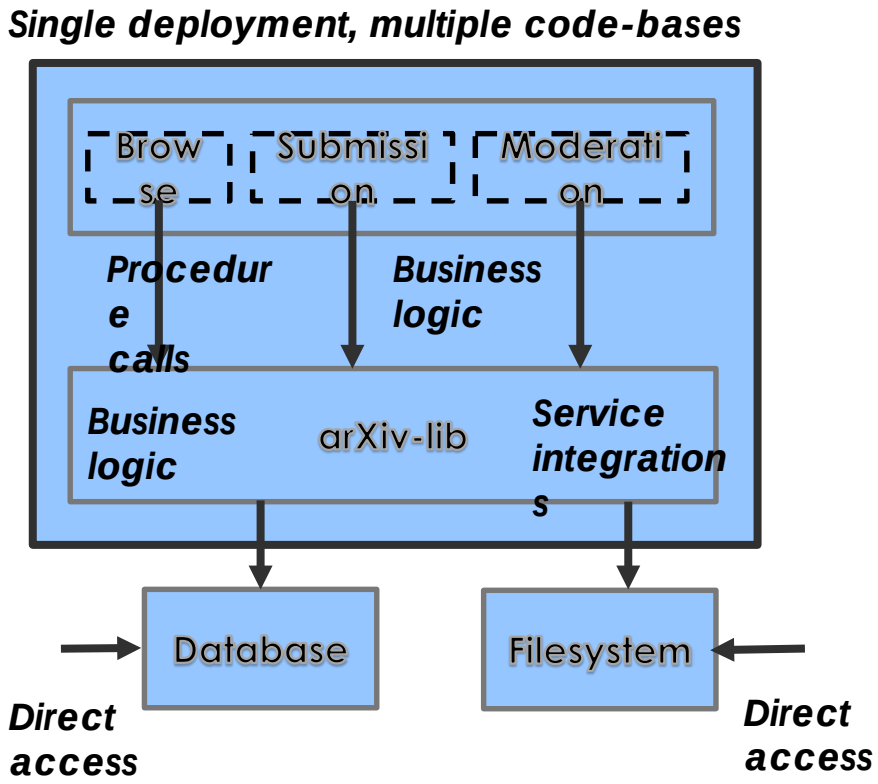
What does arXiv.org have now?

- **Legacy != broken.** "Legacy code is just code that we don't have very good tests for." -- *someone*.
  - The legacy system solved a lot of problems, and many of those drivers still exist today.
  - arXiv is stable, and users are happy.
- **Monoliths aren't necessarily evil.** Esp. for rapid prototyping of new systems, keeping everything close together minimizes unnecessary complexity.



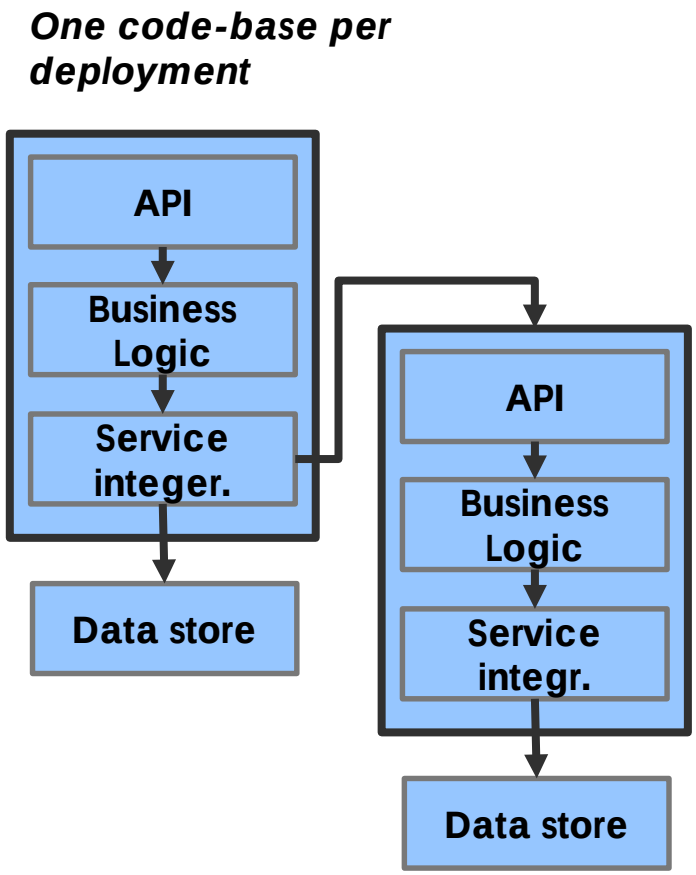
What does arXiv.org have now?

- **Poor isolation/containment** of business logic and dependencies:
  - Hard to test → hard to develop.
  - Hard to locate relevant code → slow to develop.
  - Hard to describe → hard to understand, test.
- **arXiv-lib is a ”high stress” node**: all subsystems depend on it → developers hands are tied.
- **Single/several server paradigm:** limited scaling, poor cost control.



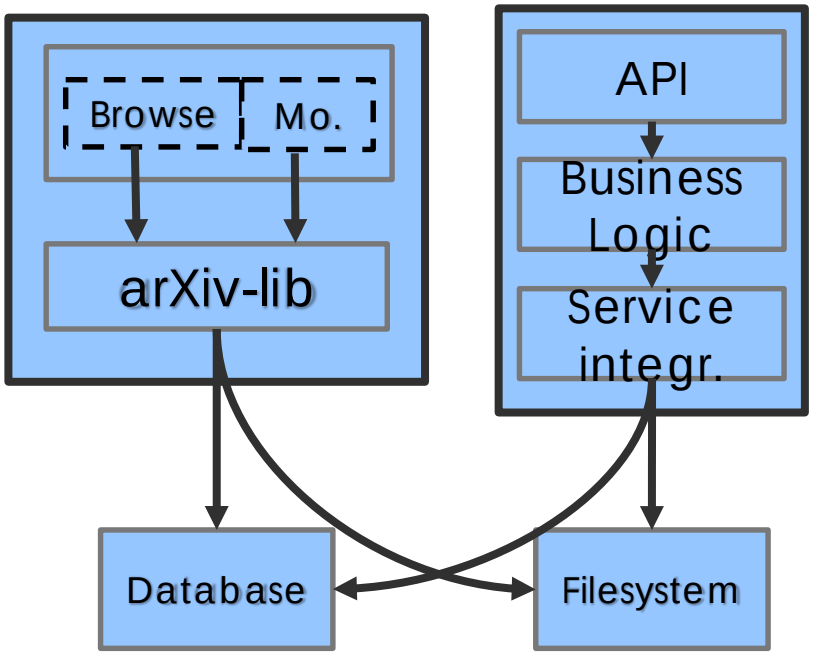
Where is arXiv.org going?

- Fine-grained **isolation with services**: **Python 3** + **Flask web** micro-framework, **Docker** containerization.
- Consistently-applied internal architecture.
- Integration via **REST APIs**, notification broker.
- **Polyglot persistence**: isolated data store, choice of technology matches service requirements.
- Independent scaling, server resources reflect demand.

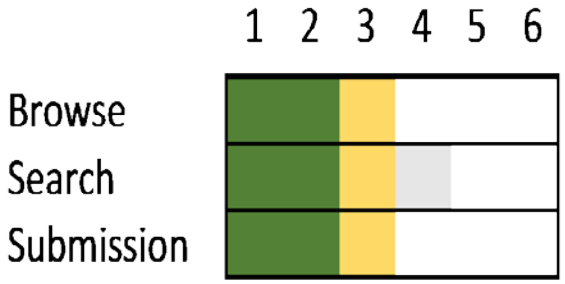


How does arXiv.org get there?

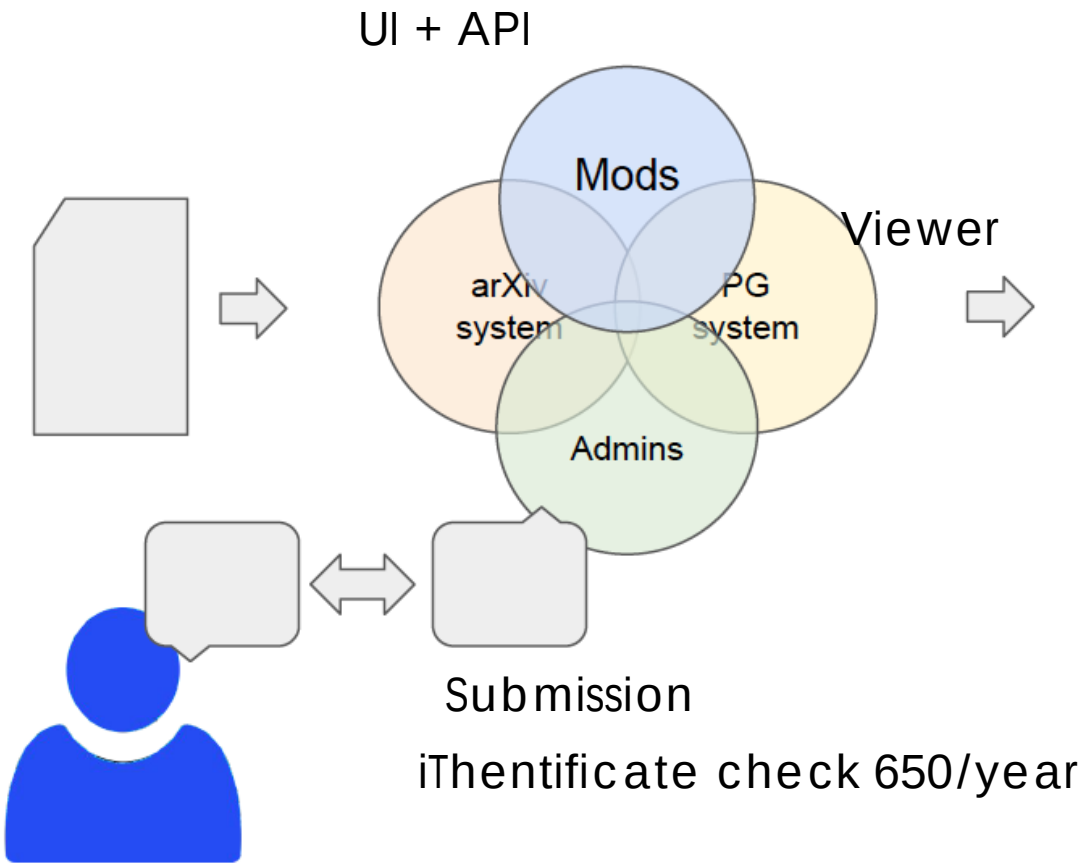
1. **Prioritization:** from the "outside" in.
2. **Identify minimum integrations:** database, filesystem.
3. **Re-engineering:** preserve behavior, but with re-architected codebase.
4. **Local deployment:** services can be deployed on existing web servers.
5. **API gateway integration:** increase access to arXiv content.
6. **Migration to AWS:** as legacy integrations drop off, services are re-deployed in AWS.

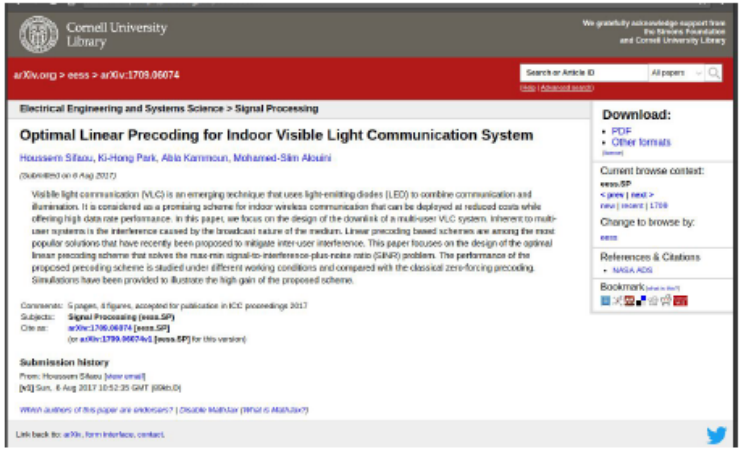


Future: Image and Data ?



Submission & Moderation





Viewer (present)

Cornell University Library

We gratefully acknowledge support from the Simons Foundation and Cornell University Library

arXiv.org > q-bio > arXiv:1708.04855

Search or Article ID | All papers

(Help | Advanced search)

Quantitative Biology > Biomolecules

Title: Entropy Transfer and Dynamics of Allostery in Proteins

Authors: Aysima Hacisuleyman, Burak Erman

(Submitted on 16 Aug 2017)

Abstract: Allostery is an intrinsic spatiotemporal property of all proteins, resulting from long range correlations in the order of several nanometers and time scales of nanoseconds. Information is carried asymmetrically from one part to another by entropy transfer. Here, we present a master equation model of allosteric communication in proteins based on the transfer entropy concept of Schreiber (PRL, 85, 465, 2000). We show how the model relates the path and velocity of asymmetric entropy transfer to conformational transitions over the rugged energy surface of proteins and how this relates to function.

Subjects: Biomolecules (q-bio.BM)
Cite as: arXiv:1708.04855 [q-bio.BM]
(or arXiv:1708.04855v1 [q-bio.BM] for this version)

Which authors of this paper are endorsers? | Disable MathJax (What is MathJax?)

Submission history

From: Burak Erman [view email]
[v1] Wed, 16 Aug 2017 11:57:38 GMT (582kb)

Download:
- PDF only

Current browse context:
q-bio.BM
< prev | next >
new | recent | 1708

Change to browse by:
q-bio

References & Citations
- NASA ADS

Bookmark

References

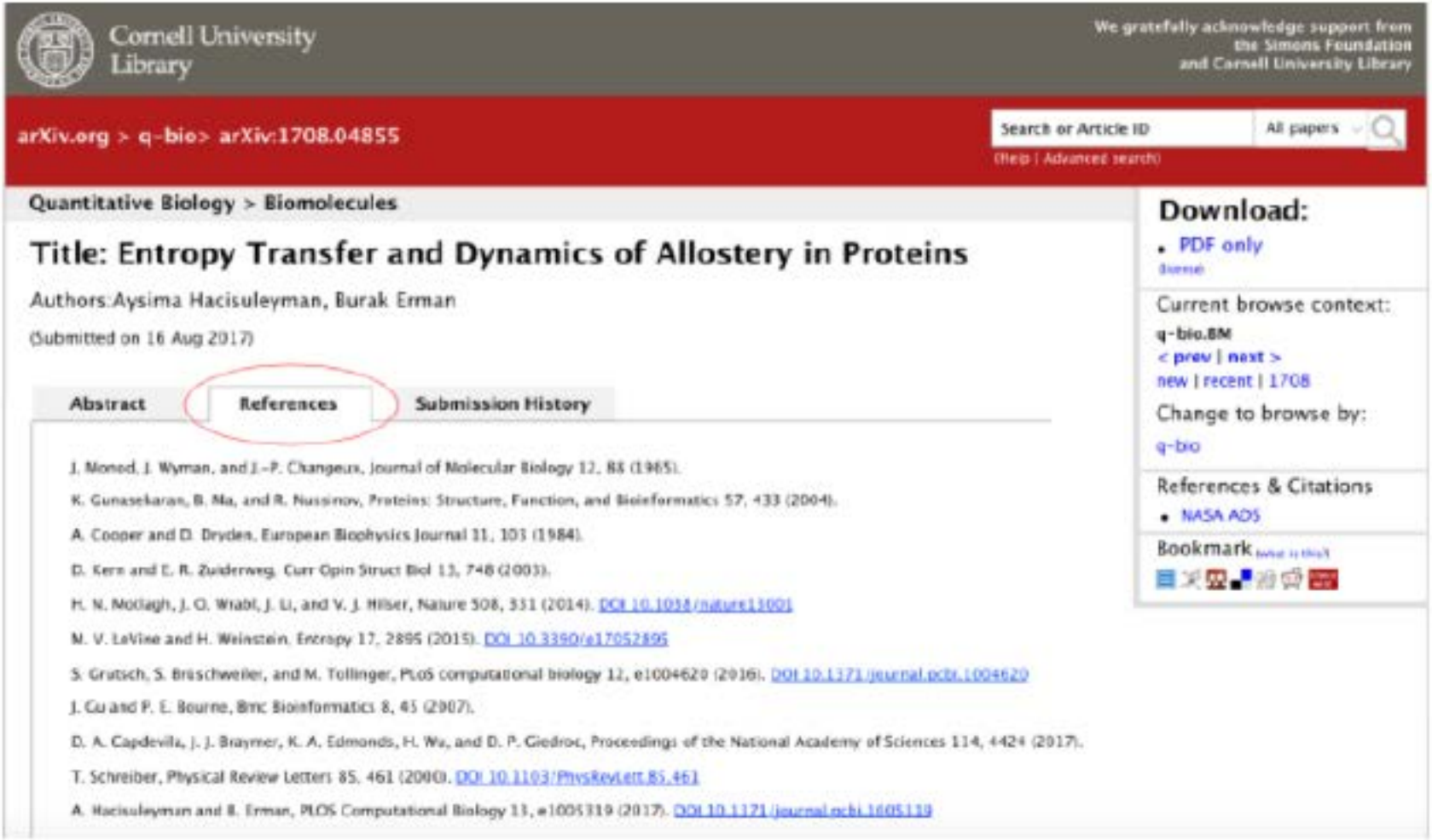
J. Monod, J. Wyman, and J.-P. Changeux, Journal of Molecular Biology 12, 88 (1965).

K. Gunasekaran, B. Ma, and R. Nussinov, Proteins: Structure, Function, and Bioinformatics 57, 433 (2004).

A. Cooper and D. Dryden, European Biophysics Journal 11, 103 (1984).

D. Kern and E. R. Zuiderweg, Curr Opin Struct Biol 13, 748 (2003).

Preferable Viewer



Based on the users' opinions

Organization

arXiv: Roles & Responsibilities

SCIENTIFIC ADVISORY BOARD:

- Provides advice and guidance pertaining to intellectual oversight of arXiv, with particular focus on arXiv's moderation system and criteria for depositing content in arXiv.
- Proposes & reviews proposals for new subject domains.
- Makes recommendations and provides feedback on development projects.

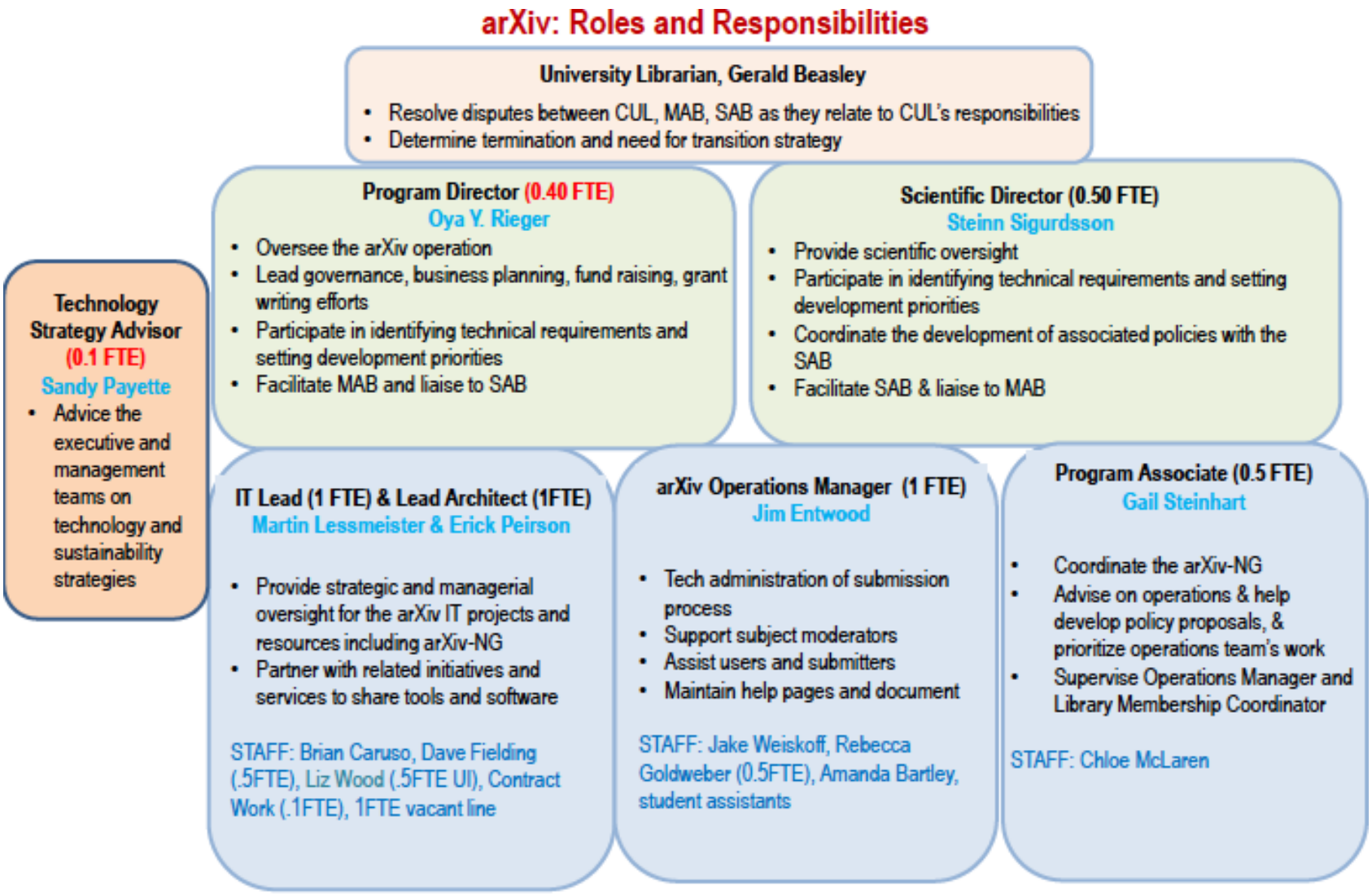
MEMBER ADVISORY BOARD:

- Represents members' interests.
- Advises CUL on development, business planning, outreach and advocacy.

CUL ADMINISTRATION:

- Assumes overall responsibility for arXiv's obligations.
- Provides institutional support and resources for arXiv (HR, business services, legal, etc.).
- Final arbiter for arXiv decisions.

Organization : Stuffs



Collaborations with societies

Societies accepting the proposal to open

MAB suggested in 2016

APS
AIP
ACS
AMS
IOP
ACM

Almost
IEEE

Membership based activities

Membership driven activities

Business model
Subscription model

Hybrid option

Green Open access

arXiv.org-NG

CORNELL UNIVERSITY LIBRARY

Search books, articles and more

Search this site

Home | About Us | Libraries and Hours | Course Reserves | Research | Services | Ask a Librarian | Help | My Account

Home / About Us / News / Story Archive / Alfred P. Sloan Foundation awards grant for arXiv upgrade

NEWS CONTACT

Melanie Lefkowitz

Staff Writer, Editor and Social Media Coordinator

(607) 254-8390

mll9@cornell.edu

libcomm@cornell.edu

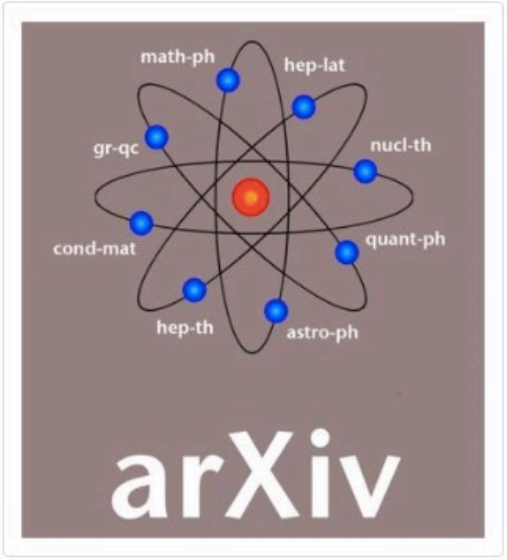
RECENT NEWS

- arXiv.org adds new subject categories
- 'Union Made' exhibit showcases labor and fashion history
- Three projects awarded 2017 digitization grants
- Freedom on the Move project awarded NEH grant
- Medieval fragments yield unexpected discoveries
- arXiv.org recognized for improving scientific research
- A century of agriculture research goes online
- Gerald Beasley named Carl A. Kroch University Librarian

Alfred P. Sloan Foundation awards grant for arXiv upgrade

11/29/16

Ithaca, N.Y. (Nov. 29, 2016) – arXiv.org, the influential open access repository for global scientific research, will begin the first phrase of a three-year overhaul and modernization with the help of a $445,000 grant by the Alfred P. Sloan Foundation.



The Sloan Foundation grant will fund the creation of a detailed blueprint for next-generation arXiv – to be known as arXiv-NG – allowing planners to establish new partnerships, identify necessary resources and build a core development team to lead the modernization process.

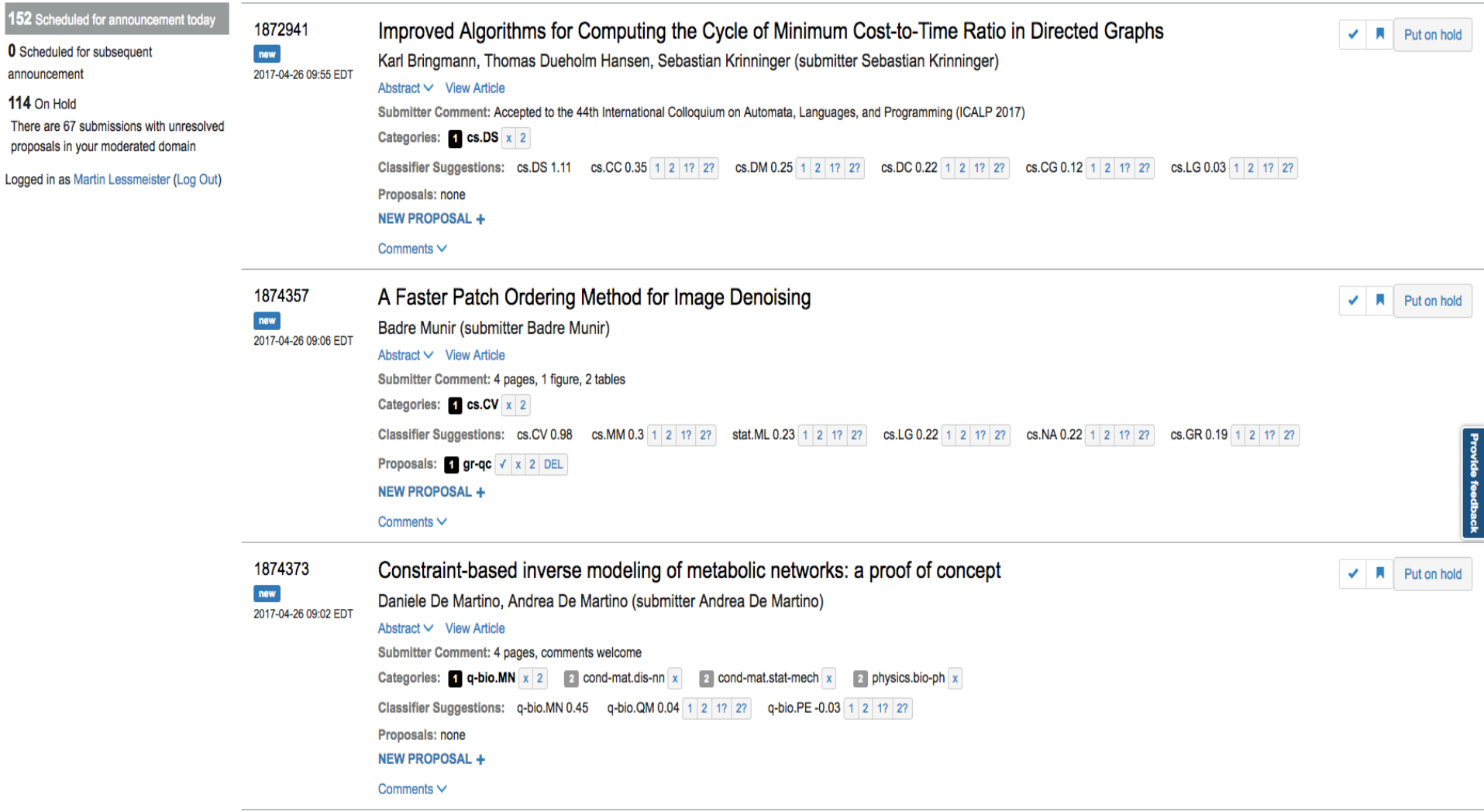
"We're gratified by the Sloan Foundation's support in helping to keep arXiv sustainable and robust," said Oya Rieger, arXiv's program director and associate university librarian for scholarly resources and preservation services at Cornell University Library. "This furthers our efforts to implement

From Part of 2017 Roadmap

- Process
- Moderation tools
- Expansion, TeX, infrastructure
- NG architecture
- Reference extraction
- Search

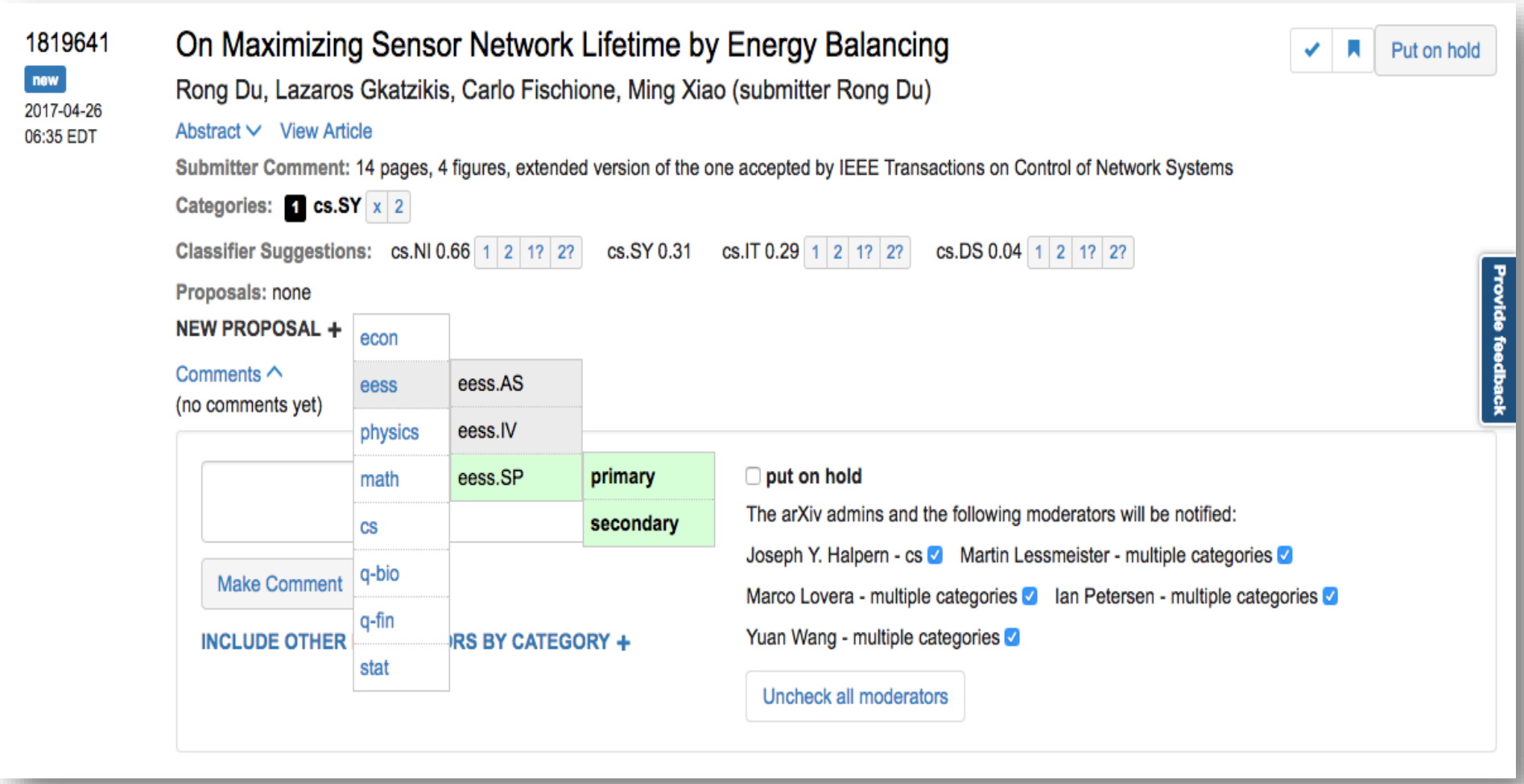
Moderation tools

Now: an entirely new UI + API



Moderation tools

- Single-click actions oriented towards proposals and reclassifications



Moderation tools

- Single-page UI backed by **RESTful** API
- UI built in **Cycle.js**
  - a first foray into modern javascript frameworks
  - valuable experience to inform future decisions around frameworks
- NG “Step 0”

TeX System

- Overhauled TeX system deployed February
- Working on repackaging as containerized service

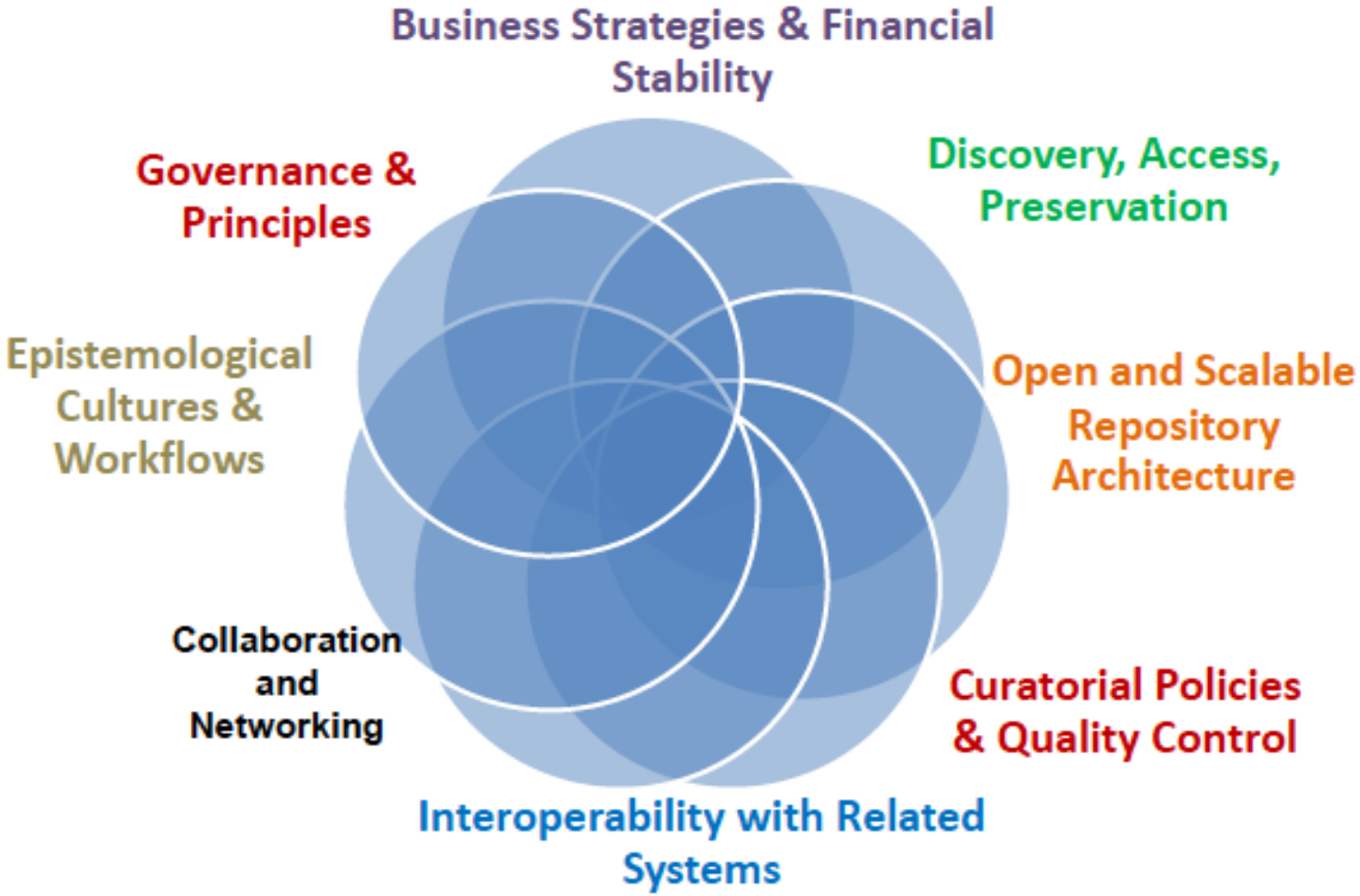
Technology review: a Highlight

- Invenio 3
  - CERN, INSPIRE
  - Highly modular framework
  - Technologies: **Python**/**Flask**, **ElasticSearch**, **Celery**/**RabbitMQ**

NG Architecture

- High- to low-level view of legacy & target architecture, and technology decisions
- Drivers: evolvability, stability, APIs
- Transition from monolith to modular: incremental isolation, re-implementation, and migration to cloud
- Technologies: Python/Flask, Docker
- Integration: REST APIs, notifications

Sustainability



SOURCE: Oya Y. Rieger. Sustainability: Scholarly Repository as an Enterprise. *Bulletin of the American Society for Information Science and Technology*. October/November 2013.; Oya Y. Rieger. Assessing the Value of Open Access Information Systems: Making a Case for Community-Based Sustainability Models. *Journal of Library Administration*. 51:485-506, 2011.

Conclusions

- License to Open License
- Hybrid after Embargo
- **Not change any publishing models, but users**
- Request institutes to accept IR, arXiv, and others