<u>arXiv.org の次世代システムの</u> 公開と戦略

<u>arXiv.org Next Generation</u> <u>-Its Opening and Strategy-</u>

> 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

<u>Video</u>

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

Origin of arXiv.org

arXiv:1108.2700

arXiv.org > cs > arXiv:1108.2700

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 2⁵ 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 (hepph); 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.



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

SOURCE: ARXIV

Preprint archive families









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

Letroscong & Johnskynin warsongens 14020148 Manag 20, 2012	@ Wide 2012
	f the HD 100546 disk seen by chel*
Evidence of a gas-rich, c	arbon-poor atmosphere?
Same Briders III, Swine F. van Dichord II	I Steven D. Doty ^{III} , and Gaugary J. KnowsgII
Son Fond Instruments Internet Arrist, Orenogle con Instrume of Astronomy, IEEE Trans. 1000 (2010), Versawing Gameric Sciences, Instrume Version, Versawing, Directory Depresent of Replan and Astronomy, Discons University, Dr.	i Jalikes, "Bet Proberingel
Tabastul os 7/Destar 2011/ Jorepublis 1484 os 5 January 20	11
Alst	BACT
interval in the strengthen of gravity density of the transmission of gravity-forwards and the transmission of the strength of the transmission of the transmission of the transmission of the transmission of the transmission of the transmission of the strength of the transmission of the transmission of the transmission of the transmission of the transmission of the transmission of the strength of the transmission of the transmission of the transmission of the transmission of the transmission of the transmission of the transmission of the transmission of the transmission of the forwards is a strength of the distance of the transmission of the transmission of the transmission of the transmission of the transmission of the transmission of the transmission of the strength of the transmission of the transmission of the transmission of the transmission of the transmission of the transmission of the transmission of the transmission of the transmission of the transmission of the transmission of the transmission of the transmission of the transmission of the transmission of the t	and one are set in particle that a character activate material and the particle on an activate device of the particle on an activate device of the particle on an activate device of the particle on a structure device of the particle of
Introduction is an important plane of a prangine same rate bills, the set an endeptic his deproduction of the set in self-corrected by in- pendic the da protogrammery deb, the list in depute it, it is a set in the set of the second second second second second of <i>Maintage</i> 2100 for second second second data with a second second second second second second second second second and the second second second second second second second d is never to make more paper fractance. The before the second d is never to make more paper fractance. The before the second d is never to make more paper fractance. The before the second secon	View (27) phases and Xeryi. Through learning of the dust type interceptors a chain wat is driven when any extended depends on a chain (Genzander et al. 1000). Through even al. (2000) for a finite dust of the control dust on the control dust on the control dust on the control dust on the control dust of the control dust on the control dust on the control dust on the control dust on the control dust of the control dust on the control dust of the control dust on the control dust of t

Stell offerter replace to Second Indexe,

ing the chemical composition of the gas by high-starty whitewas again. This idents to to probe the orders chemicaly and the physical spectrar of planes and conservationing some, which your previously macrosolide to us. The carbon budget A standard consequence of the second second

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*

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



arXiv.org

Data: Submission



arXiv New Submissions

500-600/day



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



Left: number of new submissions/year as a function of calendar year for "hep" = High Energy Physics (hep-th+hep-ph+heplat+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



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
- \bigcirc 24,588 | journal reference
- \bigcirc 128,958 | new submission
- \bigcirc 75,693 \mid repeated submission
- \bigcirc 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
- \checkmark Stable unique identifier
- Publishers has no more cost

on Baumol's cost disease

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

- ✓ 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.

Single deployment, multiple code-bases



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.

Single deployment, multiple code-bases



Where is arXiv.org going?

- Fine-grained isolation with services:
 Python 3 + Flask web microframework, 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.





Browse Search Submission 1 2 3 4 5 6



Submission & Moderation





800 papers/day

Viewer (present)



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

Preferable Viewer

Cornell University Library	We gratefally acknowledge support from the Simons Foundation and Carnell University Library					
arXiv.org > q-bio> arXiv:1708.04855	Search or Article ID All papers 🤟 🔾					
	Olega (Advanced search)					
Quantitative Biology > Biomolecules	Download:					
Title: Entropy Transfer and Dynamics of Allostery in Proteins	PDF only dense					
Authors:Aysima Hacisuleyman, Burak Erman (Submitted on 16 Aug 2017)	Current browse context: q-bio.8M < prev next > new recent 1708					
Abstract References Submission History	Change to browse by: 9-bo					
J. Monod, J. Wyman, and JP. Changeux, Journal of Molecular Biology 12, 88 (1965). K. Gunasekaran, B. Ma, and R. Nussinov, Proteins: Structure, Function, and Bioleformatics 57, 433 (2004).	References & Citations					
A. Cooper and D. Dryden, European Biophysics Journal 11, 103 (1984). D. Kern and E. R. Zuiderweg, Curr Opin Struct Biol 13, 748 (2003).	Bookmark www.interation					
H. N. Motlagh, J. O. Wrabi, J. Li, and V. J. Hilber, Nature 508, 331 (2014). <u>DCF 10.1058/nature13001</u> N. V. LeVine and H. Weinstein, Entropy 17, 2895 (2015). <u>DOI 10.3390/e17052895</u>						
S. Grutsch, S. Brüschweiler, and M. Tollinger, PLoS computational biology 12, e1004620 (2016). DOI 10.1371/journal.pcbi.10	04620					
J. Gu and P. E. Bourne, Bmc Bioinformatics 8, 45 (2907).						
D. A. Capdevila, J. J. Braymer, K. A. Edmonds, H. Wa, and D. P. Giedroc, Proceedings of the National Academy of Sciences 114	, 442+ (2017).					
T. Schreiber, Physical Review Letters 85, 461 (2000). DOI 10.1103/PhysRevLett.85.461						
A. Hacisuleyman and B. Erman, PLOS Computational Biology 13, e1005319 (2017). DOI 10.1171/journal.ocbi.1005119						

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



University Librarian, Gerald Beasley

- Resolve disputes between CUL, MAB, SAB as they relate to CUL's responsibilities
- Determine termination and need for transition strategy

Program Director (0.40 FTE) Ova Y. Rieger

- Oversee the arXiv operation
- Lead governance, business planning, fund raising, grant writing efforts
- Participate in identifying technical requirements and setting development priorities
- Facilitate MAB and liaise to SAB

Provide strategic and managerial

resources including arXiv-NG

Partner with related initiatives and

STAFF: Brian Caruso, Dave Fielding

Work (.1FTE), 1FTE vacant line

(.5FTE), Liz Wood (.5FTE UI), Contract

oversight for the arXiv IT projects and

services to share tools and software

Scientific Director (0.50 FTE) Steinn Sigurdsson

- Provide scientific oversight
- Participate in identifying technical requirements and setting development priorities
- Coordinate the development of associated policies with the SAB
- Facilitate SAB & liaise to MAB

Technology Strategy Advisor (0.1 FTE) Sandy Payette

 Advice the executive and management teams on technology and sustainability strategies

IT Lead (1 FTE) & Lead Architect (1FTE) Martin Lessmeister & Erick Peirson Jim Entwood

- Tech administration of submission process
- Support subject moderators
- Assist users and submitters
- Maintain help pages and document

STAFF: Jake Weiskoff, Rebecca Goldweber (0.5FTE), Amanda Bartley, student assistants

Program Associate (0.5 FTE) Gail Steinhart

- Coordinate the arXiv-NG
- Advise on operations & help develop policy proposals, & prioritize operations team's work
- Supervise Operations Manager and Library Membership Coordinator

STAFF: Chloe McLaren

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							and more	Q Search this site	
Home	About Us 🗸	Libraries and Hours	Course Reserves 🗸	Research 🗸	Services 🗸	Ask a Librarian 🗸	Help 🗸		
My Accour	nt								

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 nextgeneration 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



152 Scheduled for announcement today	1872941	Improved Algorithms for Computing the Cycle of Minimum Cost-to-Time Ratio in Directed Graphs	✓ ✓ Put on hold
0 Scheduled for subsequent announcement	new 2017-04-26 09:55 EDT	Karl Bringmann, Thomas Dueholm Hansen, Sebastian Krinninger (submitter Sebastian Krinninger)	
114 On Hold		Abstract V View Article	
There are 67 submissions with unresolved proposals in your moderated domain		Submitter Comment: Accepted to the 44th International Colloquium on Automata, Languages, and Programming (ICALP 2017) Categories: 1 cs.DS x 2	
Logged in as Martin Lessmeister (Log Out)		Classifier Suggestions: cs.DS 1.11 cs.CC 0.35 1 2 1? 2? cs.DM 0.25 1 2 1? 2? cs.DC 0.22 1 2 1? 2? cs.CG 0.12 1 2 1? 2? cs.LG 0.03 1 2 1? 2? Proposals: none	
		NEW PROPOSAL +	
		Comments V	
	1874357	A Faster Patch Ordering Method for Image Denoising	✓ ■ Put on hold
	new 2017-04-26 09:06 EDT	Badre Munir (submitter Badre Munir)	
		Abstract V View Article	
		Submitter Comment: 4 pages, 1 figure, 2 tables	
		Categories: 1 cs.CV x 2	
		Classifier Suggestions: cs.CV 0.98 cs.MM 0.3 1 2 1? 2? stat.ML 0.23 1 2 1? 2? cs.LG 0.22 1 2 1? 2? cs.NA 0.22 1 2 1? 2? cs.GR 0.19 1 2 1? 2?	3
		Proposals: 1 gr-qc 🗸 x 2 DEL	Provide
		NEW PROPOSAL +	feed
		Comments V	back
	1874373	Constraint-based inverse modeling of metabolic networks: a proof of concept	Put on hold
	new 2017-04-26 09:02 EDT	Daniele De Martino, Andrea De Martino (submitter Andrea De Martino)	
		Abstract V View Article	
		Submitter Comment: 4 pages, comments welcome	
		Categories: 1 q-bio.MN x 2 2 cond-mat.dis-nn x 2 cond-mat.stat-mech x 2 physics.bio-ph x	
		Classifier Suggestions: q-bio.MN 0.45 q-bio.QM 0.04 1 2 17 2? q-bio.PE -0.03 1 2 17 2?	
		Proposals: none	
		NEW PROPOSAL +	
		Comments V	

Moderation tools

18

201 06:3 • Single-click actions oriented towards proposals and reclassifications

19641 7-04-26 5 EDT	On Maximizing Sensor Network Lifetime by Energy Balancing ✓ ■ Put on hold Rong Du, Lazaros Gkatzikis, Carlo Fischione, Ming Xiao (submitter Rong Du) Abstract ∨ View Article Submitter Comment: 14 pages, 4 figures, extended version of the one accepted by IEEE Transactions on Control of Network Systems Categories: 1 cs.SY x 2								
	Classifier Suggestions: cs.NI 0.66 1 2 1? 2? cs.SY 0.31 cs.IT 0.29 1 2 1? 2? cs.DS 0.04 1 2 1? 2?								
	Proposals: none								Nou
	NEW PROPOSAL + Comments ^ (no comments yet)	econ							ide fe
		eess	eess.AS						Provide feedback
		physics	eess.IV						×
		math	eess.SP	primary	□ put on hold				
		CS		secondary	The arXiv admins and the following moderators will be notified:				
	Make Comment	q-bio			Joseph Y. Halpern - cs Martin Lessmeister - multiple categories Marco Lovera - multiple categories Ian Petersen - multiple categories		2		
	INCLUDE OTHER	q-fin	RS BY CATEG	OBX T	Yuan Wang - multiple categories 🗹				
		stat	NO DI CAILO						
					Uncheck all moderators				

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"





 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



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