The future of arXiv
and knowledge discovery in open science

First Workshop on Scholarly Document Processing (SDP 2020)

November 19, 2020
Steinn Sigurðsson, Scientific Director
Phys Rev - David Mermin noted that the shelf space will soon be expanding faster than the speed of light, but... will not violate Relativity as no information will be transmitted!
Moved to Cornell Tech
  - spring #2020…

Vice Provost and Dean Prof. Greg Morrisett

new Executive Director - Dr Eleonora Presani
  - 12 staff - 4 part time…
  - ~ 200 volunteer moderators
  - Subject Advisory Committees
  - Science Advisory Board, Member Advisory Board

Funding:
  - Members == Institutional and University Libraries
  - Simons Foundation
  - Cornell
  - Assorted donations and foundations
Wait, all the papers in your field are posted as free PDFs on arXiv? That must be killing big scientific journals, since they charge such huge subscription/publication fees.

Nah, we've been doing it since the 90s and nobody seems to care.

That makes no sense at all!!

Shhh, you'll jinx it!

https://xkcd.com/2085/
.arXiv.org

speed of research

- receive research e-prints email and/or check web
- every morning at coffee or arriving in office
- clean simple interface
- "technical" and $T_E$ — authors vs readers
- source and/or printable
- stable arXiv identifier
- papers on arXiv cited approx. twice as much
Total number of submissions as of November 18, 2020 = 1,796,491.
by the numbers

- 1,796,336 e-prints in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance, Statistics, Electrical Engineering and Systems Science, and Economics
  - ~750+ tagged EMNLP 2020
- over 1,800,000,000 downloads
- 4,800,000 per week
  - == ~10 downloads per second
- over 150,000 new submissions per year
  - ~1/4 in CS and growing
Total number of downloads through October 2020 = 1,835,035,817
Cs.* - big picture

Graph showing the growth of submissions per year in various fields from 1991 to 2019.
The arXiv is not the Internet

arXiv.org

- arXiv moderation
  - light touch, registered users
- curated collection!
- provides identifier, indexing, archive
  - version retention
- heterogenous standards by subject/category...
  - interesting or correct?
- Hard to scale well.
  - over-reliant on friends-of-friends
- **automation and ML/AI**
  - classifier for categorization
  - currently 3 operational; PwC classifier primary closes loop to authors
- **Need to automate most flow if expanding**
  - *edge cases always there and not important*
  - well, except to the authors…
- Holds… Need to head off logjams
- Normative issues - eg Covid-19 papers
- **Beware Gatekeeping**
Classification

- critical for arXiv to be interesting
  - edge cases not uncommon
  - Sturgeon’s Law - keep S/N up
- judicious cross-listing invaluable
  - auto cross-lists
  - vulnerable to blocking interdisciplinary knowledge transfer
  - or overwhelming information transfer
- lateral knowledge transfer
- choice of categories may trigger Arrow’s Theorem… 😃
  - or equivalent…
What’s Next

- refactor code base in situ
  - dockerized python, portable and robust
  - lift to cloud - google cloud
- streamline submission interface
- improved moderation tools
- provide updated API for metadata and e-prints
- open source modules
  - ~ 1/3 code base now open source
- https://github.com/arXiv
Working with:
- ADS, INSPIRE-HEP
- Semantic Scholar & Google Scholar
- Papers with Code
- CORE
- Kaggle
- and others

expanded options outside core arXiv function
Soliciting partners
- through arXiv Labs
third party and arxiv source
- eg. hepth.io arxiv-sanity.com, arxiv-vanity.com
arXiv Labs

arXiv is surrounded by a community of researchers and developers working at the cutting edge of information science and technology. While the arXiv team is focused on our core mission—providing rapid dissemination of research findings at no cost to readers and submitters—we are excited to be experimenting with a small number of collaborators on projects that add value for our stakeholders and advance research.

Here are some of the projects that our collaborators are working on right now.

**arXiv Links to Code**

Collaborators:
- Robert Stojnic
  - Papers with Code / Facebook AI Research
- Viktor Kerkez
  - Papers with Code / Facebook AI Research
- Ludovic Vlaud
  - Papers with Code / Facebook AI Research

**Code:** https://github.com/arXiv/arxiv-browse/tree/develop/browse/static/js/paperswithcode.js

**arXiv Links to Code** aims to provide an easy and convenient way to find relevant code for a paper. It is using data from Papers with Code - a free resource that links papers, code and results in Machine Learning. Papers with Code is the biggest such resource and is licensed under an open license.

**CORE Recommender**

Explore relevant open access papers from across a global network of research repositories while browsing arXiv. Research

https://labs.arxiv.org
Looking to provide additional functionality outside core services
- Metadata and full text search - Kaggle - for NLP
- https://www.kaggle.com/Cornell-University/arxiv
- software - Papers with Code, preliminary Code Ocean
- Knowledge Discovery - CORE, ADS ++
- Author links and bibliography services ++: Semantic, Google Scholars

Planned
- DOIs
- improved metadata
- Supplemental data links
- dark archive
- ++
Future

- **Ambitions**
  - content type
  - customized content service and interfaces
  - personalized discovery
  - smart discovery - “the unknown unknowns…”
  - subject area expansion
  - overlay journals
  - third party comment and discussion options

Caveat: $$$
We need people and resources
The Future of arXiv
• Did I forget anything?!
