The Transformation of Scholarly Communication

Simeon Warner

simeon@cs.cornell.edu

http://www.cs.cornell.edu/people/simeon

ALPSP, 21st International Learned Journals Seminar, London, 8th April 2005

Acknowledgements

This work in collaboration with:

- Carl Lagoze (Cornell)
- Sandy Payette (Cornell)
- Herbert van de Sompel (LANL)
- John Erickson (HP Labs)

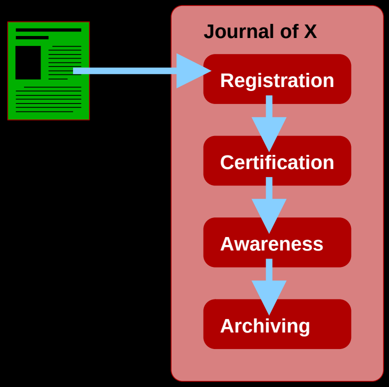
DLib paper: *Rethinking Scholarly Communication: Building the System that Scholars Deserve* (September 2004)

Access

Currently very active debate about open access to scholarly material and possible business models. Two observations:

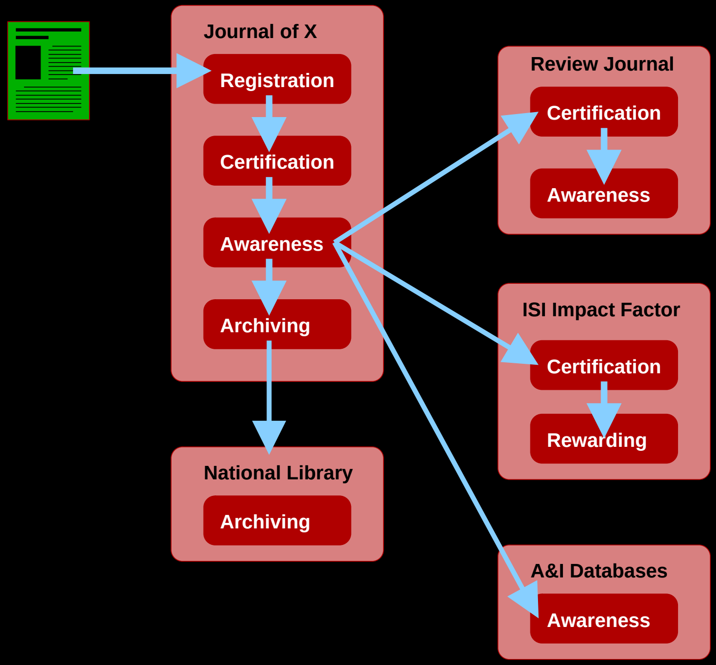
- Subscription journals make data available to Google etc.
  ⟶ Demonstrates understood value of global services.
- Strong evidence that openly available articles have greater impact (citations)
  ⟶ Is this because of readers who have access only to openly available material or because openly available material is just easier to find?

A stovepipe view of scholarly publishing



Simplistic but shows how journal publishing neatly packages the *value chain* of functions identified by Roosendaal and Guerts.

The stovepipe with bells and whistles



Add: A&I services for greater awareness; review journals for second level of certification; and *rewarding* function, most commonly via ISI *impact factor*.

Interoperability now: Of PDF and DOI

Currently limited interoperability. Two examples:

PDF

A standard electronic format. Users need only one viewing technology for all journals. Positive side effect is that tools indexing PDF manage to provide quite effective search tools based on harvested content.

DOI (Digital Object Identifiers)

DOIs link between different systems via a mutually trusted and fair intermediary. They can also provide a new identifier space: Google Scholar uses DOI-based URLs to index and link to APS articles.

The old problem of information overload

> *It is certainly impossible for any person who wishes to devote a portion of his time to chemical experiment, to read all the books and papers that are published in connection with his pursuit; their number is immense, and the labour of winnowing out the few experimental and theoretical truths which in many of them are embarrassed by a very large proportion of uninteresting matter, of imagination, and of error, is such that most persons who try the experiment are quickly induced to make a selection in their reading, and thus inadvertently, at times, pass by what is really good.* [Michael Faraday, 1826]

Information overload

I argue that this problem is best tackled with the help of automated tools and agents.

The better the information available — metadata, full text, citation data, certification information etc. — the better that our agents will be able to help with this selection.

This is a powerful motivation for the creation of a machine traversable and understandable network of scholarly information.

Changing practices

As the practice of scholarship changes, so are scholarly communication practices:

> *The traditional, linear, batch processing approach is changing to a process of continuous refinement as scholars write, review, annotate, and revise in near-real time using the Internet.* [NSF Cyberinfrastructure report, 2004]

Communication mirroring changing practices

To mirror practices the communication system must:

- be closely coupled to the scholarly endeavor;
- include data, simulations and informal results alongside formal peer-reviewed documents;
- facilitate collaboration and varying degrees of access and sharing; and
- enable the scholarly record to be preserved.

Datasets and the Grid

The Grid is an set of technologies that provide for large-scale distributed data storage and computation in disciplines such as:

- genomics, high-energy physics, astronomy and climate modeling.

Storage is now measured in *petabytes* (millions of gigabytes) and even the collaborations are huge (hundreds or thousands of people).

At present, grids exist separate from much of the rest of the scholarly communication infrastructure and there is a need to provide interoperability so that data, code and visualizations can be effectively included in the scholarly record.

Recording scholarship

Consider a paper presenting an analysis of several terabytes of data stored by the US National Virtual Observatory. A complete record of this work should include the software and dataset. It is clearly not feasible to store a copy with the paper, instead it must be clearly and unambiguously included by reference.

Challenges include:

- facilitating early registration of communication units,
- integration of heterogeneous data streams,
- recording and exposing provenance,
- ensuring integrity of complex documents.

Preservation as one of many services

Expanded unit of communication $\Longrightarrow$ expanded notion of preservable unit.

Early registration allows the possibility of preservation separate from the conventional sequence of scholarly publication.

View preservation as another service that can be achieved through various pathways.

What can we learn from the web?

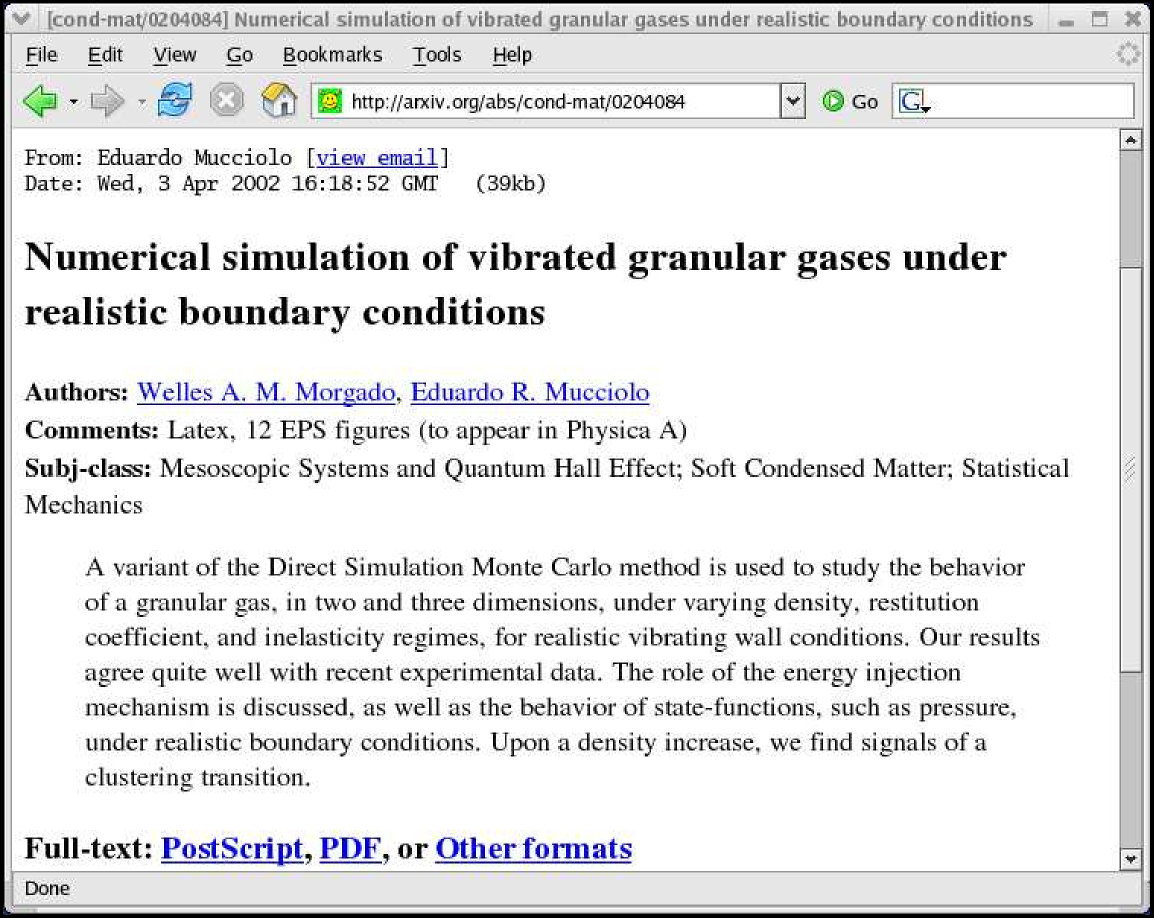
Two key architectural elements:

- **URLs** - uniform linking strategy (*http://.../*)
- **HTML** - standard formats

Add simple-to-use browsers and *voilà*, the web.

Next, add robots and spiders, build indexes, and create search engines. For certain types of search, web search engines are wonderful yet for others they are dreadful. What determines success or failure?

Searching for `Morgado AND Mucciolo` (cond-mat/0204084)


[cond-mat/0204084] Numerical simulation of vibrated granular gases under realistic boundary conditions
File Edit View Go Bookmarks Tools Help
http://arxiv.org/abs/cond-mat/0204084
Go
Done

From: Eduardo Mucciolo [view email]
Date: Wed, 3 Apr 2002 16:18:52 GMT   (39kb)

Numerical simulation of vibrated granular gases under realistic boundary conditions

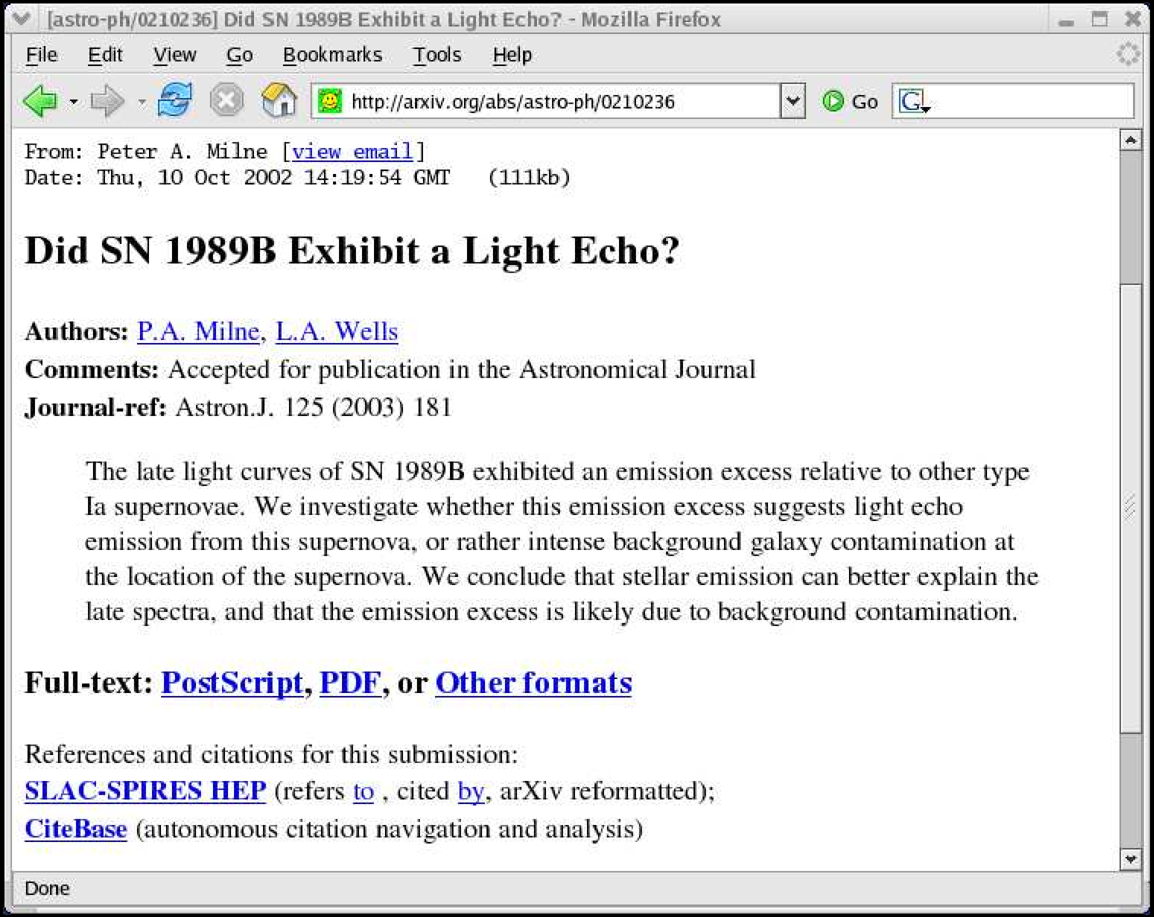
Authors: Welles A. M. Morgado, Eduardo R. Mucciolo
Comments: Latex, 12 EPS figures (to appear in Physica A)
Subj-class: Mesoscopic Systems and Quantum Hall Effect; Soft Condensed Matter; Statistical Mechanics

> A variant of the Direct Simulation Monte Carlo method is used to study the behavior of a granular gas, in two and three dimensions, under varying density, restitution coefficient, and inelasticity regimes, for realistic vibrating wall conditions. Our results agree quite well with recent experimental data. The role of the energy injection mechanism is discussed, as well as the behavior of state-functions, such as pressure, under realistic boundary conditions. Upon a density increase, we find signals of a clustering transition.

Full-text: PostScript, PDF, or Other formats

Items with unusual strings are easy to find with web search.

Searching for `Milne and Wells` (astro-ph/0210236)



[astro-ph/0210236] Did SN 1989B Exhibit a Light Echo? - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://arxiv.org/abs/astro-ph/0210236
Go
Done

From: Peter A. Milne [view email]
Date: Thu, 10 Oct 2002 14:19:54 GMT (111kb)

Did SN 1989B Exhibit a Light Echo?

Authors: P.A. Milne, L.A. Wells
Comments: Accepted for publication in the Astronomical Journal
Journal-ref: Astron.J. 125 (2003) 181

The late light curves of SN 1989B exhibited an emission excess relative to other type Ia supernovae. We investigate whether this emission excess suggests light echo emission from this supernova, or rather intense background galaxy contamination at the location of the supernova. We conclude that stellar emission can better explain the late spectra, and that the emission excess is likely due to background contamination.

Full-text: PostScript, PDF, or Other formats

References and citations for this submission:
SLAC-SPIRES HEP (refers to , cited by, arXiv reformatted);
CiteBase (autonomous citation navigation and analysis)

Items with common strings are hard to find with web search. To improve this, search engines try to interpret data that was formatted for humans, they try to recover semantics.

The semantic web

Creating a machine understandable version of the web where information is structured as assertions:

- the string “Fred Bloggs” is the author’s name

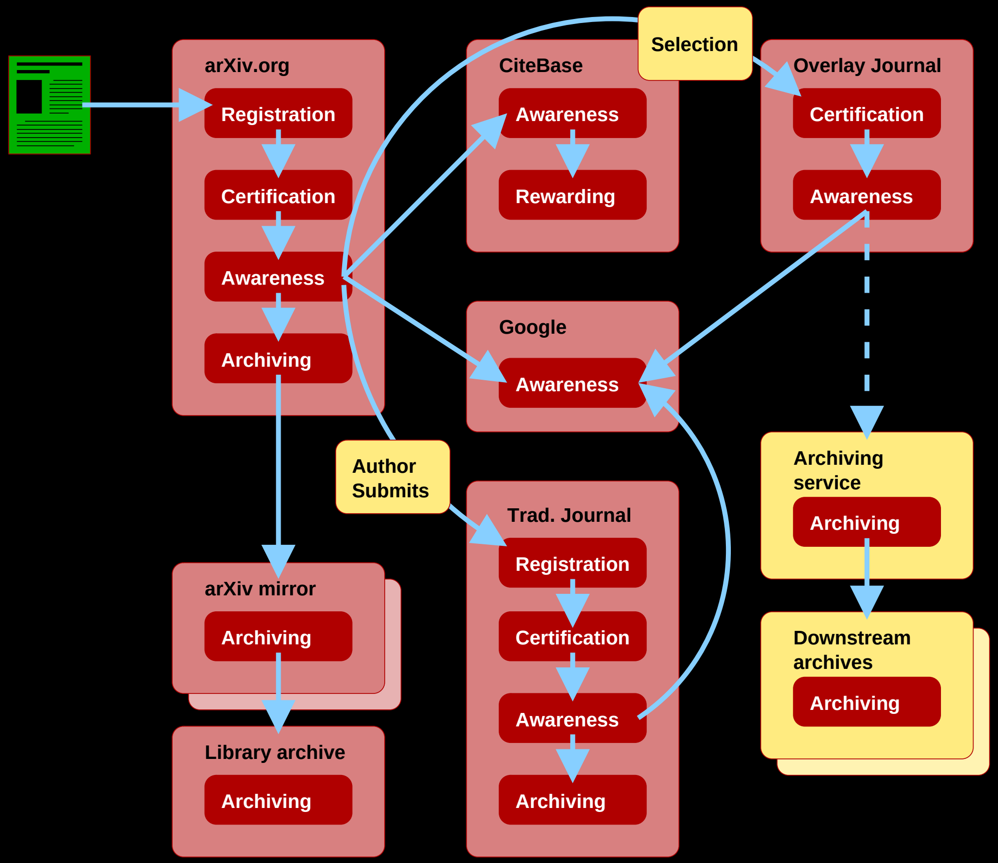
actually something more specific:

- the string “Fred Bloggs” is the author’s name in the sense defined by the British Library

add thesauri and ontologies:

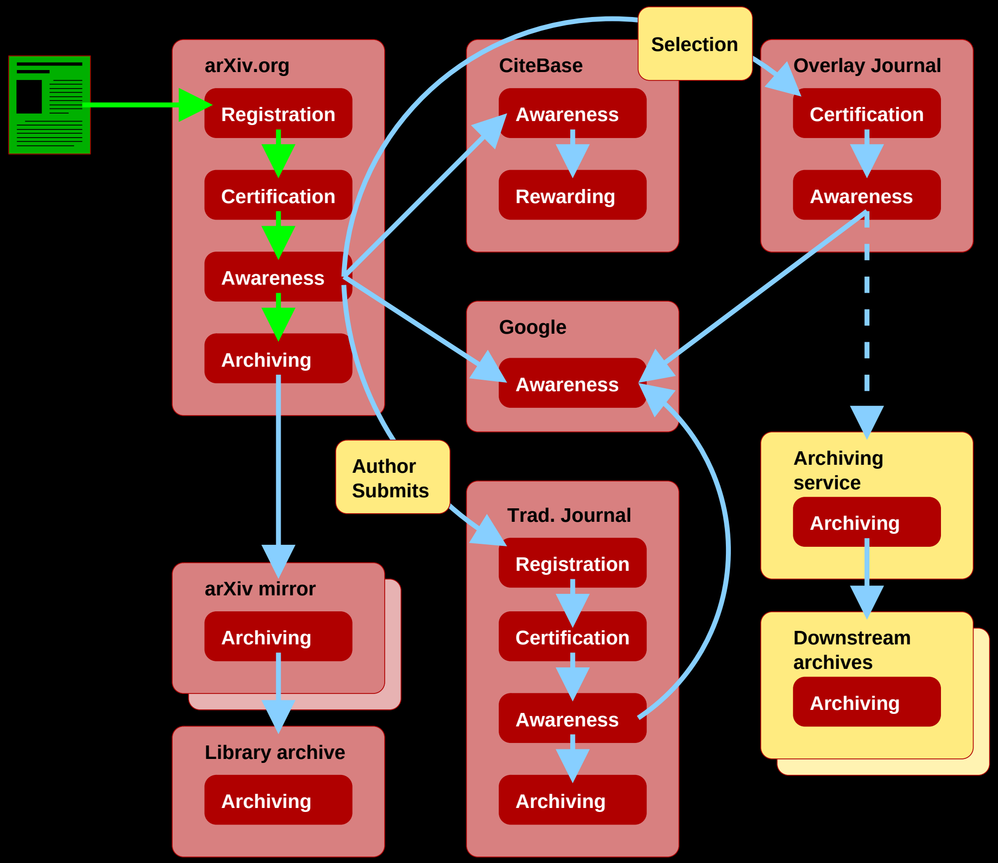
- author in the sense defined by the British Library is the same as author in the sense defined by the Library of Congress
- author is a type of person

arXiv ecology



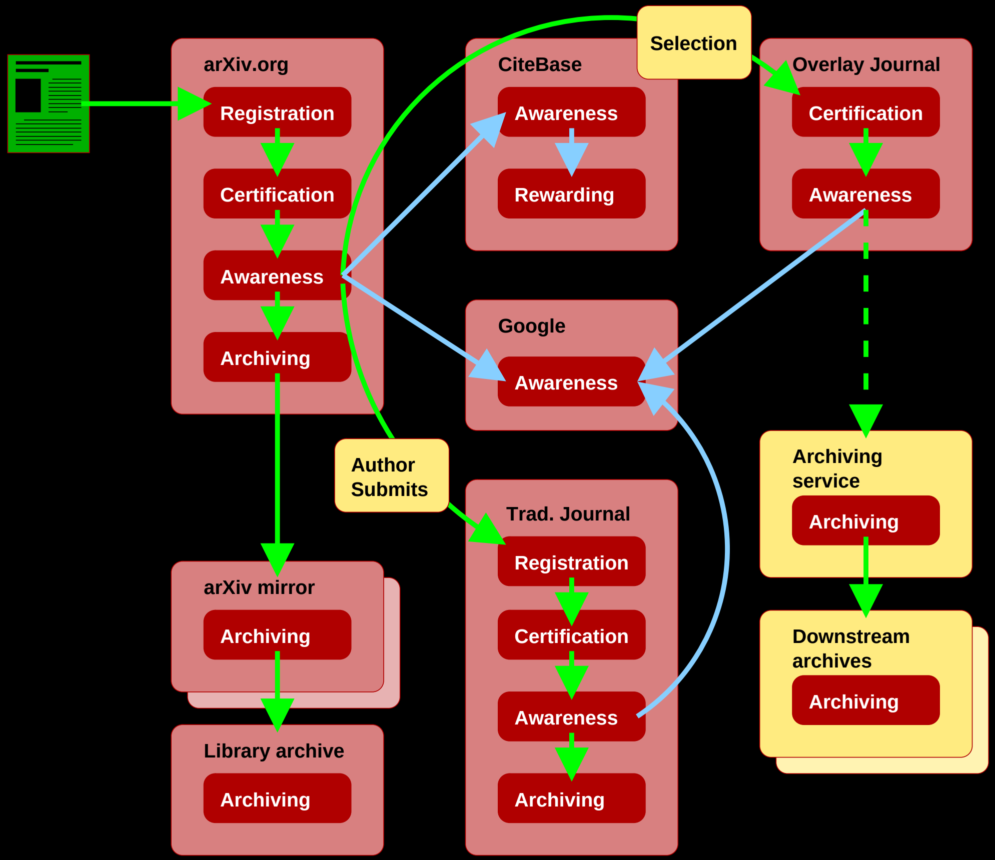
Rich ecology because of data sharing with other services.

Pathways within arXiv



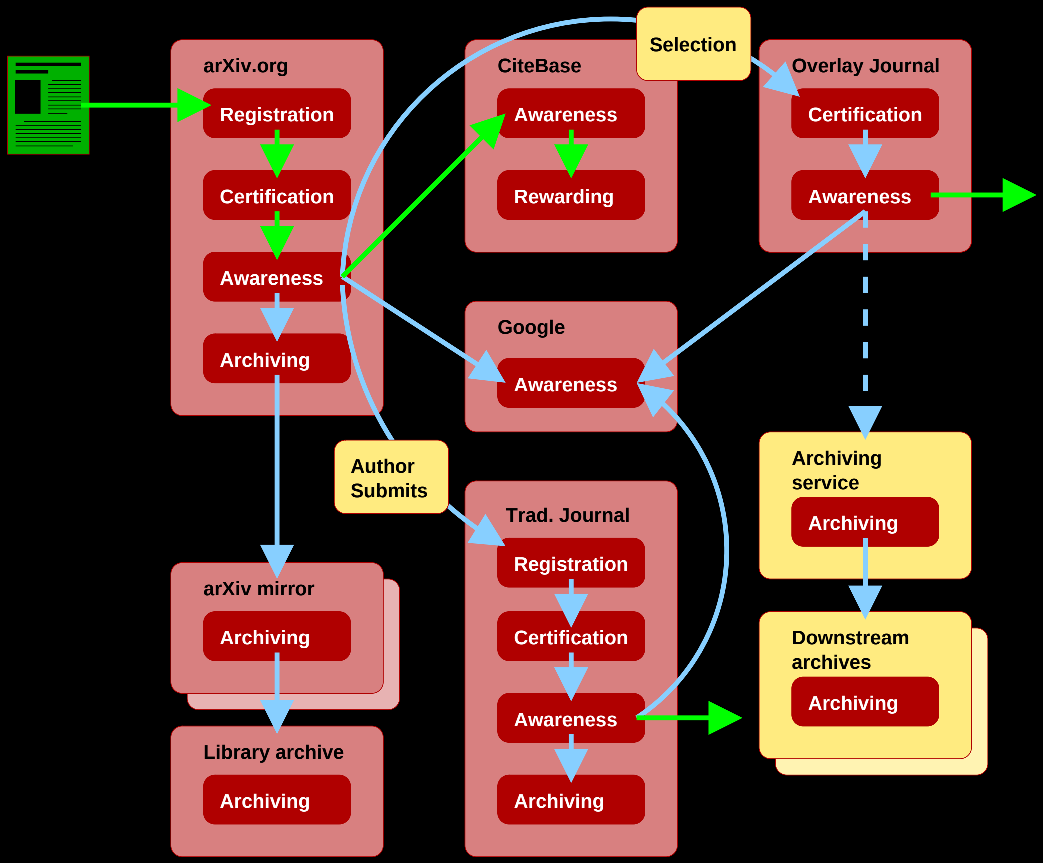
Registration on submission → (weak) certification → awareness (website, alerts) → archiving.

Archiving pathways



Many archiving pathways: local archiving, mirror network, Bibliothéque Nationale, through traditional or overlay journals. Archiving services to come?

Rewarding pathways



Novel rewarding pathways through CiteBase: citation analysis and readership estimates.

Concluding remarks

The scholarly communication system must adapt with developing research methodologies. It should support collaborative, network-based and data-intensive practices.

- the scholarly communication system must be innately digital
- it must support a much expanded unit of communication that may be heterogeneous and distributed
- it must provide for many different pathways that fulfill some or all of the necessary communication functions

The challenge

The challenge for publishers is to identify appropriate functions that really add value, and to implement them in a truly networked fashion that will best serve the community.

That's all folks...