Cross-Repository Interoperability

Simeon Warner

simeon@cs.cornell.edu

Open Scholarship 2006: New Challenges for Open Access Repositories

The University of Glasgow, 18–20 October 2006

Acknowledgements

This work in collaboration with: Carl Lagoze (Cornell), Sandy Payette (Cornell), Herbert Van de Sompel (LANL), Xiaoming Liu (LANL), Jeroen Bekaert (Ghent).

Based in part on a vision described in:

Rethinking scholarly communication: Building the system that scholars deserve. Herbert Van de Sompel, Sandy Payette, John Erickson, Carl Lagoze, and Simeon Warner. *D-Lib Magazine*, 10(9), 2004. doi:10.1045/september2004-vandesompel.

Background: 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.

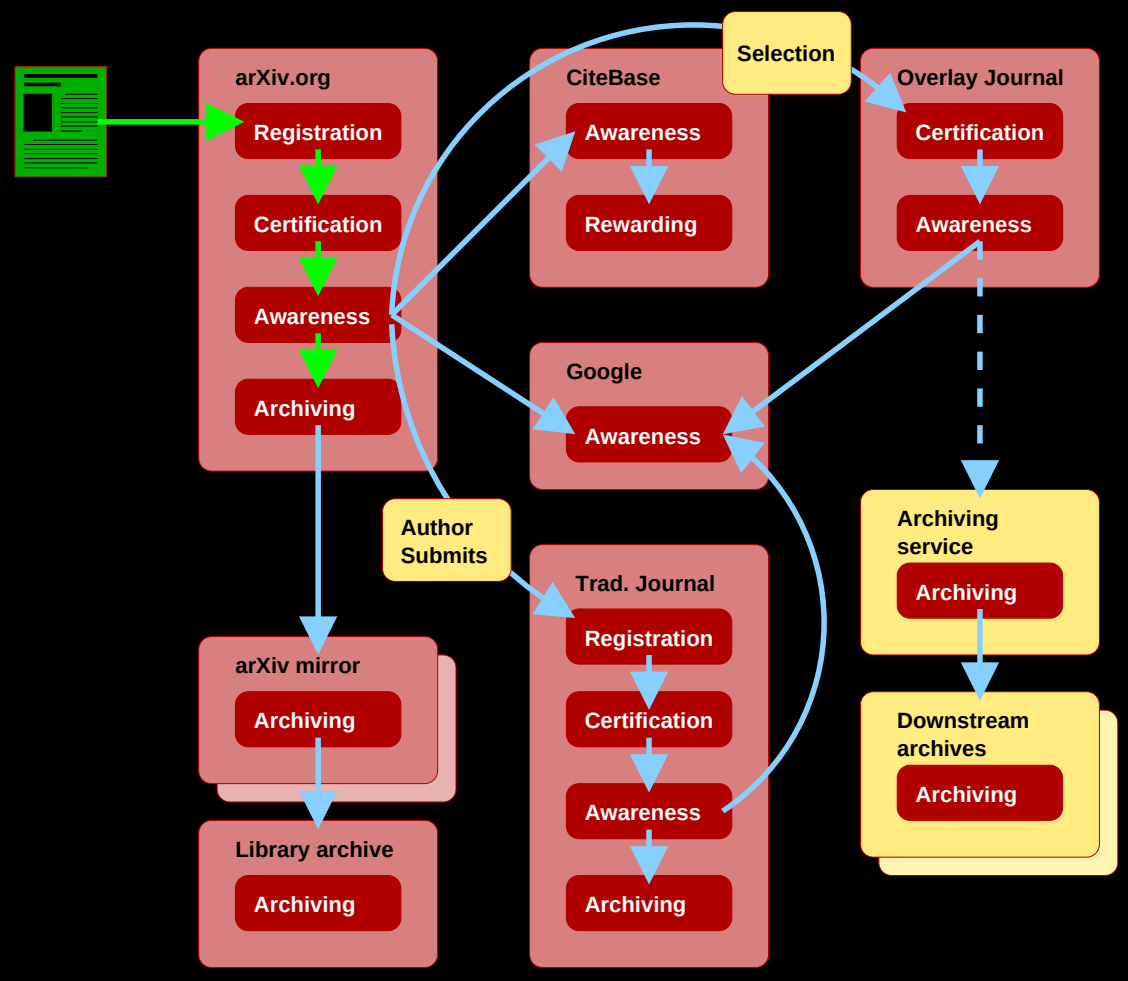
Recording scholarship

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 (by-reference).

Challenges include:

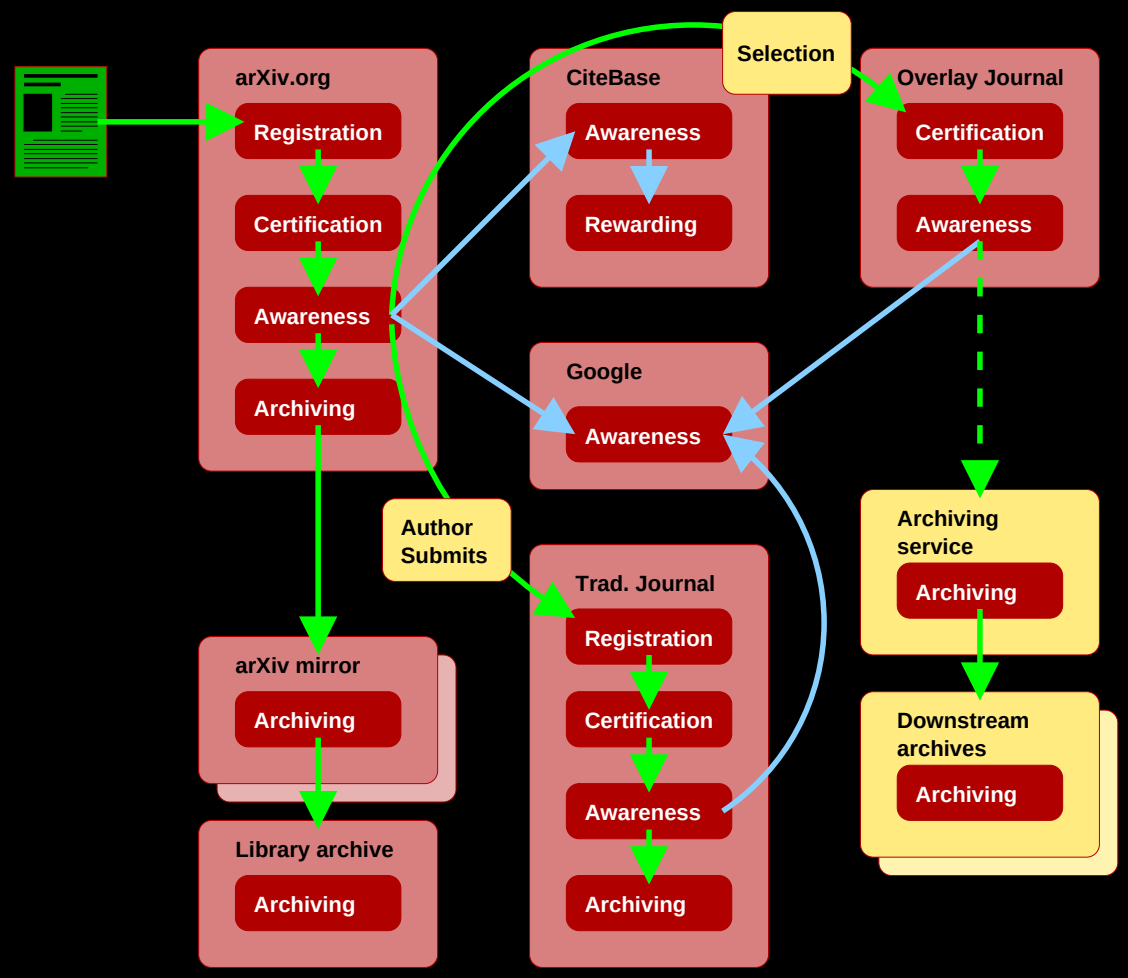
- facilitating early registration of communication units,
- integration of heterogeneous data streams,
- recording and exposing provenance,
- ensuring integrity of complex documents.

Pathways within arXiv



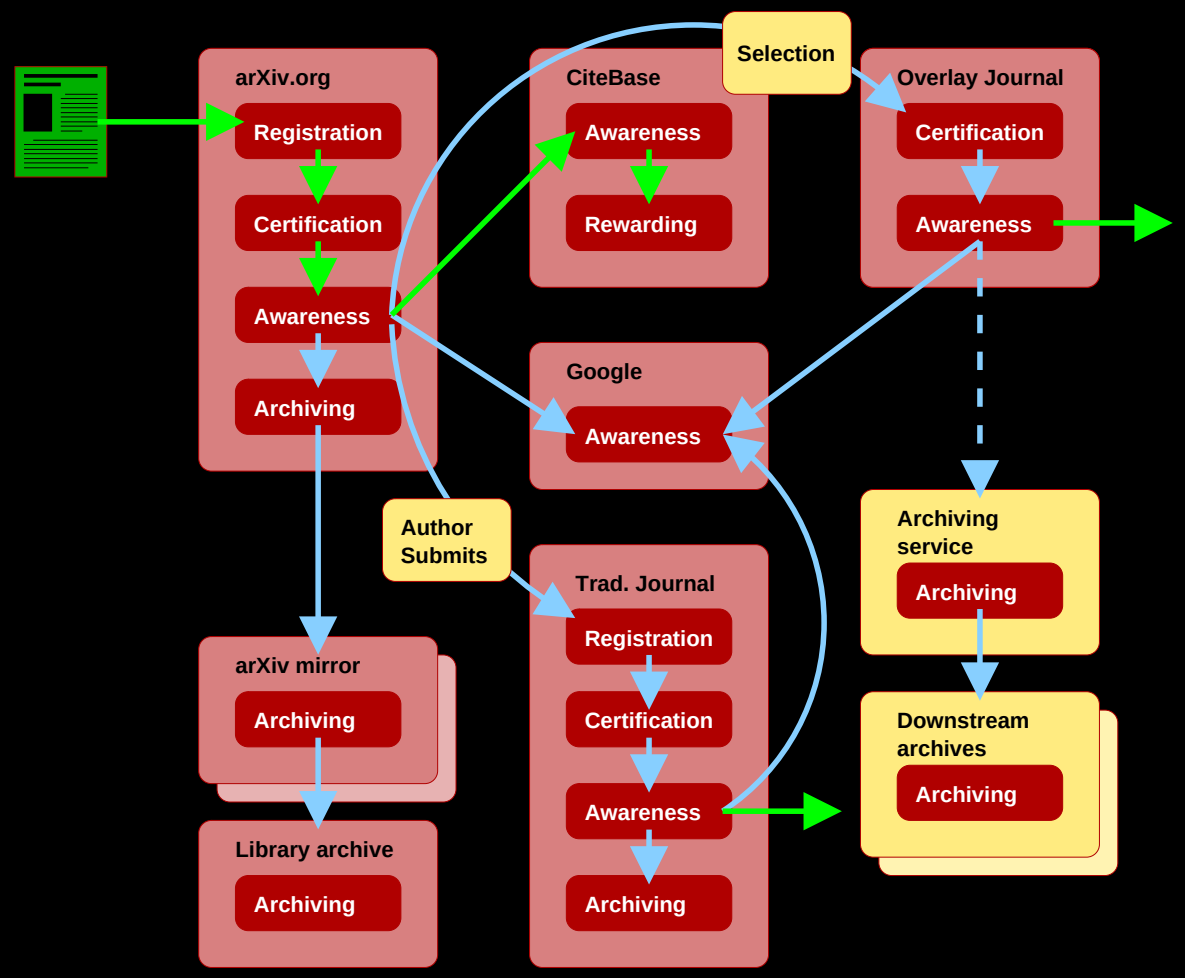
Registration on submission → (weak) certification → awareness (website, alerts) → archiving.

Archiving pathways



Local archiving, mirror network, traditional and overlay journals. Archiving services to come?

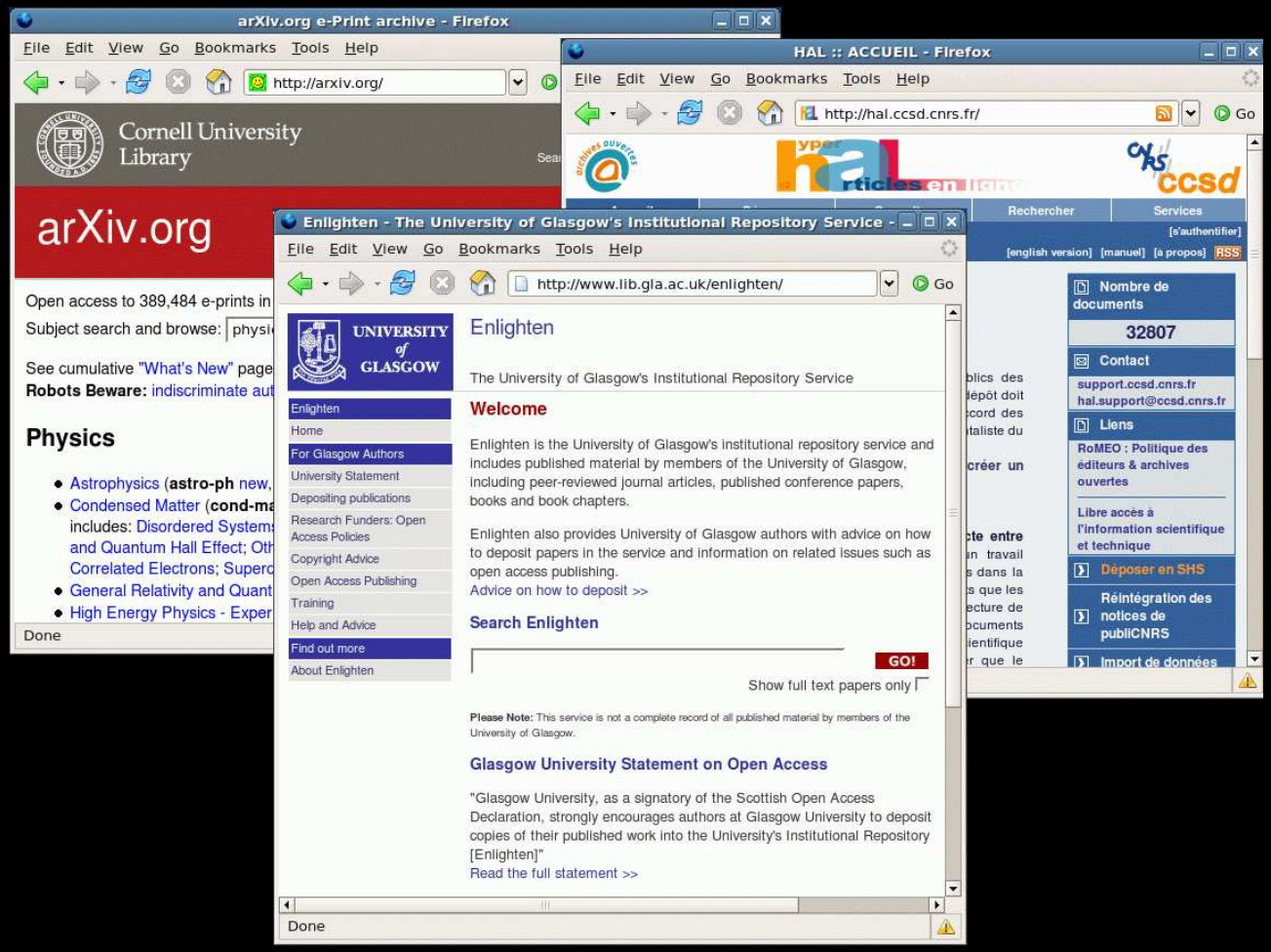
Rewarding pathways



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

Q. What is the current state of repository interoperability?

Web UI — pervasive, if limited, interoperability



TCP ... HTTP ... HTML ... PDF ... Browser.

Openly available standards with free implementations.

Web harvesting and search engines

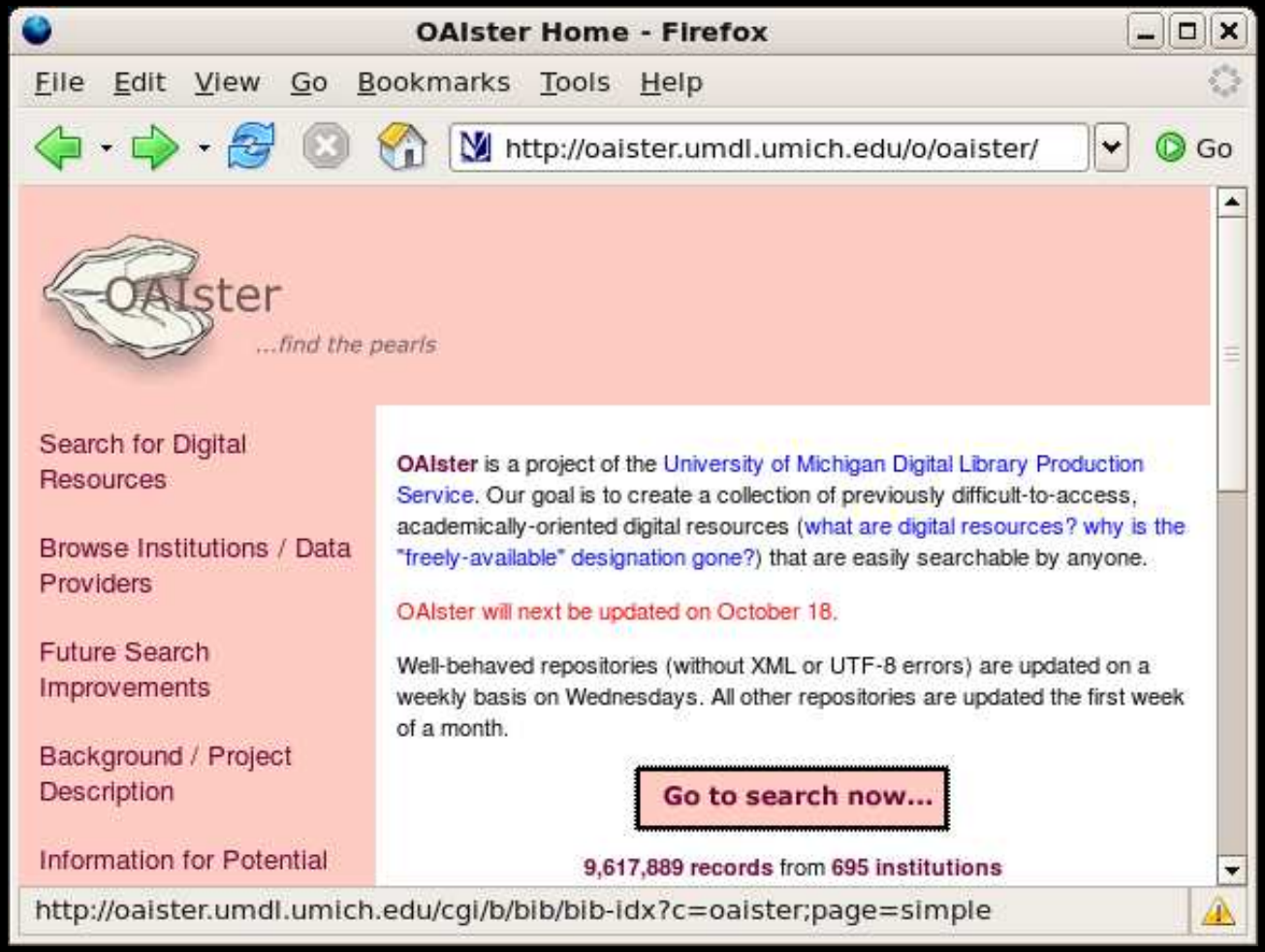
Google, Yahoo!, MSN... pretty good.

Recover semantics $\longrightarrow$ Google Scholar, Citeseer.

OAI-PMH

Share/harvest metadata (or any XML data).

Search and other services over distributed repositories.



Other interoperability elements

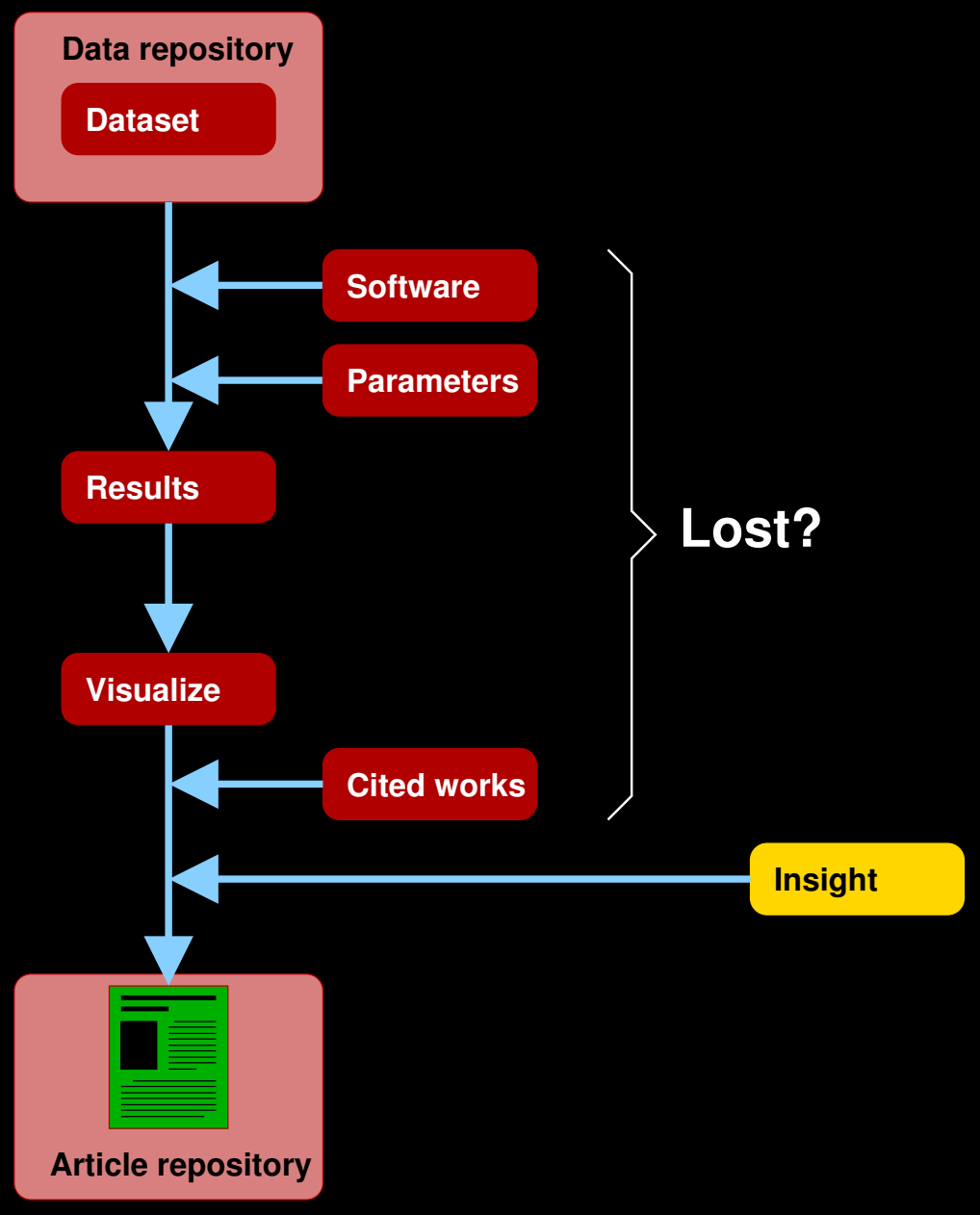
- **XML, Unicode** — done deal.
- **RSS, Atom** — similar mechanics to OAI-PMH, different use.
- **Identifiers and resolution** — URLs, Handles (DOI especially), info URI $\longrightarrow$ URI done deal?
- **Beyond e-paper?** — XML document formats (NLM dtd).
- **Rights** — Creative Commons, GFDL...
- **Usage data** — valuable but some dangers.
- **Format registries** — PRONOM and GDFR.

Q. **What should interoperability mean?**

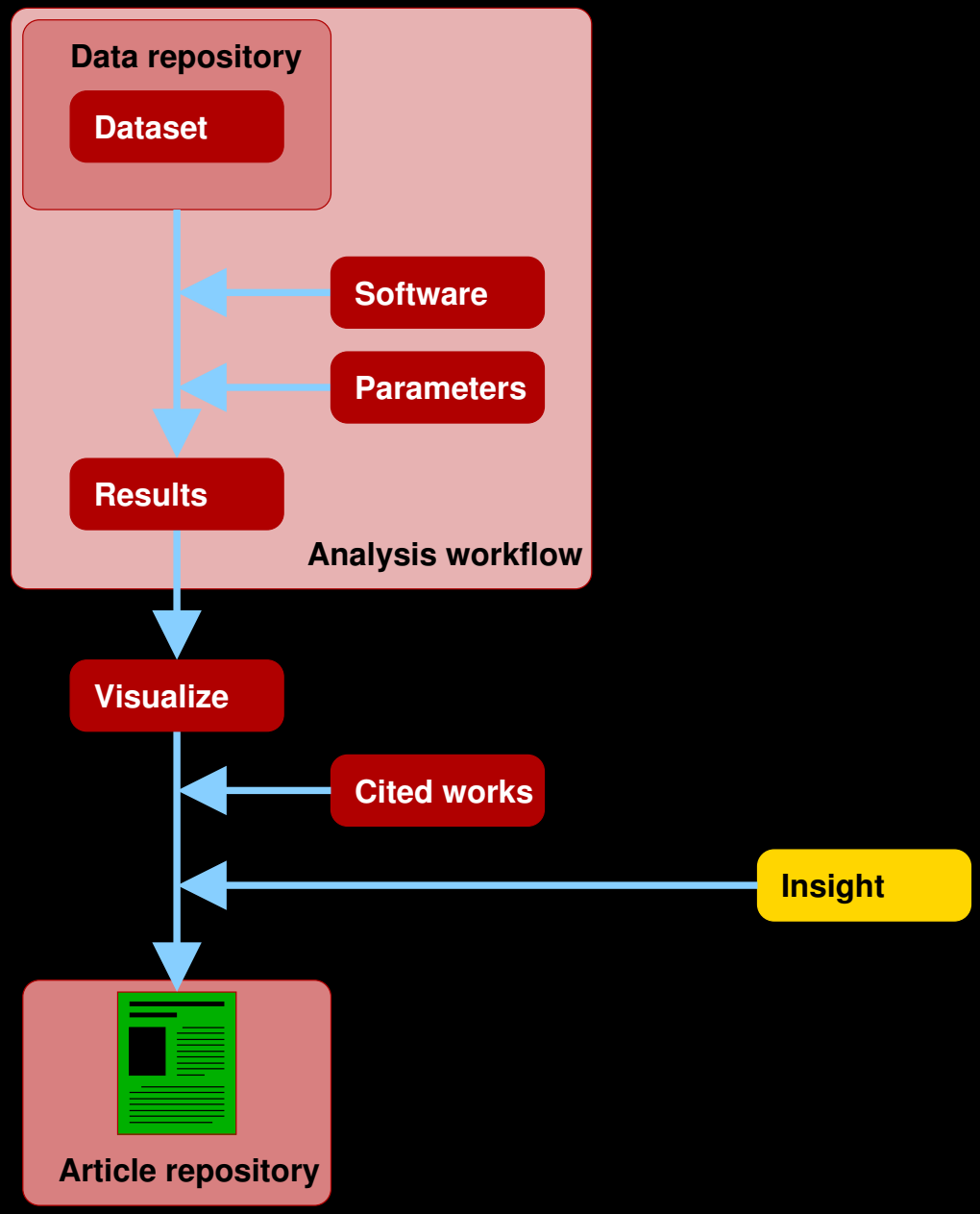
- **Improved linking** — between document repositories and between document, image and data repositories. E.g. US NLM linking between literature and bioinformatics databases, astronomical community linking to image and data catalogues.
- **Better discovery across repositories** — search in context, browsing and ranking based on many metrics, combined document/actor networks, similarity measures..
- **Overlayed tools** — can't base everything on harvesting, need service interfaces (e.g. Entrez). *For this we have to get over the idea of holding repository content hostage in return for UI traffic ransom.*
- ...

- **Provenance** — Key notion within scholarly communication!
  - Citation — currently necessary to recover citations from plain text. This is nuts.
  - Article creation — how to link within-system workflows (e.g. Taverna for data collection, curation, processing, analysis and inference), all the way up to articles which are then cited and re-used.

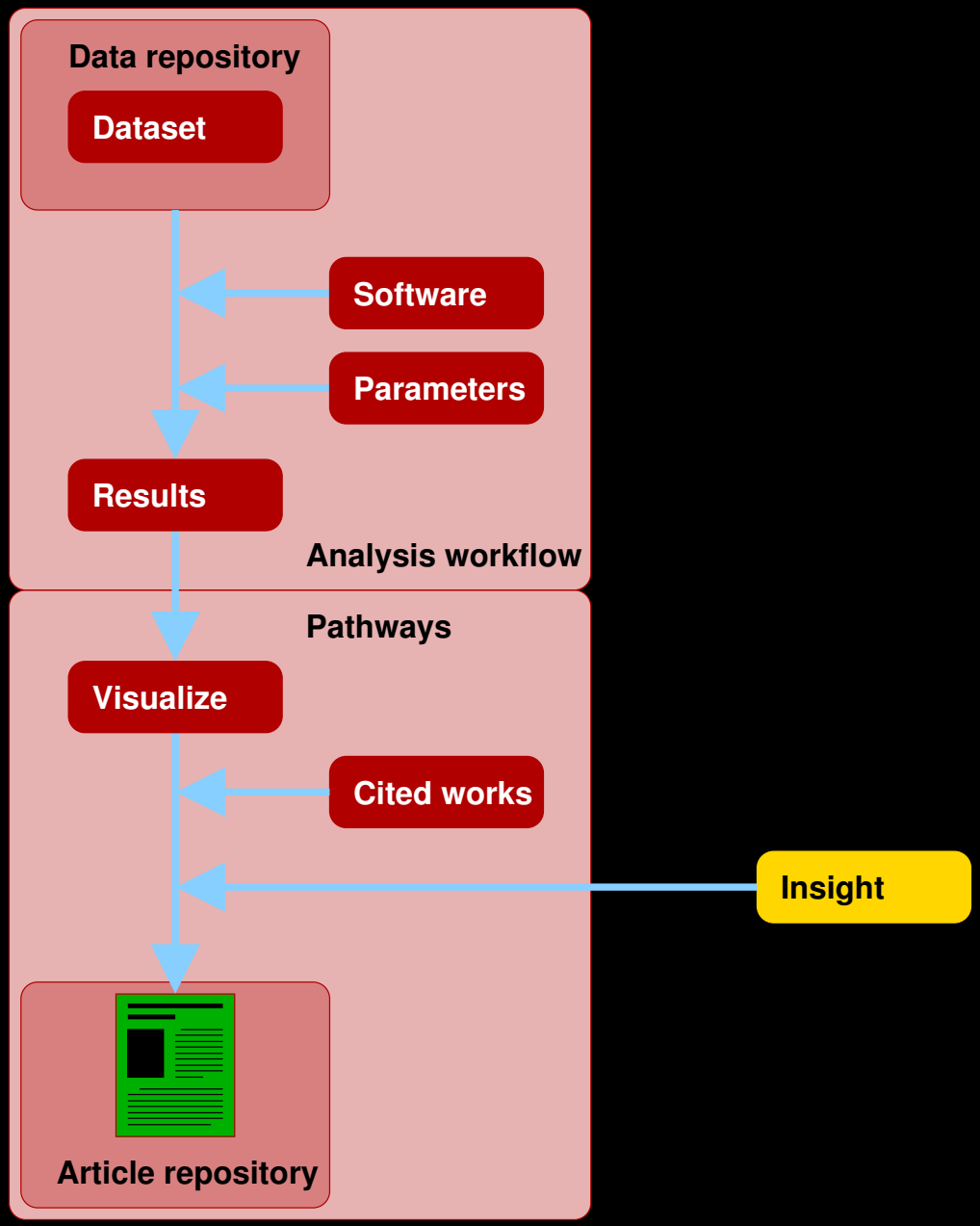
Lost information



Lost information.



Lost information..



Heterogeneity

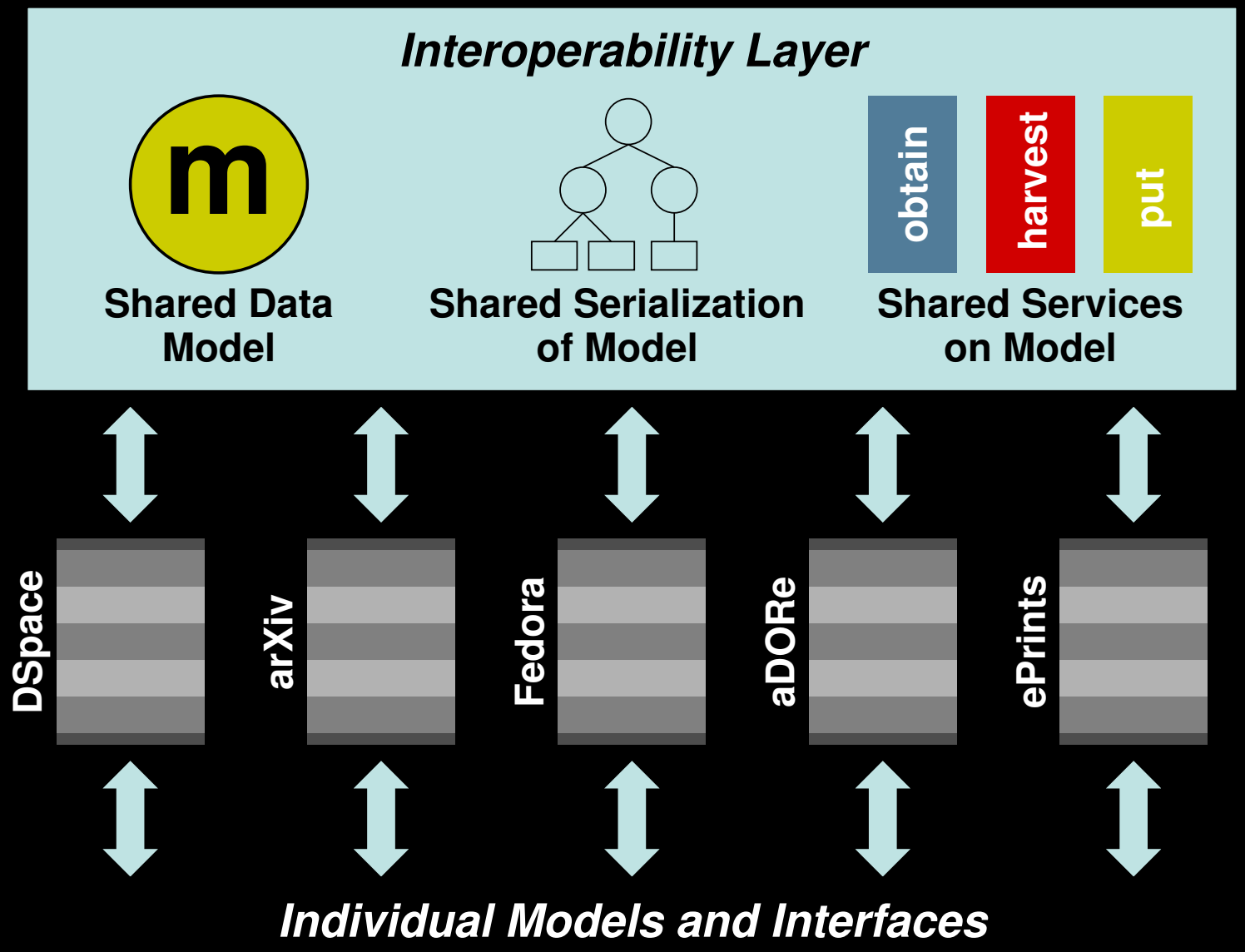
"variety is the spice of life"

Many repository types and architectures — Fedora, aDORe, DSpace, ePrints, arXiv, CDSware, Archimède, PubMed Central, data repositories (all different?) — this will not, and probably should not, change!

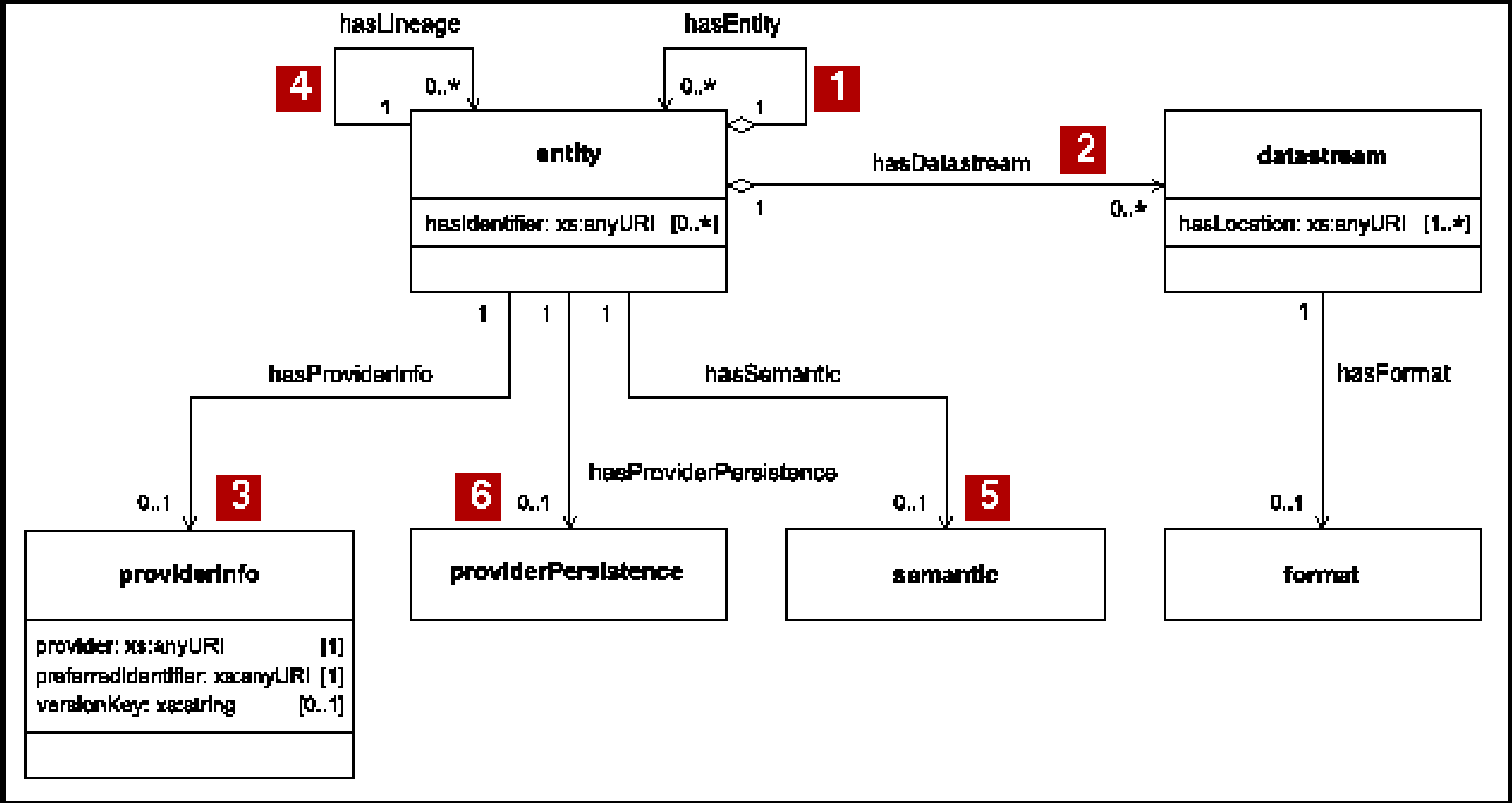
Q. Can we find a meaningful mapping between enough elements of the data models of these different systems to *overlay* services on top?

Q. Can we make these services sufficiently simple so as to be widely adopted?

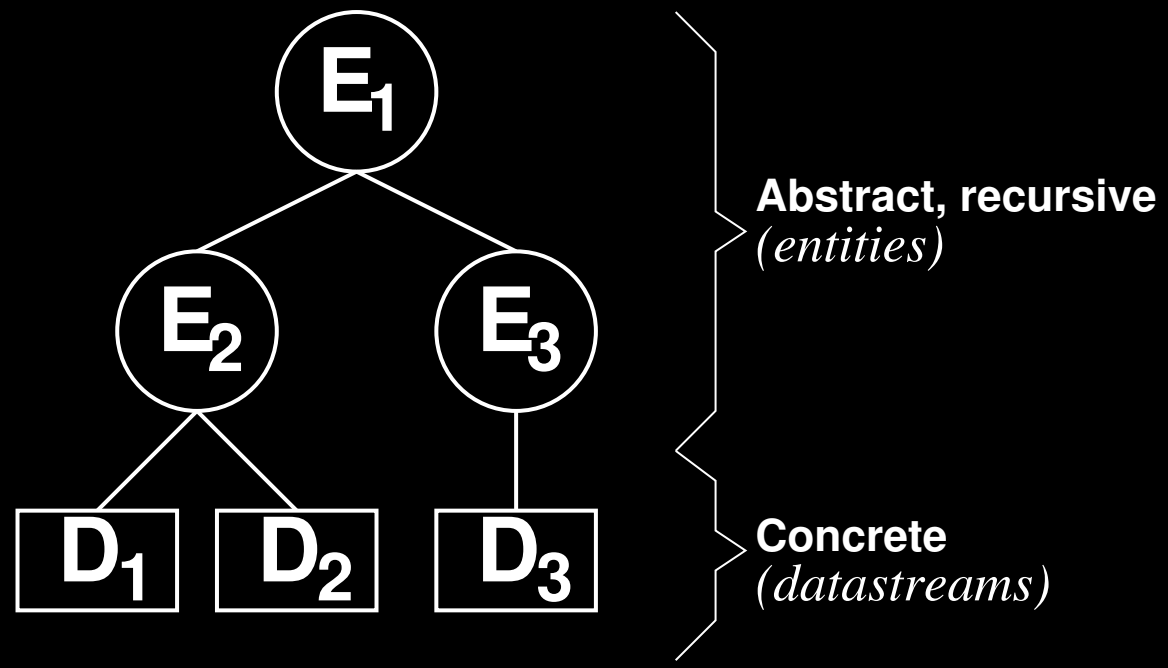
The Pathways interoperability fabric



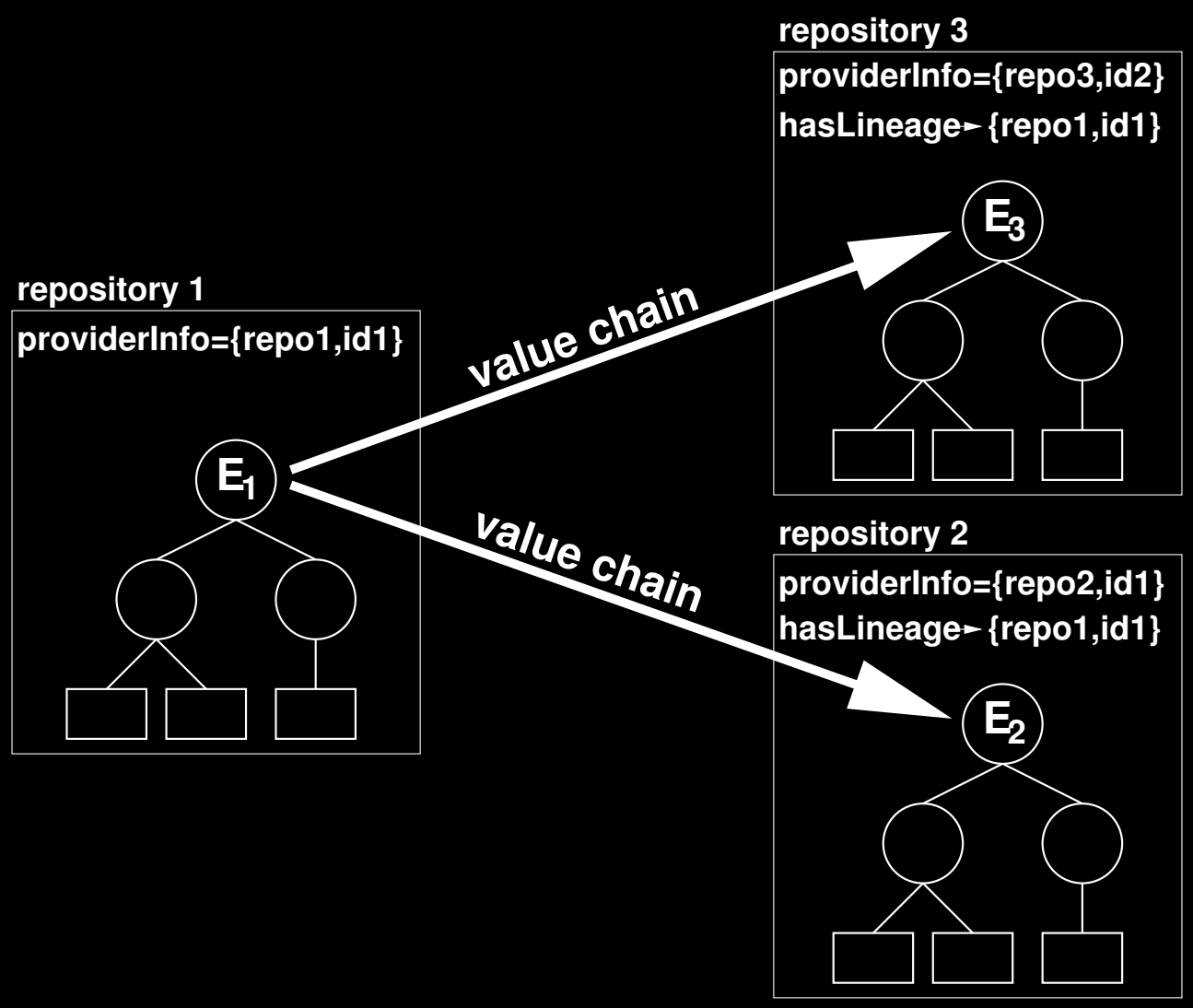
Data model


hasLineage
hasEntity
4
1
0..*
0..*
1
1
entity
hasIdentifier: xs:anyURI [0..*]
hasDatastream
2
1
0..*
datastream
hasLocation: xs:anyURI [1..*]
1
1
1
1
hasProviderInfo
hasSemantic
hasFormat
hasProviderPersistence
0..1
3
6
0..1
0..1
5
0..1
providerInfo
provider: xs:anyURI [1]
preferredIdentifier: xs:anyURI [1]
versionKey: xs:string [0..1]
providerPersistence
semantic
format

Entities and datastreams



Lineage



Broad notion of lineage, expect to sub-type.

Repository centric identification

There will continue to be many identifier schemes.

In Pathways, the identifier is a triple:

1. provider — identity of repository, key to look up service interfaces in registry.
2. preferredIdentifier — identity of entity in repository, key to request services. Syntax and semantics may be local to repository.
3. version [optional] — key to parameterize service requests according to local version semantics.

Serialization — a surrogate


```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:core="info:pathways/core#" xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <core:entity rdf:about="info:pathways/entity/info%3Asid%2Flibrary.lanl.gov..."
    <core:hasSemantic rdf:resource="info:pathways/semantic/journal-article"/>
    <core:hasIdentifier>info:doi/10.1016/j.dyepig.2004.12.010</core:hasIdentifier>
    <core:hasProviderPersistence rdf:resource="info:pathways/persistence/persistent"/>
    <core:hasProviderInfo>
      <core:providerInfo>
        <core:preferredIdentifier>info:doi/10.1016/j.dyepig.2004.12.010</core:preferredIdentifier>
        <core:provider>info:sid/library.lanl.gov:pathways</core:provider>
      </core:providerInfo>
    </core:hasProviderInfo>
    <core:hasEntity>
      <core:entityrdf:about="info:pathways/entity/info..."
        <core:hasSemantic rdf:resource="info:pathways/semantic/bibliographic-citation"/>
        <core:hasIdentifier>info:lanl-repo/ssm/doi-10.1016/j.dyepig.2004.12.010</core:hasIdentifier>
        <core:hasProviderPersistence rdf:resource="info:pathways/persistence/persistent"/>
        <core:hasProviderInfo>
          <core:providerInfo>
            <core:preferredIdentifier>info:lanl-repo/ssm/doi-10.1016/j...</core:preferredIdentifier>
            <core:provider>info:sid/library.lanl.gov:pathways</core:provider>
          </core:providerInfo>
        </core:hasProviderInfo>
```

Services

obtain

Request a surrogate for a single digital object (cf. GetRecord in OAI-PMH; OpenURL more general).

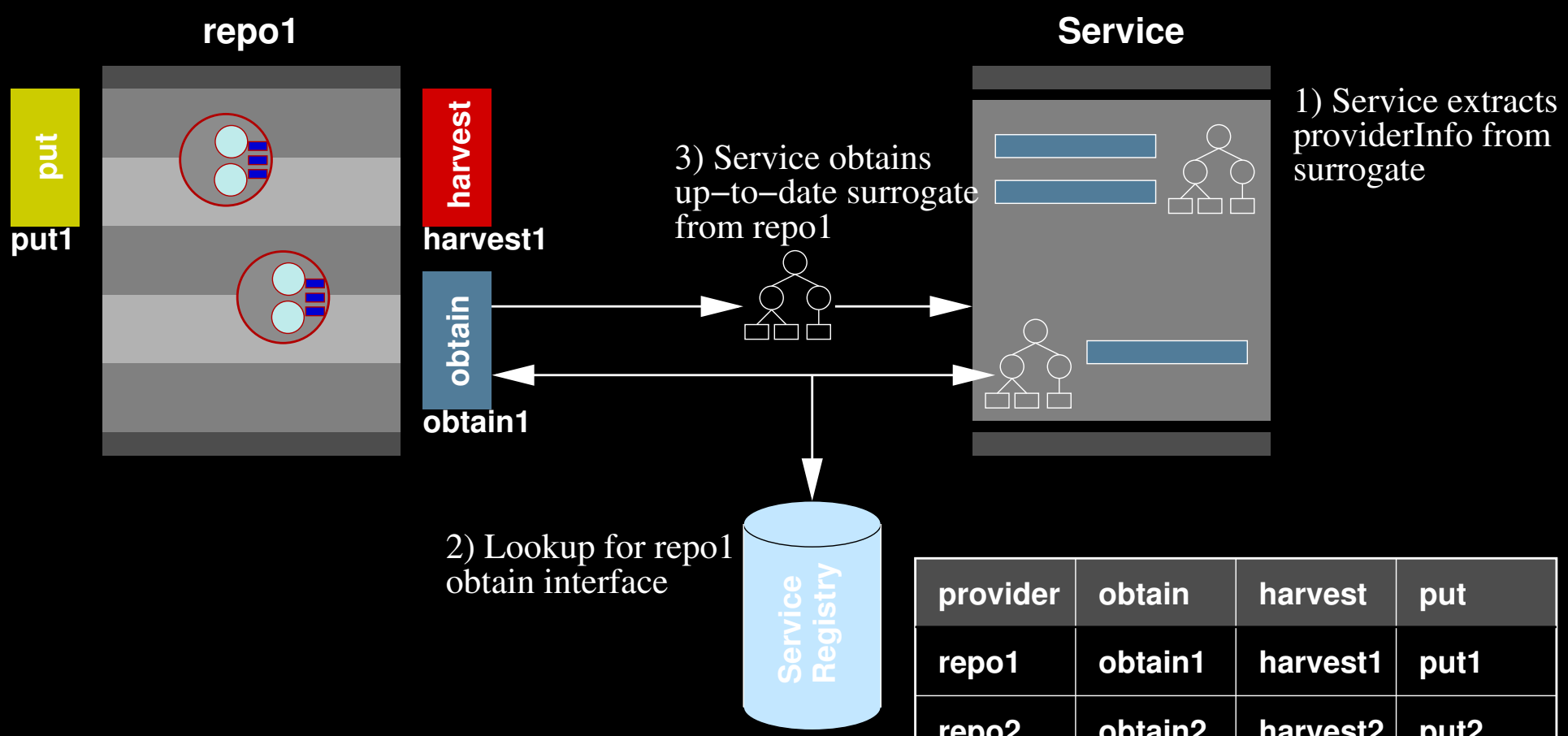
harvest

Collect batches of surrogates for several digital objects. (cf. ListIdentifiers in OAI-PMH).

put

Request ingest of one or more surrogates into a digital repository — *"request for deposit"*

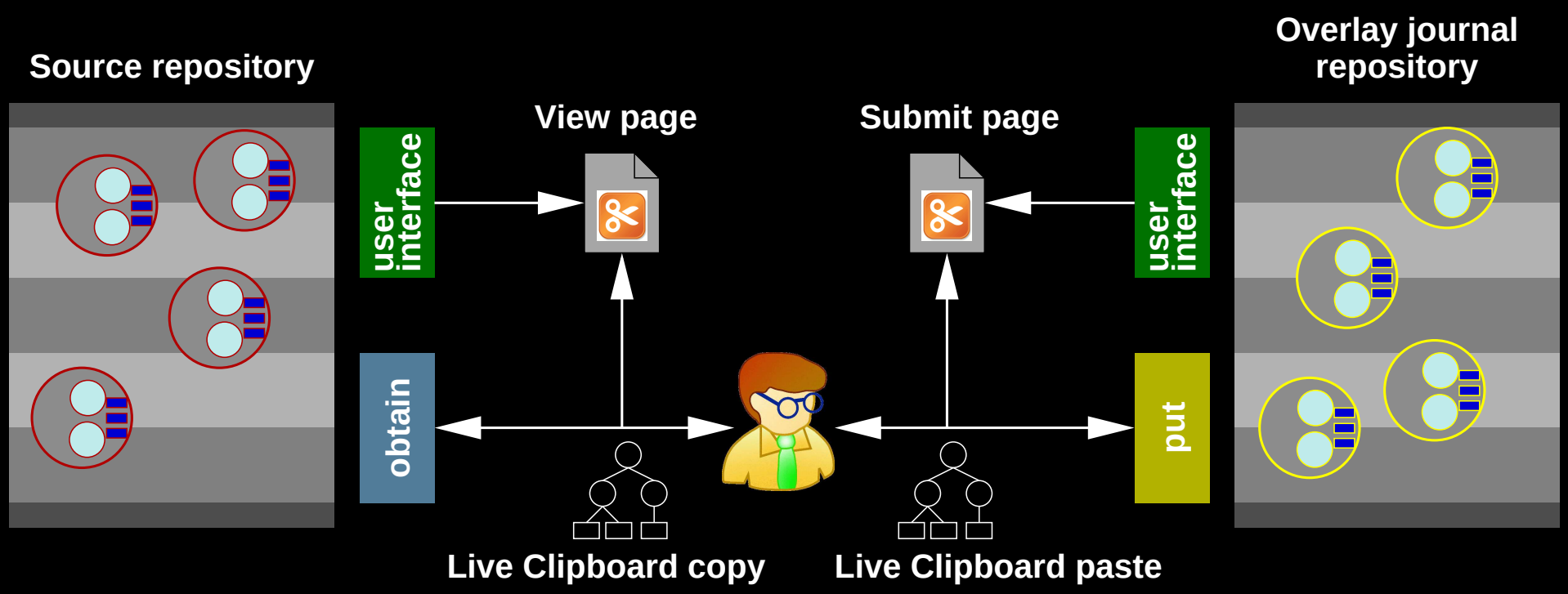
Service registry



provider	obtain	harvest	put
repo1	obtain1	harvest1	put1
repo2	obtain2	harvest2	put2

Also anticipate format and semantic registries.

An overlay journal demonstration



Summary

Scholarly communication in increasingly fluid, collaborative, network-based and data-intensive. The scholarly communication system must:

- be innately digital and interlinked
- support an expanded "unit of communication" that may be heterogeneous and distributed
- provide for many different pathways that fulfill the necessary communication functions

In this work we have demonstrated a relatively simple approach that allows construction of scholarly value chains across heterogeneous repositories.

Further reading

Pathways: Augmenting interoperability across scholarly repositories. Simeon Warner, Jeroen Bekaert, Carl Lagoze, Xiaoming Liu, Sandy Payette, Herbert Van de Sompel. *IJDL Special Issue on Digital Libraries and eScience*. arXiv:cs/0610031

An Interoperable Fabric for Scholarly Value Chains. Herbert Van de Sompel, Carl Lagoze, Jeroen Bekaert, Xiaoming Liu, Sandy Payette, Simeon Warner. *D-Lib Magazine*, 12(10), 2006. doi:10.1045/october2006-vandesompel

News: OAI-ORE — Object Re-use and Exchange — announced last Friday with Mellon funding, led by Carl Lagoze and Herbert Van de Sompel. http://www.openarchives.org/ore

That's all folks...