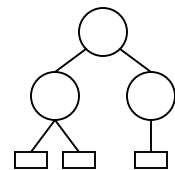


Context, motivating examples, introduction to other presentations



Obtain

Harvest

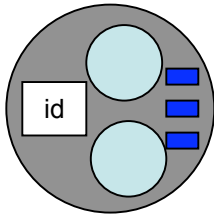
Put

Herbert Van de Sompel
Research Library
Los Alamos National Laboratory, USA

This work was supported by NSF award number IIS-0430906 (Pathways)



Terminology



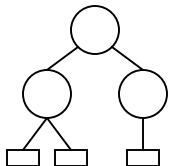
Digital Object: A data structure whose principal components are digital data and key-metadata. Digital data can be a *Datastream* or a *Digital Object*, i.e. a *Digital Object* may have one or more other *Digital Objects* as nested components. Key-metadata must include an identifier for the *Digital Object*.



Datastream: An ordered sequence of bytes.



Data Model: An abstraction for *Digital Objects* such that each *Digital Object* can be seen as an instance of the class defined by a *Data Model*. Example *Data Models* include the Pathways Core model, the MPEG-21 Digital Item Declaration model, etc.



Surrogate: A serialization of a *Digital Object* according to a *Data Model*.



Terminology



Repository: a networked system that provides services pertaining to a collection of *Digital Objects*.



Obtain interface: a *Repository* interface that supports the request of services pertaining to individual *Digital Objects* (including their component *Datastreams*).



Harvest interface: a *Repository* interface that exposes *Surrogates* for incremental collecting/harvesting.



Put interface: a *Repository* interface that supports submission of one or more *Surrogates* into the *Repository*, thereby facilitating the addition of *Digital Objects* to the collection of the *Repository*.



Augmenting interoperability across *Repositories*

- Motivations: Enable value chains that start in *Repositories*
 - Facilitate emergence of richer cross-*Repository* services
 - Facilitate scholarly communication workflow across *Repositories*
- Through our presentations, we will:
 - Share a vision
 - Introduce some concrete straw man ideas
 - Show some demonstrations of ideas
 - Leave you with a lot of food for thought, discussion and criticism

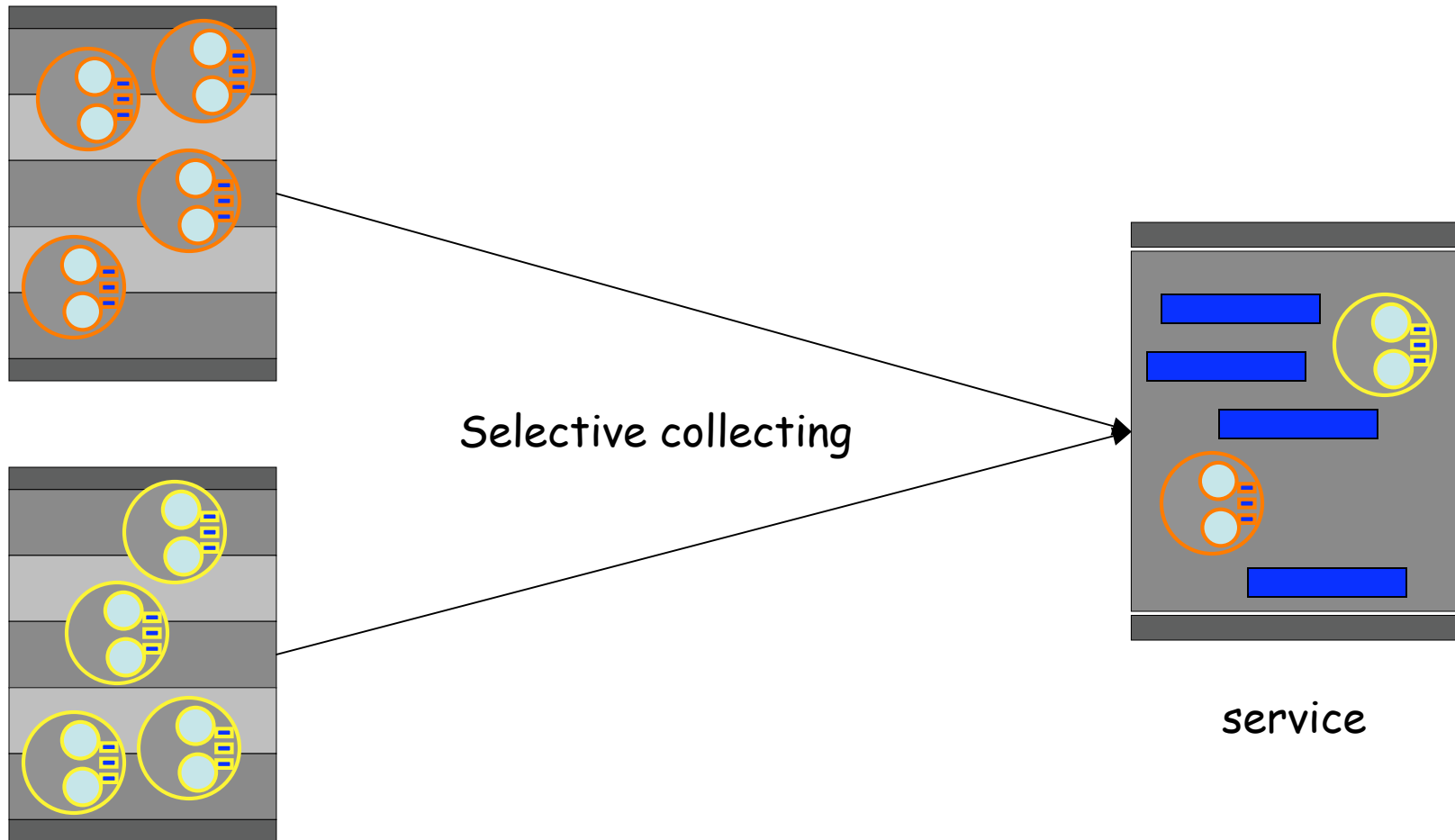


Richer cross-*Repository* services

- Distributed *Repositories* provide source materials for cross-*Repository* overlay services such as discovery services
- Manner in which those materials are exposed must allow for the seamless emergence of rich and meaningful services



Richer cross-Repository services



Richer cross-Repository services : Scenario

Scenario 1: Chemical search engine

- A search engine monitors scholarly repositories but is only interested in making machine-readable chemical structures contained in *Digital Objects* available from those repositories searchable.
- This constitutes re-use of the (part of) the *Digital Objects* by a service overlaid upon the monitored repositories.
- And, of course, a chemical compound discovered via the search engine can be cited in some new paper, i.e. the value chain does not stop here

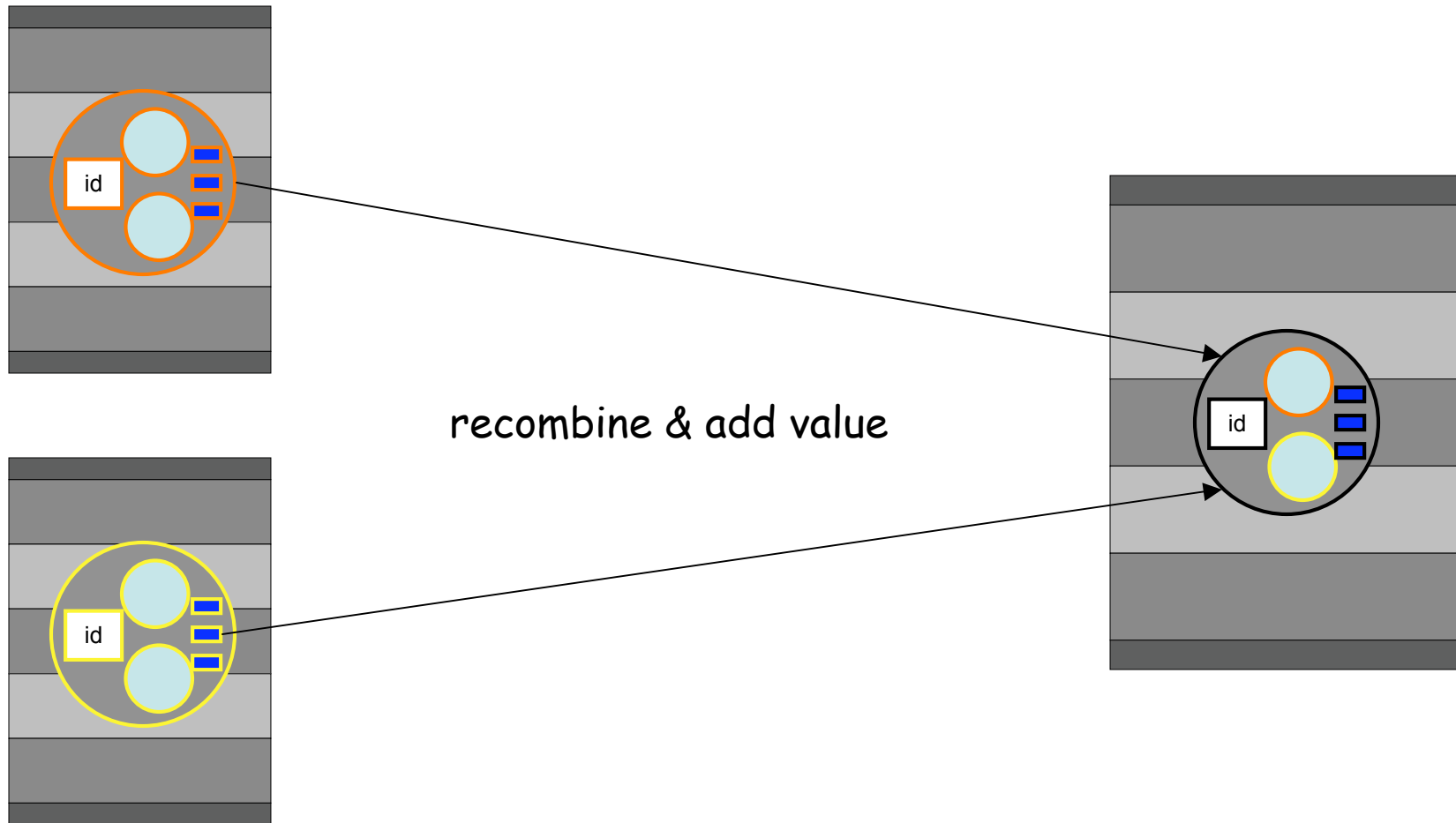


Scholarly communication workflow across Repositories

- Distributed *Repositories* at the basis of a digital scholarly communication system
- Scholarly communication as a global workflow (value chain) across those *Repositories*
- *Digital Objects* from *Repositories* are the subject of the workflow; they are used and re-used in many contexts.



Scholarly communication workflow across Repositories



Scholarly communication workflow : Scenarios

Scenario 2: Citation

- An author writes a paper (to be *Put* into her institutional repository) and cites 10 papers available from other repositories.
- A citation to a paper is a type of re-use of the cited paper in a new context.
- And, of course, the new paper can be cited too, i.e. the value chain does not stop here.



Scholarly communication workflow : Scenarios

Scenario 3: Overlay journal

- The editor of an overlay journal selects papers from 3 different repositories for inclusion in the next issue of the overlay journal.
- Each of those articles is being re-used in a new context, with value being added.
- And, the overlay journal can be mirrored for preservation purposes, i.e. the value chain does not stop here.



Scholarly communication workflow : Scenarios

Scenario 4: eScience

- A researcher uses datasets from 2 different dataset repositories, performs operations on those, and creates a publication that contains a resulting new dataset and an accompanying paper, and deposits this publication in her institutional repository.
- This constitutes re-use of the origin datasets, and value added through the creation of the new publication.
- And, of course, the new dataset can be re-used too, i.e. the value chain does not stop here.



Value chains starting in *Repositories* : Considerations

Currently:

- The Scenarios can only be realized in idiosyncratic ways.
- Our existing *bibliographic* infrastructure (typical work-level citation) is:
 - not natively machine readable/actionable
 - not at the level of granularity required to allow establishing an appropriate "chain of evidence"



Value chains starting in *Repositories* : Considerations

Lacking are:

- *Repository* interfaces: interoperable *Repository* interfaces to support these kind of workflows.
- ProviderInfo ~ machine actionable citation ~ chain of evidence:
 - means to express origin *Repository* of a *Digital Object*
 - means to express identity (identifier, version) of a *Digital Object* in its origin *Repository*
 - means to express granularity of re-use of a *Digital Object*
 - this is the parallel of a bibliographic citation at the work level (paper world)



Value chains starting in *Repositories* : Considerations

Lacking are:

- **Lineage:** means to unambiguously express the workflow-relationship between a new *Digital Object* and the one(s) it builds on. Lineage is ProviderInfo in a previous life.
- **Persistence:** a level of commitment regarding persistence of *Digital Objects* in the scholarly communication system, and a means to express that level of commitment.
- **Semantics:** means to express the type of content available from *Repositories*.



Value chains starting in *Repositories* : Considerations

Scholarly communication is not shallow online chatting:

- need **long-term perspective** when devising approaches
- for example:
 - need **abstract definitions** of *Repository* interfaces that can be instantiated on the basis of various technologies as time goes by
 - interfaces need to work with **whichever type of identifier** (current and future) because *Repositories* will use whichever type of identifier

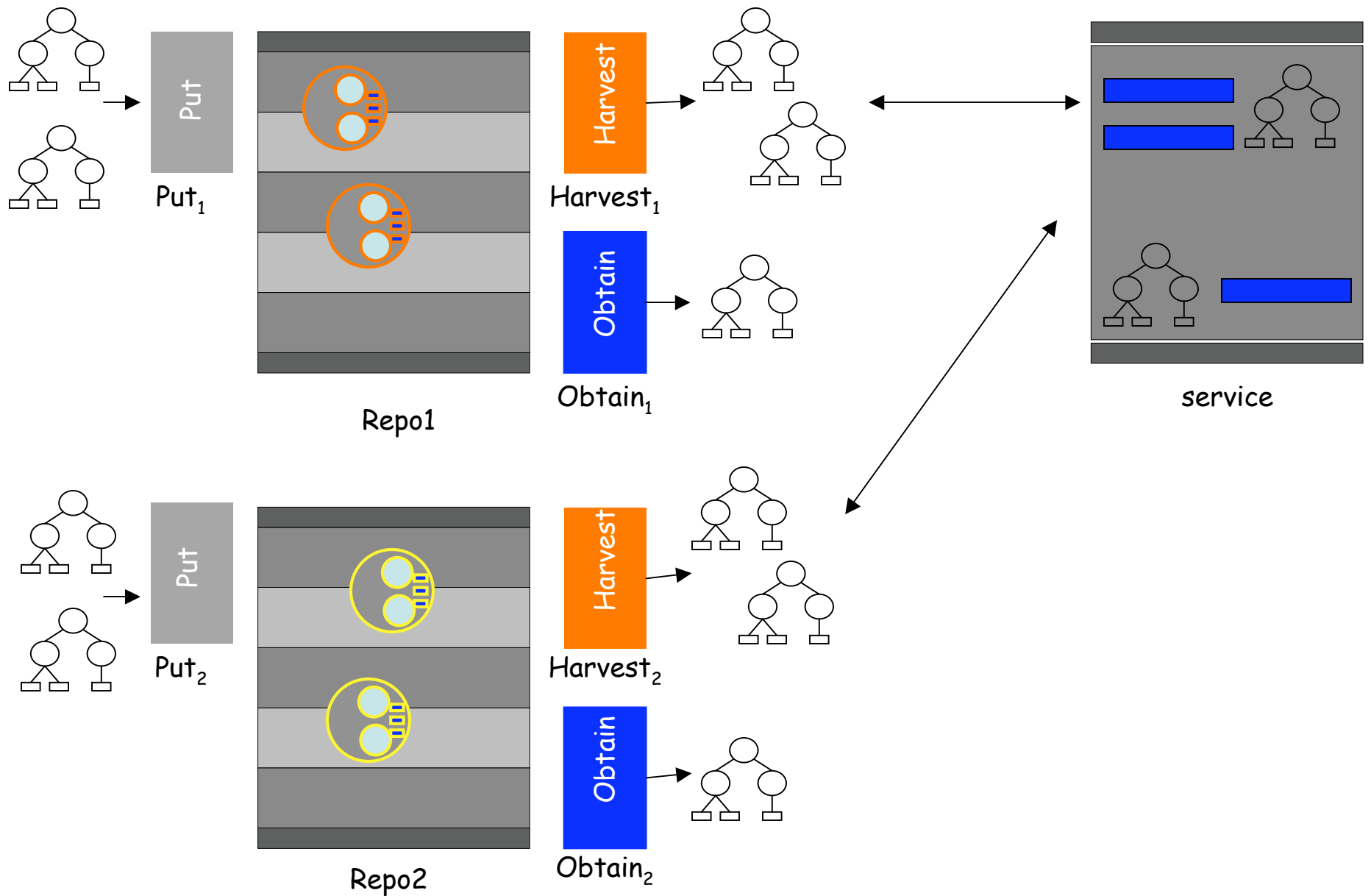


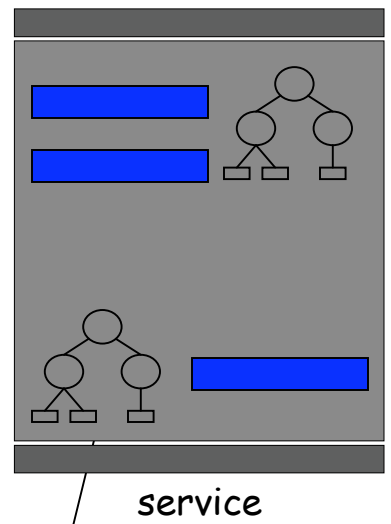
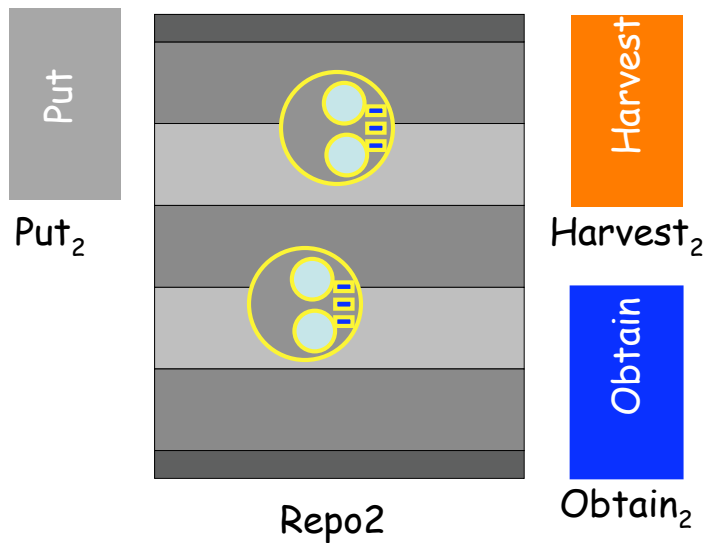
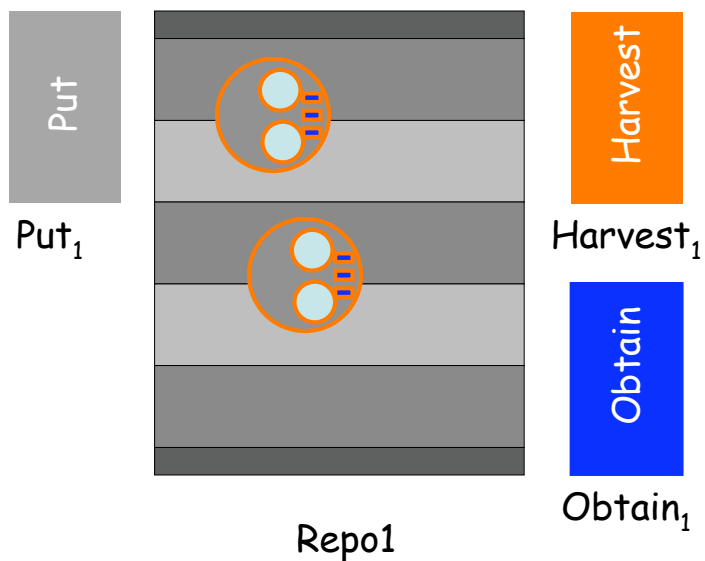
Value chains starting in *Repositories*: kick-start

This combination might kick-start the emergence of value chains:

- An appropriate *Data Model* supported across *Repositories*.
- 3 core *Repository* interfaces supported across *Repositories*: *Obtain, Harvest, Put*.
- A *Surrogate* format (compliant with the *Data Model*) supported across the *Repository* interfaces.
- Some shared infrastructure, i.e. a *Service Registry*, *Semantic Ontologies*, *Format Registries*, etc.
- These will be discussed in the other presentations.

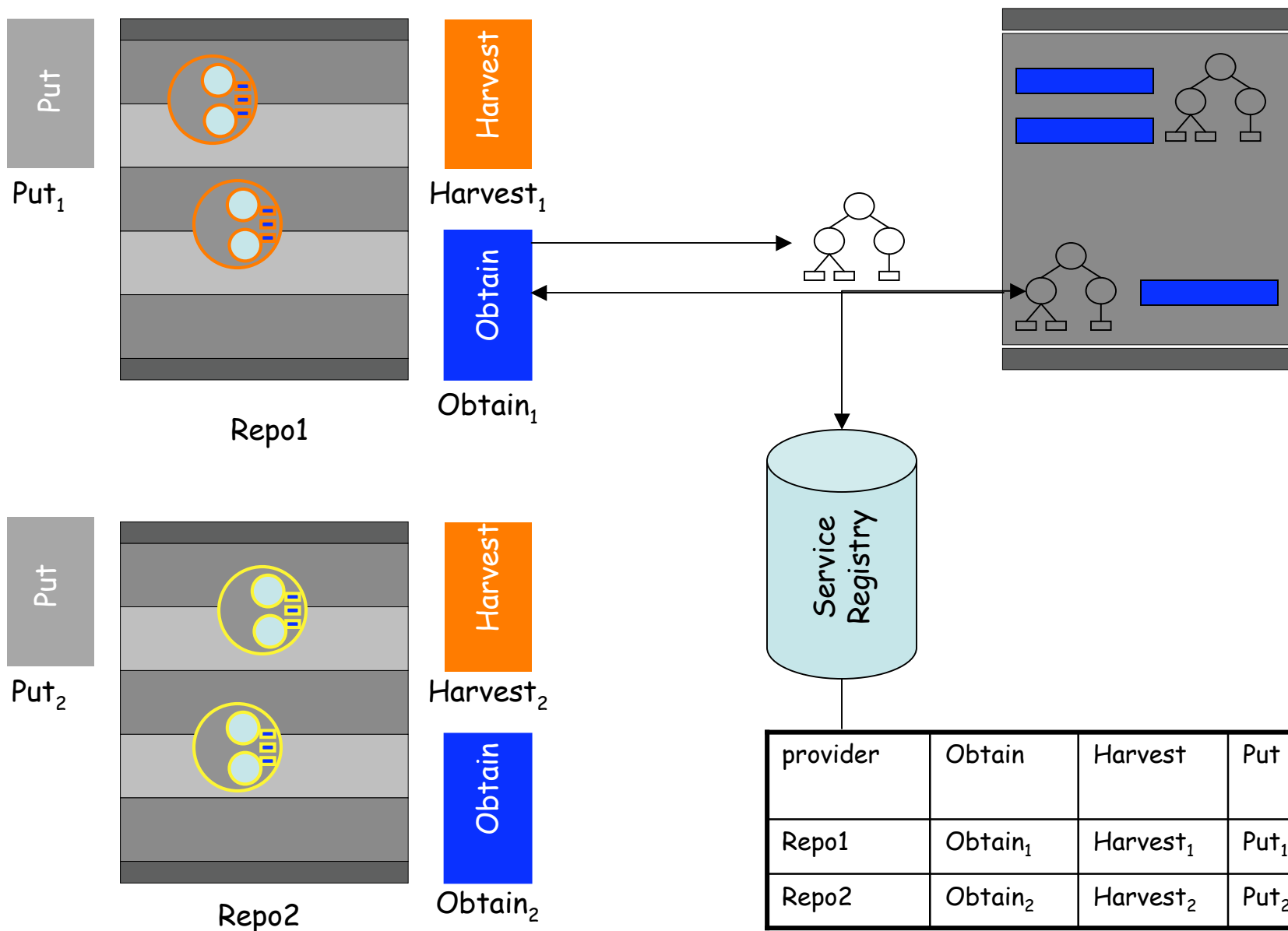






Via providerInfo a *Surrogate* contains information on how to *Obtain* another *Surrogate* at some later point in time.





Demonstration

- Overlay journal Scenario combined with Search engine Scenario
- *Surrogates* compliant with Pathways Core Data Model, expressed in RDF/XML.
- *Obtain* interfaces at:
 - an aDORe repository
 - arXiv
 - a DSpace repository
 - a Fedora repository
- *Harvest* interfaces at:
 - an aDORe repository
 - arXiv
 - a Fedora repository
- *Put* interface at a Fedora repository
- Live Clipboard functionality in user interfaces of arXiv, Fedora, overlay search engine



Demonstration

- Acknowledgments:
 - Carl Lagoze, Sandy Payette, Simeon Warner, Chris Wilper at Cornell University
 - Rob Tansley at HP
 - Luda Balakireva, Xiaoming Liu, Herbert Van de Sompel, Zhiwu Xie at the Los Alamos National Laboratory



Demonstration

