

# Transitions & Thresholds

Data Transfer & Bridging Infrastructure in Data Curation

Ingrid Mason, Deployment Strategist  
Dr Frankie Stevens, Research Engagement Strategist



# What happens “in between”...

the stages of this [curation] lifecycle model, the where and the how, of data packaging and movement, that supports



# Propositions

- By bringing forward the transitions, the “in between” stages (or thresholds) of the DCC lifecycle model we see critical points in data curation as part of research workflows, and the complex relationships around and technologies that support those activities
- By inspecting data transfer and bridging infrastructure we begin to understand how that **binding matter\***, as with the high-speed networks, reflect the processes, partnerships and communities, in the Australian research landscape.

\*Advanced research networks and their layered services now including cloud storage have sometimes been referred to as: *the dark matter binding the research universe*

# By shifting the lens this way...

The less evident data transfer and bridging infrastructure layer (implicit in the [curation] lifecycle model) that enables data use and curation, is highlighted.

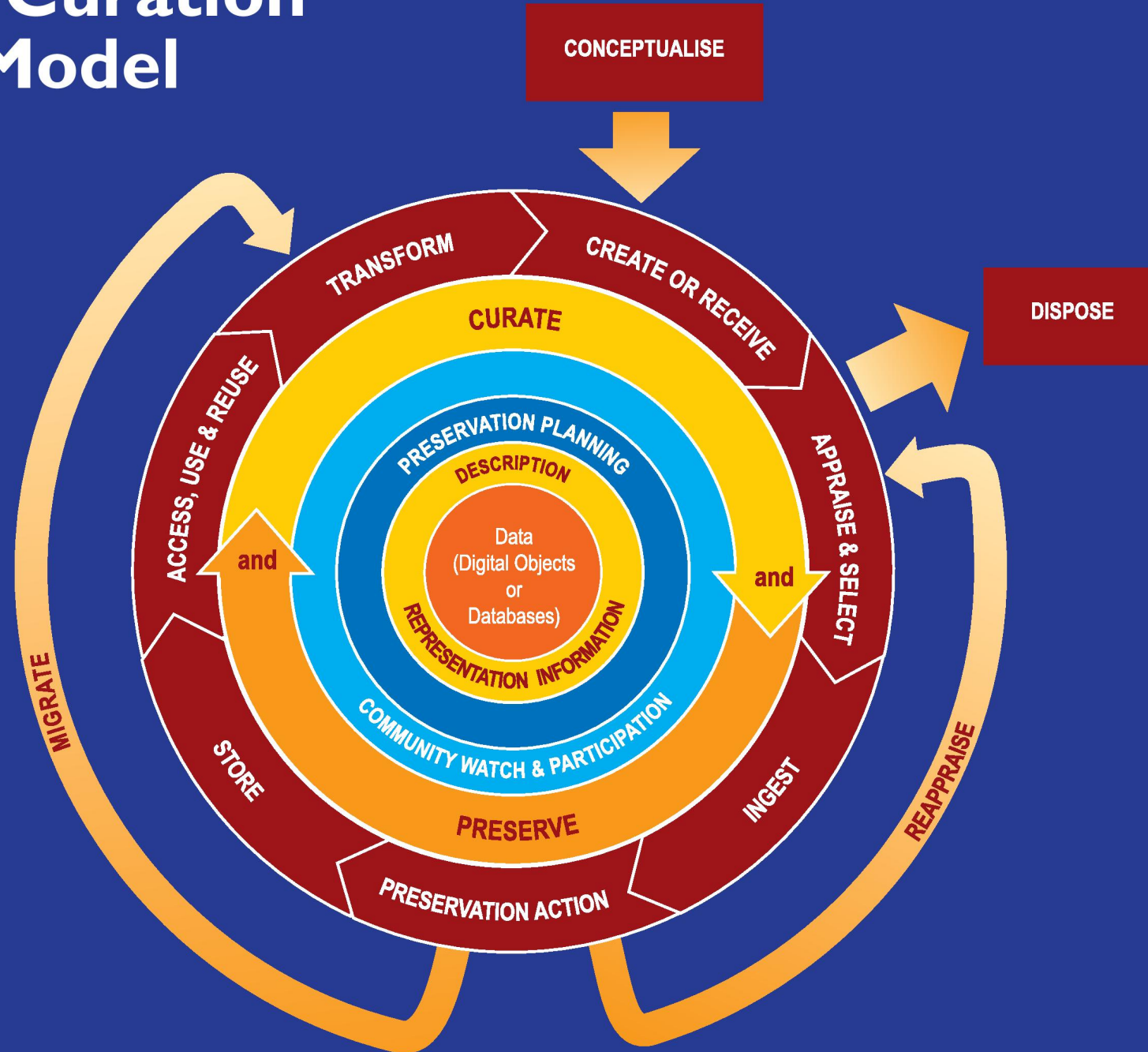
# Assertions

- Many researchers will be moving their data in/out research infrastructures and into a “third space” and “binding matter” and bridges enable this
- Curation support for the bulk of data and technology intensive research is undertaken in ad hoc, semi-manual processes rather than enabled with guidance and systematic technology enabled workflows
- A panoptic view of multiple disciplinary research workflows (between institutional boundaries) is needed to ascertain where existing research support and infrastructure\* can be broadly applied and reused, and where new research support and infrastructure is needed

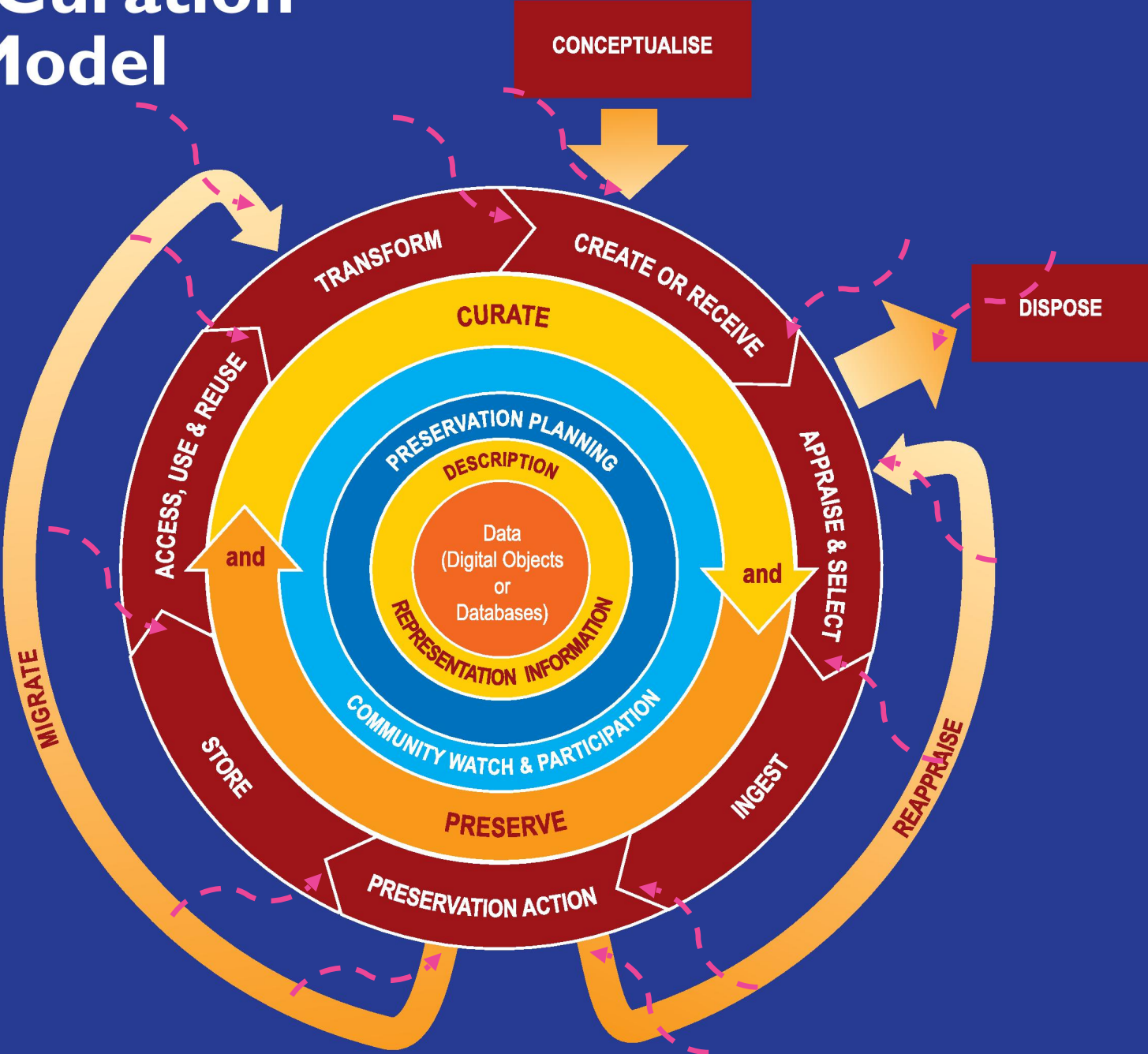
\*standards, procedures, policies, processes, guidelines, people, applications, systems, services etc



# The DCC Curation Lifecycle Model

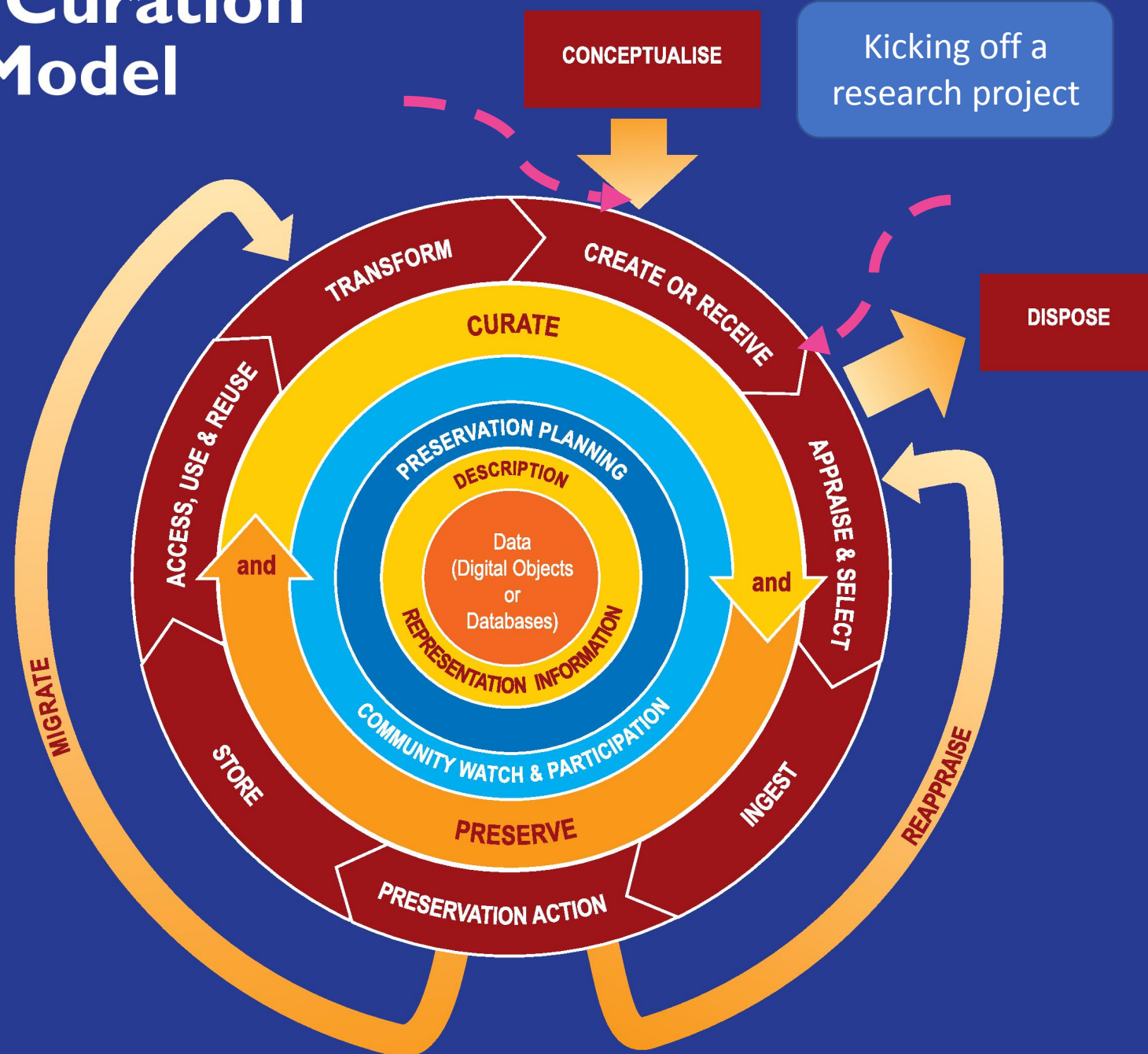


# The DCC Curation Lifecycle Model



# The DCC Curation Lifecycle Model

- 1. Packaging
- 2. Movement

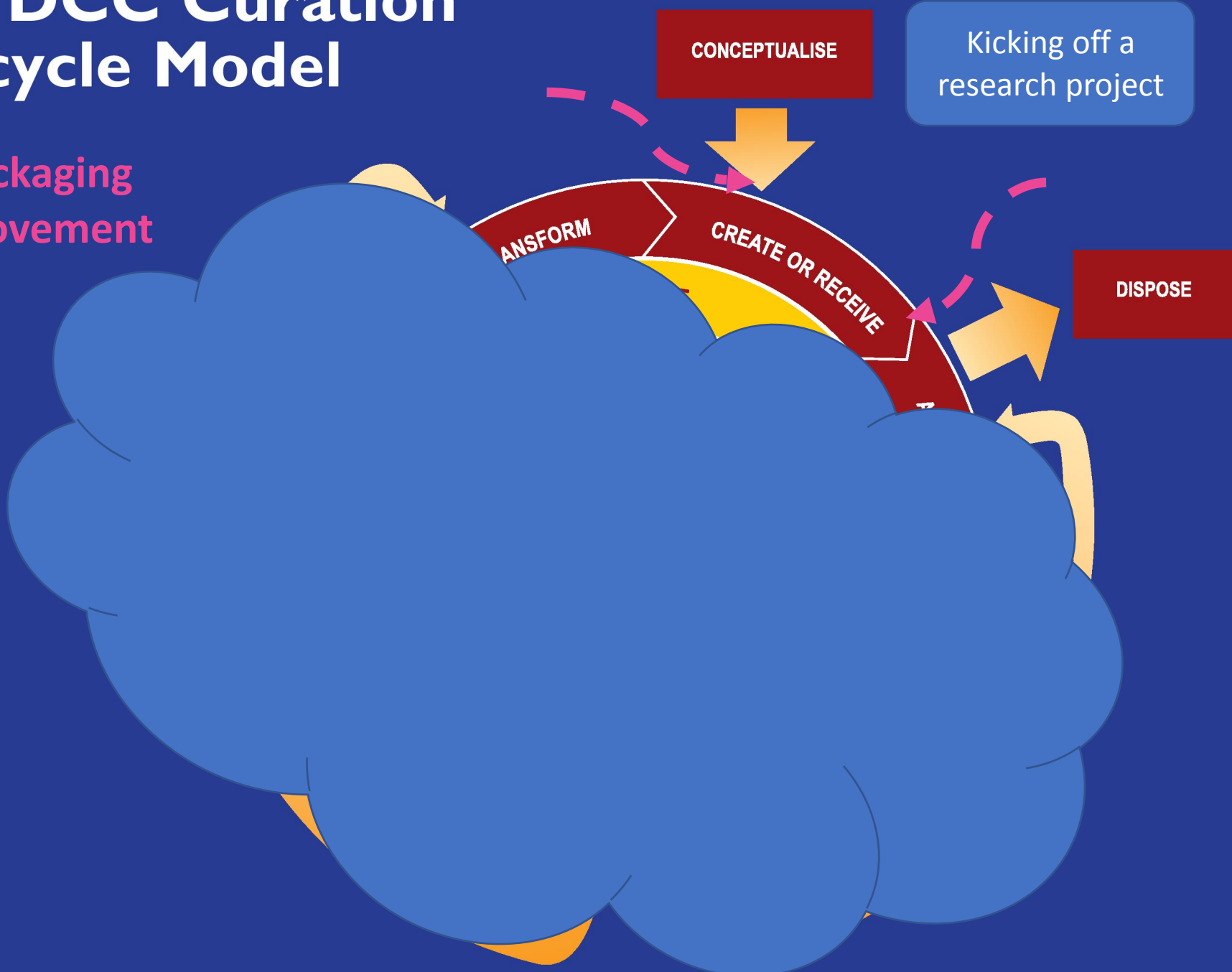


# We begin to understand...

how that binding matter, as with the high-speed networks, reflect the collaborations, partnerships and communities, in the Australian research and education landscape.

# The DCC Curation Lifecycle Model

- 1. Packaging
- 2. Movement



# Transitions & Relationships

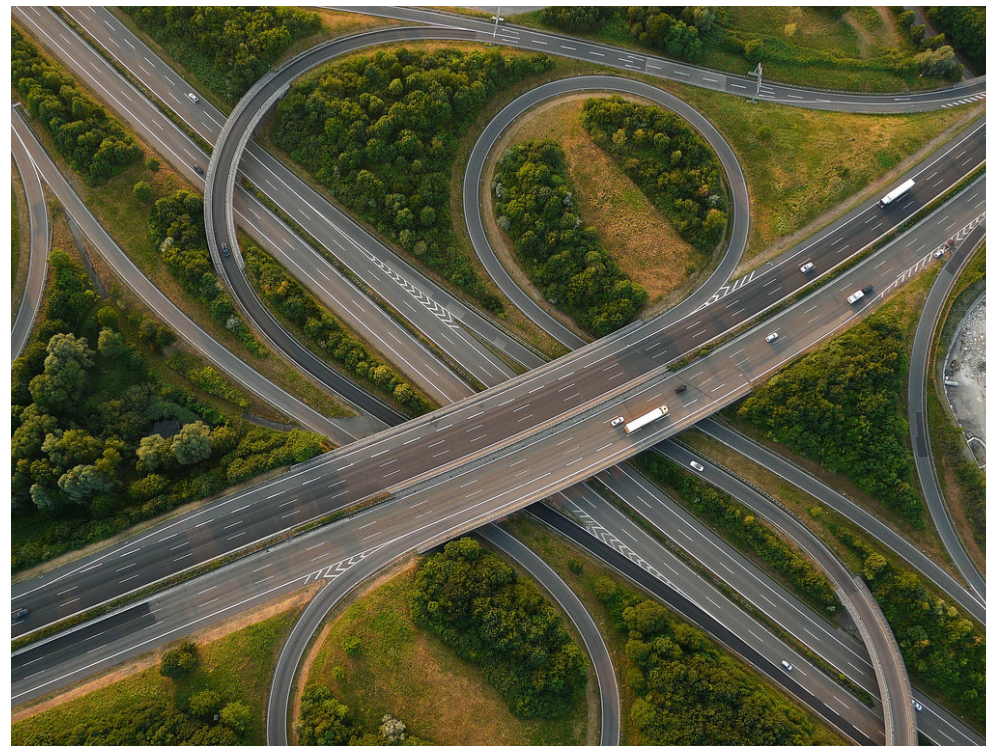
- Create

- Facility rep and researcher
- Intra/extra institutional
- Local/remote access

- Receive

- Data store rep and researcher
- Intra/extra institutional
- Local/remote access

# Data Packaging and Movement



<https://www.flickr.com/photos/neuwieser/4827571943/> CC-BY-SA 2.0

<https://www.flickr.com/photos/stevendepolo/5162503281/> CC-BY 2.0

# Analysis of lifecycle models

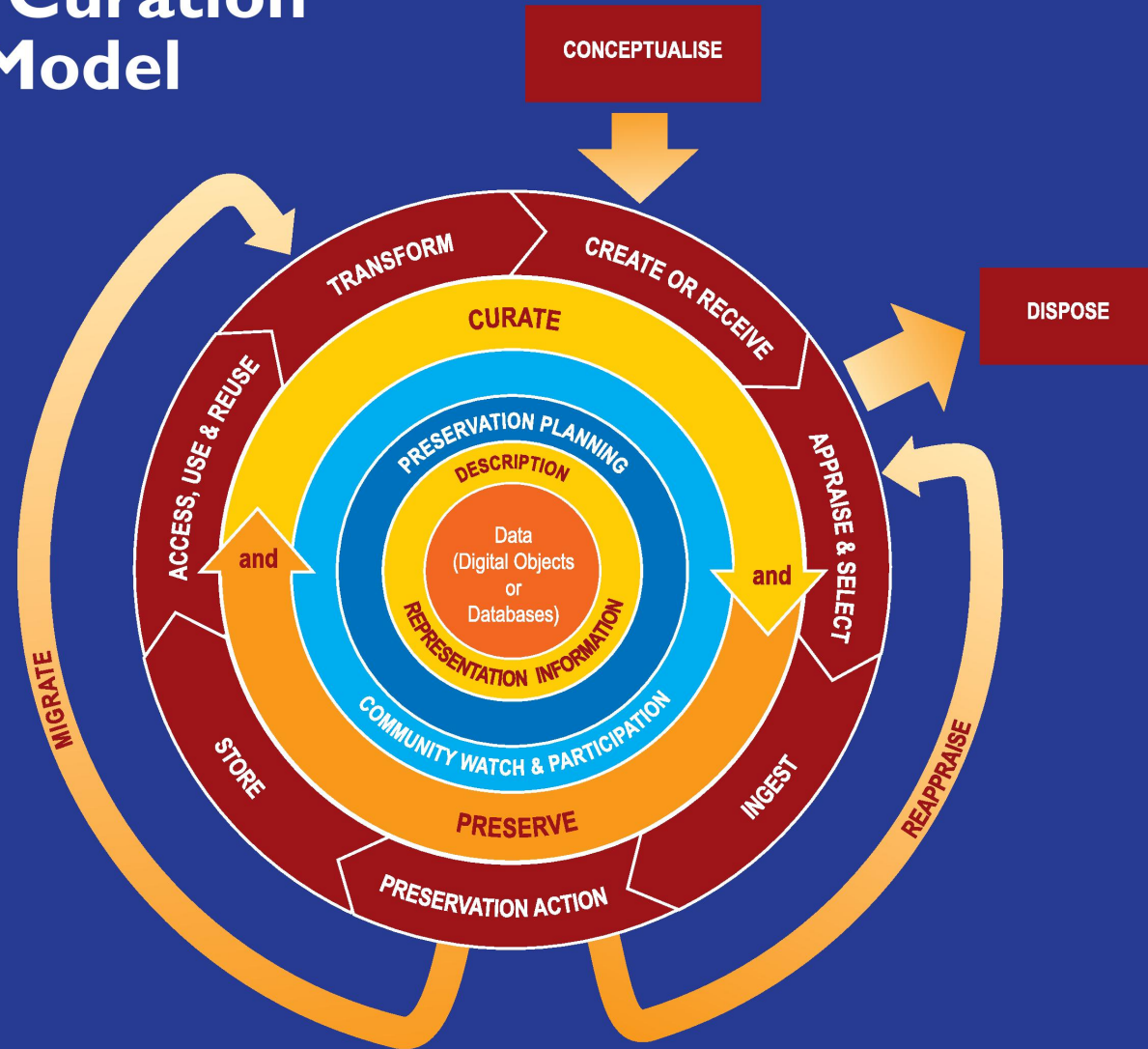
[they] encourage thinking that research processes are highly purposive, unidirectional, serial and occurring in a closed system. Research is often not like this,...

[they are] used to explain service offerings, the analysis shows that this may not always reflect researchers' own understanding of the research process. In failing to do so they can alienate potential users...

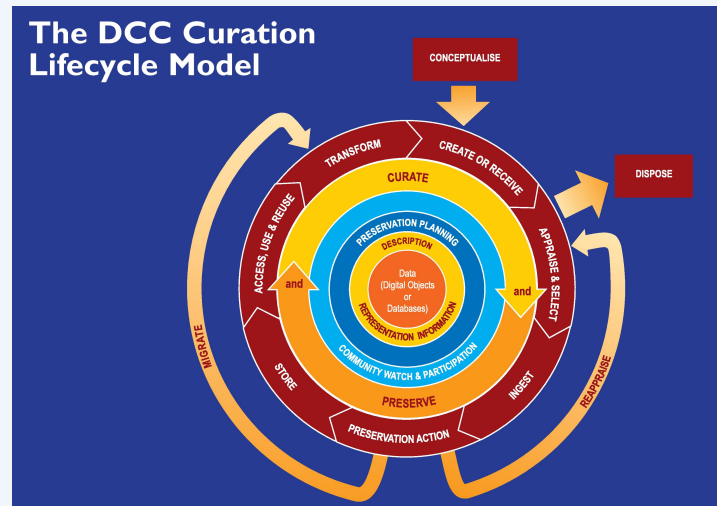
Andrew Martin Cox, Winnie Wan Ting Tam, (2018) "A critical analysis of lifecycle models of the research process and research data management", *Aslib Journal of Information Management*, Vol. 70 Issue: 2, pp. 142-157, <https://doi.org/10.1108/AJIM-11-2017-0251>

# The DCC Curation Lifecycle Model

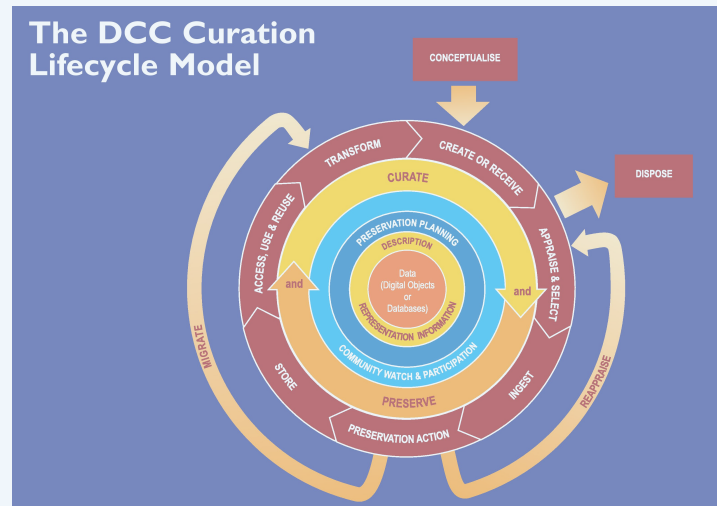
Research workflows



Research  
workflows



Research  
workflows



# Breaking it all down...

- Examine and collate common research workflows where they cross and connect personal and institutional boundaries in and out of shared cloud services
- Identify the wider network of stakeholders involved to understand the working relationships and technologies
- Identify the critical points in research workflows where data moves between specialised and underpinning research infrastructures (and the context in which they operate)

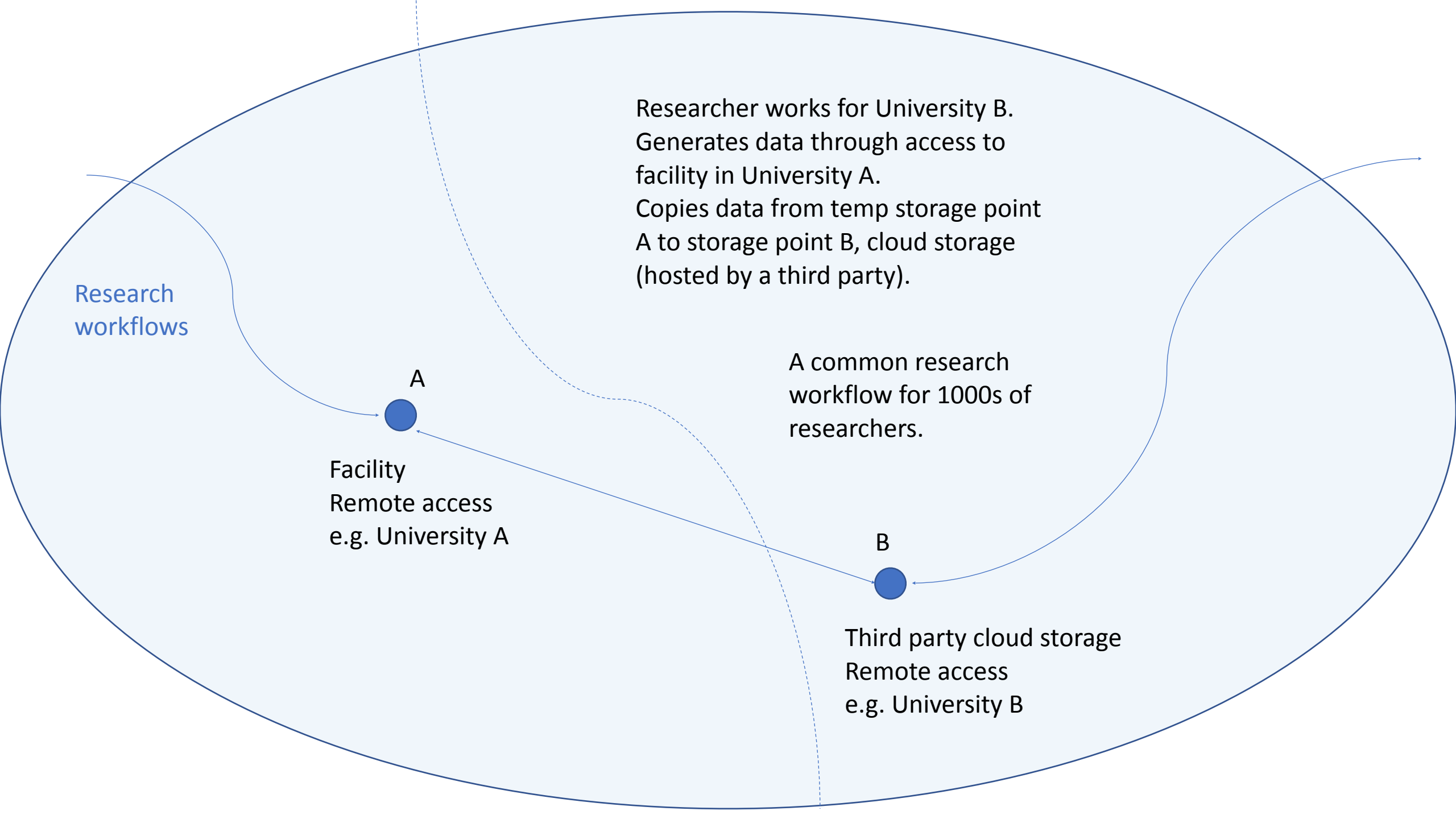
# Movement & change

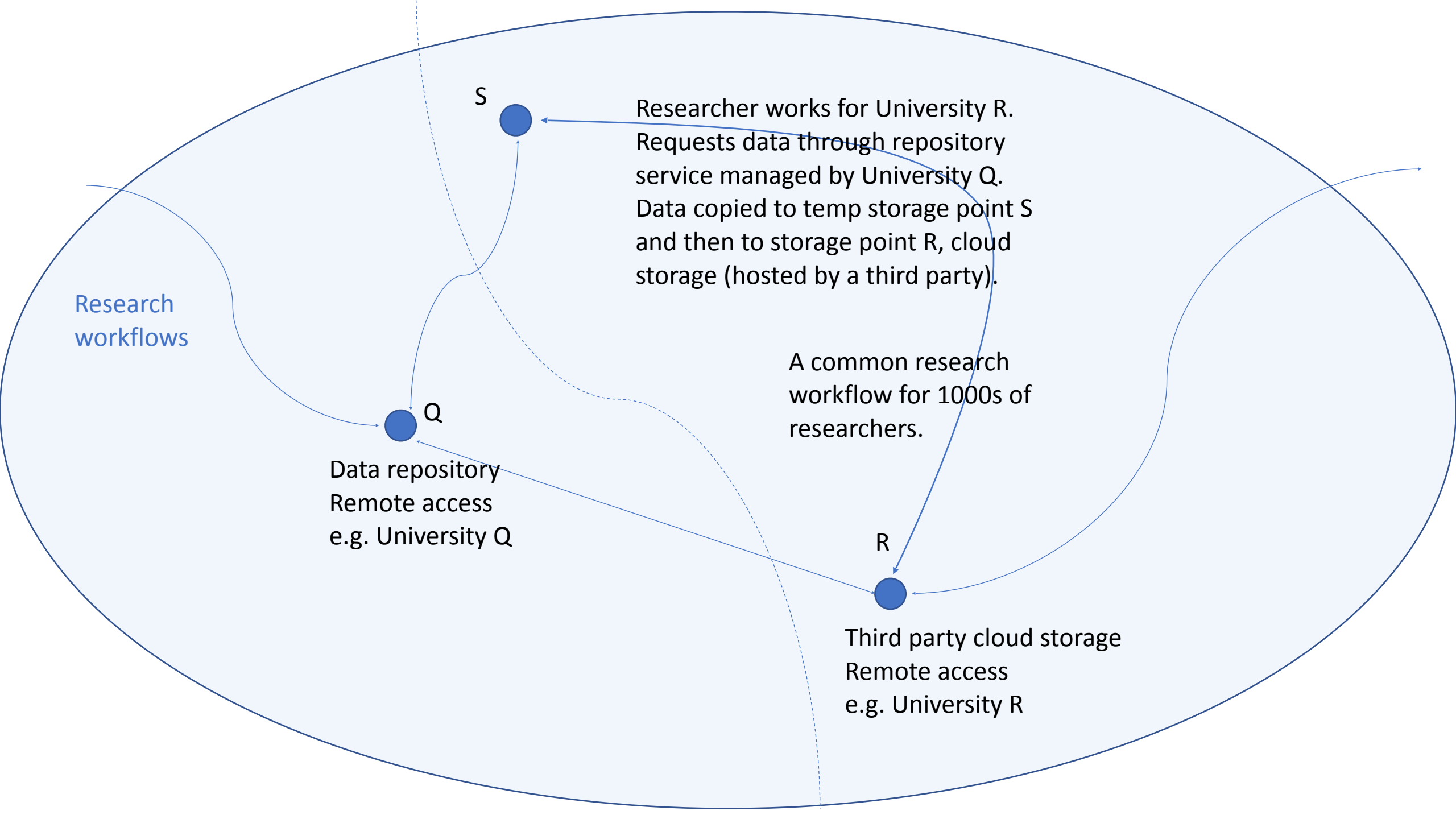
- Create

- Data is generated and stored temporarily at storage point A
- Data is copied from storage point A to storage point B
- At storage point B data is duplicated and processed
  
- Point A is facility storage
- Point B is research storage

- Receive

- Data copied from storage point Q and sent to storage point R where it can be received
- At storage point R data is copied to storage point S where it is duplicated and processed
  
- Point Q is a data store
- Point R is temp storage
- Point S is research storage





Research workflows

Q

Data repository  
Remote access  
e.g. University Q

S

Researcher works for University R.  
Requests data through repository  
service managed by University Q.  
Data copied to temp storage point S  
and then to storage point R, cloud  
storage (hosted by a third party).

A common research  
workflow for 1000s of  
researchers.

R

Third party cloud storage  
Remote access  
e.g. University R

# Breaking it all down...

- Examine key transition points and relationships in research workflows that improve support for:
  - Data sharing and collaboration 🙌 👤 ✅
  - Data movement and packaging 🚀 📦 ✅
  - *Data processing, versioning and provenance* 🖥️ 📅 🖨️ ⚒️
- Embed data curation (as a consideration) into the beginning of a research project and throughout (*velocity and viscosity*)
- Promote or develop curation systems and approaches where they are broadly applicable for wide reuse

# What about the “dark binding matter”?

- Sharing 🙌 ✅
- Collaboration 👥 ✅
- Movement 🚀 ✅
- Packaging 📦 ✅
- Processing 💻 ✅
- Versioning 📅 🛠️
- Provenance 📄 🛠️
- Evolution of the FileSender service into an integral function CloudStor
- Additions: encryption, vouchers, notifications, Rocket, Sync client, FileSender API
- Recently: *Collections plugin, tenant portal, Archivematica*
- Data in active state (kicking off a research project) and transitioning into an archive (ready for reuse)

