

PASIG 2019

February 12-14, 2019 | El Colegio de
México | Mexico City

PRESERVATION ACTION REGISTRIES

Sarah Romkey, Artefactual Systems
Carl Wilson, Open Preservation Foundation
Jon Tilbury, Preservica

#PASIG2019



Preservation Action Registries

Carl Wilson (OPF), Jon Tilbury (Preservica), Sarah Romkey (Artefactual)

JISC

Arkivum, Artefactual, Preservica, Open Preservation Foundation

<http://parcore.org/presentations/>

Agenda

- Part 1: Background & Motivation
- Part 2: Project outcomes
- Part 3: Proof of concept implementation
- Part 4: Long Term Vision
- Part 5: How can you get involved

Agenda

- **Part 1: Background & Motivation**
- Part 2: Project outcomes
- Part 3: Proof of concept implementation
- Part 4: Long Term Vision
- Part 5: How can you get involved

Background: The problem

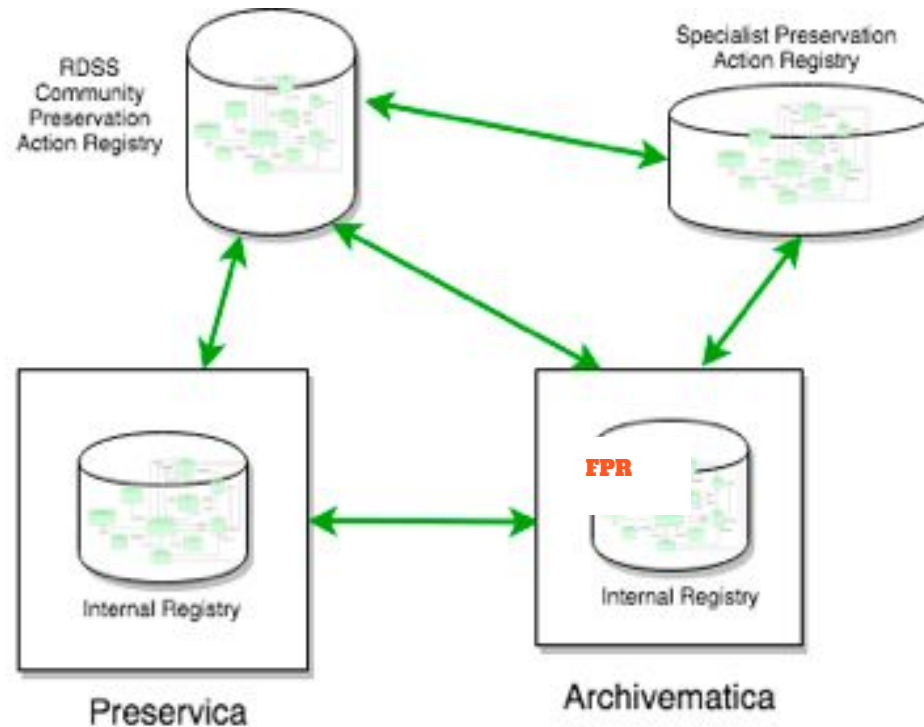
- Users want the best advice, wherever it comes from
 - Identification, property extraction, validation, migration, rendering, tools
- Multiple parallel initiatives research and advise on best practice
 - Products such as Preservica & Archivematica
 - Practitioners
 - Academics
 - Specialists
- but they don't talk to each other effectively

Background: Motivation and Objectives

- Want to
 - Improve the quality and ease of use of advice sent to practitioners as soon as it is available
 - Improve research cooperation and reduce repetition
- Expected outcome
 - Provide a mechanism to exchange information between all parties regardless of which system they use
- Exclusions
 - Protocols for prioritising and authorising which advice applies to which user / system / intent
 - One registry to rule them all

Background: Jisc RDSS Project

Development of a multi-vendor shared services platform drove discussions of interoperability of format policies (i.e. “preservation actions”) between preservation systems.



Background: Project Conception

A JISC funded project to initiate the process to deliver benefits to RDSS users

Arkivum, Preservica and Artefactual as RDSS product suppliers

Open Preservation Foundation as respected independent shared DP technology supplier

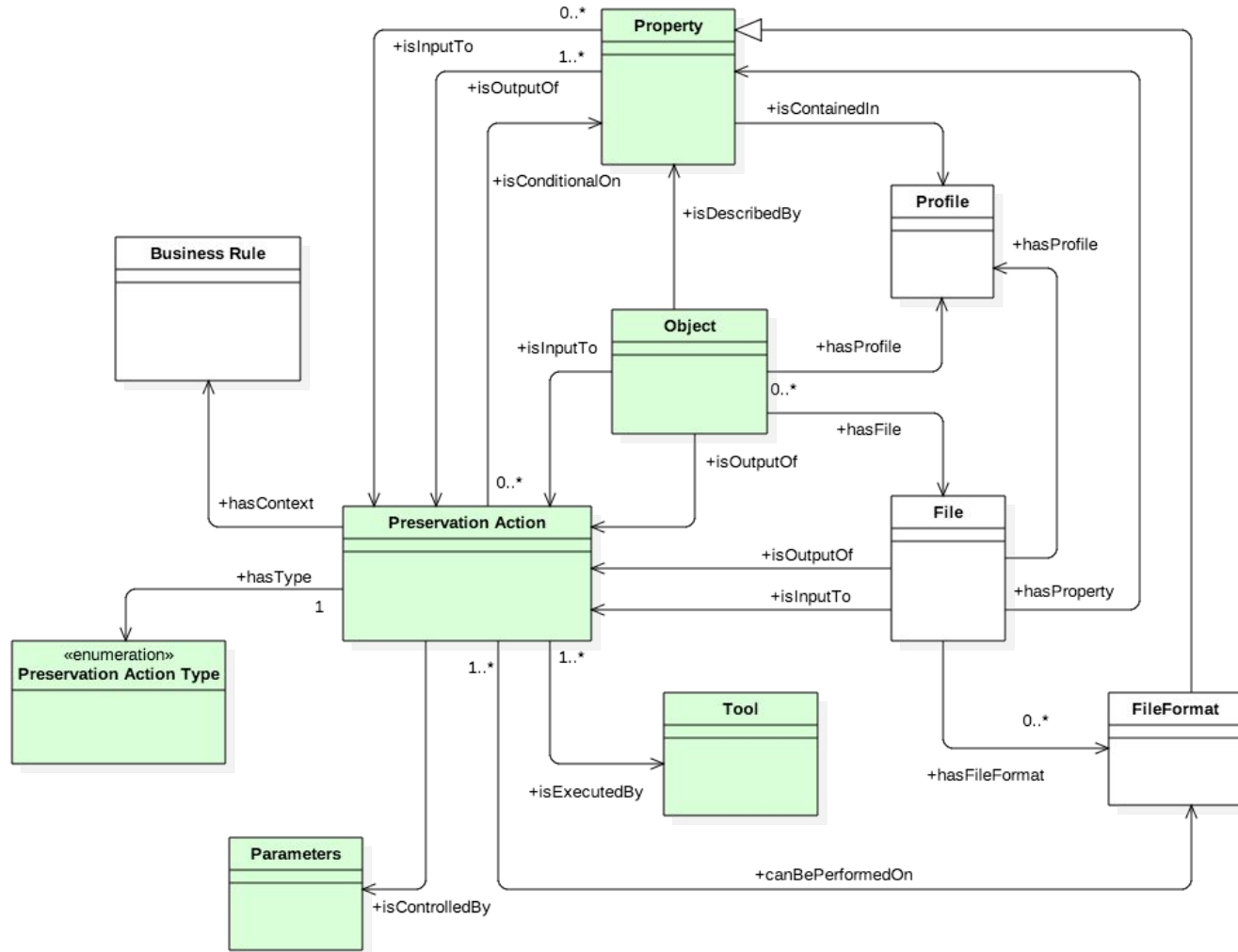
Agenda

- Part 1: Background & Motivation
- **Part 2: Project outcomes**
- Part 3: Proof of concept implementation
- Part 4: Long Term Vision
- Part 5: How can you get involved

What have we produced and why?

Conceptual Model	<ul style="list-style-type: none">• Common framework for everyone to work to• Something to argue about and agree upon!• Interlingua between preservation systems
Json Schemas	<ul style="list-style-type: none">• Formal definition of the PAR model• Machine readable, used in API payloads• Used to test and validate interoperability
API	<ul style="list-style-type: none">• Common interface for preservation systems• Well defined way to exchange information
Executable DP Actions	<ul style="list-style-type: none">• Cross-platform way to deploy/run tools• Unambiguous and vendor independent
Proof of Concept	<ul style="list-style-type: none">• Prove PAR is possible!• Not just a talking shop or paper exercise• Reference implementation to share

PAR Conceptual Model



<https://doi.org/10.6084/m9.figshare.6628418>

<https://doi.org/10.6084/m9.figshare.6628418>

Core concepts: summary

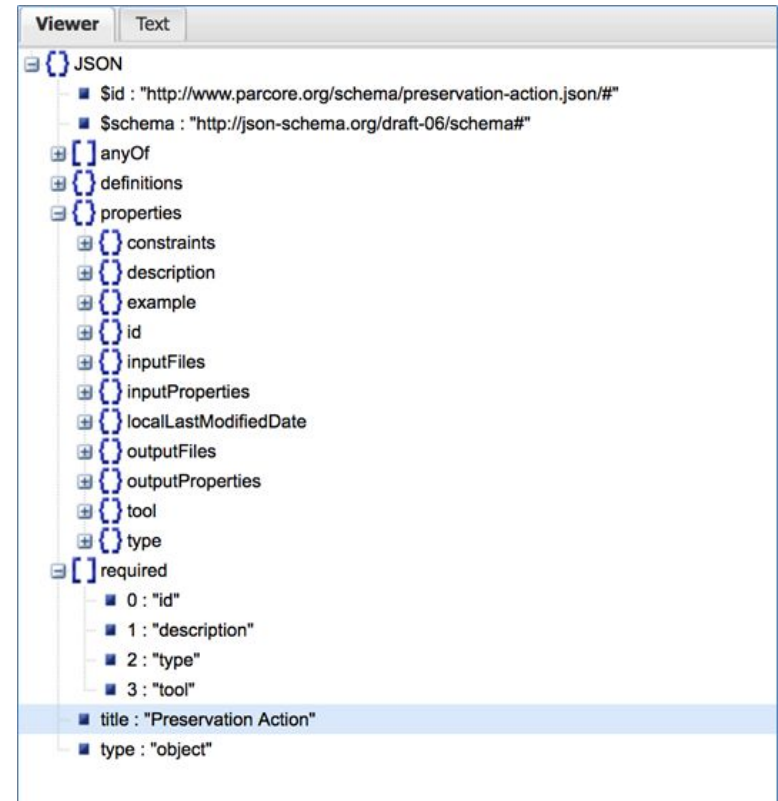
- A Preservation **Action** is something **done** as part of DP
- An **Action** has an **Action Type** defined by PREMIS
- An **Action** acts upon an input **Object** or **File**
- May take **Properties** as inputs
- Executed using one or more **Tools**
- Controlled/configured by a set of **Parameters**
- May create an output **Object** or **File**
- May create/extract **Properties** and provide as outputs
- Context on when/why/who/how is in **Business Rules**

Concrete Examples (why it can get a bit tricky!)

- Property extraction
 - E.g. MediaInfo using EBUCore to get height, width, frame rate
 - Same Tool, same File, multiple Properties, multiple Schema
- Checksum generation and validation
 - E.g. md5sum to generate a MD5 checksum for a file: type and value
 - E.g. md5sum to get PASS/FAIL for given MD5 checksum
 - Same Tool, same File, different Action, Properties go input↔output
- File format identification
 - E.g. Droid, Siegfried, Fido, File, Tika
 - Business rules on which tools work best and when
 - Business rules on priority/ranking/consensus
 - Is file format a first class entity or just another property?

JSON schemas

- Tool
- Action
- Action Type
- Format
- Property
- Business Rule
- Lots still to do, e.g.:
 - Objects
 - Containers
 - Linked Data
 - Execution by people as well as machines
 - Avoid combinatorial explosions: actions, properties, tools



<https://github.com/JiscRDSS/rdss-par/tree/master/api>

APIs

The screenshot displays the Jisc RDSS API documentation for the 'Preservation Actions' endpoint. The left sidebar contains a navigation menu with categories: AUTHENTICATION, PRESERVATION ACTION REGISTRY, BUSINESS RULES, FILE FORMATS, PAR PROPERTIES, PRESERVATION ACTION TYPES, and PRESERVATION ACTIONS (which is expanded). Under 'PRESERVATION ACTIONS', there are five items: 'Retrieve all preservation actions' (GET), 'Create a new preservation action' (POST), 'Delete a preservation action' (DELETE), 'Retrieve a preservation action' (GET), and 'Update an existing preservation action' (PUT). The main content area is titled 'Preservation Actions' and 'Retrieve all preservation actions'. It includes a description: 'Allow to retrieve the details of all the preservation action'. Below this, there are sections for 'QUERY PARAMETERS' and 'HEADER PARAMETERS'. The 'QUERY PARAMETERS' section lists: 'limit' (string, integer to limit results), 'offset' (string, integer to specify offset), 'modifiedAfter' (string, filter by localLastModifiedDate), and 'modifiedBefore' (string, filter by localLastModifiedDate). The 'HEADER PARAMETERS' section lists: 'tool' (string, filter by tool ID). On the right side, there is a 'Response samples' section showing a '200' status code and a JSON response sample for 'application/json'. The JSON sample includes fields like 'constraints', 'description', 'example', 'id' (with 'guid', 'name', 'namespace'), 'inputFiles', 'inputProperties', 'localLastModifiedDate', 'outputFiles', and 'outputProperties'.

https://52.209.71.78/#tag/Preservation-Actions

Preservation Actions

Retrieve all preservation actions

Allow to retrieve the details of all the preservation action

QUERY PARAMETERS

limit	string	An integer to limit the number of preservation action returned. Default value is zero, which will no filter the result.
offset	string	An integer to specify the offset of the first element of the list of preservation action returned. Default value is zero, which will no filter the result.
modifiedAfter	string	Filter the preservation action to return only the ones having a localLastModifiedDate value GREATER than the one passed in. Allowed datetime formats are YYYY-MM-DD\T\hh:mm:ssTZD and YYYY-MM-DD. In the second scenario, when the time is not provided, it will be assumed to be 00:00:00 UTC time.
modifiedBefore	string	Filter the preservation action to return only the ones having a localLastModifiedDate value LOWER than the one passed in. Allowed datetime formats are YYYY-MM-DD\T\hh:mm:ssTZD and YYYY-MM-DD. In the second scenario, when the time is not provided, it will be assumed to be 00:00:00 UTC time

HEADER PARAMETERS

tool	string	Filter the preservation actions by the Tool they use, provided its ID. Multiple IDs can be passed in within the same string, linked together using an ampersand symbol as follow: ID_1&ID_2&ID_3. When more than one ID is provided, the endpoint will return all the preservation actions that use one of the IDs in the list. When used together with the preservation_action_type header parameter, the endpoint will return only those preservation actions satisfying both filters' criteria. A list of Tool IDs cannot be combined with another list of Preservation Action Type IDs.
------	--------	---

GET /preservation-actions

Response samples

200

application/json

```
{
  "constraints": [
    + { ... }
  ],
  "description": "string",
  "example": "string",
  "id": {
    "guid": "string",
    "name": "string",
    "namespace": "string"
  },
  "inputFiles": [
    + { ... }
  ],
  "inputProperties": [
    + { ... }
  ],
  "localLastModifiedDate": "2018-07-31T16:03:25Z",
  "outputFiles": [
    + { ... }
  ],
  "outputProperties": [
    + { ... }
  ],
}
```

<https://github.com/JiscRDSS/rdss-par/tree/master/api>

Executable Tool Definitions



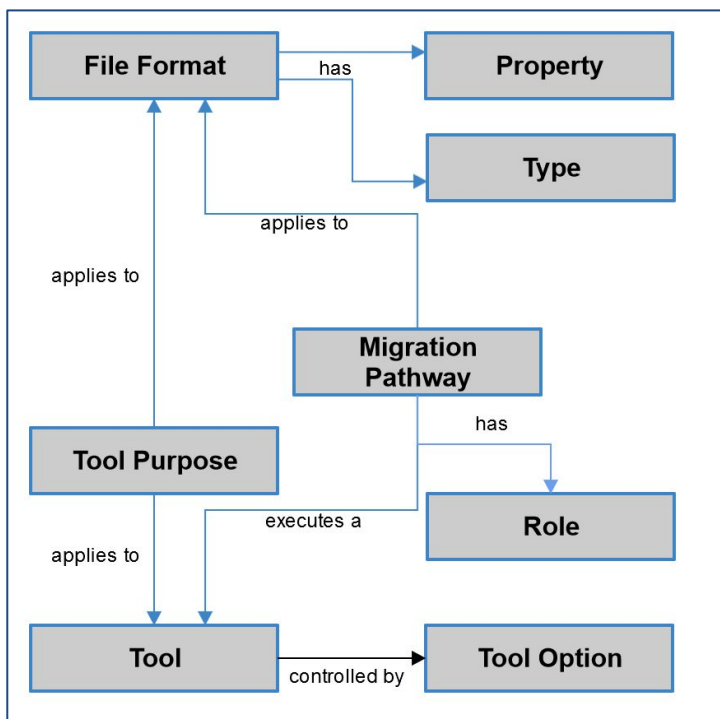
- Machine readable spec for running a tool
 - Tool command line
 - Parameters and flags
 - Inputs and outputs
 - Pre and post processing
 - Mapping to command lines
 - e.g. extract EBUCore elements/attributes from MediaInfo
 - e.g. extract PRONOM ID from Fido
 - e.g. supply checksum value and filename to md5sum
 - stdin, stdout, stderr, input/output files, parameters and flags, error codes
 - Tools, libraries, dependencies, containers
 - Mapping to web services
 - e.g. validators, file format convertors, metadata extractors
 - GET, PUT, POST, sync and async interaction, XML, json, http codes
- <https://github.com/JiscRDSS/rdss-par/tree/master/examples/cwl>

Agenda

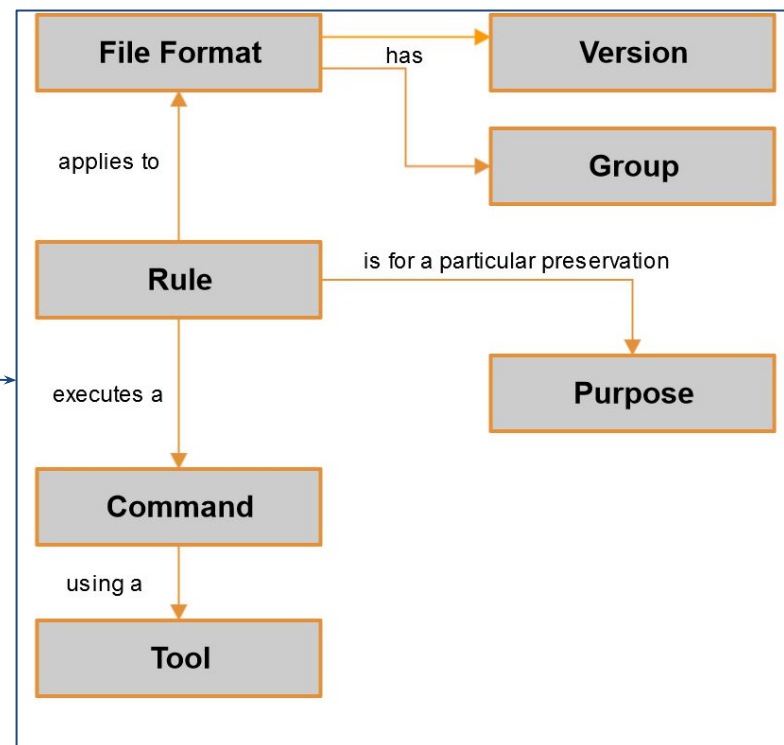
- Part 1: Background & Motivation
- Part 2: Project outcomes
- **Part 3: Proof of concept implementation**
- Part 4: Long Term Vision
- Part 5: How can you get involved

Registry (In)compatibility

Preservica Registry

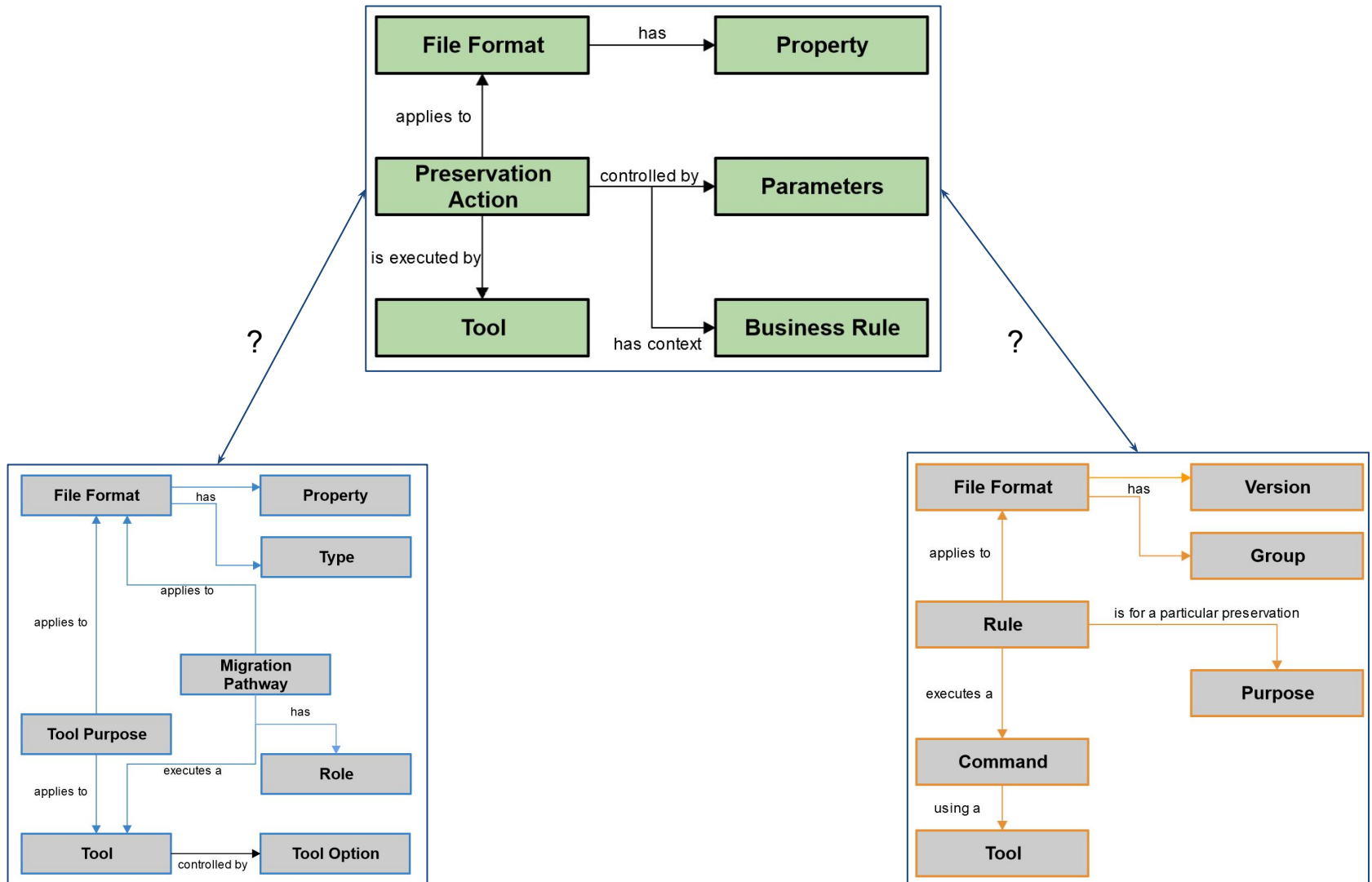


Archivematica FPR



?

Common Language



Extract Metadata Use Case

User Bob at Modern Institution is responsible for preserving AV content.

Modern Institution has decided to adopt EBUcore as the metadata standard for its AV holdings.

Bob discovers that a recent update to the MediaInfo tool allows him to extract metadata about AV files in EBUCore XML format.

He introduces this change into his Preservica Preservation System and exports this information as a Preservation Action expressed in PAR Schema.

Extract Metadata Use Case

User Alice at Post-Modern Institution imports this newly published information into her Archivematica preservation system using its PAR API.

Each PAR entry is stored in her installation as JSON files.

She does not have to read it in JSON or even PAR schema format or have any knowledge about how Bob's system works.

She instead reviews it in a PAR Form that appears under Archivematica's Preservation Planning tab.

Extract Metadata Use Case

Alice decides that this new capability is something she wants to implement.

She selects the 'Convert to FPR' option to enter it as an active 'Metadata Extraction' Rule and Command in her Archivemata instance.

At this point the PAR information is entered into the Archivemata database without affecting any existing Commands or Rules.

The new Rules and Command are linked to the PAR Preservation Action.

Demo

Demo Export and Import of Preservation Action

<https://youtu.be/ELjbgksjBmU>

Agenda

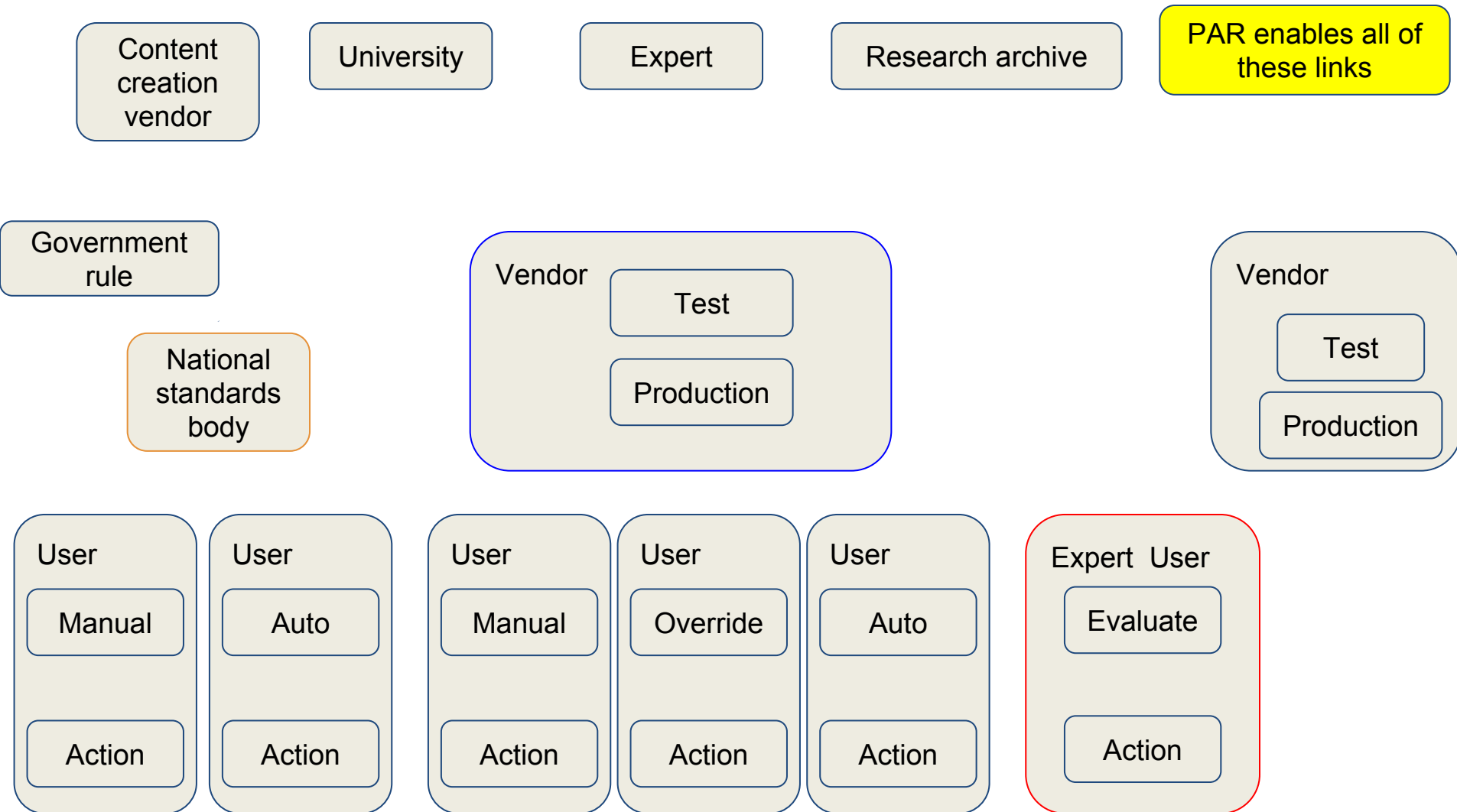
- Part 1: Background & Motivation
- Part 2: Project outcomes
- Part 3: Proof of concept implementation
- **Part 4: Long Term Vision**
- Part 5: How can you get involved

Long Term Vision

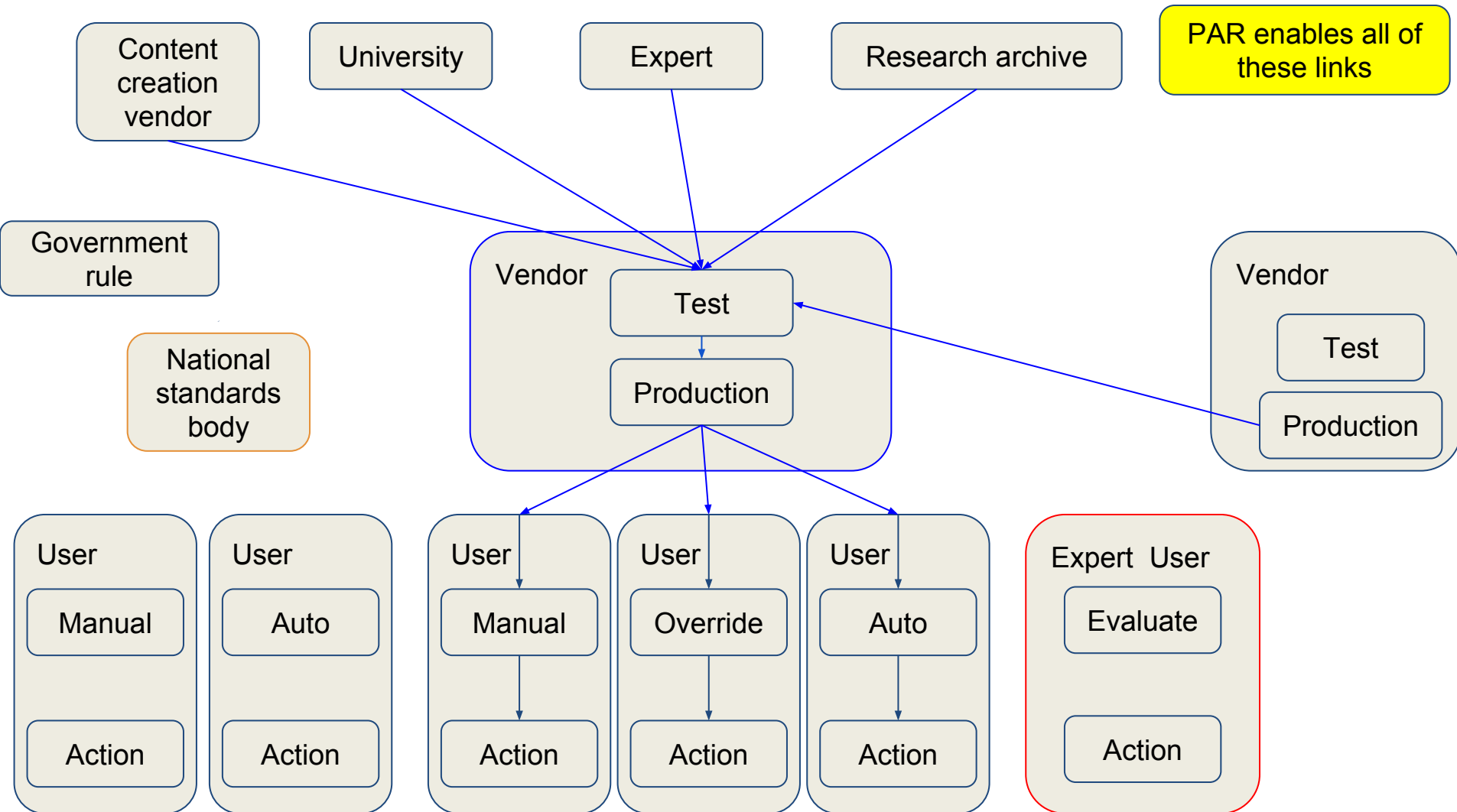
- All users want trusted information
- Some users want to make their own choices
- Some countries or communities have local rules for preservation formats
- Recommended action depends on user context
e.g. public access vs minimum storage cost
- Many users want it to just work for their product in their context

PAR must be able to help in all of these situations

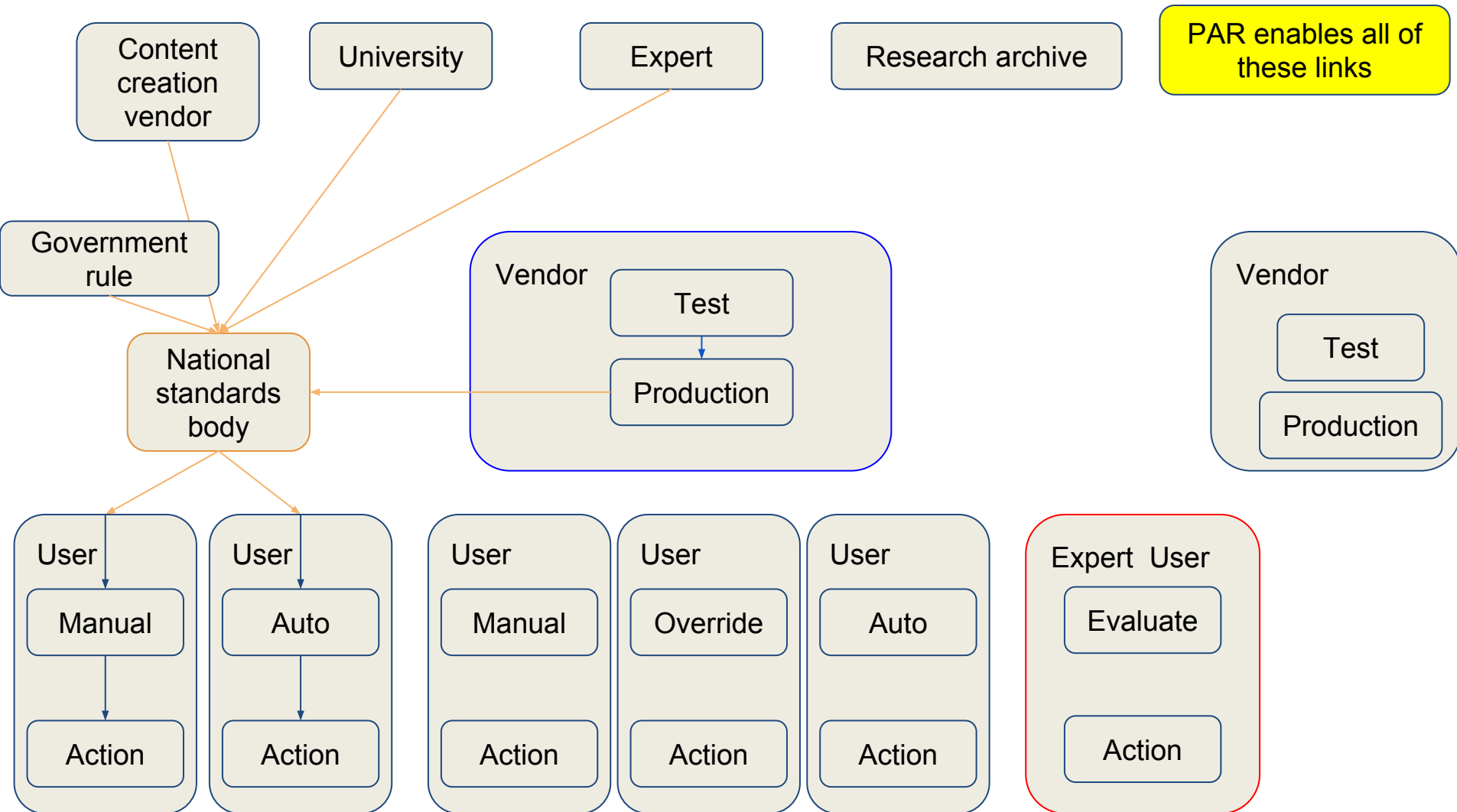
Example registry network



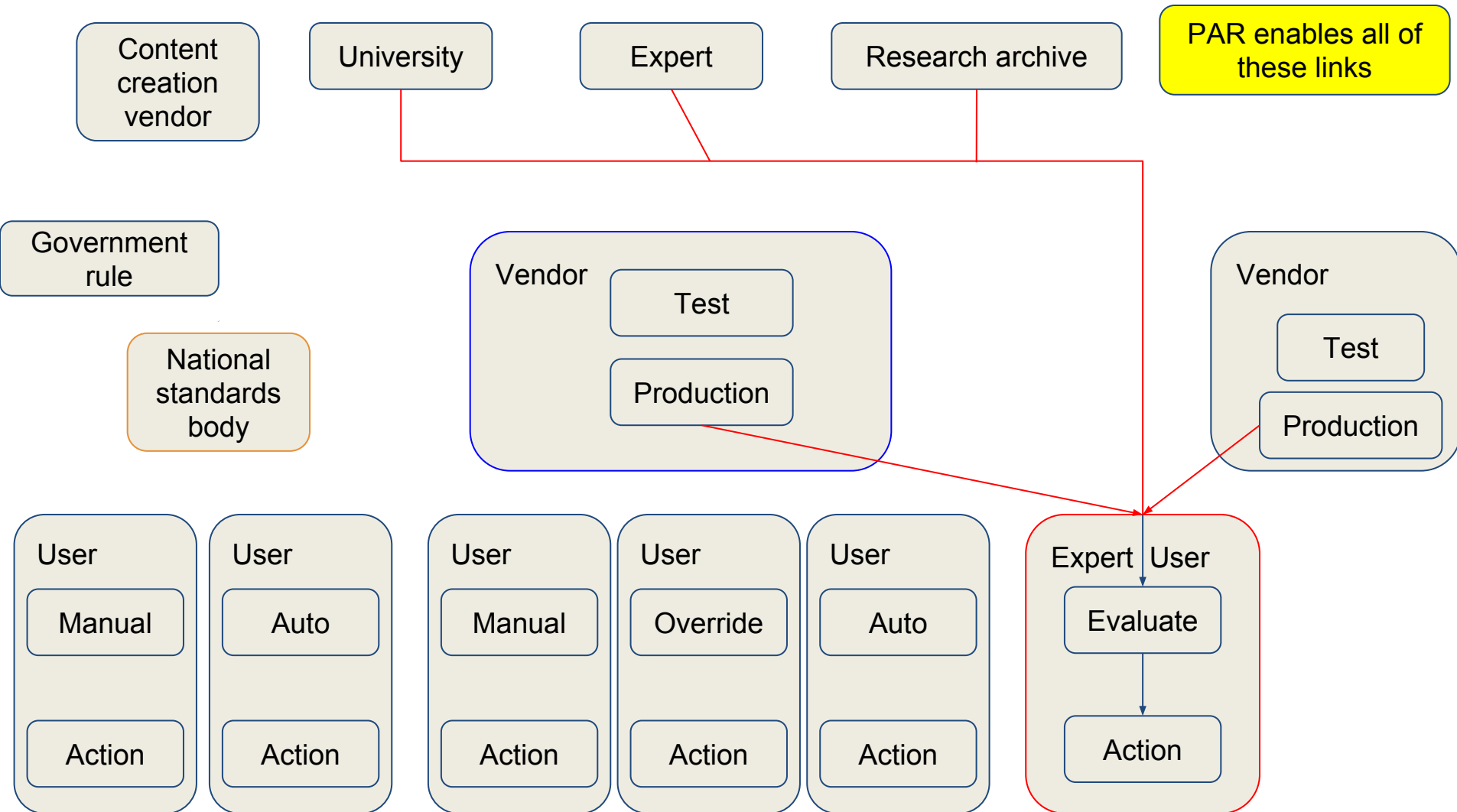
Example registry network



Example registry network



Example registry network



Agenda

- Part 1: Background & Motivation
- Part 2: Project outcomes
- Part 3: Proof of concept implementation
- Part 4: Long Term Vision
- **Part 5: How can you get involved**

Putting PAR V1 into Production

Technical

- Complete the data model
- Complete the API
- Implement in real systems to do real actions
- Tools to synchronize registries

Community

- Gather user examples
- Identify funders
- Adding endpoints to your registry
- Help code some of the components

Interested?

*PAR could enable a global exchange of best practice
between all stakeholders in DP*

It could define format preservation for the next decade

Further PAR conversation and information to be held in
the Library area immediately post this presentation

For further information and to register your interest:

<http://bit.ly/2tauvlu>

OR

<https://openpreservation.org/about/projects/par/>

Resources

- Project pages
 - <http://www.parcore.org/>
- Github repo
 - <https://github.com/JiscRDSS/rdss-par/>
- iPRES paper
 - <https://doi.org/10.6084/m9.figshare.6628418>
- DPC blog post
 - <https://www.dpconline.org/blog/a-new-era-in-collaboration-in-digital-preservation-research>
- Project announcement and contacts
 - <http://openpreservation.org/news/arkivum-artefactual-the-open-preservation-foundation-and-preservica-collaborate-on-new-jisc-initiative-for-sharing-preservation-action-best-practice/>
- Webinar
 - <http://openpreservation.org/event/introducing-preservation-action-registries/> (OPF login required)