



# Open Preservation Foundation and The Preservation Action Registry

Martin Wrigley, Executive Director, OPF

## Martin Wrigley



30+ years experience delivering software and solutions - mostly in Mobile Telecoms

10+ years experience of managing a membership driven open source association

OPF Executive Director since September 2017

Expanding my knowledge of the finer points of Digital Preservation

# Who is OPF?

- A not for profit, global membership association providing stewardship of open-source tools for the digital preservation community.
- Founded in 2010 to sustain the results of the EU PLANETS project
- The OPF reference toolset now includes veraPDF, JHOVE and more

# What is OPF's purpose?

## **OPF Vision**

Open sustainable digital preservation

## **OPF Mission**

Enabling shared solutions for effective and efficient digital preservation; the Open Preservation Foundation leads a collaborative effort to create, maintain and develop the reference set of sustainable, open source digital preservation tools.

This set of tools (including software and standards) enables organisations to evaluate, validate, document, mitigate risk, and process digital content to be preserved in line with desired policies and community best practice.

## **Values**

- Open
- Member driven
- Collaborative & Inclusive
- Innovative

# Who are OPF members?

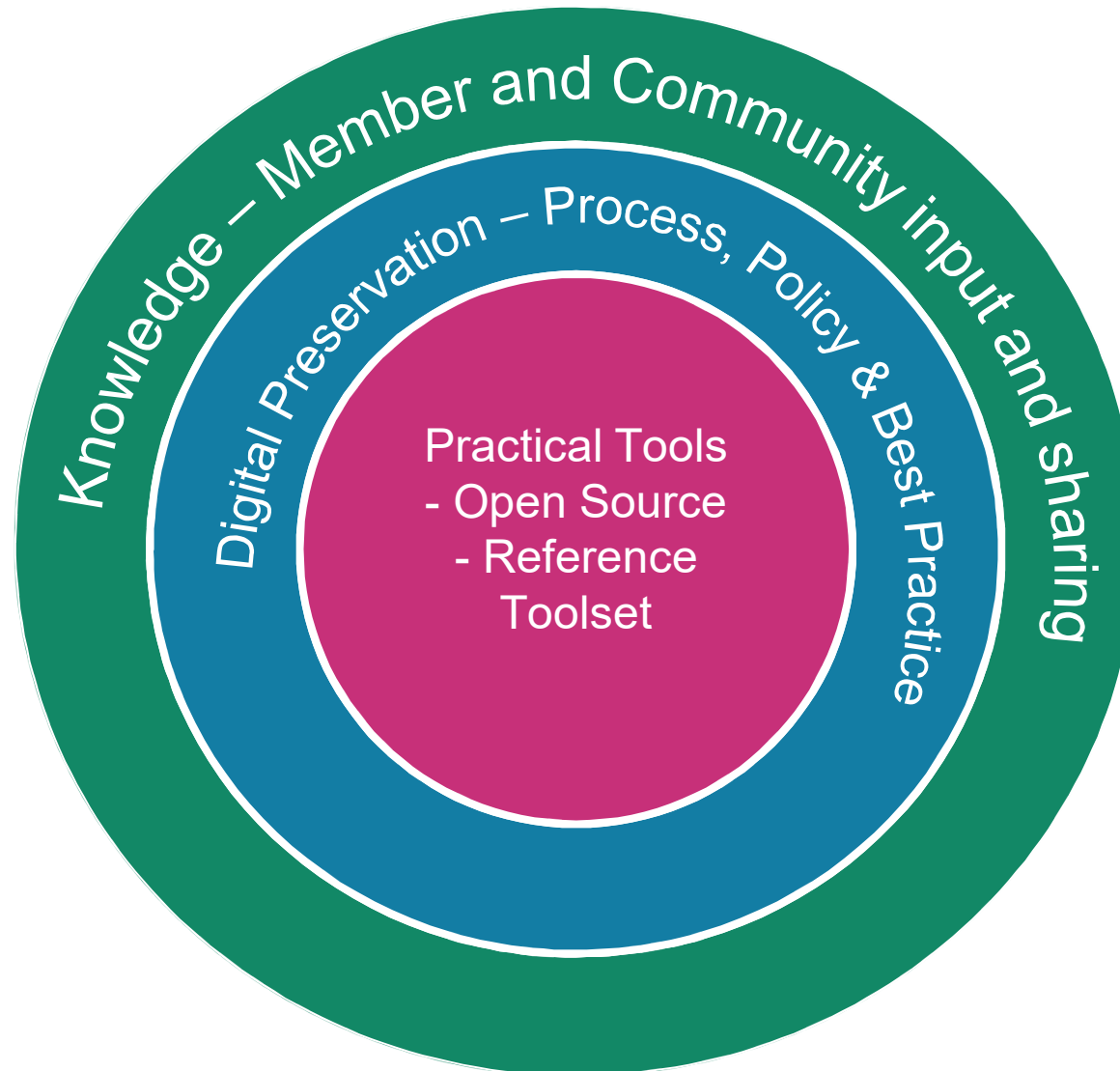
Austrian Institute of Technology  
British Library  
Bibliothèque Nationale de France  
Goportis  
International Atomic Energy Archives  
Jisc  
Koninklijke Bibliotheek  
Det Kgl. Bibliotek  
Nationaal Archief  
The National Archives UK  
Nasjonalbiblioteket  
Rigsarkivet  
Ex Libris  
Rahvusarhiiv

Latvijas Nacionālā bibliotēka  
Österreichische  
Nationalbibliothek  
Preservica  
Yale University Library  
Albert-Ludwigs Universität  
University of North Carolina  
Portico  
PSNC (Poznan Supercomputing &  
Networking Centre)  
Artefactual  
Biblioteca Nacional de Portugal  
Arcsys Software

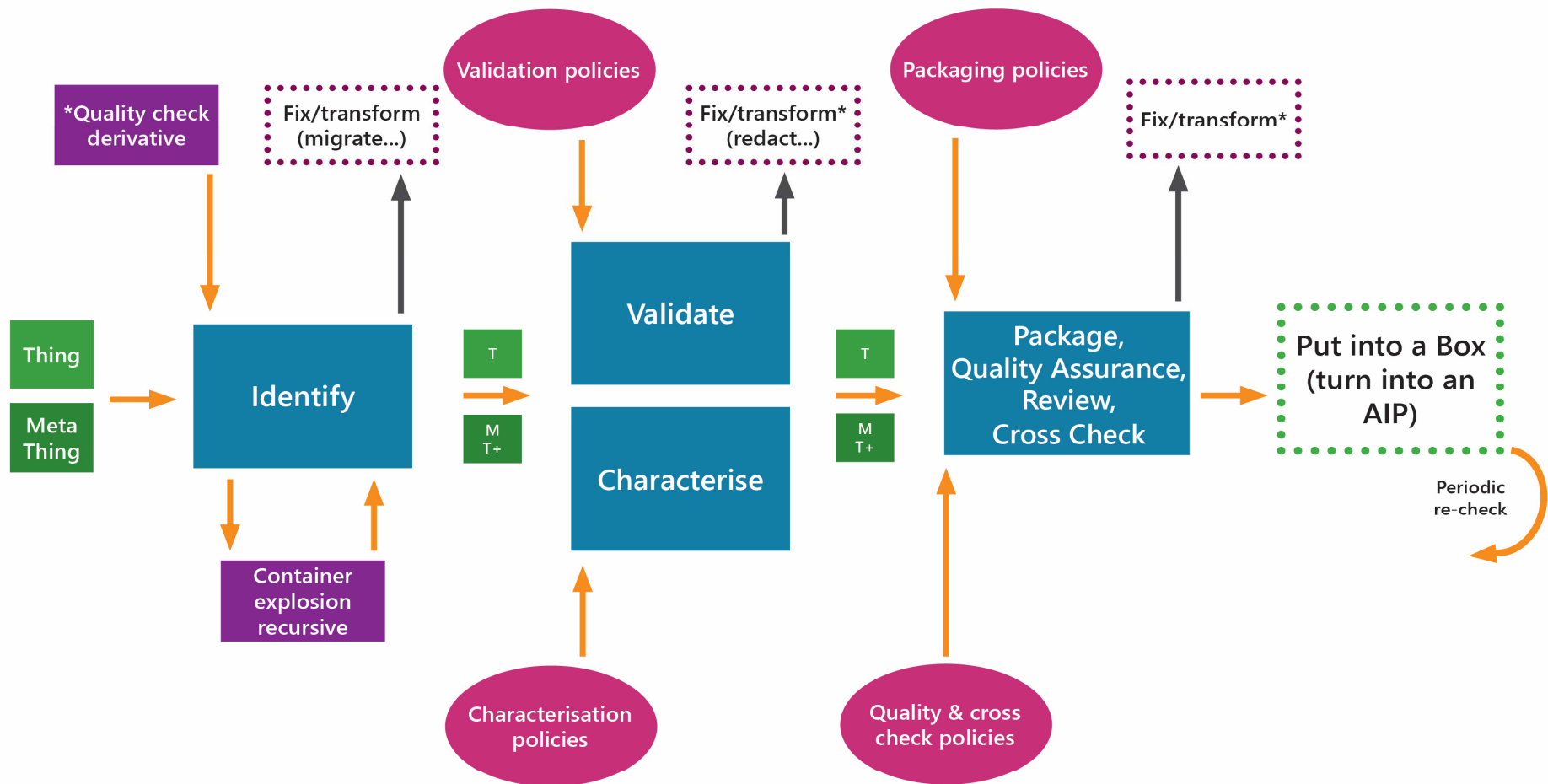
# What does OPF do?

- Community Knowledge
- Sharing knowledge
- Develop the OPF reference toolset
- Deliver to development roadmaps
- Community engagement
- Webinars and training
- Interest Groups and Tech Clinics
- OPF Software Maturity Model
- Hosting community services e.g. COPTR
- Website, blogs, events

# OPF – Digital Preservation Knowledge and Tools



# OPF Reference Toolset – generic process



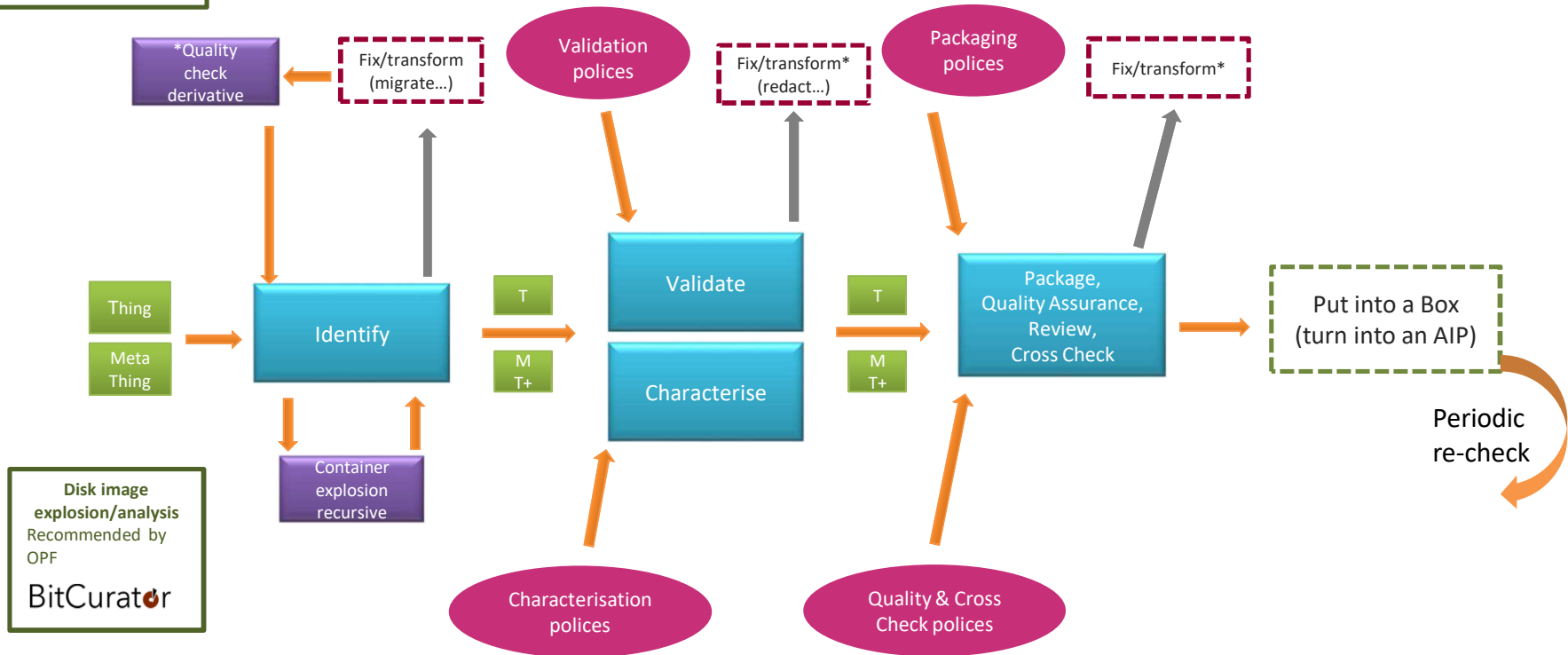
Thing is (or is becoming) a Submission Information Package



# OPF Tool Mapping

**Transform**  
Database archiving /  
Extraction tools  
Recommended by OPF  
**SIARD (SQL database to  
XML format)**

**Derivative check tools**  
Maintained through OPF  
**xcorsound** WAV, MP3



**Disk image  
explosion/analysis**  
Recommended by  
OPF  
**BitCurator**

**Identification tools**  
Maintained through OPF  
**fido**  
**Format Sniff**  
Recommended by OPF  
**DROID**  
**PRONOM**  
**FILE**

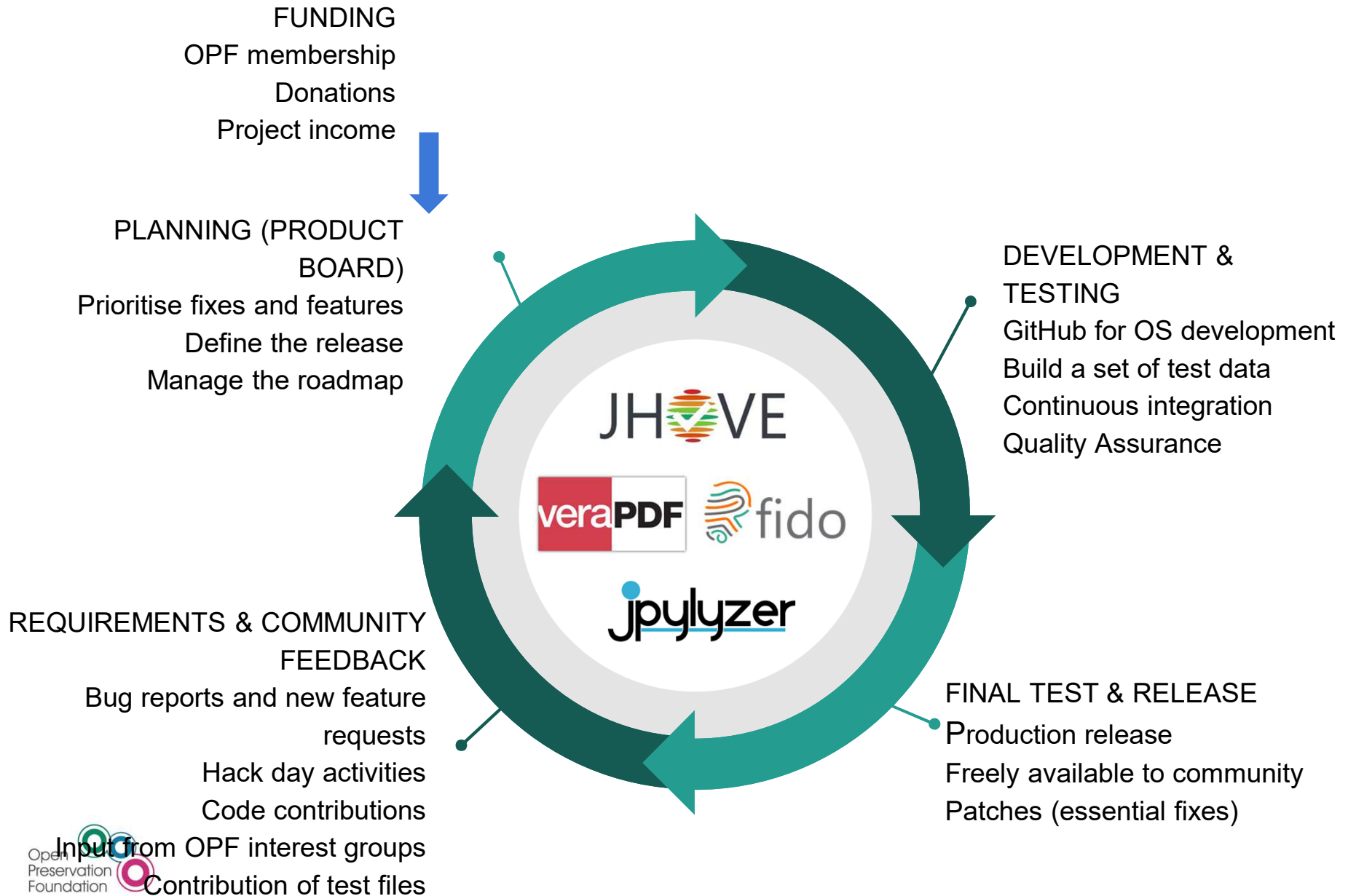
**Validation and Characterisation  
tools**  
Maintained through OPF  
**veraPDF** PDF/A  
**JHove** PDF, JPEG, WAV,  
PNG, WARC, AIFF,  
UTF8 TEXT, XML,  
HTML, GZIP, ASCII  
TEXT, MP3, GIF,  
JPEG2000  
**(DPF Manager)**  
**TIFF module** TIFF  
**Jpylyzer** JPEG2000

**Information Packaging  
tools**  
**TBA**

**Cross Check tools**  
**TBA**

**Quality check tools**  
**E-ARK CEF SIP validator**

# How do OPF projects work?





# Preservation Action Registry

# PAR Background: The problem

- Users want the best advice, wherever it comes from
    - Identification, property extraction, validation, migration, rendering, tools
  - Many sources for current ‘best practice’
    - Products such as Preservica & Archivematica
    - Practitioners
    - Academics
    - Specialists
- but they don't talk to each other effectively

# Background: Motivation and Objectives

- To provide a mechanism to exchange good practice information between organisations and preservation system suppliers regardless of which system they use.
- Explicitly: To provide compatibility/ interoperability between JISC RDSS project systems.

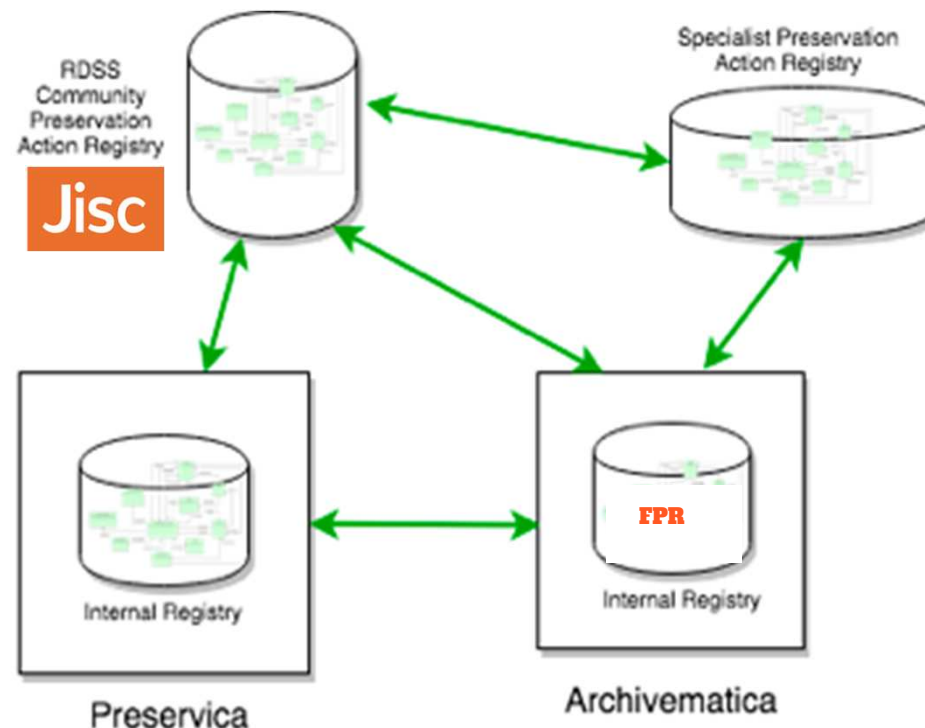
However:

It is not a single 'Best Practice'

It is not 'one registry to rule them all'

# Background: Jisc RDSS Project

Development of a multi-vendor shared services platform led to 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

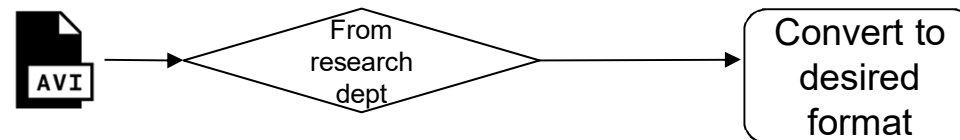
# Digital Preservation Actions

Preservation is not just about file formats, it's about making sense of data



The specific action depends on the context, and the policies.

– what action is being taken and why? What is the business rule?

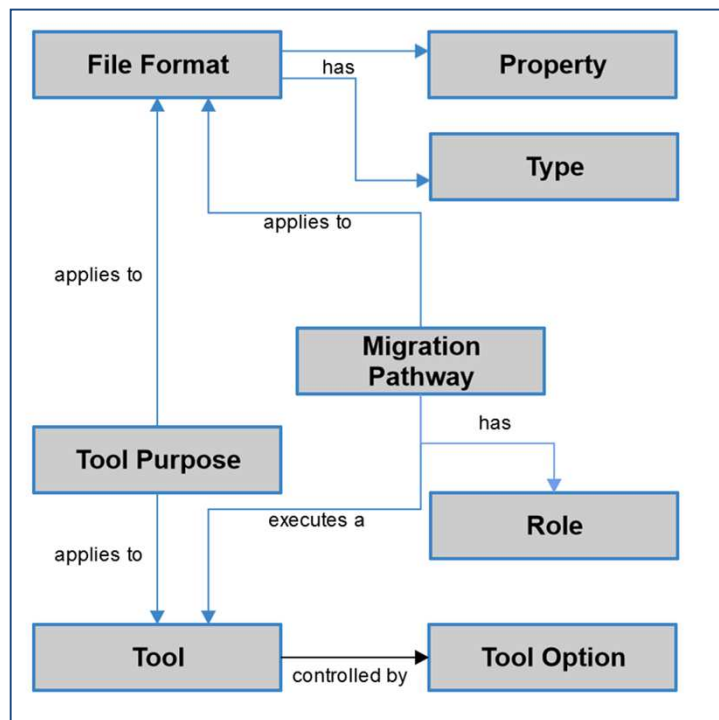


Today - preservation actions are not portable across systems  
(e.g. Archivematica, Preservica, others)

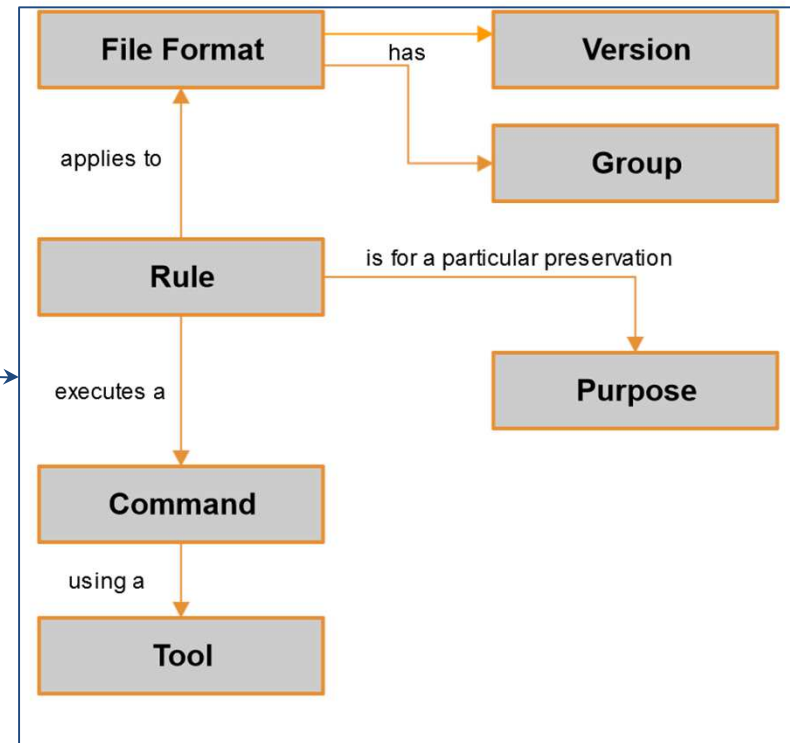


# Current Registry (In)compatibility

Preservica Registry

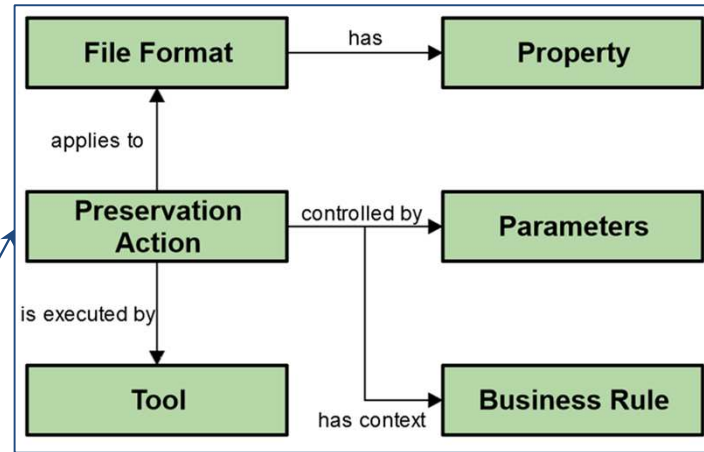


Archivematica FPR



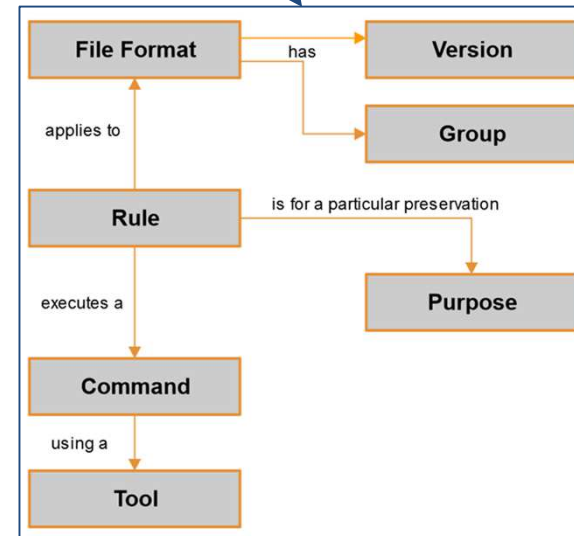
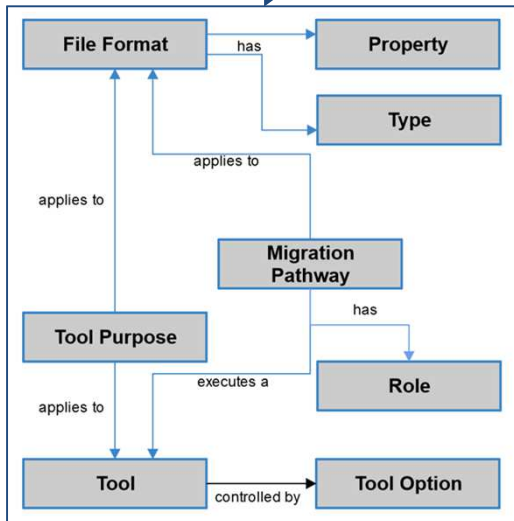
?

# Common Language



?

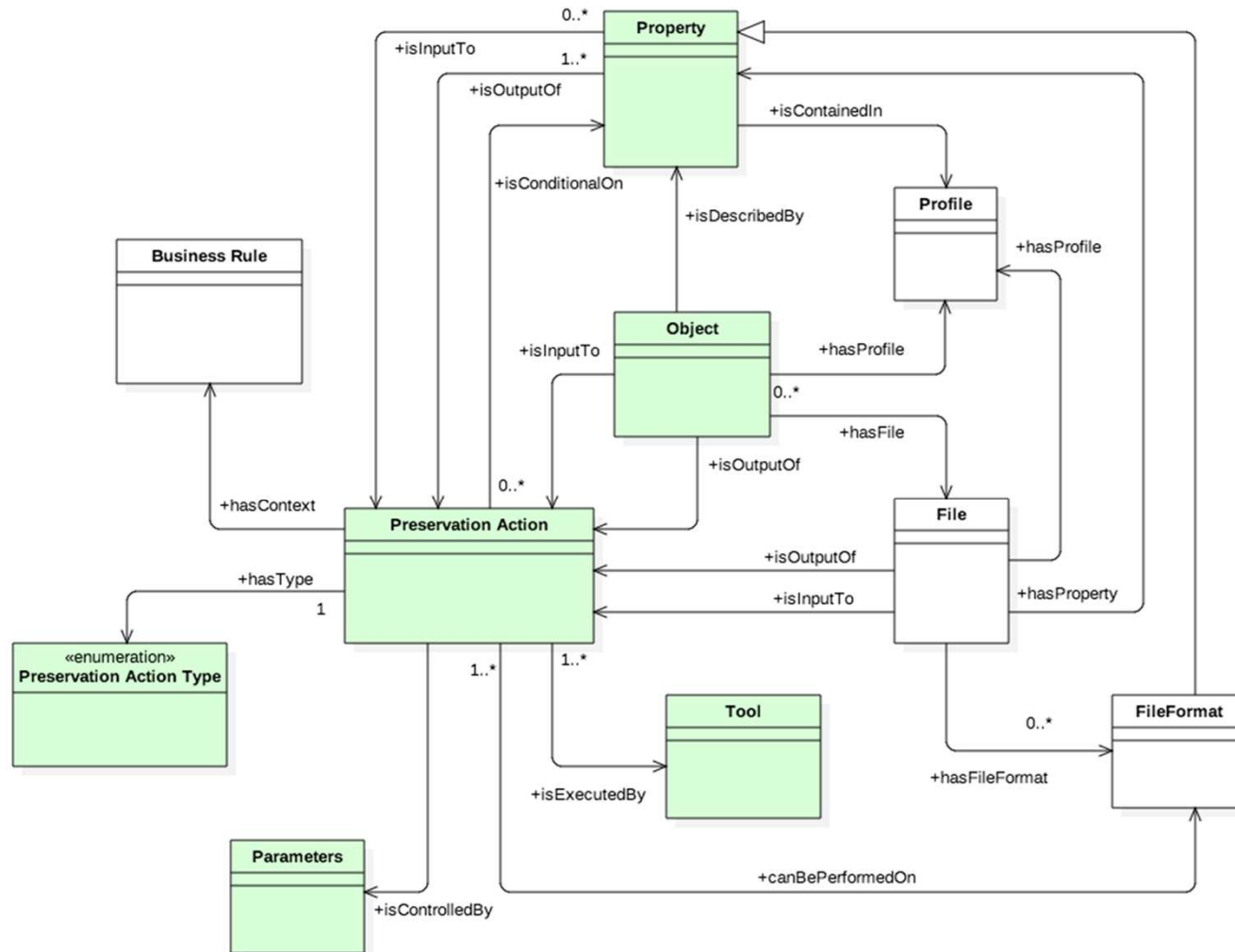
?



# What have we produced and why?

Conceptual Model	<ul style="list-style-type: none"><li>● Common framework for everyone</li><li>● Language between preservation systems</li><li>● Still under definition...</li></ul>
Json Schemas	<ul style="list-style-type: none"><li>● Formal definition of the conceptual model</li><li>● Machine readable, used in API payloads</li><li>● Used to test and validate interoperability</li></ul>
API	<ul style="list-style-type: none"><li>● Common interface for preservation systems</li><li>● Well defined way to exchange information</li></ul>
Executable Digital Preservation Actions	<ul style="list-style-type: none"><li>● Cross-platform way to deploy/run tools</li><li>● Unambiguous and vendor independent</li></ul>
Proof of Concept	<ul style="list-style-type: none"><li>● Reference implementation to share</li><li>● Make the idea really work between Preservica and Archivemata</li></ul>

# PAR Conceptual Model



# JSON schemas

- Tool
- Action
- Action Type
- Format
- Property
- Business Rule

The screenshot shows a JSON Schema viewer interface with a tree view of a schema. The root node is 'JSON', which has several properties: '\$id', '\$schema', 'anyOf', 'definitions', 'properties', 'required', 'title', and 'type'. The 'properties' node is expanded, showing a list of properties: 'constraints', 'description', 'example', 'id', 'inputFiles', 'inputProperties', 'localLastModifiedDate', 'outputFiles', 'outputProperties', 'tool', and 'type'. The 'required' node is also expanded, showing an array of property names: 'id', 'description', 'type', and 'tool'. The 'title' property is 'Preservation Action' and the 'type' property is 'object'.

```
Viewer | Text
├── JSON
│   ├── $id : "http://www.parcore.org/schema/preservation-action.json/#"
│   ├── $schema : "http://json-schema.org/draft-06/schema#"
│   ├── anyOf
│   ├── definitions
│   ├── properties
│   │   ├── constraints
│   │   ├── description
│   │   ├── example
│   │   ├── id
│   │   ├── inputFiles
│   │   ├── inputProperties
│   │   ├── localLastModifiedDate
│   │   ├── outputFiles
│   │   ├── outputProperties
│   │   ├── tool
│   │   └── type
│   ├── required
│   │   ├── 0 : "id"
│   │   ├── 1 : "description"
│   │   ├── 2 : "type"
│   │   └── 3 : "tool"
│   ├── title : "Preservation Action"
│   └── type : "object"
```

# APIs

The screenshot displays a REST client interface for the 'Preservation Actions' API. The main panel shows the endpoint 'Retrieve all preservation actions' with a description: 'Allow to retrieve the details of all the preservation action'. Below this, the 'QUERY PARAMETERS' section lists four parameters: 'limit' (string, default 0), 'offset' (string, default 0), 'modifiedAfter' (string, filter by localLastModifiedDate), and 'modifiedBefore' (string, filter by localLastModifiedDate). The 'HEADER PARAMETERS' section lists one parameter: 'tool' (string, filter by Tool ID).

The right-hand panel shows a successful GET request to '/preservation-actions' with a 200 status code. The response is in application/json format and contains a list of preservation actions. The JSON structure is as follows:

```
{
  "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



COMMON  
WORKFLOW  
LANGUAGE



docker

```
[job mediaInfo2.cwl] completed success
{
  "width": "1280",
  "bitrate": "748253",
  "height": "720"
}
Final process status is success
```

Property extraction

Fixity check

```
[job md5check2.cwl] completed success
{
  "fixity_report": "PASS"
}
Final process status is success
```

# Next steps

- OPF coordination
  - Define project deliverables and stages in more detail
- More use cases demonstrating real benefits
- Looking for more organisations to be involved
- Extend the conceptual model to more practical cases that involve more organisations

Make PAR useful to communicate good practice between systems and organisations



# Join OPF today!

For more information get in touch...

 [martin.wrigley@openpreservation.org](mailto:martin.wrigley@openpreservation.org)

 <http://openpreservation.org/>

 <https://github.com/openpreserve>

 [@openpreserve](https://twitter.com/openpreserve)

Newsletter: [www.openpreservation.org/subscribe/](http://www.openpreservation.org/subscribe/)

For more info on PAR go to

**[www.openpreservation.org/about/projects/par](http://www.openpreservation.org/about/projects/par)**