Preservation Action Registries Workshop

iPres Event Workshop September/2019

JISC, Arkivum, Artefactual, Preservica, Open Preservation Foundation

http://parcore.org/presentations

V01 18/07/19

Introduction & Overview

- Welcome to the Preservation Action Registries Workshop!

- Today we will:
  - Introduce you to PAR
  - Work in groups where we will try out PAR together
  - Discuss improvements and next steps

Agenda

Time	Session	Facilitator/Speaker
13:30 - 13:50	**PAR Background and overview**	**Matthew Addis. Arkivum**
13:50 - 14:10	Workshop overview	Justin Simpson, Artefactual
14:10 - 16:10	PAR Action Rules Workshop	All
16:10 - 16:30	Feedback, Q&A	All

Why do we need PAR?

- Users want good advice on doing DP in practice
  - Identification, property extraction, validation, migration, rendering, emulation, packaging, safeguarding ...
  - What works, what doesn't, when to use a particular approach, who's done it before, why they did it …
- There is plenty of advice out there
  - Vendors, practitioners, academics, specialists
  - Tools, policies, examples, case studies, blogs, forums
- But advice can be really hard to find and use
  - Fragmented, hard to find, inconsistent terminology
  - Not always easy to establish trust, lack of context
  - Not precise or lacks detail, can't be used directly in DP systems

The benefits of the PAR approach

Today's Workshop!

- **Descriptions of how to do digital preservation**
  - Human and machine readable
  - Enough detail to actually do something in the real world
  - Knowledge sharing through a common language and terminology

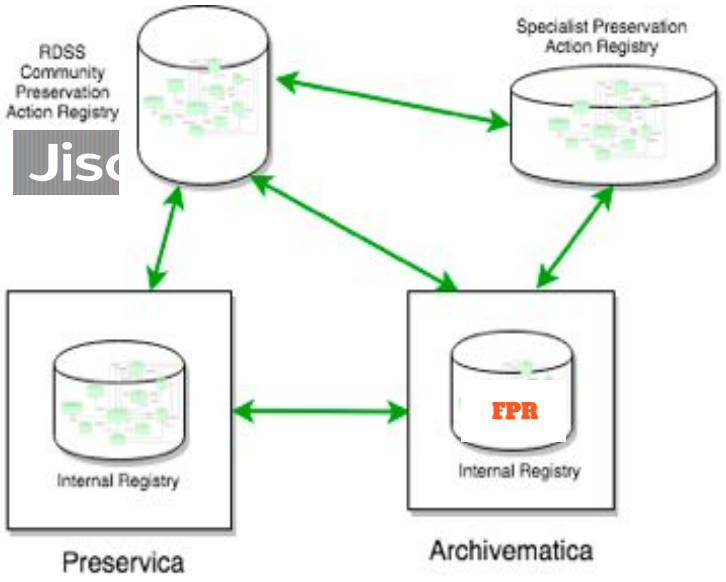
- **Registries of good preservation practice**
  - Within an institution, e.g. so staff can implement organisational policies
  - Supporting a community, e.g. DP as part of research data management
  - Discipline specific, e.g. how to preserve AV content in practice
  - National standards and guidelines, e.g. common practice across regional archives
  - Helping everybody to getting started and develop expertise in DP

- **Interoperability of DP systems**
  - Import and export from registries and exchange between systems
  - Supports transparency and trust
  - Allows comparison and migration

Where did PAR come from?

- Jisc Open Research Hub
  - Multi-vendor shared services platform for Research Data Management

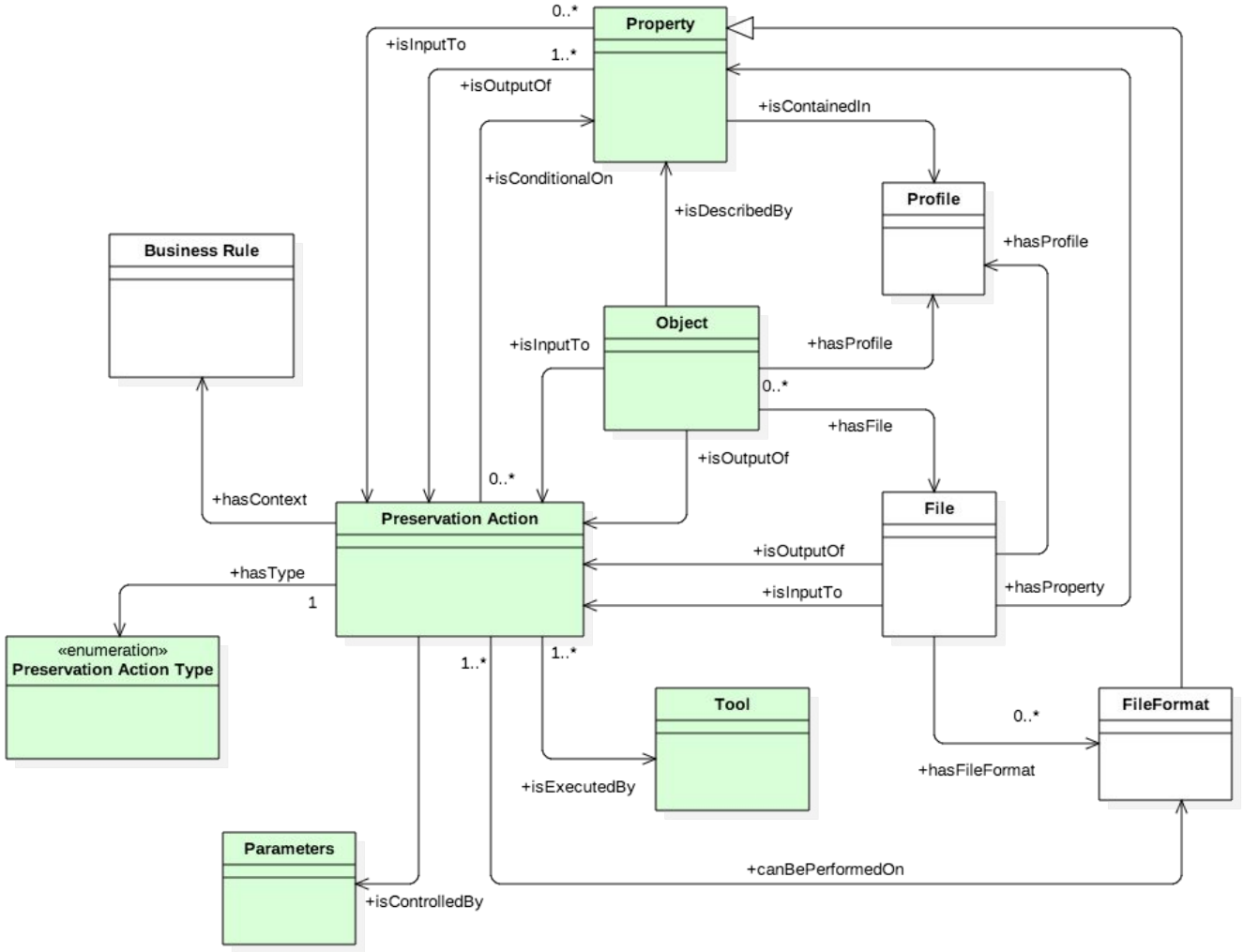
- Discussions of interoperability between the DP solutions
  - No common format policies, hard for users to decide which one to use



The PAR Team

- Justin Simpson - Managing Director; Artefactual
- Sarah Romkey - Archivematica Program Manager; Artefactual
- Matthew Addis - CTO; Arkivum
- Jack O'Sullivan - Senior Software Engineer; Preservica
- Jon Tilbury - CTO; Preservica
- Carl Wilson - Technical Lead; OPF
- Becky Mcguinness - Community Manager; OPF
- Martin Speller - Project Manager; OPF
- Martin Wrigley - Executive Director; OPF
- Paul Stokes - Senior Co-design Manager; JISC

What sort of things does PAR describe?

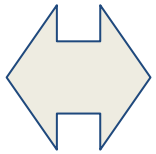


https://doi.org/10.6084/m9.figshare.6628418

Two main concepts

Preservation Actions

Something that needs to be done as part of Digital Preservation



Business Rules

Context for Preservation Actions

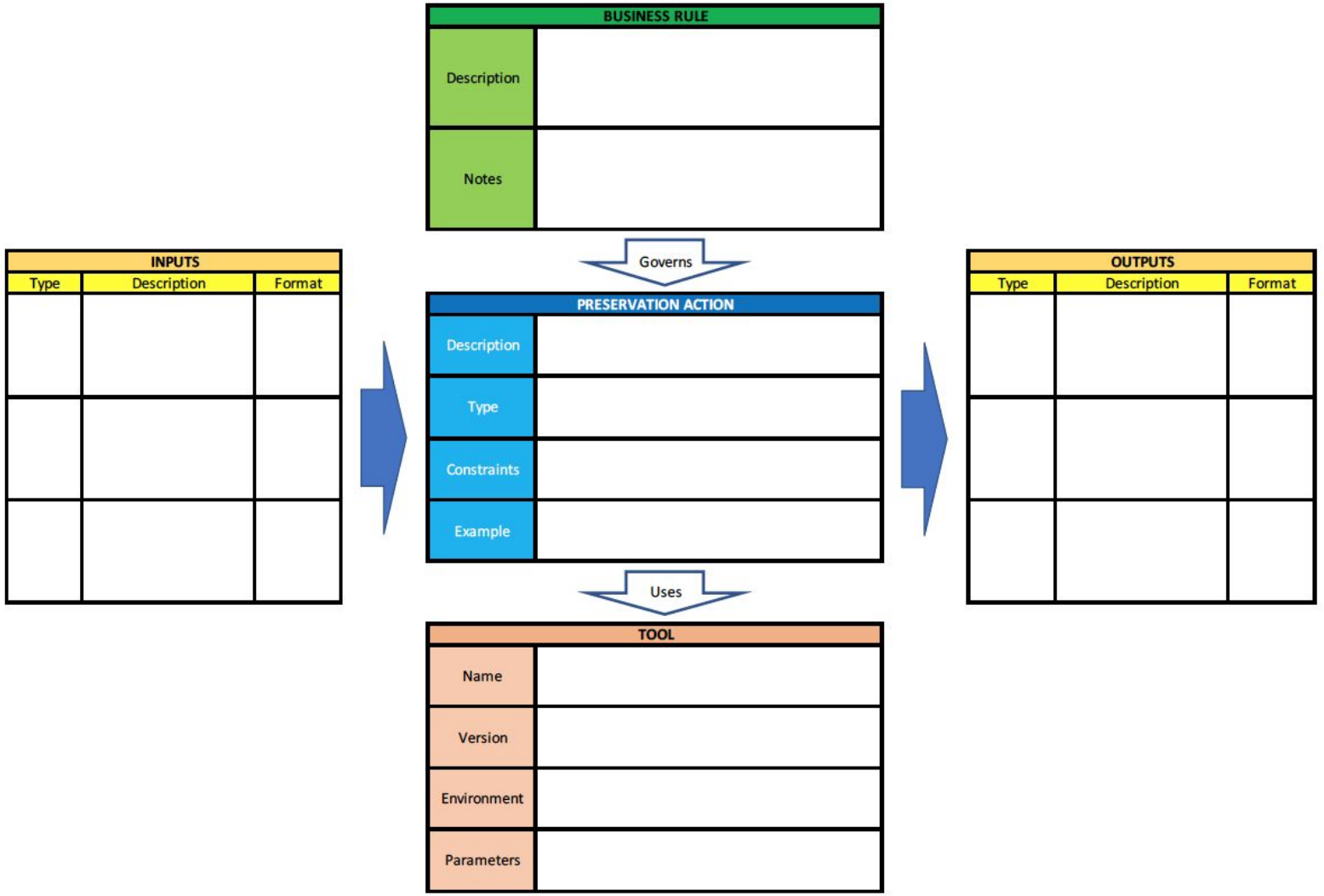
Identification
Property extraction
Validation
Migration
Rendering
Emulation
Packaging
Fixity checking....

What works best in a given scenario
Why do one thing rather than another
What options to chose and when
What happens if something doesn't work
What content types to apply an action to
Alignment with organisational policies
Alignment to community good practices

PAR is a set of core concepts

- A Preservation **Action** is something **done** as part of DP
- Context on when/why/who/how is in **Business Rules**
- 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 them as outputs

PAR template



Some examples of Preservation Actions

- A **Preservation Action** is *one or more steps* that are *executed* as part of performing *digital preservation* on *digital content*

- Some simple examples:
  - Identify file format
  - Create or check checksums
  - Convert a file from format A to format B
  - Extract properties X, Y, Z from a file
  - Validate a file against a format specification
  - Create an Archive Information Package
  - Make replicas and store in different locations

Preservation Action example: property extraction

Description: Use MediaInfo to extract AV properties in EBUCore format

Tool: MediaInfo

Input: AV file

Output: set of AV Properties

Parameter: --Output=EBUCore

```
Video
ID                                       : 1
Format                                   : AVC
Format/Info                              : Advanced Video Codec
Format profile                           : Main@L3
Format settings                          : CABAC / 3 Ref Frames
Format settings, CABAC                   : Yes
Format settings, ReFrames                : 3 frames
Codec ID                                 : avc1
Codec ID/Info                            : Advanced Video Coding
Duration                                 : 13 s 139 ms
Bit rate                                 : 385 kb/s
Width                                    : 854 pixels
Height                                   : 480 pixels
Display aspect ratio                     : 16:9
Frame rate mode                          : Variable
Frame rate                               : 23.976 FPS
Minimum frame rate                       : 23.974 FPS
Maximum frame rate                       : 23.981 FPS
Color space                              : YUV
Chroma subsampling                       : 4:2:0
Bit depth                                : 8 bits
```

mediainfo video.mp4

```
<ebucore:videoFormat videoFormatName="AVC">
  <ebucore:width unit="pixel">854</ebucore:width>
  <ebucore:height unit="pixel">480</ebucore:height>
  <ebucore:frameRate factorNumerator="999" factorDenominator="1000">24</ebucore:frameRate>
  <ebucore:aspectRatio typeLabel="display">
    <ebucore:factorNumerator>16</ebucore:factorNumerator>
    <ebucore:factorDenominator>9</ebucore:factorDenominator>
  </ebucore:aspectRatio>
  <ebucore:videoEncoding typeLabel="Main@L3"/>
  <ebucore:codec>
    <ebucore:codecIdentifier>
      <dc:identifier>avc1</dc:identifier>
    </ebucore:codecIdentifier>
  </ebucore:codec>
  <ebucore:bitRate>385204</ebucore:bitRate>
```

mediainfo --Output=EBUCore video.mp4

Preservation Action example: fixity

Action 1 (digest calculation): use md5sum to generate an MD5 checksum for a file

Action 2 (fixity check): use md5sum can check a file against an MD5 checksum

What is the fixity of parcore.ppt?

md5sum parcore.ppt: 92bc215089ccf35a8384fc25f9be1bd3 parcore.ppt

md5sum -c manifest.md5: parcore.ppt: OK

Same Tool, same File, same Property

A Parameter changes the Action and Action Type (fixity check, digest calc)

A Fixity check is Property extraction followed by Property comparison

Preservation Action example example: file format identification

Identify the file format of a file using Droid, Siegfried, Fido, FITS, File, Tika

What is the file format of parcore.ppt?

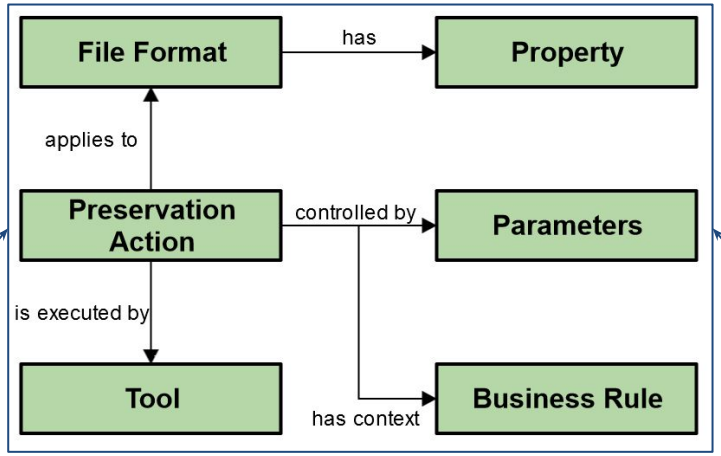
File:	Composite Document File V2 Document
Fido:	fmt/111, “OLE2 Compound Document Format”
Siegfried, DROID:	fmt/126, “Microsoft Powerpoint Presentation”
Tika:	Content-Type: application/vnd.ms-powerpoint

Different schemas: Mimetype, PRONOM ID

Business Rules needed on which Tools work best and when

Business Rules needed on priority/ranking/consensus/conflict resolution

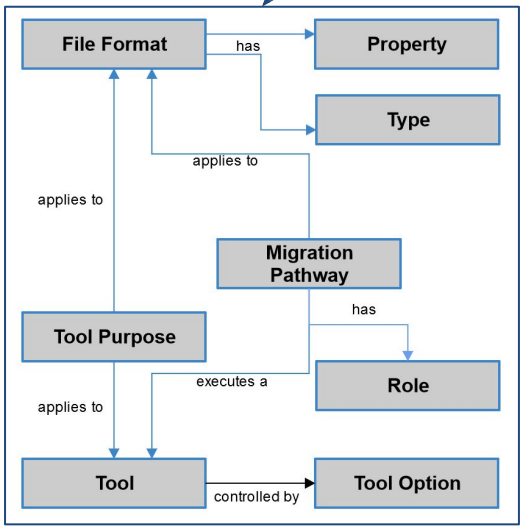
Interoperability between Preservation Systems

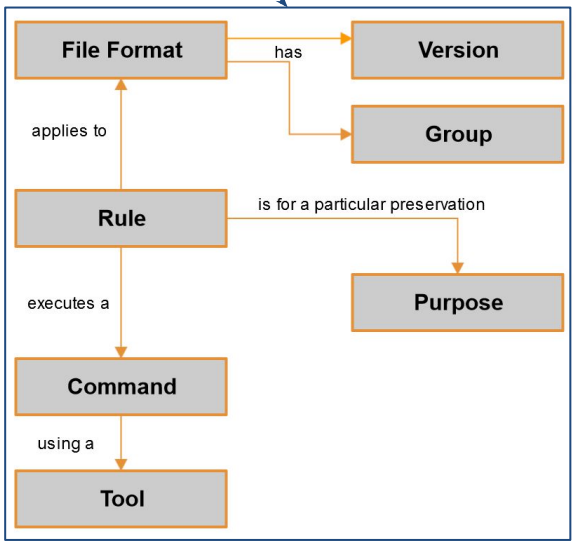


Preservica

PAR

Archivematica





How much of PAR currently exists?

Glossary	• Definition of the PAR core concepts
Conceptual Model	• Common framework for everyone to work to • Something to argue about and agree on! • Interlingua between preservation systems
Json Schemas	• Formal definition of the PAR model • Machine readable, used in API payloads • Used to test and validate interoperability
API	• Common interface for preservation systems • Well defined way to exchange information
Executable DP Actions	• Cross-platform way to deploy/run tools • Unambiguous and vendor independent
Proof of Concept	• Prove PAR is possible! • Not just a talking shop or paper exercise • Reference implementation to share

Plenty more to come!

- Beyond files:
  - Content in wrappers and containers, e.g. zip, MXF, PST
  - Complex objects, e.g. digitised books, websites, research datasets
  - Structured data, e.g. databases, spreadsheets
  - Interactive content, e.g. lab notebooks, games, eBooks, installation art
  - Preserving user experiences, e.g. multiplayer games, social media
  - Preservation workflows, e.g. multiple steps needed for complex objects

- PAR Registries:
  - Tools for building and hosting registries
  - How to reconcile multiple sources of information
  - How to curate, version and manage registry content
  - How to search across multiple registries and preservation systems
  - Deploying and trying out registries for real

We're starting with the simple stuff and working our way up!

Open Preservation Foundation

arkivum Bringing archived data to life

Preservica Active digital preservation

artefactual

Jisc

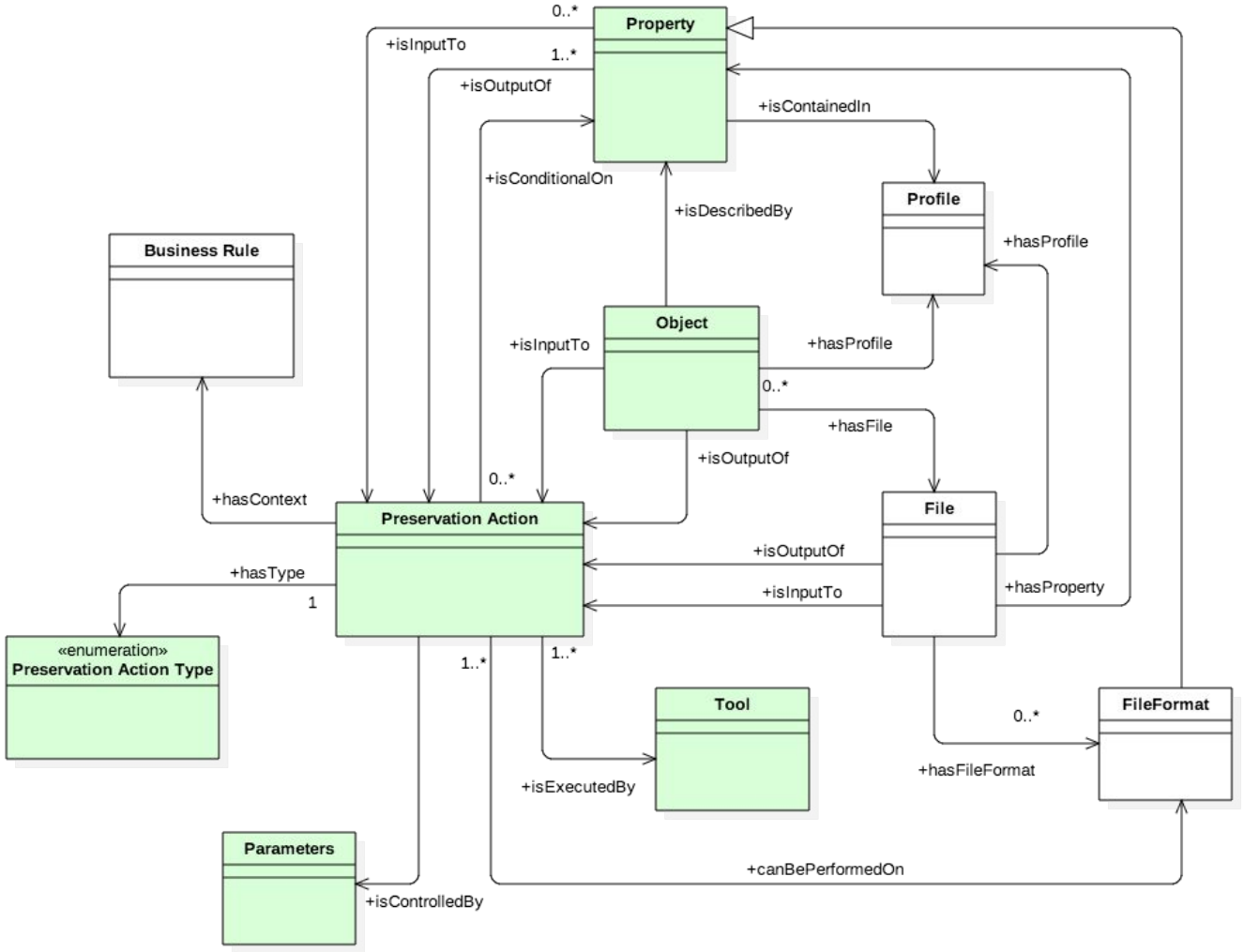
Agenda

Time	Session	Facilitator/Speaker
13:30 - 13:50	PAR Background and overview	Matthew Addis. Arkivum
13:50 - 14:10	**Workshop overview**	**Justin Simpson, Artefactual**
14:10 - 16:10	PAR Action Rules Workshop	All
16:10 - 16:30	Feedback, Q&A	All

PAR Workshop Overview

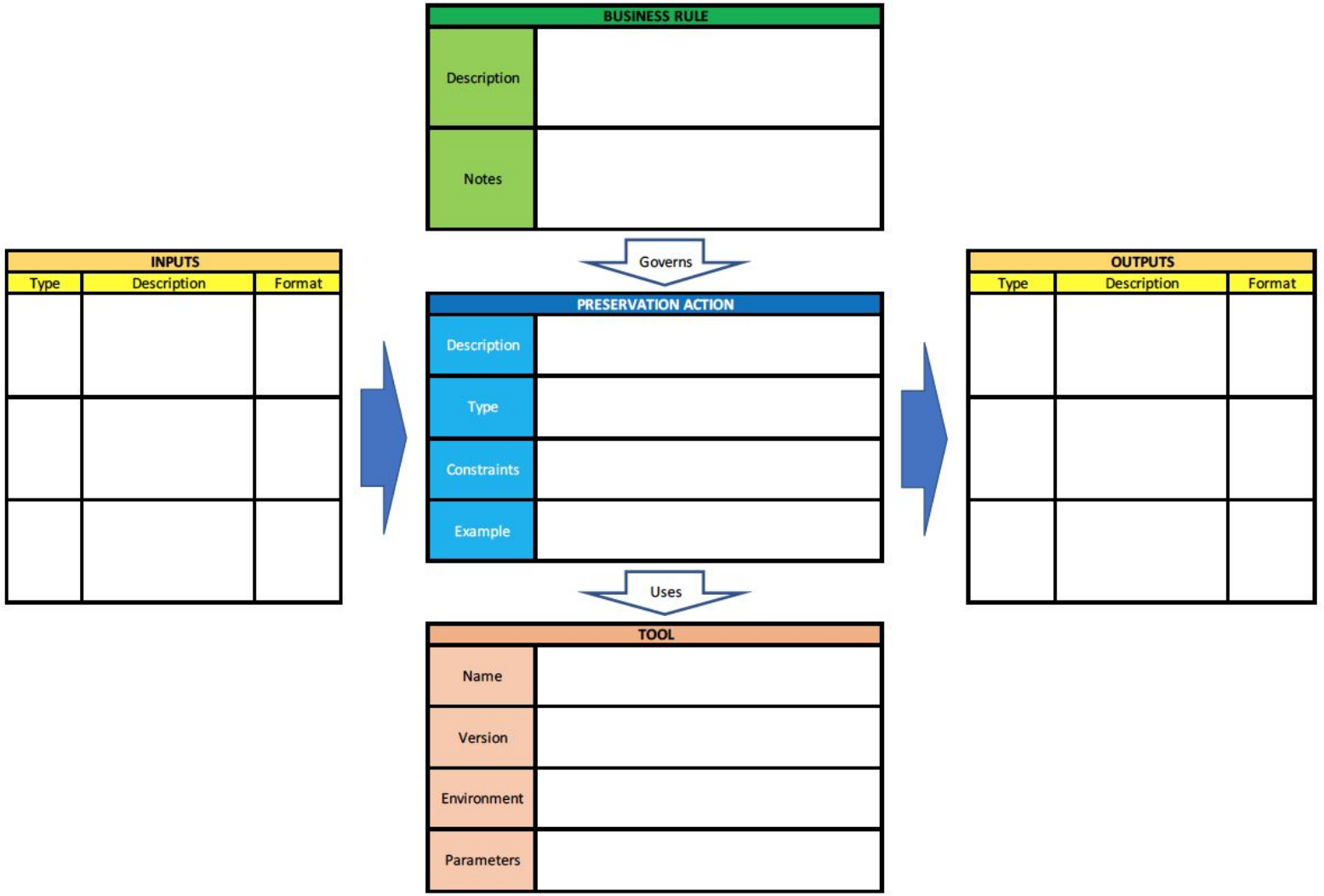
- See one
  - Some examples explained
- Break into groups
- Do one
  - Write down new examples as a group
    - Business Rules
    - Preservation Actions
    - Tools
  - About 1 hour
- Teach one
  - Each Group presents to the room
    - Explain your groups work
    - Listen to the explanations of other groups work
    - About 8 minutes per group

What sort of things does PAR describe?



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

PAR Core


BUSINESS RULE
Description
Notes
Governs
INPUTS
Type
Description
Format
PRESERVATION ACTION
Description
Type
Constraints
Example
OUTPUTS
Type
Description
Format
Uses
TOOL
Name
Version
Environment
Parameters

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.

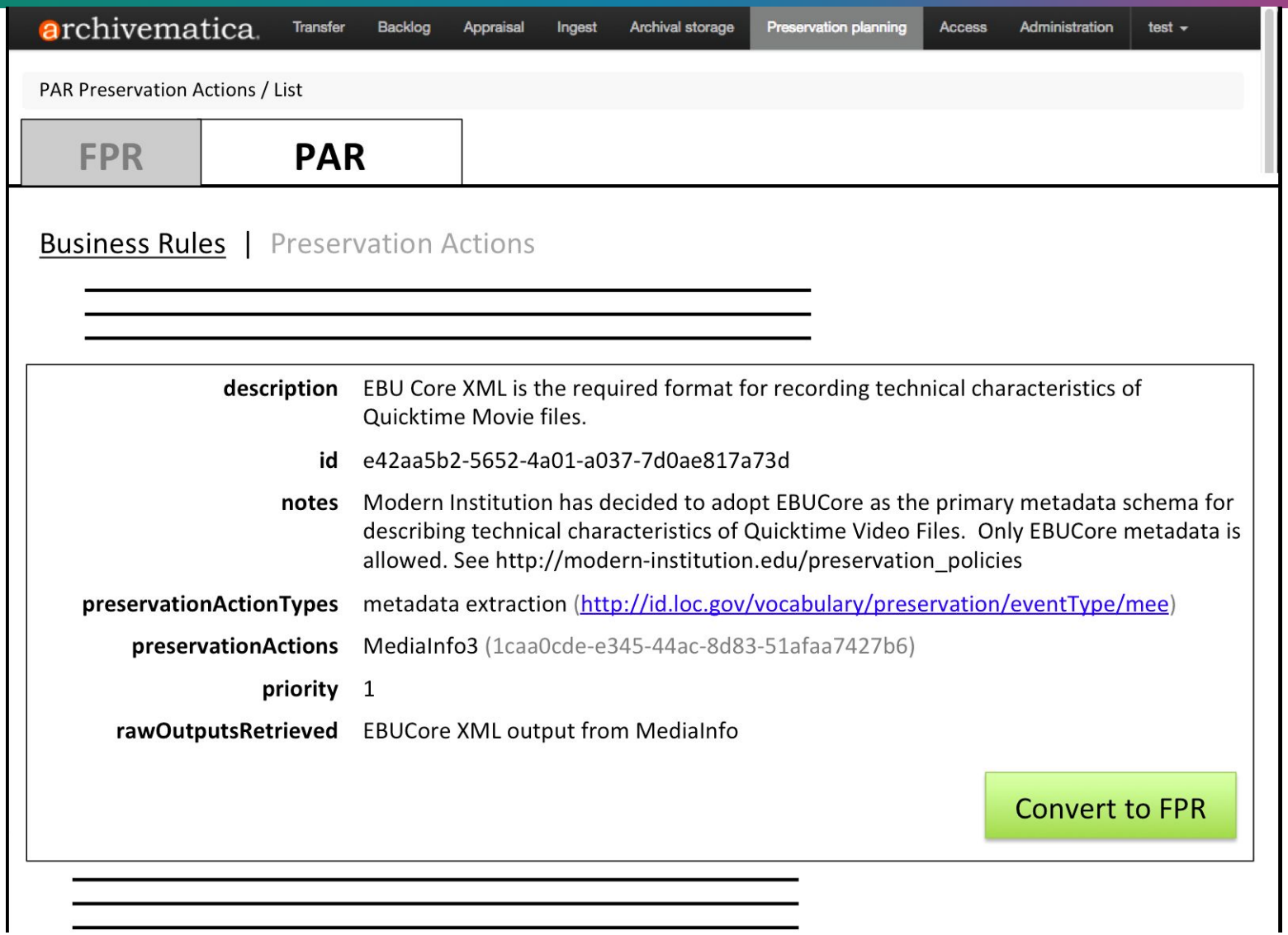
Extract Metadata Use Case

- 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.

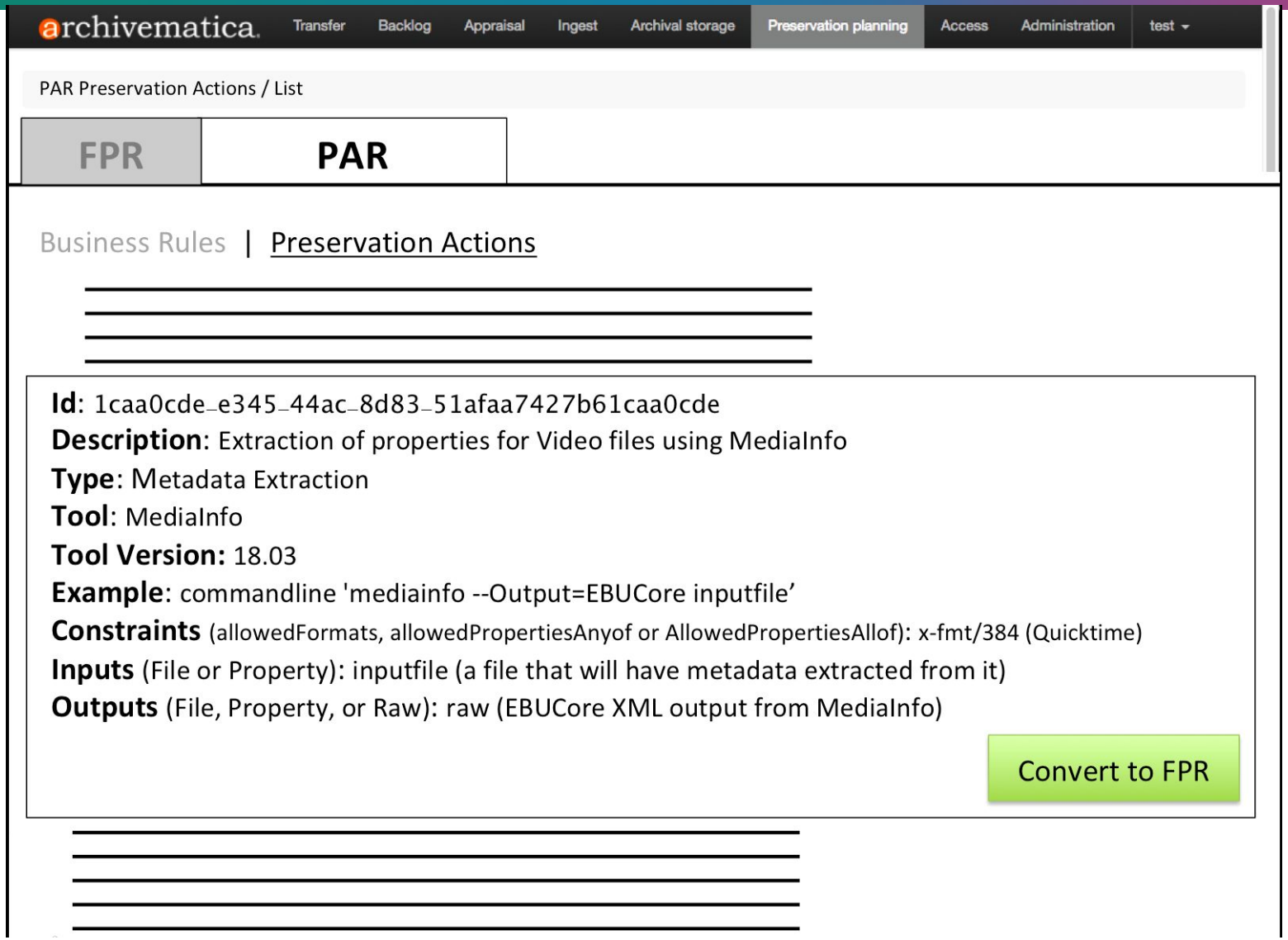
Extract Metadata Use Case

- 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 Archivematica instance.
- At this point the PAR information is entered into the Archivematica database without affecting any existing Commands or Rules.
- The new Rules and Command are linked to the PAR Preservation Action.

PAR Workshop



PAR Workshop



PAR Workshop Overview

Business Rules

- Description
- Notes
- Preservation Action Types
- Formats
- Format Families
- **ID**
- Preservation Actions
  - Priority
  - Inputs
  - Outputs

Preservation Actions

- **Description**
- **Preservation Action Type**
- **Tool**
- Tool Version
- Example
- Constraints
- Inputs
- Outputs
- **ID**

PAR Workshop Overview

Business Rules

- Description
- Notes
- Preservation Action Types
- Formats
- Format Families
- **ID**
- Preservation Actions
  - Priority
  - Inputs
  - Outputs

Business Rules explain why particular Preservation Actions are performed in particular contexts

Mostly narrative, but they provide a way to

a) explicitly link Preservation Actions to policies
b) Prioritize Preservation Actions
c) Specify required inputs
d) Specify expected outputs

PAR Workshop Overview

Business Rules Details

- Description
  - *A short human readable explanation/display name for the Business Rule*
- Notes
  - *A free text field for providing additional context. This may be used to record the decision making process that led to the formulation of this rule, details of real-world experience in applying the rule, or any other text.*
- Preservation Action Types
  - *One or more LoC Preservation Event Types*
  - *A list of Preservation Action Types that this Business Rule should be applied to. This might be a subset of those that the Preservation Actions themselves apply to.*
- Format Families
  - *A list of format families that this Business Rule should be applied to*
- Formats
  - *A list of file formats that this Business Rule should be applied to*

PAR Workshop Overview

Business Rules Details

- Preservation Actions
  - *A list of specific actions. These are defined internally to reference the Core Preservation Action, the priority order in which it should be performed, and any inputs and outputs that should be used.*

Actions

- Preservation action
  - Optional input properties
  - Output files retrieved
  - Output properties retrieved
  - Raw outputs retrieved
- Priority

PAR Workshop Overview

Preservation Actions

- **Description**
- **Preservation Action Type**
- **Tool**
- Tool Version
- Example
- Constraints
- Inputs
- Outputs
- **ID**

Preservation Actions are processes that are run as part of performing digital preservation.

- Classified by Preservation Action Type
- Executed by one or more Tools
- Constraints define when to use this action
- Inputs define files or properties required to execute the action
- Outputs define objects or properties created by the action

PAR Workshop Overview

Preservation Action Details

- Description
  - *A short human readable explanation/display name*
- Type
  - *One LoC Preservation Event Type*
  - *A Preservation Action Type that this Preservation Action performs.*
- Tool
  - *A PAR Tool*
- Tool Version
  - *The specific version of the tool this Action requires*
- Example
  - *A human readable explanation of how to execute the Preservation Action.*

PAR Workshop Overview

Preservation Action Details

- Constraints
  - *Defines limitations of when to perform this Action*
    - Limit to files of specific formats, or
    - Limit to objects with specific properties
  - AllowedFormats, AllowedPropertiesAnyOf, AllowedPropertiesAllOf
- Inputs
  - *A list of expected inputs (files or properties)*
  - *File, Property*
- *Outputs*
  - *A list of outputs created by this Action*
  - File, Property, Raw

Business Rule example 1 - Choice of file format ID tools

- Description: `For File Format Identification of any type of file, use DROID (PAR-tool/1) as the first preference, and FIDO (PAR-tool/2) as the second`
- Notes: `My order of preference/priority when using multiple tools are available for file format identification is DROID then FIDO`
- Preservation Action Types: `format identification`
- Formats:
- Format Families: `ALL`
- **ID**: `a593036d-7427-54aa-b8c0-50a5fb7bd50b`
- Preservation Actions
  - action/1
    - Priority: 1
    - Inputs
    - Outputs
  - action/2
    - Priority: 2
    - Inputs
    - Outputs

https://github.com/artefactual-labs/rdss-par/blob/am-characterize-1/examples/br-1.json

Business Rule example 2 - Migrating AVI files

- Description: `FFMPEG is preferable to HandBrake for migrating AVI files to WebM`
- Notes: `We have found that the HandBrake process occasionally hangs when performing this type of migration. FFMPEG displays no similar behaviour and therefore should be used in preference`
- Preservation Action Types: `migration; normalization`
- Formats: `fmt/5`
- Format Families: `AVI`
- **ID**: `bc6c7498-2b8d-5314-bb0d-c773e72e9ff2`
- Preservation Actions
  - action/123
    - Priority: 1
    - Inputs
    - Outputs
  - action/456
    - Priority: 2
    - Inputs
    - Outputs

https://github.com/artefactual-labs/rdss-par/blob/am-characterize-1/examples/br-2.json

Business Rule example 3 - Resizing for access

- Description: `Use resizing parameters on ImageMagick to perform generation of consistent size access copy JPEGs from TIFFs`
- Notes: `We want all JPEGs for presentation to fit into a 250x250px box so that they display properly in our access system, where the template for displaying images has a 300px div for the image itself, however, we only want to resize where the image itself is larger than that box.`
- Preservation Action Types: `migration`
- Formats: `fmt/353; fmt/155; fmt/154 [etc]`
- Format Families: `TIFF`
- **ID**: `7a71b0d2-d6f3-5d49-a132-9b6776ec6243`
- Preservation Actions
    - action/25
        - Priority: `1`
        - Inputs: `Resize to fit in 250px square box [etc]`
        - Outputs:

https://github.com/artefactual-labs/rdss-par/blob/am-characterize-1/examples/br-3.json

Business Rule example 4 - Interactive Website Capture

- Description: When capturing websites with interactive content or a small amount of content, use a manual capture process.
- Notes: In cases where an automated capture process is not able to record all of the significant properties of the website, for example with very interactive content such as user triggered content, a manual capture process is preferable.
- Preservation Action Types: capture
- Formats:
- Format Families:
- **ID**: 7a71b0d2-d6f3-5d49-a132-9b6776ec6243
- Preservation Actions
    - action/42
        - Priority: 1
        - Inputs: a website with interactive content
        - Outputs: warc file and documentation of the operators capture method.

Preservation Action Example 1: Characterize with MediaInfo

- **Description** `Extraction of properties for Video files using MediaInfo`
- **Preservation Action Type** `metadata extraction`
- **Tool** `mediainfo`
- Tool Version `18.03`
- Example `commandline 'mediainfo --Output=EBUCore inputfile`
- Constraints `allowedFormats`
- Inputs `inputfile`
- Outputs `EBUCore XML output from MediaInfo`
- **ID** `1caa0cde-e345-44ac-8d83-51afaa7427b6`

https://github.com/artefactual-labs/rdss-par/blob/am-characterize-1/examples/preservationAction/mediainfo3.json

Preservation Action Example 2: Checksum Validation with md5sum

- **Description** `Validation of an MD5 checksum on a File using md5sum`
- **Preservation Action Type** `fixity check`
- **Tool** `md5sum`
- Tool Version
- Example `commandline 'md5sum -c manifest.md5', where manifest.md5 contains a MD5 checksum for a file called inputfile`
- Constraints
- Inputs `Manifest file containing the MD5 and name of the file to be checked`
- Outputs `Fixity PASS or FAIL`
- **ID** `f2e953e4-425e-5ed1-a65e-efd0b2e061be`

https://github.com/artefactual-labs/rdss-par/blob/am-characterize-1/examples/md5check1.json

Agenda

Time	Session	Facilitator/Speaker
13:30 - 13:50	PAR Background and overview	Matthew Addis. Arkivum
13:50 - 14:10	Workshop overview	Justin Simpson, Artefactual
14:10 - 16:10	**PAR Action Rules Workshop**	**All**
16:10 - 16:30	Feedback, Q&A	All

Agenda

Time	Session	Facilitator/Speaker
13:30 - 13:50	PAR Background and overview	Matthew Addis. Arkivum
13:50 - 14:10	Workshop overview	Justin Simpson, Artefactual
14:10 - 16:10	PAR Action Rules Workshop	All
16:10 - 16:30	**Feedback, Q&A**	**All**

- Q &A
- Now please complete the event feedback form - this has just been emailed you

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)