CSC

ICT Solutions for Brilliant Minds

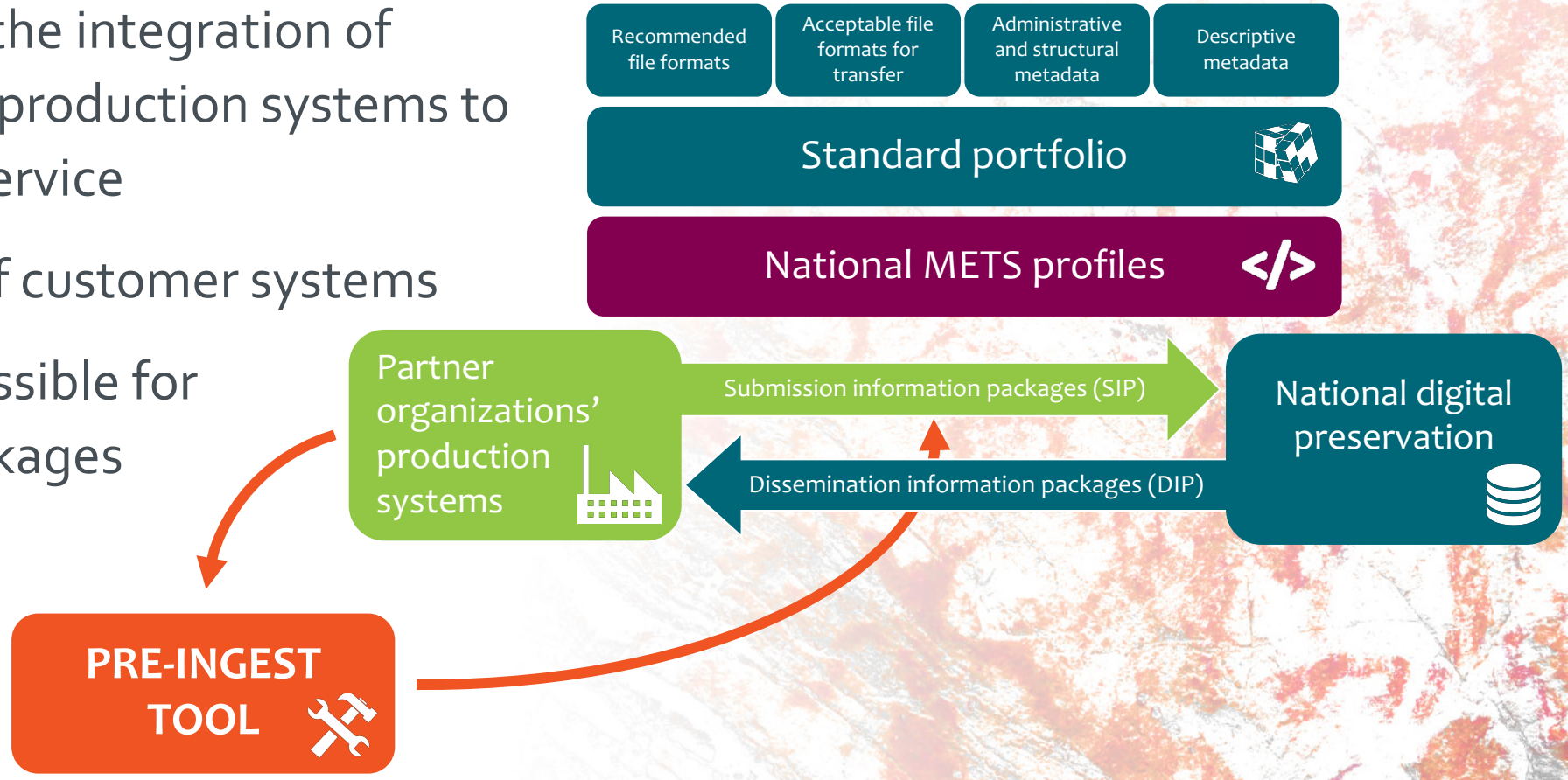
Pre-Ingest Tool

Juha Lehtonen

16.09.2019 as part of the METS Tutorial at iPRES 2019

Pre-Ingest Tool Context

- A tool to create SIPs for national digital preservation
- Conforms to our national preservation specifications
- The aim is to simplify the integration of partner organizations production systems to Digital Preservation Service
- Challenge: Diversity of customer systems
- Also manual usage possible for small information packages



Pre-ingest Tool Architecture

- Flexible and modular, customizable for various needs
  - METS document creation (with PREMIS, MIX, AudioMD, VideoMD…) is the major function of the tool
  - We have divided the packaging into several pieces, into scripts
  - We keep the scripts as independent as possible, the information is given as arguments
  - One can also use some pieces and implement the other pieces other way
  - Supplementary functionalities can be implemented in the needed pieces
- Available at GitHub
  - https://github.com/Digital-Preservation-Finland
  - License: LGPLv3
  - Does not fully cover all possible use cases, but works in many cases
  - Software development is an ongoing process, new versions will be published

Pre-ingest Tool Components (1/2)

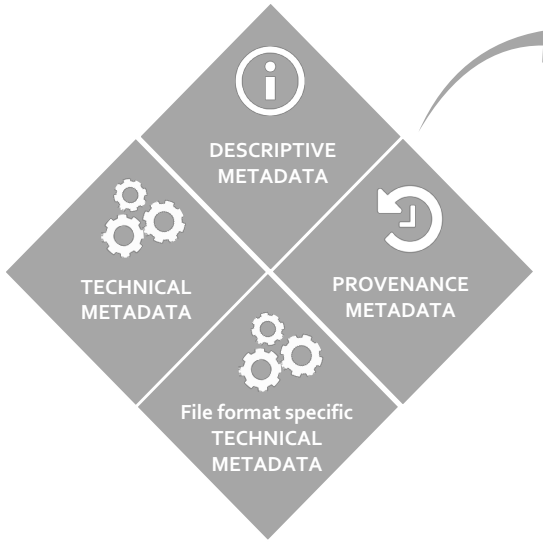
Technical metadata
Basic technical metadata, optionally file format validation

File format specific technical metadata
For images, audio, video, structured text (CSV)

Descriptive metadata
Import already existing descriptive metadata

Provenance information
Create events and agents to describe the provenance information

Pre-ingest Tool Components (2/2)



File section and structural map
File list, structure, and linkings between metadata and files

METS document
Compile METS document based on the previous steps

Sign and package SIP
Sign and package as SIP to be submitted to the DP service

Demonstrations

- See repository *siptools-workshop-2019* form: https://github.com/Digital-Preservation-Finland
    - This includes five different tutorial exercises for our Pre-Ingest Tool
    - Originally used in a national pre-ingest tool workshop on April 2019

Additionally: Metadata Libraries for Python

- We have implemented various general metadata libraries for Python
  - METS, PREMIS, AudioMD, VideoMD, ADDML…
  - Do not fully cover all the properties of the metadata formats, development ongoing…
  - https://github.com/Digital-Preservation-Finland
- Example: METS Library usage with Python

```
> import mets
> import premis
> object = premis.object(premis.identifier(
>     'my-id-type', 'my-id-value'), 'file.txt')
> xmldata = mets.xmldata(child_elements=[object])
> mdwrap = mets.mdwrap('PREMIS:OBJECT', '2.2', child_elements=[xmldata])
> techmd = mets.techmd('my-id', child_elements=[mdwrap])
> amdsec = mets.amdsec([techmd])
> root = mets.mets(child_elements=[amdsec])
```

Metadata Libraries for Python

```
> lxml.etree.dump(root)
```

```
<mets:mets xmlns:mets="http://www.loc.gov/METS/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xmlns:xlink="http://www.w3.org/1999/xlink" xsi:schemaLocation="http://www.loc.gov/METS/
http://www.loc.gov/standards/mets/mets.xsd" PROFILE="local" OBJID="899c7e80-40f4-464a-9c7f-
9f6f91e0fec7">
  <mets:amdSec>
    <mets:techMD ID="my-id" CREATED="2019-07-15T07:30:26.859016">
      <mets:mdWrap MDTYPE="PREMIS:OBJECT" MDTYPEVERSION="2.2">
        <mets:xmlData>
          <premis:object xmlns:premis="info:lc/xmlns/premis-v2" xsi:type="premis:file">
            <premis:objectIdentifier>
              <premis:objectIdentifierType>my-id-type</premis:objectIdentifierType>
              <premis:objectIdentifierValue>my-id-value</premis:objectIdentifierValue>
            </premis:objectIdentifier>
            <premis:originalName>file.txt</premis:originalName>
          </premis:object>
        </mets:xmlData>
      </mets:mdWrap>
    </mets:techMD>
  </mets:amdSec>
</mets:mets>
```

Thank you!

http://www.loc.gov/standards/mets/
http://digitalpreservation.fi/
Twitter @dpres_fi
https://github.com/Digital-Preservation-Finland/