CSC

ICT Solutions for Brilliant Minds

Workflows in National Digital Preservation Services of Finland

2024-02-05 DPC Workflow Webinar Series

CSC – IT Center for Science ltd.

CSC Provides National Preservation Services In Finland

- CSC produces centralized digital preservation services for the Ministry of Education and Culture
  - Digital Preservation Service for Cultural Heritage (in production since 2015)
  - Digital Preservation Service for Research Data (in production since 2019)
- The ownership of digital assets remains with the organizations which preserved them
- Digital Preservation Services utilize CSC's ISO/IEC 27001:2013 certified ISMS
- International cooperation including OPF, DPC, METS editorial board, several EU projects and other preservation organizations

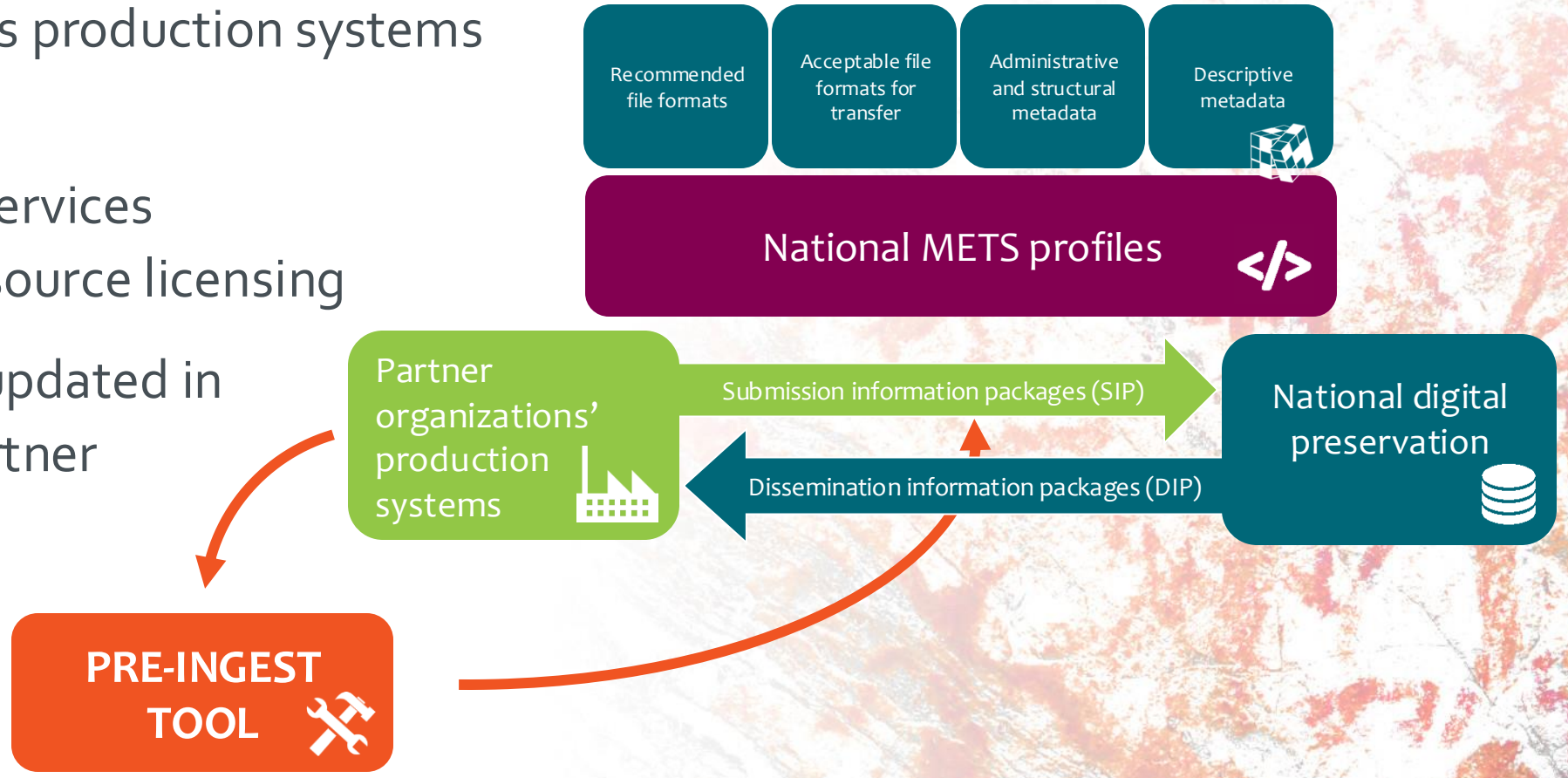
Currently we preserve 3.3PiB in over 5.5 million archival information packages

Digital assets are stored on three different media types (disk and two tapes)

… and a dark archive…

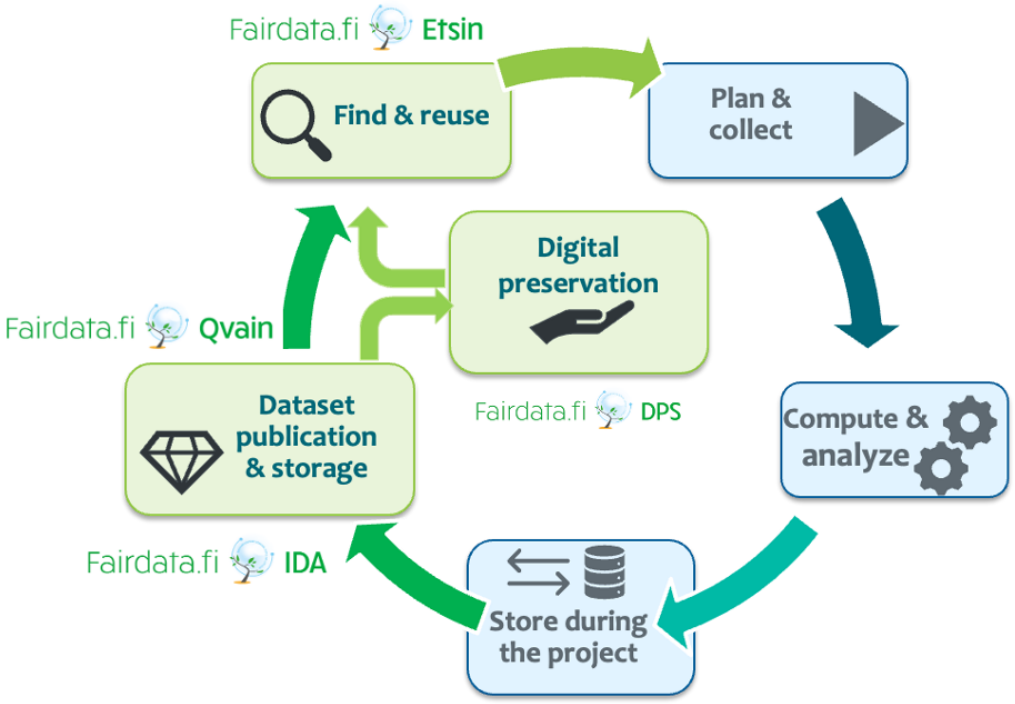
Common Standards for Preserved Data

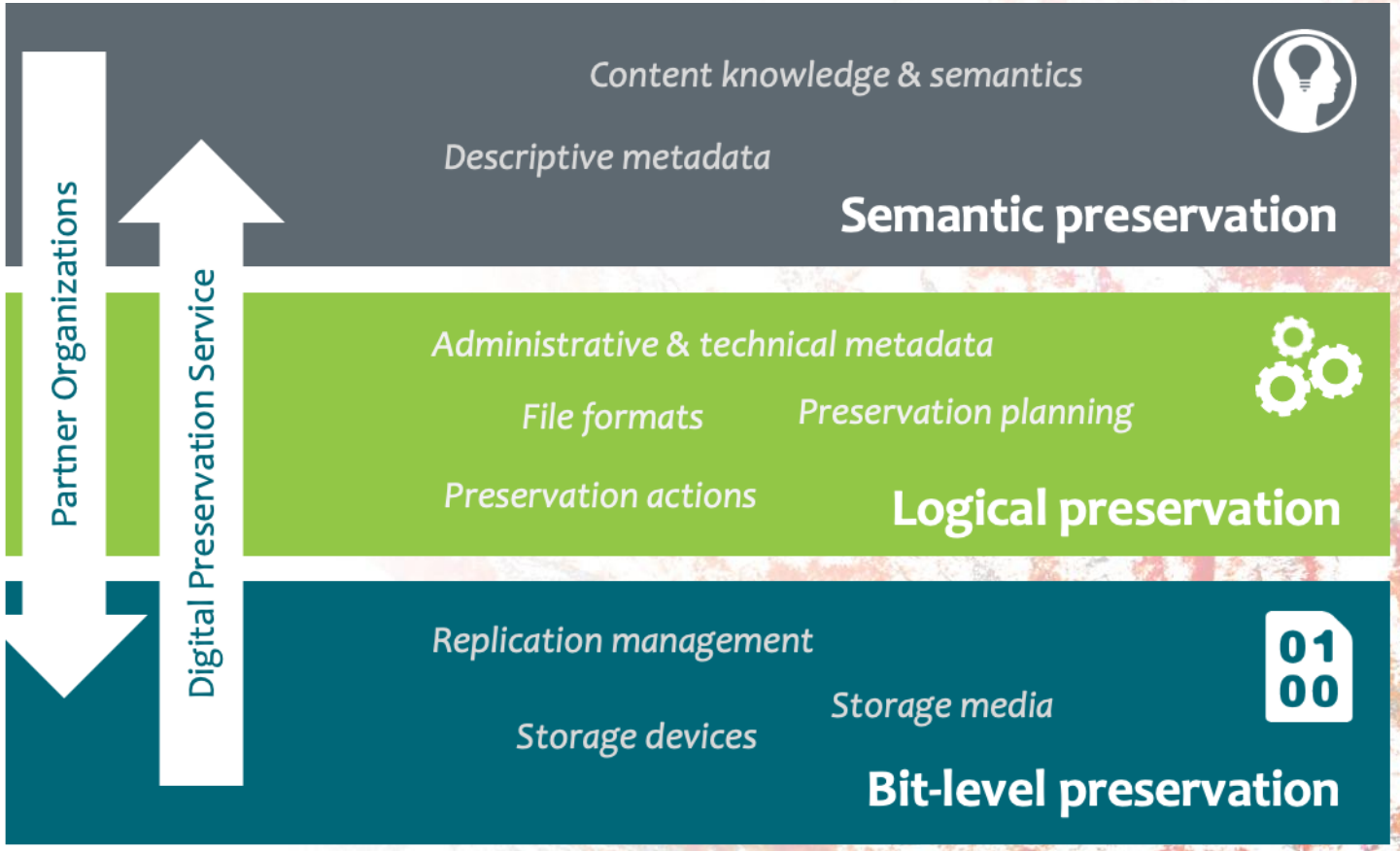
- All preserved data follows national standard portfolio / specifications
- The aim is to simplify the integration of partner organizations production systems to the DPS
- Standard tools and services provided with open source licensing
- Standards annually updated in cooperation with partner organizations



CSC provides national services supporting FAIR principles

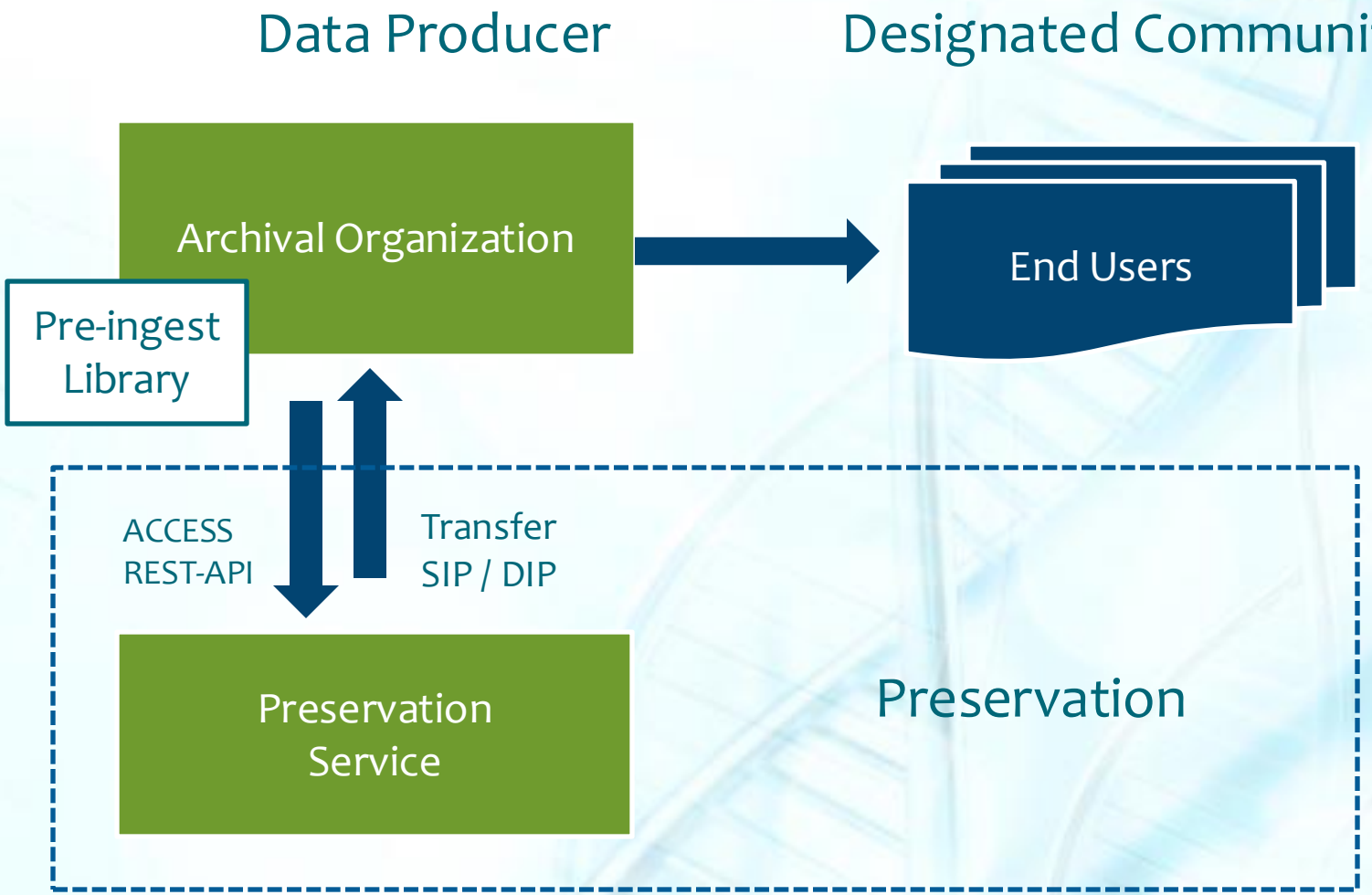
Collaboration enables long-term usability of content



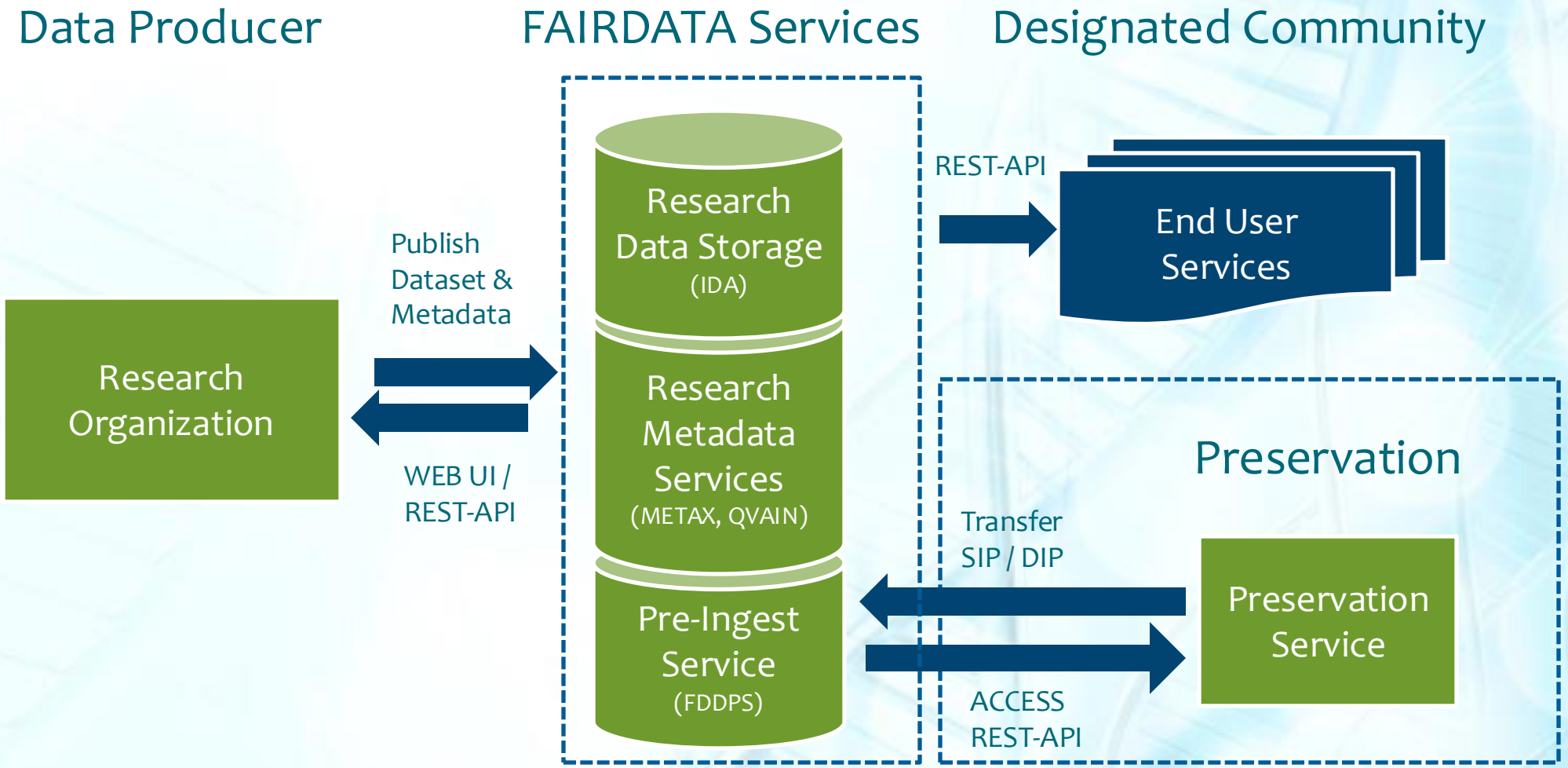


https://fairdata.fi/
https://digitalpreservation.fi/

Preservation Services 1/2 – Culture and Archival Organizations



Preservation Services 2/2 – Research Organizations



Pre-ingest Workflow – Automating Packaging and Metadata 1/2

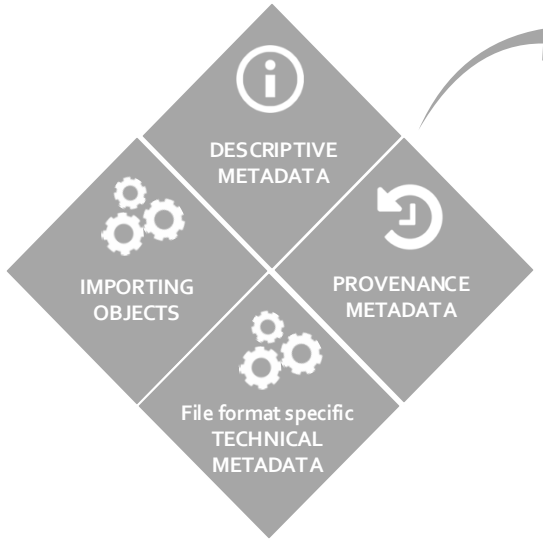
Importing objects
Basic technical metadata, optionally file format validation

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

Descriptive metadata
Import existing descriptive metadata

Provenance information
Create events and agents to describe the provenance information

Pre-ingest Workflow – Automating Packaging and Metadata 2/2



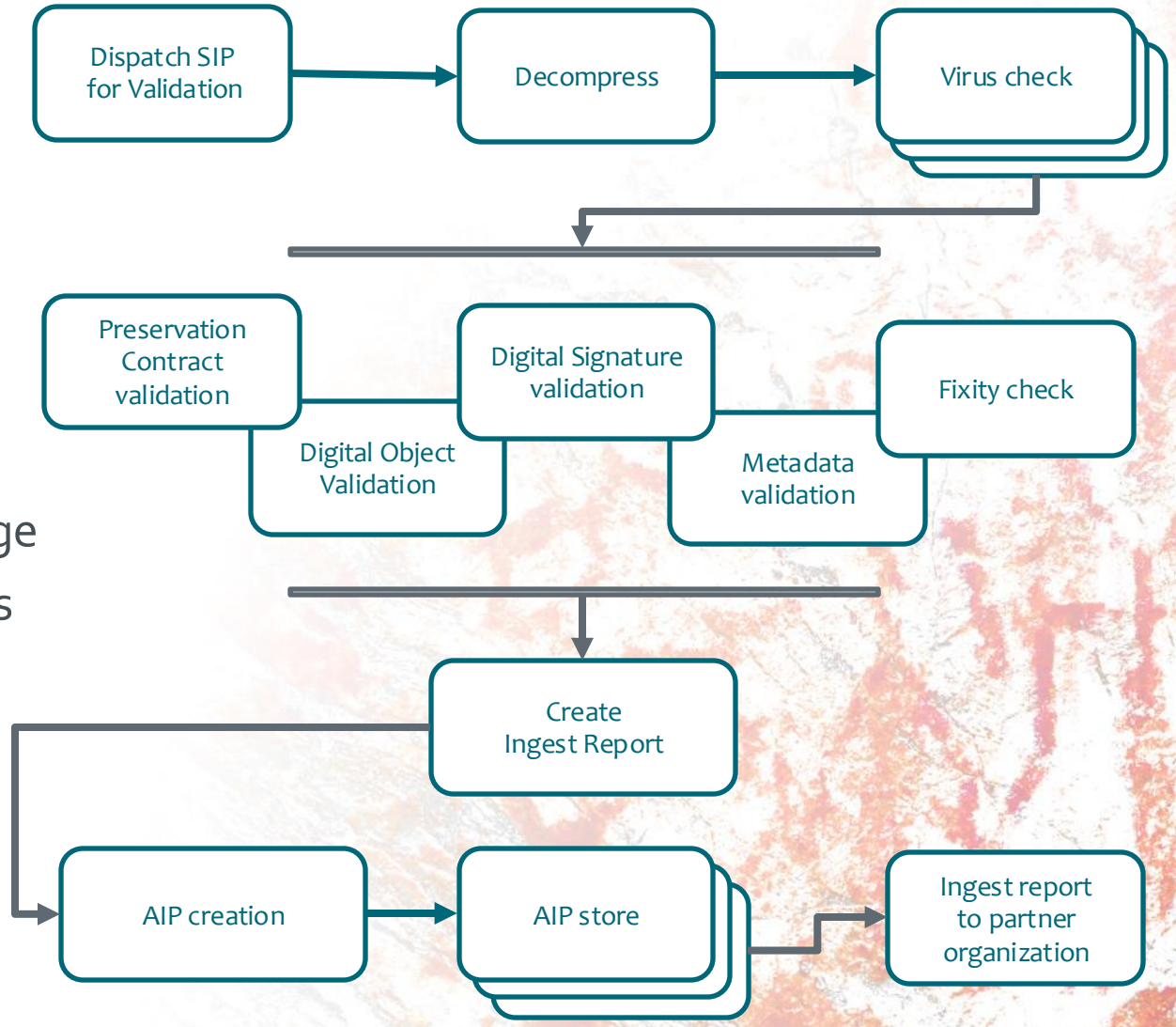
File section and structural map
File list, structure, and links between metadata and files

METS document
Compile a METS document based on the previous steps

Sign and package SIP
Sign and package metadata + contents to a SIP (to be submitted to the DPS)

Ingest Workflow – Automating Validation and Preservation

- Our ingest process is fully automated
  - Individual task for each workflow step
  - Both serial and parallel processing
  - Machine readable metadata
- The ingest process:
  - SIP validation – standards conformance
  - AIP generation – unified data blobs for storage
  - AIP storage – Distributing copies over medias
  - Reporting – Returning reports to users
- Same principles for all workflows
  - Tape storage, media refreshments
  - Data Dissemination
  - File format migrations, etc.



Ingest Workflow – Tasks Explained 1/5

Preparing for Ingest

- Dispatch
  - Enqueue uploaded SIP for distributing processing over computing cluster
  - Makes SIP available for **worker process** which processes SIP through the **ingest workflow**
  - Single validation server processes the ingest workflow
- Decompress
  - Decompress uploaded ZIP / TAR archive file to workspace directory
- Virus Check
  - Ensure uploaded files does not contain any potentially harmful code
  - After detected virus the SIP is quarantined from processing for further inspection

Ingest Workflow – Tasks Explained 2/5

Validation Steps

- Digital Signature Validation
  - Ensure provided digital S/MIME signature in the manifest file is valid
  - Ensure METS XML checksum matches checksum in the manifest file
  - Implemented using common manifest file formatsigned with S/MIME signature
- Preservation contract validation
  - Ensures METS file contains valid contract ID and the organization has quota for preservation
  - Up-to-date disk usage and rights are maintained in contracts database
- Fixity Check
  - Ensure checksum of each digital object matches provided metadata in METS XML file
  - Standard MD5 / SHA-1 / SHA-245 signatures may be used

Ingest Workflow – Tasks Explained 3/5

- Metadata Validation
  - Ensure METS XML complies with national METS XML specifications / profile
  - XML Schema validation
    - Technical XML format and structure checks, required metadata fields, allowed embedded XML formats etc.
  - Schematron rules validation
    - Additional checks ie. Internal linking, object counts etc.
- Digital Object Validation
  - Ensure each digital object is well-formed
  - Ensure technical metadata of each digital object matched METS XML / techmd records
  - Implemented using open sourced **file-scraper** library and multiple 3rd party tools / libraries
  - Usual tools for validation: Jhove, VeraPDF, LibreOffice, Ghostscript, ffmpeg, libxml2, etc.

Ingest Workflow – Tasks Explained 4/5

Storing the Data

- Create AIP
  - Package all ingested files and report into single TAR archive
  - Contains manifest in Bagit format
  - Self-sufficient single package for bit preserving the SIP data, everything included
- AIP Store
  - Copy AIP to three storage systems, geographically distributed
  - One copy to GlusterFS volume using local POSIX filesystem mounts
  - Two IBM LTFS tape copies to remote **storage servers** using storage REST API
  - Storage server use separate workflow for writing AIPs on the tape media

Ingest Workflow – Tasks Explained 5/5

Finalization and Reporting

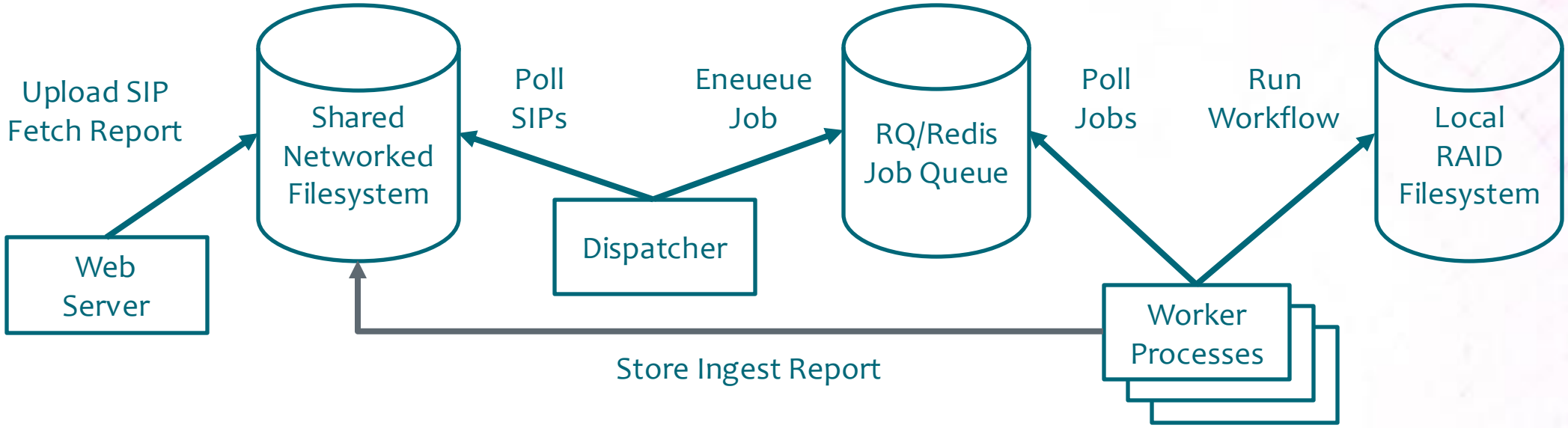
- Creating Ingest Report
  - Aggregated report of ingestion / validation events
  - Events stored in PREMIS XML format and vocabulary
  - NOTE: Ingest report is generated before AIP and included in AIP archival file
- Ingest Report to Partner Organization
  - After successful ingest, report is stored on shared GlusterFS volume for fast retrieval
  - Partner organization uses access REST API for polling and downloading the report

Workflow Engine – Luigi Workflows

- Each workflow and tasks are implemented as pure Python code
  - Workflows visualized as dependency graphs
  - Relatively easy to understand and test complex workflows
- Each workflow and tasks are Atomic and Idempotent
  - With given input produces always the same result
  - Previous outputs are overwritten using deterministic filenames, identifiers etc.
  - On success task writes output - on error does not write output (atomic state)
- Workflow/tasks are repeated until task succeeds (with graceful retry intervals)

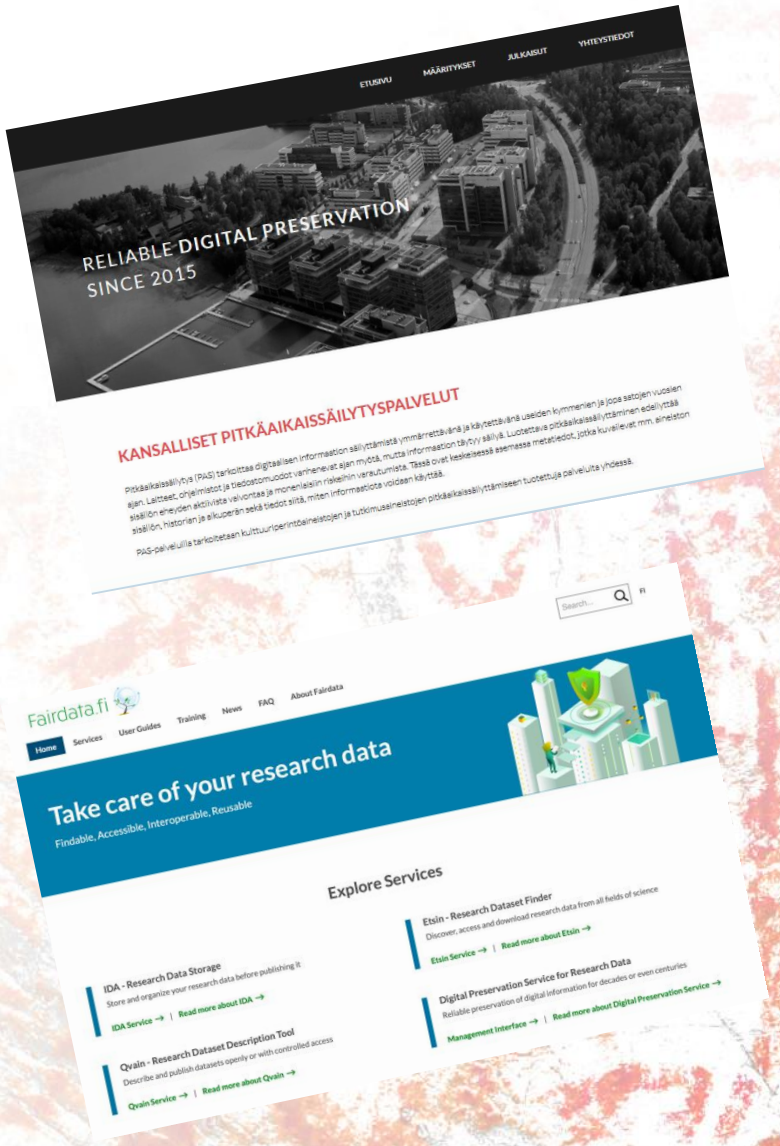
Workflow Engine - Distributed RQ Workers

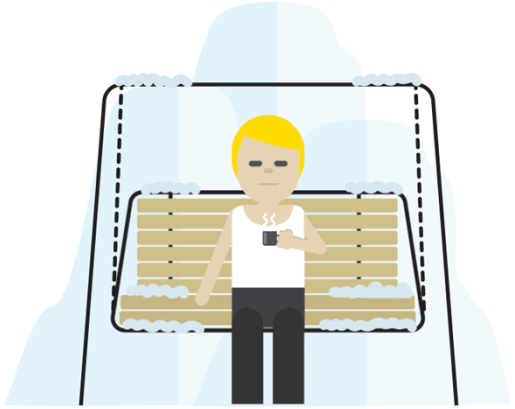
- **Dispatcher** (RQ) polls filesystem for new SIPs and enqueues job to Redis **job queue**
- Each job is processed by RQ **worker** which runs a complete Luigi **workflow**
- Worker processes utilize CPU and RAID volume on dedicated validation servers



More Information Available Online

- Digital Preservation Services
  - API specifications, METS profiles, file formats, standards, etc.
  - Publications also available in English
  - http://digitalpreservation.fi/en/
- FAIRDATA Services for Research Data
  - Documentation and services for managing research data
  - https://www.fairdata.fi/en/
- Python libraries, schemas & schematron rules available at GitHub
  - https://github.com/Digital-Preservation-Finland/





Mikko Vatanen

Digital Preservation Services
CSC – IT Center for Science Ltd

Contact the team at **pas-support@csc.fi**

facebook.com/CSCfi

twitter.com/CSCfi

youtube.com/CSCfi

linkedin.com/company/csc---it-center-for-science

github.com/CSCfi