

# Curation in the pre-ingest phase @ Meertens Institute now & near future

---

Menzo Windhouwer

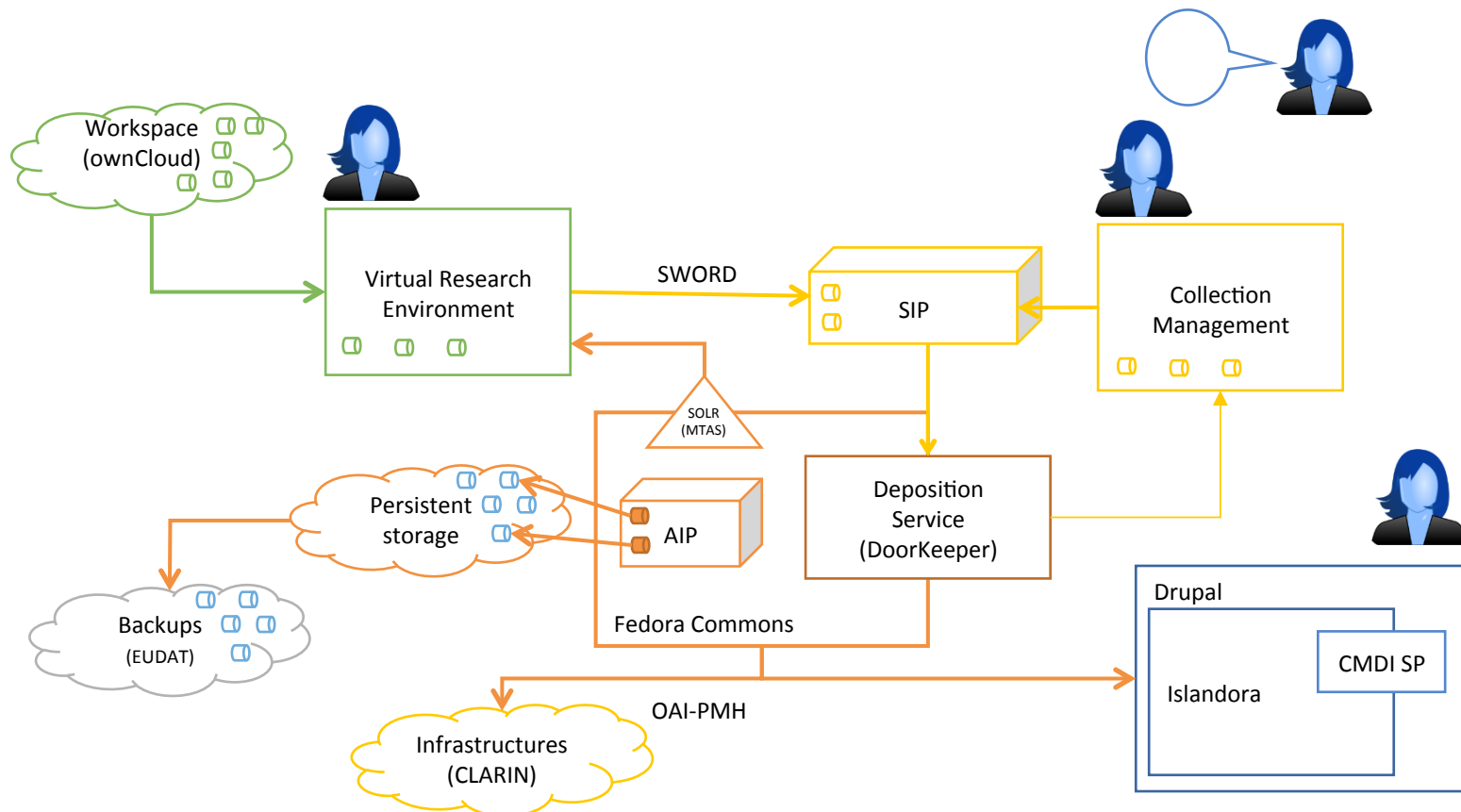
[menzo.windhouwer@meertens.knaw.nl](mailto:menzo.windhouwer@meertens.knaw.nl)

Meertens Institute, TLA

Technische Ontwikkeling



# (New) Workflow at the Meertens Institute



# Pre-ingest phase at the Meertens Institute

---

1. Do the resources fit in the collection profile?
  - a. Match MI mission statement (Dutch language and culture) or a researcher
  - b. How unique are the resources? Do we have this (type) already?
2. Check the completeness of the metadata
3. Check the resources
  - a. ...

# Pre-ingest phase at the Meertens Institute

---

## 3. Check the resources

### a. [Preferred formats](#)

1. No, **curation** by the depositor or, if needed, the MI

### b. Size

### c. Rights

1. Does the depositor own the rights?
2. Are there legal restrictions?
  - IPR
  - Personal data

### d. License

1. Donation (= MI becomes owner) or deposit?
  - An external depositor always needs to sign a donation or [license-agreement](#)
2. The use of CC-0 is actively encouraged

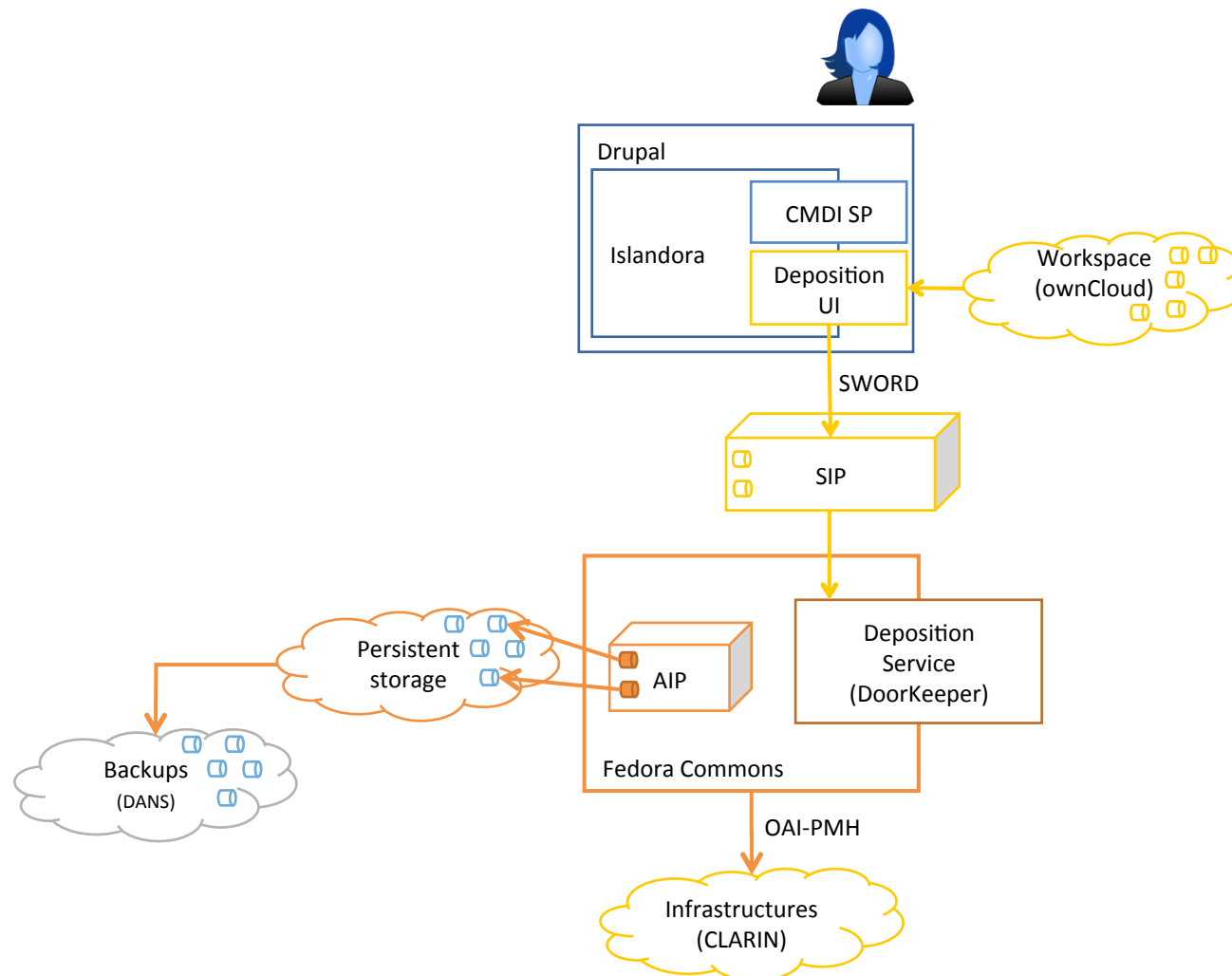
# Data curation

---

- New data
  - Starting point: depositor provides the resources in a preferred format
    - If needed conversion support
    - Completeness checks together with the depositor
  - Exceptions: unique data in a unique format
- [Digitalisation](#) of, especially, audio
  - Audio studios with extensive expertise and possibilities of recording older media
- (VRE) Projects together with MI [Technische Ontwikkeling](#)
  - Scripted conversions, checks and bulk import
  - Connection between the VRE and the repository
- Migration to a new repository ([TLA/FLAT](#))
  - In pre-ingest phase curation of the legacy, with special attention for licenses and metadata ([CMDI](#))



# (New) Workflow at the MPI/TLA



# Technical support in the pre-ingest phase

---

- LAMUS/DoorKeeper (tools which support the depositor/manager)

- Validate the metadata
  - Technical correct?
  - Quality score
    - [Scoring tool/library](#) developed by the CLARIN Metadata Curation TF
    - Not conclusive, but a low score hints at a problem
- Check the formats of resources:
  - Do they have the type/format they claim?
  - Are they valid instances of this format?
- Reject/flag resources in a unpreferred format
  - **Curation** is needed

- Connections to VREs (MI)

- Will deliver known formats and partially checked metadata
- Still interaction between depositor and collection management is needed, e.g., to check a match the collection profile

# DoorKeeper

---

- Java CLI/Servlet/library which executes a sequence of actions on a SIP
  - Action fails? pre-ingest is cancelled (or a partial ingest is rolled back)
    - Run a subsequence, e.g., the pre-ingest checks
  - Dedicated user and developer logs
- Actions can be generic (XSD validation, Schematron, FITS, derivatives, ...)
  - Share code 😊
- But also specific (institute specific policies, repository interaction, ...)
  - Actions are dynamically loaded, e.g., from an institute specific JAR
- <https://github.com/TLA-FLAT/DoorKeeper>
- Beta release, 1.0 with the upcoming FLAT release