Dolgotrajno ohranjanje raziskovalnih podatkov v manjših raziskovalnih infrastrukturah: Uporaba odprtokodne aplikacije Archivematica

Andrej Pančur (Inštitut za novejšo zgodovino) in Bogomir Rožman (UniPort — DR)

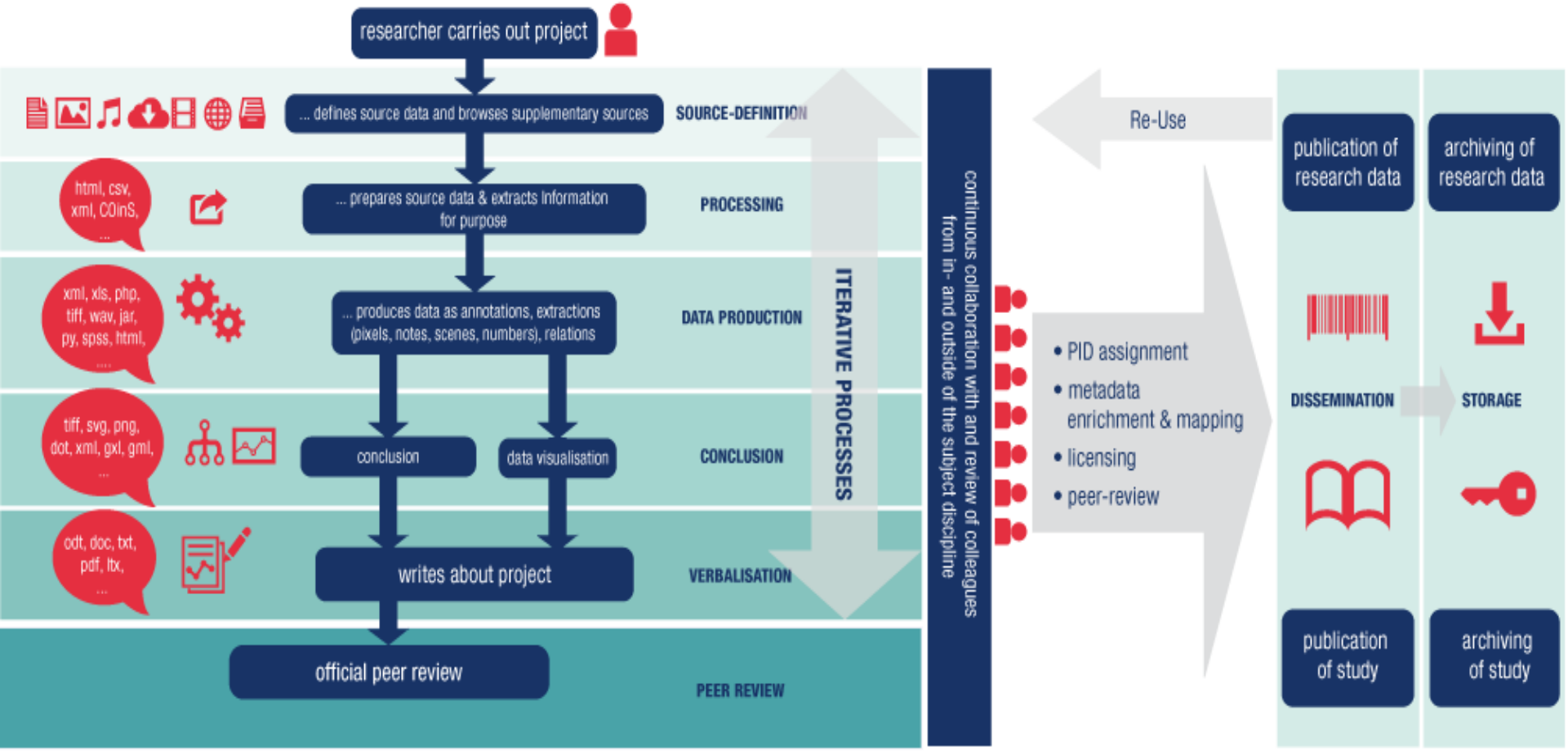
Mednarodna konferenca Jezikovne tehnologije in digitalna humanistika

Ljubljana, 29. 9. — 1. 10. 2016

Odprti dostop do raziskovalnih podatkov

- Obzorje 2020.
- Nacionalna strategija odprtega dostopa do znanstvenih objav in raziskovalnih podatkov v Sloveniji 2015—2020: pilotni projekti bodo morali raziskovalne podatke predati institucionalnim, splošnim ali področnim repozitorijem raziskovalnih podatkov:
  - Arhiv družboslovnih podatkov
  - CLARIN.SI
  - …

Življenjski cikel raziskovalnih podatkov



Zaupanja vreden repozitorij raziskovalnih podatkov

- je tisti repozitorij, katerega poslanstvo je zagotoviti zanesljiv, dolgoročen dostop do upravljanja digitalnih virov znotraj svoje skupnosti, tako danes kot v prihodnosti.
- Digitalna hramba raziskovalnih podatkov:
  - priprava metapodatkov in dokumentacije,
  - hramba podatkov v priporočenih formatih,
  - na priporočenih medijih,
  - priprava varnostne kopije in hrambe podatkov,
  - arhiviranje podatkov.

Raziskovalna infrastruktura Slovenskega zgodovinopisja Inštituta za novejšo zgodovino (RI INZ)

Dolgoročna hramba: Arzenal in Archivematica

portal Zgodovina Slovenije — SIstory

XML (TEI) → DARIAH Fedora Commons repozitorij

- publikacije (digitalna knjižnica)
- popisi prebivalstva
- žrtve 2. vojne
- ZIC in SICI
- zbirke raziskovalnih podatkov
- parlamentarni dokumenti (SlovParl)
- strankarski programi
- krajevni repertoriji
- znanstvene publikacije

Dolgoročna hramba

Arzenal

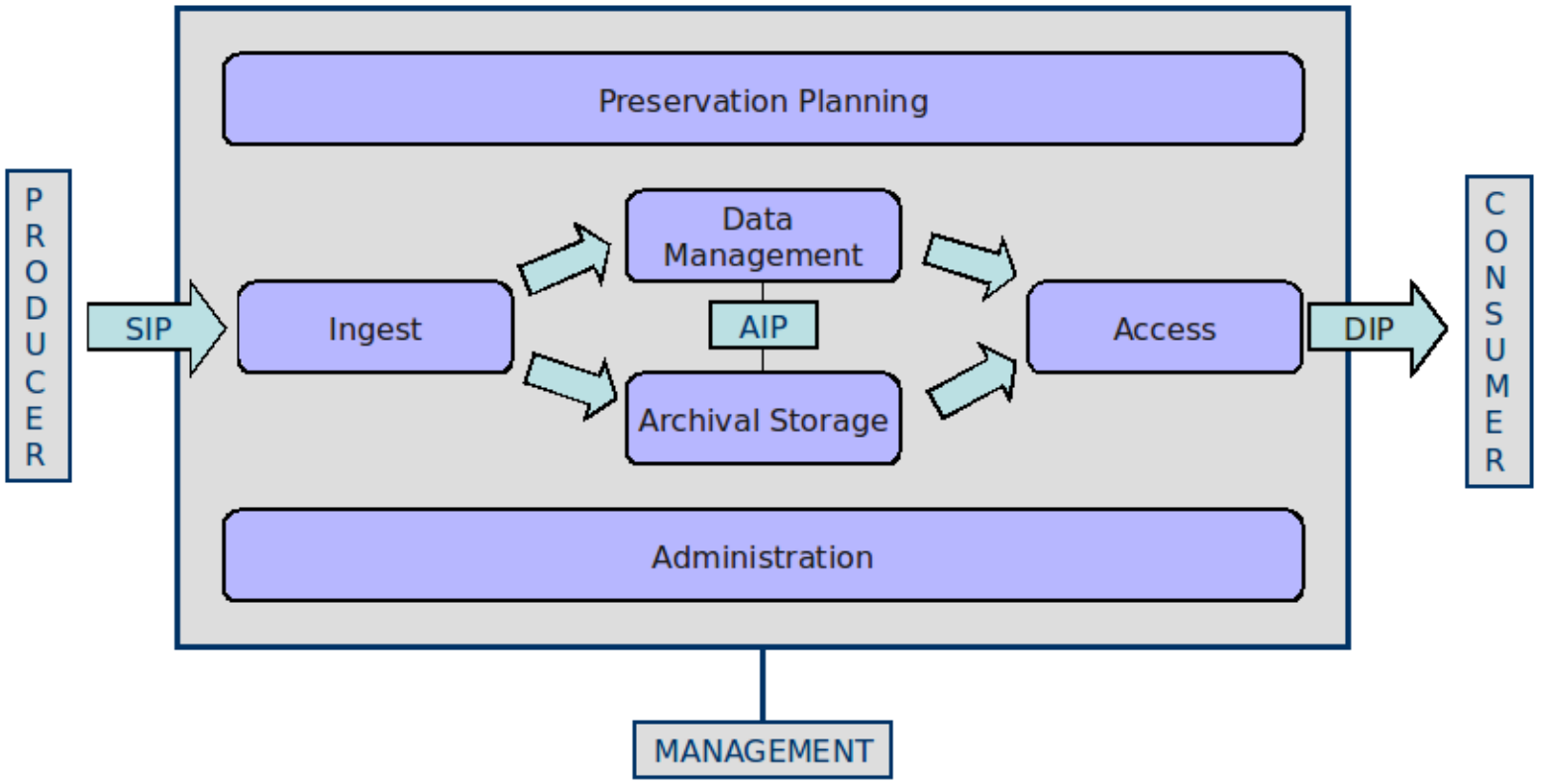
- Hramba digitaliziranega gradiva s področja dejavnosti knjižnic, arhivov in muzejev.
- Diskovna polja NAS (windows strežnik, NTFS): menijska struktura, datoteke + shadow copy.

Archivematica

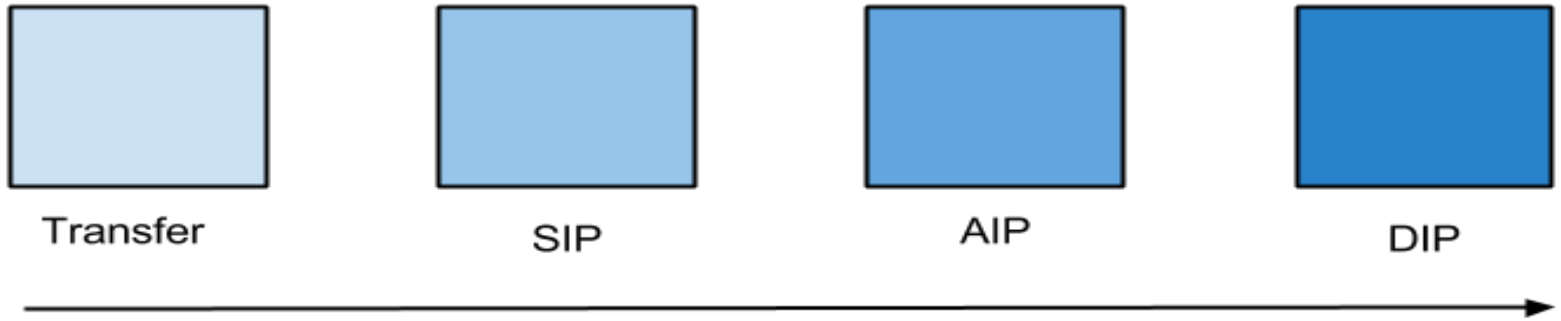
- Hramba:
  - večjih projektnih zbirk (popisi prebivalstva, parlamentarni dokumenti);
  - zbirke raziskovalnih podakov (v načrtu).
- Integralna zbirka odprtokodnih programskih orodij, ki uporabniku omogočajo obdelavo digitalnih objektov v skladu s funkcionalnim modelom OAIS.
- Izkoriščanje obstoječe IT infrastrukture.

- Arhiviranje na drug strežnik + shadow copy;
- Arhiviranje na več kot 50 km oddaljeno lokacijo (do konca leta).

Open Archival Information System (OAIS) reference model (ISO-STD 14721)

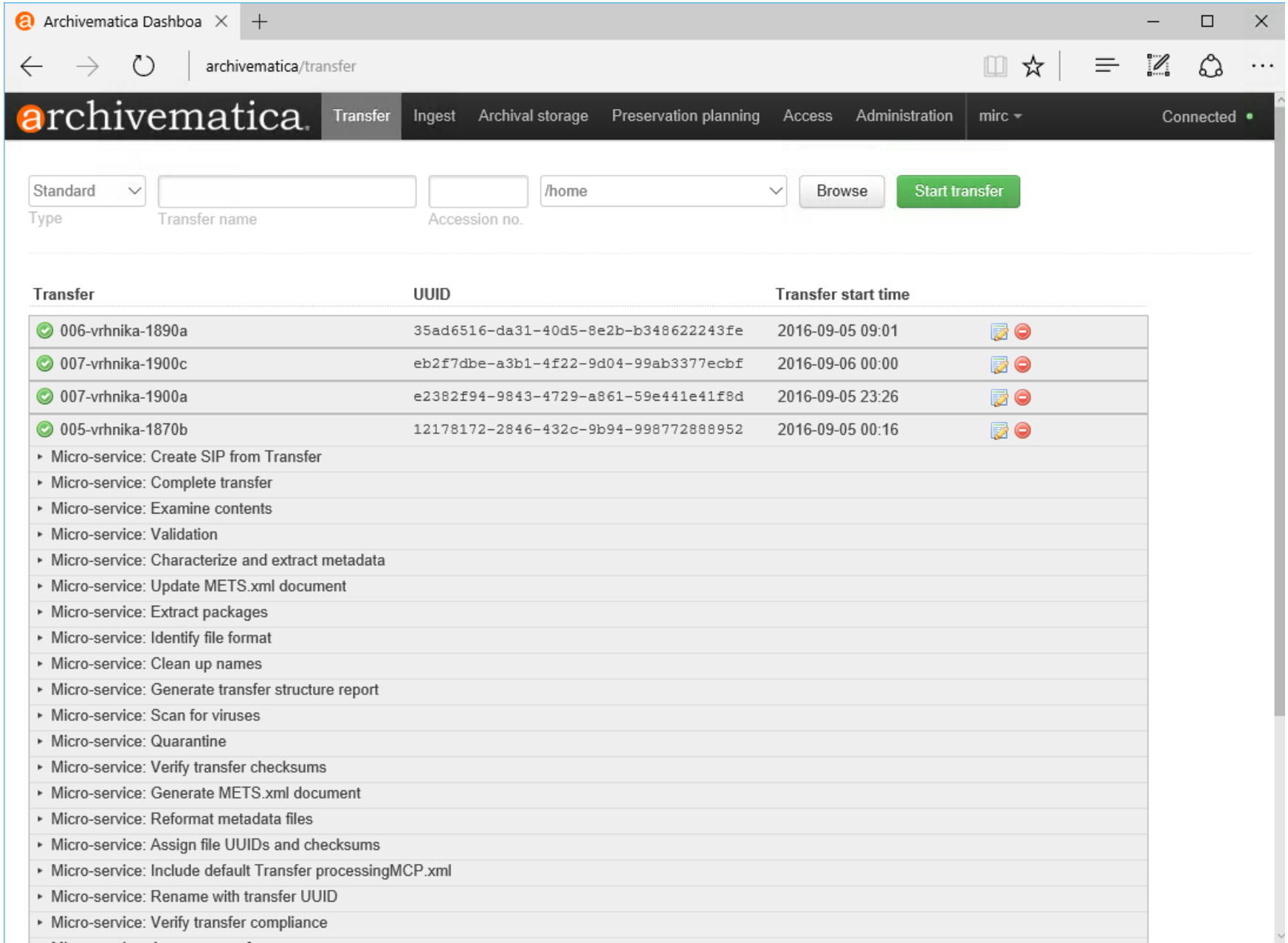


Vir: www.archivematica.org

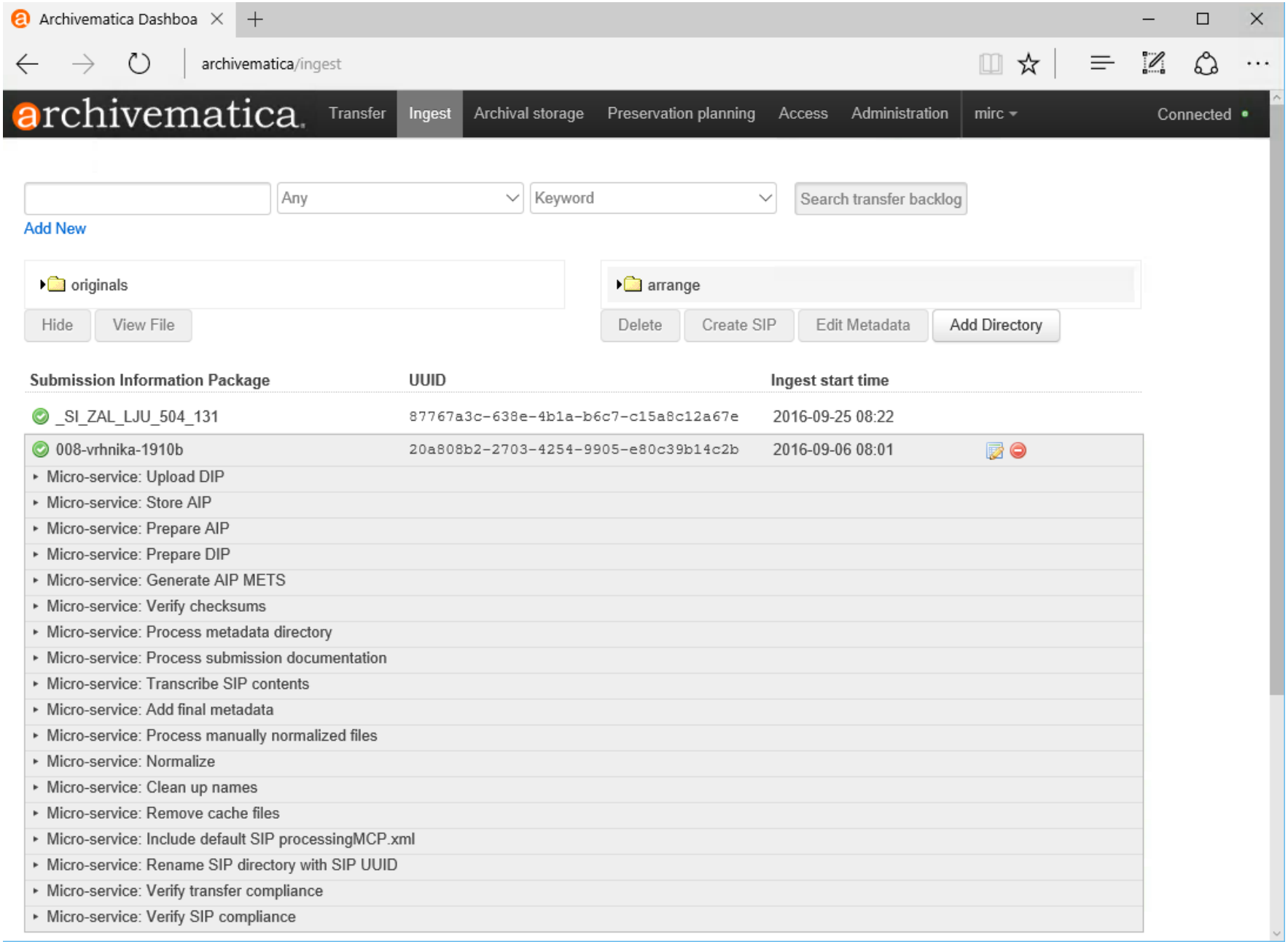


- Transfer: vhodni podatki:
  - Datoteke (PDF, PDFa, JPG, JP2, TIFF, XML, MP4, DOCX, XLSX, PPTX ...)
  - metapodatki na nivoju map in datotek (CSV).
- SIP (Sprejemni informacijski paket): strukturirani vhodni podatki za nadaljno obdelavo
- AIP (Arhivski informacijski paket): zip datoteka za dolgotrajno arhiviranje.
- DIP (Dostavni informacijski paket): zip datoteka za dostop, npr. na spletu:
  - lokalna spletna stran Atom za pregled DIP paketov;
  - uvoz datotek in metapodatkov iz DIP v Sistory.

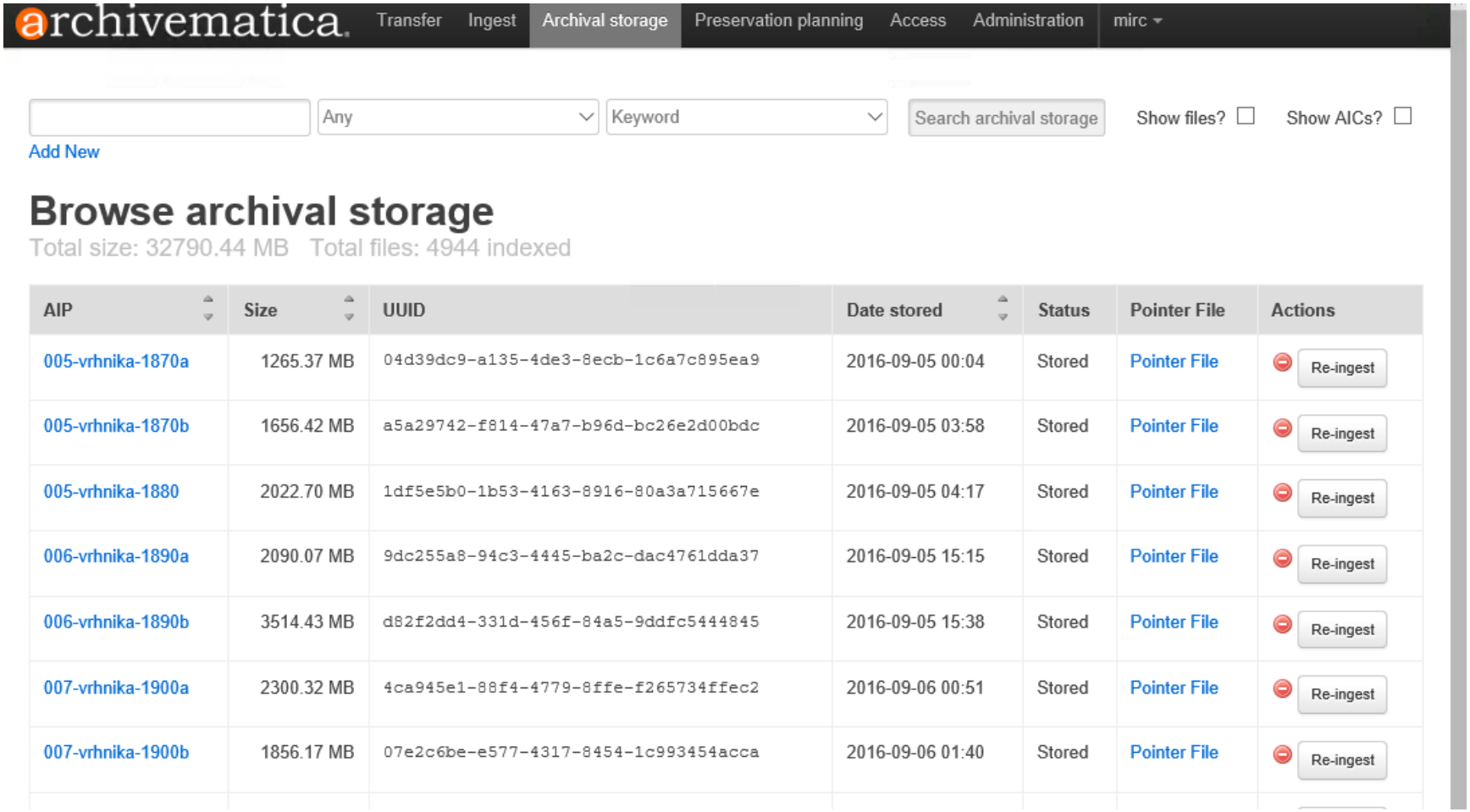
Transfer


Archivematica Dashboa
archivematica/transfer
archivematica
Transfer
Ingest
Archival storage
Preservation planning
Access
Administration
mirc
Connected
Standard
Type
Transfer name
Accession no.
/home
Browse
Start transfer
Transfer
UUID
Transfer start time
006-vrhnika-1890a
35ad6516-da31-40d5-8e2b-b348622243fe
2016-09-05 09:01
007-vrhnika-1900c
eb2f7dbe-a3b1-4f22-9d04-99ab3377ecbf
2016-09-06 00:00
007-vrhnika-1900a
e2382f94-9843-4729-a861-59e441e41f8d
2016-09-05 23:26
005-vrhnika-1870b
12178172-2846-432c-9b94-998772888952
2016-09-05 00:16
Micro-service: Create SIP from Transfer
Micro-service: Complete transfer
Micro-service: Examine contents
Micro-service: Validation
Micro-service: Characterize and extract metadata
Micro-service: Update METS.xml document
Micro-service: Extract packages
Micro-service: Identify file format
Micro-service: Clean up names
Micro-service: Generate transfer structure report
Micro-service: Scan for viruses
Micro-service: Quarantine
Micro-service: Verify transfer checksums
Micro-service: Generate METS.xml document
Micro-service: Reformat metadata files
Micro-service: Assign file UUIDs and checksums
Micro-service: Include default Transfer processingMCP.xml
Micro-service: Rename with transfer UUID
Micro-service: Verify transfer compliance

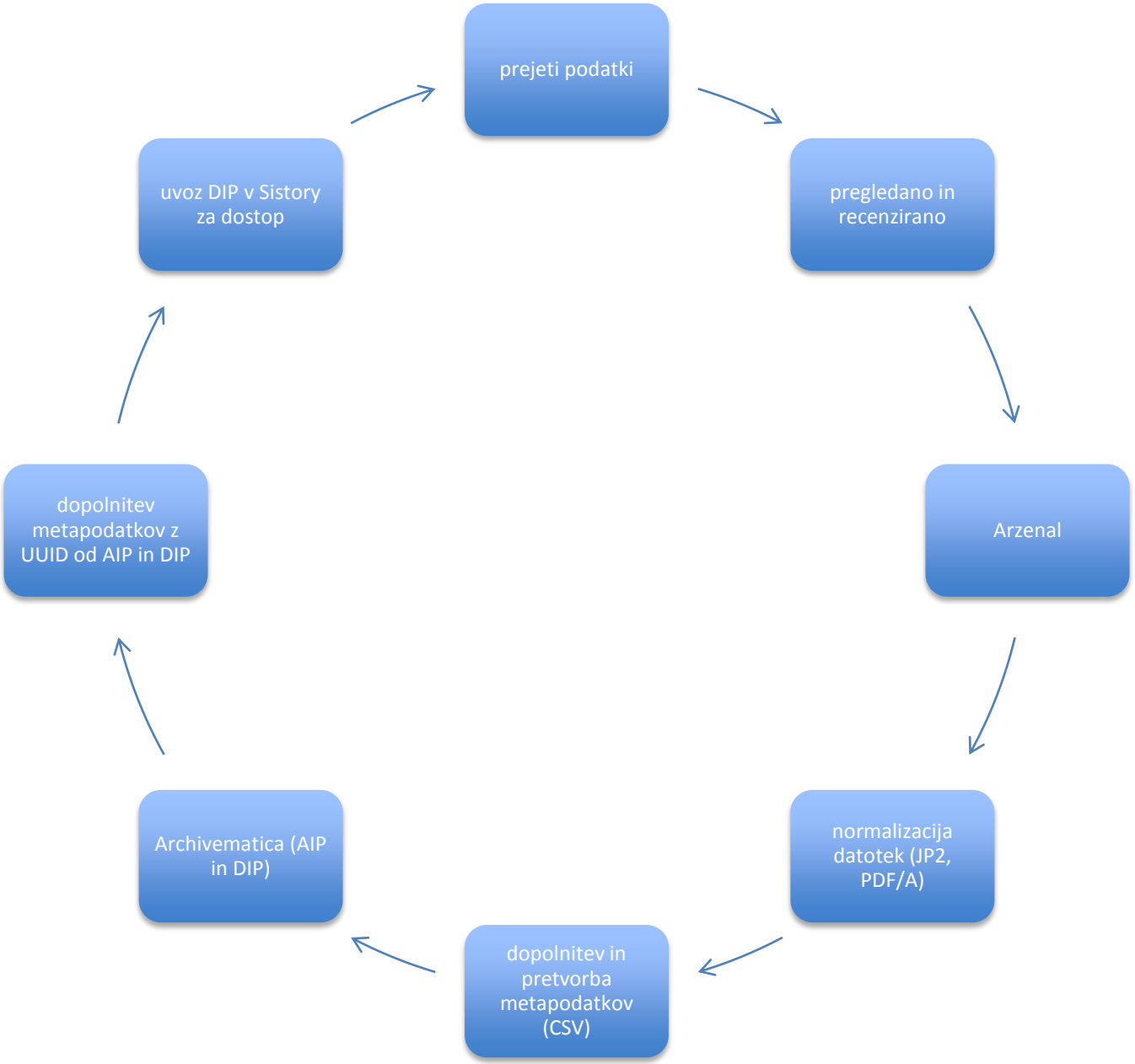
Ingest SIP


Archivematica Dashboa
archivematica/ingest
archivematica
Transfer
Ingest
Archival storage
Preservation planning
Access
Administration
mirc
Connected
Any
Keyword
Search transfer backlog
Add New
originals
arrange
Hide
View File
Delete
Create SIP
Edit Metadata
Add Directory
Submission Information Package
UUID
Ingest start time
_SI_ZAL_LJU_504_131
87767a3c-638e-4b1a-b6c7-c15a8c12a67e
2016-09-25 08:22
008-vrhnika-1910b
20a808b2-2703-4254-9905-e80c39b14c2b
2016-09-06 08:01
Micro-service: Upload DIP
Micro-service: Store AIP
Micro-service: Prepare AIP
Micro-service: Prepare DIP
Micro-service: Generate AIP METS
Micro-service: Verify checksums
Micro-service: Process metadata directory
Micro-service: Process submission documentation
Micro-service: Transcribe SIP contents
Micro-service: Add final metadata
Micro-service: Process manually normalized files
Micro-service: Normalize
Micro-service: Clean up names
Micro-service: Remove cache files
Micro-service: Include default SIP processingMCP.xml
Micro-service: Rename SIP directory with SIP UUID
Micro-service: Verify transfer compliance
Micro-service: Verify SIP compliance

Arhivska hramba



Življenski krog podatkov



Izkušnje in priporočila

- 2 leti spremljanja razvoja Archivematice, 2 testni instalaciji (verzija 1.3 in 1.4), produkcijska verzija 1.5
- normalizacija:
  - po normalizaciji se AIP paket lahko poveča za več kot 2x od uvoženega materiala (original + normalizirano za shrambo): JPG → TIFF → JPG
  - zato uporabljamo normalizirane formate JP2 in PDFa še pred Archivematico
- migracija AIP paketov med verzijami bo lažja šele z novimi verzijami
- stalna nadgradnja strojne opreme, ko so diskovna polja premajhna
- priporočamo snapshot pred vsako transakcijo, saj je drugače težko pospraviti napako
- poimenovanje paketov
- velika pozornost namenjena kreiranju metapodatkov (CSV)
- Problem: METS nima šumnikov (samo ASCII). Ta problem bo odpravljen v novih verzijah.
- Tabela hitrosti kreiranja AIP paketov:
  - po velikosti
  - po izboljšavah procesorskih moči (RAM, CPU)

Meritve

velikost materiala	št datotek	normalizacija	strojna oprema	čas	AIP velikost
9,5 GB	800	brez	4 CPU in 10 GB RAM	2,5 ure	10 GB
9,5 GB	800	brez	1 CPU in 10 GB RAM	8,8 ure	10 GB
Sklep: več procesorjev zelo pohitri obdelavo					
600 MB	300	z normalizacijo	4 CPU in 10 GB RAM	1,2 ure	4,8 GB
600 MB	300	z normalizacijo	1 CPU in 10 GB RAM	3,8 ure	4,8 GB
Sklep: normalizacija JPG slik naredi 8x večji AIP paket (pri pretvorbi JPG v TIFF)					