SSH Open Marketplace aggregation and ingestion pipelines

This document gathers minutes of the meetings supporting the SSH Open Marketplace data ingestion pipelines (i.e. DACE and custom scripts). It also includes relevant pointers.

Main participants/contacts:

- DARIAH: Laure
- OEAW: Michael, Klaus, Matej, Dalibor
- PSNC: Ola, Tomasz
- CLARIN: Alex

See also:

- https://harvester-manager.acdh-dev.oeaw.ac.at/
- DACE repo: https://gitlab.pcss.pl/dl-team/aggregation/dace > https://github.com/SSHOC/data-ingestion/tree/sshoc-dace-to-acdh-ch-k8s
    - Esp. list of sources processed: https://gitlab.pcss.pl/dl-team/aggregation/dace/-/wikis/SSHOC-Sources-Configuration
    - Specs (json/jolt mappings): https://gitlab.pcss.pl/dl-team/aggregation/dace/-/tree/develop/processors/sshoc-records-processor/src/main/resources/specs
    - Jolt tutorial https://gitlab.pcss.pl/dl-team/aggregation/dace/-/wikis/JSON-to-JSON-Mapping-tutorial
- SSH Open Marketplace Github: https://github.com/SSHOC
    - esp. data ingest repo
    - https://github.com/SSHOC/data-ingestion/tree/sshoc-dace-to-acdh-ch-k8s
- Manual mapping Gsheet: mappings - sources to MP data model
- Editorial_Board_SSHOpenMarketplace_minutes
- MP_provision&maintenance
- SSHOC T7.3 minutes Minutes_T7.3_Mapping_interoperability_meetings

Decisions summary

⇒ ingest workflow:

1. ingest first on the stage instance with status “ingested” and “dace_importer” (=system_importer) user role. Once reviewed,

2. ingest on the production instance with status “ingested” with “dace_importer” (=system_importer) user role
3. have a bulk approval via notebook

Minutes of the meetings

2024-11-28, 15h - EOSC dump ingest

Michael, Laure

2024 Autumn meetings

Minutes and decisions: Ingest and curation sprint 2024- SSH Open Marketplace

2024-09-09

PH in Campus and MP - with Michael and Vicky

- https://campus.dariah.eu/source/programming-historian/page/1 - 10 published and 10ish in the pipeline, but the idea would be to have everything (currently 110 in English)
- English content only - but mention of other languages too
- Manual entries created for Campus - tested ingest as batch for austrian how to but didn’t work.
- PH would like to update what they have in the MP - via Campus or not? Via Campus yes
- If PH team wants to create items on the MP that are not in Campus, they can do that by hand too
- MP and not Campus:
    - Multilingual records - MP possible, Campus not yet but will be
    - Use collections to gather the PH multilingual lessons, as an option

2024-08-22 16:00 - CRF ingest

Michael, Alex, Laure

https://dariah.zoom.us/j/87958421976?pwd=qbWLyU7LwsRvUy2av3s86m7XNuv5sq.1

1. **Reingest of the CRF: https://github.com/SSHOC/SSHOMP-Ingest/issues/44**

https://github.com/clarin-eric/clarin-resource-families

Exact copy from csv to JSON. Except for the publications, we ignored them and they were citation strings. Now Zotero links that we could ingest in parallel

DACE or Python scripts?

- Python scripts would be easier and Alex could keep control - https://github.com/SSHOC/SSHOMP-Ingest
- Look into the Zotero pipeline - based on https://www.zotero.org/groups/562080/clarin/collections/GDJJGIBW - maybe postponed after the conference
- Check with Ola how easy it would be to "update the existing CRF pipelines" - TO DO Laure >> done: https://github.com/SSHOC/SSHOMP-Ingest/issues/44

Delete all the CRF from production, but check before any changes that happened on production, esp. regarding relations.

- As continuous ingest - pay attention to duplicates

Timeline: finish by the CLARIN conference (15 October) would be ideal, but might be too tight

Priority list:

- CH - finish stage and push to prod.
- TDT - actors still tricky - some tests
- And then, CRF

2. **CRF citation strings**

Possibility to get the info from source? For some, yes. But we don't know

VLO can't provide all info needed for citation

3. **ATRIUM, CLARIN DOG and MP curation**

Contacts about the DOG: Dieter directly. Dev is Michał

Theoretically, take all externalIDs from MP and/or all the URLs from CRF, check via DOG sniff and fetch what is possible

https://alpha-dog.clarin.eu/?pid_field=http%3A%2F%2Fhdl.handle.net%2F11858%2F00-246C-0000-0023-8CF9-6&functionality_field=sniff

Assuming, we can fetch something, the process would be to add info in CRF records >> MP.

Laure TODO : write to Dieter

Alex's CRF plan:

- Everything in JSON
- Fetch can be fetched from VLO
- DOG for other sources/repo >> could we fetch enough to rebuild the citation strings?

2024-05-23 11:30 CEST

Michael, Laure

https://github.com/orgs/SSHOC/projects/3/views/8

TDT, CH

- identification of the possible merges: log files at the moment. In the future, flag raising
- Merge items: sources identification of merged items in the new one - how to proceed for continuous ingest? Could we not keep the mention of two sources in merged items?
    - Conceptually, all items should be attached to their sources, even when they already exist in the MP - merges after
- Iana-mime type vocab. is limited!!
- Source ActorId - how does it work to create?

TO DO:

- Laure - Ola and his colleague to check Dariah.Lab
- Laure - contact CLARIAH-NL
-

2023-12-18 15:00

https://dariah.zoom.us/j/88616493484?pwd=32mhOkI1mZybdUA40aqLor3yrsLlvc.1

2023-12-11 15:00

https://dariah.zoom.us/j/88616493484?pwd=32mhOkI1mZybdUA40aqLor3yrsLlvc.1

Agenda:

Status Dariah Campus reingestion

Ola mentioned issue with linked media, to be checked

Status Dariah campus - what to do with items that have been curated since last ingest? -> go ahead with ingest, MP curation team can compare and decide. Use

Keyword clean-up - CRF issue - leave as is because a lot of changes are planned for CRF in the near future

Dariah.Lab - first tests with ingest on dev

JOLT tests to plan (EOSC MP could be an option, although this source is on hold)

Need spreadsheet mapping/review mapping	***Need JOLT mapping***	***Need DACE ingest or re-ingest***	***Need review ingest***
TDT - source API needed		CAMPUS mappings - s... No new relevant fields, this can be next	CRF - latest status discussed on the Slack #ingestion channel - still keywords issue Natalia working on it Mess with the sourceItemID during reharvesting
		Dariah.Lab - is the JOLT mapping started? Add Ewa to GitHub	
Conversion Hub - source API needed One-time ingest through MP API, MK	PH	CESSDA Training	TAPoR ingested on dev, which status? Accepted?? Don't see any TAPoR items with activity on June 1 22 harvester here Can we aggregate on stage?
EOSC Marketplace - Michael started the "new" mapping. Need decision and clear workflow how to handle			LRS ingested on dev, which status? Can we aggregate on stage?

initial source and records already ingested			

2023-12-04 15:00

https://dariah.zoom.us/j/88616493484?pwd=32mhOkI1mZybdUA40aqLor3yrsLlvc.1

Michael, Matej, Alexander, Ola

- Status of EOSC mapping mappings - sources to MP data model "new" mapping ready to be discussed, some fields still tbd from "old" mapping. -> on hold due to expected substantial changes in the technical setup on EOSC side (procurement, new catalogue(?))
- CRF -> similar to EOSC, pending further changes to source
    - check use of white-list https://github.com/SSHOC/dace-ingestion/issues/39 Current hypothesis: Whitelist is not always applied. Whitelist should be applied *before* items are passed to the 'keyword' field at the target side
- Dariah Campus (see existing mapping mappings - sources to MP data model), good candidate for reingestion?
- Important to have Alexander join Curation TF meetings again, so that he is informed about the work on keyword curation.

To do:

- Priorities Ola:
    - Review mapping Dariah.Lab and ingest (in joint session with ACDH-CH)
    - check use of white-list https://github.com/SSHOC/dace-ingestion/issues/39
- Try reingest DARIAH Campus (has been imported initially still with Poolparty) #40 Michael check mapping
- Michael (with Klaus) check TDT, and ConversionHub - availability of the API

Need spreadsheet mapping/review mapping	***Need JOLT mapping***	***Need DACE ingest or re-ingest***	***Need review ingest***
TDT - source API needed		CAMPUS mapping...	CRF - latest status discussed on the Slack #ingestion channel - still keywords issue Natalia working on it Mess with the sourceItemID during

			reharvesting
		Dariah.Lab - is the JOLT mapping started? Add Ewa to GitHub	
Conversion Hub - source API needed	PH	CESSDA Training	TAPoR ingested on dev, which status? Accepted?? Don't see any TAPoR items with activity on June 1 22 harvester here Can we aggregate on stage?
EOSC Marketplace - Michael started the "new" mapping. Need decision and clear workflow how to handle initial source and records already ingested			LRS ingested on dev, which status? Can we aggregate on stage?

-

2023-11-27 15:00

https://dariah.zoom.us/j/86517468206?pwd=G1Z55PZtbQZembKeXW6fhPultWZ9BY.1

Michael, Matej, Laure, Ola

- New repo - https://github.com/SSHOC/dace-ingestion
  - Move relevant code branches in the new repo - which ones?? Ola and Dalibor know. https://github.com/SSHOC/data-ingestion/tree/sshoc-dace-to-acdh-ch-k8s ?
  - All DACE and MP sources related issues moved in the new repo. All tool extraction tasks closed
  - 3 main labels: "MP sources" (one umbrella issue per source to be ingested); "DACE enhancement" and "bug"
  - Can we close the following?
    - Deploy DACE on ACDH-CH servers #18 ?
    - Handle records updates in DACE properly #12
  - Reviewing open issues - time allocation
    - @ola to look into for DACE enhancements
    - For MP sources, have a generic time estimation for all sources, and assess case by case depending on specificities

        - 1 working week per task - ideally 1 month per source. Synch points, maybe 2 per source
    - Review ingest workflow as described above in this document with Klaus (user roles behind system importer??) > Michael and Klaus
- MP sources ingest status (adding the 4 phases in the table below as labels for the MP sources issues)?

Need spreadsheet mapping/review mapping	***Need JOLT mapping***	***Need DACE ingest or re-ingest***	***Need review ingest***
TDT - source API needed		CAMPUS	CRF - latest status discussed on the Slack #ingestion channel - still keywords issue Natalia working on it Mess with the sourceItemID during reharvesting
		Dariah.Lab - is the JOLT mapping started? Add Ewa to GitHub	
Conversion Hub - source API needed	PH	CESSDA Training	TAPoR ingested on dev, which status? Accepted?? Don't see any TAPoR items with activity on June 1 22 harvester here Can we aggregate on stage?
EOSC Marketplace - Michael started the "new" mapping. Need decision and clear workflow how to handle initial source and records already ingested			LRS ingested on dev, which status? Can we aggregate on stage?

- Aggregation sprint by the end of the year.
    - What to prioritise? Sources, enhancements, documentation?
    - How long? When?
        - December possible - January more time

Dec sprint 1 - 2 sources: CRF bugs, DARIAH.Lab ingest stage possible (EOSC MP in January, but maybe first assessment)

Starting next week, couple of days per week. Meeting every week, every Monday at 15h. New colleague next year.

DACE roadmap - some developments of applications not used for sshoc depend on further funding for Dariah.Lab

AOB: date for the dev telco (without Stefan?)

2023-08-30, 11h cest - Dariah.Lab ingest

https://dariah.zoom.us/j/85831452710?pwd=ZnRqUjNsUDUrT2lHTmM0QmsyVDhVUT09

Participants: Ewa, Ola, Alex, Laure

Agenda:

- Dariah.lab spreadsheet mapping: mappings - sources to MP data model - last open questions (cells highlighted in yellow)
    - Dariah.lab name @Ewa
    - Check relations mapping @Laure
- First steps to start the JOLT mapping
    - Structure challenging
    - Ingest of the relations will also be challenging too
- Agreement on a timeline
    - No strict deadlines from dariah.lab
    - Until the end of september for the first JOLT mapping
    - October review
    - Aiming at publishing November

- CLARIN Resource Family ingest status
    - Points 3, 4, 5 of https://github.com/SSHOC/data-ingestion/issues/58#issuecomment-1598406990 solved?
    - Lexical resource ingest tests - https://github.com/SSHOC/data-ingestion/issues/58#issuecomment-1633971460 - first point solved? Second point is solved = issue in the URL encoding used in the query
- @Klaus - double-check roles and rights and status for ingest system importer can currently ingest items with approved status on dev instance
- Natalia to confirm everything ok
- 3 records missing in the Lexical resources
- Then, delete everything clarin rf related on dev
- Fresh ingest
- Then, Alex to check random records

2023-07-18 - ingest status update

	Need spreadsheet	***Need DACE mapping***	***Need DACE ingest or re-ingest***	***Need review ingest***

	mapping/review mapping			
1				CRF
2	Dariah.Lab			
2	TDT			
2	Conversion Hub			
3	EOSC Marketplace			TAPoR ingested on dev, which status? Accepted?? Don't see any TAPoR items with activity on June 1 22 harvester here Can we aggregate on stage?
4		PH	CAMPUS	
5			CESSDA Training	LRS ingested on dev, which status? Can we aggregate on stage?

2023-07-17, 14h30

Klaus, Laure

- DARIAH-Lab mapping : https://docs.google.com/spreadsheets/d/17E6oJ_mXZFXNhQWhvOMmPLu5reJKImwI5OP44yudfe8/edit#gid=907596661
- Comments&Discussion - comments added directly in the table. Here is just a summary of our conversation
- :
    - Short resources vs. resources
    - En endpoints?
    - Resource types issue: several types per issue.
        - Rules for ingestion types (at the bottom of the table)
- Johd paper: from ssk to mp
- From the rancher about the CRF failure:

- /api/sources/353/items/Lexical_Resources/Dictionaries/Dictionaries%2520in%2520the%2520CLARIN%2520infrastructure/1-Monolingual%2520resources.csv%23L20
- https://sshoc-marketplace-api.acdh-dev.oeaw.ac.at/api/sources/353/items/Lexical_Resources/Dictionaries/Dictionaries%2520in%2520the%2520CLARIN%2520infrastructure/1-Monolingual%2520resources.csv%23L20

2023-07-13

Alex, Laure

CRF https://github.com/SSHOC/data-ingestion/issues/58

2023-06-26, 14h-16h - Creating a new pipeline

https://dariah.zoom.us/j/89213109643?pwd=Mi9YY3pOUmJKV2RKRFcyVHpwRER5dz09

Ola, Klaus, Matej (tbc), Laure, Tomasz & Ewa. Alex excused!

- DARIAH-PL catalogue mapping & ingest preparation

Ewa prepared the mapping on spreadsheet

- Accessmode
- Accesstype
- OrderInfo

Careful in the identification of the categories:

- What is “document” ? Always a dataset or always a publication?
- A resource can have different types in DARIAH-Lab

For the JOLT and ingest: PSNC on DARIAH-PL project and hours: name of the source “DARIAH Lab” “DARIAH-PL” > DARIAH Lab Poland

Possibility to add new properties on-demand

API is publicly open.

~less than 100

- Initial mappings are in the spreadsheet: Lexical Resources, Tools > how to create two new pipelines? On the basis of the existing CLARIN one

199 tools on dev

- See with Alex: what kind of licenses are in. If mapped to keywords

- From Tomasz (question): EOSC future - when is the deadline for OAI-PMH and aggregation routines? We know the project is extended until 12.2023, but not sure if the work can be done after 09.2023 from your perspective.
- EOSC-Future AAI topic too - https://github.com/SSHOC/sshoc-marketplace-backend/issues/374

To follow-up:

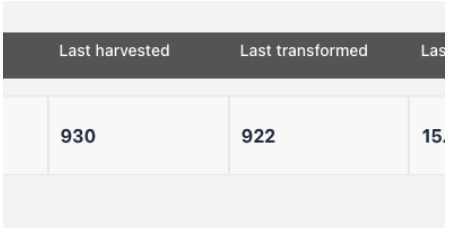
- JOLT - what to do when something new? Keep in mind the tuto: https://gitlab.pcss.pl/dl-team/aggregation/dace/-/wikis/JSON-to-JSON-Mapping-tutorial

2023-06-21, 10h-12h - Modifying an existing pipeline

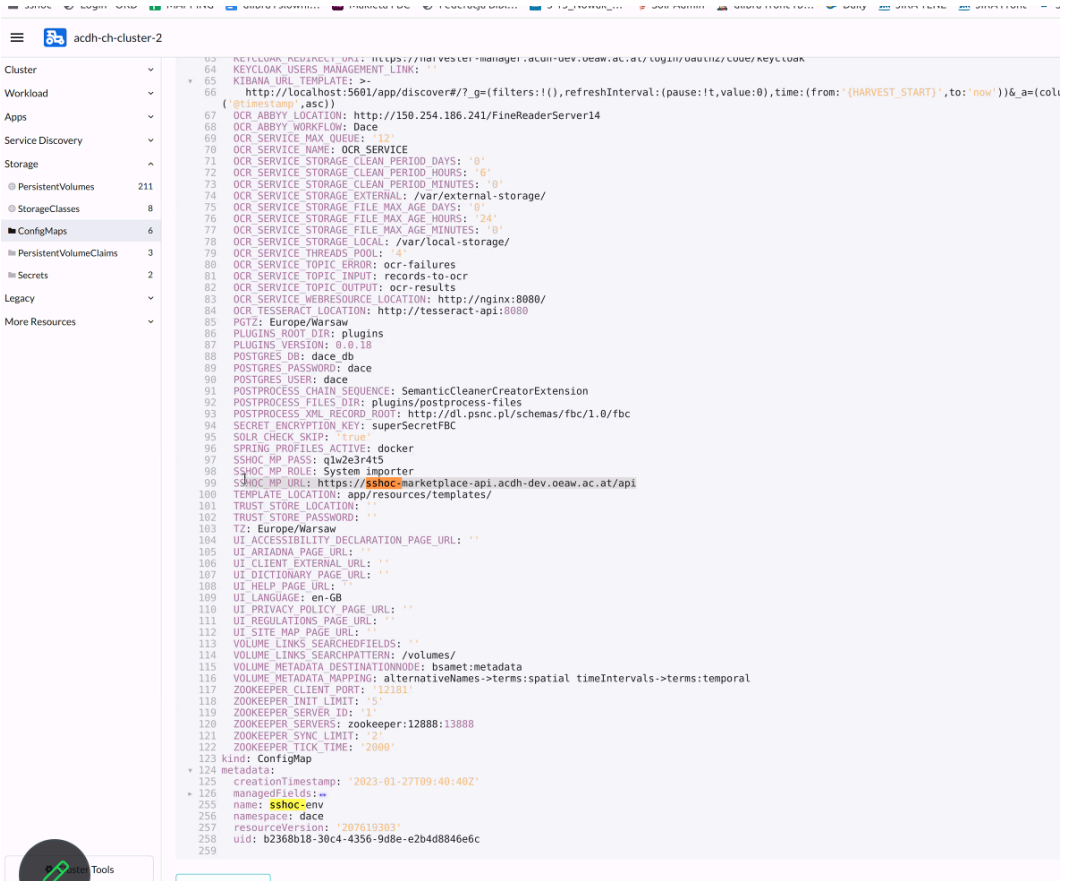
https://dariah.zoom.us/j/88071222344?pwd=Ujg2bG5aV2trT1hyQWFhcWFjU0I3Zz09

Alex, Ola, Klaus, Matej, Laure

- Last issues with the deployment? Cf. https://github.com/SSHOC/data-ingestion/issues/97
  - Deployed yes. With upgrades. But still upgrades to come.
  - Some applications are currently disabled. 3 containers not running.
- CLARIN RF - https://github.com/SSHOC/data-ingestion/issues/58
  - Re-run - 8 records missing



  - More details on the missing records? Via kibana (not implemented yet) - but log files via rancher can be consulted
  - Which instance of the MP? In the config. A separate harvesting set up. In github or via rancher



- https://github.com/SSHOC/data-ingestion/issues/58#issuecomment-1598406990

1. ok. Solved!

To test just on a few records > create a new source and just add one-two records in the “queries” field

Verify the mapping changes via JOLT, getting rid of the custom transformers. Note that operations are run in order in the jolt spec.

Rancher and Kouncil to see the logs

2. solved! Because use of an outdated branch from the clarin source

3. Not solved! If concept is not found in the given vocabulary, a concept in keyword should be created.

4. Not solved. Concept doesn’t exist at source, not in the white list, but still is created.

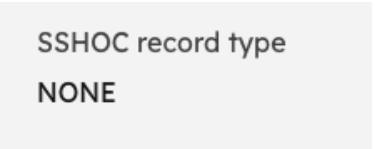
- add hidden/curation property with the DACE record ID > https://github.com/SSHOC/data-ingestion/issues/108

- CLARIN RF other mappings:

-if duplicates: the lookup searches by source and by sourceID. > Maybe for the future, change

Take one record as example in the JOLT demo

Entity type: defined either in the mapping or in DACE/source



What about the custom transformers?

2023-06-02, 13h-15h - DACE deployment

https://dariah.zoom.us/j/84600831020?pwd=YS9CWW9zWmJRSEFyR1oxZHlFbWJwdz09

Alex, Ola, Klaus, Matej, Laure, Tomasz

- https://github.com/SSHOC/data-ingestion/issues/97
- Currently update/upgrade of the code base to solve the sshoc specific issues
- The latest harvest on acdh instance worked with clarin - singular problems > Klaus and Ola worked on that together this week.
    - 19 out 20 records successful
    - Tomasz fixed a few things on dace code.
- Is there an issue with Keyword - some keywords still available, although should be eliminated (cf. https://github.com/SSHOC/data-ingestion/issues/58#issuecomment-1544029175)
- https://github.com/clarin-eric/resource-families-html-generator/blob/master/resource_families/Corpora/Academic%20corpora/1-Corpora%20of%20academic%20texts%20in%20the%20CLARIN%20infrastructure/1-Monolingual%20corpora.csv#LL14C7-L14C11
    - "<ul<li>soci"
    - Fixed with this commit
- List several csv in one source - but reaggregation will then be done for all the csv listed.
- In case of update: new entries, changes in existing entries, complete new families.. Everything could change
-

=> test on dev before the meeting, and we do from scratch together on stage next time

2023-05-11, 15h - CLARIN Resource Families ingest

https://dariah.zoom.us/j/82540636600?pwd=cHo5V0lLanpZK3hmRkhSRy9NNkg5Zz09

Klaus, Alex, Laure, Matej

Suggested agenda

Comments received about clarin rf

- Comment from Martin (Finland). Some things outdated in both MP and CLARIN RF - update to do in CLARIN RF. 2nd comment: VLO instead of CLARIN RF??? To be discussed during clarin meeting in June. Alex and Martin
- From Jörg Knappen: issue on page /dataset/CwMoRs. Please describe: This corpus should also be tagged "greek" for its language. The same goes for the other corpora from CLARIN resource families. The language information is actually available in the file https://github.com/clarin-eric/resource-families-html-generator/blob/licenseurls/resource_families/Corpora/Newspaper%20corpora/1-Newspaper%20corpora%20in%20the%20CLARIN%20infrastructure/1-Monolingual%20corpora.csv#L13 (it is the 7th column, counting from 1) so it can be easily added automatically.

DACE status

- https://harvester-manager.acdh-dev.oeaw.ac.at/sources
- https://github.com/SSHOC/data-ingestion/issues/97 + Slack ingestion channel
    - Kibana > does it include log files (for ex. If no match with a specific concept (greek missed for ex.)
    - Test processing - stop after the first record processed
- Intensive DACE meeting - before the end of May - Ola+Tomasz+Klaus+Alex+Matej > Laure to do

Tasks & Planning

- bulk rejection of the status ingested items on production instance. Because as you've noticed @Alex, at least for one item, there is a very wrong keyword value coming out of nowhere... >> within the next two weeks
- cleaning/updating the pipeline to support
    - proper keyword mapping -
        - take only the keywords that don't match the white list
        - Is it implemented already? Where to add the white list? Probably line 6 https://harvester-manager.acdh-dev.oeaw.ac.at/mappings - update of https://github.com/SSHOC/data-ingestion/issues/58 with Q for Ola
    - Language seems only problem for Greek: because there is no label for "Greek" in the language vocabulary (just "Modern Greek (1453 -)")
- Updating the existing approved records:
    - Re-aggregating
    - Bulk approval
- Then, there was also the plan of mapping and ingesting the two other folders (tools and lexical_resources).
    - Initial mappings are in the spreadsheet: Lexical Resources, Tools
        - Ola to show us how to adapt existing mapping for these two new ones.
    - an additional topic in the discussion could also be the relations between RF items

        - There is a plan to properly link Zotero items to most RF items > https://www.zotero.org/groups/562080/clarin/collections/GDJJGIBW
- duplication/merging issue: some items can be in multiple RFs
- Jakob contact for the curation and Alex for the tech part.

2023-02-14, 12h

https://dariah.zoom.us/j/87367464521?pwd=S2FOQXVJQ2xHTmMrMlVxTHJ5aFdIUT09

- Klaus, Matej, Tomasz, Ola, Dalibor, Laure

Agenda:

- Deployment di#97
    - Deployed.
    - How much resources needed?
    - Keycloak issue still open but not blocking. The harvest manager can be used already?
        - Keycloak to work with eosc aai?
        - Agree for shibboleth metadata to be sent to clarin and dariah
- Configuration in Gihub in the pipeline that delivers set-up.
    - https://github.com/SSHOC/data-ingestion but in a dedicated branch > https://github.com/SSHOC/data-ingestion/tree/sshoc-dace-to-acdh-ch-k8s
    - Public availability still a problem because of password exposure. Encryption or github secret or?? Hide variables, or make the repo private?
    - https://github.com/SSHOC/data-ingestion/blob/sshoc-dace-to-acdh-ch-k8s/envs/sshoc.env#L86
    - No change on dace side needed.
    - No password in plain text in config file + private repo
    - Harvester UI:
    - A new source with a new type would need some java implementation
    - Reharvest = continuous ingest
        - It works if IDs at source did not change.
- Where to store new JOLTs?
    - actually let's use this branch for now https://gitlab.pcss.pl/dl-team/aggregation/dace/-/tree/sshoc-in-docker/etc/aggregator-config/mappings/json
    - import/export to DACE via the interface
- Update CLARIN RF
- Continuous ingest di#100 & handle records update di#88
    - Review mode via items to moderate interface

2023-02-09, 11h30

https://dariah.zoom.us/j/87889396893?pwd=dUdhRGYrZEJwYmJFNVQ3eHR1TUJDZz09

Klaus, Matej, Laure

- Documenting the mapping definition step of the ingest workflow
  - More details than in d7.2 and d7.3? Tuto/documentation moderators
  - useful to have a template? mappings - sources to MP data model >> NO, this will change.
  - With JOLT, maybe spreadsheet not needed.
    - Conceptual vs. implementation: 1 or 2 steps?
    - Where are the other JOLT mappings?
    - Where to store new JOLTs
    - One basic fictive mapping together with Ola
  - Documenting now would be the workflow
    - JOLT interface and JSON exports??
    - Inline documentation could be added
    - JSON export as template?
- Sources waiting list - mappings - sources to MP data model
  - CLARIN RF:
    - Deleting keywords: need for Ola to wait for curation cleaning or not?
    - Two other folders at source > two other mappings - Alex
  - EOSC catalogue update mapping
  - SSHOC training toolkit
    - Klaus - API changes needed. Good use case to test the JOLT stuffs.
  - SSHOC WP3 conversion hub
  - DARIAH-PL catalogue
  - JOHD - needs investigation
    - Dblp check? pipeline
    - Journals & tools more than journals and data? Tools?
  - CLARIN training
    - UPSKILLS - Iulianna
    - TDT > curation there > MP. But double-check

New sources workflow via Github, new issue per source. Where? Data-ingestion repo or dace reop? Where/when to add the JOLT mappings (JSON)?

1. DACE running at acdh
2. Adapt an existing mapping: CLARIN RF, EOSC, SSK (missing standards)
3. Create a new one: TDT

Other dev pbms:

- Backend actions

- Eosc aai - credentials - shibboleth.xml >> Laure to check if something from Yoann

2023-01-16, 15h

- Klaus, Matej, Tomasz, Ola, Dalibor, Laure
- https://dariah.zoom.us/j/81168309646?pwd=ZnlIT09mb0tmaXM3ZGxCYzQ0a1FjQT09

Suggested agenda:

- Deployment di#97
  - https://github.com/SSHOC/data-ingestion/issues/107 but in the meantime, keycloak container
  - docker-template/docker-compose need to be adjusted to be usable by kubernetes
    (esp: "extends"/"profiles" must be eliminated)

Relevant configuration files for authentication:

- https://gitlab.pcss.pl/dl-team/aggregation/dace/-/blob/develop/managers/harvester-manager/src/main/resources/application.yml
- https://gitlab.pcss.pl/dl-team/aggregation/dace/-/blob/sshoc-in-docker/etc/sshoc.env > lines 45-48

- Continuous ingest di#100 & handle records update di#88
  - To test also with clarin resource family on stage
- CLARIN resource di#58
  - Cesare to delete existing keywords, except the ones from the white list. Add the wrong keywords in our list.
  - Then, continuous ingest against the white list as dummy source on stage

>> Jan 30, 15h (invit to be shared also with Dalibor if deployment questions to discuss)

2022-12-09, 14h-15h30

- Klaus, Matej (leaves at 15h), Tomasz (leaves at 15h), Ola, Laure
- https://dariah.zoom.us/j/85083186102?pwd=S3Yxc3lMVThLcXZza2VLOHlBYXV1dz09

- Suggested agenda

Based on DACE labels = https://gitlab.gwdg.de/groups/sshoc/-/issues/?sort=updated_desc&state=opened&label_name%5B%5D=dace&first_page_size=20

- Deployment at ACDH and work/status on the DACE pipeline
    - Deployment di#98 (discuss the priority) - information: move to GitHub
      Hosted on psnc-gitlab - https://gitlab.pcss.pl/dl-team/aggregation/dace - so not affected by gitalb-gwdg migration
    - Status of the "new" DACE developments
    - Continuous ingest di#101 & handle records update di#89

- Sources mapping and ingestion
    - CLARIN resource di#58
        - Keyword mess: Alex to prepare a list of useful values tokenised - then Ola could use the list and ignore the rest OR new column in the source csv. Depending on Ola's assessment, we can also ignore it.
            - Skip keywords? Pre-processing before dace? Or white list
            - White list from Alex => YES, needed in any case.
            - Option to remove bad keywords in post-processing via notebooks.
        - 1. Ola CLARIN data to ingest on stage instance using the current pipeline
        - 2. Cesare to dev script to remove keywords
        - 3. Alex to provide the white list
        - 4. Ola to adapt the clarin pipeline to "filter" keywords.
            - Better to have this step included in the mapping
        - Re-ingest/continuous needed
        - Ingest the missing sections of the source

16.01 > meeting yes.

Create dedicated issue for Ola to work on the white list on clarin

—--

- EOSC MP updated mapping available and re-ingest needed. First ingest with DACE.
    - (need for a second pipeline to harvest the EOSC "data sources")
- CESSDA training di#67
    - Continuous ingest needed
- TAPoR

    - If no issues with re-ingest of CLARIN and CESSDA - TAPoR as the first automatic
  - TDT
  - Conversion Hub

Issues not important for this meeting (maybe later on)

- (how does it look in the pipeline > already implemented?) Custom values mapping (di#80)
- (not necessary to discuss now) Disambiguation of actors (di#77)
- (not necessary to discuss now) Candidate concepts (di#78) > only makes sense to test it after the clean version of the keyword vocab
- (not urgent) Dealing with leading whitespaces during ingest (di#100)

2022-11-29, 15h

Klaus, Laure

https://dariah.zoom.us/j/85839056322?pwd=amRyekNrQlVJRm1DczhZTXNwMFZadz09

- Prep. clean/prioritise DACE related issues
- EOSC MP new mapping
  - Discrepancies between profiles - https://wiki.eoscfuture.eu/display/PUBLIC/B.+v4.00+EOSC+Resource+Profile - and API outcomes

2022-07-26, 13h cest

Participants: Klaus, Ola, Laure

https://dariah.zoom.us/j/86881011583?pwd=aXYwK0pkSmd4U1Jwd0VPT2ZaUWdVUT09

Suggested agenda

Finalised - no need to discuss

1. Ingested via PoolParty - no continuous ingest - no need for a DACE pipeline
   a. SSK
   b. SSK Zotero
   c. SSHOC service catalogue (SSHOC website)
2. Ingested via DACE - no continuous ingest
   a. Humanities Data (missing on SSHOC sources configuration but no need for continuous ingest, so just adding for the sake of documenting)

Ongoing

3. Ingested via DACE - need for continuous ingest / update of the ingest at some point
    a. DH Publications from dblp di#8 > plan another ingest/update once DH2022 conf abstracts will be added to dblp
    b. DARIAH-Campus di#19 > plan another ingest when possible (before end of 2022?)
    c. DARIAH contrib tool di#74 > possibly no continuous ingest (tbc)
4. Ingested via PoolParty - continuous ingest required - need for a DACE pipeline
    a. TAPoR di#84 > review of data ingest on dev > continuous ingest directly on prod?
        i. Continuous ingest with identifiers at source
        ii. need for action from reviewer: Q for Klaus - property statistic notebook queries only "approved" items, correct? To do Laure > check again the notebooks
            1. Also check if identifiers are the same. And check if items look the same on dev than on prod. 2 items on each side. Klaus and laure
        iii. Klaus to give Ola the dump of the db for prod. to check before
    b. EOSC Catalogue and Marketplace di#2
        i. Data model changed at source (be sure about the identifiers)
    c. Programming Historian - no emergency - history available di#5 (PP) need a DACE issue
        i. Pbm with languages/translations and relations to pay attention to with DACE ingest. Check how we could deal with that within DACE.
    d. CLARIN Switchboard tools - di#6 (PP) and di#83 (DACE)
        i. Status? To check with Natalia
5. DACE ingest - no previous ingest done
    a. CESSDA Training Resources di#67
        i. Links changed - ready to be processed again
    b. CLARIN Resource Families di#58
        i. Alex side to change something
    c. TDT di#64 > waiting for API endpoint
    d. Conversion Hub > need Gitlab issue, mapping, API endpoint

=>> continuous ingest relies on identifiers at source. Apparently still a few issues to solve, when identifiers URIs with spec. characters (CLARIN resources families). As long as identifiers at source are stable, it should be fine. Only tested on dev/staging

⇒ where to see the sources already processed by a specific pipeline - could be useful to create a new user and to have a DACE user "dace_importer" - Klaus to do

⇒ ingest workflow:

- In the case of continuous ingest: 1/ ingest first on stage as "ingested" with "dace_importer" (=system_importer) user role. 2/ ingest on the prod instance as "approved" with moderators role

- In the case of new ingest: 1/ ingest first on stage as “ingested” with “dace_importer” (=system_importer) user role. 2/ingest on the prod instance as “ingested” with “dace_importer” (=system_importer) user role ; 3/ have a bulk approval via notebook ;

Priority order > see mappings - sources to MP data model ToC column G - 1st 3 ones

1. **CESSDA Training Resources di#67**
2. **CLARIN Resource Families di#58**
3. **EOSC Catalogue and Marketplace di#2**
    a. More complicated - because nothing in place in DACE
    b. we talk about this endpoint: https://providers.eosc-portal.eu/api/resource/all?
    c. We don’t know how the filter looks like that Sotiris applied, we need to analyse how we came to the current ingest (it is limited to SSH domain)

Organisation post-project

6. Contract between DARIAH and PSNC to support this line of work (1-2 new sources per year)
7. Communication/meetings
8. DACE deployment at ACDH di#98
    a. The end is near :)
9. Continuous ingest workflow ok? mp#101
    a. After the three priority sources > testing it with Tapor
10. Still lacking a proper ingest workflow A to Z even though rationale described in D7.2 p. 28 and foll. - open questions mp#84 & mp#89 (bulk actions implementation postponed)
    a. Can we see which pipeline has been used in the MP data?
    b. Where to ingest for tests (dev or stage instance)?
    c. Which status for the items? (ingested vs. approved)
    d. Running stats on tests ingest - I (Laure) didn’t find a way to query ingested items via the API
11. Where to further document the whole workflow (from manual mapping to rules in the pipeline (cf. di#77 and #78))? Options: static pages, sshoc gitlab, ingest repo readme, dace sshoc wiki?
    a. Tasks via gitlab
    b. Static pages to describe the workflow
    c. DACE installation on the DACE
    d. Sshoc sources described on the sshoc dace wiki
12. Status
    a. di#89
    b. di#80