SeaDataCloud

WP9 - CDI Replication: Status of the new ingestion process

Peter Thijsse on behalf of WP9 team, GA Barcelona
8/9 Nov 2018

sdn-userdesk@seadatanet.org – www.seadatanet.org

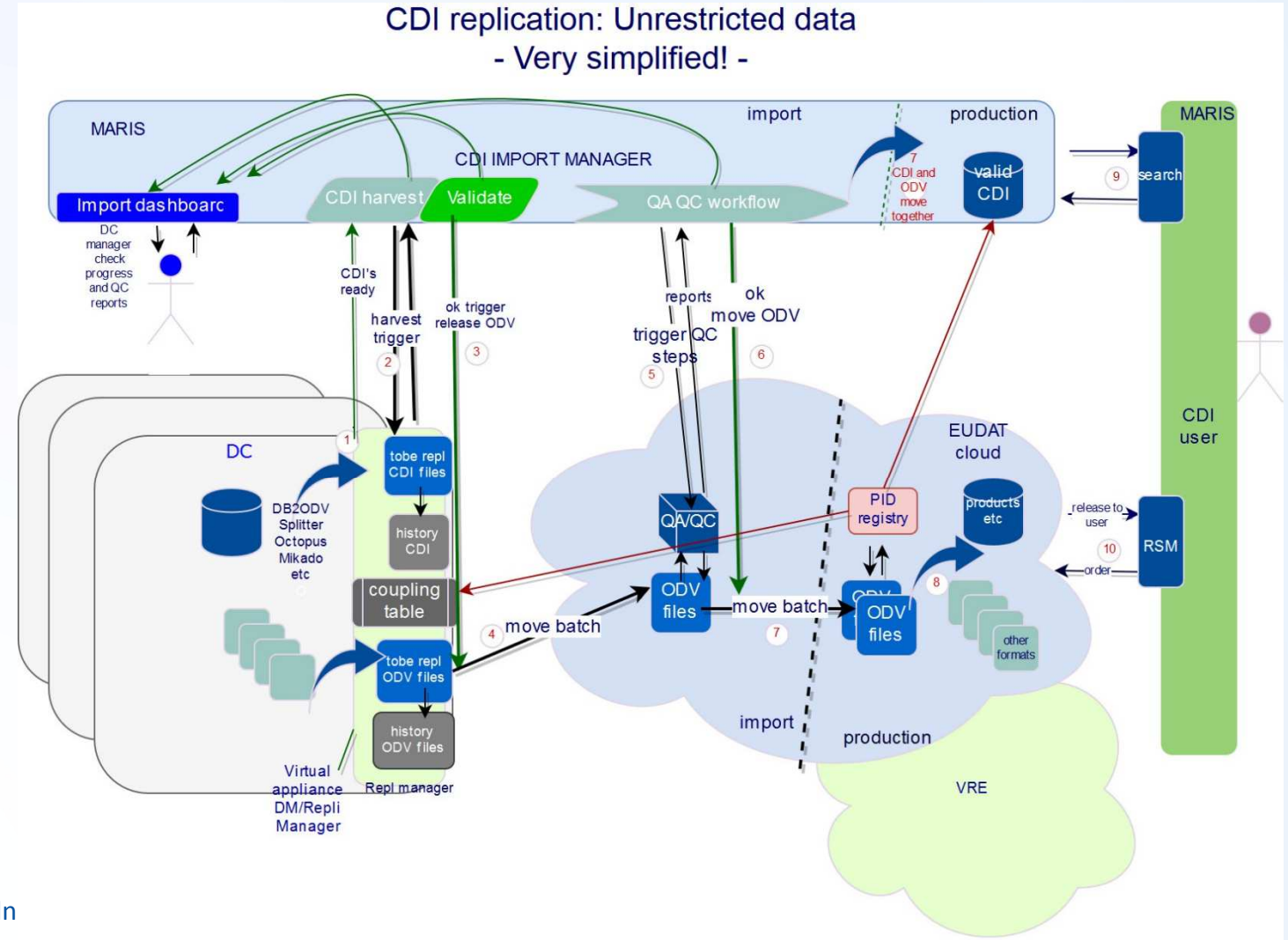
SeaDataCloud

Main components

- Replication manager(s)
- Import manager
- EUDATs HTTP API
- B2SAFE (Cloud storage)
- Authentication B2ACCESS => B2SAFE local accounts

SeaDataCloud

Overview unrestricted dataflow



Remarks unrestricted

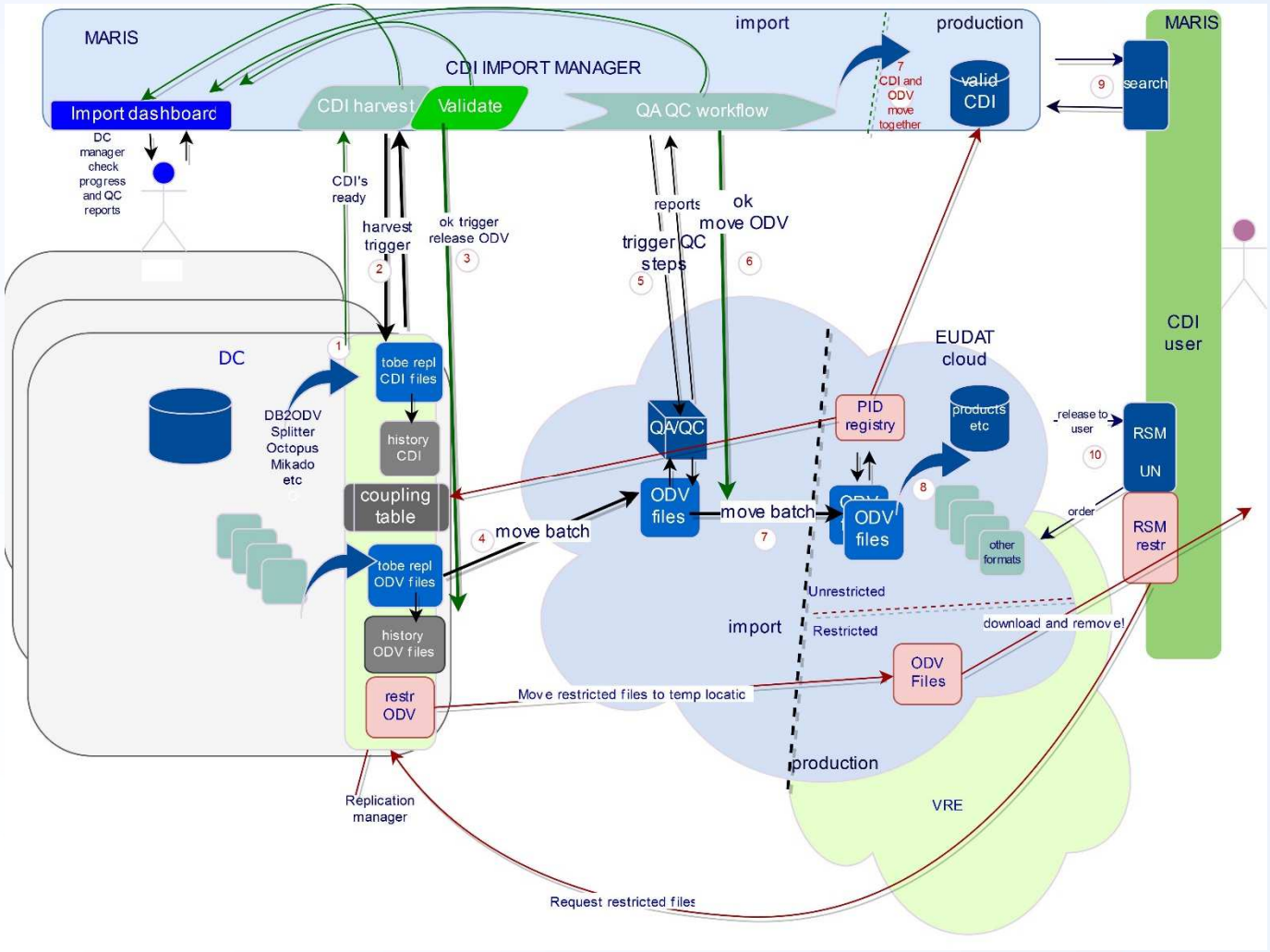
The main solution is designed for Unrestricted (UN/LS) data where the process of QA/QC and file format conversions in the cloud are involved.

Important steps:

- The CDI import manager is guiding the process of CDI and ODV release and QA/QC towards production.
- At data center there is production of CDI and ODV (or one other accepted exchange data format) using the stack of tools produced at IFREMER: DB2ODV, Splitter, Mikado, Octopus etc.
- Replication Manager at the Data Center is triggered by import manager to:
  - Move a batch of CDI files to CDI central server
  - Follow up requests to move a batch of ODV files triggered by import manager. One batch at a time handled by the system.
- After moving a batch it is archived by the RM (history of full situation).

- The EUDAT cloud consists of a area for Import and an area of Production. During import QA/QC processes will be triggered by the CDI import manager after which reports flow back. After an ok the datasets will be made available in production. Again moved after a trigger from the import manager.
- When a dataset moves to production the PID is reported back to import manager and DC (coupling table)
- The DC data manager (and masters) can follow progress and if necessary take actions to get to a next stage, or e.g. discard the set and release a new updated batch.

Overview restricted dataflow



Remarks restricted

Replication Manager at the Data Center is for restricted data triggered by RSM to:

- Follow up requests for releasing and moving a set of restricted data
  - Only when allowed by the DC via confirmation in the RSM
  - Directly to a “user section” in a restricted part of the cloud where it will be available to the user to download it for limited time.
  - After download it will be removed.

Replication of RS data is a sensitive step:

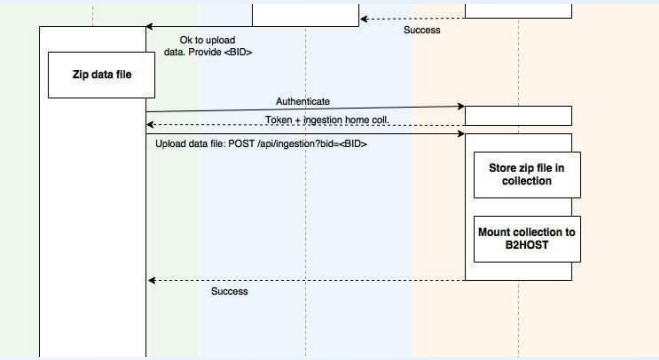
- Trust is needed from the DC in the secure part of the cloud to temporarily store the data
- But, working with the same “trigger and move” process/protocols for each release of data makes it simpler
- The set will be removed in the cloud once downloaded, so it is ready for the next batch.

6 steps in the unrestricted data ingestion process

1. Authentication
2. Publishing metadata
3. Checking metadata
4. Uploading a databatch
5. Unzip and run data checks
6. Move validated metadata and data in production

Step 1: Authentication

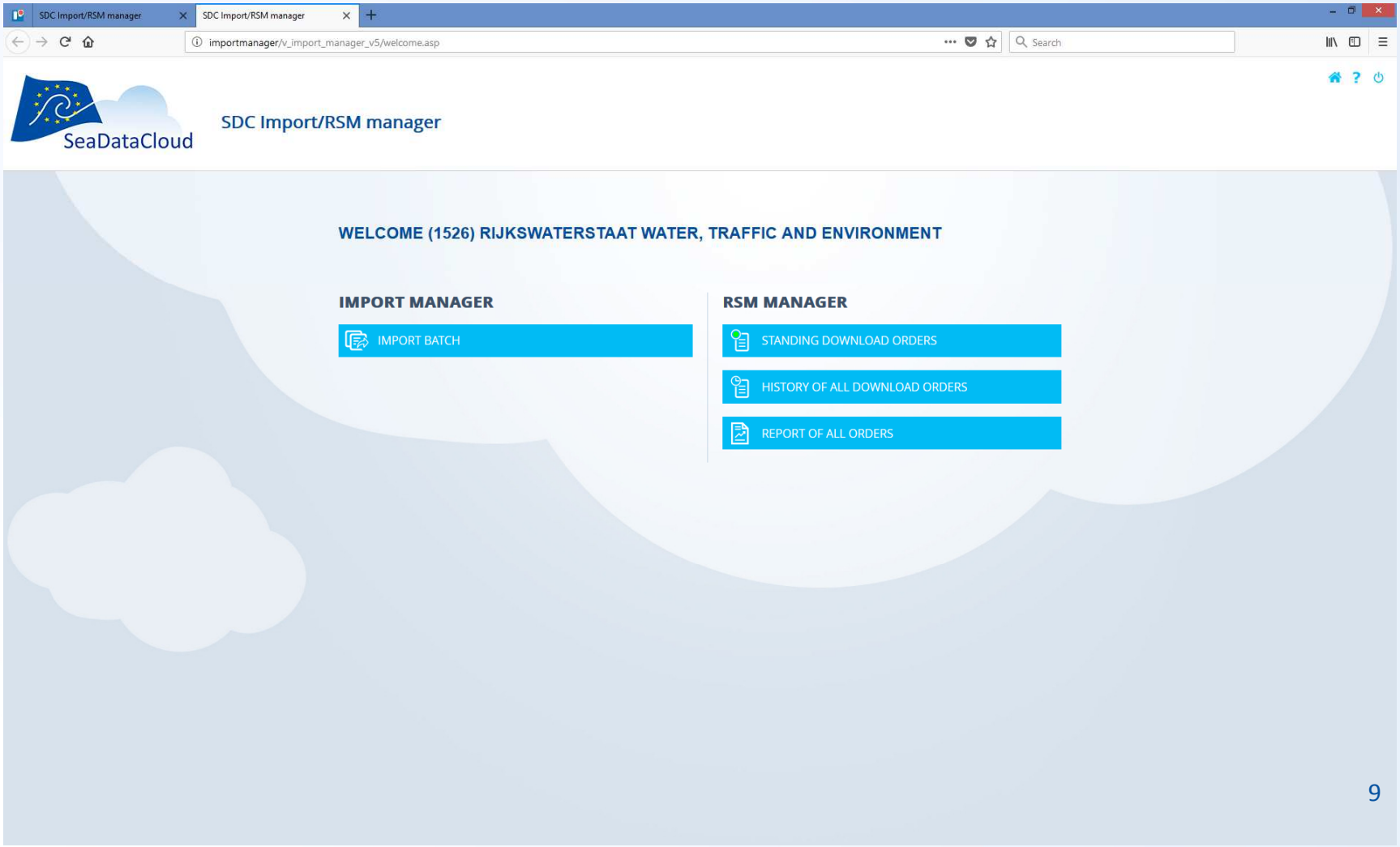
The problem:
RM needs to access the Import Manager and the EUDAT B2STAGE HTTP-API through a software (RM).
Data center contacts needs to access the IM dashboard.



The solution

1. Every RM registers with a B2SAFE local account. MARIS prepared neutral “@sdn-taskmanager.org” addresses. These are forwarded to the RSM contacts marineID’s.
2. DC will configure the RM with the credentials
3. RM calls /auth/b2safeproxy with credentials
4. Data center contacts linked to the IM via their MarineID can contact the IM dashboard

Login to the Import Manager for process overview



Step 2: Publishing the CDI metadata

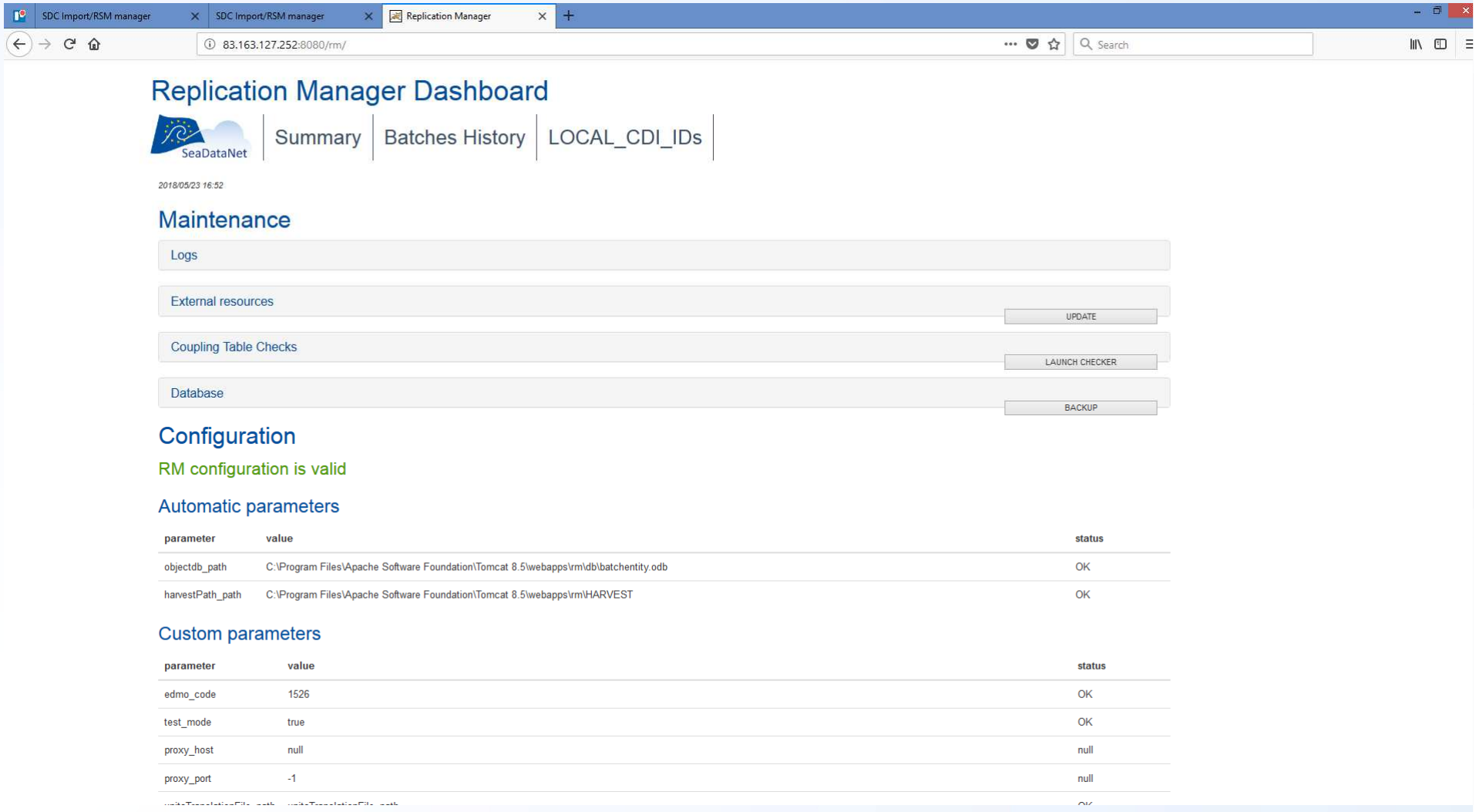
The problem

- RM needs to notify the Import Manager a new batch is ready
- IM will download when ready

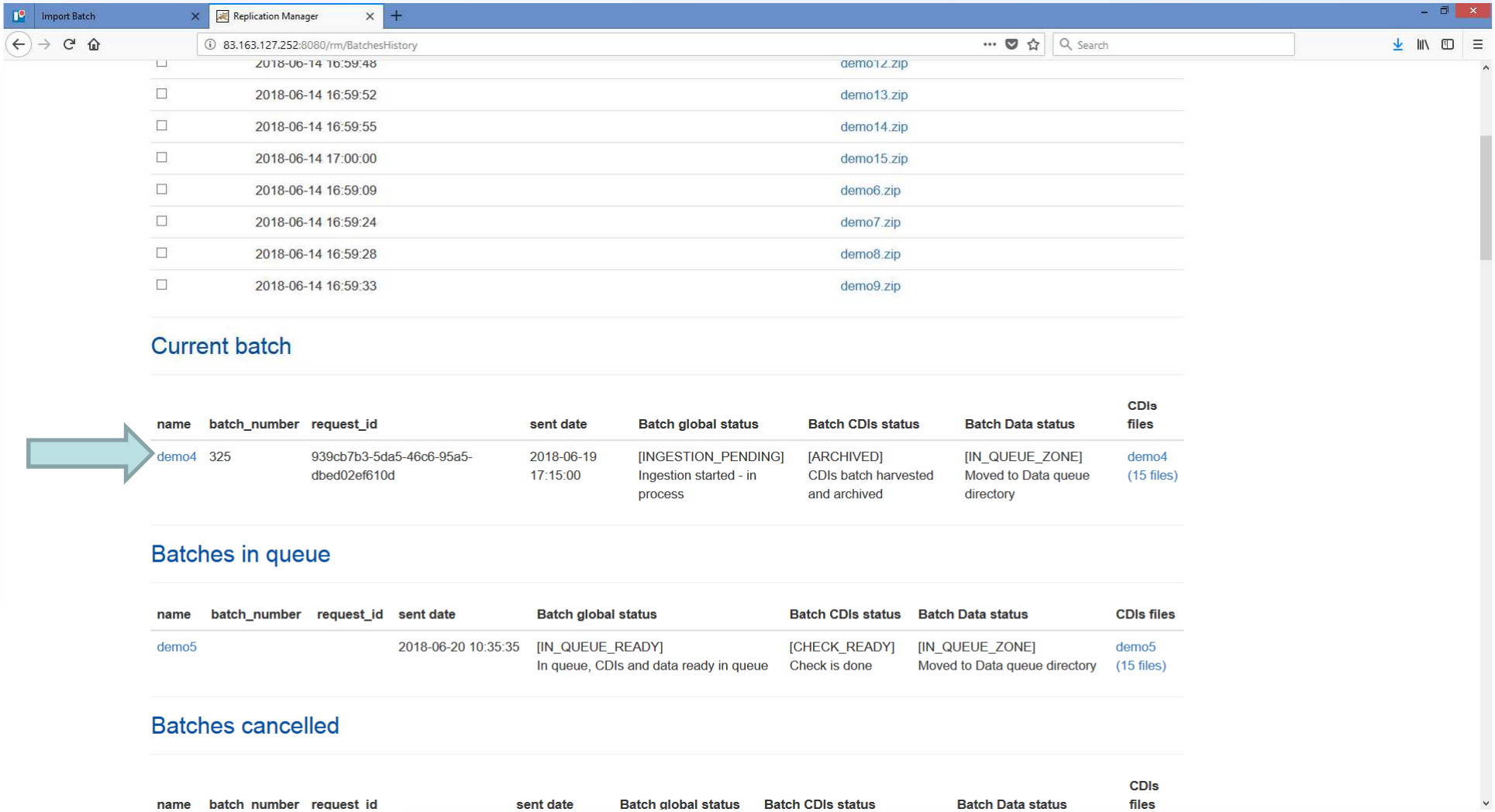
The solution

1. DC manager puts one or more batches in publish-directory
2. RM triggers
3. IM downloads the CDI metadata batch
4. One batch at a time in the process

View of RM dashboard



RM: Batch received by IM - Now indicates batchnumber



Step 3: Checking the CDI metadata

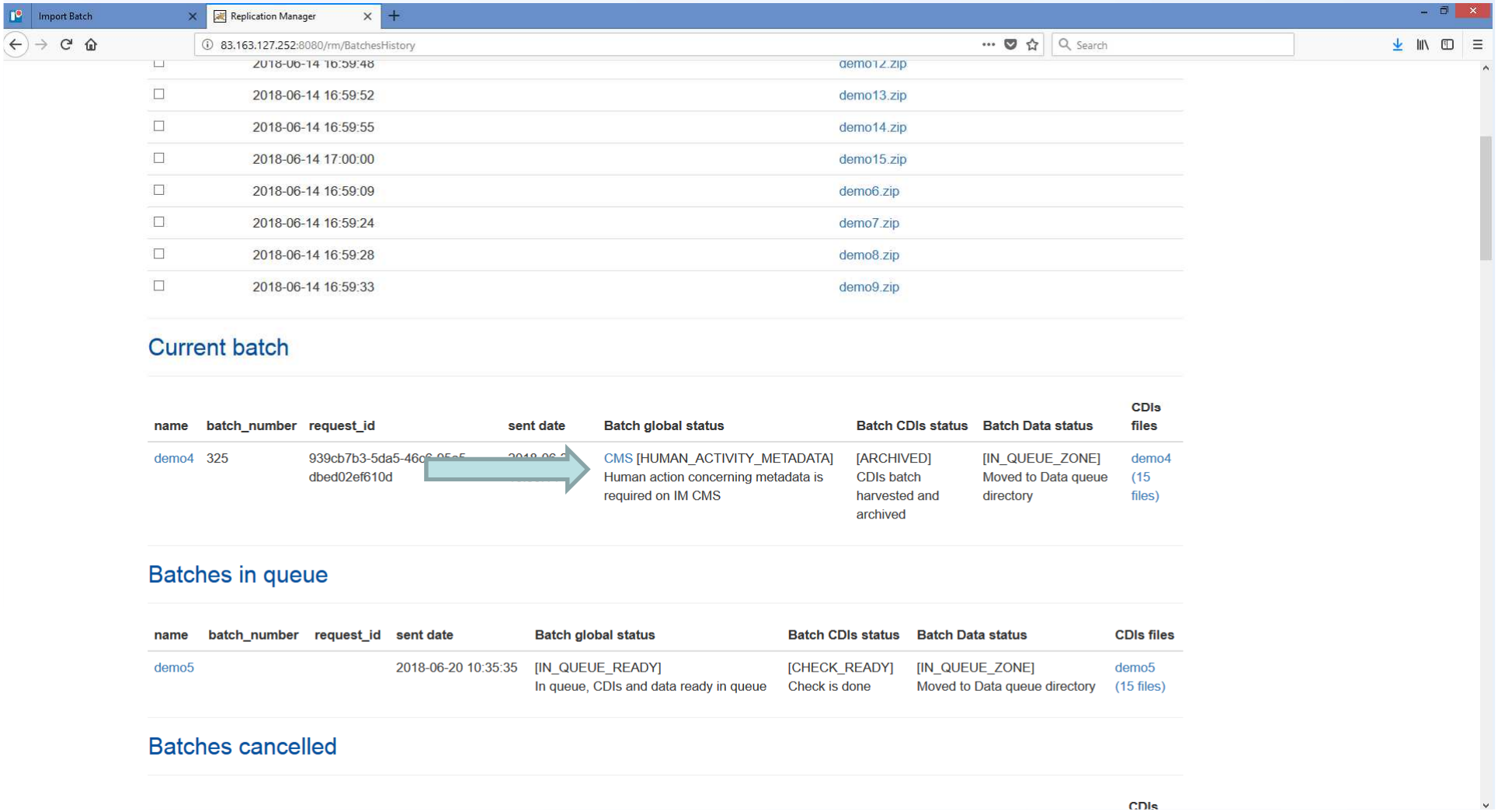
The problem

- Import Manager to check the CDI XML and its content
- RM manager to check also visually

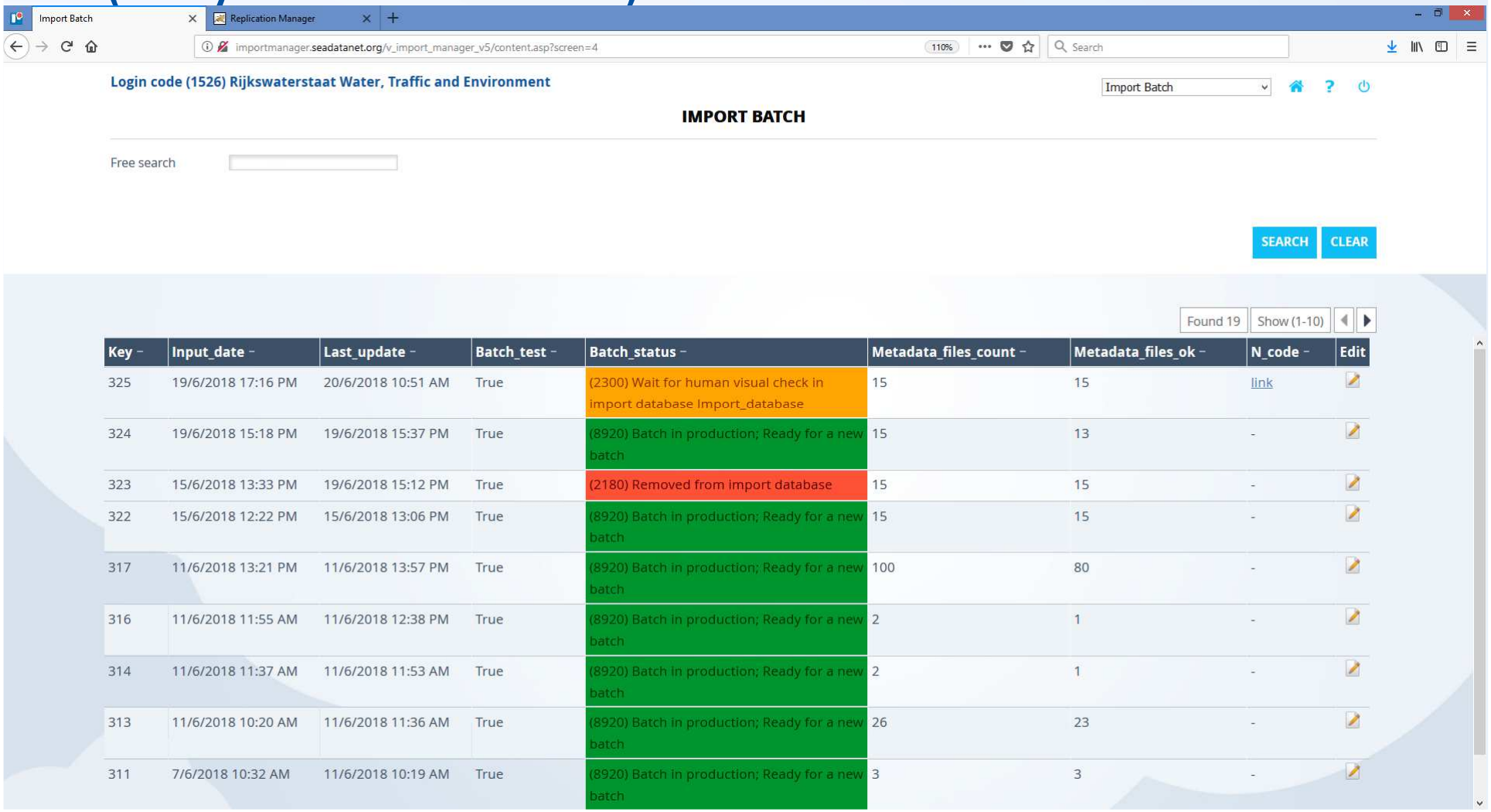
The solution

1. Unzip the batch
2. Run sequence of metadata checks (same as current)
3. All logs added to IM CMS
4. Metadata loaded to import database
5. RM notified visual check in import is needed
6. RM manager decides which files continue to next steps

RM: Notification calls datamanager to action

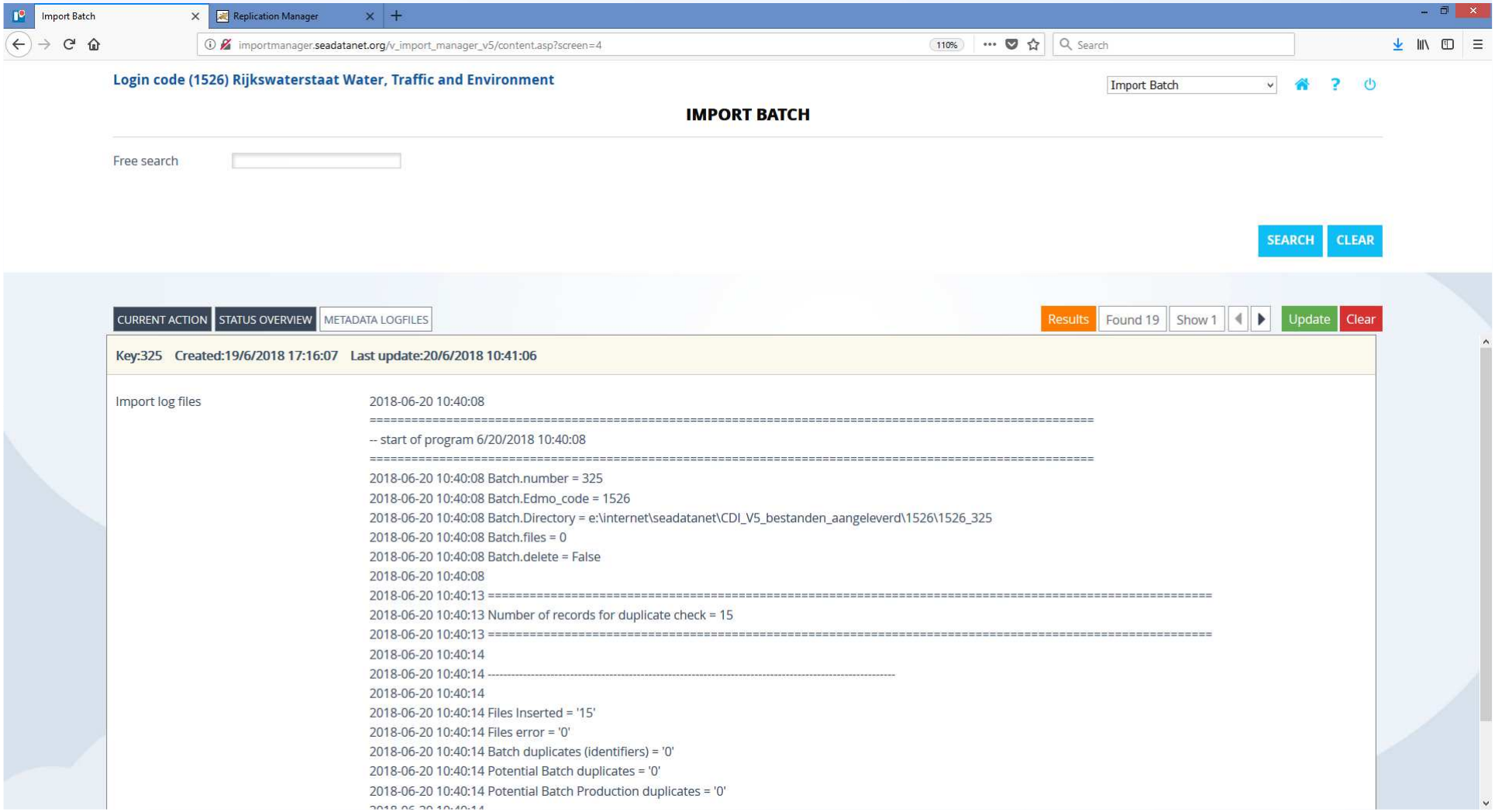


IM: Ready to visually check logs and interface (only human action)

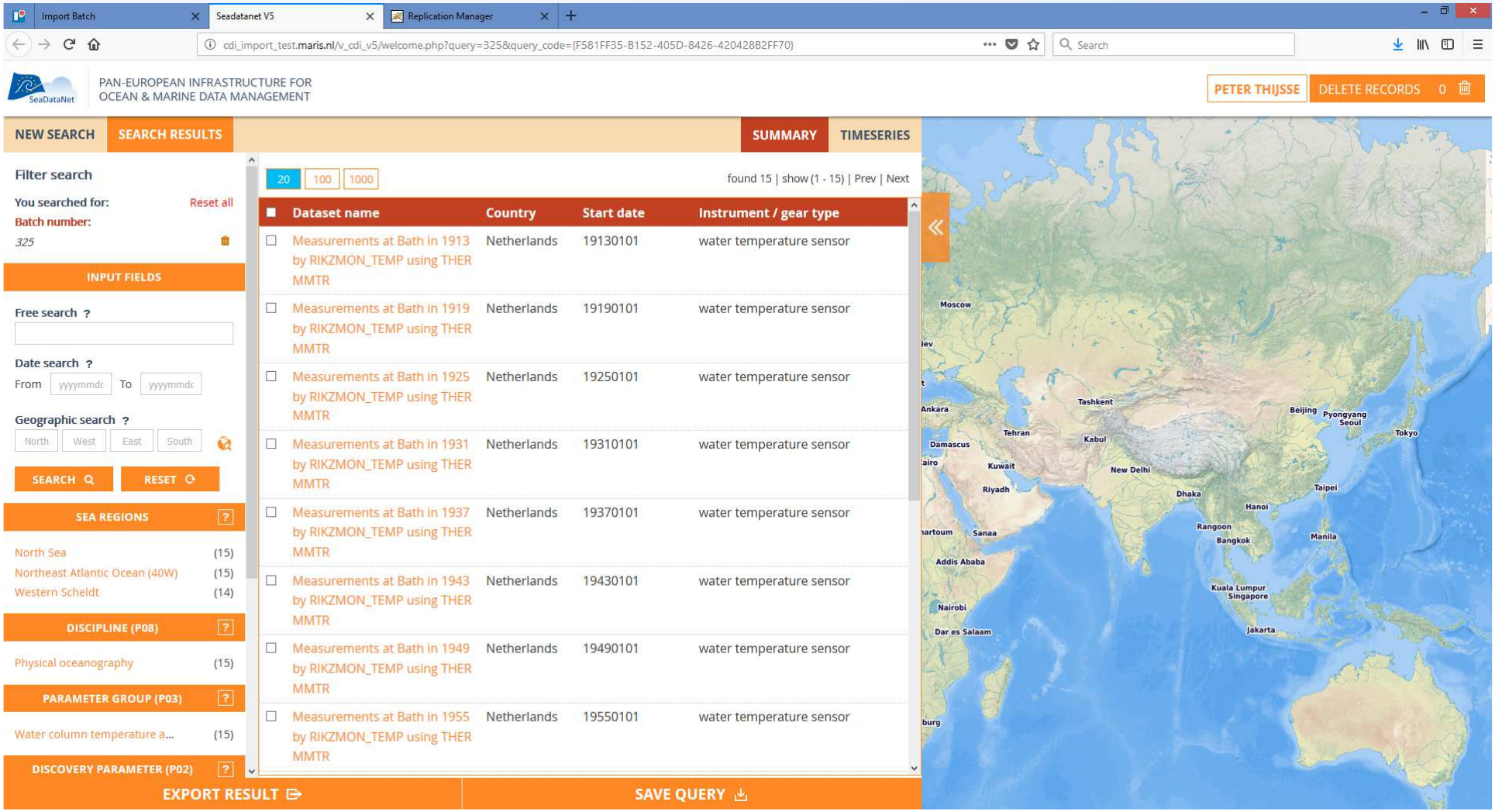


Key	Input_date	Last_update	Batch_test	Batch_status	Metadata_files_count	Metadata_files_ok	N_code	Edit
325	19/6/2018 17:16 PM	20/6/2018 10:51 AM	True	(2300) Wait for human visual check in import database Import_database	15	15	link	
324	19/6/2018 15:18 PM	19/6/2018 15:37 PM	True	(8920) Batch in production; Ready for a new batch	15	13	-	
323	15/6/2018 13:33 PM	19/6/2018 15:12 PM	True	(2180) Removed from import database	15	15	-	
322	15/6/2018 12:22 PM	15/6/2018 13:06 PM	True	(8920) Batch in production; Ready for a new batch	15	15	-	
317	11/6/2018 13:21 PM	11/6/2018 13:57 PM	True	(8920) Batch in production; Ready for a new batch	100	80	-	
316	11/6/2018 11:55 AM	11/6/2018 12:38 PM	True	(8920) Batch in production; Ready for a new batch	2	1	-	
314	11/6/2018 11:37 AM	11/6/2018 11:53 AM	True	(8920) Batch in production; Ready for a new batch	2	1	-	
313	11/6/2018 10:20 AM	11/6/2018 11:36 AM	True	(8920) Batch in production; Ready for a new batch	26	23	-	
311	7/6/2018 10:32 AM	11/6/2018 10:19 AM	True	(8920) Batch in production; Ready for a new batch	3	3	-	

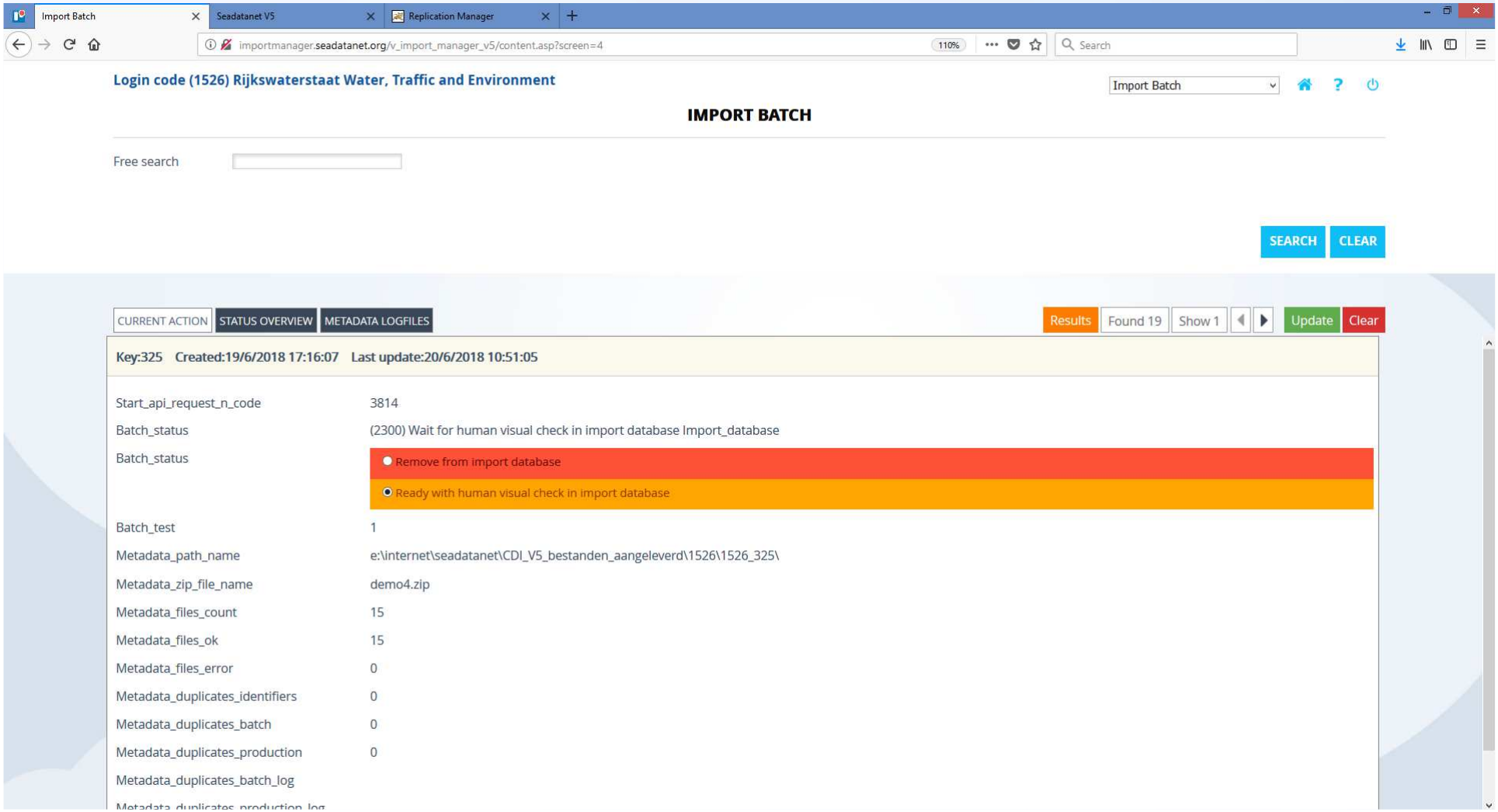
IM: Status progress and logs XML validation available



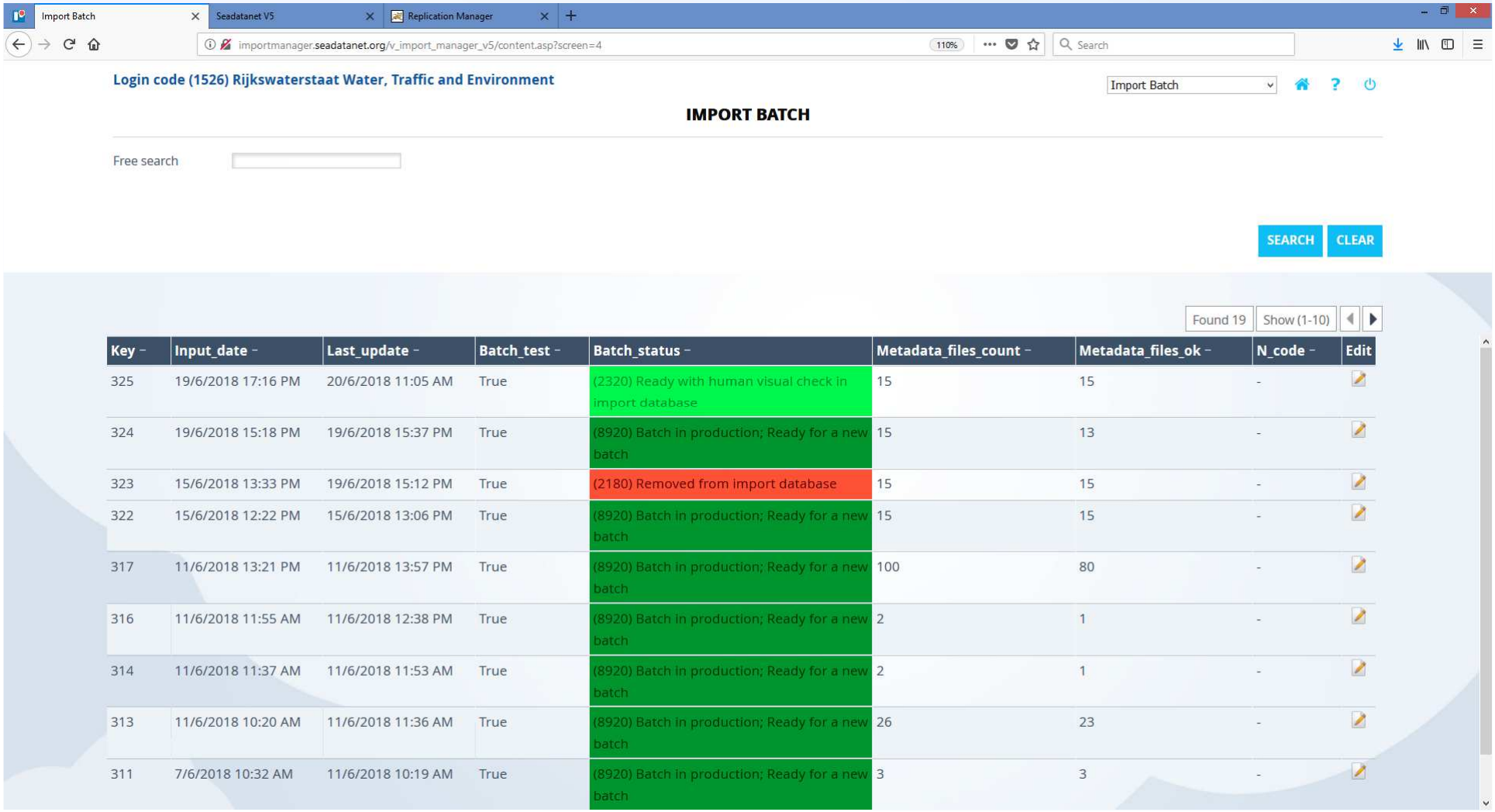
IM: Visual check - Individual records can be removed



IM: RM manager indicates check is done (go on or cancel batch)



IM: Confirmed and programmes take over again



Key	Input_date	Last_update	Batch_test	Batch_status	Metadata_files_count	Metadata_files_ok	N_code	Edit
325	19/6/2018 17:16 PM	20/6/2018 11:05 AM	True	(2320) Ready with human visual check in import database	15	15	-	
324	19/6/2018 15:18 PM	19/6/2018 15:37 PM	True	(8920) Batch in production; Ready for a new batch	15	13	-	
323	15/6/2018 13:33 PM	19/6/2018 15:12 PM	True	(2180) Removed from import database	15	15	-	
322	15/6/2018 12:22 PM	15/6/2018 13:06 PM	True	(8920) Batch in production; Ready for a new batch	15	15	-	
317	11/6/2018 13:21 PM	11/6/2018 13:57 PM	True	(8920) Batch in production; Ready for a new batch	100	80	-	
316	11/6/2018 11:55 AM	11/6/2018 12:38 PM	True	(8920) Batch in production; Ready for a new batch	2	1	-	
314	11/6/2018 11:37 AM	11/6/2018 11:53 AM	True	(8920) Batch in production; Ready for a new batch	2	1	-	
313	11/6/2018 10:20 AM	11/6/2018 11:36 AM	True	(8920) Batch in production; Ready for a new batch	26	23	-	
311	7/6/2018 10:32 AM	11/6/2018 10:19 AM	True	(8920) Batch in production; Ready for a new batch	3	3	-	

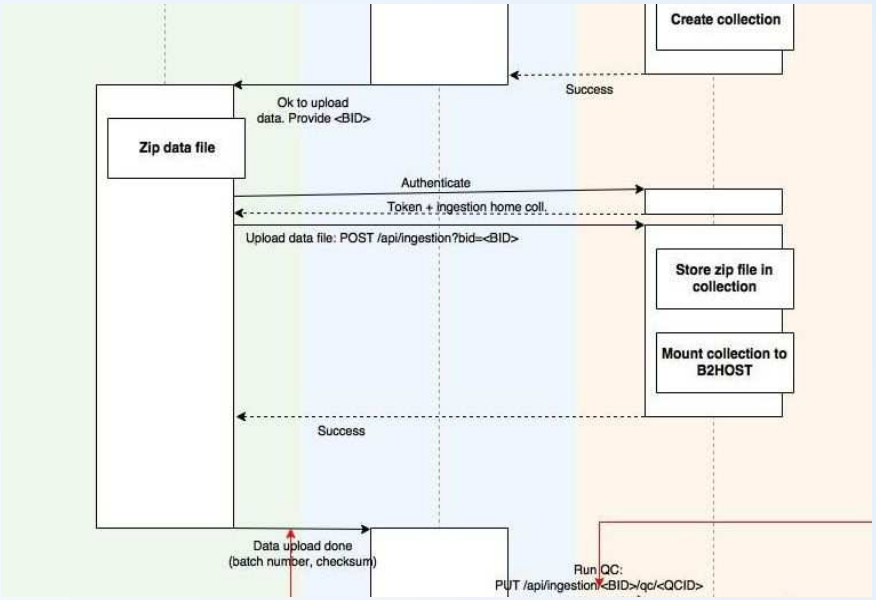
Step 4: Uploading a databatch (Unrestricted only!)

The problem

IM requests to RM to upload specified datafiles (matching CDI's with datafiles). RM has one zip file with data (!) to upload

The solution

1. RM holds files in a folder
2. RM receives from IM a Batch ID (BID), CDI_identifiers, and formats to use for upload
3. IM creates BID and creates the virtual folder on cloud
4. RM uploads the zip file to the folder in cloud
5. HTTP API confirms upload completed to RM
6. RM notifies Import Manager that upload is ready.



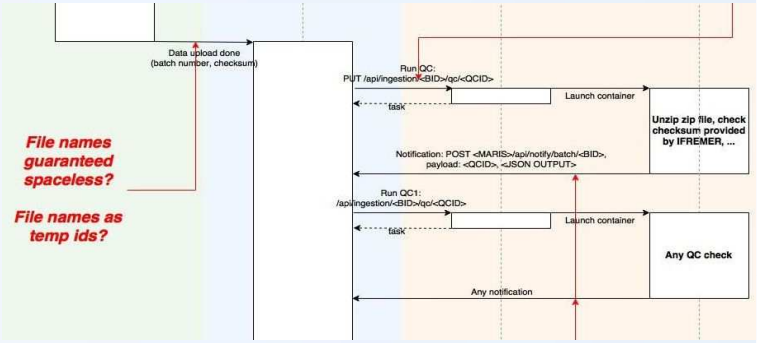
Step 5: Unzip and run datachecks

The problem

QC developer wants to run an action (unzip/quality check) on every uploaded batch

The solution

1. QC Developer writes a program to test each file of a specific folder
    1. QC developers pushes the container for this binary to EUDAT HUB
    2. QC identifier available in the end (QCID)
2. Import Manager calls PUT on /api/ingestion/BID/qc/QCID
3. HTTP API runs an asynchronous task T for QCID mounting BID folder
4. T ends and calls POST on MARIS API with:
    1. The JSON output from the task/JAVA code incl QCID
5. Next action



Technical QC implementation status

- Checks implemented:
  - Compare checksum of file
  - Compare length of file
  - Unzip file
  - Compare nr of files
  - Check length of files > 0
  - Check for each CDI_identifier if requested files and formats exist
- All errors reported back to IM, logs are user and shown in CMS.
- Erroneous records removed from batch, but batch moves forward as long as ok records exist.
- No more records, or batch cancelled by data manager, start again.

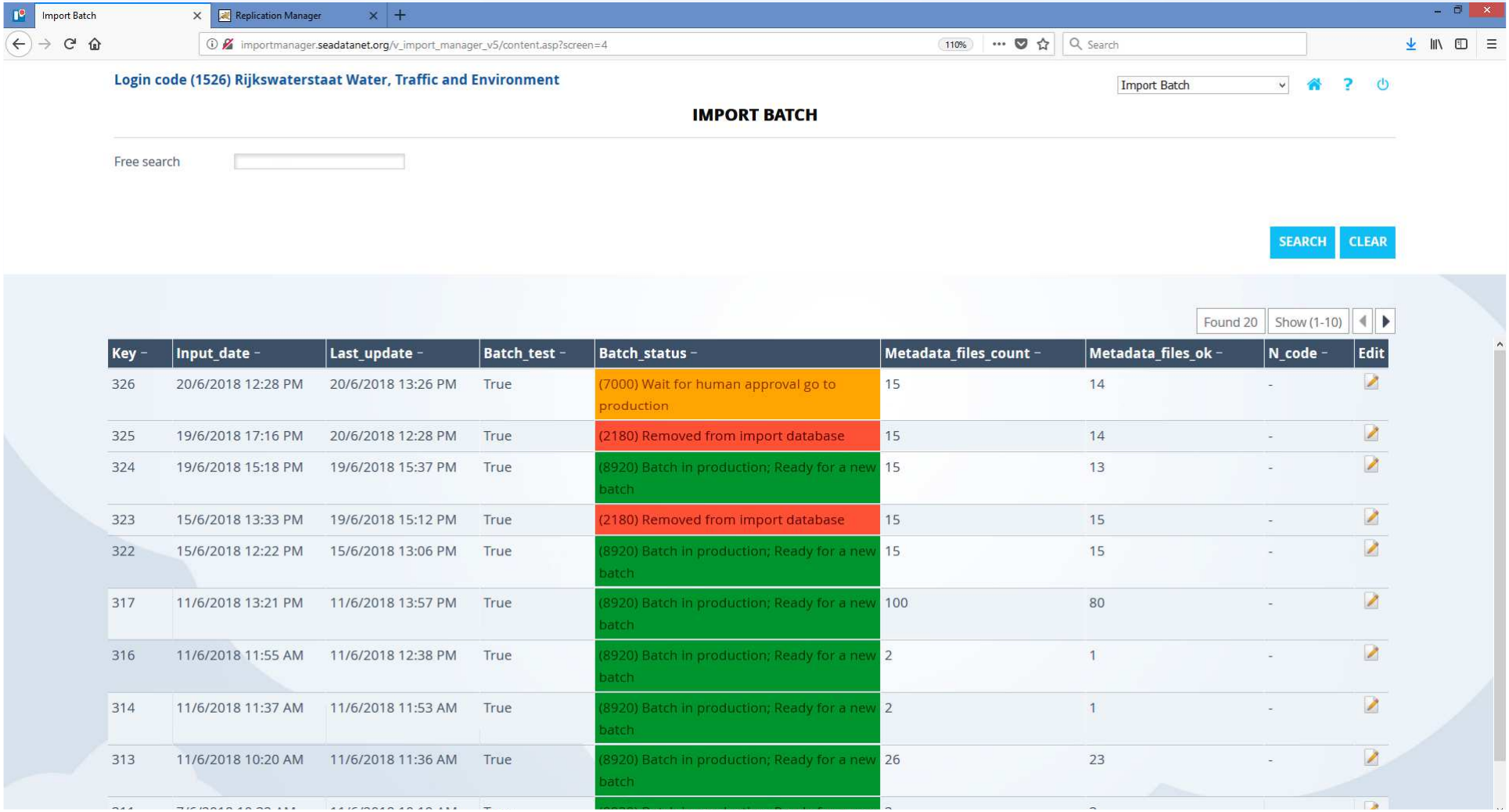
Followed by a sequence of additional “QC checks”

- ODV quality checks
- Conversion to other formats (if not available yet)
- P01 Extraction
- Format syntax check

Facilitated by Octopus, plus additional programmes

SeaDataCloud

IM: Unzip and checks are done, left-over files ready to be moved. RM manager should check logs and confirm

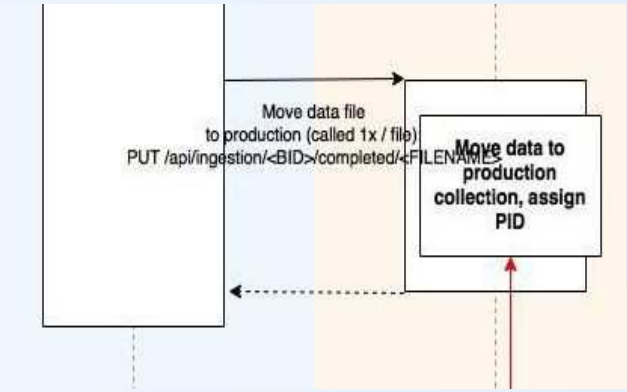


Key	Input_date	Last_update	Batch_test	Batch_status	Metadata_files_count	Metadata_files_ok	N_code	Edit
326	20/6/2018 12:28 PM	20/6/2018 13:26 PM	True	(7000) Wait for human approval go to production	15	14	-	
325	19/6/2018 17:16 PM	20/6/2018 12:28 PM	True	(2180) Removed from import database	15	14	-	
324	19/6/2018 15:18 PM	19/6/2018 15:37 PM	True	(8920) Batch in production; Ready for a new batch	15	13	-	
323	15/6/2018 13:33 PM	19/6/2018 15:12 PM	True	(2180) Removed from import database	15	15	-	
322	15/6/2018 12:22 PM	15/6/2018 13:06 PM	True	(8920) Batch in production; Ready for a new batch	15	15	-	
317	11/6/2018 13:21 PM	11/6/2018 13:57 PM	True	(8920) Batch in production; Ready for a new batch	100	80	-	
316	11/6/2018 11:55 AM	11/6/2018 12:38 PM	True	(8920) Batch in production; Ready for a new batch	2	1	-	
314	11/6/2018 11:37 AM	11/6/2018 11:53 AM	True	(8920) Batch in production; Ready for a new batch	2	1	-	
313	11/6/2018 10:20 AM	11/6/2018 11:36 AM	True	(8920) Batch in production; Ready for a new batch	26	23	-	

Step 6: Move validated data in production

The problem

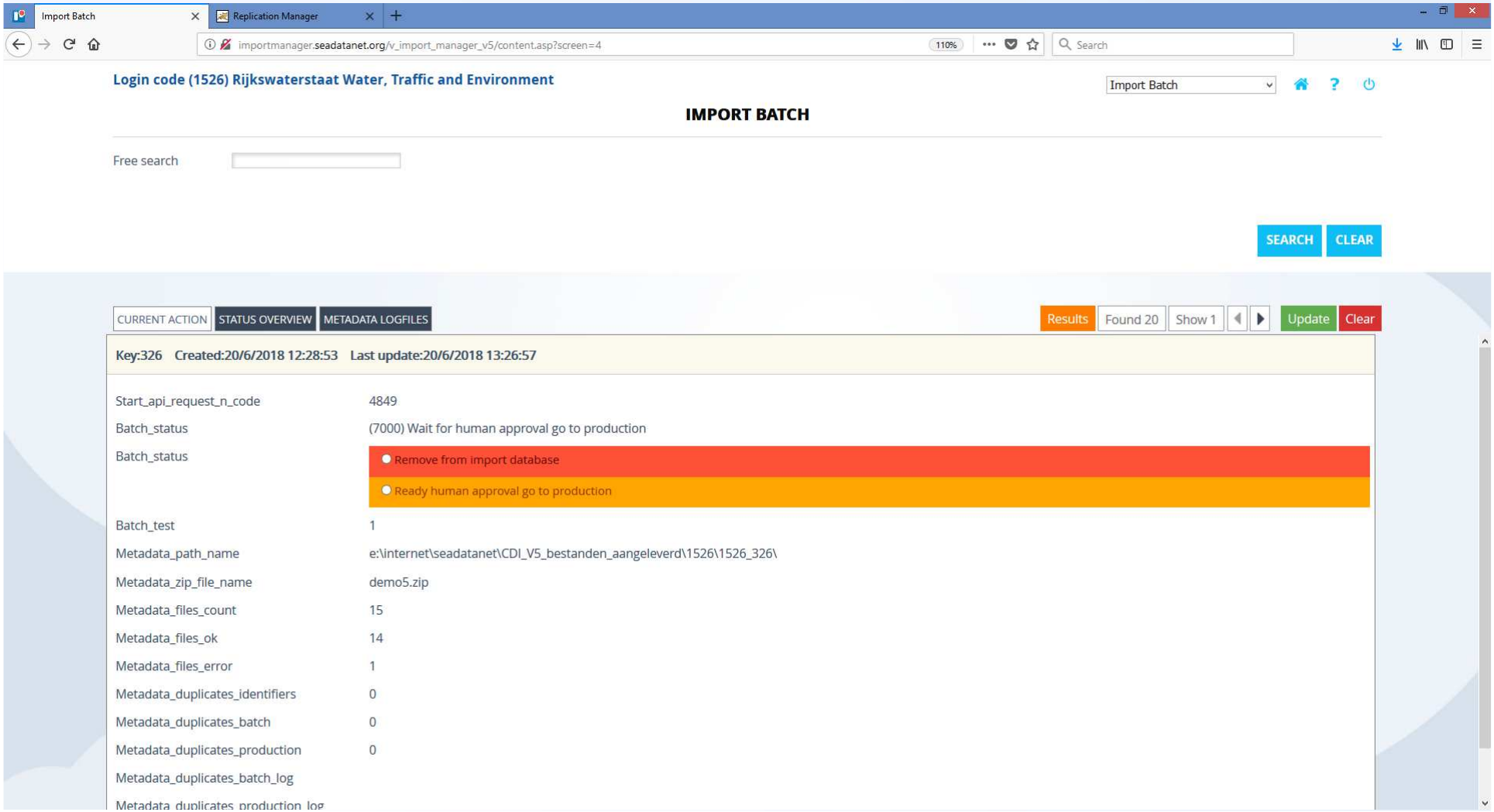
- Import manager has a matrix of Quality checks for the batch
- All batch records that passed all quality checks is to move in production



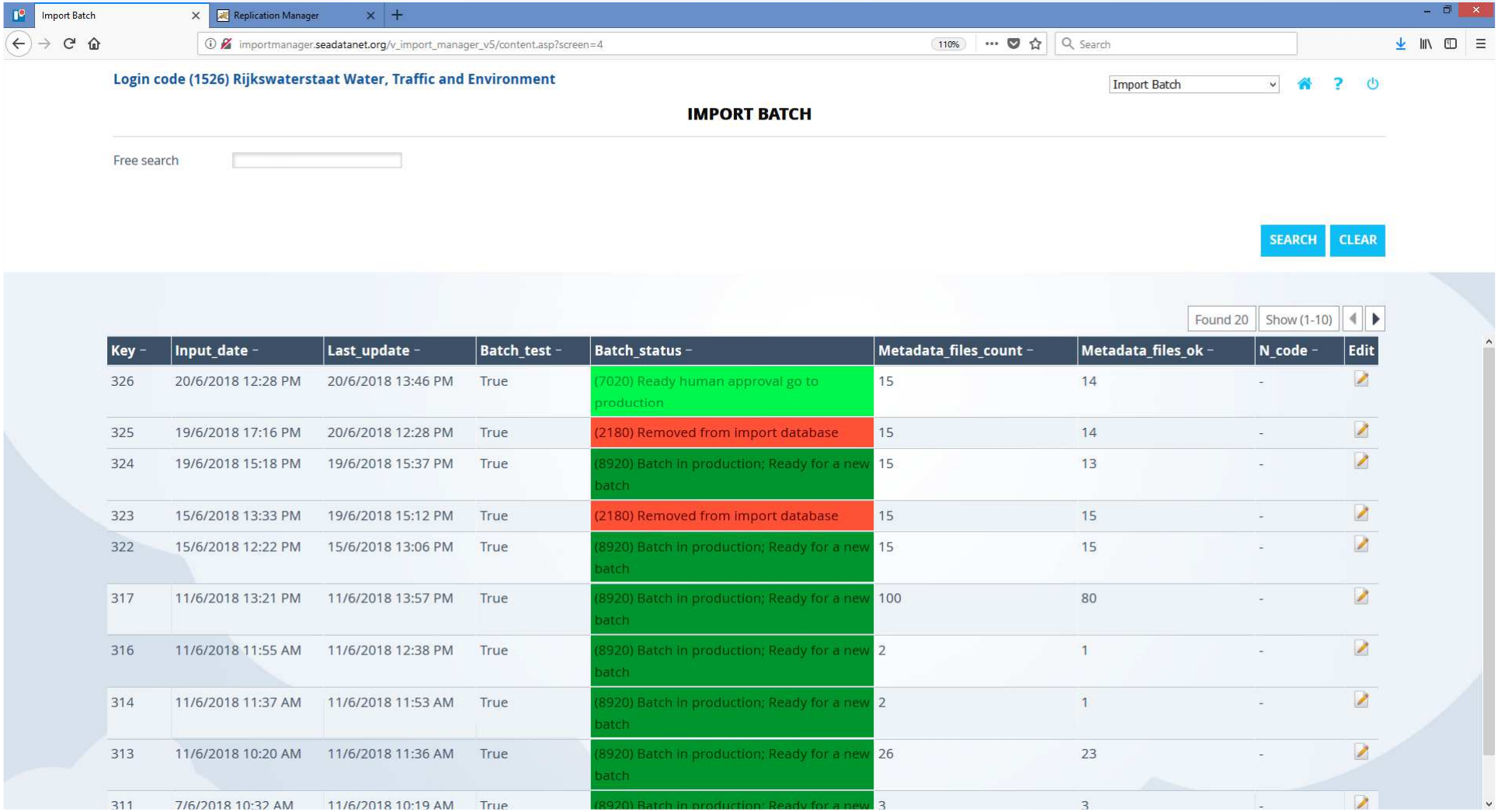
The solution

1. IM calls HTTP API with the filename for each approved file
2. HTTP API asynchronously triggers the ingestion process of IRODS
3. iRODS moves the collection to Production
4. Automatic irules are triggered for each file to generate the PID
5. HTTP API supplies IM a JSON with filenames and their PIDs (per format)

IM: Human approval to move to production

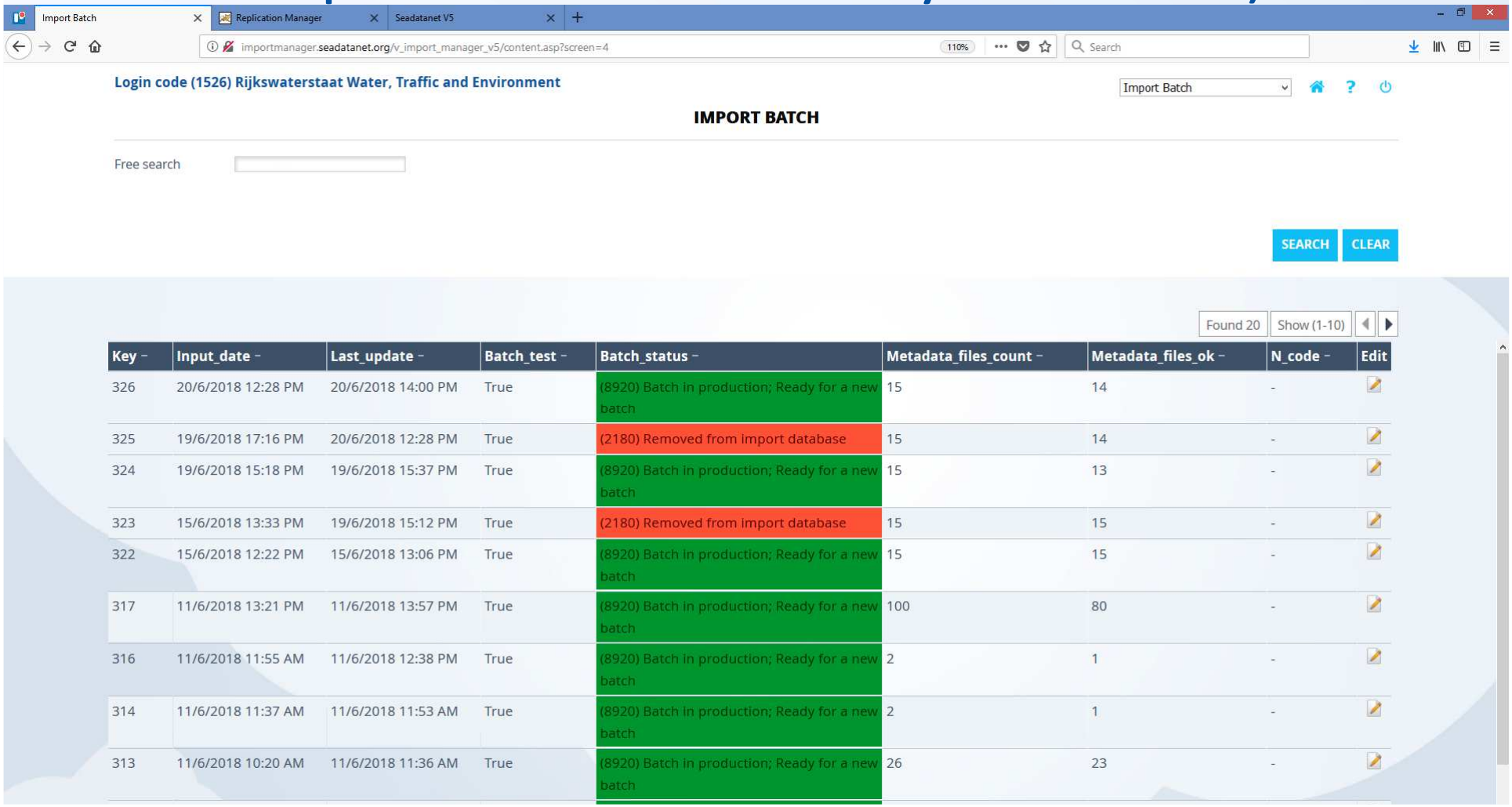


IM: Ready to go for final steps



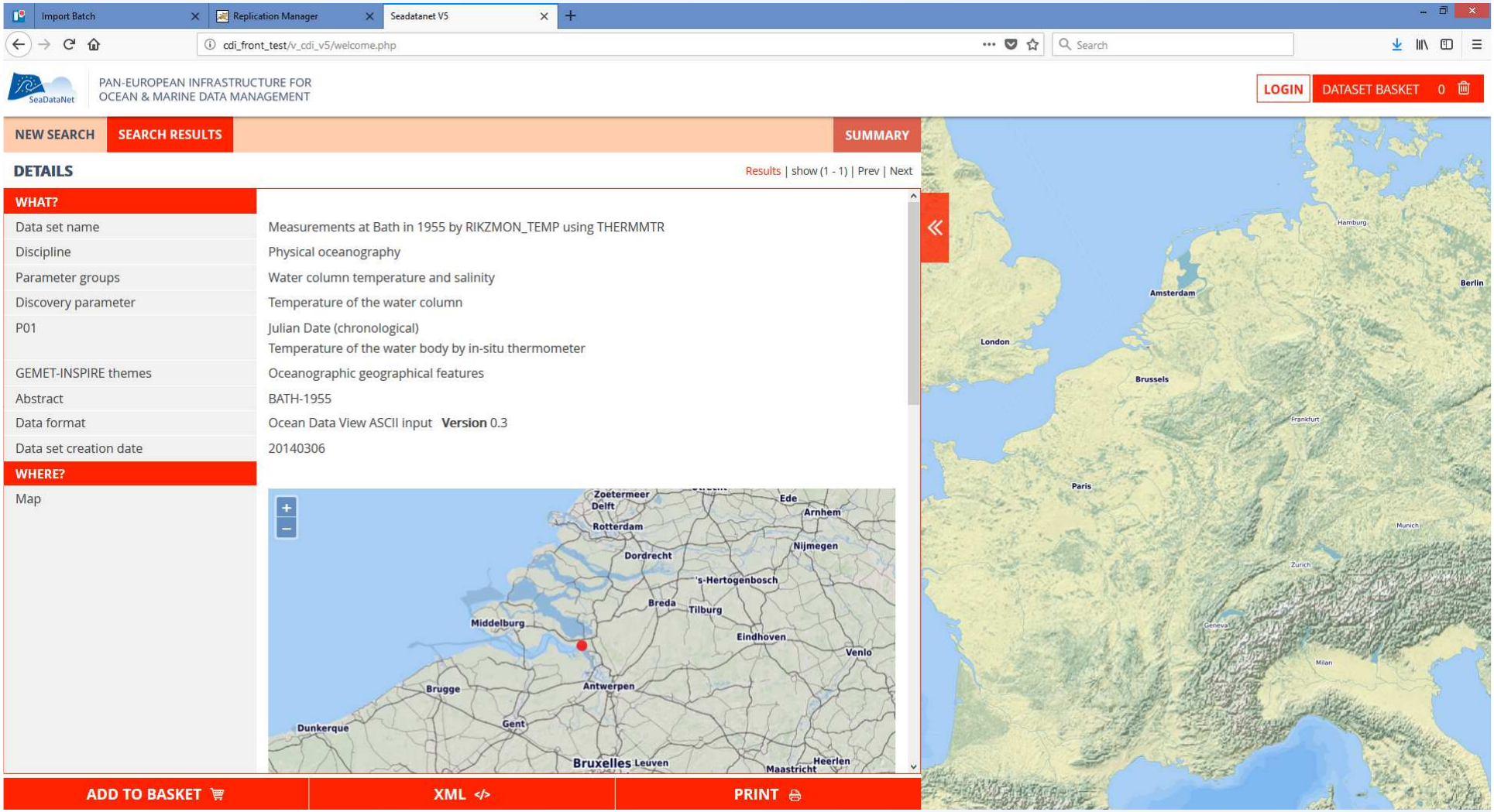
Key	Input_date	Last_update	Batch_test	Batch_status	Metadata_files_count	Metadata_files_ok	N_code	Edit
326	20/6/2018 12:28 PM	20/6/2018 13:46 PM	True	(7020) Ready human approval go to production	15	14	-	
325	19/6/2018 17:16 PM	20/6/2018 12:28 PM	True	(2180) Removed from import database	15	14	-	
324	19/6/2018 15:18 PM	19/6/2018 15:37 PM	True	(8920) Batch in production; Ready for a new batch	15	13	-	
323	15/6/2018 13:33 PM	19/6/2018 15:12 PM	True	(2180) Removed from import database	15	15	-	
322	15/6/2018 12:22 PM	15/6/2018 13:06 PM	True	(8920) Batch in production; Ready for a new batch	15	15	-	
317	11/6/2018 13:21 PM	11/6/2018 13:57 PM	True	(8920) Batch in production; Ready for a new batch	100	80	-	
316	11/6/2018 11:55 AM	11/6/2018 12:38 PM	True	(8920) Batch in production; Ready for a new batch	2	1	-	
314	11/6/2018 11:37 AM	11/6/2018 11:53 AM	True	(8920) Batch in production; Ready for a new batch	2	1	-	
313	11/6/2018 10:20 AM	11/6/2018 11:36 AM	True	(8920) Batch in production; Ready for a new batch	26	23	-	
311	7/6/2018 10:32 AM	11/6/2018 10:19 AM	True	(8920) Batch in production; Ready for a new	3	3	-	

IM: Success! Ready for a new batch! (if next one in queue it automatically continues)



Key	Input_date	Last_update	Batch_test	Batch_status	Metadata_files_count	Metadata_files_ok	N_code	Edit
326	20/6/2018 12:28 PM	20/6/2018 14:00 PM	True	(8920) Batch in production; Ready for a new batch	15	14	-	
325	19/6/2018 17:16 PM	20/6/2018 12:28 PM	True	(2180) Removed from import database	15	14	-	
324	19/6/2018 15:18 PM	19/6/2018 15:37 PM	True	(8920) Batch in production; Ready for a new batch	15	13	-	
323	15/6/2018 13:33 PM	19/6/2018 15:12 PM	True	(2180) Removed from import database	15	15	-	
322	15/6/2018 12:22 PM	15/6/2018 13:06 PM	True	(8920) Batch in production; Ready for a new batch	15	15	-	
317	11/6/2018 13:21 PM	11/6/2018 13:57 PM	True	(8920) Batch in production; Ready for a new batch	100	80	-	
316	11/6/2018 11:55 AM	11/6/2018 12:38 PM	True	(8920) Batch in production; Ready for a new batch	2	1	-	
314	11/6/2018 11:37 AM	11/6/2018 11:53 AM	True	(8920) Batch in production; Ready for a new batch	2	1	-	
313	11/6/2018 10:20 AM	11/6/2018 11:36 AM	True	(8920) Batch in production; Ready for a new batch	26	23	-	

Batch active in interface (example record – in test)



Restricted data (Recently finalised)

- ✓ CDI metadata same flow
- ✓ Data files extracted at same time as metadata. Copy stored (for each version)
- ✓ Files will be temporarily released to cloud directory of user upon request RSM
- ✓ When released: Added all to same order-zipfile in the cloud

Other recent actions completed

- ✓ Updates of RM and IM software
- ✓ Development version still at CINECA
- ✓ More stable test (pre-production) environment with TEST and IMPORT-PUBLIC versions at CSC environment.
- ✓ Risk analysis and completed calls/error responses to make the developed system processes full proof

Ongoing actions – Replication manager

- Check minimum RAM: Now 4GB, ok for 2 installations sid by side (test and production)
- Large file upload test (split if too large)
- Large batches test
- Manual/documentation update
- Consistency/integrity checks - what if files in order/request not available => Monitoring plan
- First stable version release to 5 test partners – expected around 20 November

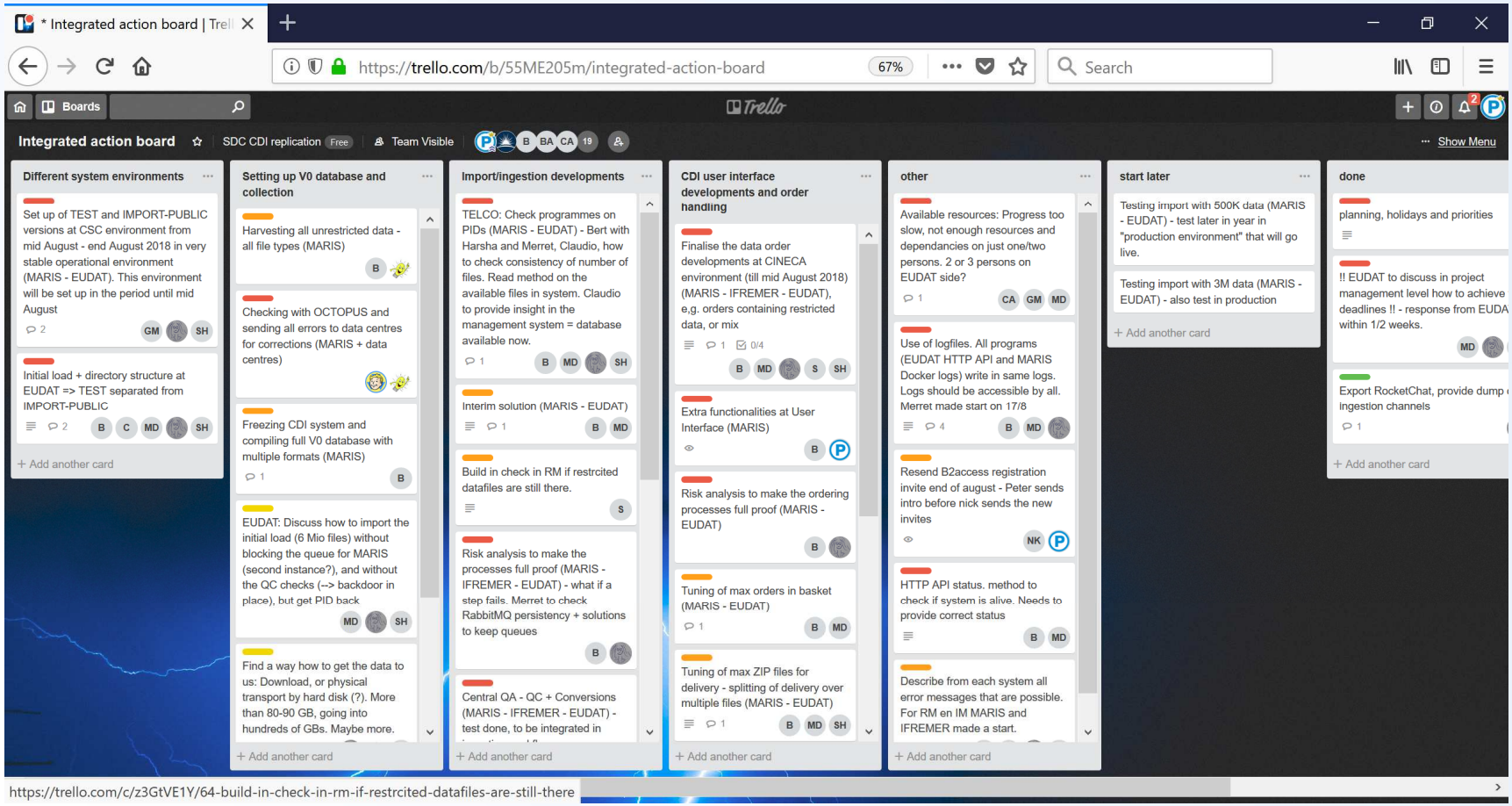
Ongoing actions – EUDAT cloud

- Stability improvement
- Authorisation hick-ups
- Test with workers – now 5 (ingestion and orders combined?), is this sufficient?
- Test and production environment:
  - RM and IM API calls have a switch Test/Prod, needs balancer:
    - If test => go to test cloud
    - If prod => go to production
- Performance of ingestion and order handling optimization
- Initial load of files (500K first, then 5M files)
- Consistency/integrity checks - what if files in order/request not available => Monitoring plan

Ongoing actions - Import Manager

- Test vs Production calls testing
- Consistency/integrity checks - what if files in order/request not available => Monitoring plan
- Integration tests of ingestion and order workflows

In more detail on our action board



Plus:

Create documentation

Further planning towards mid March:

- See implementation plan

SeaDataCloud

- END -