Status of CDI harvesting and ingestion service

By
Dick M.A. Schaap – Technical Coordinator

Brest – France, 17 September 2015,
SeaDataNet II Plenary Meeting

Harvesting and ingestion of CDI XMLs

- Objective is to upgrade the submission and processing of CDI entries from data centres to the CDI portal service by means of **harvesting and ingestion**

- IFREMER has adopted GeoNetwork and provided guidance document for supporting CDI XML output of MIKADO and making it available by means of local CS-W service

- Testbed has been set-up with **IFREMER, BSH and IEO** that have installed and configured the GeoNetwork software as provided by IFREMER.

- MARIS has set-up and tested well functioning of central harvesting of CDI XML from the GeoNetwork CS-W service (could also be provided by other local software than GeoNetwork)

- A test page was set-up and shared with the pilot partners to check the number of entries for harvesting using a date criterium, which worked out fine => **customised GeoNetwork package and MARIS central harvesting work ok**

Challenge: central CDI ingestion taking into account the staging process and relational model CDI – coupling table – local data

- MARIS applies a staging process for populating new and updated CDI entries, received from data centres:
  - Validation of syntax and semantics

    if ok
  - Duplicates check => report to data centre for check

    if ok
  - Import of CDIs incl GML validation

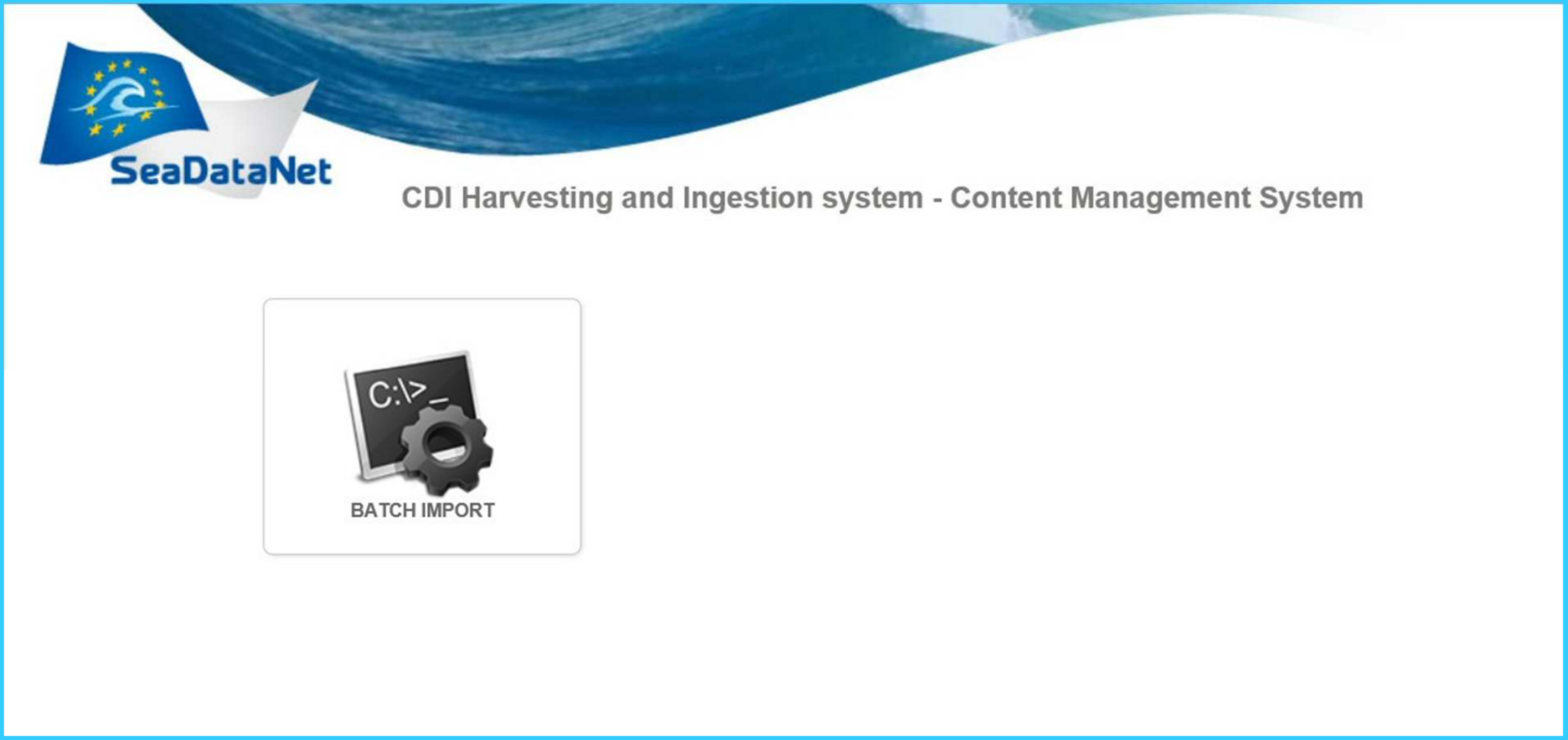
    if ok
  - CDIs in Import CDI service and user interface for visual check by data centres

    if ok
  - Data centres must update Coupling Table and arrange Local Data sets

    if ok
  - CDIs moved to production CDI service for public use

Online CMS for ingestion workflow

- An online CMS system has been set-up whereby data centres can manage themselves => establishing data centre self responsibility + 24 / 7
- Logon by AAA service for data provider contacts


SeaDataNet
CDI Harvesting and Ingestion system - Content Management System
C:\>_
BATCH IMPORT

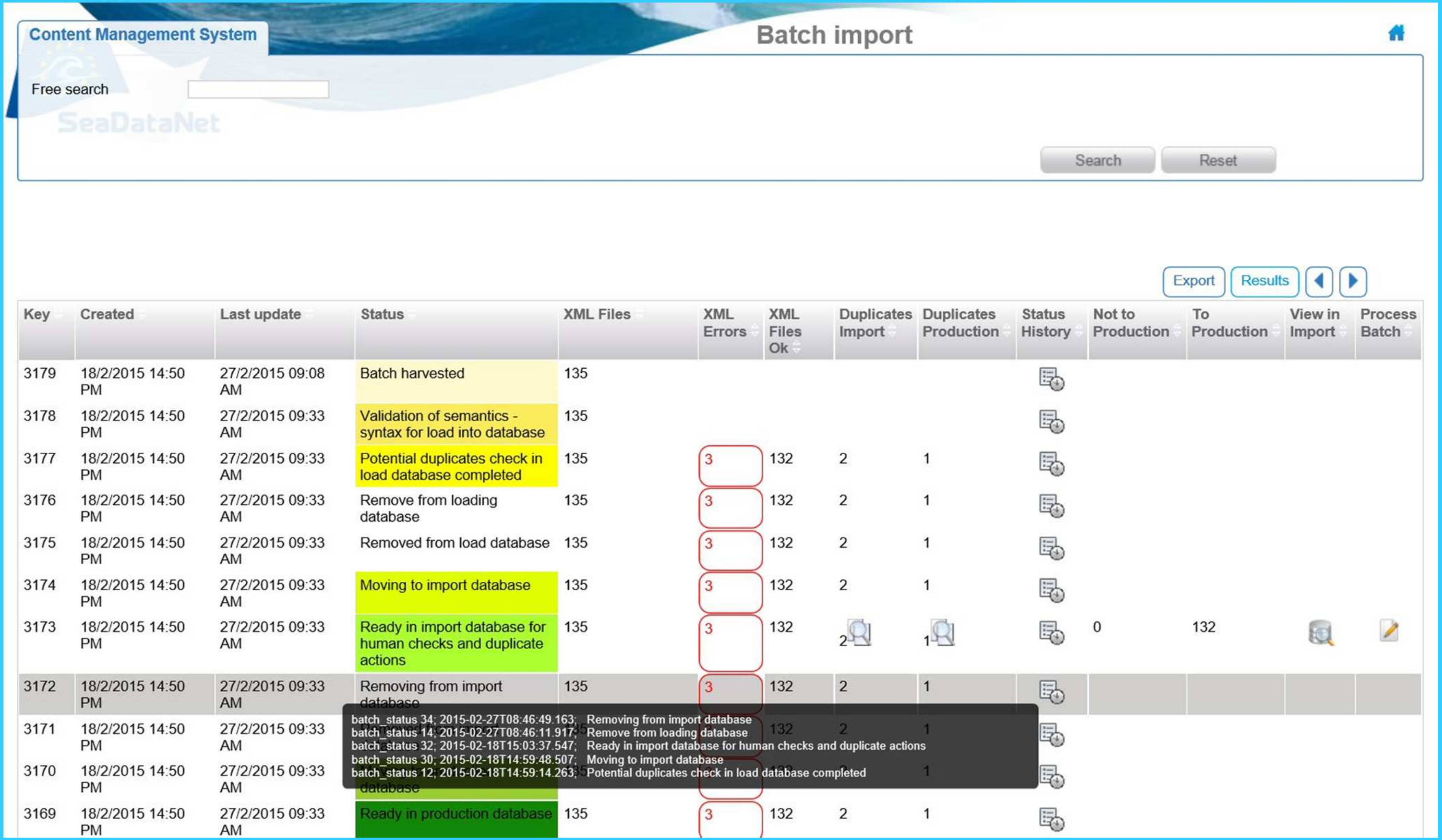
online CMS for CDI ingestion by data centres

- Principle is that per data provider each time only ONE harvested batch will be processed; when ready, then next harvest will take place at regular intervals (e.g. each week)

- System runs in steps through a batch process in which data provider is asked to interact only a few times:
  - Check identified XML errors (syntax – semantics) for possible repair in next batch
  - Check identified potential duplicates (against import and production) and undertake action to delete real duplicates from the import database
  - Check overall remaining CDI import and submit action for moving to production (after confirming up-to-date coupling table and new data availability)
- The overall process works in batch mode: for example moving from import to production takes place each night because also requires a lot of indexing.

- Note: CDI deletions continue as specials by email from data centres to MARIS

Online CMS for CDI ingestion by data centres

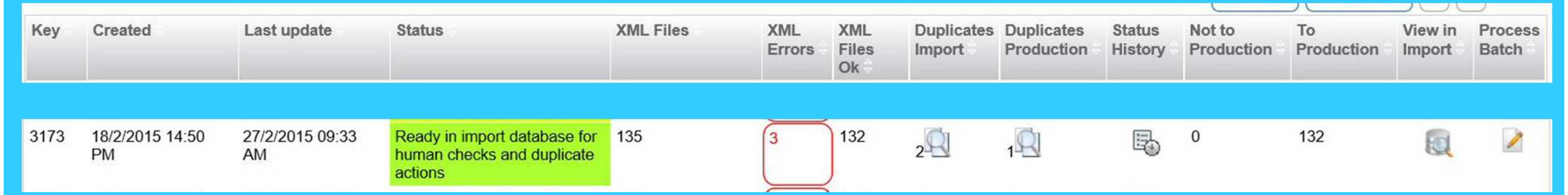


Key	Created	Last update	Status	XML Files	XML Errors	XML Files Ok	Duplicates Import	Duplicates Production	Status History	Not to Production	To Production	View in Import	Process Batch
3179	18/2/2015 14:50 PM	27/2/2015 09:08 AM	Batch harvested	135									
3178	18/2/2015 14:50 PM	27/2/2015 09:33 AM	Validation of semantics - syntax for load into database	135									
3177	18/2/2015 14:50 PM	27/2/2015 09:33 AM	Potential duplicates check in load database completed	135	3	132	2	1					
3176	18/2/2015 14:50 PM	27/2/2015 09:33 AM	Remove from loading database	135	3	132	2	1					
3175	18/2/2015 14:50 PM	27/2/2015 09:33 AM	Removed from load database	135	3	132	2	1					
3174	18/2/2015 14:50 PM	27/2/2015 09:33 AM	Moving to import database	135	3	132	2	1					
3173	18/2/2015 14:50 PM	27/2/2015 09:33 AM	Ready in import database for human checks and duplicate actions	135	3	132	2	1		0	132		
3172	18/2/2015 14:50 PM	27/2/2015 09:33 AM	Removing from import database	135	3	132	2	1					
3171	18/2/2015 14:50 PM	27/2/2015 09:33 AM	Removed from import [illegible]	135	[illegible]	132	2	1					
3170	18/2/2015 14:50 PM	27/2/2015 09:33 AM	[illegible] database	135	[illegible]	132	2	1					
3169	18/2/2015 14:50 PM	27/2/2015 09:33 AM	Ready in production database	135	3	132	2	1					

```
batch_status 34; 2015-02-27T08:46:49.163;   Removing from import database
batch_status 14; 2015-02-27T08:46:11.917;   Remove from loading database
batch_status 32; 2015-02-18T15:03:37.547;   Ready in import database for human checks and duplicate actions
batch_status 30; 2015-02-18T14:59:48.507;   Moving to import database
batch_status 12; 2015-02-18T14:59:14.263;   Potential duplicates check in load database completed
```

CMS with all possible ‘status’ examples

online CMS for CDI ingestion by data centres



Key	Created	Last update	Status	XML Files	XML Errors	XML Files Ok	Duplicates Import	Duplicates Production	Status History	Not to Production	To Production	View in Import	Process Batch
3173	18/2/2015 14:50 PM	27/2/2015 09:33 AM	Ready in import database for human checks and duplicate actions	135	3	132	2	1		0	132		

Batch no

Prevailing status

Option to retrieve log of XML errors

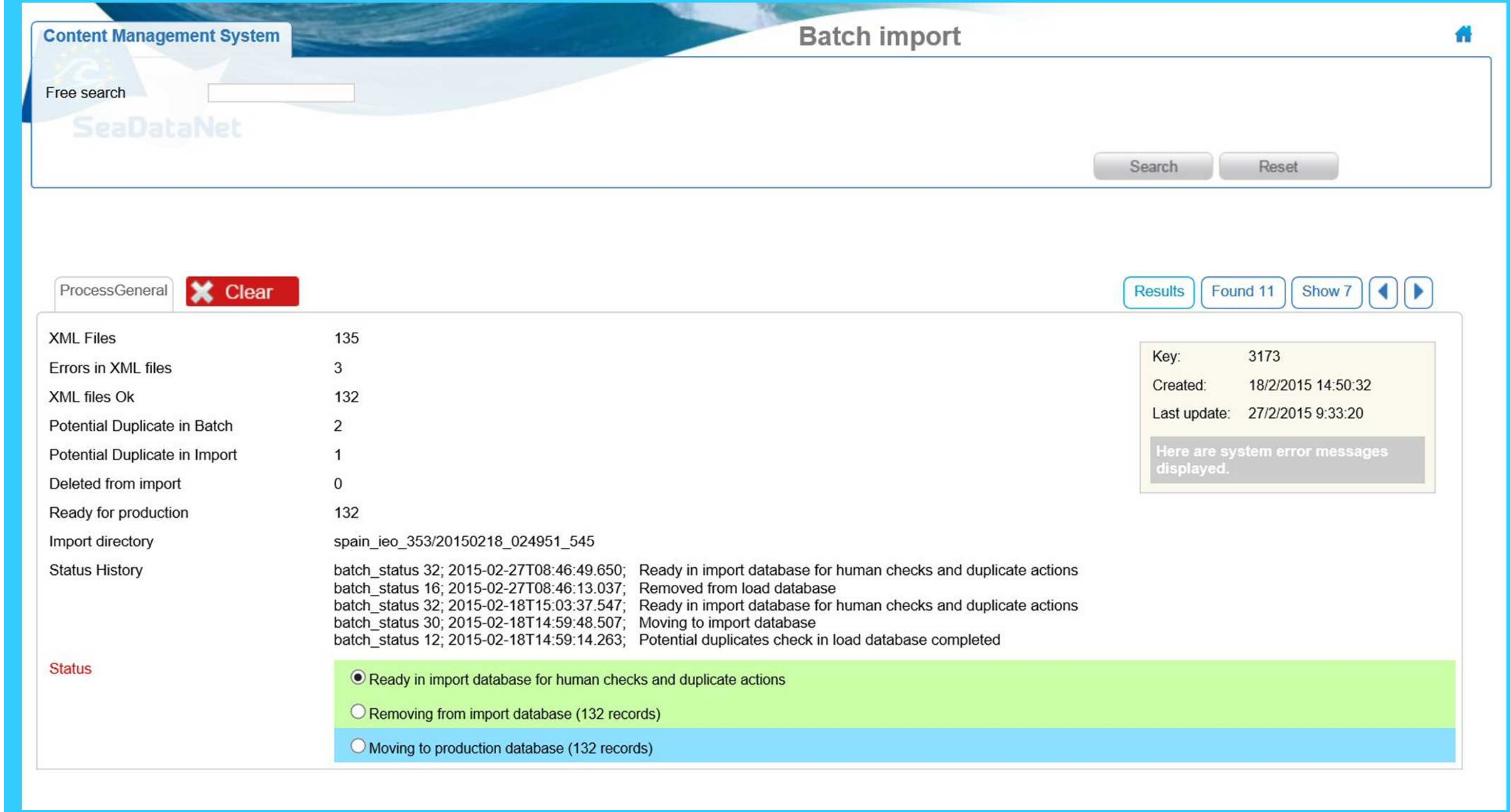
Options to retrieve logs, to inspect CDIs and to mark CDIs in import for deletion

Option to inspect CDIs in import

Click to mark for removal or moving to production (see next page)

Possible interactions for data provider

online CMS for CDI ingestion by data centres



Batch import

Free search

XML Files	135
Errors in XML files	3
XML files Ok	132
Potential Duplicate in Batch	2
Potential Duplicate in Import	1
Deleted from import	0
Ready for production	132
Import directory	spain_ieo_353/20150218_024951_545
Status History	batch_status 32; 2015-02-27T08:46:49.650; Ready in import database for human checks and duplicate actions batch_status 16; 2015-02-27T08:46:13.037; Removed from load database batch_status 32; 2015-02-18T15:03:37.547; Ready in import database for human checks and duplicate actions batch_status 30; 2015-02-18T14:59:48.507; Moving to import database batch_status 12; 2015-02-18T14:59:14.263; Potential duplicates check in load database completed
Status	● Ready in import database for human checks and duplicate actions ○ Removing from import database (132 records) ○ Moving to production database (132 records)

Key:	3173
Created:	18/2/2015 14:50:32
Last update:	27/2/2015 9:33:20

Here are system error messages displayed.

Final action: Removing remaining Batch or moving to Production

CDI ingestion test in 3 rounds

- Pilot partners for testing: BSH, IFREMER, and IEO
- With each data centre 3 rounds are undertaken:

1) MARIS harvests CDI XML and also receives CDI by email as check for conformity of the CDIs

CDI are loaded in online CMS and MARIS pushes it through

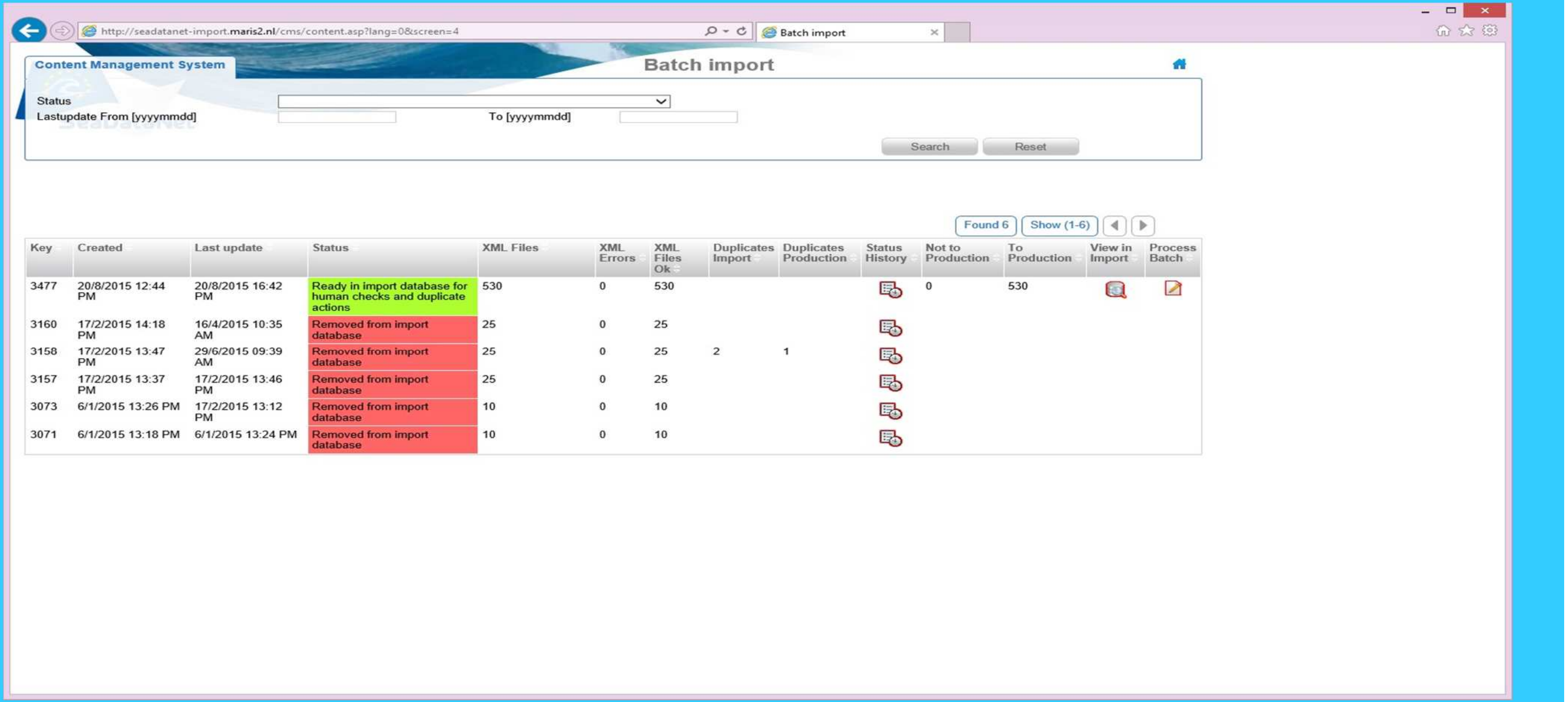
2) MARIS harvests 2nd batch CDI XML and also receives CDI by email as check for conformity of the CDIs

CDI are loaded in online CMS, MARIS pushes it through and data centre is invited to monitor the online CMS and workflow passively

3) MARIS harvests 3rd batch CDI XML and also receives CDI by email as check for conformity of the CDIs

CDI are loaded in online CMS, data centre is invited to undertake all steps itself via the online CMS while MARIS is watching and giving guidance where needed

CDI ingestion test with BSH


http://seadatanet-import.maris2.nl/cms/content.asp?lang=0&screen=4
Batch import
Content Management System

Batch import

Status

Lastupdate From [yyyymmdd]

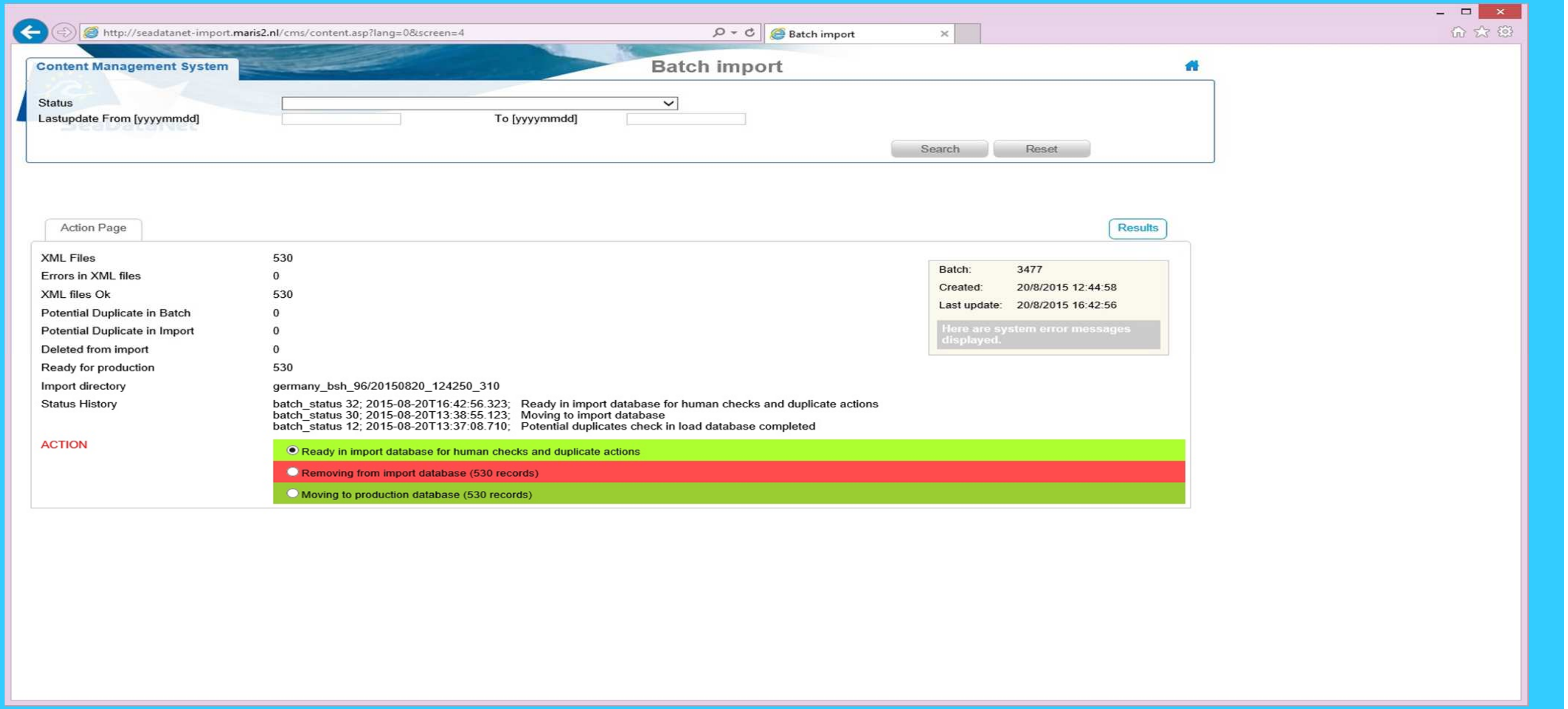
To [yyyymmdd]

Search
Reset
Found 6
Show (1-6)

Key	Created	Last update	Status	XML Files	XML Errors	XML Files Ok	Duplicates Import	Duplicates Production	Status History	Not to Production	To Production	View in Import	Process Batch
3477	20/8/2015 12:44 PM	20/8/2015 16:42 PM	Ready in import database for human checks and duplicate actions	530	0	530				0	530		
3160	17/2/2015 14:18 PM	16/4/2015 10:35 AM	Removed from import database	25	0	25							
3158	17/2/2015 13:47 PM	29/6/2015 09:39 AM	Removed from import database	25	0	25	2	1					
3157	17/2/2015 13:37 PM	17/2/2015 13:46 PM	Removed from import database	25	0	25							
3073	6/1/2015 13:26 PM	17/2/2015 13:12 PM	Removed from import database	10	0	10							
3071	6/1/2015 13:18 PM	6/1/2015 13:24 PM	Removed from import database	10	0	10							

Harvested batch 3477 loaded into CMS for BSH

CDI ingestion example with BSH


http://seadatanet-import.maris2.nl/cms/content.asp?lang=0&screen=4
Batch import
Content Management System
Batch import
Status
Lastupdate From [yyyymmdd]
To [yyyymmdd]
Search
Reset
Action Page
Results

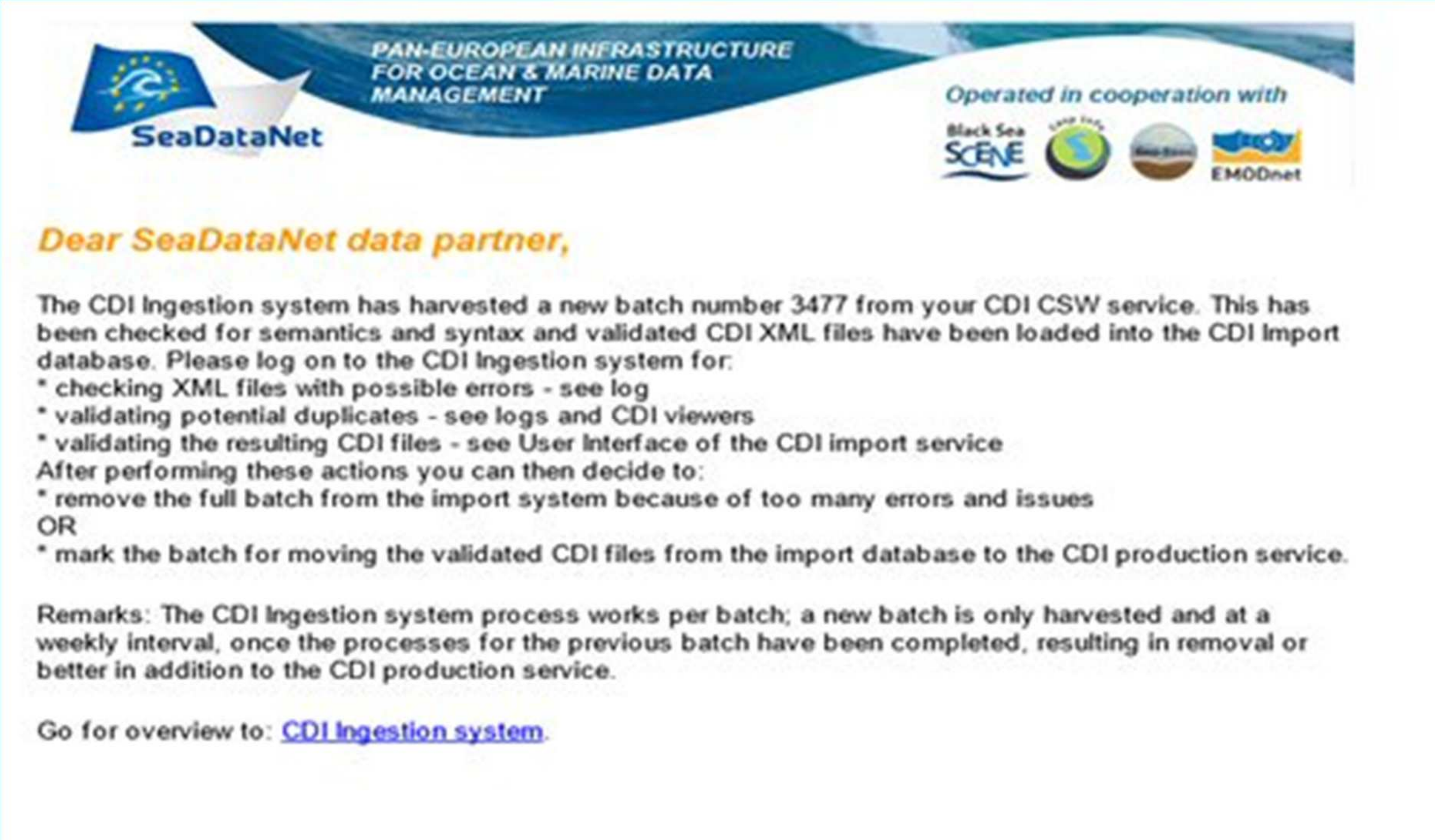
XML Files	530
Errors in XML files	0
XML files Ok	530
Potential Duplicate in Batch	0
Potential Duplicate in Import	0
Deleted from import	0
Ready for production	530
Import directory	germany_bsh_96/20150820_124250_310
Status History	batch_status 32; 2015-08-20T16:42:56.323; Ready in import database for human checks and duplicate actions batch_status 30; 2015-08-20T13:38:55.123; Moving to import database batch_status 12; 2015-08-20T13:37:08.710; Potential duplicates check in load database completed
ACTION	● Ready in import database for human checks and duplicate actions ○ Removing from import database (530 records) ○ Moving to production database (530 records)

Batch:	3477
Created:	20/8/2015 12:44:58
Last update:	20/8/2015 16:42:56

Here are system error messages displayed.

Details of batch 3477: number, errors, xml ok, potential duplicates agains batch and against production

CDI ingestion example with BSH



Dear SeaDataNet data partner,

The CDI Ingestion system has harvested a new batch number 3477 from your CDI CSW service. This has been checked for semantics and syntax and validated CDI XML files have been loaded into the CDI Import database. Please log on to the CDI Ingestion system for:
* checking XML files with possible errors - see log
* validating potential duplicates - see logs and CDI viewers
* validating the resulting CDI files - see User Interface of the CDI import service
After performing these actions you can then decide to:
* remove the full batch from the import system because of too many errors and issues
OR
* mark the batch for moving the validated CDI files from the import database to the CDI production service.

Remarks: The CDI Ingestion system process works per batch; a new batch is only harvested and at a weekly interval, once the processes for the previous batch have been completed, resulting in removal or better in addition to the CDI production service.

Go for overview to: CDI Ingestion system.

E-mail message to data centre about harvesting and successful loading and invitation to logon to online CMS for further processing of workflow

CDI ingestion pilot on-going

- Full cycle of 3 rounds has been successful with BSH
- Full cycle of 3 rounds has also been successful with IFREMER; however IFREMER has introduced a 4th round with a harvest of > 40.000 CDI XML. This gives some issues with GeoNetwork taking long time for indexing and this is now being analysed
- Next is IEO
- If results stay as successful as so far, then the process and guidance notes for GeoNetwork installation and use of Ingestion CMS will be documented in an update of D5.4 and it can be adopted by other data centres in communication with MARIS for initialisation
- MARIS will monitor these ingestion processes and also keep on processing CDI batches received from other data centres in the traditional way by email and possibly FTP
- Note: the online CMS is anyway used and has been used by MARIS in the last 6 months for all CDI packages received