# Testing catalogue consistency

## Abstract

The following paper will describe a low impact method to compare the catalogue consistency between the different GISCs. The results are presented on the example of the OAI set “WIS-GISC-Offenbach” and further findings are documented.

### Catalogue consistency – a low impact algorithm

Until now most catalogue consistency checks compared the identifiers in each single OAI set between the different GISCs. This check is helpful to detect cases where metadata records are missing or a harvesting GISC missed a deleted message. But this check can’t detect the case when a provider updates a record and the harvester missed the update message.

The following method describes a low impact algorithm which avoids harvesting all records and then dealing with the question when 2 metadata records are different. For example an XML document is equivalent to another XML document if the order to the attributes inside of an element is different.

The method is based on OAI ListIdentifer request and the OAI dateStamp. Every time when a record is inserted, updated or deleted on the OAI provider the OAI dateStamp is updated. When a GISC harvests the record from the provider and inserts, updates or deletes the record the OAI dateStamp is also updated. Therefore an older OAI dateStamp in a harvesting GISC for the same identifier is an indicator that an update was missed.

**Examples**

**Successful update**

Authoritative source (GISC Offenbach):  
de.pangaea.dataset676755 2015-09-19 17:40:04.0

Harvesting GISC:  
de.pangaea.dataset676755 2015-09-19 18:38:38.0

**Missed update**

Authoritative source (GISC Offenbach):  
de.pangaea.dataset676755 2015-09-19 17:40:04.0

Harvesting GISC:  
de.pangaea.dataset676755 2015-07-24 21:17:59.0

**Possible inaccurate OAI datestamp handling**

Authoritative source (GISC Offenbach):  
de.pangaea.dataset676755 2015-09-19 17:40:04.0

Harvesting GISC:  
de.pangaea.dataset676755 2015-09-19 17:40:04.0

### Tests

To verify the above presented method I harvested from all operational GISCs the OAI set “WIS-GISC-Offenbach”. The tests were performed with the GISCs from BOM, CMA, IRIMO, JMA, KMA, MF, PME, Roshydromet and UKMO.

### Results

I could test the method only with 7 of the 9 above listed GISCs because in one case a GISC always throws an internal error if queried by an OAI ListIdentifier request while in the other case I received too many connection errors to complete the tests.

In all other 7 cases the method was successful. I identified in 1 GISC 11 outdated records while in all other cases the records where up-to-date. Furthermore I found that in 1 GISC 60 records of the “WIS-GISC-Offenbach” OAI Set are missing.

### Notes

4 GISCs had in most cases the same OAI dateStamp as the GISC Offenbach provider. Excluding the rare possibility of a coincidence that always in the same second the GISC harvested GISC Offenbach’s provider this issue was already addressed in last TT-GISC meeting 2014 with the topic inaccurate OAI dateStamp [1].

Limitations of the algorithm: the presented algorithm can’t detect cases when a GISC modifies the harvested records or when only the OAI dateStamp is updated and not the record.

## References

[1] Final Report TT-GISC 2014 http://wis.wmo.int/file=759

## Recommended text for report

The meeting noted the presented report from Germany and recognizes with appreciation the overall good consistency of the catalogues. The meeting encourages all GISCs to implement similar tests and to validate the catalogue consistency. Furthermore the meeting emphasizes the importance of accurate handling of the OAI dateStamp and urges the GISCs to fix these issues by the next ET-WISC meeting. Also the meeting urges the GISCs to fix issues with OAI ListIdentifier requests as soon as possible.

--------------------