

Introduction to the Aggregation Database

Søren Roug, IT Project manager

What is the problem?

The Aggregation Database is being developed to solve some new problems occurring when data is distributed in SEIS:

- How to *discover* the datasets
- How to *search* for the datasets
- How to *track* updates to the datasets
- How to *bookmark* found datasets
- How to *merge* datasets
- How to *trust* the dataset
- How to trust the *trust*

The system is going to be an extension to EEA's Reportnet system

Discovering the data

Datasets tend to be only numbers. Unlike Google, that mainly harvests webpages, a SEIS search engine can't get information from the content of the files.

The aggregation database will use two mechanisms to discover relevant datasets.

1. A collaborating node provides a *manifest* of the relevant files existing on its own website
2. Users register the files at a new website called QAW along the principles of del.icio.us and digg.com

Discovering with the manifest file

- The manifest lists location and any metadata the provider has on each file
- The format can handle any type of metadata property

```
<?xml version="1.0" encoding="utf-8" ?>
<rod:Delivery rdf:about="http://cdr.eionet.europa.eu/hu/colrzxujg/envrzxuuw">
  <dc:title>Bathing water report, 2006</dc:title>
  <rod:released>2007-11-15T16:17:27Z</rod:released>
  <rod:locality rdf:resource="http://rod.eionet.eu.int/spatial/17"/>
  <rod:period>2006</rod:period>
  <rod:obligation rdf:resource="http://rod.eionet.eu.int/obligations/21"/>
  <rod:file rdf:resource="http://cdr.eionet.europa.eu/hu/bwd.xml"/>
</rod:Delivery>
```

Registering a SEIS dataset

Quality Assessment Workbench

SERVICES | REPORTNET | TOOLS | TOPICS

You are here: Eionet » QA Workbench

» [Main page](#)
» [Personal page](#)
» [Add resource](#)
» [List saved resources](#)
» [Add QA Report](#)
» [Dataflow search](#)

User name: roug [Logout](#) [Personal page](#)

Register new resource

URL:

Recent resources

1. [River data Cyprus](#)
Discovered : 14/10/2008 @ 19:22
Related tags:
2. [2006_September](#)
Discovered : 14/10/2008 @ 19:22
Related tags:
3. [EUMM_2005_Mar05](#)
Discovered : 14/10/2008 @ 19:22
Related tags:
4. [Ozone Data for 2001 - annual statistics and exceedances](#)
Discovered : 14/10/2008 @ 19:22
Related tags:
5. [Groundwater Quality Data 2004_corr](#)
Discovered : 14/10/2008 @ 19:22
Related tags:
6. [2005 - May](#)
Discovered : 14/10/2008 @ 19:22
Related tags:
7. [ES122 - FUERTEVENTURA](#)

Discovered via manifest files and manual registration

Breaking news!

Now you can use a [bookmarklet](#) to simplify the process of adding resources to the system. Follow this link to [add the bookmarklet to your browser](#).

Adding metadata

Quality Assessment Workbench

SERVICES | REPORTNET | TOOLS | TOPICS

You are here: Eionet » QA Workbench » Add/edit resource

- » Main page
- » Personal page
- » Add resource
- » List saved resources
- » Add QA Report
- » Dataflow search

Edit resource

Info (hide)

1. Resource was successfully added

Main data | Attributes | Methodology | Obligations | Users QA reports

URL  <http://uba.at/riverdata/delivery2008.xml>

Type

Title

Description

Tags

Space separated

Save

Clear entries

Breaking news

Now you can use a [bookmarklet](#) to simplify the process of adding resources to the system. Follow the link to [add the bookmarklet to your browser](#).

Bookmarking and searching the dataset

EIONET
Quality Assessment Workbench



SERVICES | REPORTNET | TOOLS | TOPICS



You are here: [Eionet](#) » [QA Workbench](#) » Personal page



» [Main page](#)
» [Personal page](#)
» [Add resource](#)
» [List saved resources](#)
» [Add QA Report](#)



Personal page



[My history](#)



[Bookmarklet - Wikipedia, the free encyclopedia](#)  
Related tags: [bookmarklets](#) [wikipedia](#)



[afoe | A Fistful of Euros | European Opinion](#)  
Related tags:



[Paul Krugman - Op-Ed Columnist - New York Times Blog](#)  
Related tags: [news](#) [opinion](#)



http://www.eionet.europa.eu/gis/docs/EEA_GISguide_v2.doc  
Related tags: [maps](#) [gis](#) [guidelines](#)



[Estonia](#)  
Related tags:

[What Programming Languages Should You Know?](#)  
Related tags: [programming](#)

[The New York Times - Breaking News, World News & Multimedia](#)  
Related tags: [news](#)

http://www.sun.com/software/star/odf_plugin/index.jsp  
Related tags: [sun](#) [odf](#) [openoffice](#)

[The Register: Sci/Tech News for the World](#)  
Related tags: [news](#)

<http://www.tietoanator.com>  
Related tags: [software](#) [company](#)

Breaking news
Now you can use a [bookmarklet](#) to simplify the process of adding resources to the system. Follow the link to [add the bookmarklet to your browser](#).

- [news](#)
- [bookmarklets](#)
- [company](#)
- [gis](#)
- [guidelines](#)
- [maps](#)
- [odf](#)
- [openoffice](#)
- [opinion](#)
- [programming](#)
- [software](#)
- [sun](#)
- [wikipedia](#)

Working with files vs. records

- Now we know where the files are in the SEIS universe
- But we can do *more*:
 - We can read the content of XML files
 - Example of an XML snippet:

```
<stations xmlns:xsi=http://www.w3.org/2001/XMLSchema-instance
xsi:noNamespaceSchemaLocation="http://water.eionet.europa.eu/stations.xsd">
  <station>
    <local_code>32301</local_code>
    <name>St. Pölten</name>
    <longitude>15.63166</longitude>
    <latitude>48.21139</latitude>
    <altitude>270</altitude>
    <station_type>Industrial</station_type>
    <area_type>urban</area_type>
  </station>
  ...
</stations>
```

Merging principles

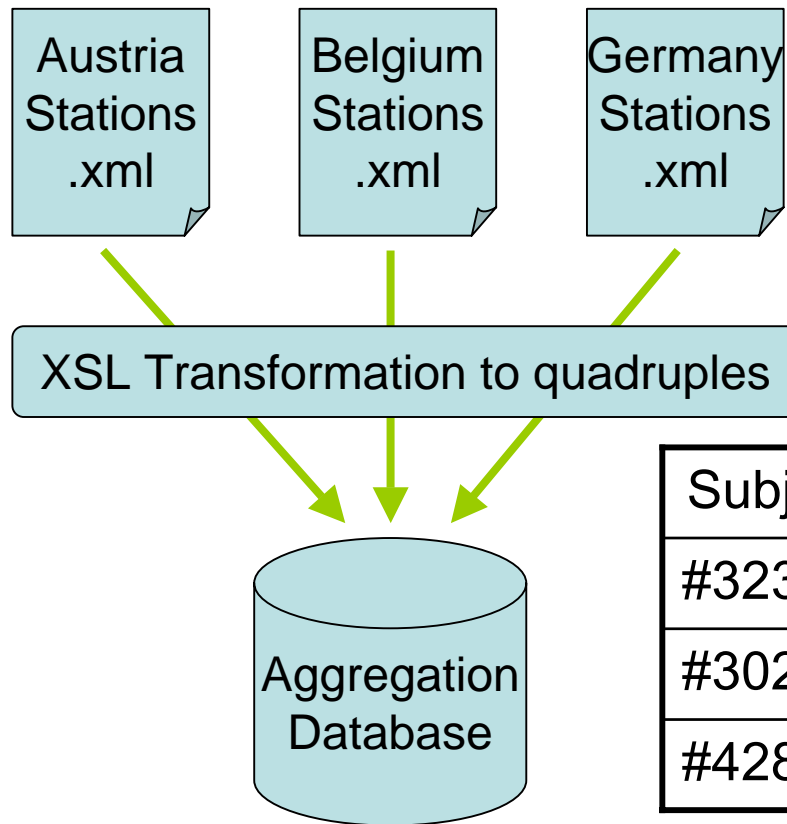
Station structure as a table (austria.xml)

| Identifier | local code | name | ... |
|-------------------|-------------------|-------------|------------|
| #32301 | 32301 | St. Pölten | ... |
| #32302 | 32302 | Linz | ... |

Quadruple structure

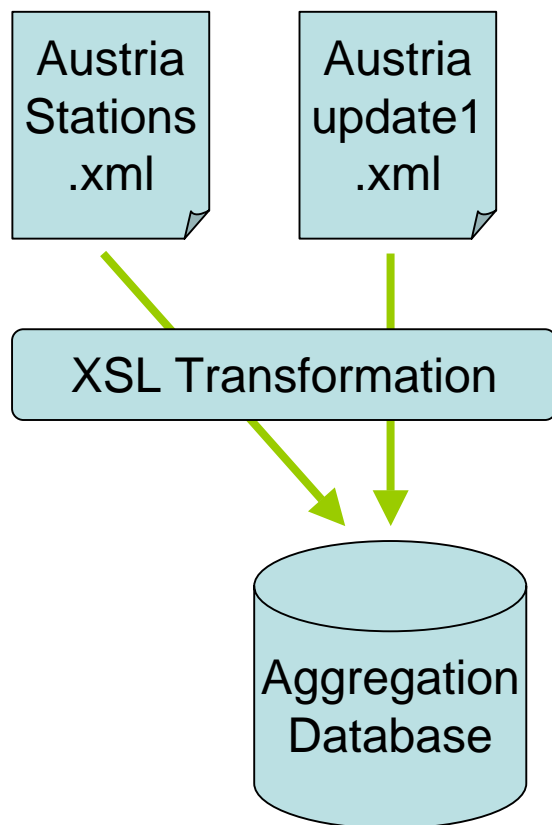
| Subject | Predicate | Object | Source |
|----------------|------------------|---------------|--------------------|
| #32301 | type | River Station | <i>austria.xml</i> |
| #32301 | local code | 32301 | <i>austria.xml</i> |
| #32301 | name | St. Pölten | <i>austria.xml</i> |
| #32302 | type | River Station | <i>austria.xml</i> |
| #32302 | local code | 32302 | <i>austria.xml</i> |
| #32302 | name | Linz | <i>austria.xml</i> |

Merging the datasets



| Subject | Predicate | Object | Source |
|---------|-----------|------------|---------|
| #32301 | name | St. Pölten | Au..xml |
| #30299 | name | Gent | Be..xml |
| #42882 | name | Köln | Ge..xml |

Merging the datasets (with later updates)



| Subject | Predicate | Object | Source |
|---------|-----------|------------|-----------------|
| #32301 | name | St. Pölten | Au..xml |
| #32301 | date | 2005-10-8 | Au..xml |
| #32301 | name | Spratzern | Au..update1.xml |
| #32301 | date | 2008-6-18 | Au..update1.xml |

Searching

- To find all river stations in Europe you search for subjects with the *type="River Station"*

| Identifier | Local_code | Name | Date | Longitude |
|------------|------------|------------|------------|-----------|
| #32301 | 32301 | St. Pölten | 2005-10-8 | 15.63166 |
| #32301 | | Spratzern | 2008-6-18 | |
| #30299 | 30299 | Gent | 2004-11-12 | 3.733333 |
| #42882 | 42882 | Köln | 2001-4-14 | 6.966667 |

- The query will format it as a table for you
- Obviously you get duplicates because 32301 has been updated

QA work

- Let's first colour the cells by their source

| Identifier | Local_code | Name | Date | Longitude |
|------------|------------|------------|------------|-----------|
| #32301 | 32301 | St. Pölten | 2005-10-8 | 15.63166 |
| #32301 | | Spratzern | 2008-6-18 | |
| #30299 | 30299 | Gent | 2004-11-12 | 3.733333 |
| #42882 | 42882 | Köln | 2001-4-14 | 6.966667 |

QA work

- Then we merge by letting the newer sources overwrite the older:

| Identifier | Local_code | Name | Date | Longitude |
|------------|------------|-----------|------------|-----------|
| #32301 | 32301 | Spratzern | 2008-6-18 | 15.63166 |
| #30299 | 30299 | Gent | 2004-11-12 | 3.733333 |
| #42882 | 42882 | Köln | 2001-4-14 | 6.966667 |

QA work

- Don't trust one source?
- Turn it off before you merge

| Identifier | Local_code | Name | Date | Longitude |
|-------------------|------------|----------------------|----------------------|-----------|
| #32301 | 32301 | St. Pölten | 2005-10-8 | 15.63166 |
| #32301 | | Spratzern | 2008-6-18 | |
| #30299 | 30299 | Gent | 2004-11-12 | 3.733333 |
| #42882 | 42882 | Köln | 2001-4-14 | 6.966667 |

QA work

- Then we merge

| Identifier | Local_code | Name | Date | Longitude |
|------------|------------|------------|------------|-----------|
| #32301 | 32301 | St. Pölten | 2005-10-8 | 15.63166 |
| #30299 | 30299 | Gent | 2004-11-12 | 3.733333 |
| #42882 | 42882 | Köln | 2001-4-14 | 6.966667 |

QA work

- Gapfilling?
- Add your own source as a layer
- The layer is stored on QAW

| Identifier | Local_code | Name | Date | Longitude |
|------------|------------|------------|------------|-----------|
| #32301 | 32301 | St. Pölten | 2005-10-8 | 15.63166 |
| #32301 | | Spratzern | 2008-6-18 | |
| #30299 | 30299 | Gent | 2004-11-12 | 3.733333 |
| #42882 | 42882 | Köln | 2001-4-14 | 6.966667 |
| #32301 | | | 2008-11-27 | 15.65000 |

Hermann's gapfilling layer created 2008-11-27

QA work

- Then we merge

| Identifier | Local_code | Name | Date | Longitude |
|------------|------------|-----------|------------|-----------|
| #32301 | 32301 | Spratzern | 2008-11-27 | 15.65000 |
| #30299 | 30299 | Gent | 2004-11-12 | 3.733333 |
| #42882 | 42882 | Köln | 2001-4-14 | 6.966667 |

- And we export to our working database for production...

Trusting the dataset and trusting trust

- Datasets and values can be evaluated by looking at the source
- Is the source URL from a reliable organisation/person?
- Is the methodology described?
- Are there reviews on QAW?
 - Who wrote the reviews?
- Are there others who have used the data?
 - Who are they?

Summary

These new tools intend to solve the use of the Reportnet deliveries:

- Aggregation/Merging
- Manual QA and gap-filling
- Traceability to the sources
 - Noticing when the source has been updated/deleted
- Review of the source for inclusion
 - That was no problem before because only authorised parties could upload to CDR
 - With SEIS now anyone can participate