Protocol data uploading

Project **BwN DM 1.1**, Deliverable **2009**

⊘ Printing: If you want to print this document select tools -> export to pdf from the top right corner

Table of contents

Introduction

The OpenEarth data uploading protocol is a refinement of the following mission statement:

"All environmental data gathered with tax-payer money should be freely available on the internet according to international standards to all fellow scientists in specific and all citizens in general for the overall benefit of mankind."

A number of specific criteria are posed to meet this mission statement.

Data (products) collected with tax-payer money should:

- Be open to everyone according to a standardized, internationally accepted copyright statements as provided by for example GNU, or Creative Commons.
- Available on the internet through web-services as envisioned by the inventor of the www. Web-services avoid the risk associated with the widespread practise of creating non-synchronized copies.
- Have meta-data stored *with* the data in standard, internationally accepted manner. At the very least data should have units and a quantity name. Files without units and a quantity may not be considered as data at all (e.g. ascii arcgis files).
- Be fully traceable to its rawest data and its processing tools, to both of which should apply the same conditions as to data (products). All data products should have a version number by which every single number can be tracked down to a raw data value (measured voltage or count) and a processing routine. Dara altered by any undocumented, manual modification may not be considered as data any more.
- Should be easily viewable by the general public.
- Be be archived and disseminated through a system which is open too so that anyone can have a copy of the complete system.

OpenEarth has designed a system with various open source components that fulfills the above criteria.

Raw data server

All raw data and all processing tools to transform the raw data into usable date products are to be stored in a repository. The main function of this repository is archiving, not dissemination. The contents of the repository should be sufficient to regenerate data products from the raw data. This processing tools should be fully open to allow for thorough peer-review as described by Popper. The raw data repository should assign version numbers to raw data + scripts, such that any state of the data products has a unique version number.

- All raw data and processing routines should be uploaded to the OpenEarthRawData Subversion repository: https://svn.oss.deltares.nl/repos/openearthrawdata/trunk/.
- The processing script should contain the feature to read the data, to write the data as netCDF file (see below) and as kml file (see below).
- Generic routines may be uploaded to the OpenEarthTools Subversion repository: https://svn.oss.deltares.nl/repos/openearthtools/trunk/. For the protocol that applies to scripts/functions please refer to Protocol Matlab programming style.
- All modifications in the repository should have a unique version number.
- All modifications in the repository should be traceable to a specific person at a specific institute. The username of each repository user is equal to its institutional emailadress.
- In the repository small caps shall be used.
- All datasets should be stored in the main folder of the above mentioned repository under the name of the institute that holds the copyright(e.g. tno). Add a *.url file that contains a link to the institute webpage.
- All datasets should be stored in the institute folder under the free-to-choose (tree of) name(s) of the dataset(e.g. ahn100).
- All datasets should at the lowest level be divided into a folder for the raw data (raw), the scripts (scripts) and optionally a cache (cache) for downloaded data and optionally processed data (processed) for data that take a long time to process.
- To make data understandable for others and to conform to the European Inspire guideline extra metadata should be provided. The inspire guideline follows the ISO19115 metadata standard. This should be done by filling in the forms in the inspire metadata editor. The resulting file should be saved next to the data set at this level (ahn.url). The proprietary inspire metadata editor can be used to describe datasets. Alternatively you can use the mig editor which runs local and is open source.
- At this level optionally each dataset can have a *.url file with a relevant weblink (e.g. ahn.url).
- On windows computers the Subversion repository server can be accessed by use of the tortoise svn client , through the add network wizard (webdav) or through a web browser.

In summary the lay-out of raw data repository looks like:

```
https://repos.deltares.nl/repos/OpenEarthRawData/trunk/...
\_your_organization
    \_weblink_to_your_organization.url
    \_your_dataset_title
      \_your_dataset_sub_title
        \_inspire_description.xml
        \_relevant_weblink.url
        \_raw
          \_data file_1
          \_data file_2
          ...
          \_data file_n
        \_scripts
          \_script_read_raw_data
          \_script_to_make_netCDF_file (see below)
          \_script_to_make_kml_file (see below)
        \_cache
        \_processed
```

Processed data format

netCDF

- To facilitate the use of data the open file format netCDF is adopted. For this format not only the description if fully open, but also the libraries to read them are open.
- For each dataset a processing script (in any language of choice) shall be made, and uploaded to the repository, that (i) transforms the data into netCDF format (e.g. https://svn.oss.deltares.nl/repos/openearthrawdata/trunk/tno/ahn100m/scripts/ahn2nc.bat) and (ii) adds all relevant meta-information.
- For the netCDF file structure the CF convention shall be used. The CF convention comprises
    - standard names for quantity names.
    - standard names for unit names.
    - standards for description of spatio-temporal coordinates.
    - standards for description of spatio-temporal statistical operations.
    - standards for description of discovery information (title, name , institution)
- In cases where the CF convention does not hold, additional conventions can be used, provided they are shared via the OpenEarth portal. Any additional convention shall be considered as temporary, and all effort shall be made to have the addition accepted in a standard, internationally accepted convention.
- The netCDF shall have the name of the script that produced it appended as meta-information to facilitate the tracking down of every number. The Subversion keywords 'Id' and 'Headurl' shall be used for that in the OpenEarthRawData and OpenEarthTools repository.
- OpenEarth requires each netCDF file to contain a disclaimer and a terms_for_use statement.
- Examples of how to make netCDF files, and detailed requirements, can be found in the tutorial section on the OpenEarth.nl portal.

PostgreSQL/PostGIS

- Besides netCDF there is a number of reasons to store observations in a database. Therefore a PostgreSQL/PostGIS database is set up. Data should be submitted to the repository as csv files with a description in some format (doc, pdf, txt). Data should be unambigous. Check this document which contains all necessary information on the data, the formats and the standards that can be used.
- The standards named under the CF convention can also be used for single observations to be stored in the PostgreSQL database.

Processed data server

netCDF

- The netCDF file shall be uploaded to the OpenEarth OPeNDAP server: http://opendap.deltares.nl. Please see the Protocol OPeNDAP uploading.
- The main function of this is OPeNDAP server is dissemination, not archiving.
- The structure of the server is the same as the structure of the repository. At the level where the directories \raw and \scripts reside in the repository, netCDF files (*.nc) reside in the OPeNDAP server.

In summary the lay-out of the OPeNDAP server looks like:

```
https://opendap.deltares.nl/...
\_your_organization
   \_your_dataset_title
     \_your_dataset_sub_title
        \_data file_1.nc
        \_data file_2.nc
        ...
        \_data file_n.nc
```

PostgreSQL/PostGIS

Data store in PostgreSQL/PostGIS database is only visible through the use of a client like pgAdmin. Software is freely available at http://pgadmin.org/. There is a readonly OpenEarth account available.

The postgres database can be reached by the URL postgresx03.infra.xtr.deltares.nl. username is OpenEarth, password is oet@BWN, database is BWN.

Visualized data and server

netCDF

- For each dataset a visualisation shall be made in the open *.kml format, such that it can be viewed directly in Google Earth and other programs that support the open kml format. The script that generates the kml shall be uploaded to the repository, just as the script that generated the netCDF file.
- A kml file shall be made that contains deep links to the associated netCDF files on the OPeNDAP server.

In summary the lay-out of the kml server looks like:

```
https://opendap.deltares.nl/...
\_your_organization
   \_your_dataset_title
     \_your_dataset_sub_title
        \_data file_1.kml
        \_data file_2.kml
        ...
        \_data file_n.kml
```

PostgreSQL/PostGIS

Data in the database can be disseminated through the geoserver. The geoserver can be found at http://pmr-geoserver.deltares.nl. In the document presented above there is a paragraph on the way to disseminate data (information). The functions in the database enable various ways of aggregation of the data.

Status BwN datasets

The current status of the Building with Nature datasets can be found in this excelsheet