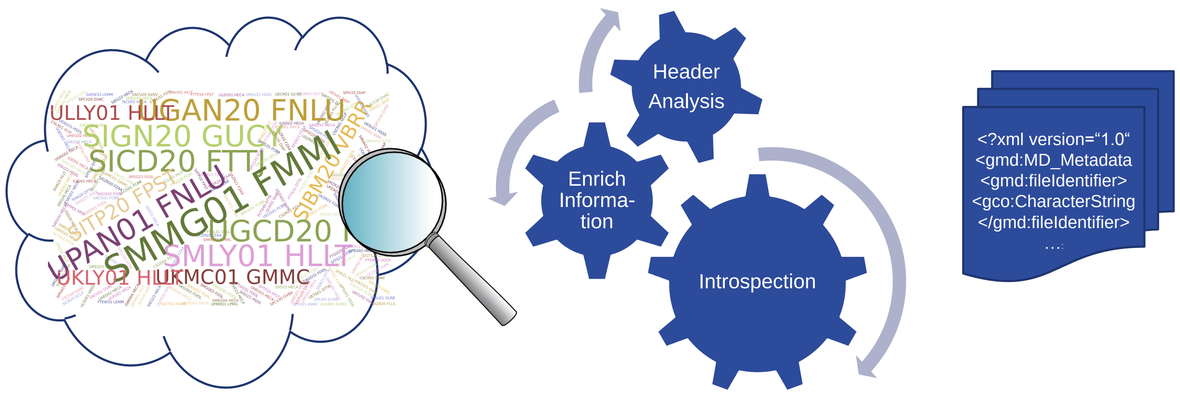
PA13A-1971: Automatic Metadata Generation for Dark Data to Support Information Systems

M. Heene (DWD[1]), T. Büßelberg (DWD), D. Schröder (DWD), A. Brotzer (ask[2]) and S. Nativi (CNR-IIA[3])
[1] Deutscher Wetterdienst – Germany, [2] ask – Germany, [3] Consiglio Nazionale delle Ricerche – Italy

More than 50% of all data and products shared over the Global Telecommunication System (GTS) owned by the weather services are without metadata. As a consequence thereof, these data and products are neither discoverable, accessible nor retrievable by users of the WMO Information System (WIS), GEOSS or National Data Infrastructures even so that most of these data and products are intended for global exchange to support a broad range of users and decision-making processes. These data and products are only accessible within the limited scope of the GTS while they become dark data for Information Systems. The following paper presents a concept and implementation how meaningful metadata for Information Systems can be derived automatically for these dark data. Data and products shared over the GTS follow a formal naming concept the so-called Abbreviated Heading Line (AHL). Based on the AHL the implementation conducts the type of data (e.g. forecast, observation, …), data format and the originating centre of the dark data. In a second step this information is enriched by additional data from knowledge repositories like contact details for each weather service, geographical information for stations, territories and countries, mappings of data format and AHL with theme keywords, and the introspection of the dark data. In a third step the incoming data are monitored for one week to determine the time pattern. The so collected information is used in a last step to generate ISO 19139-compliant metadata for the Information Systems.



Initial Situation

User searches meteorological data – "I want all precipitation data for South-East Europe"

- The Information System only "knows" data and products described by metadata
- More than 50% of all data and products shared over the GTS are without metadata → dark data for Information Systems
- The User can only discover a subset of all available data and products due to a lack of metadata while a significant part of the data and products is not discoverable with Information Systems

Other Usage Scenarios

The approach presented here also works for other scenarios – e.g. we applied it to products which we were initially described in a ticket-system. The ticket-system provides structured data like: theme keywords, geographic bounding box, … . After extracting the relevant information and enriching it with additional knowledge resources like contact details and geographical details we generated ISO 19139-complinant metadata records.

Results

- The suggested procedure is a support tool for NMHS without metadata tools
- It enhances user experience by making more data discoverable and accessible
- It provides Increased interoperability with other systems

Outlook

- The presented approach is easily extendable to other scenarios and data types
- The introspection feature is extendable for other formats like GRIB, NETCDF, …

Automatic Metadata Generation

Data and products shared over the GTS follow a formal naming concept (AHL) – Example "SMMK10 LYPG"

- Step 1 - Derive from the abbreviated heading:
  - SM → Product: Main Synoptic Hour → WMO Formatted Code: FM 12
  - MK → Country: Montenegro
  - LYPG → Originating Centre: NMHS Montenegro
- Step 2 – Enrich initial information with additional knowledge sources
  - WMO Formatted code FM 12 → list of meteorological keywords like: temperature, precipitation, …
  - Country Montenegro → geographical bounding box (first guess)
  - Originating Centre → contact details like address, focal point, web site, …
- Step 3 – Introspection
  - Collect data and products for one week → derive time pattern → temporal keywords
  - Extract station identifier (e.g. 13463) if applicable → station name and coordinates → geographical bounding box
- Step 4 – Generate metadata
  - Use collected information to generate a ISO 19139-compliant metadata record for Information Systems

Contact
Email: markus.heene@dwd.de