# Abstract

The following document outlines two methods for analyzing the search behavior of a GISC. Method 1 analyzes the user input and groups the user input into categories. The users of GISC Offenbach are searching for geographical search terms (cities, countries, …) , domain terms (meteorological keywords, station ids, …) and terms in German language. Method 2 defines four test cases with search terms which were issued against each operational GISC. Method 2 demonstrates that the GISCs produce on the same metadata different result sets. Furthermore method 2 shows that the result set vary between GUI and SRU search.

The document discusses briefly browsing and demonstrates in a next step the data availability on basis with a few control samples. As an appendix JMA and DWD prepared an overview about the message format found in the 24h caches.

## Methods

I used two different methods to analyze the search behavior:

* Log files – GISC Offenbach logs the search terms in anonymized files
* Test cases with search terms issued at each operational GISC

The tests were run two times – first time in November 2012 and beginning of July 2013. The 2012 results were presented at the GISC Seoul Workshop on Jeju Island - at this time only 5 GISCs were operational.

### Method: Log files

A detailed analysis of GISC Offenbach’s full text search shows that users search for the following categories:

* Geographical search terms
	+ Cities: Toulouse, Pretoria, Beijing, Canberra, Offenbach
	+ Countries: Germany, France, Turkey
	+ Europe, Sea, Pacific, …
* Domain terms
	+ Tsunami, Temperature, Biodiversity, Hydrology, Satellite
	+ Station IDs
	+ TTAAii CCCC
* Output formats and codes
	+ GRIB, BUFR
	+ TAF, SYNOP, METAR
* Language Mix & Typos
	+ German “Temperatur” vs English “Temperature”
	+ German “Schneehöhe”, “Sonnenscheindauer”, “Strassenzustand”
	+ “EUEMTSAT” vs. “EUMETSAT”

In contrast to the analysis of November 2012 in July 2013 wildcards like ‘\*’, ‘?’ and logical operators like ‘AND’ & ‘OR’ were more often used.

Furthermore I recognized a significant increase in external monitoring searches – sometimes we have around 500 requests per day where a user/process checks if specific metadata are in the catalogue.

### Method: Test cases

I issued the following searches at each operational GISC

* Pljevlja – City in Montenegro
* Bar – City name (Bar, Marble Bar, …), unit of pressure, …
* Space Weather – the literal “Space Weather” and not somewhere space and weather in the text
* Wild card search: Temperatur\_, Temperatur%, Temperatur\*, Temperatur?

with the following results:

* Pljevlja: 6 GISCs find 4 hits, 1 GISC finds 0 hits (same results for SRU) – 2012 results: 3 GISC find 4 hits, 1 GISC finds 1 hit, 1 GISC find 0 hits
* Bar: 5 GISCs find 280 hits, 1 GISC finds 273 hits, 1 GISC finds 9 hits – 2012 results: 1 GISC 12 hits, 1 GISC 9 hits, 2 GISC 5 hits, 1 GISC 2 hits. SRU: 1 GISC 7914 hits, 3 GISCs 280 hits, 1 GISC 273 hits, 1 GISC 9 hits, 1 GISC 5 hits. The SRU results are different to the GUI results. It seems like that the GISC with 7914 hits masks the search term with wildcards (‘\*Bar\*’). Please note that I haven’t found an easy way to use wildcards in SRU.
* Space Weather: 1 GISC finds 24 hits, 2 GISCs find 23 hits, 1 GISC finds 22 hits, 1 GISC finds 2 hits, and 2 GISC finds 1 hit. 4 GISCs found the record “OSCAT L2A ISRO PRODUCT – OCEANSAT” which contains the word “Space” and “Weather” but not the literal “Space Weather”. 1 GISC holds a record with “Space Weather” which is not synced with the other GISCs. SRU: 2 GISCs 53 hits, 1 GISC 52 hits, 1 GISC 51 hit, 1 GISC 2 hits and 2 GISC 1 hit
* It seems like that 5 GISCs do not support wild card search – at least I was not successfully with my tests and I couldn’t find a hint in the documentation. 2 GISCs support wild cards but with a different notation ‘\*’ and ‘?’ vs ‘%’ and ‘\_’. I haven’t found an easy way to search with wildcards in SRU.

The test cases above were issued against 7 operational GISC. The GISC implementations use 3 different search engines. Already these basic tests show how different the systems react on the same set of metadata (test case 1). A possible reason could be the configuration of the tokenizer used by the search engine – for example Oracle Text requires the configuration of word separators like ‘,’.

Furthermore the tests demonstrate that the same implementation provides different results (test case 2) – a possible explanation is that one GISC holds still an old version from 2012 in its catalogue while the other GISCs perform the search on the latest version of the records. This finding is mentioned in the Metadata Sync paper 5.1 too.

While a search for wildcard is maybe not the most common use case, a search for a literal seems essential. I found only at 2 GISCs the possibility to use wildcards and at 3 GISCs the possibility to issue a literal-search (‘Space Weather’).

The test cases performed with SRU provide two findings:

* The formats of the SRU responses are different. One implementation (4 GISCs) uses a compact format with system internals like db ids, harvested time while the other implementations return the whole metadata record. In case of the compact format it can happen that the user retrieves a response without the search term.
* Test case 2 and 3 show a difference in the result sets of GUI vs. SRU.

### Browsing categories

As highlighted in the document “5.7 User Interface” from Baudouin Raoult, ECMWF only one GISC offers browsing. In addition GISC Tokyo provides a “Select and Browse” interface where a user creates with different predefined categories a basket which he can subscribe in a next step.

GISC Beijing solution demonstrates that building a browse-able tree is possible in a control environment like the GTS-Bulletins but in a heterogeneous environment soon boundaries are reached. The usage of the element <gmd:hierarchyLevel> is used by some DCPCs to build a browse-able tree. The DCPC controls in this case the usage and can build a browse-able tree. In the next version of ISO 19115-1 the element will be deprecated and building up a global tree for a GISC needs manual interaction to map the DCPC categories into the global tree.

GISC Offenbach and GISC Tokyo provide a provider search – the user can limit with this criterion the search in one specific set of metadata.

### Data availability

With a search a user finds data and products he is interested in. A GISC maintains a 24h cache with all data and products intended for global exchange. To get an overview about the data availability I checked the following bulletins (SM<TT>01 <CCCC> and US<TT>01 <CCCC>, with <TT> geographical designator of the GISC and <CCCC> Location Identifier) at each operational GISC. The following graphic shows the results recorded at July 4 2013.

5 GISCs have got all bulletins of the control sample, while at GISC Seoul the bulletins of AMMC, BABJ and EGRR of the control sample are missing. For some reason I found only EGRR bulletins at UK MO.

### Message format

The message format is discussed in detail in the paper from DWD and JMA.

## Recommended Text

Mr Markus Heene (DWD) presented the paper Search (Internal and SRU). The document outlines two methods for analyzing the search behavior of a GISC. Method 1 analyzes the user input and groups the user input into categories. The users of GISC Offenbach are searching for geographical search terms (cities, countries, …) , domain terms (meteorological keywords, station ids, …) and terms in German language. Method 2 defines four test cases with search terms which were issued against each operational GISC. Method 2 demonstrates that the GISCs produce on the same metadata different result sets. Furthermore method 2 shows that the result set vary between GUI and SRU search.

The document discusses briefly browsing and demonstrates in a next step the data availability on basis with a few control samples. As an appendix JMA and DWD prepared an overview about the message format found in the 24h caches.

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