# Semantic web for data infrastructure

## An ontology for metadata mediation

## Introduction

**Earth observation** data are highly **heterogeneous**; anyway, **NetCDF** self-described format is commonly employed for representing those scientific in-situ or remote sensed data, resolving **syntactic** and **structural** heterogeneity problems.

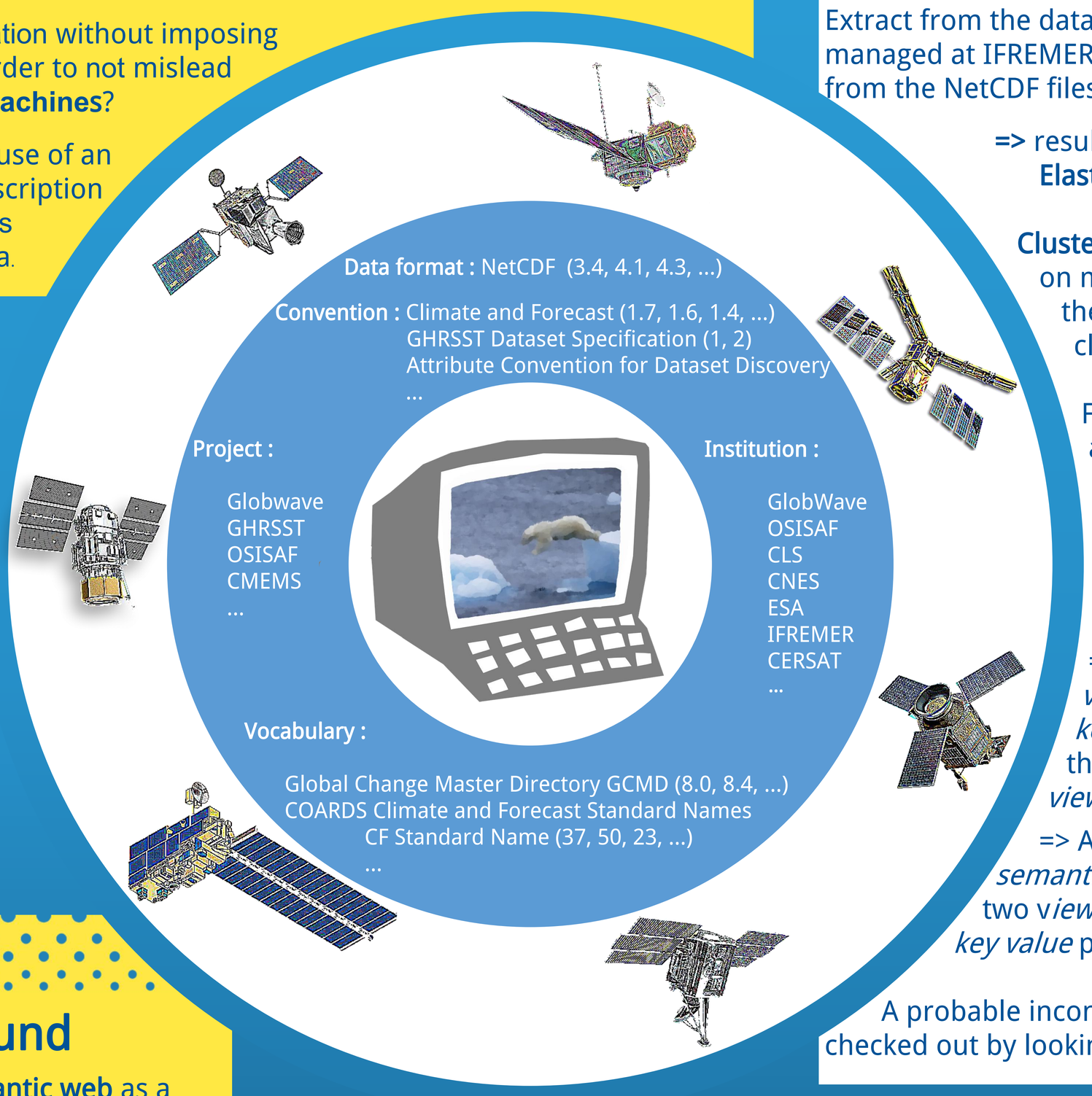
Working groups normalized how and which **built-in metadata** are to be registered : **CF**, **ACDD** conventions, among other project specific ones like ARGO or GDS.

**However**, it still happens that from one project to another the same vocabulary is used for slightly different things, e.g. the processing levels in remote sensing.

This **semantic heterogeneity** is disturbing for multi-project final users and data managers, e.g. to set up an **automatic processing** chain or build a **data catalog**.

How to harmonize information without imposing new conventions and in order to not mislead end users, **humans and machines**?

The hypothesis is that the use of an **ontology** can ease the description of semantic inconsistencies in NetCDF built-in metadata.




## Methodology

To build the ontology, **best practices** of the **semantic web** are followed, among them :

- reuse existing ontology and vocabulary
  - FOAF for person, organization, group
  - GCMD, COARS, CF standard names voc.
- publish online the ontology

To populate the ontology, the focus is made on **satellite observation data** using NetCDF data format. Steps are :

Extract from the datasets produced and/or managed at IFREMER the built-in metadata from the NetCDF files : **attributes** and **values**.

=> results are stored in an **Elasticsearch** database

**Clusters** are built depending on metadata homogeneity, they correspond to the class **Viewpoint.**

For every viewpoint, attributes and values populate the classes **Key** and **Value**

**Rules** are defined to **infer new knowledge** :

=> A *value* is a probable *viewpoint* when the *key value* pair is always the same for a same *viewpoint*.

=> A value has a probable *semantic inconsistency* when two v*iewpoints* use the same *key value* pair.

A probable inconsistency has to be checked out by looking at the definitions.

## Background

The W3C defined the **semantic web** as a web where data are linked one another in order to be more comprehensible for **humans and machines.** It encompasses solutions such as languages, protocols, tools and best practices.

A **graph** of resources is made using **Triples** :

| *subject* | *predicate* | *object* |
|---|---|---|
| IMDIS | is a | Conference |
| IMDIS | hasEdition | 2018 |

Shared knowledge can be modeled using different vocabulary (classes, properties) depending on the complexity of the domain, from a **taxonomy** to an **ontology. OWL** is a much more expressive language than **RDFS**.

## An ontology for viewpoints mediation

### Classes

Viewpoint
Key
Value

### Properties

uses
hasOrigin
isInconsistentWrt

### Usage examples

Viewpoint uses Key => *CF-1.6* uses *Keywords*
Viewpoint hasOrigin foaf => *OSISAF* hasOrigin *IFREMER*
Key hasValue Value => *processing_level* hasValue *L3*
Viewpoint uses Viewpoint => *OSISAF* uses *CF-1.6*

### Inference

Viewpoint isInconsistentWrt Viewpoint
=> *OSISAF* isInconsistentWrt *GlobCurrent*


IFREMER

Contact : dominique.briand@ifremer.fr

Ifremer