ECMWF Data Governance

Maintaining consistency and reproducibility in our meteorological data archive using Data Governance

Sebastien Villaume

Senior analyst in the products team @ ECMWF

Sebastien.Villaume@ecmwf.int

Data Governance workflow: what’s that all about?

- an agreed process to manage decisions about data
- Provide expert guidance when data related decisions need to be made
- collaborate closely with relevant international bodies (such as WMO, OGC and netCDF/CF) to support current standards and develop those which will be used in the future

- **Increased consistency and reproducibility**
- **Increased data interoperability within the organisation and partners**
- **Improved data discovery**
- **Increased usability and reusability of data sets**

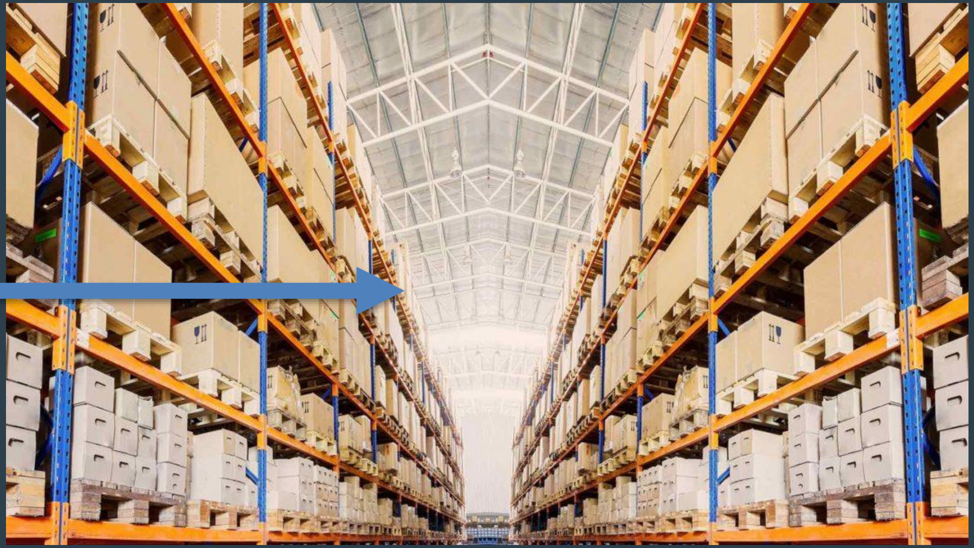
The ECMWF meteorological data archive: MARS

- ~300 PB of data (without accounting for the backups)
- Roughly 400 billions individual records
- In “only” 11.5 millions files
- Few more hundreds of TB produced daily
- Deliver several tens of TB outside ECMWF

Without careful curation and use of controlled vocabularies, the archive would become a mess

© Charlie Cartwright / SWNS.com

Any single record can be retrieved using 10-15 key/values pairs



Retrieving data from the archive

```
Retrieve,
    class=od, # operational data
    stream=enfo, # ensemble forecast
    expver=1, # versioning
    type=pf, # perturbed forecast
    levtype=pl, # pressure level
    levelist=700/850/925/1000, # values of the pressure levels
    date=20191001, # date
    time=12, # UTC time
    step=1/TO/24, # 24 first hours of the forecast
    number=1/TO/50, # 50 ensemble members
    parameters=t/u/v # temperature, u/v components of wind
    target=theFilenameYouLike.out
```

what are the challenges?

- Quite often, researchers do not realise how important it is to define properly the metadata.

- Metadata should be consistent across the archive

- Metadata should stay generic and avoid using community terms

- Metadata should integrate well with other standards to allow interoperability and format conversion

Metadata scoping

- Metadata can be categorised into 3 groups:
  - Core metadata (anything mandatory to decode and use the data):

    parameter, units, grid resolution, validity date and time, packing, projection

  - Indexing metadata (anything that helps categorise the data):

    data producer, forecasting system, model version

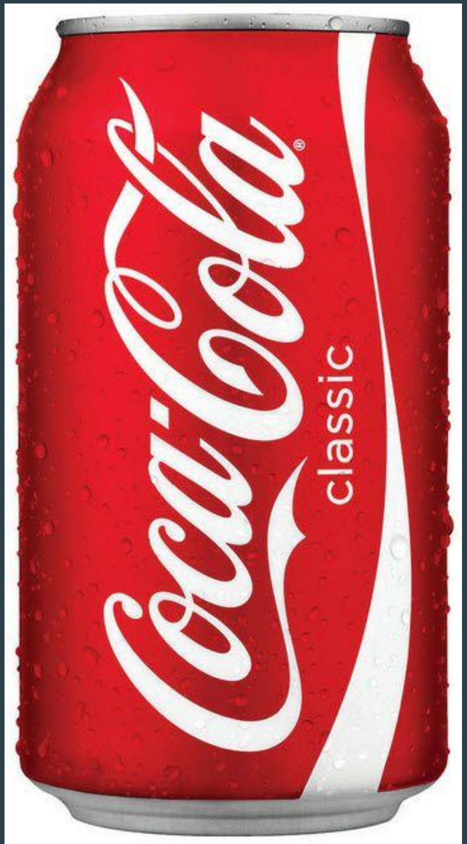
  - Superfluous metadata:

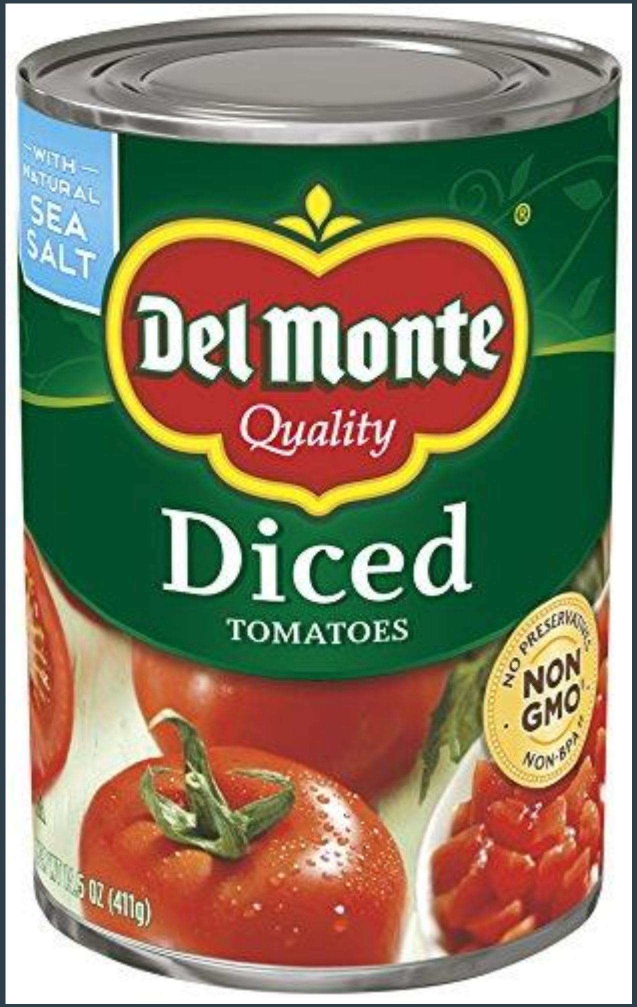
    Any information that is not used to describe nor index the data

Data without metadata is useless !

Data with incomplete metadata is not really better!

data with all the metadata!





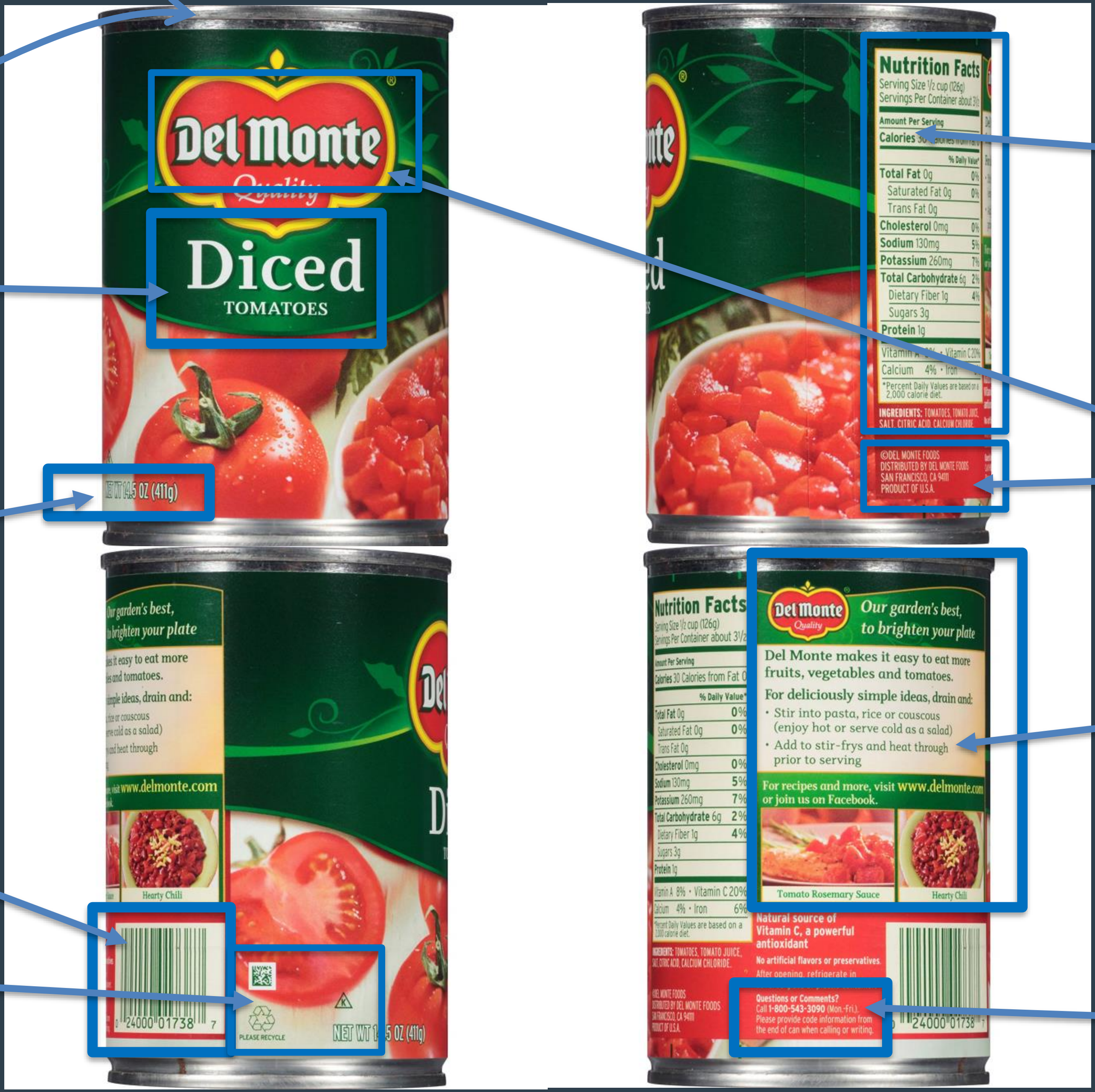
Production date
Expiry date
Batch number

Product name

“units”

identifier

ISO, norms, standards


Del Monte
Quality
Diced
TOMATOES
NET WT 14.5 OZ (411g)
Nutrition Facts
Serving Size ½ cup (126g)
Amount Per Serving
Calories 30
Total Fat 0g
Saturated Fat 0g
Trans Fat 0g
Cholesterol 0mg
Sodium 130mg
Potassium 260mg
Total Carbohydrate 6g
Dietary Fiber 1g
Sugars 3g
Protein 1g
INGREDIENTS: TOMATOES, TOMATO JUICE, SALT, CITRIC ACID, CALCIUM CHLORIDE
©DEL MONTE FOODS
DISTRIBUTED BY DEL MONTE FOODS
SAN FRANCISCO, CA 94111
PRODUCT OF U.S.A.
Our garden's best, to brighten your plate
Del Monte makes it easy to eat more fruits, vegetables and tomatoes.
For deliciously simple ideas, drain and:
• Stir into pasta, rice or couscous (enjoy hot or serve cold as a salad)
• Add to stir-frys and heat through prior to serving
For recipes and more, visit www.delmonte.com or join us on Facebook.
Tomato Rosemary Sauce
Hearty Chili
Natural source of Vitamin C, a powerful antioxidant
No artificial flavors or preservatives.
Questions or Comments?
Call 1-800-543-3090 (Mon.-Fri.).
Please provide code information from the end of can when calling or writing.
24000 01738
PLEASE RECYCLE

Quality Control, Information, Composition

producer

Description or Marketing crap?

Contact info

Indexing metadata: really like in a normal store!

- Indexing works the same way

Departments (Grocery)

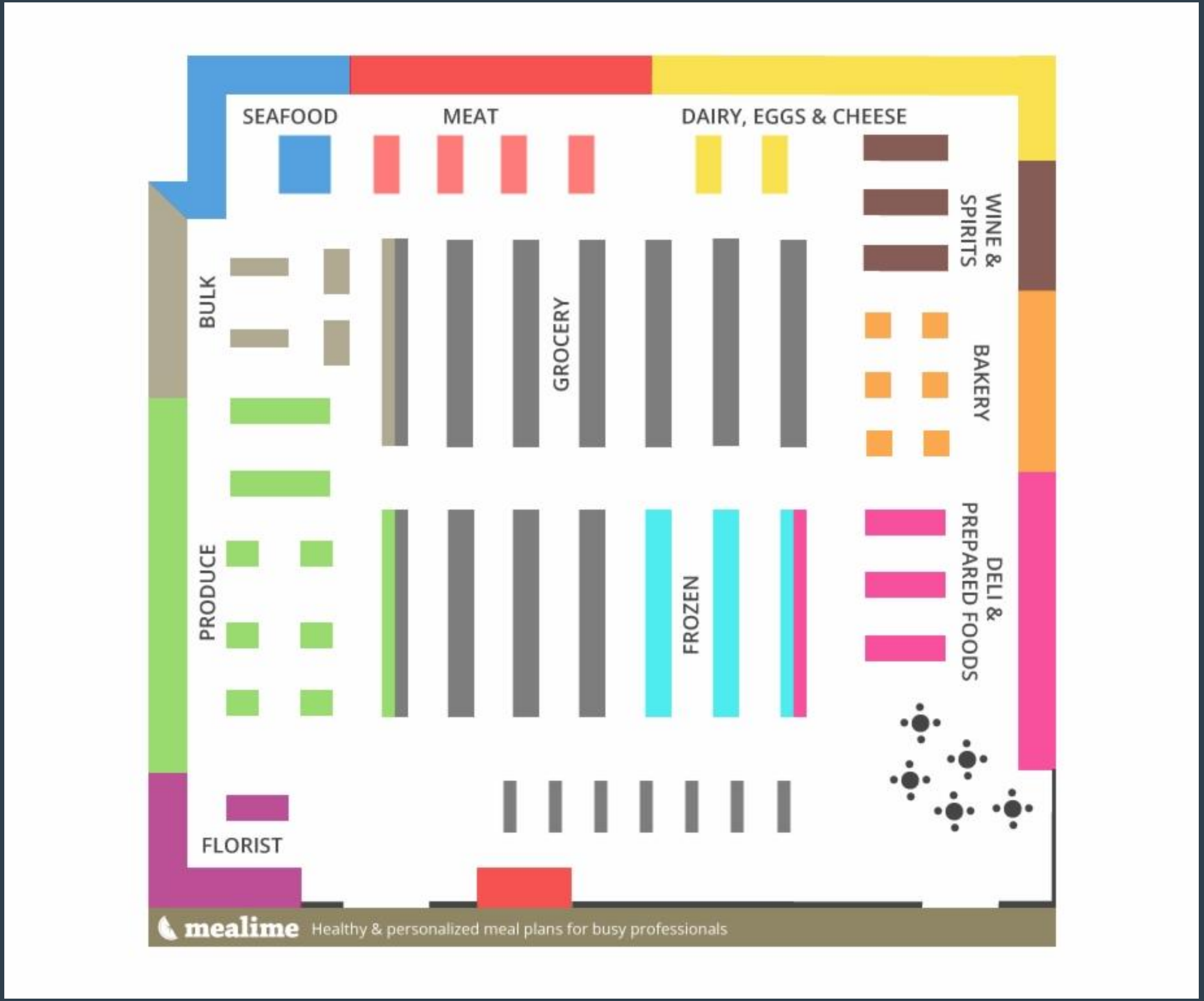
Aisles (canned food)

Shelves (canned vegs)

Category (canned tomatoes)

Sub-Category (canned diced tomatoes)

product (canned diced tomatoes from a specific brand)



Back to the Data Governance Workflow

- Example of requests we receive:
  - New parameters:

    “Probability of Precipitation for the following thresholds 25mm/24h, 50mm/24h and 100mm/24h”

    “mixed layer convective available potential energy in the lowest 50 hPa”

  - New concepts:

    Possibility to encode source/sink for the atmospheric composition model (source of emissions, carbon sinks, etc.)

    possibility to encode parameters on soil, snow and sea-ice multilayers

  - New datasets:

    allocate identifiers for new datasets or new types of data

Encoding in GRIB2 : “probability of total precipitation of at least 25mm”

Discipline	0	Meteorology
parameterCategory	1	Moisture
parameterNumber	52	Total precipitation rate
typeOfStatisticalProcessing	1	Accumulation
typeOfFirstFixedSurface	1	surface
productDefinitionTemplateNumber	9	Probability forecasts at a horizontal level or in a horizontal layer in a continuous or non-continuous time interval
probabilityType	3	Probability of event above lower limit
scaledValueOfLowerLimit	25	Threshold value
scaleFactorOfLowerLimit	0	No scaling

Encoding in GRIB2: source/sink for atmospheric composition

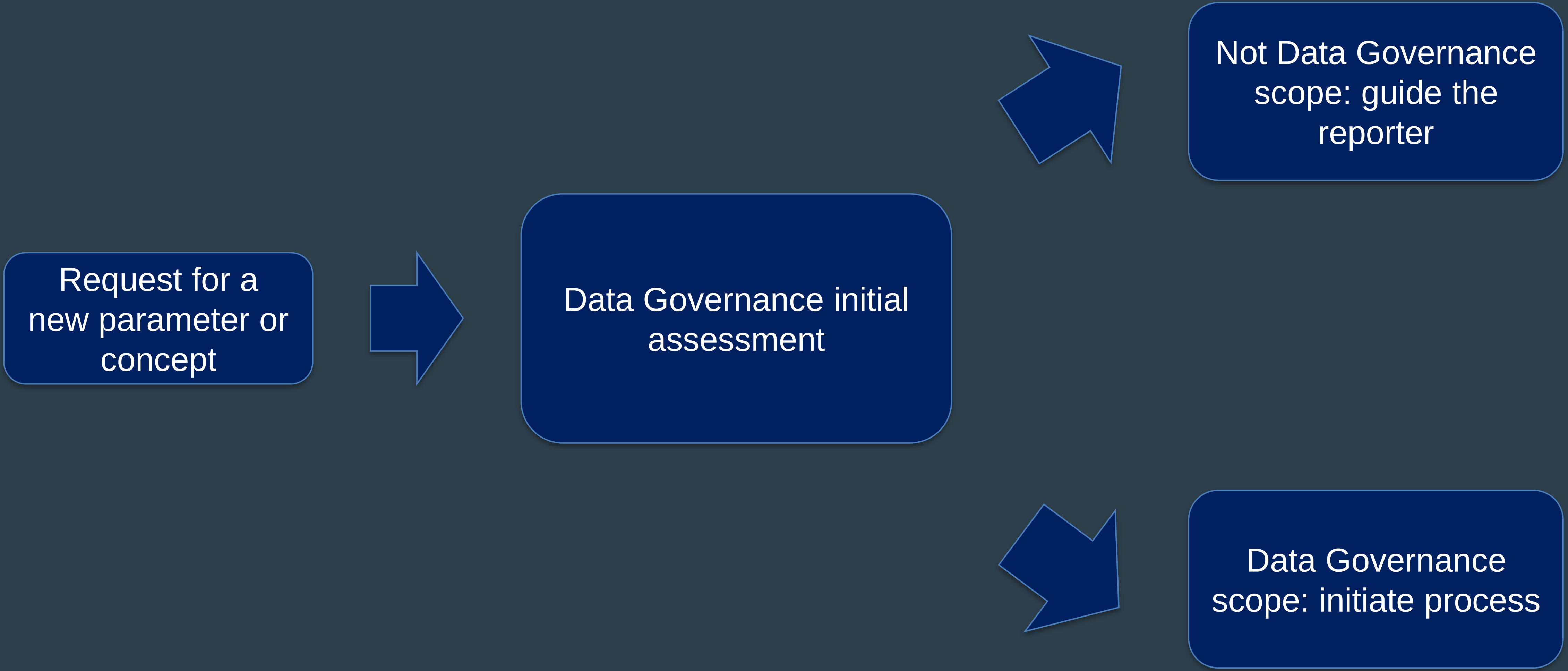
Octet No.	Contents
10	Parameter category (see Code table 4.1)
11	Parameter number (see Code table 4.2)
12–13	Atmospheric chemical constituent type (see Code table 4.230)
14	source, sink or chemical/physical process (see Code table 4.238)
15	Type of generating process (see Code table 4.3)
16	Background generating process identifier (defined by originating centre)
17	Analysis or forecast generating process identifier (defined by originating centre)
18–19	Hours of observational data cut-off after reference time (see Note)
20	Minutes of observational data cut-off after reference time
21	Indicator of unit of time range (see Code table 4.4)
22–25	Forecast time in units defined by octet 20
26	Type of first fixed surface (see Code table 4.5)
27	Scale factor of first fixed surface
28–31	Scaled value of first fixed surface
32	Type of second fixed surface (see Code table 4.5)
33	Scale factor of second fixed surface
34–37	Scaled value of second fixed surface

Encoding in GRIB2: source/sink for atmospheric composition

Code table 4.238 - source, sink or chemical/physical process

Code	Name
0	Reserved
1	aviation
2	lightning
3	biogenic sources
4	anthropogenic sources
5	wild fires
6	natural sources
7	volcanoes
8	bio-fuel
9	fossil-fuel
10	wetlands
11	oceans
12-191	Reserved
192-254	Reserved for local use
255	Missing

The Data Governance Workflow : initiating the process



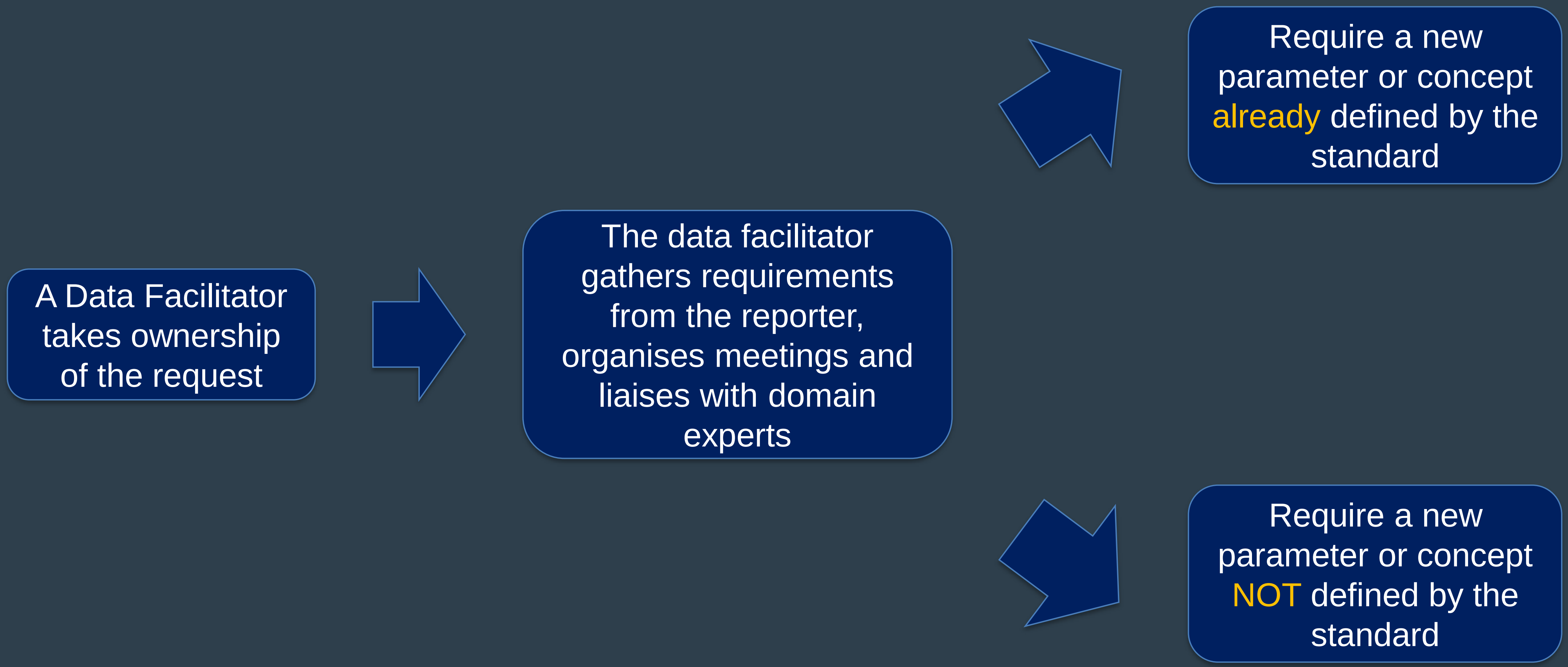
Request for a new parameter or concept

Data Governance initial assessment

Not Data Governance scope: guide the reporter

Data Governance scope: initiate process

The Data Governance Workflow: core work



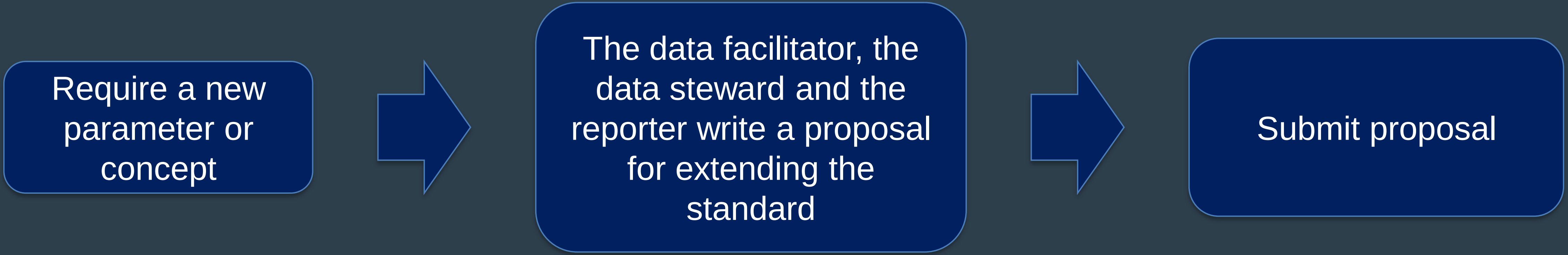
A Data Facilitator takes ownership of the request

The data facilitator gathers requirements from the reporter, organises meetings and liaises with domain experts

Require a new parameter or concept already defined by the standard

Require a new parameter or concept NOT defined by the standard

The Data Governance Workflow: extend the data format



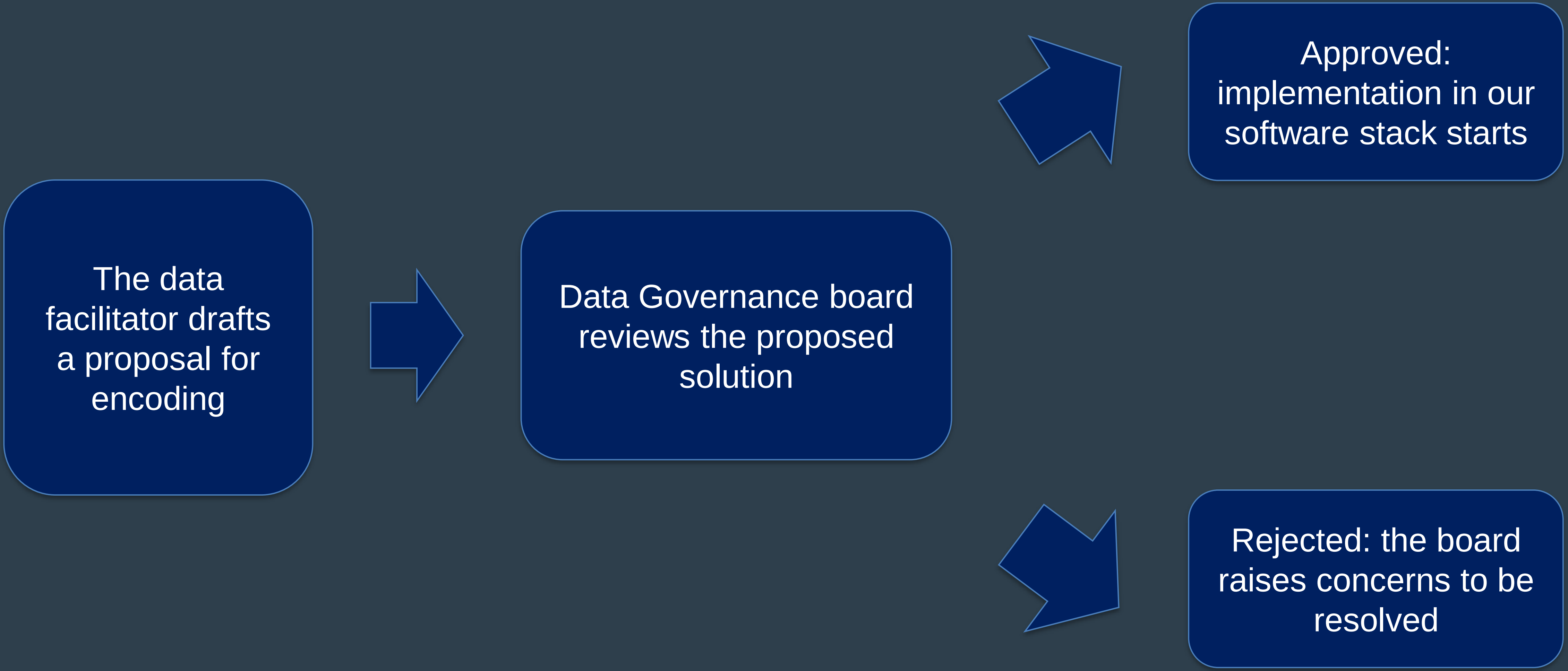
Require a new parameter or concept

The data facilitator, the data steward and the reporter write a proposal for extending the standard

Submit proposal

GRIB and BUFR formats are maintained by WMO team “Inter Programme Expert Team on Code Maintenance” (IPET-CM).
Marijana and myself are members of the team.

The Data Governance Workflow: propose encoding



The data facilitator drafts a proposal for encoding

Data Governance board reviews the proposed solution

Approved: implementation in our software stack starts

Rejected: the board raises concerns to be resolved

The Data Governance Workflow: implementation

- The solution is implemented is our software stack:
  - Encoding/decoding in ecCodes
  - Entry in the Parameter Database: paramDB
  - Interpolation in MIR
  - Plotting with Magics
  - Archiving in MARS (clients and servers)
  - model output (IFS, NEMO, etc.)

  - Once this is done and tested, the case is closed ☺

Concluding remarks

- Metadata is as important as the data

- Implementing the Data Governance process has proven to be very valuable for ECMWF.

- 200+ cases looked at since we started in January 2017