

DATA description : formal vs informal

Authors : Béatrice LARZUL and Denis MINGUILLON

CNES/DCT/PS/TVI

18 avenue E. Belin

31401 Toulouse Cedex 4

E-Mail: beatrice.larzul@cnes.fr denis.minguillon@cnes.fr

CONTENTS

STATE OF THE ART

NON SPECIALIZED TOOLS

Advantages

Drawbacks

DEDICATED TOOL

Advantages

Drawbacks

CONCLUSION

LINKS

STATE OF THE ART

Concerning data compliant with a standard (FITS for Astronomy data, PDS for planetary data, CDF for spectrometry data etc...) the data format is more or less imposed by the standard. Tools are available to create and access such data. So, there is no need for an additional description.

For other kinds of data, not compliant with a published standard, a description is needed do be able to create or access such data.

Very often this description is elaborated with usual known tools such as Microsoft Word or Excel, etc...

Sometimes this description is achieved via dedicated tools (designed for data description).

The question is : "what is better?..." use of these non dedicated tools or use of dedicated tools

The present paper tries to compare these two ways to produce this description.

The following pages will try to highlight the advantages and/or drawbacks of each approach.

NON SPECIALIZED TOOLS

Advantages.

1 -Everything is allowed.

As much particular data could be, it will be possible to write a free text description.

The accuracy of this description is up to the user, no constraint.

All data properties (name, meaning, range, size, owner, context, reliability, etc) can be formulated as wished.

2 - Everybody has access to such tools and, usually, masters the user interface. No specific skills are required.

The description documents are easy to exchange, with the readers using the same tools.

Drawbacks

1 - With non specialized tools no help is provided to describe complex data.

Variable formats such as

- arrays with a dynamic size deduced from other data.,
- machine dependent data,
- optional data the existence of which depends on other data
- lists of an unknown number of item

2 - No control of data types (e.g. : range vs. size).

No help to associate enumeration values to their encoding

3 - No control of the rules constraining the variability
(rules expressed in free text cannot be checked)

4 - No help to take into account data changes (painful updating).

(e.g. : reordering data generates many changes on the output document : no help to reorder it)

5 - No way to deduce read/write software functions from free text.

DEDICATED TOOL

The study of the advantages and the drawbacks cannot be done on the basis of an ideal an virtual dedicated tool. All the knowledge presented hereafter has been acquired while using an existing tool. This tool is a workbench named BEST. BEST is an open source software that can be downloaded at the following address : <http://logiciels.cnes.fr/BEST/EN/best.htm>

Advantages

1 - Early controls

Some controls can be done on the fly.

- an attempt to define an 8 bits field with a maximum value superior to 255 will be rejected (avoiding a later error in a user software).
- defining a record with no component will generate a warning (completeness check)
- defining a data size located behind the concerned data will be rejected
- correctness of variability rules will be ensured by the rule elaboration process
- deletion of a data involved in a variability rule will be rejected (ensures the consistency of the description)
- etc...

2 - Easy updates

- Reordering data can be done (easily using a scrollbar) at any moment. This reordering is checked. (e.g. the tool will check the user is not moving a size behind the data it sizes).
- Adding a data structure similar to one already described will be easy with copy and paste.
- Reuse of an existing description can be made via the notion of named types. A type gathers all the description attributes of a data field. Giving the same type to another field will affect the same description to this other field. A modification of the type will lead to the modification of all the fields belonging to the type.
- Adding a level of hierarchy (regrouping data) to make the description easier to understand can be achieved by capturing with a lasso the fields to be gathered. The field appearing as the parent of the regrouped fields can be easily deleted to come back to the previous state .

3 – Early data simulation

As soon as the formal data description is available another tool of BEST (named DPE standing for Data Producer Editor) allows to run a data simulation driven by this formal description. At the end of this simulation process, data produced is compliant with its description (by construction). This permits to anticipate integration tests, before the operational process supposed to deliver real data is available.

Depending on how much representative simulated data are wished two generation modes are proposed.

- Random . the values are compliant with the description (coding, range, location) but not meaningful with regard to an operational process)
- Driven by directives : a generation directive can be associated to each field of data. Simple laws can be embedded in the description, more complex behaviours can be delegated to an external function (java)

For degraded test purposes, directives can force incorrect values.

4 - Interface Control Document (ICD) production

Once a formal data description is available, it can be used to deduce the corresponding ICD that remains always up-to-date.

As soon as an updated formal description is available, a new version of the document can be automatically generated (much less tedious to maintain).

The tool that produces the ICD from the formal description is named SCRIBE.

5 - Reading and writing data

As soon as a formal data description is available an application programming interface (API), also provided by BEST can be used to read and/or write the described data.

This prevents the user application to have to deal with the data format.

Checks are applied at each request through the API. (e.g. data out of range, conditional data not available, etc...).

The API is available for C, Java, FORTRAN and Ada applications.

Drawbacks

The only difficulty may be the change in habits for users.

People who have to describe data may be worried by the fact of having to master tools they don't know.

This can easily be bypassed with a few demos and coaching during early use.

CONCLUSION

Reading the advantages and drawbacks mentioned here above one may be tempted to try the dedicated tools and compare with his own experience generally based on non dedicated tools. During these trials, support can be obtained via the hotline : east@cnes.fr (as well as for downloading, installation or use)

LINKS

HOTLINE : east@cnes.fr

BEST download : <http://logiciels.cnes.fr/BEST/EN/best.htm>

BEST user manual : http://logiciels.cnes.fr/BEST/EN/best_fichiers/20100125_Best_UserManual_V2.3.1.pdf

SCRIBE download : http://debat.c-s.fr/product/licence_scribe.html