

SFB1182 – Data management

Working with iRODS

Author: Marc P. Hoeppner

Version: 1.0

Date of last revision: 2017-03-17

1	ABSTRACT	2
2	OVERVIEW	2
3	APPLYING FOR DATA ACCESS	2
4	ACCESSING IRODS FROM THE COMMAND LINE (RZCLUSTER)	2
4.1	SETTING UP IRODS	3
4.2	UNDERSTANDING THE IRODS FILE SYSTEM	4
4.3	TYPICAL IRODS USE CASES	4
1)	IGET - GETTING DATA FROM IRODS	5
2)	IPUT - PUTTING DATA INTO IRODS	5
3)	BUNDLING FILES USING IBUN	5
4)	IMETA - ADDING META DATA TO IRODS OBJECT	6
5)	IMETA – QUERYING META DATA FROM AN IRODS OBJECT	6
6)	SEARCH IRODS OBJECTS BASED ON AVUS	7
7)	MOUNT AN IRODS FOLDER	7
5	USING THE WEB INTERFACE	7
5.1	BASIC NAVIGATION	7
5.2	SEARCHING DATA AND CREATING RE-USABLE QUERIES	10
6	RELATED DOCUMENTS	10
7	REVISION HISTORY	10

SFB1182 – Data management

Working with iRODS

Author: Marc P. Hoepfner

Version: 1.0

Date of last revision: 2017-03-17

1 Abstract

This document outlines the usage of the CRC data management solution iRODS.

2 Overview

iRODS is a data management middle ware (sitting between a lower-level hardware infrastructure and some user-accessible front end). In essence, it is a storage application that tightly controls data access and provides powerful tools for enriching data objects with meta data. It exists in parallel to a typical UNIX file system and emulates most of the expected behavior of such an environment. More on that below.

So, in practical terms – what IS iRODS in the context of the CRC1182? For the CRC, iRODS is used primarily to store raw data (i.e. sequencing reads) and enrich it with meta data. This way it should be easier to perform data discovery and push data to downstream sources like public databases and the like. All sequencing data produced within the CRC should be deposited in iRODS and if data is being generated at the IKMB, this will happen automatically. If not, members of the CRC are expected to take care of data archival using the information included in this document as well as the data policy of the CRC or seek help from the INF project.

What iRODS is not: A convenient additional storage space for your projects. iRODS has a fairly poor read/write performance and should really only be used for archive storage. If you wish to work with data that lives in iRODS, you will have to copy it to your local folder. Again, more on that below.

3 Applying for data access

Following the data policy of the CRC1182, data is organized by CRC project and access is granted based on project membership. In order to be added to the access list of a project, please contact Marc Höppner (m.hoepfner@ikmb.uni-kiel.de) and provide both your username on the RZCluster and the project(s) you wish to get access to. If you are not obviously associated with a given project, we will have to check with the PIs of that project to verify that you are eligible for access.

4 Accessing iRODS from the command line (rzcluster)

Disclaimer: The following instructions are fairly use-case oriented. If you are more of a purist and want concise, technical explanations, the iRODS reference manual can be found here:

<https://docs.irods.org/master/icommands/user/>

iRODS integrates with Unix-based operating systems; many iRODS commands in fact mimic typical Unix bash commands, but with an added 'i' in the beginning. In order to start using

SFB1182 – Data management

Working with iRODS

Author: Marc P. Hoeppner

Version: 1.0

Date of last revision: 2017-03-17

the iRODS command line tools on RZCluster, you first need to make sure that iRODS is setup for your cluster account.

4.1 Setting up iRODS

Check if you have an iRODS folder in your home directory:

```
ls $HOME/.irods
```

If yes, great. If not, create it.

```
mkdir $HOME/.irods
```

The iRODS client fetches information about your user credentials and the “zone” you are using. At the CAU there is only one zone (CAUzone), so that keeps things simple.

```
nano $HOME/.irods/irods_environment.json
```

The iRODS configuration file must have to following format (add your RZCluster username where indicated):

```
{
  "irods_host": "irods-icat.rz.uni-kiel.de",
  "irods_zone_name": "CAUZone",
  "irods_port": 1247,
  "irods_user_name": "<YOUR_USERNAME>",
  "irods_authentication_scheme": "PAM",
  "irods_default_resource": "BaseResc-irods-R1",
  "irods_client_server_negotiation":
"request_server_negotiation",
  "irods_client_server_policy": "CS_NEG_REQUIRE",
  "irods_encryption_key_size": 32,
  "irods_encryption_salt_size": 8,
  "irods_encryption_num_hash_rounds": 16,
  "irods_encryption_algorithm": "AES-256-CBC",
  "irods_ssl_verify_server": "none"
}
```

Once the file is updated/created, you can initialize iRODS:

```
iinit
```

And that should do the trick. You are now in your cluster home directory as well as in your iRODS home directory. Let's discuss this some more...

SFB1182 – Data management

Working with iRODS

Author: Marc P. Hoepfner

Version: 1.0

Date of last revision: 2017-03-17

4.2 Understanding the iRODS file system

If you are familiar with a typical Unix file system, you already know that ‘ls’ shows you the content of the current directory and that you can navigate directories with the ‘cd’ command. iRODS adds a second set of commands, including among other things, ‘ils’ and ‘icd’. These commands do exactly what their well-known i-free siblings do, but in a file system that lives side-by-side with that default file system. This creates a situation where you can be in your Unix \$HOME directory, but in some totally different location of your iRODS file system.

```
pwd
/home/sukmb352

ipwd
/CAUZone/home/sukmb352

icd data

ipwd
/CAUZone/home/sukmb352/data
```

But still, when you do a normal pwd:

```
pwd
/home/sukmb352
```

So basically, you will have to keep track of two locations at the same time. Since you won’t be using iRODS all that much, this shouldn’t be problem.

It is also worth keeping in mind that most tools (like bioinformatics pipelines and such) assume file objects to live within the “old-school” Unix file system and use the related commands to navigate around. So in essence, you cannot use iRODS as a file system for analytical purposes. This statement is essential true, but not really because iRODS folders can be mounted into the Unix file system – but the performance is pretty abysmal. More on that later.

4.3 Typical iRODS use cases

First of all, the CRC iRODS folder obviously does not live in your home folder, but directly under the CAU zone:

```
icd /CAUZone/sfb1182
```

You can now list the CRC projects:

SFB1182 – Data management

Working with iRODS

Author: Marc P. Hoeppner

Version: 1.0

Date of last revision: 2017-03-17

```
ils
/CAUZone/sfb1182:
C- /CAUZone/sfb1182/A1
C- /CAUZone/sfb1182/A2
C- /CAUZone/sfb1182/A3
C- /CAUZone/sfb1182/A4
C- /CAUZone/sfb1182/B1
C- /CAUZone/sfb1182/B2
C- /CAUZone/sfb1182/C1
C- /CAUZone/sfb1182/C2
C- /CAUZone/sfb1182/C3
```

(the C at the beginning of each line indicates this object to be a collection, which is iRODS speak for “folder”, more or less).

1) iget - Getting data from iRODS

Say you want to get raw data out of the B2 project in iRODS to perform some analysis on the cluster, you can use the iget command to copy it out of iRODS and onto the Unix file system:

```
iget /CAUZone/sfb1182/B2/raw_data/some_file.tar .
```

This would copy some_file.tar directly to the current Unix directory. Alternatively, you can copy it to some other folder you are not currently in (the same applies to the iRODS locator, by the way).

```
iget /CAUZone/sfb1182/B2/raw_data/some_file.tar
/home/sukmb352/some_other_folder/
```

Why not icp, you ask? Well, ‘icp’ works of course like our old friend ‘cp’, but unfortunately only within iRODS.

2) iput - Putting data into iRODS

Data can be copied into iRODS using the iput command.

```
iput -Dtar some_file.tar /CAUZone/sfb1182/B2/raw_data/
```

Here the “D” flag tells iRODS that the object is of the data type “tar”, which is special to iRODS. You can also omit this flag if the object is not a “tar” archive. But ideally it will be, unless we are talking about actual single files. Bundling a collection of files into a tar archive makes iRODS transactions a lot faster and can be used in conjunction with theibun command, which we will discuss next.

3) Bundling files using ibun

SFB1182 – Data management

Working with iRODS

Author: Marc P. Hoeppner

Version: 1.0

Date of last revision: 2017-03-17

The value of bundling files into tar archive lies in the performance gains when checking data into or out of iRODS. However, once in iRODS, it would be nice to “de-bundle” the files again – or conversely, to bundle a list of files in iRODS for faster transfer to the local file system. This is whereibun can help.

```
tar -chlf mydir.tar -C /x/y/z/mydir .  
iput -Dtar mydir.tar .  
ibun -x mydir.tar mydir
```

In this example, we first bundle an entire folder into one tar archive. Next, the tar archive is copied into iRODS using iput and then, once in iRODS, extracted into the original folder again.

The same thing works in reverse:

```
ibun -cDtar mydir.tar mydir  
iget mydir.tar  
tar -xvf mydir.tar
```

Note that full paths to the various files have been omitted in our examples - we are basically assuming that we are both in the iRODS folder where the folder lives that we wish to tar as well as in the Unix directory where that folder should be copied to. Otherwise you can of course add full path names to these commands and run them from wherever.

4) imeta - Adding meta data to iRODS object

One of the main jobs of iRODS to manage not only data but also metadata – information that better describes an object – how it was produced, who it belongs to and anything else that could be of use at a later time. Meta data is defined as “AVU” – Attribute/Value/Unit triplets. The CRC1182 data management has defined a standard of attributes that should be followed as closely as possible to enable easier data discovery.

Adding data to an iRODS object works like so:

```
imeta add -d /CAUZone/sfb1182/<your_project>/<your_file>  
MAIN_CONTACT_NAME "Klaus Müller" "string"
```

This specifies that the attribute “MAIN_CONTACT_NAME” should be added to the file, using the value “Klaus Müller”, which we define to be a string (i.e. a bunch of text).

5) imeta – querying meta data from an iRODS object

We can now retrieve meta data from our object:

SFB1182 – Data management

Working with iRODS

Author: Marc P. Hoeppner

Version: 1.0

Date of last revision: 2017-03-17

```
imeta ls -d /CAUZone/sfb1182/<your_project>/<your_file>
```

6) Search iRODS objects based on AVUs

Say we want to get the raw data for the sequencing library “F13511”:

```
imeta qu -d LIBRARY_ID = F13511
```

And that will return:

```
collection: /CAUZone/sfb1182/B1/raw_data  
dataObj: F13511-L1_S272_L001.tar
```

7) Mount an iRODS folder

So yes, copying things in and out of iRODS seems a bit cumbersome. Keep in mind though that iRODS is meant as an archival storage and therefore should not be used on actively changing data. However, if for some reason you really want to work directly on the iRODS folder without moving things around, you can do this:

```
irodsFs <local_folder> -o max_readahead=0
```

This will mount the iRODS folder you are currently in to the empty directory <local_folder>.

To unmount the iRODS folder, do:

```
fusermount -u <local_folder>
```

But be warned, working directly on iRODS mounts is very, very slow and not recommended.

5 Using the web interface

5.1 Basic navigation

The compute center provides an easy-to-use web frontend to iRODS for users not comfortable with the Unix command line or who like a less technical platform for browsing their data.

The frontend can only be accessed from within the CAU network, so your computer either needs to be physically located in this network or connect to it through the University VPN (https://www.rz.uni-kiel.de/en/hints-howtos/vpn?set_language=en).

SFB1182 – Data management

Working with iRODS

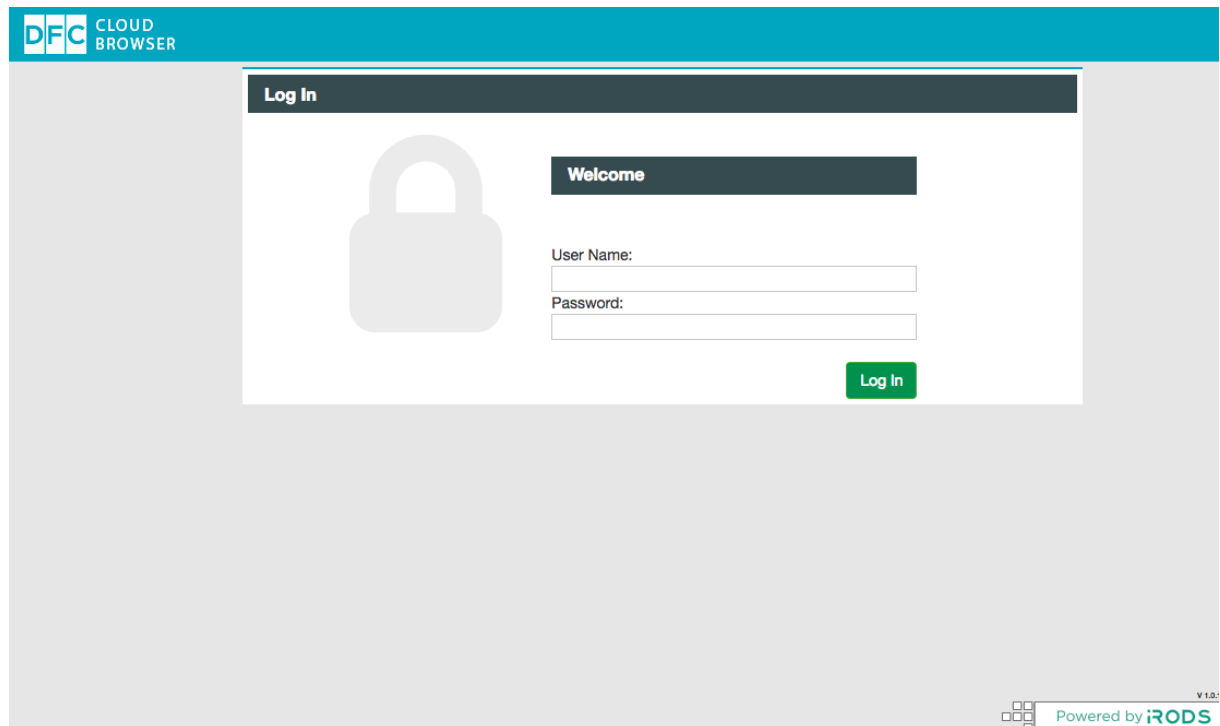
Author: Marc P. Hoeppner

Version: 1.0

Date of last revision: 2017-03-17

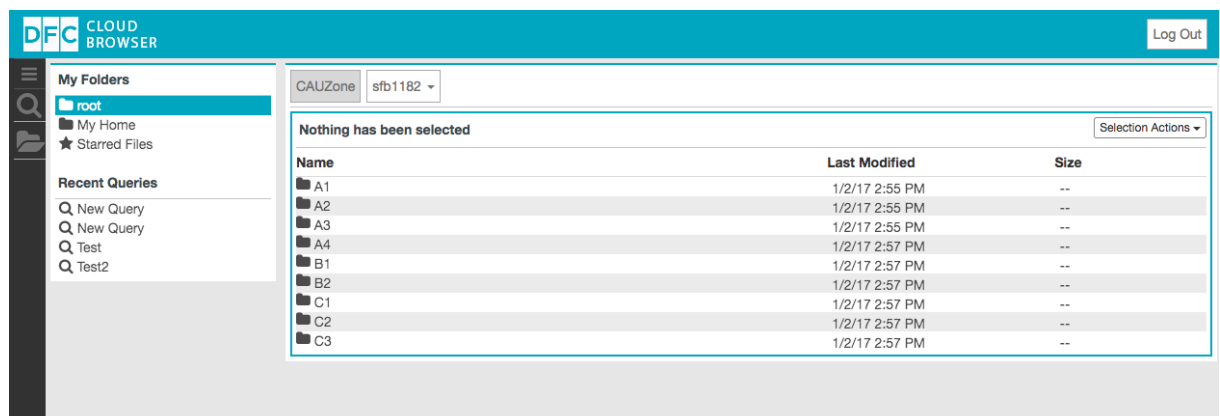
The URL of the front end is:

<https://irods.rz.uni-kiel.de/irods-cloud-frontend>



The username and password are the same as the one that you use for e.g. logging into the cluster or check your emails.

Once logged in, you will be presented with a simple navigation and a list of files and folders in your home directory. Moving from here to any other place within the iRODS system works exactly like any other file browser.



Data objects can be both up- and downloaded through the frontend. To upload, check the menu next to the folder name:

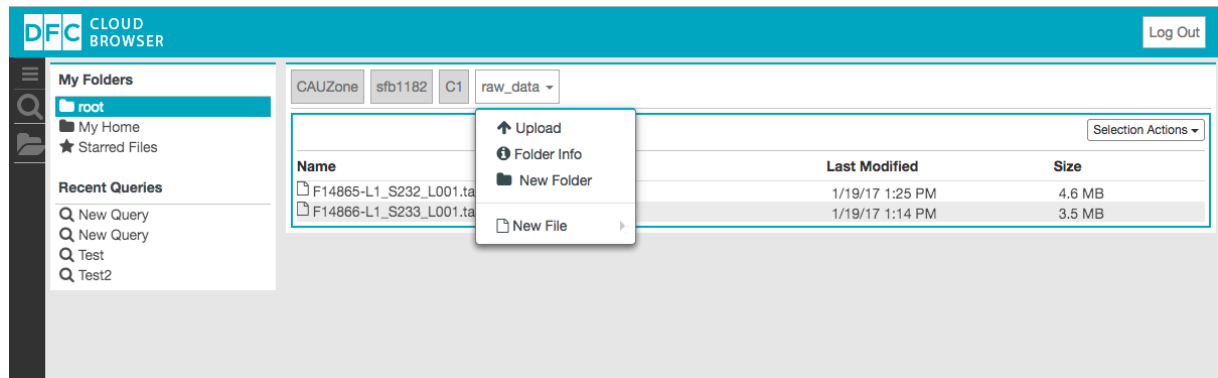
SFB1182 – Data management

Working with iRODS

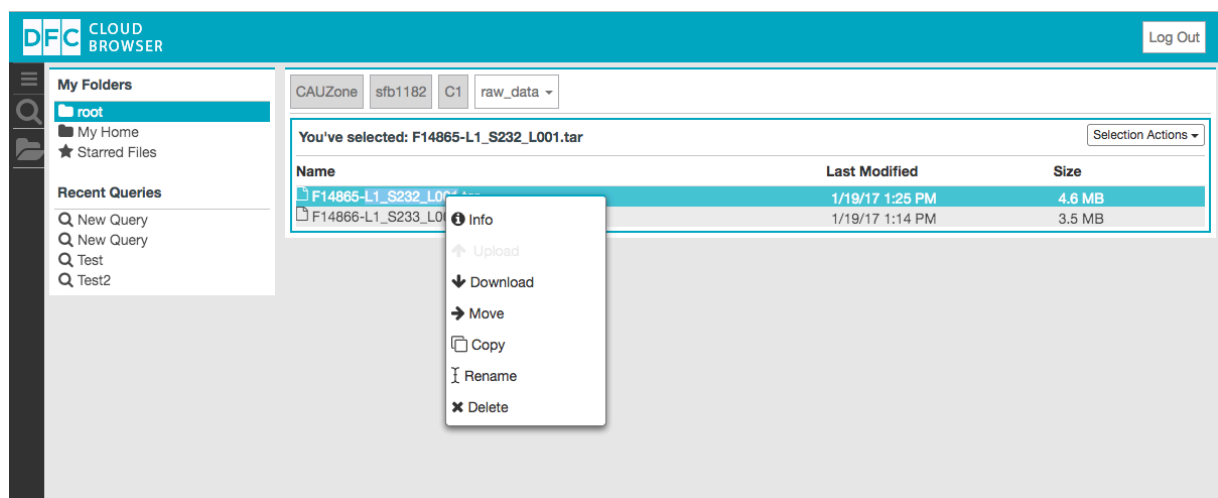
Author: Marc P. Hoepfner

Version: 1.0

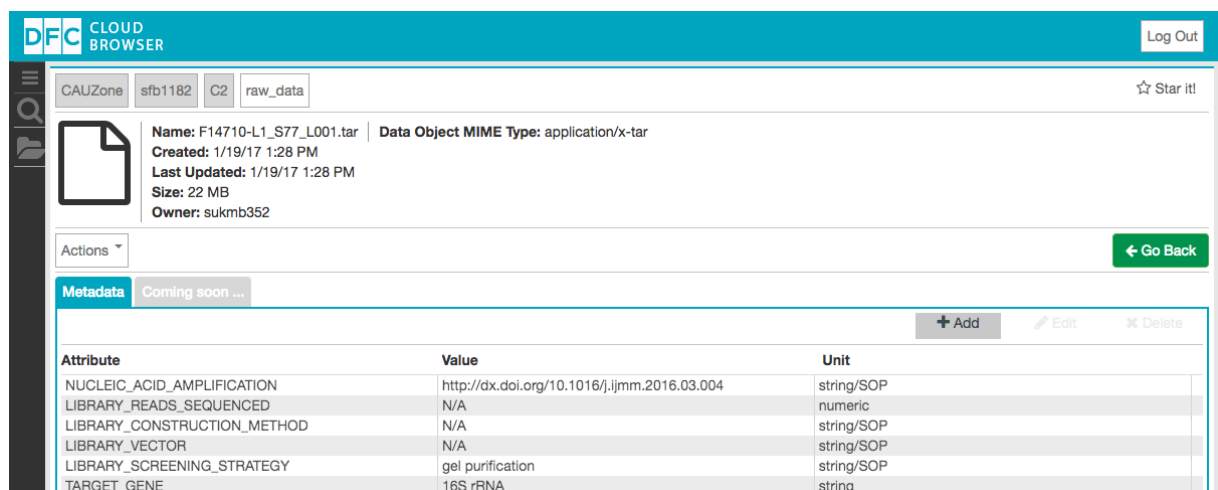
Date of last revision: 2017-03-17



If you like to download files, simply right-click on them.



Finally, the frontend can be used to add meta data to objects via the “info” option and then the “Add” button:



SFB1182 – Data management

Working with iRODS

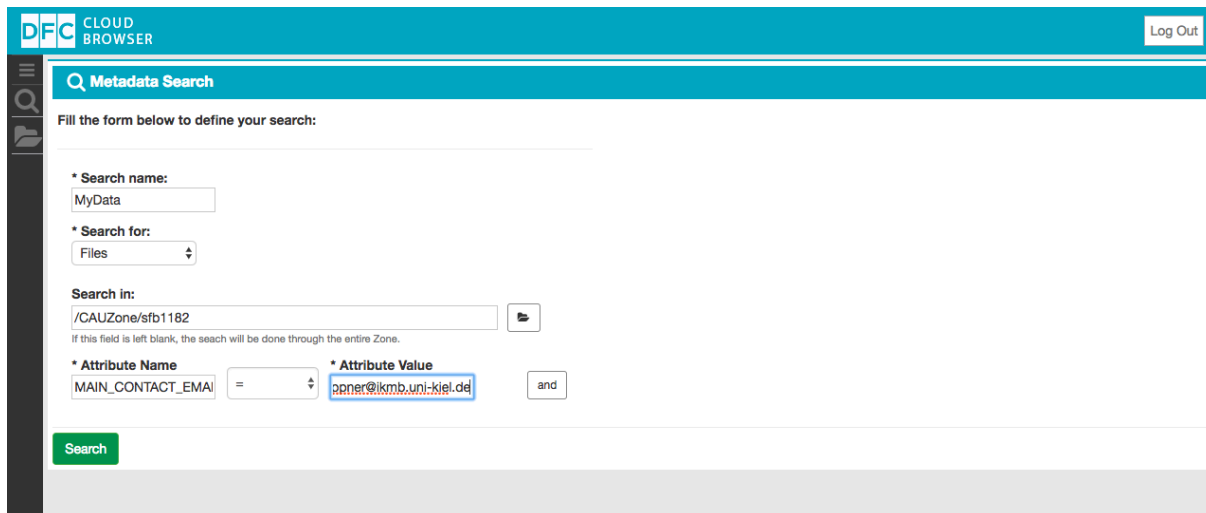
Author: Marc P. Hoepfner

Version: 1.0

Date of last revision: 2017-03-17

5.2 Searching data and creating re-usable queries

The frontend has a separate tab for running searches – indicated by a little magnifying glass on the left side of the screen.



Cloud Browser Log Out

Metadata Search

Fill the form below to define your search:

* Search name:
MyData

* Search for:
Files

Search in:
/CAUZone/sfb1182

If this field is left blank, the search will be done through the entire Zone.

* Attribute Name
MAIN_CONTACT_EMAIL

=

* Attribute Value
hoeppner@kimb.uni-kiel.de

and

Search

Here, we have defined a new query with the name “MyData” and used a meta data key as the sole search criterion (here: The Email address of the sample owner). If we run this query, we will get a list of all files in the CRC folder tree that have this Email address attached to it. Not only that, but the query will automatically be saved so that you can simply re-run it at a later time if more data has been added.

6 Related Documents

See the CRC1182 data policy for reserved meta data keys and descriptions.

7 Revision history

Created by M. Hoepfner, January 23rd 2017