Infrastructure and Design for the 1st CALIFA Data Release

Document created by the DR working group

Released on the 13th of September 2011

Abstract

This document is the result of the first meeting of the CALIFA data release working group in Almeria from the 30th to the 31st of August. The purpose of the meeting was to establish a scheme of the Data Release (DR) infrastructure as the basis for the activities within our working group. We tried to precisely define the goals and expectations of the 1st CALIFA data release (DR1), how we want to achieve them and assigned priorities. The principle design of the user interfaces to access the CALIFA data has been discussed and we will present some ideas that go already beyond DR1. At the end of the document we list the work packages assigned to the members of the working group with a tentative time frame.

1. Members of the DR working group

Current members of the working group are

- Sebastian F. Sánchez (CAHA)
- Bernd Husemann (AIP)
- Jochen Klar (AIP)
- Harry Enke (AIP)
- Sergio Albiol Peres (Teruel Polytechnic School of Engineering)
- Gabriel Fuertes Munoz (independent software engineer)
- Guillermo Palacios Navarro (Teruel Polytechnic School of Engineering)
- Markus Demleitner (ARI Heidelberg)

All members except Harry, Guillermo and Markus attended our first meeting in Almeria.

2. The scope of the first and future data releases

The primary aim of the CALIFA DR1 is to provide public access to the reduced science ready data of the CALIFA survey as well as to the raw data. Whether and when any additional data and high-level products eventually become also public will depend on the decision of the board. However, since we want to use the same infrastructure to provide the collaboration immediate access to important high-level and catalog data, these features will already be taken into account for the DR infrastructure. Thus, it will be easy to permit also public access to advanced data products in future data releases.

3. Infrastructure, work flow and data management

Given the different expertise of the working group members we agreed on a division into certain responsibilities, which matches with the infrastructure and work flow as sketched in Fig. 1. The infrastructure for the CALIFA data can be divided into a few categories:

- **CALIFA Pipeline**
  Concerning the CALIFA data, the pipeline is responsible for processing the raw data to produce science ready datacubes that will be distributed to the public and the collaboration.

- **External catalogs and ancillary data**
  External catalogs of galaxies properties need to be provided by the collaboration for the entire CALIFA sample, which will be crucial for the selection of specific galaxy samples by the users.

- **Database Management**
  A dedicated database system is needed to store not only the actual information and data references for each galaxy, but also some meta information how the information and data was acquired and produced, respectively, the so-called Provenance Management.

- **User interface**
  A sophisticated user interface is needed to allow intuitive access to the CALIFA data based on the information in the databases via dedicated web pages.

- **Virtual Observatory tools**
  A second option to access the data should be the Virtual Observatory (VO) that has is own standards, databases and search tools along which the science ready data will be published.

Within this framework we assigned the following responsibilities to the group members:

1. Group Coordinator: Bernd
2. Pipeline development: Bernd and Sebastian
3. External catalogs: Sebastian
4. Database structure and management: Jochen and Harry
5. User interfaces of DR1 web pages: Sergio, Gabriel and Guillermo
6. Virtual Observatory: Markus

Besides the highest data quality to deliver, the most crucial parts of the data release is the database system and the user interfaces to the data. During our meeting we spend most of the time extensively discussing these two aspects.

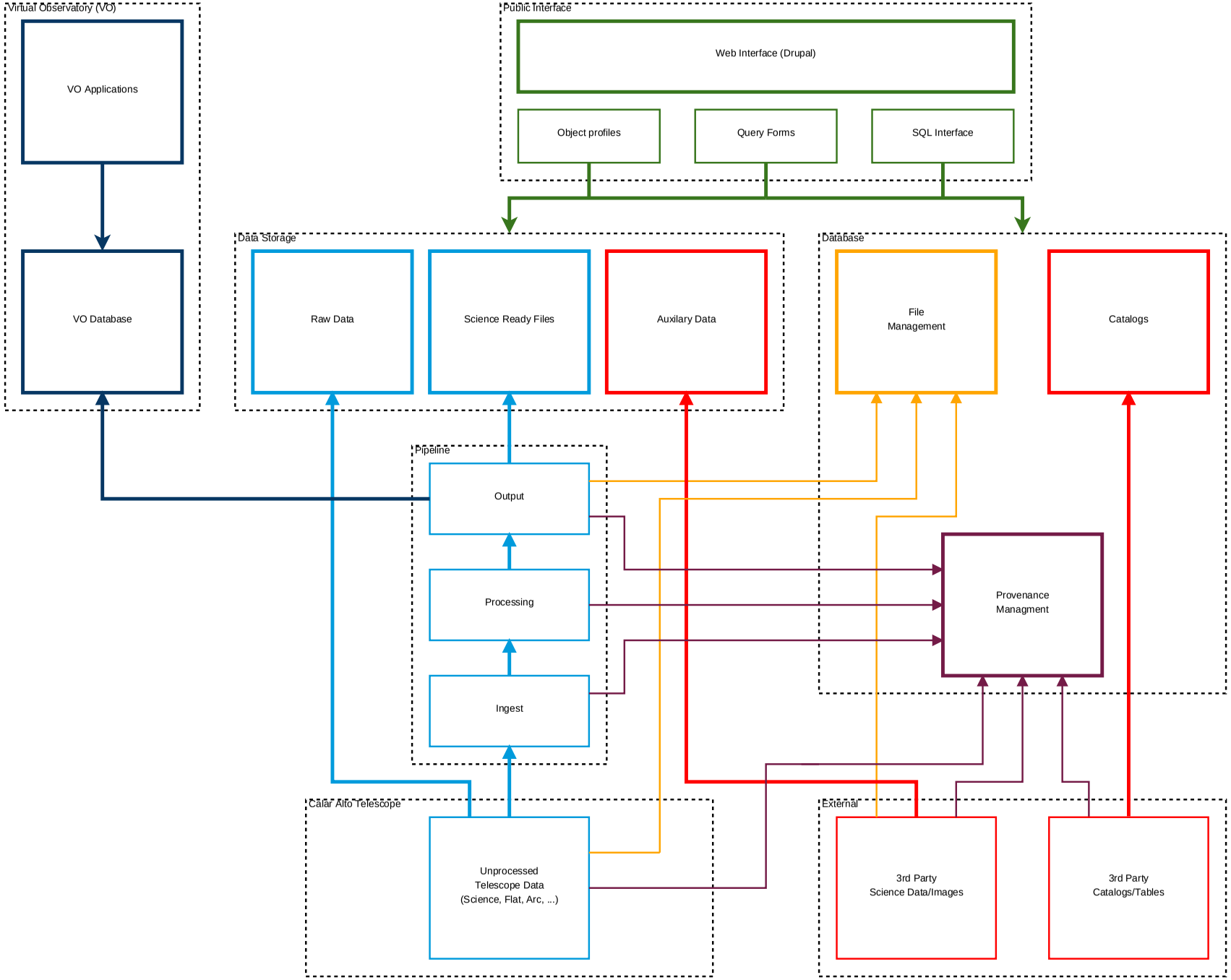

Figure 1: Diagram of the main constituents involved in the Data-Management for CALIFA DR1 including the corresponding flow of data. The different colors depict associated parts of the data flow: *Light Blue*: CALIFA Data; *Red*: External Data; *Green*: Public Interface; *Dark Blue*: Virtual Observatory; *Orange*: File Management; *Purple*: Provenance Management The thick arrows indicate the actual flow of science data, while the thin arrows indicate database ingest or logging operations.

4. Outline of the database

The main scope of the database is to store all the information we acquired related to the CALIFA data itself, and about all the ancillary data collected for the CALIFA sample of galaxies. While the public should not have full access to all kind of data in the first place, some information will be required or useful in order to allow a proper search for certain galaxies according to their properties.

One problem in the database design is that its structure needs to be dynamic as new tables of data will be subsequently added to the database. Thus, we decided to split the database into three parts which we describe more detailed below. While the technical details of each database still needs to be figured out, their purposes and relations are well characterized.

4.1. Provenance management

At present there is already a database for the pipeline reduction which stores the full FITS header information of the reduced CALIFA data in which certain parameters and information are stored during the pipeline reduction process. We decided to continue this scheme and to integrate this in a common provenance database until a different solution is demanded. However, the pipeline needs to make sure that no important information concerning the raw file association as well as process and quality control parameters are missing in the FITS header. Otherwise, the database could not be rearranged without creating a new version of the pipeline to provide a proper "book keeping".

Additionally, the collaboration will generate a lot of external catalogs to characterize the CALIFA sample based on other surveys and later on the CALIFA data itself. The provenance management will keep the references to available tables, their content description like the meaning and units of the table column, how the information has been generated and if the table is public or only accessible for the collaboration. Therefore, new catalog tables can dynamically be added to the database at any time together with assigned permission for different kind of users.

4.2. Catalog database

All the CALIFA sample characterization will be stored in several tables within the database. The content of each table should be focused either on a specific data-set/survey, e.g. NVSS radio information, GALEX photometry, or SDSS photometry or the results of a particular data analysis carried out within the collaboration like a dedicated morphological, spectral or photometric analysis. The creation of a large table collecting information from different kind of information is *strongly disfavored*. However, views that display the content of various related columns of different tables can always be created for convenience.

In order to allow an easy and almost automatic ingest of catalog information into the database and to prevent the loss of meta information, we agreed on a dedicated table format as described in appendix A, which is mandatory to use by the collaboration *from now on*. Although this may introduce some additional work, we think that these specific table formats will rather guide the members of the collaboration to produce tables with complete information.

Currently, we expect that the PI of CALIFA will make the final decision when a particular catalog table is ready to become publicly available. Within the collaboration, tables will be

visible much earlier so that any problem with the content can already be figured out by the collaboration itself and particularly by the sample characterization group.

4.3. File database

The ancillary data like images or spectra from other surveys like SDSS, NVSS, WISE, etc. that will or have already been collected by the collaboration for CALIFA needs to properly archived. Currently, they are placed in separate directories on the FTP which can be retrieved through the WIKI. However, the naming of these files is not always very convenient and data for particular objects can hardly be identified.

Within the DR framework this problem will be solved with a dedicated file database containing tables for each ancillary data set. These tables link at least the path of a particular ancillary data file to the corresponding CALIFAID of the galaxy. This is then completely independent of the file naming and an automatic delivery system can always be adjusted later on to change the name for the download file to a certain file name convention.

However, this means that the collaboration needs to prepare a table using the same convention as for the catalogs which provide the proper linking between CALIFAIDs and files of any ancillary data set.

5. File Server and potential CALIFA data centers

Since we will have a large amount of data that should be accessible only by the collaboration and not by the public, we decided to run two different file servers for security reasons. The file server containing only the data for the collaboration can be protected by authentication so that an anonymous access is impossible even in the case that the link to certain files became public.

During the meeting we discussed the possibility to create a few dedicated CALIFA data centers at distinct places, like Madrid or Potsdam, in addition to Calar Alto. On the one hand, this would ensure a fast and continuous access to the data even if one data center is temporarily not available. On the other hand, the data centers would serve as backups of the CALIFA data for each other. The decision about the creation of CALIFA data centers will be taken by the Board, but a strong preference for it was expressed by the DR working group. However, a special database of the reduced CALIFA data needs to be created for the VO service anyway.

6. Data release web pages and user interfaces

The new CALIFA website based on the Content Management System (CMS) "Drupal", based on `PHP`, replaced the old site by now. We decided that dedicated web pages will be added in the CMS to access the CALIFA data instead of creating a special web server for the DR. How these additional web pages can be added to the CMS depends on the capabilities of Drupal to communicate with an external database. In case this is possible, the web pages will be dynamically filled based on the information in the database given a certain page layout predefined in the CMS. Otherwise, `PHP` scripts inserted into the CMS will do the job, which would be slightly less elegant and requires more work for maintenance and future development.

6.1. Authentication

Authentication for the DR is quite important, since the public and the CALIFA collaboration will share the same web sites to access public and restricted information, respectively. The following option were discussed:

- personalized authentication for everyone with different permissions
- anonymous access for the public and personalized authentication for collaboration members
- anonymous access for the public and a common authentication for all collaboration members

The first option was withdrawn cause the public should have the most easiest access to the data in the spirit of a legacy survey. Whether collaboration members need to sign up individually or a common account should be created was strongly debated. Finally, we decided to use a *common authentication for the collaboration* for the sake of simplicity.

User roles will be assigned to each content within the CMS to govern their visibility depending on the user permissions. The displayed information on the DR web sites will then automatically depend on the authentication of user being one big advantage of a CMS. Therefore, the authentication procedure can completely be handled within the framework of the CMS without additional work.

In addition to these two user groups for the DR access, there will be two other kind of users. Obviously, there will be one or two person signed as *Administrators* for the CMS with a personalized account and unrestricted permission to change and add content. A reduced number of people will serve as *Editors* with personalized accounts and restricted permissions to change content in certain areas.

6.2. Explore pages for each CALIFA galaxy

To present the CALIFA data and galaxies, a dedicated page with the characteristics of each galaxy will be laid out. The layout of this page and its actual content, e.g. which maps will be shown as thumbnail images and which catalog information should be displayed, needs to be defined separately based on the available information until the first data release.

A dedicated group of a few collaboration members will be created specifically for this task. The content of this page maybe extended for each subsequent data release and information shown to public users and collaboration members should be different. Several links to CALIFA and ancillary data will be available in a menu at the left side of the explore page for direct download of single files. The CMS will handle which data links are available for public and collaboration users, respectively.

6.3. Search interfaces

To search for specific CALIFA data, we agreed on a few basic search interfaces that need to be available for the CALIFA DR1:

- Single object search by name or position
- Multiple object search by uploading an ASCII file

- A web-form guided search on specific galaxy parameters
- A flexible search using SQL queries (low priority)

Those different kind of searches will be available on one web page in different tabs.

When searching on the object name(s), the reference CALIFA name needs to be resolved via NED or Simbad. Our group will explore the possibility to create a corresponding table for the relatively small number of CALIFA galaxies to minimize the dependencies on other web services for our own DR. Otherwise the position is matched with the position of the CALIFA galaxies.

A dedicated web form will be created where constraints on certain predefined galaxy parameters, e.g. morphology, stellar mass, absolute magnitude, color, etc. can be given to find matching CALIFA galaxies. A more flexible search on the entire database via SQL queries should be available at least for the collaboration at some point. Given the probably still restricted amount of secure catalog information by the time of DR1 and considering the required security features to protect the database, we assigned a rather low priority to it. This SQL search interface may thus be postponed to DR2.

If the search on a single galaxy is successful, the user will be forwarded directly to the corresponding galaxy explore page. In all other cases, a web page with a list of the matching CALIFA galaxies will be returned. Each row consists of a tick box for data-set selection, thumbnail images of a reconstructed broad-band PPAK image and the integrated PPAK spectrum, the name of the galaxy as a link to the explore page of each galaxy, followed by links with accompanied tick boxes for CALIFA science files and ancillary data files (depending on the user role). With select all and unselected all button on top of each column the user can easily select all galaxy data for download if no specific sub-selection is demanded. At the bottom of the page the user can press a button to retrieve an ASCII file used for mass download of the selected data with `wget`. Depending on the users permissions, the CMS will create the proper data links in the ASCII file that point to one of the FTP file servers.

A static result page of this kind will be in the system to allow selection and retrieval for all CALIFA galaxies without requiring one of the described queries.

6.4. Future development

The above queries are quite standard, but to explore the data of the CALIFA survey even before downloading it requires interactive tools within the DR web page. Our group has pointed out two services that we think would significantly improve the presentation of the CALIFA data:

1. A *Cube Viewer* on the web page to explore the content of the data cube
2. An *interactive diagram* to visually select specific galaxies or galaxy samples within a 2D parameter space

In order to integrate these tools into the web page calls for new Drupal plug-ins to increase its functionality. The `HighCharts` plug-in was already identified to provide the needed functionality. We emphasize that these services do not have a high priority and will certainly not be available for DR1, since the technical feasibility needs to be explored first.

7. Access through the Virtual Observatory

Access to the CALIFA science data through the VO is a separate topic and was extensively discussed in our meeting since the infrastructure to deliver the data via the VO is already there. A test case for one of the current CALIFA data sets was presented and how the data can be retrieved and displayed. From the point of an astronomer we have identified some major disadvantages and advantages of the current functionality that the VO provides.

Advantages:

- Complex SQL queries within data set are possible
- Sophisticated VO tools allow an easy quick look and exploration of the data
- Cross-matching with other VO databases

Disadvantages:

- Each spectrum of a CALIFA data set is stored individually, so that no RSS or cube FITS files can be retrieved so far
- The available VO tools are designed for individual spectra and do no allow scientific analysis of 3D data

Why do we want to publish the CALIFA data also through the VO? After a short discussion we agreed that the VO will become one of the most important data mining centers for astronomy in the future. A legacy survey like CALIFA should therefore be part of it. It is clear that the current capabilities do not meet with the scientific needs, especially for 3D data. We anticipate that a close collaboration would be a chance for the VO to improve the infrastructure and tools also for these kind of data that will become more common in the future.

An immediate advantage of the VO is that we have a second channel of visibility for the CALIFA data. Whenever users search the VO for galaxies and find matching CALIFA galaxies, they immediately find a link to our DR web pages. Although the VO does not yet provide the ability for downloading a complete data-set it will at least give a reference. Another long term aspect of the VO is that the data will be kept available for decades, so that access will be possible even when the CALIFA web pages are not maintained anymore.

8. Raw data access

For the CALIFA DR it is also mandatory to make the raw data publicly available. Calar Alto started to run a web service for all the raw data taken from the observatory. This service was initially designed by Sebastian. For the CALIFA DR1 we therefore decided to create a copy of the service including only the raw data of CALIFA and integrate it into the DR1 web pages.

9. Work packages and time line

At the end of the meeting we defined reasonable work packages and assigned them to individuals with a tentative time frame for each task.

- Writing up of the DR Design Document and inform the collaboration (Bernd, 3 weeks)
- Details of database structure + data ingest scripts (Jochen, 4 weeks)
- Transform existing ancillary data to the new standard (Sebastian+Bernd, 4 weeks)
- Create a new server structure at Calar Alto (Sebastian, 2 weeks)
- Check and test the interface between Drupal and databases (Gabriel+Sergio, 3 weeks)
- Create a small working group for the Explore page (Sebastian, 2 weeks)
- Define the content of the Explore page (Sebastian, 4 weeks)
- Implementation of the Explore page into Drupal (Gabriel+Sergio, 2 weeks)
- Integration of the search and download services (Gabriel+Sergio, 4 weeks)
- Copy raw data system into the CALIFA DR1 web service (Sebastian, 2 weeks)
- Produce VO compliant CALIFA data sets (Bernd+Markus, 3 weeks part of pipeline development)
- Discuss the issue of dedicated CALIFA Data Centers with the CALIFA Board (Sebastian)
- Check technical feasibility of interactive diagrams (Gabriel+Sergio)

Additionally, interested people within the collaboration are encouraged to add or improve content for the general CALFIA survey web pages.

The time frame for individual work packages are such that at least a test service of the DR1 web pages are going to be introduced during the 2nd Busy Week to the collaboration. We therefore agreed on the following time line:

- First presentation of DR web pages on 2nd Busy Week to get feedback from the collaboration
- Official release of DR1 in the first quarter of next year (may depend on the pipeline progress)
- DR group meeting shortly after DR1 to plan for DR2

Some of the tasks reach already beyond DR1 so that some part of the work already for DR2 has been started.

A. Format specification for CALIFA tables

A.1. Aims

For the exchange of important information on the characterization of the CALIFA sample as well as to create a dedicated database system for the public CALFIA data release a standard for the format of tables need to specified. The two common table formats widely used in astronomy are ASCII and FITS tables. Here we define some rules and standards for both

table formats the collaboration must follow to ensure a minimum level of homogeneity, prevent loss of information and allow an easy ingest to a database.

To create a table for CALIFA either the FITS or ASCII format may be used. It will automatically be converted to the other format by the Data Release working group. Together with the table an associated ASCII file needs to be created that describes in detailed how the content of the table was generated. Until a dedicated data management system is running for CALIFA both tables formats will be posted on the WIKI.

A.2. ASCII tables

ASCII tables must consist of a header providing required meta data followed by the data section with the actual content of the table.

A.2.1. Header section

The header of the ASCII files must have a form similar to the following example:

```
# AUTHOR: Carlos C. Califa
# SOURCE: CALIFA Collaboration
# DATE: 2011-08-24
# VERSION: 1.0
# COLAPRV: J. Walcher
# PUBAPRV: None
# COLUMN1: CALIFAID, int, , the ID of the CALIFA galaxy
# COLUMN2: CALIFAName, string, , the CALIFA name of the galaxy
# COLUMN3: Name, string, , the NED name of the galaxy
# COLUMN4: RA, float, degrees, right ascension J2000.0
# COLUMN5: DEC, float, degrees, declination J2000.0
```

The different keywords are:

- `AUTHOR`: creator of this particular file
- `DATE`: date on which the table was created (the format is `yyyy-mm-dd`)
- `VERSION`: version of this particular table
- `SOURCE`: sources of the data, e.g. a survey, a collaboration, a paper
- `COLAPRV`: Name of the person who approved the table for internal release, (the default is "None" and will be later changed by the person in charge)
- `PUBAPRV`: Name of the person who approved the table for public release, (the default is "None" and will be later changed by the person in charge)
- `COLUMNn`: description of the column `n`. It must contain 4 fields:
  - name of the column
  - data type of the column (see below for the possible types)
  - physical units of the column (if no unit applies this field is left blank)
  - brief description of the column

A.2.2. Data section

The header is followed by the data section were the different values for each column are separated by commas in a row. Based on the header example above the data section would look like this:

```
1, CALIFA001, IC5376, 0.33241081, 34.52566909
2, CALIFA002, UGC00005, 0.77351248, -1.91383457
3, CALIFA003, NGC7819, 1.10210525, 31.47200775
4, CALIFA004, UGC00029, 1.14060366, 28.30172348
5, CALIFA005, IC1528, 1.27240324,-7.09338998
...
```

A.3. Binary FITS tables

A.3.1. Table FITS header

The FITS header of the table must contain the following keywords similar to those specified for the ASCII header:

- `AUTHOR`: creator of this particular file
- `DATE`: date on which this particular file was created (the format is `yyyy-mm-dd`)
- `VERSION`: version of this particular table
- `SOURCE`: sources of the data, e.g. a survey, a collaboration, a paper
- `COLAPRV`: Name of the person who approved the table for internal release, (the default is "None" and will be later changed by the person in charge)
- `PUBAPRV`: Name of the person who approved the table for public release, (the default is "None" and will be later changed by the person in charge)
- and for each column (`n` represent the number of a particular column):
  - `TTYPEn`: name of the column
  - `TFORMn`: data type of the column
  - `TUNITn`: physical units of the column (if no unit applies this keyword can be omitted)
  - `TCOMMn`: brief description of the column
  - `TZEROn`: NULL value for column (only for integer types), see below

A.4. Reference column

Each table must contain a reference column with a unique identifier in each row. For tables describing primary galaxy properties like morphology or magnitudes for each CALIFA galaxy, the reference column must be the CALIFAID column as the first column in a table.

A.5. Data types

We support six different data types for the columns. In ASCII files they must be specified in the header using the C/C++/Java like identifiers given in the table below. In FITS files the usual formats in the header apply.

ASCII	FITS	Explanation
short	I	16-bit signed integer
int	J	32-bit signed integer
long	K	64-bit signed integer
float	E	32-bit single precision floating point
double	D	64-bit double precision floating point
string	?A	String in ASCII format/Number of characters for FITS

A.5.1. Units

In principle any system of units is accepted for a given column, as long as the values can be expressed by integers or floating point numbers. Right ascension and declination and any other angle must be given in degrees. Hours, minutes, and seconds will not be supported. The coordinates should be provided for the J2000 epoch. In any case indicate the epoch of the coordinates in the unit field within parenthesis, so that it can always be converted if required.

A.5.2. NULL values

If a particular value cannot be provided, a special marker must be used to indicate a `NULL` value:

- for ASCII files: the value is omitted, i.e. two commas appear next to each other
- for FITS files: the used marker depends on the data type of the column:
  - for strings: an empty string is used
  - for float and double: the `NaN` value is used
  - for integer and long integer: a user-give value is used, which must be specified using the `TZEROn` keyword in the FITS header. Of course the chosen value needs to be far outside the range of the other values for this column.

A.5.3. Special characters

In ASCII format, strings must not contain commas or colons, unless the whole string is quoted by double quotes.

A.6. Associated ASCII file

Together with the table an associated ASCII file needs to be created which describes more detailed the creation and origin of the table. Since this will very much depend on the information content of the table, no particular format is defined except that it should start with the three keywords `# AUTHOR:`, `# Date:`, `# Version:` like in the ASCII table file.

The rule should be that it contain as much information to independently reproduce the content of the table, i.e. state the SQL Query to extract information from an external database,

source of data and analysis steps that yield the provided data. Each time a new version of the table is generated this file must to be extended to describe any changes regarding the previous version.

A.7. File naming

The name of the two table formats and the associated ASCII file need to have an identical file name to ensure a unique association except of the extension:

- `.csv` for ASCII tables
- `.fits` for FITS tables
- `.txt` for the associated files