

D3.14: Reliable and efficient access to raw and processed data from *Kepler*: Report on how processed data can be accessed

Introduction

Observational and processed data from *Kepler* are distributed via the KASOC (for Kepler Asteroseismic Science Operations Center) database. In addition, it serves for the collaboration on and distribution of scientific publications based on these data (see D3.12), access to other types of observational data on the stars observed by *Kepler* (see D3.11), and access to the results of the analysis of inferred stellar properties and models (see D3.13). Moreover, a preliminary interface for accessing these data through the Seismic Plus portal has been implemented and a dedicated server for the so-called Virtual Observatory format will go into operation next year.

The concept of the KASOC database has been to implement a reliable and easily accessible interface for Kepler data by using state-of-the-art database technology and backup strategy. The data redundancy was achieved by a multi-path backup strategy providing complete copies of all observational and processed data at different physical locations, in preparation also for the long-term archive of Kepler data (see D3.15).

The detailed implementation of data access of *Kepler* data is outlined below.

The Kepler Asteroseismic Science Operations Center (KASOC)

The primary purpose of the KASOC database was to distribute observational data from the *Kepler* mission, at various levels of processing to the members of the Kepler Asteroseismic Science Consortium (KASC). This work has also continued in the new K2 mission where the Kepler spacecraft has been refurbished after the loss of a second of its reaction wheels to point along the ecliptic plane.

A complete list of currently provided data products by KASOC is given below in Table 1. Additionally to the original target pixel data and associate light curves, provided by MAST (Mikulski Archive for Space Telescopes), KASOC includes various corrected timeseries and power spectra specifically processed for asteroseismic analyses.

Data Product	Type	Description	Formats
Kepler Target Pixel Files	Image cubes	Original Kepler data produced by Science Office.	FITS
Kepler Lightcurves	Timeseries	Original Kepler data produced by Science Office.	FITS, ASCII
WG1 corrected data	Timeseries	Working Group 1 corrected data (Solar-like stars). No longer produced.	ASCII
WG2 corrected data	Timeseries	Working Group 2 corrected data (Clusters). No longer produced.	ASCII
WG4 corrected data	Timeseries	Working Group 4 corrected data (Delta Scuti stars). No longer produced.	ASCII
WG8 corrected data	Timeseries	Working Group 8 corrected data (Red Giants). No longer produced	ASCII

WG10 corrected data	Timeseries	Working Group 10 corrected data (Gamma Doradus stars). No longer produced.	ASCII
KASOC corrected timeseries	Timeseries	KASOC Corrected and concatenated timeseries.	FITS, ASCII
KASOC unweighted power spectra	Power spectra	KASOC unweighted power spectra of corrected and concatenated timeseries.	FITS, ASCII
KASOC weighted power spectra	Power spectra	KASOC weighted power spectra of corrected and concatenated timeseries.	FITS, ASCII
Red Giants corrected data	Timeseries	Raw concatenated timeseries.	FITS, ASCII
Red Giants corrected data	Timeseries	Corrected concatenated timeseries.	FITS, ASCII
Red Giants corrected data	Power spectra	Power density spectra of Red Giant corrected timeseries.	FITS, ASCII
Red Giants corrected data	Timeseries	Corrected and filtered timeseries.	FITS, ASCII
Red Giants corrected data	Power spectra	Power density spectra of Red Giant corrected and filtered timeseries.	FITS, ASCII
KaDACS corrected and inpainted timeseries	Timeseries	KaDACS corrected and inpainted timeseries.	FITS, ASCII
KaDACS corrected and inpainted power density spectra	Power spectra	KaDACS corrected and inpainted power density spectra.	FITS, ASCII

Table 1: Data products available in KASOC (KaDACS is the Kepler asteroseismic Data Analysis and Calibration Software)

Efficient access to these KASOC data products are offered via an easy-to-use Web interface described below in greater detail.

Web interface

The main philosophy behind the KASOC website has been to create a contemporary and modern website, employing some of the newer web technologies to create a system that is easy to use, and still contain powerful functionality. Many other systems for scientific data distribution that are commonly in use often seem to be more an obstacle than a useful tool for exploring scientific data. One of the design goals therefore has been to put emphasis on providing efficient tools for easy data access.

The main interaction between the KASOC members and the KASOC is through the dedicated KASOC website

<http://kasoc.phys.au.dk> .

A typical user will have access to the following sections on the website:

- Information.
- Kepler Data Search.
- Bundles.
- Star Catalog.
- Ground-Based observations.
- Results.
- Publications.
- User account settings.

In the following the general philosophy behind the website and some of the most important sections mentioned above are discussed.

Another clear philosophy behind the design of the website has been that it should be possible for a user with very limited computer and programming skills, or being limited to a very restrictive system set-up, to use the interface. The typical person that we have had in mind is a person using a standard Microsoft Windows installation with only the standard Notepad editor to open files.

First of all this means that we must provide data in an easily readable format. We have therefore chosen to provide the *Kepler* light-curve data in ASCII format as an alternative to the standard binary FITS format. Special care has been taken for generating ASCII files so that we provide the Windows-style line-format endings¹. This is because Unix-like, Linux and Mac system editors, in general, can handle all kinds of line endings, whereas the default Windows editor (Notepad) can only handle Windows-style endings.

Secondly, it means that it should be simple for the user to perform relatively complex searches on the system, without having to write any form of scripts or complex search queries. The website should therefore provide an easy-to-use graphical search interface which will automatically generate, invisible for the user, the appropriate complex query for the underlying database.

The user interface, and indeed the entire website, is built with well-formatted and validated XHTML 1.1 with CSS (Cascaded Style Sheets) for describing the page layout. Moreover, much of the graphical user interface is generated by using extensive JavaScript functionalities with the aim of creating dynamical webpages. For this we use the libraries jQuery and jQuery UI² and employ their extensive AJAX capabilities. As far as possible we have attempted to employ a policy of *graceful degradation* in regards to the JavaScript functionality, meaning that the website is built in such a way that if the JavaScript is disabled in the user's browser (either by the user himself or because of unforeseen errors) the website will still function and will allow for basic usage, even if the user's web browser could not enable the full feature-set of the interface. On the server side the PHP scripting language is used for generating the XHTML, CSS, and JavaScript code which is sent to the client browser.

Because we are dealing with sensitive and proprietary information and data we have made sure that the information is secure and that access to the website and its information are not accessible without a valid user account at KASOC. This means that extra care was taken in avoiding SQL injections, encrypting information whenever possible and in general ensuring a secure system.

Data Search and bundles

The main cornerstone of the KASOC website is the data search facility. It enables the KASOC members to download *Kepler* data and processed data products listed in Table 1. As explained above, the search interface is aimed to be of easy use, yet providing a powerful search facility.

An example of the standard search interface can be seen in Figure 1. The set-up consists of a number of different fields to provide search constraints, sorted into a number of tabs. Several parameters have easy-to-use sliders and other helpful JavaScript functionalities.

¹ On Microsoft Windows lines in text files are ended with a Carriage return plus a line feed character (CR+LF), whereas on Linux and Mac OS X only LF is used.

² Available at <http://jquery.com> and <http://jqueryui.com>

Once the user clicks the “Search” button, the website will parse the user’s input and automatically constructs complex SQL commands to be sent to the database. The so-obtained results from the database are then parsed and rendered into a well-arranged and highly-configurable XHTML table.

The screenshot displays the 'KIC parameters' search interface. It includes a search bar with 'V* V2213 Cyg' entered, and various filters for RA, Dec, Magnitude, Teff, log G, Radius, [Fe/H], E(B-V), A(V), Contamination, and multiple magnitude bands (u, g, r, i, z, gred, D51, J, H, K). A 'Search' button is located at the bottom left of the search area.

Below the search area, a table shows search results. The table has columns for Filename, KIC-10 number, Cadence, Season, Magnitude, Teff, Radius, SIMBAD, RA, and Dec. Two rows are shown, with the second row selected.

Filename	KIC-10 number	Cadence	Season	Magnitude	Teff	Radius	SIMBAD	RA	Dec
kplr007529123-2009166043257_llc	7529123	Long	Q1	14.762	5673	0.766	V2213 Cyg	19:28:58	43:06:25
kplr007529123-2009259160929_llc	7529123	Long	Q2	14.762	5673	0.766	V2213 Cyg	19:28:58	43:06:25

2 rows returned (1 selected).

Download format:

- ASCII
- FITS
- Target Pixel data
- Above table with KIC parameters
- Plots of timeseries and amp. spectra

[Download Selected](#)

Figure 1: The standard search interface to the database on KASOC.

Below the generated table with results (see Figure 1) the user can select the format with which the data should be retrieved. The following options are available:

- *FITS light curve* – Light curve file in FITS format.
- *ASCII light curve* – Light curve file in ASCII format.
- *Target pixel data* – Target pixel data file.

- *Table with parameters* – An ASCII version of the generated search results as a comma-separated file which is easy to parse in any data analysis program.
- *Plots of time series and spectra* – PDF document with plots of selected time series and amplitude spectra.
- *wget script (FITS)* – Shell script for downloading FITS light curve data using the `wget`-program.
- *wget script (ASCII)* – Shell script for downloading ASCII light curve data using the `wget`-program.
- *wget script (TPD)* – Shell script for downloading target pixel data using the `wget`-program.

Once the user clicks the “download selected” button the list of selected files is submitted to the website for download. The user can also simply click the individual data file names and the browser will download the specific file in accordance with the selected format.

The first three format options will result in the *Kepler* light curves or target pixel data to be downloaded. If multiple files were selected, the website will automatically bundle and compress the data into a ZIP archive before sending it to the client browser. In this way the user instantly gets the data he or she has selected without having to go through further time-consuming steps. This can lead to quite a substantial load on the web server if many large ZIP files need to be generated at the same time. Our remedy for overloading the server is to also provide the files in large pre-bundled ZIP archives or via the `wget`-script options. In the latter case the website will not actually start the data transfer to the user but instead generate a Linux shell script which, when executed, will download the selected files one at a time, eliminating the need for creating the ZIP archive. The script calls the standard `wget` programme, which is available on both Linux and Windows. But for now, the generated script cannot be executed directly on a standard Windows machine without additional (free) software installation³.

Another possibility of downloading data is the pre-bundled data files. If a pre-bundled data file is requested, the website will first control the contents of the available file for correctness before serving it to the user, instead of simply serving the file instantly. The website will generate an up-to-date list of what data files should be in the bundle of interest. This list is then checked against the stored list of files which are currently bundled in the ZIP file. If the two lists are identical the file is sent to the user and the list of files in the bundle is written to the download-log. If, on the other hand, the two lists are not identical the file will not be downloaded, but instead the user will be presented with an error-message explaining the problem. This effectively ensures that the contents of the bundles are always up-to-date and not out of sync with the files available through the data-search interface. Otherwise, this could be a serious concern, because the user expects the bundles to contain the newest data. With this set-up data consistency is automatically guaranteed.

The two remaining options in the list of download formats (*Table with parameters*, *Plots of time series and spectra*) will either generate an ASCII table of the search results (as shown in Figure 1) or generate a PDF file with plots of the time series and amplitude spectra. The plots can also be viewed directly on the web page by clicking the image-icon in front of the file name.

Whenever data are downloaded from the database a precise log is kept of each downloaded file. This has, from the beginning of the *Kepler* mission, been a requirement from the *Kepler* project management. A record of the download is kept in the database and consists, in short, of a list of the downloaded files, the

³ See e.g., <https://www.cygwin.com/>, <http://gnuwin32.sourceforge.net/packages/wget.htm>

format, the user who downloaded them, and a time-stamp of when the download occurred. Great care is taken to ensure that the log is written before the file is sent to the user. This is achieved by using SQL transactions and careful error-handling.

When new derived data files (e.g., corrected time series or power spectra) are ingested into the KASOC database, care is taken to also store information about which original data files were used to generate the new derived data product. This information is also stored in the database and enables complete *traceability* of all files. This, for example, means that a potential problem with an original data file can easily be traced to all derived data products that will be affected.

Access to Kepler/K2 data through Virtual Observatory (VO) and the Seismic Plus Portal

In addition to providing reliable and efficient data access through the KASOC web interface we are in the process of implementing an additional interface for the Seismic Plus Portal (SPP) using the VO format. In a first step we provide plain ASCII tables of the requested metadata needed by the SPP according to the specifications by the SPP administrators (document “*Seismic Plus Portal observation metadata and stellar parameters data*”, Draft 0.1 by C. Renié). The table is generated automatically by the KASOC database and made available to the SPP through an https link. This table contains metadata of all data products and formats as listed in Table 1. SPP periodically fetches this table from KASOC and incorporates the metadata into the SPP database.

Except for the first two entries in Table 1 (original Kepler data products) user authentication on the KASOC web site is still required for downloading the actual data contents (note the metadata describing the actual data is not password protected). Every user can sign up for free on the KASOC web site subject to the KASC policies.

In a next and final step the metadata will be made available through the VO Table Access Protocol (TAP) using a dedicated VO server at Aarhus University.