D3.15: Long term KASOC Archive

Rasmus Handberg[1], Günter Houdek[1], Jørgen Christensen-Dalsgaard[1], Anders S. Conrad[2] and Michael Svendsen[2]

1) Stellar Astrophysics Centre, Aarhus University, Denmark.
2) Royal Library, Copenhagen, Denmark.

1. Introduction

Since the beginning of the Kepler Asteroseismic Investigation, the Kepler Asteroseismic Science Operations Center (KASOC) has been tasked with storing and archiving all Kepler data as well as distributing the data to the members of the Kepler Asteroseismic Science Consortium (KASC) (Kjeldsen et al. 2010). This has been done through the KASOC website[1], which provides the members of KASC easy access to both the original Kepler data, but also to several unique data-products created by the members of KASC. Records of all scientific publications by KASC are also stored as well as high-level data products like derived stellar properties and stellar models.

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, an interface for accessing these data through the Seismic Plus portal has been implemented and a dedicated server for the so-called Virtual Observatory has been put into operation. 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 (D3.14).

As part of the SpaceINN work-package 3 ("Data Handling and Archiving"), KASOC has been tasked with constructing the long-term archive for the *Kepler*/K2 data. In this context, we have started a close collaboration with The Royal Library in Copenhagen, Denmark, where we have created formal Data Management Plans, strategies and requirements for how such an archive should operate.

2. The "living" archive

Currently the KASOC is maintaining an archive of *Kepler*/K2 data as well as derived data, produced by KASC, and a database of derived stellar properties. The goal is that we should create a permanent archive of *all* these data.

In terms of the future of this archive, it is important to realise that even though there are several ongoing and future missions and projects (e.g. SONG, BRITE, TESS, PLATO) devoted to time-domain astronomy, none of them is likely to achieve the same long time-coverage of many stars over many years. Therefore, *Kepler* datasets are in many ways unique and will be useful for active research for many decades to come.

Traditionally, a data archive can be thought of as the digital equivalent of putting all data, notes and publications in neatly labelled boxes and locking them in a storage room for safe keeping.

This is not what we want! Instead, we think that *the future is a living archive*.

[1] http://kasoc.phys.au.dk

This means that we have the following list of requirements to the archive:

- The archive should, at a minimum, be available the next 50 years.
- Data are always freely available on-line.
- Data will continue to be used in active research.
- Extendable: New information should be added to the archive as our knowledge grows.
- Data should be stored in formats that are easily readable by both humans and computers.
- Understandable and useful for future researchers – No matter the science case of the researcher.

3. Discoverability

What will future researchers be interested in and how will they use *Kepler* data? Naturally, we have no idea about the answer to this question. However, when we are thinking on timescales on more than 50 years, we have to take into account that all people having a direct hands-on experience with the data and worked on them when they were taken are no longer around. Therefore, we have to make sure that data are packaged and searchable in a way that makes it obvious to a future researcher what they are looking at.

How will a future researcher, possibly with a very different research topic in mind, best *discover* the *Kepler*/K2 data? To try to get closer to an answer to this, we conducted some thought-experiments of how future researchers in astronomy would search for information. Which parameters are the most important in order to discover the data and their usefulness in future research? Besides general descriptions of *Kepler*/K2 and the different data products (measurements of stellar flux as a function of time), we found that the key parameters to search for is *the astronomical object*, in this case meaning the star, and/or its position on the sky. This is such a fundamental set of properties in astronomy, that this should be our main parameters for data discovery. This has an impact on the way we have chosen to bundle the information (see §4).

4. Data formats

The requirement of having data available on long timescales also requires that we ensure that the structure and formats of the data are readable, useful and extendable.

Files are finally put into a "bag" defined by the BagIt-format, a standard format defined by a collaboration between prominent libraries and universities, including California Digital Library and The Library of Congress (USA) (Kunze et al. 2016). A bag will be self-contained, holding all information about the star, plus all data products for the star (incl. original *Kepler*/K2 data, processed data and power spectra), stellar evolution and structure models, auxiliary files, descriptions and documentation.

A "BagIt" bag is essentially a hierarchical file structure with a number of descriptive text files in the root directory, identifying the format and contents of the bag. The bag also contains a manifest of all files, including cryptographic file-hashes making it possible to verify the integrity of the bag. The "payload" of the bag is provided in the "data" subdirectory. In our case, the payload consists of a generic "readme.txt" file explaining the

```
Kepler_10.zip
|   bag-info.txt
|   bagit.txt
|   fetch.txt
|   manifest-sha1.txt
|
└──data
     |   bundle.xml
     |   readme.txt
     |
     ├──datafiles
     |   └──...
     |
     ├──additional_files
     |   └──...
     |
     ├──documentation
     |   └──...
     |
     └──stellar_models
         └──...
```

Example 1: Structure of a BagIt package.

structure of the payload and an XML file. We have opted for using an XML (Extensible Markup Language) format for storing all results and meta-data for a star. One of the advantages is that format is easily readable to both humans and computers, and very widespread, meaning that parsers are available in most programming languages in use today. A very basic example of such a file is shown in Example 2.

The XML format will also be used on the existing KASOC websites for exchange of results between users and KASOC.

```
<star kic="12345678">
  <numax value="3100" error="20" unit="uHz" />
  <mass value="1.0" error="0.01" unit="solar" />
  <radius value="1.0" error="0.01" unit="solar" />
  <datafiles>
    <datafile uid="1" path="datafiles/original/kplr12345678_llc.fits" />
    <datafile uid="2" path="datafiles/kasoc.ts/kplr12345678_kasoc.ts.fits">
      <dependency datafile="1" />
    </datafile>
  </datafiles>
  <model path="stellar_models/kic12345678/" />
</star>
```

Example 2: Example of XML format used to store results and meta-data. The actual files contain much more information, but this should convey the general ideas of the format. The file format contains records of all available data files, stellar models and other stellar parameters. It also contains dependencies between records.

5. Metadata

Each "bag" of data will be have an associated metadata document, which will follow the DataCite schema for metadata (Starr et al. 2016). This will contain all relevant information about the package, who created the data, who is responsible for its maintenance, references to relevant documentation, traces of different versions of the bag and other general information. All of this is written in an XML format defined by the DataCite schema, which therefore is easily readable by both humans and computer indexers. See §9 for an example of such a file.

Since the DataCite schema does not allow non-standard "fields" to be put directly into the DataCite metadata document, the DataCite document will explicitly reference an additional metadata file that will contain all relevant astronomical metadata. This file will follow the VOTable definition from the Virtual Observatory (Ochsenbein et al. 2013; Demleitner et al. 2010) and contain e.g. resolvable object name and sky positions of the target in question. See §10 for a basic example of such an astronomical metadata file.

6. "Proof-of-concept" archive

We are currently in the process of setting up a "proof-of-concept" version of the long-term archive, which will demonstrate the capabilities and act as a test of the different systems involved.

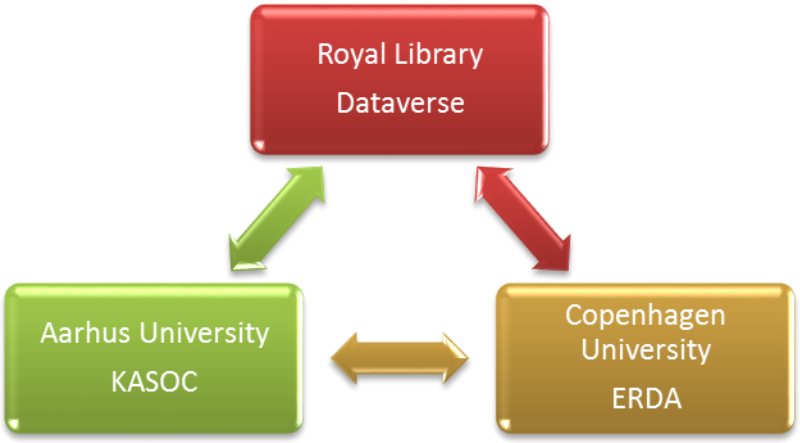
Finding an institution that are willing to and capable of running storing hundreds of terabytes of data, and provide a service layer on top granting access and search-interfaces to these data has proven very difficult in practice. The digital-preservation institutions (e.g. libraries and national archives) are not capable of

handling the amount of data in question[2], and anyone capable of storing the data (e.g. universities) are not able to supply any services on top of the storage or are simply too expensive.

We have therefore chosen to set up the archive as a non-centralized system of three main components: The archive storage system is provided by the Electronic Research Data Archive (ERDA)[3] at the University of Copenhagen. The service-layer is set up by the Royal Library, which will provide the user-interface, and API to communicate with the archive and the actual creation and maintenance of the data packages will be handled by KASOC at Aarhus University.

The reason why we insist on having the physical location of the archive storage being somewhere other than Aarhus University is a simple matter of enhanced security through so-called "geo-replication", meaning ensuring that data are kept at separate geographical locations to minimise risks of data loss due to catastrophic events potentially destroying an entire data centre.

The service layer provided by the Royal Library is build up around the Dataverse[4] system, which is a set of open source research data repository software. The service is currently being run in a virtual machine in a cloud-based computing (Amazon EC2), but could be moved to any of the partners (or new partners) without any issues.



We note that the Royal Library has undergone substantial reorganization in the past year, somewhat delaying this project. The reorganization is continuing in an ongoing merger with the State and University Library, Aarhus. However, there is a continuing commitment to this project, and we expect that the collaboration between KASOC and the reorganized library will continue also in future.

7. The Future

First, it is important to clarify that KASOC will continue running, providing access to *Kepler*/K2 data. Regarding the actual implementation and hosting of a long-term archive, we are as mentioned above

[2] The data amount in KASOC alone is (currently) more than twice the data amount of the Danish national tax authorities (SKAT).
[3] http://www.erda.dk/
[4] http://dataverse.org/

setting up a proof-of-concept archive, with a collaboration between Aarhus University, The Royal Library and University of Copenhagen, utilizing different expertise in long-term file storage, discovery services and astronomical knowledge. Another important issue that is currently under discussion is who should be responsible for the long-term funding of such archives. This question is not yet fully resolved, and may require political decisions on faculty, university or even national level in Denmark.

On the long term, the plan is that the *Kepler*/K2 data and the KASOC archives are copied into this new configuration, and everything we have done in this context could easily be applied to other missions and datasets in the future (e.g. SONG, TESS, PLATO). This will allow researchers of the future who wants to continue to use the *Kepler*/K2 data in active research to discover new things we have not even begun to think of…

8. References

Kjeldsen, H., J. Christensen-Dalsgaard, R. Handberg, T.M. Brown, R.L. Gilliland, W.J. Borucki, and D. Koch. 2010. "The Kepler Asteroseismic Investigation: Scientific Goals and First Results." *Astronomische Nachrichten* 331 (9–10): 966–71. doi:10.1002/asna.201011437.

Kunze, John A., Justin Littman, Liz Madden, Ed Summers, Andy Boyko, and Brian Vargas. 2016. "The BagIt File Packaging Format (V0.97)." https://tools.ietf.org/html/draft-kunze-bagit-14.

Starr, Joan, Jan Ashton, Jan Brase, Paul Bracke, A Gastl, and F Ziedorn. 2016. "DataCite Metadata Schema for the Publication and Citation of Research Data (Version 4.0)." doi:10.5438/0012.

9. Example of metadata file

```
<?xml version="1.0" encoding="UTF-8"?>
<resource xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://datacite.org/schema/kernel-4" xsi:schemaLocation="http://datacite.org/schema/kernel-4
http://schema.datacite.org/meta/kernel-4/metadata.xsd">
   <creators>
      <creator>
         <creatorName>Handberg, Rasmus</creatorName>
         <givenName>Rasmus</givenName>
         <familyName>Handberg</familyName>
         <nameIdentifier schemeURI="http://orcid.org/" nameIdentifierScheme="ORCID">0000-0001-8725-
4502</nameIdentifier>
         <affiliation>Aarhus University, Denmark</affiliation>
      </creator>
   </creators>
   <titles>
      <title xml:lang="en-us">Kepler Asteroseismic Science Operations Center - Data Bag: KIC
12345678</title>
   </titles>
   <publisher>Aarhus Astronomy Data Centre</publisher>
   <publicationYear>2016</publicationYear>
   <subjects>
      <subject xml:lang="en-us" schemeURI="http://udcdata.info/" subjectScheme="UDC"
valueURI="http://udcdata.info/027427">524.3 Stars</subject>
      <subject xml:lang="en-us" schemeURI="http://www.ivoa.net/rdf/Vocabularies/vocabularies-
20091007/IVOAT/">astrophysics</subject>
      <subject xml:lang="en-us" schemeURI="http://www.ivoa.net/rdf/Vocabularies/vocabularies-
20091007/IVOAT/">variableStar</subject>
      <subject xml:lang="en-us" schemeURI="http://www.ivoa.net/rdf/Vocabularies/vocabularies-
20091007/IVOAT/">oscillation</subject>
      <subject xml:lang="en-us" schemeURI="http://www.ivoa.net/rdf/Vocabularies/vocabularies-
20091007/IVOAT/">asteroseismology</subject>
      <subject xml:lang="en-us" schemeURI="http://www.ivoa.net/rdf/Vocabularies/vocabularies-
20091007/IVOAT/">extrasolar planet</subject>
      <subject xml:lang="en-us" schemeURI="http://www.ivoa.net/rdf/Vocabularies/vocabularies-
20091007/IVOAT/">CcdPhotometry</subject>
   </subjects>
   <contributors>
      <contributor contributorType="ContactPerson">
         <contributorName>Handberg, Rasmus</contributorName>
         <familyName>Handberg</familyName>
         <givenName>Rasmus</givenName>
         <nameIdentifier schemeURI="http://orcid.org/" nameIdentifierScheme="ORCID">0000-0001-8725-
4502</nameIdentifier>
         <affiliation>Stellar Astrophysics Centre, Aarhus University, Ny Munkegade 120, Bldn. 1520,
8000 Aarhus C, Denmark</affiliation>
      </contributor>

      <contributor contributorType="DataCollector">
         <contributor>NASA Kepler Mission</contributor>
         <affiliation>NASA, USA</affiliation>
      </contributor>

      <contributor contributorType="DataCollector">
         <contributor>NASA K2 Mission</contributor>
         <affiliation>NASA, USA</affiliation>
      </contributor>

      <contributor contributorType="ResearchGroup">
         <contributor>Kepler Asteroseismic Science Consortium (KASC)</contributor>
      </contributor>
   </contributors>
   <language>en-us</language>
   <resourceType resourceTypeGeneral="Dataset">Dataset/BagIt</resourceType>
   <formats>
      <format>application/x-gzip</format>
   </formats>
   <rightsList>
      <rights rightsURI="https://creativecommons.org/licenses/by/4.0/">BY 4.0 International</rights>
   </rightsList>
   <descriptions>
```

```
      <description xml:lang="en-us" descriptionType="Abstract">
         Collection of all scientific data available in the Kepler Asteroseismic Science Operations
Center (KASOC)
         on the star KIC 12345678. Includes original data from the NASA Kepler spacecraft, derived
data products and
         high-level scientific results.
      </description>
      <description xml:lang="en-us" descriptionType="TechnicalInfo">
         ########################################################################
           Methodology
         ########################################################################

         The NASA Kepler spacecraft, launched in March 2009, measured the
         brightness of thousands of stars to search for extrasolar planets and
         study the stars through the techniques known as "asteroseismology".

         ########################################################################
           Directory overview
         ########################################################################

         "bundles.xml"
            Main file with records of all available data products, records of
            high-level scientific analysis results, stellar models and other
            files.

            Full documentation on the format in this file is found in
            "documentation/xml-format.pdf".

         "datafiles"
            Contains original and corrected Kepler/K2 data products.
            This constitutes time-series observations of the given star
            (pixel images), light curves extracted from these and calculated
            frequency power spectra.
            All files are provided in either FITS or ASCII format,
            and in many cases both.

         "documentation"
            Contains documentation on all the included data products.

         "stellar_models"
            Stellar structure and evolutionary models available for the given
            star. Stellar structure is stored in the FGONG format, and
            evolutionary tracks in CSUM format (see "documentation/models").

         "additional_files"
            Auxiliary files uploaded by the members of the Kepler Asteroseismic
            Science Consortium (KASC) for this given star.
      </description>
   </descriptions>

   <!-- DOI OF THIS BAG -->
   <identifier identifierType="DOI">10.10.1038/nphys1170</identifier>

   <relatedIdentifiers>
      <!-- Astronomical metadata: -->
      <relatedIdentifier relatedIdentifierType="URL"
relationType="HasMetadata">http://example.com/example_bag_metadata_votable.xml</relatedIdentifier>
      <!-- General documentation: -->
      <relatedIdentifier relatedIdentifierType="DOI"
relationType="IsDocumentedBy">10.10.1038/nphys1170<!-- DOI OF DOCUMENTATION PACKAGE --
></relatedIdentifier>
      <relatedIdentifier relatedIdentifierType="URL"
relationType="IsDerivedFrom">http://kasoc.phys.au.dk/</relatedIdentifier>
      <relatedIdentifier relatedIdentifierType="URL"
relationType="IsDerivedFrom">https://archive.stsci.edu/</relatedIdentifier>
      <!-- Documentation bibcodes: In practive we will grab ALL bibcodes in bundle.xml -->
      <relatedIdentifier relatedIdentifierType="bibcode"
relationType="IsDocumentedBy">2014MNRAS.445.2698H</relatedIdentifier>
      <relatedIdentifier relatedIdentifierType="bibcode"
relationType="IsDocumentedBy">2011MNRAS.414L...6G</relatedIdentifier>
   </relatedIdentifiers>

   <dates>
      <date dateType="Created">2016-11-21</date>
```

```
      <date dateType="Updated">2016-11-22</date>
   </dates>

   <version>2.0</version>

   <sizes>
      <size>3 KiB</size>
   </sizes>
</resource>
```

10. Example of astronomical metadata file

```
<?xml version="1.0" encoding="UTF-8"?>
<VOTABLE
   xmlns="http://www.ivoa.net/xml/VOTable/v1.3"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xsi:schemaLocation="http://www.ivoa.net/xml/VOTable/v1.3 http://www.ivoa.net/xml/VOTable/v1.3"
   version="1.3">
   <DESCRIPTION>
   Astronomical metadata for Kepler-10 bag.
   </DESCRIPTION>
   <RESOURCE>
      <TABLE nrows="1">
         <GROUP utype="stc:CatalogEntryLocation">
            <PARAM name="href" datatype="char" arraysize="*"
               utype="stc:AstroCoordSystem.href"
               value="ivo://STClib/CoordSys#TDB-ICRS-BARY"/>
            <PARAM name="URI" datatype="char" arraysize="*"
               utype="stc:DataModel.URI"
               value="http://www.ivoa.net/xml/STC/stc-v1.30.xsd"/>
            <FIELDref ref="ra" utype="stc:AstroCoords.Position2D.Value2.C1"/>
            <FIELDref ref="de" utype="stc:AstroCoords.Position2D.Value2.C2"/>
         </GROUP>
         <FIELD datatype="char" arraysize="*" name="target_name" ID="target_name"
ucd="meta.id;meta.main" />
         <FIELD datatype="double" name="ra" ID="ra" ucd="pos.eq.ra;meta.main" unit="deg" />
         <FIELD datatype="double" name="dec" ID="dec" ucd="pos.eq.dec;meta.main" unit="deg" />
         <DATA>
            <TABLEDATA>
               <TR>
                  <TD>KIC 12345678</TD>
                  <TD>7.2854</TD>
                  <TD>69.1300</TD>
               </TR>
            </TABLEDATA>
         </DATA>
      </TABLE>
   </RESOURCE>
</VOTABLE>
```