Backup to the Future

Creating the *Kepler*/K2 long-term "living archive"

Rasmus Handberg[1,⋆], *Anders S.* Conrad[2], and *Michael* Svendsen[2]

[1] *Stellar Astrophysics Centre, Department of Physics and Astronomy, Aarhus University, Ny Munkegade 120, DK-8000 Aarhus C, Denmark.*
[2] *The Royal Library, Copenhagen, Denmark.*

Abstract. What if something horrible happens in the future? Great Scott! Do we have a backup? Is a backup even enough? At the Kepler Asteroseismic Science Operations Center (KASOC) we have a goal that all data and information from *Kepler* and KASC is preserved for the future. The benchmark is that the data should be useful for, at least, the next 50 years. But how do we ensure that hundreds of terabytes of data are understandable or even readable in half a century?

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) [1]. 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.

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 may ways unique and will be useful for active research for many decades to come.

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.

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*. 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. But 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 it when it was taken are no longer around. So we have to make sure that data is packaged and searchable in a way that makes it obvious to a future researcher what they are looking at.

But 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

⋆e-mail: rasmush@phys.au.dk
[1] http://kasoc.phys.au.dk

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.

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 both to humans and computers, and very wide-spread, meaning that parsers are available in most programming languages in use today. A very basic example of such a file is shown below.

```
<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" />
</star>
```

Example 1: Example of XML format used to store results and meta-data.

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

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) [2]. A bag will be self-contained, holding the XML data, 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. The bag also contains a manifest of all files, including cryptographic file-hashes making it possible to verify the integrity of the bag.

5 The Future

First of all, 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 currently in dialogue with national facilities in Denmark that can ensure its operation on timescales of many decades. Currently the most likely configuration looks like being a collaboration between Aarhus University, The Royal Library and Copenhagen University, utilizing different expertises in long-term file storage, discovery services and astronomical know-how. 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 (decades) the plan is that the *Kepler*/K2 data and the KASOC archives are copied into this new configuration, allowing researchers of the future who wants to continue to use the Kepler/K2 data in active research to discover new things we haven't even begun to think of...

Acknowledgements

Funding for the Stellar Astrophysics Centre is provided by The Danish National Research Foundation (Grant DNRF106). The research was supported by the ASTERISK project (ASTERoseismic Investigations with SONG and Kepler) funded by the European Research Council (Grant agreement no.: 267864).

References

[1] H. Kjeldsen, J. Christensen-Dalsgaard, R. Handberg, T.M. Brown, R.L. Gilliland, W.J. Borucki, D. Koch, Astronomische Nachrichten **331**, 966 (2010), `1007.1816`

[2] J. Kunze, J. Littman, L. Madden, E. Summers, A. Boyko, B. Vargas, Internet-Draft draft-kunze-bagit-13, IETF Secretariat (2016)