IJDC | *General Article*

Reuse for Research: Curating Astrophysical Datasets for Future Researchers

Anders Conrad
Copenhagen University Library
Royal Danish Library

Rasmus Handberg
Department of Physics and Astronomy
Aarhus University

Michael Svendsen
Copenhagen University Library
Royal Danish Library

Abstract

"Our data are going to be valuable for science for the next 50 years, so please make sure you preserve them and keep them accessible for active research for at least that period."

These were approximately the words used by the principal investigator of the Kepler Asteroseismic Science Consortium (KASC) when he presented our task to us. The data in question consists of data products produced by KASC researchers and working groups as part of their research, as well as underlying data imported from the NASA archives.

The overall requirements for 50 years of preservation while, at the same time, enabling reuse of the data for active research presented a number of specific challenges, closely intertwining data handling and data infrastructure with scientific issues. This paper reports our work to deliver the best possible solution, performed in close cooperation between the research team and library personnel.

Received 23 February 2017 ~ *Accepted* 8 December 2017

Correspondence should be addressed to Anders Conrad, Royal Danish Library, P.O. Box 2149, 1016 Copenhagen K, Denmark. Email: asc@kb.dk

An earlier version of this paper was presented at the 12th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

 doi:10.2218/ijdc.v12i2.516

Introduction

The Kepler Asteroseismic Investigation is part of the scientific organisation of the NASA Kepler mission[1] to observe stars for the possible discovery of extrasolar planets (Kjeldsen et al., 2010). The Kepler satellite, which was launched in 2009, is sampling thousands of stars and transmitting image data of their light emissions to the Earth. Apart from playing an important role in the mission's scientific process, the Kepler Asteroseismic Science Consortium (KASC), based at Aarhus University, Denmark, offers a setup for broader asteroseismic studies[2], based on Kepler data.

The KASC, consisting of hundreds of members worldwide and several working groups, has organised the collection and publication of data in the Kepler Asteroseismic Science Operations Center, KASOC[3]. KASOC consists of a website where data and documentation from NASA can be retrieved alongside with the derived data products, documentation, journal articles and related materials produced by the KASC members. KASOC is currently running at and being funded through Aarhus University.

At the time of writing, the Kepler satellite is still operating and transmitting data, although on different operating conditions after a technical malfunction. The extended mission has been named K2, and the KASC and KASOC remain active in the processing and scientific exploration of new data.

All this is well and good, but what will happen the day KASC and KASOC no longer exist, and nobody in the research community has a first-hand memory of the Kepler/K2 mission? Will somebody take care of the data? Will future researchers be able to interpret and use the data outside today's operational context? These were questions behind the case, driven by the researcher's belief that this body of data will remain of immense value for researchers for decades to come. Out of this came the requirement for a long-term preservation, extracting data from KASOC and depositing them onto some sort of preservation archive, in a way that allows for continued research.

We have tried to meet this challenge, as a pilot case in a Danish data management project[4] made us consider all steps of a data life cycle, possibly providing new services, if required[5]. As a lot of work regarding data collection and analysis was already taken care of by the research team, we focused our efforts on how to establish a long-term archive that would allow for continued use for active research.

The project being reported will remain active into the spring of 2017, so certain new developments may be reported at the IDCC17 conference which are not covered here. The specifics of the preservation case for KASOC have been reported as deliverable to the EU SpaceINN project (Handberg, Houdek, Christensen-Dalsgaard, Conrad, and Svendsen, 2016). For this reason, we will, in this paper, focus mainly on the curation considerations and only provide a brief resume of the technical solutions.

1 Kepler and K2. Mission overview: https://www.nasa.gov/mission_pages/kepler/overview/index.html
2 Asteroseismology is the study of stellar oscillations, which plays an important part in the Kepler/K2 search for planets, but is also a research field on its own.
3 KASOC Data Release: http://kasoc.phys.au.dk/index.php
4 The Data Management in Practice project, with participation from Danish research libraries, as well as The Danish National Archives.
5 This is truly a practice paper, and we apologise for possibly having overlooked valuable work by others in the field, as we have been working with limited time allocations.

The Challenge

In the report accompanying the SpaceINN delivery mentioned above, the goal of a living archive was specified by the following requirements:

- The archive should, at a minimum, be available for 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 what the science case of the researcher may be.

We decided to take a closer look at the exact meaning of these requirements and how they work together in defining our task of curating the data. Apart from the requirements listed here, the Data Management in Practice project that we are part of, required all datasets to be supplied with a DOI and to be searchable in a national research database.

Available for a Minimum of 50 Years: Long-term Stewardship

As a matter of fact, we have not been able to find an institution or a method that allows us to claim the 50 years of preservation to be definitively solved. 50 years is a lot longer than the usual five or ten years often required by data policies. The most likely places we might have achieved this would have been the Royal Danish Library or the Danish National Archives, which both have a very strong institutional sustainability and a commitment to preserve materials for a very long time. Unfortunately, neither of these institutions are able to accommodate the needs at the present time.

There were serious talks with the National Archives, who showed a sincere interest in taking on the task. The obstacles were twofold, one being the size of data, anticipated to reach about 110 terabytes; the other being the lack of a method to guarantee preservation of this kind of data in the strict sense of the word, that is, preserving the full intellectual content and readability of data by means of format migrations, etc as media and formats change.

This led us away from thinking about our task as a preservation challenge and rather, to consider it as a curation challenge: how to curate the data for availability and reuse in the short term, while trying to organise storage, data, and documentation in a way that best allows for long-term stewardship, even if we do not know who may eventually take over the responsibility for the data or in which institutional and technical context that might happen.

 doi:10.2218/ijdc.v12i2.516

Freely Available Online: Licensing

This requirement is closely linked to the NASA's open data policy that places published Kepler/K2 data into the public domain[6]. KASC has adopted an open policy in the latest version of its data policy for KASOC (Kjeldsen, Christensen-Dalsgaard, Kawaler, and Gilliland, 2012). This openness is a requirement for publication on KASOC and should be reflected in very open licensing, CC0 or similar, when re-publishing the data outside of KASOC.

Useful for Future Researchers: Dataset Content

This heading sets the direction for a large part of our work: how might a future researcher discover, retrieve and interpret our data, and – related to that – for what kind of research purpose? And what implications would that necessarily impose on the way we structure and document our datasets?

Although, in principle, the data could be interesting for purposes and research we cannot even imagine today, we decided to restrict the scope of our project to include research that is related to the current purpose of the Kepler/KASOC data, i.e. astronomy. We conducted future-workshop sessions, where KASC researchers envisioned non-Kepler research questions for which the KASOC data could be relevant. According to the data management plan[7], the data is comprised of five types, representing an increasing level of processing:

- Kepler original pixel data (retrieved from the NASA MAST[8] archive).
- Light curves (time series) originating from NASA or from KASC.
- Power-frequency spectra.
- Analysis data, such as measured oscillation frequencies and stellar properties.
- Physical models of stars.

We envisioned that if a future researcher is interested in continuing KASC-like research, the high-level data products of KASC would be considered trustworthy and valuable. But also, that it should be possible to repurpose the data for other astronomical research by re-calibrating lower-level products for the study of different signals, or for re-processing or further refining the high-level models. This pointed towards the fact that the original NASA data would be important as well, also emphasised by the fact that a lot of other data products build directly upon pixel and light curve data from NASA.

The question was then, whether the Kepler original data from NASA should be included in our datasets, or if they should be referred as external resources located in the NASA archives. Communication with NASA, revealed that the contract with MAST would run for ten years after the end of the mission and that only the final processing of each file would be archived. NASA recommended archiving data in different places for the long term (Barclay, personal communication).

In the KASOC publication model, it is transparent and well-documented how each data product builds upon specific versions of other data products. So just preserving the latest version would incur a risk of breaking scientific integrity. Combined with our 50

6 Data Use Policy: https://archive.stsci.edu/data_use.html
7 KASOC Data Management Plan: http://kasoc.phys.au.dk/docs/data_management_plan_kasoc.pdf
8 Mikulski Archive for Space Telescopes (MAST): https://archive.stsci.edu/

year obligation, much longer than the ten years required by NASA, we would need to include all versions of all the NASA data files in KASOC in our archive as well. At the same time, we could serve as an alternative archive, as recommended by NASA.

In KASOC, data files are organised according to the Kepler and K2 missions' quarters and campaigns, as well as by various astrophysical classifications. We decided that for the long-term archival, the primary structure of the datasets would be according to the observed stars: one dataset for each star, comprising data files of all the five levels listed above. That is going to make data most useful to researchers.

Understandable for Future Researchers: Documentation

We started by imagining that a future researcher would somehow discover and retrieve one of the datasets we were planning to archive. Would they be able to understand what the data were? What celestial object was being observed? What processing had been done to produce the data and by whom? Could a future researcher, potentially with no or very little knowledge about the specifics of the Kepler/K2 mission, assess the credibility and usability of the data for their research purpose? Or even – related to another requirement – could a computer program?

To make data understandable, documentation is necessary. The main source of information is the data release notes published by NASA and KASC. As these documents are generally released as part of quarterly data releases, each of them is necessarily going to cover files across our datasets, which we decided are going to be organised by star, not by data release. This, in turn, means that information related to one of our datasets could be distributed across several different release notes. In order to avoid including many or all release notes with every dataset, we decided to provide a special documentation package as part of the archive, which can then be referred to by all the regular datasets.

Some basic documentation about the observations can be found in the header of each data file. Apart from that, for some of the data products, and particularly the higher-level models, the documentation of the data has been published as part of scientific articles. We need to provide this information, in the form of bibcode[9] references, as the data might be useless for research without knowledge of which processing has been applied.

In order to provide these various bits of information without forcing our future researcher to parse all the individual data files of the dataset, we thought it necessary to provide an inventory file that would gather all the basic metadata for each file in a standardised format. This file provides links to necessary documentation, both within and outside the dataset.

The combined dataset and documentation package was tested on researchers in astrophysics to verify that it was indeed understandable as a standalone data object.

Continued Use in Active Research: Discoverability

Although the requirement that data should be useful and understandable plays a crucial role for continued use, it also depends largely on the ease with which data can be identified and discovered by future scientists. This, in turn, relates to the metadata that we apply to our datasets and the discovery features that are being offered by the data repository service.

9 Bibliographic identifier related to the astronomical SIMBAD database. See documentation on http://adsabs.harvard.edu/abs_doc/help_pages/data.html

 doi:10.2218/ijdc.v12i2.516

Each dataset, containing all the available data files for one star, will be considered as one data object in the repository. So there will be one identifier – one DOI – that is covering the entire dataset, and each dataset will be provided with citation metadata in the Datacite schema[10]. This will make it slightly harder to cite a single file within a dataset. We have not been able to circumvent this limitation because it would have been a major work to treat each single file as an independent data object, while still supplying the necessary context and documentation.

It is anticipated that future researchers, who could not be expected to know about the Kepler mission, will primarily discover archived Kepler data either by the name or the position[11] of the observed star. This calls for the use of discipline-specific metadata as well as the citation metadata mentioned above. In addition, it raises the requirement for a repository system that allows for discovery based on astronomy-specific metadata. In our work with the case, this single requirement has proved to be one of the most difficult ones to solve.

Extendable: Adding New Versions and New Files

The KASOC database is still active, and new data files are being added, as well as new versions of existing files. We would, however, want to start depositing datasets into a repository or storage archive as soon as possible. The solution will be that new versions and new files will be added to datasets, without changing or removing the old ones. This will allow consistency of references to datasets and will retain all earlier versions of files that may be referred from published work as well as by other files building upon them.

Just like the individual data files are equipped with a version number added to their file name, datasets will be versioned by incremental integer numbers. We have envisioned that datasets where the content has been expanded will be re-deposited to the archive repository at regular intervals, e.g. three months. Because each incremental version of the dataset will keep all the existing files unchanged, it should be possible to overwrite old versions in the repository and just keep the newest one, under the same DOI as previous versions.

Although this strategy incurs the overhead of re-ingesting files which are already present at the repository, we anticipate that the number of datasets that will become expanded and need re-uploading will be limited. But there is another problem regarding continued active research: the archiving strategy presupposes that the KASOC database is still there to generate the datasets. This is not going to be the case forever!

We do not currently have a solution for that. Repository software may provide the option to unpack datasets and allow for versioning of individual files. This would, however, question the coherence between documentation and datasets that we plan to provide. The problem of continued active research may be one of several probable issues to be taken care of by future stewards of the data.

Readable by Both Humans and Computers: Machine Readability

Whereas current scientific practice still has a large focus on human processing and evaluation, initiatives such as FAIR principles for data management require that data, and particularly metadata, must be machine readable (Wilkinson et al., 2016). With a 50 years' horizon in mind, this requirement would seem to be indispensable.

10 DataCite Schema: http://schema.datacite.org/

11 A system of celestial coordinates has been standardised in (Rots, 2007)

We decided to accommodate this by, as best we could, supplying both datasets, as well as metadata, in standardised file formats and structures that can be easily parsed by a computer. Apart from that, we have made some attempts to supply the inventory file of each dataset with references to astronomical vocabularies, for an initial level of semantic machine processing. This could clearly be developed further in future projects.

The Proposed Solution – Proof of Concept

The practical solution that we are working on is based on the fact that all the data that we want to curate as structured datasets, are already present in the KASOC database and can be extracted from there and ingested to the chosen repository/storage in an automated fashion. Aside from trying to find infrastructure and finances for the long-term archive, we have been working on a proof-of-concept for the practical solution. Our task has been to identify and generate the specific formats needed to secure access to all the necessary services, and to develop a workflow for ingestion that is both technically feasible and possible to implement with the chosen repository/storage solution. In practice, this has been done as a combination of manual testing and scripting the data extraction from KASOC to produce the desired file and dataset formats.

Dataset Structure and File Formats

As mentioned above, for each observed star there will be just one dataset, containing all the files available in KASOC relating to that star. Internally, the datasets will be structured according to the quarter/campaign numbers of the Kepler/K2 mission, as well as types of observation[12]. This makes it easier to navigate the dataset according to the information in the quarterly release notes. The datasets will be packaged in the BagIt archive format which is simple and easy to parse for both computers and humans, and described as being suitable for digital preservation purposes (Kunze et al., 2016). BagIt archives can be stored and transmitted as compressed ZIP archives.

The main format of the data files themselves is FITS[13], a well-established and standardised format in astronomy, and therefore well-suited for long-term reuse[14]. There is an issue with respect to preservation, that this format is quite old and therefore not space-efficient. We hope to overcome this problem when the FITS files get stored as part of the compressed BagIt archives. By storing files together in BagIt archives, we intend to accommodate, in the best possible way, a situation where the dataset packages will most likely outlive current repository and discovery options.

Another benefit of the BagIt format is the possibility of adding an optional file, fetch.txt, to the archive with references to other files that need to be fetched and added for the full understanding of the archived data. We use this feature to refer to content of the documentation package, which is shared by all the individual datasets. Currently the links point to the existing KASOC site, but it is planned to generate links to the archived documentation, once the repository solution is in place.

12 This will all be documented in the documentation package.
13 The FITS Support Office: https://fits.gsfc.nasa.gov/
14 See Library of Congress assessment at Flexible Image Transport System (FITS), Version 3.0: http://www.digitalpreservation.gov/formats/fdd/fdd000317.shtml

 doi:10.2218/ijdc.v12i2.516

We have produced a number of datasets for testing, including the inventory file for each dataset, and these can now all be generated automatically. The datasets will vary in size, with the maximum size of a dataset being around ten gigabytes.

Storage and Repository Solution

As it became clear that neither the Royal Danish Library, nor the Danish National Archives could store our data, we started looking towards other stable institutions with sufficient storage and archiving facilities. We came upon the University of Copenhagen's ERDA file archive[15], based at the Niels Bohr Institute and consisting of several petabytes of storage. The staff responsible for ERDA has been very obliging about hosting the data with a reasonable price model, but they do not offer any discovery or metadata service beyond the raw file archive.

It is currently being discussed at the Royal Danish Library to set up a front-end repository service, with user interface and API's for ingest, discovery, download, etc. A proof-of-concept installation has been made by installing the Dataverse data repository software[16] on an Amazon EC2 server, mapping the ERDA file repository as the backend file store by means of the WebDAV protocol. There are some technical issues with this solution, with regards to a production setup, but it shows a way of moving forward, by distributing various parts of a data repository service to different partners.

At the moment, we are experimenting with Dataverse's API's for the creation of datasets and addition of data files, with the purpose of testing a workflow of automated ingest from KASOC into Dataverse and ERDA.

Discovery and Metadata

It is a rather indispensable requirement from the researchers that data must be discoverable by celestial object name and from its position on the sky, thus enabling future researchers to discover datasets covering a certain area of the sky. That requirement is not easily met by standard repositories or repository software! KASOC is already part of the astronomical Virtual Observatory (IVOA)[17] which does not, however, exactly accommodate the kind of archiving situation we envision.

We chose to work with Dataverse, partially because it offers some support for astronomical data (Pepe, Goodman, Muench, Crosas, and Erdmann, 2014), based on the IVOA resource metadata specification (Hanisch, The IVOA Resource Registry Working Group, and The NVO Metadata Working Group, 2007). It is possible to add some astronomy-specific metadata which are then indexed and accessible for discovery. Unfortunately, we have run into what seems to be a challenge with respect to the inability to allow searching based on a user-specified area of the sky. We are currently discussing possible solutions with the Dataverse community.

Citation metadata are rather trivial to extract from KASOC into any desired format. Although Dataverse takes care of creating a Datacite metadata record[18] during the creation of a new dataset, we specifically add a Datacite record to the BagIt package, in order to secure the future understanding of it, independent of any specific repository software.

15 Electronic Research Data Archive: http://www.erda.dk
16 The Dataverse Project: http://dataverse.org/
17 International Virtual Observatory Alliance: http://ivoa.net/
18 DataCite Metadata Schema: http://schema.datacite.org/

Conclusions

We have come quite far in obtaining our goals of curating the Kepler/KASOC data for long-term active use and reuse for research, but we are not totally there yet. We hope, in the last months of the project, to be able to resolve the remaining technical and financial obstacles to enable us to establish the proposed solution in a production setup.

We have experienced that a detailed understanding of how data might be used scientifically in the future, has been guiding decisions on how to curate it with long-term stewardship in mind. We have also come to acknowledge that an archival task of this size and complexity is challenging to existing preservation institutions and may require new partnerships across traditional institutional roles in order to be solved.

It has been essential for our work that it integrates scientific skills and deep understanding of the data with strong IT technical skills, as well as library competencies such as metadata and information handling. We consider this combination of competencies as very fruitful and productive and as an experience that could contribute to data science in the future.

It may seem excessive to go into this level of detail for just one case of data curation! We have nonetheless been lucky to have an opportunity to explore an ambitious and challenging case in-depth in a pilot-project situation. We hope to be able to transfer parts of the experience into future work, where curation and data stewardship are likely to become integrated parts of research projects, without the need or possibility to gather a team of experts each time.

Acknowledgements

The work presented in this paper was funded by Denmark's Electronic Research Library (DEFF), Aarhus University and Royal Danish Library, and was performed as part of the project Data Management in Practice.

References

Handberg, R., Houdek, G., Christensen-Dalsgaard, J., Conrad, A.S., & Svendsen, M. (2016). Long term KASOC archive. Retrieved from http://www.spaceinn.eu/wp-content/uploads/2017/01/D3.15_SpaceINN_Deliverable_KASCO_Archive.pdf

Hanisch, R., The IVOA Resource Registry Working Group, & The NVO Metadata Working Group. (2007). IVOA recommendation: Resource metadata for the virtual observatory. Retrieved from http://www.ivoa.net/documents/latest/RM.html

Kjeldsen, H., Christensen-Dalsgaard, J., Handberg, R., Brown, T.M., Gilliland, R.L., Borucki, W.J., & Koch, D. (2010). The Kepler Asteroseismic Investigation: Scientific goals and first results. *Astronomische Nachrichten, 331*(9–10), 966–971. doi:10.1002/asna.201011437

Kjeldsen, H., Christensen-Dalsgaard, J., Kawaler, S., & Gilliland, R. (2012). KASC strategies and policies in the extended Kepler mission. Retrieved from http://kasoc.phys.au.dk/docs/DASC_KASOC_0041_5.pdf

Kunze, J., Littman, J., Madden, L., Summers, E., Boyko, A., & Vargas, B. (2016). The BagIt file packaging format (V0.97). Retrieved from https://tools.ietf.org/html/draft-kunze-bagit-14

Pepe, A., Goodman, A., Muench, A., Crosas, M., & Erdmann, C. (2014). How do astronomers share data? Reliability and persistence of datasets linked in AAS publications and a qualitative study of data practices among US astronomers. *PLoS ONE, 9*(8). doi:10.1371/journal.pone.0104798

Rots, A.H. (2007). STC metadata for the VO. Retrieved from http://hea-www.cfa.harvard.edu/~arots/nvometa/STC/STC-20071030.html

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*, 9. doi:10.1038/sdata.2016.18