

Managing Contributions to the Unified Astronomy Thesaurus

Katie Frey^{1,*}, Sarah Weissman², Barbara Kern³, Jill Lagerstrom², Josh Peek², and Alberto Accomazzi⁴

¹Wolbach Library, Harvard-Smithsonian Center for Astrophysics, USA

²Space Telescope Science Institute, USA

³Science Libraries, University of Chicago, USA

⁴SAO/NASA Astrophysics Data System, USA

Abstract. The Unified Astronomy Thesaurus (UAT) project managers have long defined the UAT as "an open, interoperable, and community-supported thesaurus." How do we solicit the detailed, comprehensive, and consistent community feedback that is required to keep the UAT relevant? The Steering Committee for the UAT has developed a visual organizational tool that lets reviewers suggest new concepts and restructure the existing hierarchy. Researchers and librarians can use this "Sorting Tool" to submit contributions and feedback to the UAT. The UAT Curator adds feedback to the UAT's GitHub Issues, which allows for tracking, searching, and referencing the suggestions. The UAT Curator implements the accepted suggestions, preparing to include them in a future release of the Unified Astronomy Thesaurus. This process of continual improvement ensures that the UAT project remains community supported.

1 A Formal Vocabulary for Astronomy

One of the guiding principles for the Unified Astronomy Thesaurus (UAT) is the understanding that labels are important. If something does not have a label, you cannot talk about it. To make matters even more complicated, different perspectives on a topic can lead to very different labels and therefore a different understanding. It is equally important to realize that the organization and relationships between labels also drive how a subject is understood and discussed. The field of Astronomy has a highly publicized example of the impact that a change in understanding can have on a field of research. As scientists' understanding of the word "planet" became more refined and specific, the International Astronomical Union (IAU) realized it had to reclassify Pluto as a "dwarf planet." This new label for Pluto has fundamentally altered the way people everywhere, not just scientists, understand the Solar System. This only serves to emphasize that in order to discuss a topic cogently, there has to be a mutual basis of understanding, a common ground from which to grow.

A thesaurus provides a consistent set of keywords that classify, describe, and map the understanding of a particular topic or subject. It builds a structure that can be mapped and navigated, leading from one concept to another in a logical way. It organizes the concepts into similar groups, adding additional context to help clarify the meaning of any particular concept.

*e-mail: kfrey@cfa.harvard.edu ORCID: 0000-0001-9891-4465

Many service providers and stakeholders in astronomy have expressed interest in embedding the UAT into a variety of applications, which means that the vocabulary will act as a bridge between these systems. We anticipate use of the UAT increasing and believe that the UAT will underpin and inspire a new range of cross-system data-sharing applications by creating a formalized source of vocabulary words. Using the same set of concepts to describe datasets, instruments, and astronomical objects, it enables an increased level of connectivity between data and content, an ability to enhance literature searching and indexing, and a huge improvement to discoverability.

2 Beyond the Astronomical Subject Keywords

The Astronomical Subject Keywords initiative ([1]) is one of the most recent efforts to build a unified list of keywords across publishers and publications in astronomy. In fact, this list has been approved by editors from many of the premier astronomy journals, who all saw the inherent value in sharing keywords in order to enhance findability. However, this keyword list is currently stagnant, having not received an update since 2013. Similarly, other previous efforts to create comprehensive keyword lists have stalled: the IAU Thesaurus ([2]) was last updated in 1995, the IVOA Thesaurus ([3]) was last updated in 2009, and the Physics and Astronomy Classification Scheme ([4]) was officially retired in 2010.

The UAT began as a direct result of efforts by the Institute of Physics and the American Institute of Physics. Independently, these two publishers recognized that current keyword lists, in addition to being woefully out of date, were also not in a format suitable for new linked data applications. By coincidence, each publisher began working with Access Innovations, a content and knowledge management business, to update, modernize, and improve their keyword lists. When each learned of the other's efforts, the two publishers decided to work together, merging their projects and donating the results to the American Astronomical Society with the intention that the thesaurus be an open and community supported resource. This thesaurus, referred to as the beta version of the UAT, included 1910 concepts, 15 top-level concepts, and a maximum depth of 12 levels.

A more detailed accounting of the provenance and history of efforts that lead up to the Unified Astronomy Thesaurus is available in a paper by Alberto Accomazzi, et al from 2014 ([5]).

3 Major Updates to the Unified Astronomy Thesaurus

Although the beta version of the UAT was a huge improvement over the older and no longer supported vocabularies, it had many flaws that needed to be addressed. The original concepts were serviceable, but not organized in the way that made logical sense to researchers in the field of astronomy and astrophysics. Top-level categories in the original beta version included "Astronomical objects," "Lunar physics," and "Positional astronomy," none of which are major fields of study today.

One of the first major undertakings was to align the top-level categories of the UAT with the Divisions of the International Astronomical Union, which are widely recognized as sub-disciplines within astronomy and astrophysics. This process involved a small committee of librarians and researchers who evaluated the structure of the UAT on a broad and large-scale basis, merging related top-level categories, promoting subsections to top-level categories, and enacted other overarching changes. After the bones of the reorganization were in place, and the Steering Committee felt comfortable with the overall structure, we quickly realized that the UAT needed detailed analysis on a concept-by-concept basis.

Since the UAT still had over 1900 concepts at this time, we worked with researchers to divide the UAT into logical and discrete subsections, which could be easily examined in detail. We ended up

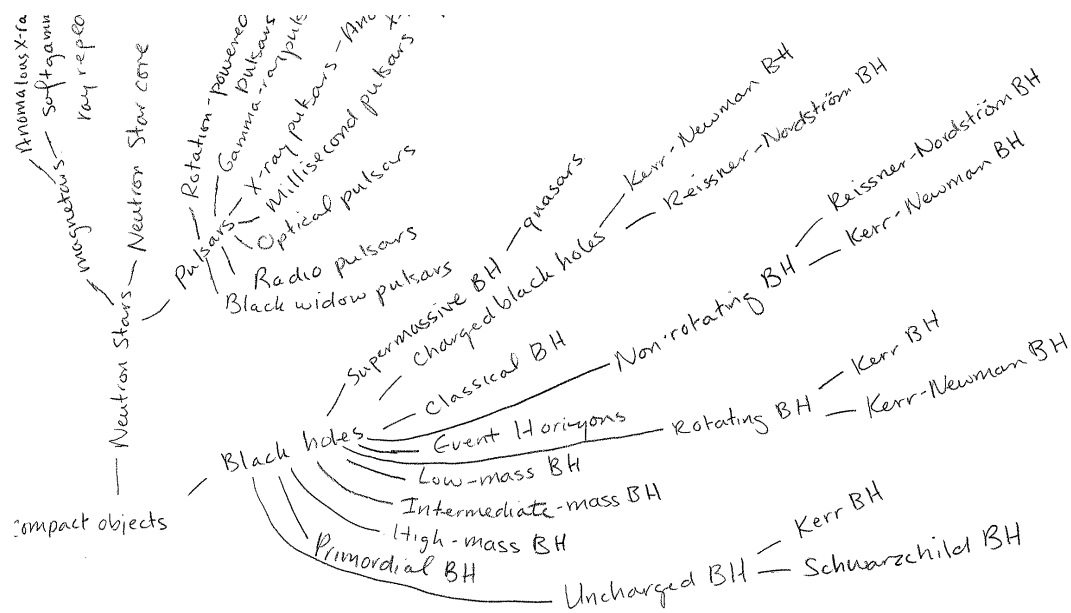


Figure 1. Partial image of a hand drawn hierarchy, describing the relationship between concepts under "Compact objects."

with 40 subsections across the 11 top-level categories, each of which had about 50 to 100 concepts. Librarians participating in the UAT project from many institutions across the United States volunteered to take on the task of working with researchers to evaluate each subsection. They examined the subsection one concept at a time, exploring the validity and usefulness of each concept, along with the inter-relationships between the concepts in their subsection.

The feedback we received from this exercise was extremely varied. Some returned with a table listing each concept, its validity, comments regarding how it would connect to other concepts in the section, and sometimes a short list of new concepts that should be included on the topic. Others returned with hand drawn thought diagrams (see Fig. 1), explicitly describing the relationships between concepts in a field of study. Still others used an early version of the Sorting Tool to return a structured, but barebones, list of suggested changes. Finally, we also received suggestions in prose form, consisting of written paragraphs describing how a section of the UAT should be modified to more accurately describe its field of astronomy.

Based on this feedback, major changes were made to the substructure of each top-level category, but we were happy to see that our high-level reorganization remained intact. By the end of this huge undertaking, we removed about 300 concepts and added approximately 200 new concepts. We also learned a lot about the kinds of feedback we would receive and the kinds of tools our users preferred to use.

The result of this entire process was version 1.0 of the Unified Astronomy Thesaurus, released on December 23, 2015. It included 1836 concepts, with 11 top concepts, a depth of 10 levels, and 329 related concept links. True to our intent, the UAT was released in RDF/XML, using format standards compatible with linked data so that it could be ingested and used by applications that also followed those standards.

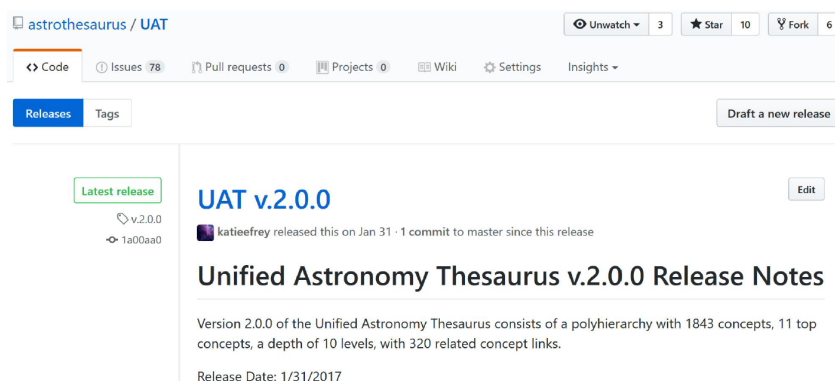


Figure 2. The current version of the Unified Astronomy Thesaurus is available on GitHub.

We decided to experiment with descriptive identifiers and released a minor version update (version 1.1.0) about six months later that also re-standardized the Uniform Resource Identifiers (URI). It was argued that descriptive identifiers would help to make the UAT more human-readable; i.e. a user would know just by looking at the URI which concept was being referred to. However, we found descriptive identifiers to be too rigid for the amount of changes and updates still being made to the UAT. We reverted to numeric identifiers in the version 2.0.0 release on January 31, 2017.

The version 2.0.0 release also included an attempt to add context, clarity, and consistency to a large number of concept labels. For example, instead of just using a label such as "A dwarf," we extended the label to "A dwarf stars," likewise "Double-mode Cepheids" became "Double-mode Cepheid variable stars," and so forth. These changes added additional information to each concept so that they could stand without the context of the UAT hierarchy, allowing them to be more useful as keyword tags. During this update we also discovered and removed 9 duplicate concepts, added 16 new concepts, and moved a few concepts from one parent to another. As of this writing, the version 2.0.0 of the Unified Astronomy Thesaurus is the current version and it includes 1843 concepts, with 11 top concepts, a depth of 10 levels, and 320 related concept links.

4 Contributing to the Unified Astronomy Thesaurus

As described above, the librarians and researchers often used a web-based visualization tool to explore the hierarchy of the UAT. It was a popular tool for analyzing the structure of the UAT since it easily displays the relationships between concepts, allowing researchers to understand how each concept fit into the greater context of the Thesaurus.

During the course of updating the UAT from beta to version 2.0.0, we made extensive updates to this web-based visualization tool. Instead of simply being a way to browse through the Thesaurus, it can now be used to directly submit suggestions and feedback. Upon launching the Sorting Tool ([6]), as we now call it, the user selects a specific branch of the UAT to work on. Then, as with the original visualization tool, they may click open a concept node to view its child concepts. Those concepts can also be expanded to show their children, and so on. The entire hierarchy of a branch can be displayed this way, within the limits of the display device. The extended capabilities now allow a user to click and drag a concept, moving it from one node to another. Changes are immediately visible on the screen, so the user can get a sense for the full context and implications of their suggestions.

UAT Sorting Tool

Choose a branch of the UAT (v2.0.0)*:

--Exoplanet astronomy ▾

Add Concept:

Add Node

Send Your Feedback

Changes made using the sorting tool will be automatically included in your feedback when you submit this form.

Your Name:

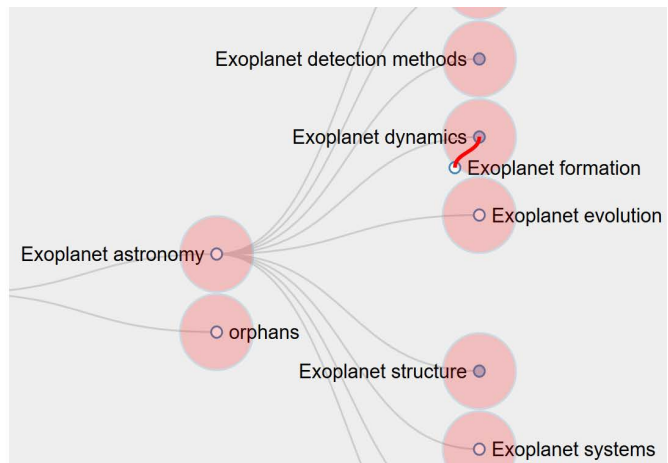


Figure 3. Example of the UAT Sorting Tool in use, concept "Exoplanet formation" is being moved to "Exoplanet dynamics."

Should a concept be deemed unnecessary, it can be moved to the "orphans" node, indicating that the user believes that this concept does not belong in this particular section of the UAT and should be considered for deprecation. Finally, by using the "Add Concept" input box, the new Sorting Tool lets users suggest new concepts. The new concepts will appear on the "orphans" node, and can then be dragged wherever in the hierarchy a user thinks they may belong. (see Fig. 3)

Once a user has finished re-arranging the concepts and has made all of the suggestions that they believe are necessary, they are asked to include some personal information, such as name, email address, and institution. We ask for this information so that we may verify the validity of the suggestions and reach out for questions if needed. A radical suggestion sent from a respected specialist in the field would garner more weight than a similar suggestion from an unknown user. Lastly, there is also a space provided for additional comments, should the user wish to make note of anything in particular or give an explanation for their suggestions. At that point they simply need to click the "submit" button and an email is sent to the Curator for the Unified Astronomy Thesaurus.

5 Tracking Suggestions to the Unified Astronomy Thesaurus

Throughout each revision of the UAT, we have been accepting and incorporating feedback, most of which was received via email. In addition to feedback generated by the Sorting Tool, we would also receive unstructured emails, often containing multiple suggestions. Although many of these suggestions were clear and easily incorporated, some suggestions were complicated, requiring further thought and consideration. Tracking which suggestions from which emails had been enacted and which were still pending quickly became very difficult. Additionally, by holding the suggestions in the personal email inbox of the UAT Curator, we noticed that we were essentially creating a "dark archive" of feedback that only one person could see, act upon, or reference.

Since we had already been using GitHub to publish release versions of the UAT, we decided to use GitHub Issues, the associated platform for managing suggestions for software updates, as our web-based platform for tracking feedback. We opened the first Issue on GitHub on November 18, 2016 by copying the contents of emailed suggestion into the web-based platform. One benefit of using GitHub

was that each suggestion now has an open discussion forum. Not only can users search Issues to see if their feedback has already been submitted, they can create an account to add to the discussion. GitHub also has functionality to collect Issues together into a "milestone," which can then be linked to a release. This helps to create and maintain a log of changes, showing the provenance of the UAT as new versions are released.

6 The Future of the Unified Astronomy Thesaurus

In addition to the pending feedback that has already been received, the Curator for the Unified Astronomy Thesaurus has identified concrete areas that need improvement in order to move the UAT forward. Specifically, the UAT has some obvious gaps in content surrounding the topics of publishing, astronomical modeling, and software. Each of these subjects stands out as something that will only become more important in the future, and yet none are well covered in the UAT. Work has already begun on a comprehensive examination of these topics to draw out relevant concepts that can be used as a starting point for building new sections within the UAT.

To make full use of its linked data capabilities, we need to build connections between the UAT and other existing vocabularies. These would include general linked data sources such as Wikidata, specific vocabularies such as the Space Object Behavior Sciences taxonomy, helmed by the University of Arizona, and a potential taxonomy of astronomical instrumentation.

Another goal for improving the UAT is to include definitions and examples for concepts that are new to the field of astronomy or somehow ambiguous.

As for soliciting feedback in the future, GitHub Issues has created a fantastic springboard for launching conversations. Many of the suggestions are not trivial, meaning that to accept the change as stated might have broader implications over the UAT as a whole. These things require conversation, discussion, and contextualization. As we have learned from our work on the Unified Astronomy Thesaurus over the last few years, the most comprehensive feedback we have received came out of these types of discussions. To that end, we will be piloting focus groups with the aim of hashing out some of the thornier suggestions for the UAT. It is our expectation that these discussions will yield not only reasonable solutions, but also additional suggestions and feedback which will need to loop back into future focus groups.

In order to remain relevant as new discoveries are made in astronomy, the UAT must be considered a continual work in progress.

References

- [1] *Astronomical Subject Keywords* (2013), <http://journals.aas.org/authors/keywords2013.html>
- [2] *IAU Thesaurus* (1995), <http://www.mso.anu.edu.au/library/thesaurus>
- [3] *IVOA Thesaurus* (2009), <http://www.astro.physik.uni-goettingen.de/~hessman/rdf/IVOAT/index.html>
- [4] *Physics Astronomy Classification Scheme* (2010), <http://journals.aps.org/PACS>
- [5] A. Accomazzi, N. Gray, C. Erdmann, C. Biemesderfer, K. Frey, J. Soles, *The Unified Astronomy Thesaurus*. (2014), Vol. 485 of *ASP Conference Series*, vol. 485
- [6] *The UAT Sorting Tool* (2017), <http://uat.altbibl.io/>