Online Discovery: Search Paradigms and the Art of Literature Exploration. Edwin A. Henneken, Alberto Accomazzi, Michael J. Kurtz, Carolyn S. Grant, Donna Thompson, Giovanni Di Milia, Jay Luker, Benoit Thiell and Stephen S. Murray, Smithsonian Astrophysical Observatory, 60 Garden Street, Cambridge, MA 02138, USA, ehenneken@cfa.harvard.edu.

Introduction: Furthering science depends to a large degree on knowledge and information transfer. This makes it critically dependent on discoverability. The universe of potentially interesting, searchable literature is expanding continuously. Besides the normal expansion, there is an additional influx of literature due to interdisciplinary boundaries becoming more and more diffuse. Hence, the need for accurate, efficient and intelligent search tools is bigger than ever. The SAO/NASA Astrophysics Data System (ADS) has been a fixture of research in astronomy and the planetary sciences for two decades. The traditional ADS interface was designed in 1992 and has served our users well, but modern information technologies allow a substantially improved experience. We have introduced a test environment, "ADS Labs" (http://adslabs.org), where we expose our users to new technologies and prototype services. This environment allows ADS users to explore the vast universe of scholarly publications with more powerful tools, and therefore find what they are looking for more efficiently. In this presentation we highlight a number of important aspects of the "ADS Labs" environment.

New Search Paradigms: Even with a sophisticated search engine, looking for information can still result in overwhelming results. A lengthy discovery process costs significantly both in terms of time and money. Ease of access and exploration is essential for efficient information discovery. In other words, new search paradigms are necessary. Different classes of users search for information in different ways. A professional scientist, for example, will use the ADS differently than a graduate student, or a librarian. It is important to provide tools that support these different modes of searching as effectively as possible. While it is important for professional scientists to stay up-to-date with the latest developments, graduate students will often look for review and influential papers on a subject with which they are acquainting themselves. The traditional ADS service offers a search experience that is comparable to shopping in a supermarket: you know what you want and how to find it. With the new search paradigm we take into account that the visitor might have a preference for organic products, with a small carbon footprint (to build on the supermarket analogy). We are giving the ADS users an option to specify beforehand what kind of information would best answer the query, and/or easily filter query results afterwards. ADS Labs should make it relatively straightforward to find answers to questions like "*I would like to learn about spectra, and find related data products, of Trans Neptunian Objects*" or "*I am interested in peculiar strontium lines in solar type spectra and I would like to get additional information based on literature published on this subject and find relevant data products*". The two major components of ADS Labs are a streamlined abstract search interface and a new full-text search (which now includes current publications).

Streamlined Search: The streamlined search provides a basic, one-box search with several options for sorting and exploring, which allow users to retrieve information from different points of view. This supports the different modes of searching, typical for different classes of users. Sorting options represent a criterium to rearrange a results list, which can be by publication date, citations or readership. We also offer a sorting option called "relevancy", which is a hybrid score based on a weighted combination of the other sort options. The modes of exploration are more than just a rearranging of the results list: these are so-called second order operations [1]. Second order operations are actions that take lists which have been generated by a database query, and from those lists form sets of other lists, which can then be merged and sorted on the basis of one or more of the attributes of the items in the lists. This can be used to e.g. find review papers: here we return the list of documents citing the most cited papers on the topic being researched. These are papers containing the most extensive reviews of the field. This mode of searching is very useful for graduate students who are familiarizing themselves with a field. We can also determine the list of documents most cited by the most relevant papers on the topic being researched. These are most likely papers cited by experts in the field. We also offer a way to find the papers that are read the most by people who are interested in the topic being researched.

The results from the streamlined search are displayed in a custom information environment, specific to the query, to further, filter, explore and learn. The main component of this environment is a system of hierarchical categories ("facets"). In addition to this, we also offer some graphical interfaces to further explore relations. These interfaces show the publications in the results list represented as a co-authorship net-

work, offering a view of the distribution of contributors and their inter-relationships, a word cloud based on the abstracts or a sky map, displaying the objects mentioned in these papers. The user can interact with these interfaces to filter the publications in the results list.

Publications in the results list link to an abstract page providing access to the full-text, references, citations, and data products. If available, it also includes a set of recommended publications [2], which are based on text similarity, citations and readership ("collabortive filtering"). The inclusion of recommendations to the usual citations and references links adds an element of serendipity to the usual activity of searching and browsing the literature, with the goal of increased discoverability.

The power of the streamlined search is best illustrated with use cases. We will take the two questions used as illustration at the end of the section "New Search Paradigms".

To explore the first question, we do a streamlined search for the terms "TNO spectra", using the option to find review papers. In the results list, the "Archives" facet (representing availability of data) is used to select publications with links to data products at ESO by selecting "ESO". Clearly, the publications in the results list provide information about what we can learn about TNOs from spectra. When we, for example, select the paper by Barucci et al. (2011Icar..214..297B), the abstract page provides a link to data products via the link 'Archival data'. These turn out to contain VLT observations using the SINFONI spectrograph. The results page makes it easy to explore additional questions like "*who published a lot on this subject?*" , "*what is the temporal distribution of publications?*" and "*which objects are discussed in these papers?*". This type of analysis would have been impossible or at least extremely difficult using the traditional ADS service.

For the second question, we do a streamlined search for the terms "peculiar strontium solar", using the default date sorting. In the results list, select "ESO" and "MAST" in the 'Archives' facet. We use the resulting set of publications to create a word cloud via the "View as…" menu. Displaying the word cloud immediately suggests that Mercury (Hg) and Manganese (Mn) are highly relevant, indicating that you should keep the contributions of these elements in mind. This is a result you could not have obtained in any other way. The publications in the results list provide links to e.g. observations made with HARPS. This is another example of an analysis that would have been impossible with the traditional ADS service.

A search for most relevant, refereed papers on HARPS in the period 2004-2011 provides a nice example for the use of the co-author network: it gives an immediate way to detect distinct groups and collaborations involved in a particular research field.

Full-text Search: The new full-text search interface allows users to find all instances of particular words or phrases in the body of the articles in our full-text archive. This includes all of the scanned literature in ADS as well as a select portion of the current astronomical literature, including *The Astrophysical Journal*, *The Astronomical Journal*, *Monthly Notices of the R.A.S., Astronomy & Astrophysics*, *Solar Physics* and additional Elsevier and Springer journals. Full-text search results include a list of the matching papers as well as a list of "snippets" of text highlighting the context in which the search terms were found.

Full-text searching provides a tool for text mining that is interesting for different classes of users. Researchers will be able to search for specific celestial objects, techniques or other concepts that might not be mentioned in either title, abstract or keywords. For people involved in the history of science, full-text searching will help finding e.g. the first use of a term or concept. Librarians will use full-text searching with systems such as TELBIB and FUSE [3] to create bibliographies for instruments, observatories and organizations. When we search for refereed publications mentioning the spectrograph HARPS in either title or abstract, we find 158 publications, starting in 2004 ("First Light" for this instrument was on February 11, 2003). A full-text search for HARPS finds 560 refereed publications, starting in 2001. Data center administrators can use the search to find references to data set identifiers. For example, searching for a PDS data set identifier like EAR-A-3-RDR-SAWYER-ASTEROID-SPECTRA-V1.2 finds the paper by Jenniskens et al in *Meteoritics & Planetary Science* (2010M&PS...45.1590J) where this data set is cited. This can be used to create links between data sets and publications, or to track how often a specific data set has been used for publications.

References:

[1] Kurtz et al. (2002), *Astronomical Data Analysis II*, *SPIE 4847*, 238. [2] Henneken et al. (2011), *Future Professional Communication in Astronomy II*, *Astrophysics and Space Science Proceedings, vol. 1*, 125. [3] Erdmann & Grothkopf (2010), *Library and Information Services in Astronomy VI*, *ASPC, vol. 433*, 81

Acknowledgement: The ADS is funded by NASA Grant NNX09AB39G.