Short note: SEP data catalog

Abdullah Al Theyab, Gboyega Ayeni and Yunyue Elita Li

ABSTRACT

Motivated by the long-recognized need for bookkeeping of the datasets in the SEP data library, we have implemented a data catalog database supported by a web-based front end. The new database facilitates searching, referencing, and, most importantly, maintenance of our data library. The database design enables direct connection between each dataset and any relevant internal and external publications. We summarize the database design, data catalog structure, project progress, and future directions.

INTRODUCTION

Being a data-oriented research organization, Stanford Exploration Project (SEP) puts significant effort into keeping its large data library organized. In the past, meta-data organization was based on text and latex log files written by previous SEP students (Clapp et al., 1999). Many years after they were introduced, most of the log files broke down as a result of system and software changes or file deletion and reorganization. This breakdown of the log system makes accuracy and consistency difficult for data maintainers and users. Because of the large number of datasets and their non-systematic organization, searching for datasets related to particular problems is cumbersome and time consuming.

The premise for the current effort is the recognition that, by linking datasets to: (1) *keywords* of geophysical problems and data types, (2) data providers, and (3) relevant documents and correspondences, data search and maintenance can be more efficent. Because soft linking using log files is non-trivial (if not impossible), a different catalog design paradigm is required.

Databases are better suited for meta-data organization. They are easy to use, maintain, and query for needed information. Because databases enable complex queries, they can save a lot of research and data maintenance time. They can also be used as a monitoring tool for meta-data consistency. Our current database-based catalog system has the following advantages:

- Easy meta-data access and maintenance.
- Direct links between datasets, providers and provider contacts.
- Automated yearly notifications of expired data licenses.

- References to online and offline copies of data files.
- Automated daily checks and reports on status of data files.
- Automated bi-yearly reread (rotate) for backup disks (tapes).
- Interrelation of different datasets using geophysical and data type keywords.
- Direct links between each dataset and internal and external publications.

In this report, we summarize the new SEP data catalog design and show snapshots of the working website for future reference.

THE ESSENTIAL SOFTWARE

The new database is hosted by the MySQL database management system. The front-end website, run by an Apache server and available only to SEP researchers, is interfaced with the database by using Python modules. Figure 1 shows the basic software elements used in building the new data catalog. Although the database is accessable in several ways, the website is the most user-freiendly.

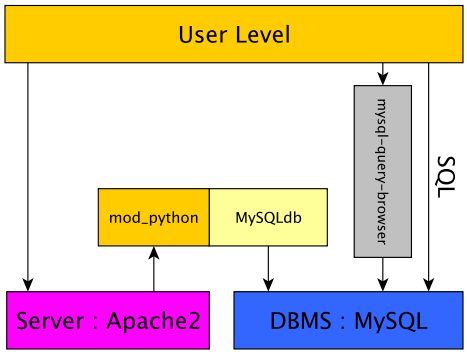

Figure 1: Software elements for the SEP data catalog.

THE CATALOG SCHEMA

The database design (Figure **??**) is composed of many entities revolving around the *dataset* entity. The *dataset* entity is a table of attributes related to each dataset in our data library. Examples of such attributes include:

- *name* : a distinctive name for the dataset.
- *proprietary_info* : information about who can use the dataset, required permissions before publishing results from this dataset, and provider contact(s) for such permissions.

- *lic_exp* : dataset license expiration year (where applicable).
- *sep_handler* : a list of previous and current SEP researchers who obtained and have maintained the dataset.

The *associated data* entity represents the subsets of a dataset. Examples include the velocity, density, and raw data files associated with a particular dataset. *Associated data* entities have both online and/or offline copies.

The *paper* entity relates each dataset to publications in which such dataset has been used.

The *document* entity holds any file that is of a value to future users and maintainers of a dataset.

Each entity in the database has a FIXME flag and log attributes (not shown in the ER-diagram). These two attributes are used to report problems and fixes to data by maintainers and users.

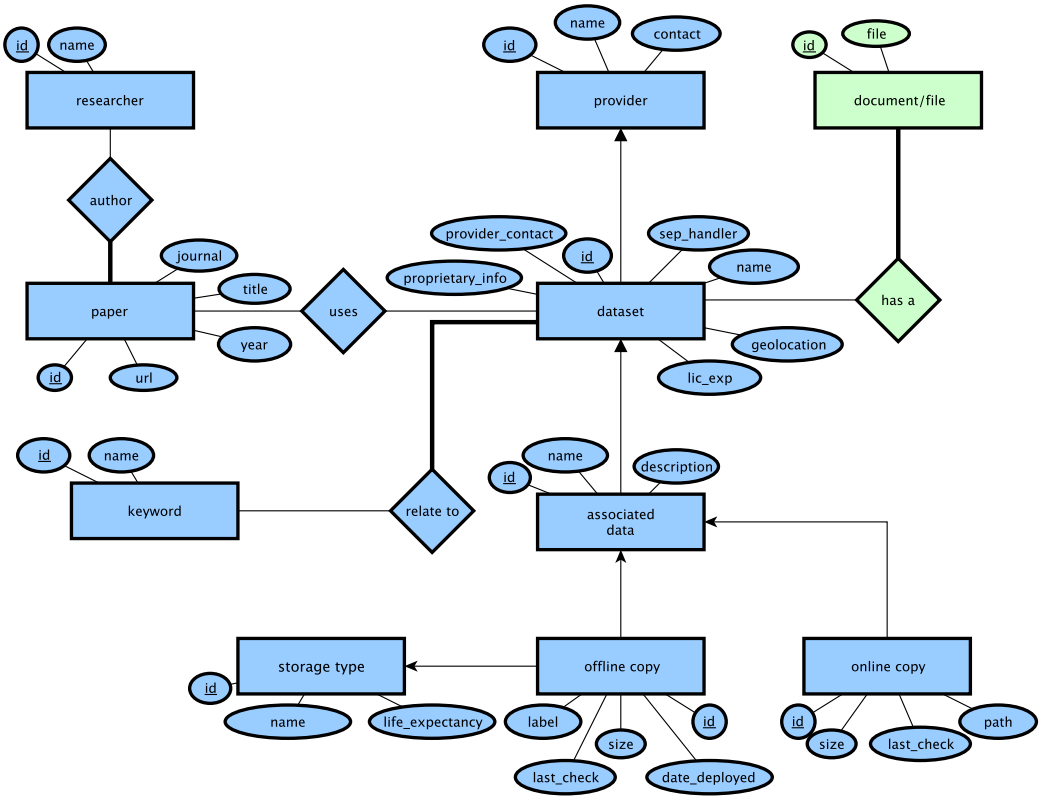

Figure 2: Entity-relation diagram of the SEP data catalog database.

THE PROJECT PROGRESS

The database and the front-end website are active. Entries for legacy and new datasets are being created and the exploratory analysis of these will be uploaded into the database. Figure **??** shows one entry of the database.

Figure 3: A snapshot of the SEP data catalog website.

FUTURE DIRECTION

Information entry— for the datasets and associated entities —into the database will soon be completed. To ensure that the data catalog up-to-date and consistent with online and offline data copies, monitoring and notification tools will be implemented. Because this database is flexible and easy to maintain, we hope that it will be continually expanded by current and future SEP students.

REFERENCES

Clapp, R. G., M. Brown, L. Vaillant, C. Mora, M. Prucha, and Y. Zhao, 1999, SEP's data library: SEP-102.