

A historical survey of technology in the UCB natural history museums

Ginger Ogle May 2005

1960s

- **1965-7** [UCMP] Grad students digitize 152,000 specimen records and 84,000 localities using 80-column punch cards, magnetic tape, and computerized paper catalogs.
- **1967** [UCMP] Rensberger & Berry publish “An automated system for retrieval of museum data.”

1970s

- **1975-8** [UCMP] re-digitizes its vertebrate data using SELGEM (Smithsonian) Cobol - IBM 360 JCL system (80-col punch cards to mag tape + paper listing) improvement: multiple cards for 1 record
- **1979** [MVZ] begins digitizing mammal records using Taxir (Fortran - IBM 360) (U. Mich)
- **1979** [UCMP] transfers its data from SELGEM to Taxir

1980s

- **1984** [UCMP] acquires two PCs for in-house use, uses dBASE/Pascal for label printing, data entry
- **1984-5** [MVZ] gets PCs, writes Pascal programs for data entry into Taxir, uploaded to mainframe
- **1984-5** [UCJeps] begins to database its type specimens using dBASE
- **1987-8** [BotGdn] begins to database its specimens using dBASE
- **1988** [IST] Advanced Technology Planning (ATP) is formed within IST
-

1990s

- **1990** [QAL] relational database evaluation: Sybase, Ingres, Oracle – Sybase is chosen
- **1991** [IST] Architectural Slide Library is developed by ATP using Ingres (later Sybase)
- **1990-2** Society of Vertebrate Paleontology proposes data model (incl. Blum, Lindberg)
- **1991** [UCMP] abandons Taxir in favor of dBASE on in-house PCs
- **1992** [UCJeps] **SMASCH starts** (DB architecture based on Arch. Slide Library Project)
- **1992** **MIP established** – takes over Arch. Slide Library, SMASCH
- **1993** [UCMP] puts type specimen data online via gopher
- **1993** [DLP] CalPhotos developed as Chabot using PostGRES, TCL
- **1993** **Mosaic appears** (first web browser) March
- **1993** [UCMP] website comes online (grad student Rob Geralnik) August
- **1994-5** [UCMP] shifts from dBase to Paradox for specimen database
- **1995** **Digital Library Project begins**
- **1995** [DLP] CalPhotos goes online as “Cypress”
- **1997** [DLP] CalFlora comes online
- **1997-8** [MVZ] transfers data from Taxir to Sybase
- **1999** [MVZ/DLP] online specimen DB goes live
- **1999** [BotGdn] dBASE database ported to Sybase with xdb interface

2000s

- **2000** [MVZ/DLP] AmphibiaWeb goes live
- **2002** [UCMP/DLP] online specimen database goes live
- **2002** **BNHM Informatics begins**
- **2003** [EME/DLP] database goes live
- **2003-4** Distributed systems PaleoPortal, HerpNet, Manis, et al

Data Management and Storage in the BNHM

Nearly all of the digitized data that currently resides in the natural history museums is contained in some type of *relational database*. A variety of relational systems has been used over the years including server-based commercial databases such as Sybase and Informix, PC-based databases such as Access, Paradox, FilemakerPro and dBASE, and open source database systems such as MySQL and PostGres. This section briefly describes the different types of databases in use.

Relational Model

The relational model was first described by E. F. Codd in 1970, and was based on relational algebra. It was developed as an alternative to traditional methods that tied data storage and retrieval to the architecture of a particular computer, such as the early databases used in the MVZ and UCMP. Relational databases organize data and the relationships among them into tables of rows and columns. Each column and row in the table is unique, the sequence of columns and rows are insignificant, and the contents of one cell are considered atomic. The relational model is used in most current database systems today from simple Excel spreadsheets to more elaborate data management systems such as Sybase and Oracle. Relational databases can be indexed efficiently so are usually much faster than hierarchical or object-oriented models. Because of their table structure, they can support a much more flexible array of queries. All the museums currently use some type of relational database.

Hierarchical and Object-Oriented Models

In this model, data is organized in a tree or web structure instead of a table structure. Nodes usually represent “parent” or “child” objects. The relationships between objects are represented by the segments that connect them. Relationships may be one-to-many, as in a hierarchical arrangement of taxonomy, or many-to-many, as in a network or web. [Hierarchical databases were popular in the 1960s and 1970s](#), and object-oriented databases enjoyed popularity in the 1990’s for their ability to more intuitively represent certain types of data such as time-sequence video. The downside of storing data this way is that the relationships between data elements must be determined in advance and are fairly static -- this type of database handles hierarchical data elegantly but is not conducive to a broader relational query such as “...where date is after 1980 and collector=”Smith”. Queries to hierarchical databases are usually much slower since comparisons must follow a tree or web structure.

Hash Tables or Look-up Tables

This model uses a key-value method of storing data, and typically does not support the expression of relationships among different data elements. Look-up tables such as BerkeleyDB are widely used in web applications where a large body of data must be searched quickly on one or a few fields, such as retrieving customer account info given the customer’s name. They are also used for storing data in small consumer appliances, because they are very fast and very lightweight. Look-up tables are useful when queries only need to be made on one data element at a time, for example, retrieving all records that match a particular taxon. UCJeps uses look-up tables for some of its online queries, taking an extract from the Sybase data and then creating a hash table for each searchable field. Look-up tables are not well suited to datasets that need to be queried on multiple fields.

Online databases

What technology is used by very large, web-based databases such as eBay, Amazon, and Google? They typically do not use traditional relational database technology at all, except perhaps for archival or permanent storage. Much of the functionality that databases traditionally provide is instead moved to the application level, such as joins, sorting, and indexing. Data is partitioned and then replicated so it can be distributed across many CPUs for load balancing, failsafe protection, and above all, fast access. Searches are performed on customized replicated indexes. For example, eBay performs both sorting and joining in the application, not in the database. They use few triggers, and have moved towards “denormalization of the schema [Vogel]”. The emphasis is on scalability and speed. A recent essay by Adam Bosworth lists three requirements that are needed for today’s databases that are not being provided by database vendors: dynamic schema, dynamic partitioning, and modern indexing.

Sources:

“How are databases used at big customers?” Werner Vogel, CTO - Amazon.com, <http://weblogs.cs.cornell.edu/AllThingsDistributed/archives/000280.html>

“Where have all the good databases gone?” Adam Bosworth, Dec 2004, <http://www.adambosworth.net/archives/000038.html>

“Peeking Into Google”, Susan Kuchinskas, internetnews.com, <http://www.internetnews.com/xSP/article.php/3487041>

PC-Based vs. Server-based

Single-user databases that run on a personal computer include Access, Paradox, and dBASE. These databases have been very popular (and still are) because they are easy to set up and do not require a great degree of skill to begin using. During the 1980's as personal computers became affordable, many of the Berkeley museums began to manage their collections using dBASE, a product that was widely available for PCs. This represented an attractive alternative to timesharing on a remote mainframe computer, not only because it was less expensive but also because the museums had more control over the data and access to the computer. Today, most of the museums continue to use PC-based databases for small research projects as well as for collection-related data such as loans, acquisitions, images, etc. One museum still houses its specimen data on PC-based database: UCMP's data resides as four separate collections on the PC database Paradox. The main disadvantage of using single-user PC-based databases is that the database is only available to one person at a time for updates and additions. Museums often work around this by creating several copies of the same so that multiple people can work on the data at once. Keeping the data synchronized when there are multiple versions is difficult and time-consuming. In the late 1980's, client-server databases began to appear, such as Ingres, Sybase and Oracle. The client-server architecture allows multiple users to query and update data at the same time. With the explosion of the World Wide Web, server-based databases also offered a way to make data available online to a wide audience.

Distributed databases

As soon as museum collections began to be digitized, interest turned to creating structures that would support queries beyond the museum's own collection. One way to do this is for all the participating collections to use the same database and/or data structure. This was suggested as early as 1972 in UCMP's "A Local System of Automated Paleontologic Data Retrieval and its Potential Contribution to an Eventual Nationwide System". UCMP was an early adopter of technology and expected that other collections would quickly follow their lead. But this was not to happen, perhaps because technology changes seem to be closely tied to funding. Similarly, SMASCH was originally conceived as a distributed database system, and the 1992 paper that described the new project announced that all the herbaria in California would share this system. Although some other UC herbaria are now actively using SMASCH to enter some of their data, SMASCH has not become the state-wide system that was envisioned. Without a large funding effort, most academic museums have either not been able to digitize their collections, as in the case of the Essig Museum, or have used more affordable methods such as PC-based databases that often do not interface well with more elaborate projects. Even when funding is not an issue, museums may not agree on what components are important for a system – systems at multiple museums rarely share enough common design elements to support this type of distributed query.

Another way to support distributed queries is for each collection to make only a subset of its data available in an agreed-upon format, and then either run an application locally that allows queries from the outside, or supply a third party with the data who will then make it available for queries. The DiGiR project uses this approach. This method is more attractive than a shared data structure since participants don't need to commit to a particular overall data structure for their entire collection management system. Another appealing feature is that they retain control over their own data. A disadvantage of this method is that some work is needed at each site to produce the data subset and to run the application that makes it available. For museums that are struggling to fund basic collection management, participating in a distributed system may not be economically feasible.

A third method for supporting distributed queries across collections is to create a centralized repository. In this model, participants contribute data in an agreed-upon format as in the previous example, or they may simply supply the data in whatever form it exists in their own system, and leave re-formatting to the repository. This method is attractive because of its very low cost for each participant. Each collection need only create a mechanism for exporting data from the local collection management system, a task that is straightforward for nearly any type of database. The repository can automatically pick up data, format it,

and finally include it in the distributed system. A distributed query system that uses this approach is the BNHM all-museum queries.

Source: "Database Models", UnixSpace <http://unixspace.com/context/databases.html>

Normalization

Data normalization is a method for removing redundancy from data, to reduce storage requirements and also to increase control over data integrity. An example of normalization that all the museums use is Collector information. Each specimen record contains information about the collector of that specimen, such as the person's name, address, and affiliation. Typically one collector appears on many specimen records. Rather than repeating all the details about the collector on every specimen record he/she collected, we instead make a separate table of unique collectors containing all their information, assign a unique number to each collector, and then store only the collector number in the specimen record. This not only reduces the amount of data in the specimen database, but it also simplifies upkeep on collector information and ensures that any corrections to collector information will be reflected identically on all specimen records. The specimen database is now said to be *normalized*. At query time, if "collector" is part of the query or should be included in the results, the database system performs a *join* on the specimen and collector tables. The collector number is the *primary key* for joining the specimen table with the collector table. The process of normalizing the main table may continue with other fields, and has been formalized as a set of rules ("first normal form", "second normal form", and so on.) A highly normalized database contains many tables, each of which has few data elements. A database that is minimally normalized has few tables, each having many data elements. All the museums normalize their databases to some degree.

Too much normalization can itself create problems, however. One problem is complexity. The more a database is normalized, the more new tables are created, and the more complex the database structure becomes. New sub-tables may be joined to the main table, as in the Collector example, but they might also be joined to other sub-tables. For example, we could normalize the collector table and create a new table of affiliations, using a number or an abbreviated code in the collector table as the primary key. Highly normalized databases may contain hundreds of tables that are related to one another in complicated ways. Only the most experienced users might understand how data is stored in the database and how the data elements are related to one another, which can impede the construction of efficient queries. Adding a new field may impact dozens of tables, making changes complicated and time-consuming. Creating new applications for previously unanticipated functionality may be slowed because of the knowledge required to understand basic relationships within a complex database structure. Flexibility and adaptability are reduced.

A second problem has to do with limitations in the database software itself and allocation of resources on the host computer. Joining two or more tables to perform a query is a computationally expensive task, since all the affected tables must be fetched from the database and loaded into memory in order to make the necessary comparisons among records to perform the query. A database with hundreds of related tables may require the joining of many tables to perform even basic queries, tying up the CPU and yielding very slow response. In the worst case, the query may be impossible -- some queries simply cannot be performed because the database vendor has placed an upper limit on the number of joins allowed. When BNHM began selecting Darwin Core elements from all the museums' databases it was found that the Sybase limit on joins was exceeded for the MVZ data model when the necessary query was performed. Even if the database system can support any number of joins, the time needed to perform the joins may render a query that is unacceptably slow for users who are accustomed to immediate results on the web.

Normalization in databases used by the museums

Collection	# Tables	Database	# Records	
AmphibiaWeb	4	MySQL	9,000	
Botanical Garden	100	Sybase	39,000	
CalPhotos	9	MySQL	87,000	
Essig	5	MySQL	18,000	
Field Notebooks	6	MySQL	6,000	
MVZ in house	230	Sybase	640,000	Including 110 data dictionary tables
MVZ online	6	MySQL	640,000	

UCJeps SMASCH	75	Sybase	475,000	
Personal Library	2	MySQL	5,600	Contains MVZ reprints, Essig journals, etc.
UCMP in house	2	Paradox	300,000	4 partitions
UCMP online	2	MySQL	300,000	

Sources: Effie Dilworth, Joyce Gross, Dick Moe, John Wieczorek

Controlled Vocabularies (aka Authority Files)

A controlled vocabulary is a list of standardized terms that are “officially” acceptable for a particular data element in the database. Examples of controlled vocabularies include countries, collectors’ names, and scientific names. All of the museum databases use controlled vocabularies to some extent in order to control data integrity and maximize query precision -- if a collector’s name is spelled multiple ways in the database, it will not be possible to retrieve all the records for that collector. The standards that are used vary from museum to museum. Some museums use international standards such as ISO-3166 for country names, while others use their own internal standards. Different standards among the museums are sometimes due to the geographical extent of the collection: for example, relatively few UCJeps specimens were collected outside California, so State and Country are less important than in other collections. UCMP and Essig contain a large number of specimens collected world-wide, so more attention must be paid to non-US regional and political boundaries. In some cases, the type of specimen being collected affects the standardization. Mammals tend to be collected on land, so a list of continents is straightforward, but fossils of microorganisms are taken from ocean floor core samples, so a list of continents for UCMP must take this into account. For the UCMP, a standardized list of epochs and periods is essential, but these are not used at all in the other museums. The museums generally exercise local control over standards used for species names. The SMASCH database looks to the Jepson Manual for its standardized list of taxa, which may include names not in use outside of California. When museums use services that are available to other museums, some negotiations and relaxations must take place. For example, the CalPhotos database uses ITIS to enforce taxonomic standards, but will accept forms that ITIS does not recognize in order to accommodate museum-specific standards that differ from ITIS. On the other hand, the taxa for all amphibian photos in CalPhotos must conform to the list of names recognized by AmphibiaWeb, because AmphibiaWeb is sufficiently authoritative for world-wide amphibian species.

How does the database enforce its controlled vocabulary? Database systems often include a facility to specify the values that can appear in a particular field. However, most of the museums perform their enforcement in applications outside of the database. For example, the UCJeps SMASCH data entry forms require the user to enter a taxon that is recognized by the database, as does MVZ’s Versata system. The UCMP is unique in that it maintains an accepted list of around 50,000 locality names (“dig sites”) in addition to the standardized county, state, country, and continent lists that most of the other museums maintain. The Essig database web form forces the data entry person to select from a list for all controlled fields. Most of the museums restrict access to the controlled vocabularies, so that only authorized staff can add new items to the list.

Integrity Constraints (aka Business Rules)

Integrity constraints (also called “business rules”) are a set of rules that data must conform to in order to ensure correctness and consistency. Constraints are applied when new data is added, or when existing data is changed. For example, if the state is “Alabama” then the country is constrained to be “United States” and users receive an error notice if a different country is entered. Most database systems support some type of integrity constraints. Even an Excel spreadsheet will attempt to force an incoming value to a standardized date form if the cell has been pre-defined as such. In an SQL-based database like Sybase or MySQL, a field can be defined as being an integer, or text, or “not null”, and an error will result if a value is entered that doesn’t conform. Integrity constraints are often implemented outside of the database at the application level. The MVZ uses Versata to enforce integrity constraints on its Sybase database, while UCMP uses those provided by Paradox for its in-house data. The DLP uses Perl functions tied into its data entry and data correction applications to prevent users from entering or correction data that is not properly

constrained. Putting integrity constraints into an application rather than relying on the DBMS usually speeds up data entry and correction, and offers more control and greater flexibility. A similar technique is the use of *triggers* – database functions that cause a field to be modified when a related field is changed. For example, if the scientific name is changed, a trigger might be defined that will automatically update the family, class, and so on. Triggers are convenient but they can result in complex interdependencies that may become difficult to maintain in a large database.

Data Models (aka Schema)

A data model is a conceptual model for how data is stored and queried. For a relational database, the data model describes the logical organization of the database – the tables and their relationships to one another. The data model is often represented graphically using a tree structure or flow charts. An E-R graph (Entity-Relationship, coined by E. F. Codd, the father of relational database theory) is one example of a graphical representation of a data model. *Schema* is often used as a synonym for data model.

The MVZ 's 1996 description of its new data model laid out two approaches: “whether to maximize the representation of business rules in a highly normalized design, or whether to anticipate database implementation by making “concessions” that improve performance of the most common tasks”. The former approach was taken. Three of the museums – MVZ, UCMP, and UCJeps - developed data models in the early to mid 1990's that are characterized by a complex, highly normalized structure, made up of many tables.

In 1990, the SMASCH proposal was submitted to the NSF, and by 1992 SMASCH in collaboration with MIP had designed a data model that it began to implement as a relational database. The UCMP model had its origins in a 1990 [Society of Vertebrate Paleontology](#) workshop attended by David Lindberg, Director of the UCMP, who formalized the model in 1992. The report from the 1990 SVP workshop had been edited by Stanley Blum, who later co-authored the MVZ model in 1996. The UCMP model was never implemented. By 1995 UCMP had begun to keep its data in-house in a Paradox database with the intent of implementing the data model, but a funding shortfall and staffing changes left museum scientists to fend for themselves. Said one of the collection managers: “we actually found managing some of the multiple tables a pain.... They were nice in principle, but annoying in practice...we realized we didn't have to break up the tables so much and found it much easier to do our jobs with the data in just a handful of tables.”

The Digital Library Project originated in the Database Research Group at UC Berkeley, which developed PostGres. When the DLP began in 1995, a photo database and a document database already existed from earlier projects. The document database was part of a distributed repository of technical documents (CSTR, 1992). Its schema was based on a standard, RFC 1357, whose goal was to link traditional library standards such as MARC to “the on-line world of digital objects”. The DLP used this model for environmental documents and adapted it over the years to accommodate new types of documents. It is presently used for the Personal Library service. The image database, then called Chabot, had originally been included in a spatial database schema called “Big Sur”, so named because of its proposed ability to include diverse data types. The photos shared few data elements with the spatial data in Big Sur, however, and mapping metadata for a photo database onto a spatial data schema proved cumbersome and awkward. Chabot next tried adapting the schema developed for documents, but this too failed to map easily, so a simple schema was designed that included only those elements relevant to the collection of images. The CalPhotos schema has been very dynamic over the years, retaining most of its original elements but tripling in size and becoming more normalized as needs changed. For example, an annotation system implemented in collaboration with the CalFlora project resulted in the creation of two new sub-tables. Other images collections such as the Corel stock photos and the Fine Arts Museums images received their own schema, modeled after CalPhotos and sharing functions and utilities, but customized for the particular application in which those were used. A union schema that included all the image collections was developed and maintained for a time, but it was rarely used so it was later discarded. When the MVZ and AmphibiaWeb databases were designed, the data model was quite different from that of the photos, but all shared application-level integrity constraint libraries, query mechanisms, and update facilities. This made it possible to quickly develop systems for new collections, since most of the

application-level functionality from existing collections could be re-used, allowing developers to focus only on the elements unique to the new collection, such as the schema. The UCMP online database went live just three months after initial talks began between DLP and UCMP. The flexibility and adaptability of the DLP data model has made it possible to support constantly changing needs within the museums and new applications that have arisen from collaborations outside the museums. The schema is assumed to be dynamic, and this has indeed proved to be the case for all the databases developed by DLP. The model has weathered several changes in database software, from PostGRES, to Illustra, to Informix, and currently to MySQL. MySQL is particularly well-suited to online databases where access is predominantly read-only, i.e., many queries but not as many updates and inserts, which is typical of the museum collection databases. Experience with the CalPhotos, AmphibiaWeb and Essig data upload and correction systems is showing good performance as well.

Sources

Conversations with Patricia Holroyd and David Lindberg

MVZ: <http://www.mip.berkeley.edu/mvz/cis/mvzmodel.pdf>

UCJeps: http://ucjeps.berkeley.edu/smasch_dist_doc/tables.pdf

UCMP: <http://www.ucmp.berkeley.edu/museum/datamodel/datamodel.html>

UC Museum of Paleontology

“Deficiencies in record-keeping as well as the knowledge that most UCMP staff were unaware of the actual breadth of the collections ... led to the decision to use a computer system...” Berry, 1972

When UCMP decided to digitize its specimens and localities in the mid-1960s, it could not afford to purchase its own computer. An IBM 7094, a mainframe for scientific applications that was introduced in 1962, cost about \$3.5 million. Timesharing was the only way most academic users could get access to a computer. UCMP decided to rent computer time from the University's Computer Center, which included an IBM 7094. The IBM 7094 had only 32K of memory. Therefore, UCMP did not expect to be able to permanently store its specimen data on the computer. Data would be entered on 80-column punch cards, one card per specimen or locality in order to save computing time. Since 80 columns did not accommodate a full specimen record, fields to be digitized were selected carefully. Abbreviations were used as much as possible to render human-readable output, but the contents of most fields were encoded so as to take up as little space as possible, and translation dictionaries were written to expand the encoded fields.

In order to query the data once it was digitized, a sequential scan of the magnetic tape would be necessary to load all the data into the computer's memory, and then perform the necessary computations to produce the desired result. Computer time was very expensive, so UCMP sought to minimize its use. They decided to print paper catalogs from the digitized data, and use these catalogs to examine and query the collections. The catalogs would be updated periodically. The printed output was carefully sorted and formatted so that selected fields could be easily scanned by eye by museum scientists and students. Decisions were made about which of the fields would be queried most often and which would produce the most useful results. For example, fields for which no data existed for most records would not be queryable since it was not cost-effective. When new records were added, the set of punch cards would be mechanically sorted to reduce computer time, and then loaded into the computer's memory along with the Fortran code. Results were written out to the paper catalogs and the data was backed up on magnetic tape. The Fortran programs tied directly to the machine language for the particular architecture. Thus a further complication was that whenever the Computer Center upgraded its hardware, all of UCMP's programs had to be modified for the new platform.

The UCMP continued to use punch cards and paper catalogs when it transitioned to SELGEM in the mid 1970s with the added benefit that more than one punch card could be used for a single specimen record. However, the paper catalogs were only updated once a year or so. When UCMP followed MVZ in adopting Taxir in 1979, it gained the benefit of being able to query data interactively rather than relying exclusively on paper catalogs, since the new mainframes had disks on which data could be stored. However the primary record for still the paper catalogs, which were updated less and less frequently. Personnel changes within the museums left few staff who were knowledgeable about the system.

Complicating matters was the physical administration within the museum due to the wide variety of holdings. When UCMP first began digitizing records, its holdings were grouped into five separate collections: vertebrates, plants, invertebrates, protozoa, and microorganisms. Protozoa has since been included in the Micro collection, but the four collections are still managed separately. Queries on one collection were often quite different from those on another. For example, vertebrate records were usually identified to genus, so queries on the taxon were common. Microorganism specimens on the other hand were rarely identified below the level of class, so queries on the taxon were rarely performed, but environmental data was very important for these specimens. Thus, when UCMP relied on paper catalogs, a different output format was needed for each of the collections, so that catalogs could be customized according to how they would be used.

In the 1980's when PCs made in-house computing affordable, UCMP began to keep some of its records in dBASE, and by the early 1990s, collection managers were using PCs to manage specimen data. UCMP did not make the jump to MIP-managed Sybase databases when MVZ and UCJeps did a few years later, and they never implemented the highly normalized relational data model they had developed in the early .

In 2002, UCMP and DLP collaborated to bring UCMP specimen and locality data online using essentially the same structure as the in-house Paradox database.. That database currently contains 250,000 specimen records and 50,000 localities.

Sources:

"A Local System of Automated Paleontologic Data Retrieval and its Potential Contribution to an Eventual Nationwide System", William B. N. Berry, Journal of Paleontology, v. 44, p. 527-535, May 1970.

"An Automated System for Paleontologic Data Retrieval – A Case History", William B. N. Berry, 24th IGC, 1972 – Section 16.

"The IBM 7094 and CTSS", Tom Van Vleck <http://www.multicians.org/thvv/7094.html>

Conversation with David Lindberg

History of the UCMP Data Model <http://www.ucmp.berkeley.edu/museum/datamodel/history.html#Ref>

Museum of Vertebrate Zoology

MVZ did not begin using a relational database until 1997-98. For nearly 20 years MVZ's specimen data had resided in the mainframe-based Taxir program. MVZ designed a new data model in anticipation of using a relational database, and it is the one used today. The data model is described in a 1996 paper authored by Barbara Stein and Stan Blum. The new data model would include the data then in Taxir as well as other non-digitized data within the MVZ such as images and field notebooks. The model was based on a concept called Object Role Modeling (ORM), created by Terry Halpin (see <http://www.orm.net/overview.html>). ORM is a method for developing databases "at the conceptual level where the application is described in terms easily understood by non-technical users." ORM was chosen for the MVZ model because a highly normalized model was desired, and "ORM is better suited to developing a fully normalized logical model". The MVZ data model is unique among the other museums in that it uses one data model to represent many different types of data, such as species records, photos, and sound files. The MVZ data model was implemented as a Sybase database which was administered by QAL, which had been managing UCJeps' Sybase database for several years. For data entry and management, MVZ developed a system using commercial software called Versata, a "comprehensive business rules platform for agile application development". The Sybase/Versata system was not conducive to online access, however. The same qualities that produced a conceptually pleasing schema proved anathema to fast online queries. Because of the complexity of the highly normalized data structure, common queries were cumbersome and in some cases, not possible because of Sybase limits on the number of joined tables. The Digital Library Project developed a data model for MVZ specimen data in 1998-99 that enabled online queries, and was based on previous data models developed for CalPhotos and CalFlora. The AmphibaWeb database was developed at the same time. Data is extracted from the Sybase database weekly, reformatted, and loaded into a DLP mysql database.

Sources:

Blum, S. D., and B. R. Stein. The MVZ Collections Information Model. pp. 1-141. Collections Information System Re-Engineering Project, Museum Of Vertebrate Zoology, University Of California, Berkeley.
<http://www.mip.berkeley.edu/mvz/cis/mvzmodel.pdf>

Digital Library Project

The Digital Library Project's initial involvement with the natural history museums had more to do with serendipity than with any intentional planning. DLP's databases and query systems were originally developed by computer scientists to provide a test bed of images and documents for computer science research. By chance, some of these test beds turned out to contain data of value to the natural science

community. Early collaborations with the museums grew from a enthusiasm for biology among computer scientists and an enthusiasm for technology among biologists.

When the NSF-funded Digital Library Project began in 1995, its research group included faculty and students who had worked on two previous Computer Science research projects: the Sequoia Project, a collaboration of computer scientists and earth scientists building database technology for environmental data, and the CS Technical Report project (CSTR), a distributed query system that had begun in 1992 among multiple universities for computer science technical documents. Sequoia had been a project of the Database Research Group, Michael Stonebraker's group at Berkeley that developed PostGres, so there was a strong emphasis on database technology. One of the Sequoia projects was an image database called Chabot developed in PostGres1993-4 by Ginger Ogle. It contained 12,000 photos from Sequoia partner California Department of Water Resources (DWR) and supported queries on image content as well as on text metadata. The nascent Digital Library Project included computer vision researchers who planned to use Chabot as a test bed for their research.

One of the partners in the Digital Library Project was CERES, a new state agency for environmental information. CERES included former DWR staff who'd worked on the Sequoia project, and they wanted to adapt the CS Technical Report Project system to scan legacy environmental documents, database them, and make them available online. Joyce Gross was in charge of this project. Over the next few years, around 2,000 environmental documents were made available online including County General Plans for all California counties. As with the image database, the document database was used by UCB computer scientists as a test bed for research (document decoding, image analysis, natural language processing, and new document models) but it represented a valuable collection of documents not available online elsewhere.

In 1996, CERES asked DLP to put 12,000 images of California wildflowers online using the Chabot image database. These photos had been taken by Brother Alfred Brousseau and were being digitized by a colleague at St. Mary's. (Interestingly, CERES had earlier asked SMASCH to partner on this but they declined.) When the Brousseau photos came online that year, they began to attract the attention of botanists who were creating online information systems about plants. A botanist at Texas A&M, Hugh Wilson, had created an online database of California plant names using data collected by a US Forest Service plant ecologist, Ann Dennis. Wilson and DLP developed a system to link individual photos at DLP and their taxonomical information at TAMU. In 1997, DLP collaborated with Dennis to create CalFlora, an online database for California plant taxonomy as well as a synonym table and a plant observation database which grew to include nearly a million records of plant occurrences. CalFlora interacted with CalPhotos, displaying a photo for each taxonomic page, as well as with the DLP-developed GIS Viewer, which displayed occurrence data on a map. CalFlora established itself as a non-profit corporation in 1993 and set up its system offsite.

In 1998, DLP staffer and herp aficionado Joyce Gross took an amphibian decline class taught by David Wake, and for her class project decided to create an amphibian decline database modeled on the CalFlora database. This class project sparked discussions between MVZ and DLP, culminating in the MVZ online specimen database, which went live in 1999. AmphibiaWeb came online early in 2000. This laid the groundwork for a collaboration with UCMP in 2002. By using existing DLP systems such as MVZ, CalFlora, and CalPhotos, DLP was able to develop an online system for UCMP in just 3 months. That same year, BNHM hired informatics coordinator John Deck, and BNHM and DLP began collaborating on data standards as Deck was building the BNHM all-museum databases. Deck facilitated a collaboration between the Essig Museum and DLP, and Essig's database came online in 2003. The Essig system is unique in that it is the only museum database that uses that same system for in-house and online databases.

Big Sur: <http://s2k-ftp.cs.berkeley.edu:8000/sequoia/schema/html/BigSur/schema.html>
CSTR: <http://www.cnri.reston.va.us/describe.html>

University and Jepson Herbaria

The botanical community has played a leading role in technology development in the Berkeley natural history museum collections. The Taxir system, developed at U. Michigan to manage botanical collections, was first used by the MVZ in 1979 to begin digitizing its collection and was adopted by the UCMP shortly thereafter. Taxir represented a great step forward in collections management since museums could now access their data interactively via terminals in the museum rather than batching data and sending it off to be processed remotely. Because data was stored efficiently on disk, live queries could be run rather than relying on infrequently-updated paper catalogs. Taxir proved to be robust enough to remain in use for nearly 20 years in the MVZ. One of the botanists associated with Taxir at U. Michigan was Tom Duncan, who came to UCJeps in the late 1970's and became the Herbaria director in 1982. By 1984, UCJeps had begun to digitize its type specimens using the PC-based dBASE database. In 1990, Duncan spearheaded the SMASCH proposal, which planned to use a relational database, Ingres, to manage "an estimated 3.5 million herbarium specimens of California plants housed in California herbaria." Data entry and query mechanisms were modeled on the applications developed by IST's ATP group for the Architectural Slide Library. As SMASCH got underway in 1991, Duncan was the impetus for the founding of MIP, which was conceived as "a collaborative effort to coordinate the application of information technology in the museum and other organized, non-book collections on the campus." MIP represented a synthesis of the technology being developed for SMASCH by Duncan and his associates, and the development projects of the ATP group within IST which had developed the image database for the Architectural Slide Library. By the end of the 1990's, SMASCH had digitized more than 300,000 of its specimen records, and continues to use the SMASCH data structure (Sybase database with X-windows GUI it developed for SMASCH.) A web-based query system for selected fields was also supported. In addition to digitizing specimen records, SMASCH photographed and digitized around 40,000 of its mounted specimens. Around 2,000 of these are accessible in CalPhotos. (Unfortunately the original quality of these digital images was not sufficient to justify the expense of retrieving them all from tape for inclusion in CalPhotos. Though the images are intact on tape housed within MIP, the hardware to read the tapes is hard to find.) In the late 1990's, the Jepson Manual, the authority for California plant taxa, was partially marked up in XML, making it possible to extract key elements from the text to build a web-based reference. Building on this, the Jepson Interchange came online, which taxa lists and other information from the Jepson Manual, to the SMASCH specimen database as well as other resources such as the CalPhotos database. The Jepson Interchange is based on formatted text extracted from the Jepson Manual. It is rebuilt periodically. A more flexible query system to the SMASCH database has recently replaced the original one, which uses highly optimized hash tables on key data elements from the database. UCJeps has plans to implement both the Interchange and the online queries in a relational database.

Sources:

"The Specimen Management System of California Herbaria as a Model for an Inter-Institutional Distributed Database System", B. Bartholomew and T. Duncan, 1992.

"Lessons From The Berkeley Museum Informatics Project", Barbara Morgan, December 1993,
<http://www.educause.edu/ir/library/text/CNC9344.txt>

"Electronic Activities of the University and Jepson Herbaria", Richard Moe, Madrono, Vol. 47 No. 4, 2000

Essig Museum of Entomology

The Essig Museum of Entomology has by far the largest specimen collection on campus (estimated at 12 million records and over 30,000 species) but until 2003 had no digital catalog system, probably due to the sheer mass of the collection. Few entomology museums have digitized their holdings. Taking advantage of the need for a new design from scratch, Essig decided to plan a collection management system that is completely online, that would integrate all aspects of data management within the museum including specimen records, photos, accessions and loans, data entry and correction, and publications. Essig had begun using the DLP Personal Library system in 2002 to scan California Insect Survey journals, which would provide a start for assembling species lists within the museum. In 2003 Essig began working with the Digital Library Project to create a web-based specimen database with data entry capabilities modeled on previous DLP projects such as UCMP online, AmphibiaWeb, and CalPhotos. In April 2003 the first specimen records came online, and a few months later, accessions and collector tables were added. Early in 2004 a set of 12,000 aphid records, along with a photo of each specimen, was added, and by the end of the year the number of digitized records had reached around 20,000. A web-based data entry form which directly updated the specimen database was added in late 2004. A few months later, a database of 42,000

unique species names came online which will be integrated into the data entry system. Records in all the databases can be corrected by any Essig staff with the proper permissions. As of this writing (May 2005), a number of museum scientists and faculty are using the data entry form to add new specimens not only for Essig but also for other entomology museums that do not have digital cataloging in place. The next challenge for Essig is to obtain funding to digitize its existing specimens – a daunting task.

Natural History Museums at Other Universities

Despite proposals over the years to unify data structures for multiple museum collections, most natural history museums today use systems that were developed in house and that are particular to their own needs. Many privately-endowed museums such as the Smithsonian and the Field Museum use commercial software like KE Emu, a popular collection management system that uses a proprietary database characterized by minimal normalization and generalized to include most kinds of specimen data. On the Berkeley campus, the Phoebe Hearst Museum of Anthropology has also turned to a commercial database, Gallery Systems. Image-based collections often look to the art museum community for software such as Cumulus, used by the California Academy of Sciences Image Library for in-house management, or 4D, used by the Fine Arts Museum of San Francisco. However, academic museums often do not have the resources needed for the large initial outlay to digitize collections and the ongoing licensing costs of these commercial systems. Few academic museums have technical people on their staff. Anecdotal evidence depicts a predominance of PC-based databases such as Access, dBASE, and Paradox created and maintained by museum scientists and students. While online access to these collections has become much more common, it is still not the norm for most academic museums. UC Berkeley museums are at the forefront in devoting resources to technical development and are often looked to for technical guidance. The U. Kansas museums have also been a leading player in the community, developing Species Analyst in the early 1990s and collaborating on many joint projects including MIP and DiGiR.

In the last few years, some interesting collaborations have developed among museums that have been made possible in large part by the Library community's Dublin Core metadata standards. The goal of Dublin Core was to develop a minimal set of core metadata elements that all collections have in common. With the advent of internet-based searches, it was becoming clear that the large, complex, and cumbersome protocols and standards of the library community, such as MARC and Z39.50, did not lend themselves well to fast searching on the Internet. In addition, few collections had answered the call in the late 1980s and early 1990s to collaborate on large all-encompassing schema to be used by all museums across disciplines. Dublin Core offered a way for disparate collections to interact with each other without the need to buy in to a particular data model. DC was originally conceived for document collections, so core elements included Title, Author, Publication Date, and so on. The biological community adapted it and created the Darwin Core, and this is the core set of metadata that makes DiGiR possible. DiGiR is a system for supporting distributed queries to the collections of participating museums. In the DiGiR model, each participant makes Darwin Core data available via specialize software called a portal. Nearly all types of relational databases have data export mechanisms that are simple and straightforward, so the Darwin Core model makes it easy and inexpensive for a variety of museums to participate. Users can then issue queries to any of the portals. Search time can be optimized by caching the data in a central repository. Using DiGiR software to access Darwin Core elements at multiple museums, multi-museum research projects are now underway including Manis (mammals), HepNet (amphibians), and Ornis (Birds). The new Geomancer project within MVZ is investigating more automated methods for georeferencing specimen data. Nearly all specimen collections entered the Internet age with little if any machine-readable fine-grained locality data (i.e., a numeric latitude and longitude). Locality information is essential to many types of research in the biological community that use large cross-institutional datasets, so there is a great need to annotate specimen records with this information. In the Paleontological community, PaleoPortal is a web-based resource that allows queries to fossil records at multiple institutions using DiGiR technology. Academic participants that are currently running DiGiR portals for PaloPortal include U. Iowa, U. Kansas, LSU, Texas Tech, U. Colorado, Carnegie-Mellon, Michigan State, UC San Diego, and others, which demonstrates the variety and facility of this system.

Acknowledgements

Many thanks to Effie Dilworth, David Lindberg, Joyce Gross, Pat Holroyd, and Dick Moe for making themselves available for lengthy interviews and providing source material. Thanks to Judy Scotchmoor for material about PaleoPortal. Thanks to Joyce and Effie for reading and editing. Thanks to google for making research so easy.

Author's Disclaimer

There are many more people I would have liked to interview if I'd only had more time, so this paper should not be viewed as either authoritative or complete. Please send corrections and comments to ginger@berkeley.edu