created: x.16.2005
revised iii.3.2006

The California Beetle Project database

This document describes the general purpose, structure, and function of the California Beetle Project database. It is intended as a guide for the project's data entry staff, and as an overview for other potential users of the database.

For instructions on using the on-line database, see the file
http://www.sbcollections.org/cbp/docs/CBP_SimpleQueryInfo.pdf

Purpose of the database

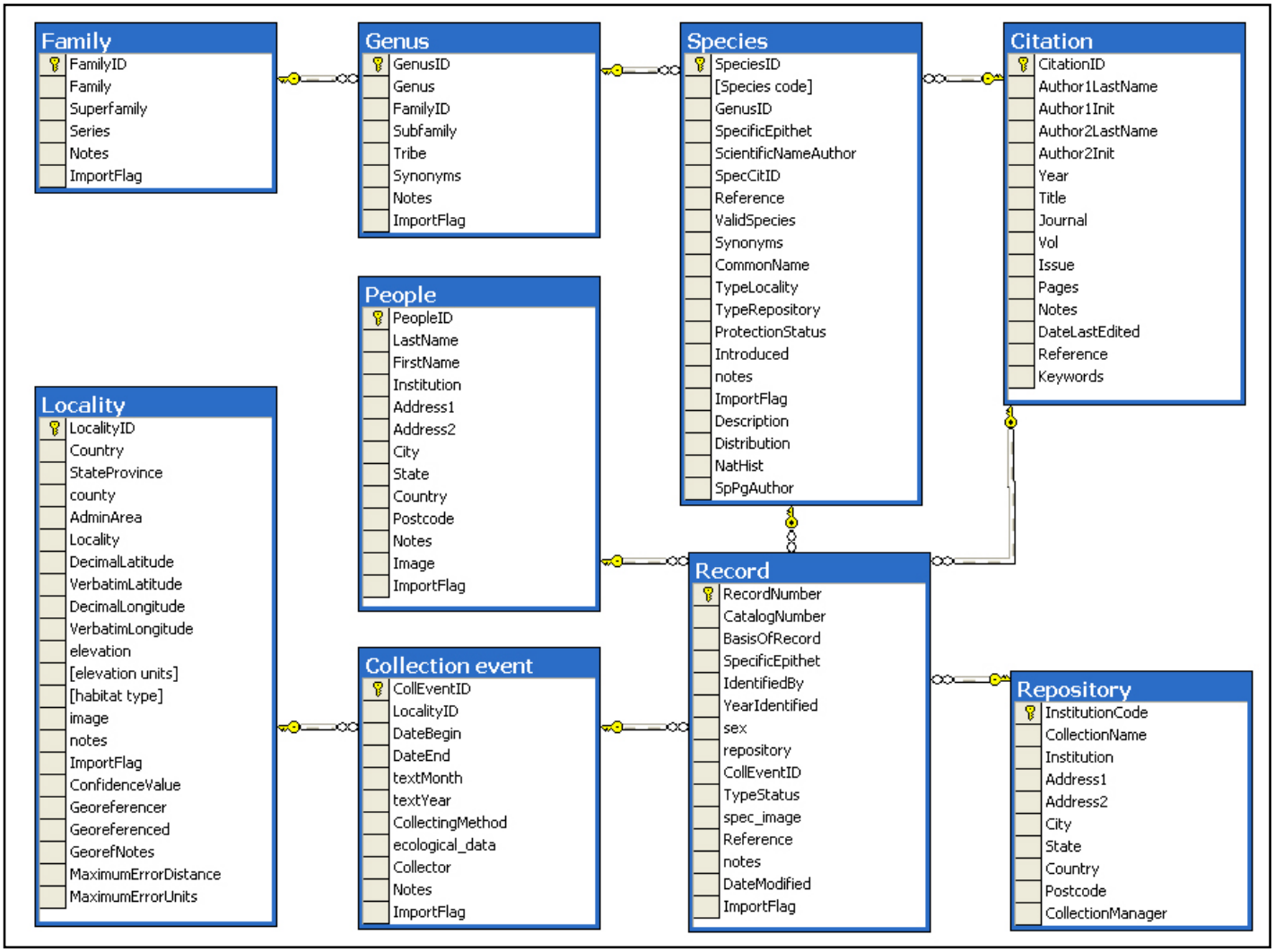
- to house any type of occurrence data pertaining to the California beetle fauna
  - vouchered records
    - catalogued specimen data from SBMNH collections
    - catalogued specimen data from other collections
  - unvouchered records
    - uncatalogued specimen data from other collections
    - occurrence records from entomological literature
    - observational records (sightings) from field studies
- to permit local and remote users to summarize occurrence data according to their needs
  - Area-specific lists of taxa
  - Compilations of distributional records (geographic or temporal) for specified taxa
  - Mapped distributional records for specified taxa
  - Lists of taxa housed in specified repositories
  - Bibliographies for published occurrence of specified taxa in California

STRUCTURE OF THE DATABASE

- **Some terminology**
  - The database is housed in a Microsoft SQL Server database on a remote computer. We view, enter, and edit data through a Microsoft Access interface called **CBP.adp**.
  - The database is relational in structure, meaning that it consists of multiple linked tables, each specialized to contain particular information on entries in other tables.
  - Each 'Table' is composed of several 'Fields'
  - Each entry in any table is referred to generally as a 'record'. (The **Record** table contains records, and the Family, Genus, Species, etc.. tables also contain records.)
  - Each table has a 'Primary Key', which is a unique identifier for each record in the table. Most Primary Keys are hidden numbers to speed indexing, while each table will also contain a unique text field that we see.
  - Links between tables are generally in the form of 'one-to-many' relationships. For example many records in the **Record** table are linked to each record in the **Species** table; many **Species** records are linked to each **Genus** record, etc..

- **General description of structure**
  - The core of the database is the **Record** table. Each record in the **Record** table corresponds to a single occurrence of a species, either a single specimen, literature record, or a sighting, and contains information specific to that particular occurrence.
  - The **Record** table is primarily linked to two other series of tables, one which describes taxonomic information (the **Species**-**Genus-Family** tables), one which describes temporal and geographic information about the occurrence (the **CollectionEvent-Locality** tables). Records are also linked secondarily (not all records have these) to the **Repository, Citation**, and **People** tables, describing where specimens are housed, and what publications have cited those records, and who is responsible for the identifications.

- **Tables and fields**
  - **Record** – all occurrence records of California beetles
    - **RecordNumber** – unique, numeric identifier for each occurrence record.
    - **CatalogNumber** – unique number assigned to specimens in SBMNH or other collections. Not all occurrence records will refer to a catalogued specimen.
      - SBMNH California Beetle Project records are 7 digits preceded by 'CBP'
    - **BasisOfRecord** – Indicates whether occurrence record is based on a 'specimen', a 'literature' record, or a 'sighting'.
    - **SpecificEpithet** – Strictly, the SpeciesID, linking the occurrence record to a single record in the **Species** table, but displayed as a name in all forms.
    - **IdentifiedBy** – Person who determined each specimen or observational record; linked to and equivalent to PeopleID in the **People** table.
    - **YearIdentified** – Year identification was made.
    - **Sex** – sex of the specimen or observed animal, if known.
    - **Repository** – Institution specimen housed, if known.
    - **CollEventID** – Linked and equivalent to CollEventID in **Collection Event** table.
    - **TypeStatus** – Kind of type (holo-, para-, lecto-, neo-, syn-), if specimen was used as basis for a species description.
    - **spec_image** – Filename(s) for linked images, if any.
    - **Reference** – Linked and equivalent to CitationID in **Citation** table, indicating any publications citing specific occurrence record.
    - **Notes** – any additional information pertaining to record.



Tables and relationships in California Beetle Project database.

- **Collection Event** – Specific information on when, how, and who acquired each field sample (each of which may have contained many specimens).
  - **CollEventID** – Unique, hidden numerical identifier for each record.
  - **LocalityID** – Linked and equivalent to LocalityID in **Locality** table.
  - **DateBegin** – Sampling date, or the beginning of a sampling interval.
  - **DateEnd** – End date of a sampling interval, if applicable.
  - **CollectingMethod** – Technical means sample was made (trap type, etc.)
  - **ecological_data** – Any natural history data associated with each sample (host plant association, brief habitat description
  - **Collector** – The person (First initials, last name) of the collector.
  - **Notes** – Any additional information on the circumstances of collection.

- **Locality** – Table containing all geographical information pertaining to occurrence records. (May be linked to many collection event records.)
  - **LocalityID** – Unique, hidden numerical identifier for each record.
  - **Country** – Default USA, but some significant MEX (Baja California) records
  - **StateProvince** – Default CA, but few records from AZ, OR, NV, or BCN.
  - **County** – County name only, without word 'County'.

- **AdminArea** – Any administrative unit containing the collecting locality, abbreviated as follows.
    - NP – National Park
    - NF – National Forest
    - NM – National Monument
    - SP – State Park
- **Locality** – Verbatim description of Locality from label
    - '10 mi W Cuyama', 'Jct. Hwy. 154 and Santa Ynez River', 'Santa Barbara', 'S shore Big Bear Lake', etc.
- **DecimalLatitude** – Decimal expression of latitude: e.g. '35.0139', without minutes or seconds, N or S.
- **VerbatimLatitude** – Non-decimal expression of latitude: 34deg12'46"N, or 34deg12.66'N. Always substitute 'deg' for '°' [degree symbol].
- **DecimalLongitude** – Decimal expression of longitude: e.g. '-119.0139', without minutes or seconds, E or W (West longitudes entered as negative).
- **VerbatimLongitude** - Non-decimal expression of latitude: 119deg12'46"W, or 119deg12.66'W. Always substitute 'deg' for '°' [degree symbol].
- **elevation** – Integer expression of elevation.
- **elevation units** – Units elevation expressed, in 'ft' or 'm'
- **habitat type** – General verbal description of habitat at collecting site.
- **image** – Filename of photograph of site, if available.
- **notes** – Any additional information on collecting site.
- **ConfidenceValue** – Codes of 1-8, corresponding to increasingly large error radii. These correspond largely to those developed by the MapStedi project (http://mapstedi.colorado.edu/).
    - **1** – GPS coordinates obtained electronically at collection.
    - **2** – Amended exact coordinates; coordinates provided with specimen, but found not to be accurate.
    - **3** – Accurate to within 100m; retrospectively assigned, but point described explicitly.
    - **4** – Accurate to within 1km.
    - **5** – Accurate to within 5km.
    - **6** – Accurate to within 10km.
    - **7** – County record only.
    - **8** – State record only.
- **Georeferencer** – Last name of person assigning coordinates to locality
- **Georeferenced** – 'Y' or 'N', to allow extraction of only georeferenced records.
- **GeorefNotes** – Software or other resource used to assign coordinates to locality; plus any other notes on coordinate assignment.
- **MaximumErrorDistance** – Explicitly calculated georeferencing error radius (from software such as Manis's online calculator (http://manisnet.org/gc.html).
- **MaximumErrorUnits** – Units for calculated error radius.

o **Species** – Information on species level names.
- **SpeciesID** – Unique, hidden numerical identifier for each record.
- **GenusID** – Linked and equivalent to GenusID in **Genus** table.
- **SpecificEpithet** – Species name (without genus).
- **ScientificNameAuthor** – Describer of species, in parentheses '()' if species no longer in original genus.
- **SpecCitID** – CitationID for citation of species in California.
- **Reference** – Verbal reference (e.g. Caterino, 2006) for SpecCitID.

- **ValidSpecies** – 'Y' or 'N', indicating whether species name is currently considered valid or invalid, respectively.
- **Synonyms** – Senior synonym, if ValidSpecies=N
- **CommonName** – Common name of species, if any.
- **TypeLocality** – State: Locality, where species was first described from.
- **TypeRepository** – Museum where type specimens are housed, generally following Arnett et al. (1993; also at http://hbs.bishopmuseum.org/codens/codens-r-us.html).
- **ProtectionStatus** – Whether 'FT' (federally threatened), 'FE' (federally endangered), or 'SSC' (California Species of Special Concern).
- **Introduced** – 'Y' or 'N'
- **notes** – Any other information about the species.
- **Description** – Short verbal description of appearance and identifying characteristics of the species, displayed in species pages.
- **Distribution** – Short verbal description of total range of species, displayed in species pages.
- **NatHist** – A short (<500 characters) verbal description of species life history, displayed in species pages.
- **SpPgAuthor** – Author of species page text.

o **Genus** – Information pertaining to genus names, including higher categories between genus and family.
- **GenusID** - Unique, hidden numerical identifier for each record.
- **Genus** – Genus name.
- **FamilyID** – Linked and equivalent to FamilyID in **Family** table.
- **Subfamily** – Subfamily containing genus, if applicable.
- **Tribe** – Tribe containing genus, if applicable.
- **Synonyms** – Senior synonym, if no longer valid.
- **Notes** – Any other information about the genus.

o **Family** – Information pertaining to family names, including higher categories between Family and Order.
- **FamilyID** – Primary key. Unique, hidden numerical identifier for each record.
- **Family** – Family name.
- **Superfamily** – Superfamily containing Family, if applicable.
- **Series** – Series containing Family, if applicable.
- **Notes** – Any other information about the family (alt. spellings, rankings, etc.).

o **Repository**
- **InstitutionCode** – Primary key, unique acronym signifying each collection, following Arnett et al. (online at http://hbs.bishopmuseum.org/codens/codens-r-us.html).
- **CollectionName** – name of Museum if different from Institution
- **Institution** – Name of parent institution
- **Address1**
- **Address2**
- **City**
- **State**
- **Country**
- **Postcode**
- **CollectionManager**

- **People –** maintains data on people identifying or borrowing specimens.
  - **PeopleID** – Unique, hidden numerical identifier for each record.
  - **LastName** – Person's surname.
  - **FirstName** – Person's given name
  - **Institution** – Person's home institution.
  - **Address1**
  - **Address2**
  - **City**
  - **State**
  - **Country**
  - **Postcode**
  - **Notes**
  - **Image** – Image of person.

- **Citation** – manages data on publications associated with occurrence data.
  - **CitationID** – Unique, hidden numerical identifier for each record.
  - **Author1LastName** –
  - **Author1Init** – probably will delete this and next two, lumping all authors in one field.
  - **Author2LastName** –
  - **Author2Init** –
  - **Year** – Year of publication.
  - **Title** – Title of publication.
  - **Journal** – Full journal title.
  - **Vol** – Volume of journal.
  - **Issue** – Issue of journal.
  - **Pages** – Page range.
  - **Keywords** – Taxonomic keywords not in article title.
  - **Notes** – Generally whether reference is in SBMNH library, and whether records have been extracted from it.
  - **DateLastEdited** –
  - **Reference** – An automatic concatenation of several of above fields to form unique informative entry that can be viewed or selected in other forms.

WEB INTERFACE TO THE DATABASE

- Server architecture
  - The website accesses a copy of the CBP SQL database housed on a server outside the Museum's firewall, in the so-called 'DMZ'. This machine is a webserver, and also houses all webpages used to interact with the database.
- Update/backup routines
  - The primary, active database resides on an internal, publicly inaccessible server, and exports a complete copy to the server outside the firewall each evening. So data are accessible to web users within one day of posting or editing.
  - The primary database also creates an additional backup copy of itself nightly. These are archived for one week before overwriting. Tape backups of the internal SQL Server machine are made weekly.
- scripting
  - ASP.NET
    - All database access pages use Microsoft's ASP.net framework.
    - The webserver runs ASP.NET (v2.0) software. This program interprets all pages with '.aspx' extensions as they are called by the internet user, sending queries to the database and integrating the information returned into a relatively standard html template.
  - Other scripting
    - Some complex queries are written in 'C#', a programming language interpretable to the .NET framework.
    - 'Stored Procedures' (written in SQL - Structured Query Language) in the database assemble commonly accessed data elements for various web views, speeding return of data from queries.
    - Minor page elements (pop-ups) use Java scripts.
    - A single Cascading Style Sheet (CSS) is called by all pages to standardize typeface usage across pages.
- Mapping
  - Specimen record mapping uses the 'Berkeleymapper' system, developed by informatics scientists with the UC Berkeley Natural History Museums consortium (http://berkeleymapper.berkeley.edu)
    - when a user chooses a set of records to map, an internal script (in our case, written in C#) writes a tab-delimited text file containing fields we specify to a location on our accessible (DMZ) server.
      - At present we include fields: Record:Catalog Number, Family:Family, Genus:Genus, Species:SpecificEpithet, Locality:Country, Locality:State/Province, Locality:County, Locality:DecimalLatitude, Locality:DecimalLongitude, CollectionEvent:DateBegin.
      - How these fields correspond to DarwinCore fields is detailed in an .xml configuration file also housed on our server.
    - Script in our 'MapResults.aspx' page sends a request to a program running on the berkeleymapper server to retrieve this text file, and map the included records.
    - Our script simultaneously tells berkeleymapper the location of our .xml config file which describes which fields it will be receiving.
    - A new browser window displays the map assembled and served by the berkeleymapper server.