25-305 ReptileDB

Project Proposal

Prepared for

Dr. Peter Uetz

Virginia Commonwealth University

By

Sanie Fragata
Joshua Grove

Nathaniel Grove

Elyse O'Rahilly

Under the supervision of

Dr. John Leonard

October 5th, 2024

Table of Contents

Section A. Problem Statement

The reptile database features decades of accumulating data detailing all known reptile species. It catalogs over 12,000 unique species and includes information like previous classifications, diagnosis, and natural ranges of each. The current raw database is hand-updated by Dr. Peter Uetz personally. It is based in FileMaker so all the information files it produces are in a format that doesn't lend itself well to normalization and relational database querying. Since this resource is the only complete collection of the species definitions of the entire class of Reptiles, it is a necessary tool for biologists all over the world to catalog newly discovered or evolving species. Though public versions of the database exist, none of them are up to date with current database technology. As it currently stands, a lot of biological research on reptiles is done directly from print media like books or by physically visiting specimens in a handful of museum collections all over the world. In a time of worldwide information disbursement, this is a disservice to biologists all over the world who should be able to have a robust and easily accessible database of this information. By creating a relational database with parsed out and easily searchable information, this tool can be hosted online and made available to biologists globally. By relating different bits of data to each other this database will allow for relational queries to study how different reptile species are similar or dissimilar and how evolution has changed them over time.

The goal of the project is to create the database off of Dr. Uetz's files and add a front-end website to host queries of the database. Some such websites already exist based on the current version of Dr. Uetz's database. Websites like the Catalog of Life (Contributors, *Col* 2024) house the information in a public forum and Dr. Uetz hopes to connect the site produced by the reptile database to this site. Last year, a group of senior Computer Science students created a converter to parse out the information from the flat filemaker files, but this is only the first iteration of a living database and there is still plenty of work to be done. The data has been parsed out into columns but is far from normalized, meaning all the information needs to be separated into smaller chunks in order to make it searchable and usable by a querying language. Additionally, the converter takes too long (more than ten hours) to build the database once the file is updated. The current version of the front end (website) is usable but written in a format that doesn't allow for a lot of options as the project evolves and scales up in the future. This front end needs to be rewritten in a format that will allow for integration of other resources such as iNaturalist (Community, *INaturalist* 2024), a website that utilizes the citizen science method for gathering data about biodiversity all over the world. iNaturalist allows citizen scientists the world over to upload their own images of species they encounter and have the images evaluated by biologists to identify the species photographed. As such, the site has a treasure trove of images that can be displayed on the site to give reference to each species entry.

The goal of the project will be to improve upon this previous design and address the most urgent issues including fixing the search function, parsing out the data even further, and reworking the front end from the ground up. Utilizing the agile method of software development, the team hopes to create the next iteration of the converter in stages and then develop a front end that allows for easy access to the database. Agile method exists as an iterative process of development that encourages multiple rounds of prototypes that can be presented to the stakeholders and then added to or adjusted as the process reiterates. It is often cited to exist in two phases, as demonstrated by the “double diamond” diagram in figure 1.

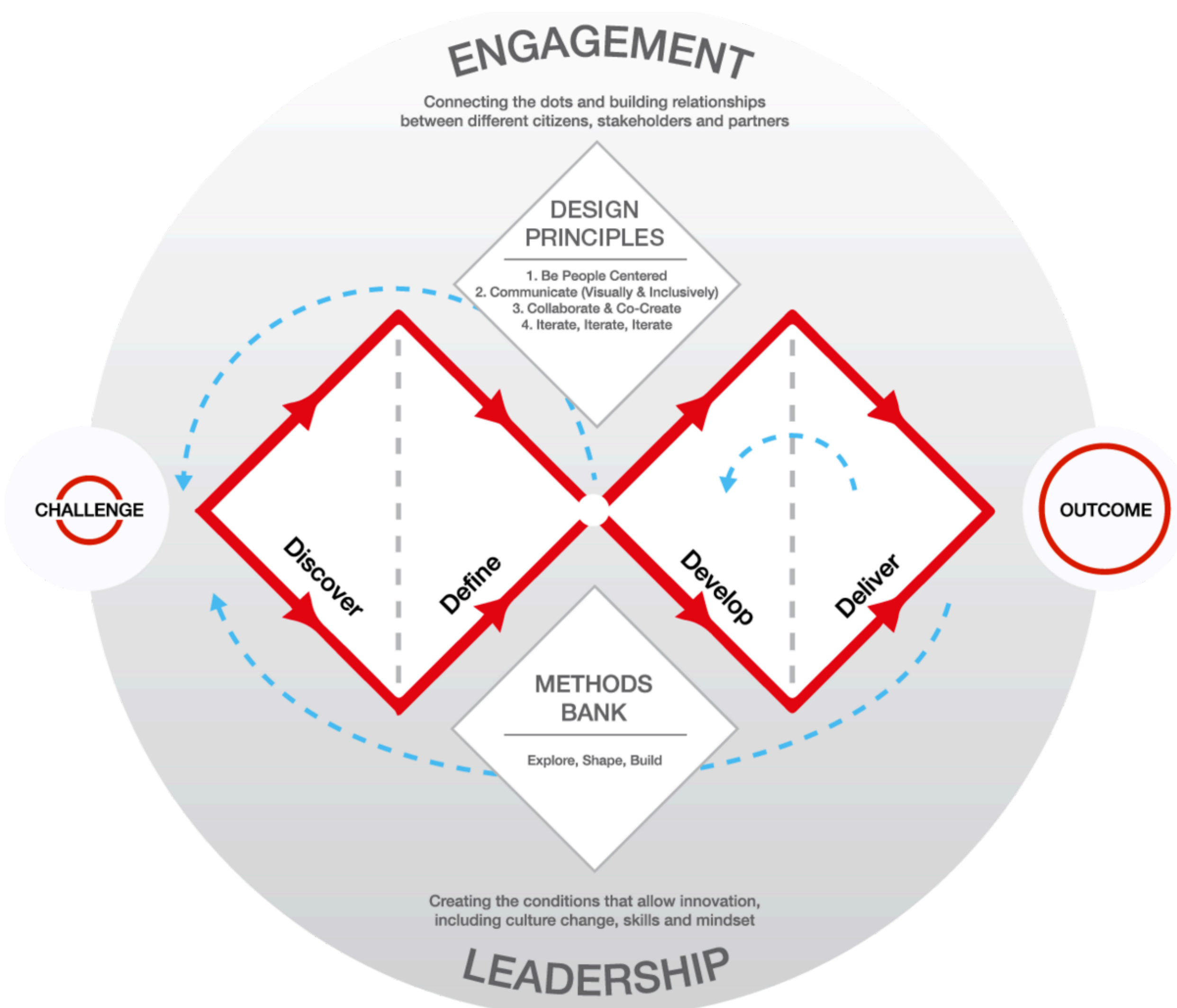


figure 1, British Design Council, 2005

This strategy of constant reiteration has proven effective in software development as a stakeholder’s needs are often changing or their expectations might not line up with the design team’s ideas. The team has outlined the timeline of each iteration of the process in the gantt chart in Appendix 1.

Citations

British Design Council. (2005). *Home*. Design Council - Design for Planet. https://www.designcouncil.org.uk/our-resources/framework-for-innovation/

Community, iNaturalist. (n.d.). *INaturalist*. https://www.inaturalist.org/

Contributors, C. (2024). *Col*. Catalogue of Life. https://www.catalogueoflife.org/

Section B. Engineering Design Requirements

B.1 Project Goals (i.e. Client Needs)

- Redesign the front end of the reptile database to provide a modern, user-friendly interface that is visually appealing and intuitive for users. (Feedback: Focus on responsive design, accessibility, and ease of navigation to improve user engagement.)
- Fix and optimize the search function to return small, concise, structured results. Implement additional features like sortable columns and the ability to download specific entries in common formats (CSV, Excel). (Feedback: Ensure search results are displayed in a user-friendly format, with quick filtering options and export functionality.)
- Allow users to contribute to the database by submitting species sightings, including geographic location data and photos. Provide an easy-to-use form for users to upload files and leave feedback on their submissions. (Feedback: Consider adding a review/approval process for user submissions to maintain data accuracy and integrity.)
- Implement functionality for users to submit suggestions to update existing entries or upload their own images to improve the accuracy and richness of species data. (Feedback: Allow users to suggest updates with clear feedback on their submissions, perhaps enabling a community moderation or verification process.)
- Develop a process to merge user edits on the web platform with the original database, ensuring consistency and accuracy across both sources of data. (Feedback: Automate as much of this as possible, but provide tools for manual review and conflict resolution.)
- Investigate the potential to integrate machine learning technology that can recognize species in submitted images and return appropriate species information. (Feedback: Explore external sources like the iNaturalist or Seek image analysis engines as possible integrations to avoid reinventing the wheel. Machine learning could enhance user experience and data accuracy.)

B.2 Design Objectives

- The database will return search results within 15-20 seconds to ensure a responsive user experience, even with large datasets or complex queries.
- The database will be scalable to accommodate new species, taxonomy updates, or changes in data, ensuring it can grow alongside evolving conservation and research needs.

- Design the database to be accessible through both an API and a web interface, allowing users to search for and retrieve reptile records based on a variety of filters such as species name, habitat, or conservation status.
- The database will feature user-friendly data reporting, enabling users to easily upload additional information for each species, including photos, research papers, geographic distribution data, and other relevant media or documentation.
- The database will include Geographic Information System (GIS) capabilities, allowing users to visualize and analyze species' habitat ranges, supporting research and conservation efforts by mapping critical habitat zones.

B.3 Design Specifications and Constraints

- The website will pull objects from the same database as the previous iteration, ensuring continuity of data while implementing enhancements.
- The database will be modernized using GraphQL APIs and Next.js, allowing for more efficient data querying and faster load times, providing a modern user experience.
- The database will retain the same data model as before, ensuring compatibility with existing tools and processes while allowing for future modifications if needed.
- The database will continue to run on the same server infrastructure until approval is granted for migration to the official VCU server, ensuring minimal disruption during development.
- The database and website will be containerized using Docker, providing a reliable and isolated environment for development, testing, and deployment, while streamlining future updates and scaling.
- Normalization techniques will be applied to reduce data redundancy, with an updated data converter to efficiently migrate and format data from the previous version of the database.

B.4 Project Design Best Practices

Data Privacy and Security:

- **GDPR Compliance** (General Data Protection Regulation, EU): If the database will collect personal information from users (e.g., location data, email addresses), it must comply with GDPR standards, which include obtaining explicit consent, allowing users to request data removal, and ensuring secure data storage.
- **CCPA Compliance** (California Consumer Privacy Act): If U.S. users from California will access or contribute to the database, CCPA regulations should be considered, particularly regarding transparency, opt-out provisions, and user rights to access or delete their data.

- **Encryption**: Secure all user-provided data, such as geographic location or personal information, using encryption in transit (e.g., HTTPS) and at rest (e.g., AES-256 encryption).

Intellectual Property and User Content:

- **Copyright and Licensing**: Ensure that any data or images submitted by users are handled according to copyright law. Provide clear guidelines that inform users that they retain ownership of their submissions but grant the database permission to use, display, or modify that content.

Biodiversity Data Standards:

- **Darwin Core Standards**: To facilitate data sharing and interoperability with other biodiversity databases, the database will adhere to **Darwin Core** standards, which provide a consistent format for sharing species and environmental data globally.

Ethical and Legal Considerations for Conservation Data:

- **CITES (Convention on International Trade in Endangered Species)**: Ensure that the database complies with CITES regulations regarding endangered species. This includes making users aware of species' protected status and ensuring that the database doesn't inadvertently facilitate illegal trade in reptiles.
- **Nagoya Protocol (Access and Benefit-sharing)**: If the database collects or shares data on genetic resources or species from indigenous lands, it must adhere to the **Nagoya Protocol**, ensuring that indigenous communities have provided consent for the use of their biodiversity data and that they benefit from its utilization.

GIS and Location Data Compliance:

- **Data Accuracy**: Ensure geographic location data provided by users is accurate and updated to reflect the latest information on species distribution.
- **Sensitive Data Masking**: If the database stores sensitive location data (e.g., precise habitats of endangered species), consider using masking or other methods to protect species from poaching or habitat disruption. Restrict access to exact location data unless users demonstrate legitimate research or conservation purposes.

Accessibility and Inclusion:

- **WCAG 2.1 (Web Content Accessibility Guidelines)**: Ensure that the web interface of the reptile database meets **WCAG 2.1** standards for accessibility, making it usable by people with disabilities, including screen reader compatibility, color contrast, and keyboard navigation.

Machine Learning and AI Compliance:

- **Ethical AI Use**: If implementing machine learning for species identification, ensures the algorithms are trained using diverse, unbiased datasets to prevent inaccuracies or misidentifications, especially for species in underrepresented regions.
- **Explainability in AI**: Provide transparency around how machine learning algorithms work and allow users to question or verify the results of species recognition models.

Section C. Scope of Work

C.1 Deliverables

The list of major expected deliverables are:

1. The Project Proposal itself.
2. The Fall Design Poster, which will highlight the results of the design, as well as showcase the background, criteria, analysis methods, and improvements of the new design.
3. The Preliminary Design Report, which will showcase the design process for the improved database, as well as showcasing the work the team has accomplished thus far, as well as the methods by which that progress was achieved.
4. Docker container framework.
5. A website that will serve as the new front-end of the reptile database, equipped with Next.js and more advanced structure.
6. A new iteration of the converter that produces a more normalized database
7. A streamlined search function.

Possible obstacles include learning how to implement the database properly, as well as potential issues with hardware.

C.2 Milestones

1. Completion of docker container framework for testing and implementation.
2. First iteration of converter
3. First iteration of website
4. Second iteration of converter
5. Second iteration of converter (Additional iterations of both the website and the converter will continue following the AGILE framework).
6. A working prototype for the primary database functions, concurrent with a completed analysis of the previously developed features.
7. A proof of concept front-end to be integrated with the database when feasible.
8. Partial Data integration with the new design.

9. Finalized Database Prototype.

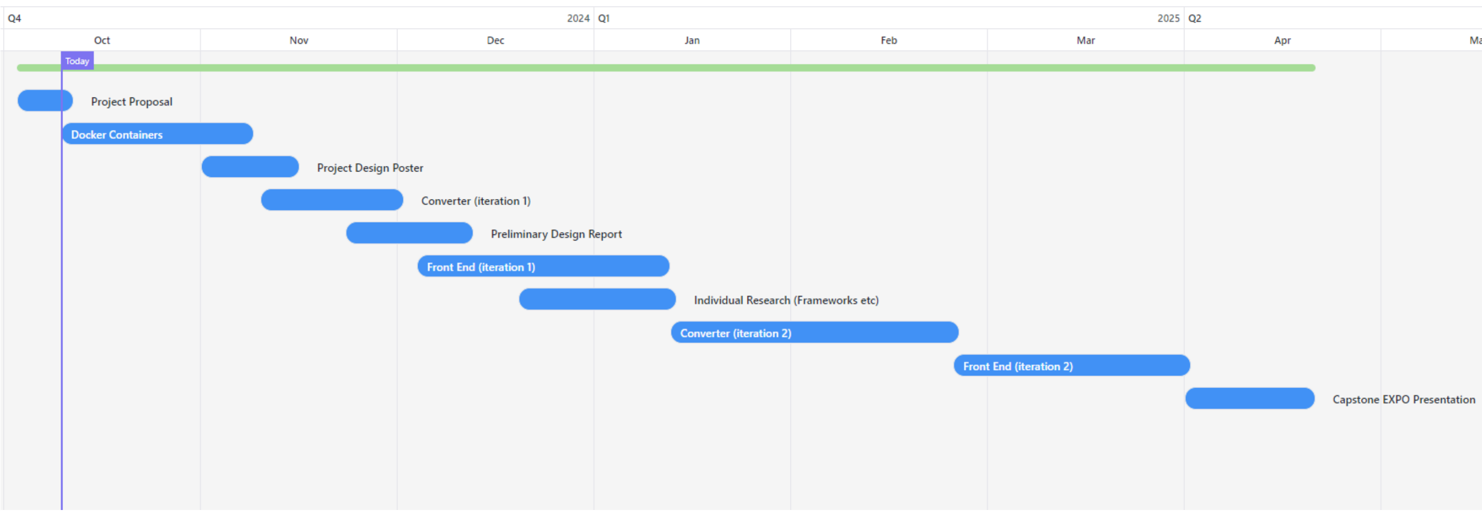
10. Finalized Product Completion.

C.3 Resources

1. A docker framework system for automating interactions between component software
2. Software to design a front-end component that reflects modern industry standards.
3. Either an overhaul of the previous converter or a redesigned iteration.
4. Access to previously developed software to assist in the redesign efforts.

Appendix

Project Timeline



Team Contract

W 25-305 ReptileDB Project Proposal.docx