



# HelixAID Project Document

08/07/2024

---

Michael Gage

UMaine Graduate School

DIG 510

## Project Description

HelixAID is a metadata system designed to organize genetic autoimmune disease research data. This tool is intended to assist in the research and production of new autoimmune disease discoveries and treatment regimens. There is limited knowledge on the genetic origin of these diseases, and the development of HelixAID provides researchers with an aggregation of pertinent, domain-specific data. This data is selected from prominent sources and standardized to ensure an organized and reliable query environment. Utilizing a comprehensive schema that is tailored to genetic research data, HelixAID applies a combination of leading metadata standards to ensure query output relevancy. Users may interact with the accessible search interface or directly survey data via robust SQL queries. With the intent of an ever-increasing data pool, HelixAID also provides institutions with a data input interface to enrich the resources available to the medical research network.

## Stakeholders

- Academic Institutions – Organizations requesting data relevant to their own research, bolstering educational resources, or contributing to the data pool within the system.
- Research Laboratories – Laboratory and research professionals utilizing/contributing to the system to support ongoing autoimmune disease research and discoveries.
- Biotechnology/Pharmaceutical Corporations – Businesses interested in supporting their research in the development of autoimmune disease treatments, devices, and pharmacological developments.
- Patient Advocacy Groups – Organizations that represent patients affected with autoimmune diseases and support research efforts in the domain.
- System Administrators – Those primarily responsible for maintenance of the data accumulated in the HelixAID system.

## Access

Role-based controls assign various users unique access depending on the scope of their permissions and data relevance to them.

A web-based administrative dashboard is available to enable system monitoring, data management, creation, updates, and deletions. This administration interface requires strict authentication and validation procedures, and will allow admin users to access security settings within the system.


End-users will have access to a secure login web interface and the ability to query data and associated files within the system. These users may include researchers, students, clinicians, etc.

Contributors from institutions that are pre-approved can access a data submission interface through a secure web portal. Data input follows structured schema-specific forms and submission status can be tracked within the portal.

HelixAID may be accessed programmatically for machine clients via API for automated data collection. Unique API keys will be assigned to each client to ensure proper authentication and role-specific access.

## Rights

Administrators will have access to the full system, including the ability to delete, produce, or edit user accounts. This right allows administrators to manage access permissions and role-specific parameters for various user types. Data submissions from contributors can be accepted or rejected by approved administrators, as well as data retrieved by the system from a variety of database sources. This includes oversight of API permissions and functionality.



Contributing institutions have the right to submit any domain-relevant data they deem necessary to the development of the systems data pool. They retain the rights to the ownership of submitted data, and may redact prior submissions into the HelixAID system. Contributors may access submission statuses and updates in the approval process.

End users have access to any data that is queried (via search interface or SQL) within the HelixAID system, assuming this data is publically available (deemed by the institution of ownership). Associated files may also be downloaded or saved within the system for future access.


All users may edit account information or attributes specific to their related roles within the system.

## Responsibilities

Administrators are responsible for placing users into the correct roles within the system, ensuring that they have the proper permissions and access. They must warrant the integrity, safety, and security of the system and user data. Administrators are under obligation to oversee technical issues and the availability of user support on an ongoing basis. Admin will set and update schema standards appropriately. Approval of data submissions and API functionality/retrieval are also essential responsibilities of the administrator.

Contributing institutions must ensure that data submissions follow the strict HelixAID metadata schema standards. Approved contributions must be continually reviewed and updated to adhere to accuracy, relevancy, and error-correction expectations. It is the contributors responsibility to ensure that redacted papers and research are removed from the system or flagged appropriately. Institutions must maintain updated account information and ensure that their data submissions follow strict legal, institutional, and federal guidelines.

End users must follow all terms of service regarding data use, property, and privacy. This



includes legal, institutional, and federal guidelines for the use of data. Users will update their account information appropriately to ensure proper role-based permissions and access assigned by administrators. Users must maintain confidentiality of their account information.

## Constraints

The HelixAID system is dependent upon a few external databases to facilitate the appropriate outputs for end users. These databases follow schemas and metadata standards utilized in the design of HelixAID, but individual data sources will inevitably have limited schematic elements. The variability in database sourcing requires tailored data-retrieval solutions. These constraints are directly linked to the development of the HelixAID API data retrieval system. Due to the comprehensive schema standards within the system, partial data matching and required field standards (i.e relation to autoimmune disease) allow for the addition of relevant data.

The contributor feedback loop is dependent on institutional submissions to bolster the data pool accessed through HelixAID. With limited research on genetic precursors to autoimmune diseases, it is essential that new research is integrated and organized effectively within the HelixAID system. This requires accuracy, schema-specific research elements, and scalable data storage capabilities. Ultimately, the development of HelixAID functions as a service to aggregate domain specific data related to genetic autoimmune diseases.

Intellectual property (IP) of genetic data and research documents are under ownership of the submitting institution. Data will only be retrieved from source databases if it is considered “open” or public to the research community. Institutions are expected to exclude or de-identify protected aspects of their research when contributing to the HelixAID system. End users who encounter query results with strict IP permissions may access the source data from the institution that holds these rights, according to that institution's requirements. In these cases, only the public metadata and study DOI will be accessible to the user through the HelixAID system. This constraint limits user's direct access to protected data within the HelixAID system, however, users may access the DOI

link to the desired research data listed in the query output and request access directly from the institutional source.

## Comparable Projects

**ADEX:** <https://adex.genyo.es/>

The Autoimmune Disease Explorer from the Genyo Bioinformatics Unit provides gene expression and methylation data from a collection of autoimmune disease studies. This data is retrieved exclusively from the NCBI GEO database and contains 50 dataset entries.

**ImmPort:** <https://www.immport.org/home>

Funded by the NIH, NIAID, and DAIT, ImmPort is a database for bioinformaticians focused on data sharing in the broader field of Immunology. ImmPort covers genetic data related to autoimmune diseases, but is expanded to include other immune-related disease data. This platform supports role based users identified as Administrators, Submitters, or Viewers, similar to the HelixAID structure.

There are few comparable projects to HelixAID, and those related are either limited in scope or expand to a far more general data pool. Those aforementioned are the closest in resemblance to HelixAID based on what systems are available today.

## Resources Required

### Human

#### Development

- Developer time to design the framework of the system
- Developer time to structure the schema and contribution requirements
- Developer time to construct the API data retrieval mechanisms
- Developer time to design the query optimization tools

- Administrative time to collect data for user role development
- Data science time to develop algorithms for data processing and quality

### Ongoing

- Administrative time to update user permissions and roles
- Administrative time to perform routine system maintenance and quality control
- Administrative time to oversee data contribution quality and schema approval
- Contributor time to actively input AID genetic research data
- Data science time to update and maintain algorithms and API functionality
- IT support time to troubleshoot user interactions with the system

## Computing

The HelixAID platform primary and web servers contain multi core processors to manage several queries, processes, API requests and traffic simultaneously. High capacity RAM is necessary to accommodate data and user interactions, as well as to support platform performance. MySQL software is utilized as an ideal database management system when integrating the relational model used in HelixAID. R computing is utilized for backend development in regards to APIs and their interactions with source databases and MySQL. R packages like “shiny” are utilized for web development and user interface design.

## Data

Metadata will be extracted from three primary databases to develop a sufficient pool of information for end users. These three public sources are: GenBank, GEO (Gene Expression Omnibus), and dbGaP (Database of Genotypes and Phenotypes). These data sources have predefined frameworks for metadata, including the MIAME, MlxS, and BioSample standards. These three standards have been combined in the development of a new HelixAID standard that determines what pertinent metadata is retrieved from source databases. This includes:

- **General Information:** Source, Title, Gene Function, Gene Location, DOI/URL, Institution
- **Experiment Information:** Platform, Type

- **Phenotypic/Genotypic Information:** Phenotype, Genotype, Autoimmune Diagnosis/Disease
- **Demographic Information:** Age, Sex, Ethnicity
- **Study and Sample Information:** Design, Sample Source, Data File Type, Dates, Description

## Workflow

### Ingestion, Manual and Automated

Ingestion of data into the workflow can be categorized into two primary methods: API ingestion (automated) and user contribution (manual).

HelixAID utilizes API's to automatically ingest data from three primary sources: GenBank, GEO, and dbGaP. With the use of R programming language, API's can be accessed to send automated requests to database sources and retrieve pertinent metadata. This data harvesting method is scheduled to ensure that a constant influx of data is ingested, leading to new information from these three sources. This information is then normalized according to HelixAID's schema and structured within the relational database for end user queries.

Manual data ingestion is primarily enacted by contributing institutions and their research teams. A web-based interface is available to institutional users with applicable roles. This interface/web form is structured according to the HelixAID system schema and is overseen by administrators to ensure that the data input contains the correct elements. After inspection and approval by administrators, this data is ingested and stored for future access by end users.



## Integration

HelixAID integrates elements of MIAME, MlxS, and BioSample standards in the development of a one-size-fits-all schema. These standards are primarily used in the genetic research domain, and all are integrated in some capacity by GenBank, GEO, and dbGaP into their own systems. Due to the consistency of these shared standards across source databases, cross-database integration within the HelixAID system is possible. This allows a query to generate related autoimmune disease data from multiple sources into a single output.

Although the metadata and schema elements themselves are integrated from external sources, the HelixAID system does exist independently and can function solely on manually contributed data. This system pulls metadata and genetic data files unidirectionally from database sources, and thus, is not currently integrated with other external systems bidirectionally.

## Quality Control

Quality control of data and metadata stored in the HelixAID system is the primary responsibility of administrative users with associated roles. Admin must ensure that all contributing data submitted by institutions fits within the schema structure of the system and includes essential metadata. They must also ensure that the retrieving API is pulling data properly and in the correct format for ingestion into the system.

## Storage

Although journal articles are not stored in the HelixAID system, their associated metadata and genetic data files do require external storage. This is essential, as the data pool regarding genetic autoimmune diseases is expected to grow exponentially. This data is stored in a cloud-based system that allows for continued expansion (Amazon S3). Though a cloud based system may not be necessary in the initial development of the system, the expansible nature of these storage systems may be necessary as genetic autoimmune



disease research data increases.

External storage is also necessary if end users are expected to access and download relevant genetic data files directly on the HelixAID platform. Once data is queried in HelixAID by a user, the system utilizes metadata tags assigned to the selected file to locate it in Amazon S3. The selected data file can then be downloaded directly in the HelixAID platform through seamless integration with Amazon S3.

## Access

Access workflows within the HelixAID system are only applicable to the administrative users. This workflow includes access specification among the three roles within the system: Administrator, Contributor, and End User. Administrators will assign access parameters to each role, depending on the intended use of these user accounts.

## Technical Description

### Scope

The primary scope of HelixAID is to aggregate metadata and associated data files from genetic research related to human autoimmune diseases. Research in this domain is limited, yet scattered widely among existing database systems. There is a significant need to assemble this data into a cohesive system that will benefit researchers and expedite access to such resources. In addition to metadata and genetic data file aggregation from external sources, HelixAID is designed for data/metadata input – an opportunity for contributors to share and store new research in a unified system designed for genetic autoimmune diseases. Input and data query is available through role-specific user search interfaces and the option to utilize SQL for advanced queries.

The scope of the collected metadata includes essential elements from widely used bioinformatics and genetic schema standards: MIAME, MIxS, and BioSample. These standards are highly integrated into the source databases (GenBank, GEO, dbGaP) and allow for transitional ease of research metadata and files into the HelixAID system.

## Database System and Query Language

HelixAID will use a relational database management system (RDBMS), a framework that is optimal for structured data with compound relationships. This model offers advantages when discussing schema flexibility, data integrity, and scalability. Many databases that contain genetic data follow a RDBMS system. Following this system will allow for ease regarding data ingestion and output. The relational model is effective when performing SQL queries and provides a structured framework that compliments the HelixAID schema. The scalability of the relational model is essential when storing and accessing large genetic data files and metadata.

The query language used in HelixAID and many other bioinformatics/genetics data platforms is the Structured Query Language (SQL). This language is widely adopted in the field, and is familiar to data scientists, bioinformaticians, geneticists, and researchers. With the complexities of genetic research data, the use of SQL provides advanced query operations to researchers or machine clients needing programmatic access to data. This language also provides administrator users with a robust toolkit for inserting, updating, and deleting data in the HelixAID system.

## Schema(s) Used

The HelixAID schema combines essential elements of the MIAME, MlxS, and BioSample standards commonly used in research today. These standards are tailored to genetic data, and the integration of these standards in HelixAID allows for a standard organization that is familiar to researchers and maintains universal attributes. Schema elements are hand picked for the organization of genetic autoimmune disease data and are included in the requirements for institutional contributors when submitting new data/metadata into the system.

Data Type	Namespaces Used
General	MIAME: <a href="https://www.fged.org/projects/miame">https://www.fged.org/projects/miame</a> MIxS: <a href="https://www.genesc.org/pages/standards-intro.html">https://www.genesc.org/pages/standards-intro.html</a> BioSample: <a href="https://www.ncbi.nlm.nih.gov/biosample/docs/attributes/">https://www.ncbi.nlm.nih.gov/biosample/docs/attributes/</a>
Experiment	MIAME: <a href="https://www.fged.org/projects/miame">https://www.fged.org/projects/miame</a>
Phenotypic and Genotypic	MIAME: <a href="https://www.fged.org/projects/miame">https://www.fged.org/projects/miame</a> MIxS: <a href="https://www.genesc.org/pages/standards-intro.html">https://www.genesc.org/pages/standards-intro.html</a> BioSample: <a href="https://www.ncbi.nlm.nih.gov/biosample/docs/attributes/">https://www.ncbi.nlm.nih.gov/biosample/docs/attributes/</a>
Demographic	BioSample: <a href="https://www.ncbi.nlm.nih.gov/biosample/docs/attributes/">https://www.ncbi.nlm.nih.gov/biosample/docs/attributes/</a>
Study and Sample	MIAME: <a href="https://www.fged.org/projects/miame">https://www.fged.org/projects/miame</a> MIxS: <a href="https://www.genesc.org/pages/standards-intro.html">https://www.genesc.org/pages/standards-intro.html</a> BioSample: <a href="https://www.ncbi.nlm.nih.gov/biosample/docs/attributes/">https://www.ncbi.nlm.nih.gov/biosample/docs/attributes/</a>

## Expression

After a query is performed, the expressed metadata/data files will be generated in a tabular human-readable format. A tabular format is easily exported and analyzed in software like R, Python, or Excel. Each row of the table will contain a distinct study, experiment, or dataset relative to the search criteria. Key metadata elements (titles, genetic markers, source links, date, accession) and associated downloadable files will be generated in columns. This allows for easy access to the expressed material and the opportunity for end users to download metadata/data files associated with it. Within the HelixAID interface, a detailed view may be selected to uncover additional metadata elements related to the study of interest.

HelixAID also provides an API output with machine readable formats (JSON, XML) for programmatic access. This output is effective for bioinformatics pipelines and data retrieval from various data analysis platforms.

## Derivatives

HelixAID does not produce derivative data and is designed as a system for research metadata aggregation. This data exists prior to system development in source databases, and it is not generated by the system itself. Any data that exists within the system is input (contributions, account information, etc) or fetched (API retrieval of metadata) but not developed algorithmically within HelixAID.

## Long-Term Management

The long-term management of HelixAID primarily incorporates data management and integrity. Administrative users are responsible for the continuous auditing of incoming data, ensuring accuracy, consistency, and relevance to the system. This also includes the active updating of role-based access among users.

As HelixAID becomes more widely used among institutions and researchers, it is anticipated that new genetic autoimmune disease research will be contributed to continually increase available data. As new data is introduced, the implementation of enhanced tools within the platform may be necessary. This may include AI-oriented processes and visualization tools for users.

## Interfaces

### Human

The human interface within HelixAID is web-based and can be apportioned into three distinct facets for utilization by the defined roles within the system. These facets include end user, administrative, and contributor interfaces. Due to the explicit role accesses within the system, it is necessary to provide various interface options for optimal, appropriate use.

## End Users

This interface is the primary facet for general use of the system. It includes a search dashboard and advanced SQL query-based interactions. User navigation is limited to account management, querying, query history, and access to general system information. Most roles within the system fall into the “end user” category, as the primary function of HelixAID is to provide a query platform for research metadata and genetic files.

## Administration

With full access to the system, the administrative interface is necessary for the continued management of HelixAID. This facet enables system monitoring, data management, creation, updates, deletions, and access controls. After the submission of data by contributors, the admin interface allows for review of submissions for approval.

## Contributors

In addition to attributes included in the end user interface, the contributor interface allows institutions to submit new research data through web forms containing schema-specific protocols. Contributors can browse existing data and track the status of prior submissions into the system. This allows for direct communication between users and admin in the approval process.

## Machine

The machine interface enables programmatic access into and out of HelixAID. The use of API allows for machine clients to retrieve data in readable formats (JSON, XML). Unique API identifiers are assigned to machine clients within the platform, ensuring that all interactions are approved and secure. API interfaces are also utilized for data ingestion, providing the system with automated data collection from external databases. API algorithms specify and digest incoming data to properly adhere to the HelixAID schema.

## Networks

HelixAID utilizes the network interface in the communication and transfer of data with Amazon S3. This intermediary is essential for the proper storage of large genetic data files, metadata, and access to this data upon user query. Using HTTPS, API requests are sent via the network interface to store data with Amazon S3.

## Evaluation

Success of the HelixAID system will be based on the following criteria:

- Data is reliable and adheres to the predetermined schema standards.
- Data is manageable, available, and relevant to the end user query.
- The interface facets are user-friendly and promote the effective use of the system.
- Automated data interactions function smoothly via proper API integration.
- Data protections and compliance are strictly followed by admin and users.
- The system maintains adequacy for expansion and continued use.
- The system ultimately promotes continued development and improvement of genetic autoimmune disease research endeavors.