**Data Preservation Case Study: The ALIEnS Project**

**Jake Carlson, Purdue University,** [jakecarlson@purdue.edu](mailto:jakecarlson@purdue.edu)

**Lisa Zilinski, Purdue University,** [lzilins@purdue.edu](mailto:lzilins@purdue.edu)

*Background*

The Advanced Life-Support for Inhospitable Environments in Space (ALIEnS) was an interdisciplinary research center headquartered at Innovation Park at State Tech University. The purpose of ALIEnS was to investigate the best way to enable humans to live in the extreme environments of outer space for extended periods of time without the ability to resupply life-supporting essentials like oxygen, water, or food. In such an environment, all resources would have to be rationed carefully and recycled wherever possible. The results of this research would be used to create a closed looped system that could control and account for all factors needed to support human life. The application of this system would be capable of supporting the development of permanent human manned outposts on the moon or other planets. The ALIEnS project was comprised of researchers from agriculture, chemistry, civil engineering, food sciences, electrical computer engineering, mechanical engineering and other fields. These researchers and their graduate students worked together to design different facets of the closed looped system with coordination from the central administrative team. Although headquartered at State Tech University, the ALIEnS project was a collaboration that included researchers from two other universities.

The ALIEnS project was funded by NASA and was expected to last for five years. However, the project has fallen victim to the budget cuts during the economic downturn and funding was withdrawn after three and a half years. This unexpected loss of funding caught ALIEnS unprepared. They were just beginning to bring the components of the project together and to test their closed loop model. The leadership of ALIEnS believes that funding could be restored in the future, and wants to reduce the risk of losing the results of the work that had already been completed. They are interested in preserving their data, documents, and other supporting materials so that when funding is restored they could easily resume their work, potentially with a new set of researchers and graduate students.

They have three core needs for the short-term preservation of their materials. First, the administration of ALIEnS needs to have control over access to the materials due to their confidential nature. However, these materials would still need to be accessible by current ALIEnS researchers. They also want to promote some of the published material about the project (such as the annual reports) and therefore this information would need to be available to the public. Second, in addition to their core materials, the data and documents that were interregnal to developing the closed loop system model, other materials may be needed for description or explanatory purposes so that the core materials, processes and decisions that led to their creation can be fully understood by others. Third, as the project was comprised of multiple areas of research, the connections between materials and how they relate to the project as a whole will need to be readily apparent to the future researchers who will take up this project.

The ALIEnS project was comprised of five focus areas, each one generating its own set of core materials. The five focus areas are: Biomass & Crop Production, Food Safety & Processing, Resource Recovery, Systems Analysis & Integration, and Education & Outreach. Each focus area was in turn comprised of a number of individual research efforts. A sixth source of materials is the administration of the ALIEnS project.

The data generated in these focus areas include:

* Sensor readings – These included measurements of light production, soil moisture, carbon dioxide composition and other factors in the growing conditions of plants and edible fungi, determining oxygen loss and other aspects of maintaining a sustainable environment. Sensor output is in a proprietary format and these files can only be accessed using the software that was designed for the sensor. Data was captured on a daily cycle and so hundreds of files were produced.
* Plant and Fungi Samples – Samples from the plants and fungi are harvested at several stages in their development for measurement and testing purposes to determine the overall health of the plant and fungi. Some of the samples are destroyed as a part of the testing and measurement process, others are stored in a freezer in the lab.
* CSV files of sensor data - The sensor data was exported to csv as it is an open format and easier to use for data analyses purposes. However, some information contained in the sensor files is lost as a result of the transfer.
* CSV files of samples data – The results of the testing and measurement of plants were entered into a csv file. Typically, instances of data collection of the plant and fungi samples are captured on a tab within the csv file.
* SAS files – Data from the csv files are imported into SAS for statistical analysis purposes.
* Visualizations – Tables and graphs are produced from the analyzed data using the capabilities in SAS.
* Presentations and other educational materials – These include power point slides, posters, and lesson plans.
* Published papers and research presentations – These are in pdf format.
* Administrative materials – These include annual reports (pdfs), emails, and other internal communications amongst project partners and stakeholders.

ALIEnS did not expect to lose their funding, so they did not engage in any sort of preservation planning in their project. ALIEnS used a content management program called QuickSpace to store, manage and share their materials amongst project personnel. Materials are organized in a hierarchical folder structure. Although they used a common and centralized content management system, project personnel did not follow uniform practices of naming files, or providing descriptive information about their content and context. Their collection of materials includes multiple iterations and drafts of the important outputs of the project.

The principal investigator and the project manager of ALIEnS have approached the libraries for assistance. Ideally, they would like to use the Libraries’ institutional repository to create a short-term preservation system for their materials. They recognize that resources will likely be required to make this work and have some remaining funding available that could be applied to developing a preservation system. The project manager has stated that this is a priority for him and that he has time to help the libraries sort through their materials; he just needs direction on what to do.

**Discussion Questions**

1. Types of data
   1. What types of data are being collected for this study?
   2. How could the researchers encourage all research staff to use the same practices with their data and data definitions?
   3. Preparing for preservation, how could you plan to preserve this data? What would this plan need to address and account for? What information would you need in order to populate this plan?
2. Contextual details
   1. Given the situation, how would you advise the project administration team on future projects?
   2. What impact would naming conventions, standard file formats, and metadata have had on a data inventory process?
   3. Given the needs of the project administration, what recommendations would you make and what actions would you take to help them achieve their goals?
   4. What contextual details need to be captured to make the data meaningful to others? How could they be connected to the data?
   5. How could you acquire contextual information?
3. Data storage, backup, security
   1. How would you determine the functional requirements of preserving this data?
   2. What kinds of capabilities should the technology provide? How would you determine what resources you need?
   3. How could researchers who manage data security across the project’s focus areas account for the different access and preservation needs? How could they ensure compliance?
   4. Where and on what media should the data from each data source be stored?
   5. How could the researchers have prepared for their need for continued and ongoing access to their data files?
4. Data protection/privacy
   1. Who owns the project’s data? What are the rights you would need to preserve this data? How would you go about getting the rights to preserve the data?
   2. What issues might arise related to copyright or intellectual property rights associated with the data? How could they be addressed?
   3. What issues might arise related to restricted research being conducted? How could they be addressed?

1. Policies for access, sharing and reuse of data
   1. Could the project’s data be made available in an Open Access Repository? Why or why not?
   2. What aspects of the data cannot be shared? Why?
   3. What process should be followed to gain future access to the data?
   4. What are potential barriers to sharing the data?
   5. What restrictions would need to be placed on re-use of the data and why? Are there any reasons not to share or reuse data in this situation?
   6. How might the data be restricted to be reused only for certain purposes or by specific people?
2. Archiving and preservation
   1. What content will need to be included in an archive? How could you develop the selection criteria?
   2. What contextual data and information would be included in the archive? How would you go about getting this data and information?
   3. How long should the data be kept beyond the life of the project? How long should the data be kept after the end of the initial 5-year short-term preservation commitment? What criteria would you use to make these decisions?
   4. How would you determine where the data could be archived?
   5. Who should be responsible for and involved in archiving and preserving the data?