Nuffield Research 2016
Lunar Mission One

Designing a Public Archive Data Structure with Wikipedia as Database

Ruihua Zhang
Mentor: Gerald Shields, J.D., LL.M.

Contents

Chapter 1

Introduction

Lunar Mission One is a public-funded project that aims to do inclusive space exploration that belongs to everyone. Besides the plan to send an international robotic lander to the south pole of the Moon, the project also plans to leave a record of all life on Earth down there, a lasting record of human existence that will endure for millions of years. There are two types of archives to be sent: public or private. This project would focus on the data structure for the public archive, which is designed to engage the public in space exploration, from collecting most representative data of human beings to engaging children into making their own humanity record, which links to education and our everyday life.

A data model is an abstract model that organizes elements of data and standardizes how they relate to one another and to properties of the real world. For both compilation and reading, the archive will need some form of data structure. The database for public archive in particular is likely to be very complex. However, as the largest encyclopedia online, Wikipedia might be helpful for forming the basis for the public archive. With the aid of Wikipedia, how could the data be organized and inclusive? How can we organise and manage a multi-layered approach to information quality - from highly qualified and officially approved material, to children's contributions with only teacher approval? This project sets these as the objectives.

Chapter 2

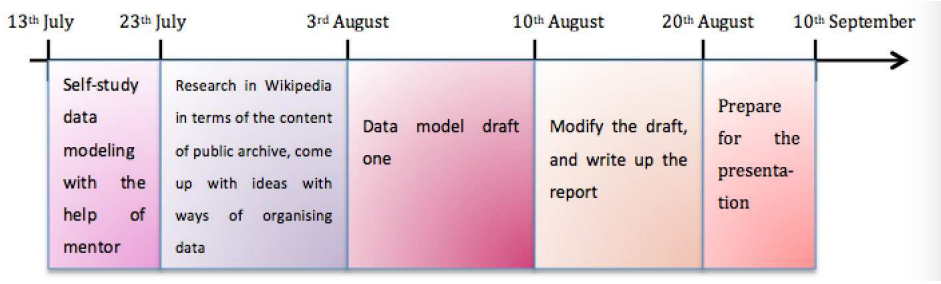
Project Plan

2.1 Approach

In order to come up with a suitable data structure for public archive, first of all, I need to investigate what a data model. Since data model is not a very familiar area for high school students, I need to self-study it by researching online and asking for help from my mentor when needed. Secondly, I need to investigate what the Lunar Mission 1 archive should contain by researching in Wikipedia and discussing with my mentor, in doing so a multi-layered data structure could be designed. For example, a public archive should contain human history and civilization, plus the science of life on Earth with a database of species. What is more, since this project is enormous, I may need to prioritise my work—I need to select an area of database that I can do in the time period given. Thirdly, I need to combine the knowledge of data modeling and the content of LM1 Public Archive to design a suitable data structure. I will need to make sure the structure is applicable by discussing it with my mentor.

2.2 Plan

According to the suggestion from David Iron, the founder of the Lunar Project One, I decided that my process of the project would be divided into five parts, as shown below:



The time is allocated in this way since the learning and designing process would be the most time-consuming parts of this project. I would carry out the full process by myself.

2.3 Resources

Gerald Shields, a self-taught expert in Wikipedia, provided me with key information about Wikipedia as my mentor. As the database of this project, I decided to use Wikipedia (the English version) as a resource where most of my resources came from. Further references which helped me learn the concepts of data structure and data modeling would be mentioned in the Bibliography section at the end of this report.

Chapter 3

Background Research

3.1 Data Modeling

Data modeling is the process of creating a data model for an information system by applying formal data modeling techniques. In case of a data modeling process, we need to define and analyze data requirements needed to support the business processes within the scope of corresponding information systems in organizations. Since Wikipedia is the largest search engine free to use and available planet-wide, I chose to use Wikipedia as the database for the public archive. There are three different types of data models, as shown below, that are produced while progressing from requirements to the actual database to be used for the information system.

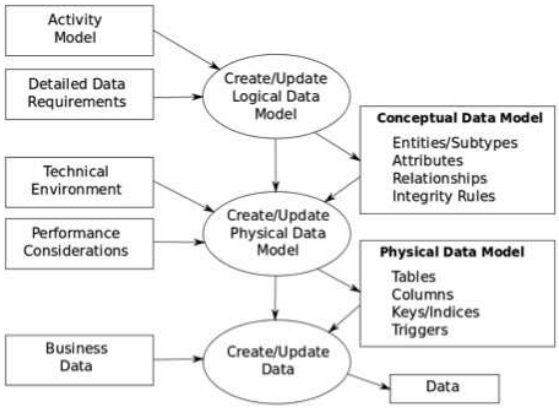


Figure 3.1: Process of Data Modeling

3.2 Data Structure

In computer science, a data structure is a particular way of organizing data in a computer so that it can be used efficiently. In this case, the efficient organizing of the data used for the public archive.
Data structures provide a means to manage large amounts of data efficiently for uses, such as large databases and Internet indexing services. The information that the public archive requires is enormous and hence needed to be managed in a logical and efficient way.
The picture below introduces simple and compound data structures:

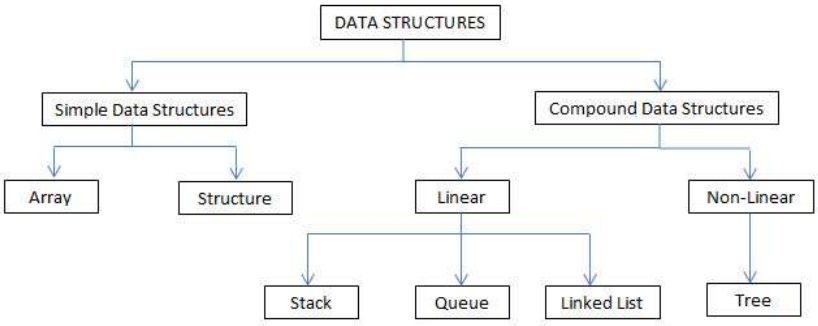


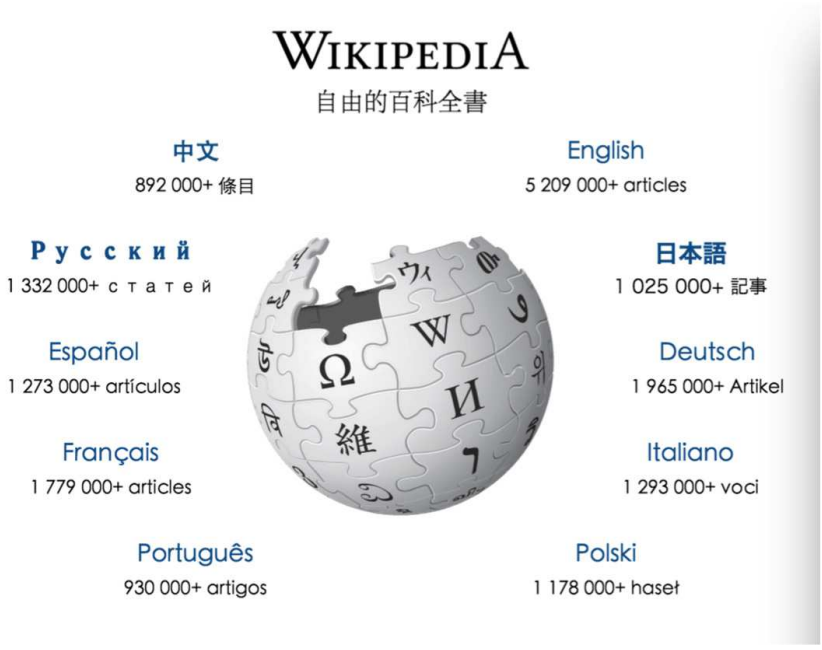
Figure 3.2: Types of Data Structures

Chapter 4

Design the Data Structure

4.1 Language

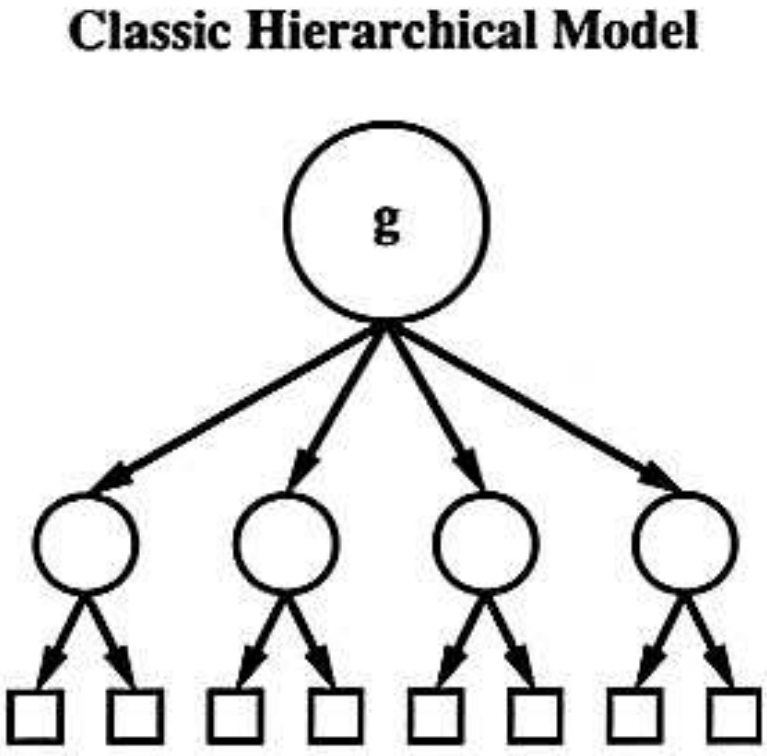
What language to be used is one of the first things comes into mind when designing a data structure. The illustration below shows the articles under each type of language versions of Wikipedia. In order to attain the widest range of information used in the database, I chose to use English Wikipedia as the database for the public archive, since the English version has the most number of articles stored.



In the future, it might be necessary to convert English into symbols recognized by intelligence, since the public archive might be read by intelligence from outer space or by Earth-descended humanity in the far future. This is to provide the far future readers a basis to translate the text into a usable format.

4.2 Data Structure

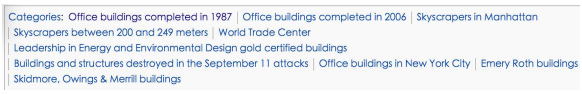
To decide which type of data structure to be used, first of all, I need to know the data structure of the database: Wikipedia.
The data structure Wikipedia is a **hierarchical model**, a multi-layered data structure where the data is organized into a tree-like structure, as shown below.



In a hierarchical database model, the data is stored as records which are connected to one another through links. A record is a collection of fields,

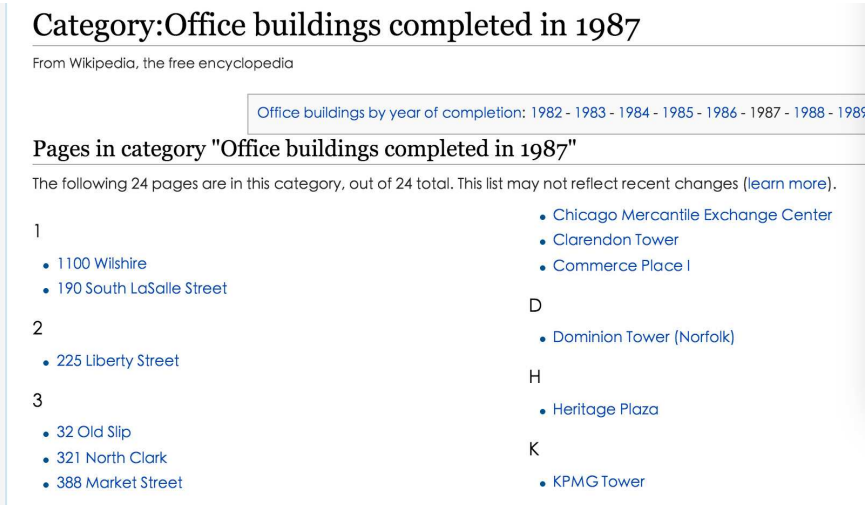
with each field containing only one value. The entity type of a record defines which fields the record contains. A record in the hierarchical database model corresponds to a row (or tuple) in the relational database model and an entity type corresponds to a table (or relation).
The hierarchical database model mandates that each child record has only one parent, whereas each parent record can have one or more child records. The way Wikipedia constructs this tree-like structure is through randomness to determine automatically the most related information, the same way as we use randomness in the Galton Board[1]. By determining the level of relevance, data are grouped into different categories and forms a hierarchical model.

This structure could be seen through the operations of searching in Wikipedia:
If you click into one passage and scroll all the way down, you would find the category it belongs to.



Then if you click on the category, the website would demonstrate all the sub-categories. If you go on clicking these, you would be directed to sub-sub-categories and so on until finally the website would direct you into an article:

[1]The Galton board is a device for statistical experiments, which consists of an upright board with evenly spaced nails (or pegs) driven into its upper half, where the nails are arranged in staggered order, and a lower half divided into a number of evenly-spaced rectangular slots. In the middle of the upper edge, there is a funnel into which balls can be poured, where the diameter of the balls must be much smaller than the distance between the nails. Each time a ball hits one of the nails, it can bounce right (or left) with some probability. This process gives rise to a binomial distribution of in the heights of heaps of balls in the lower slots, and if the number of balls is sufficiently large, the distribution will approximate a normal distribution–one type of statistic distributions.



Category:Office buildings completed in 1987

From Wikipedia, the free encyclopedia

Office buildings by year of completion: 1982 - 1983 - 1984 - 1985 - 1986 - 1987 - 1988 - 198

Pages in category "Office buildings completed in 1987"

The following 24 pages are in this category, out of 24 total. This list may not reflect recent changes (learn more).

1
- 1100 Wilshire
- 190 South LaSalle Street

2
- 225 Liberty Street

3
- 32 Old Slip
- 321 North Clark
- 388 Market Street

- Chicago Mercantile Exchange Center
- Clarendon Tower
- Commerce Place I

D
- Dominion Tower (Norfolk)

H
- Heritage Plaza

K
- KPMG Tower

From the above, we could see that the organization of data in Wikipedia is like a tree of many branches.
Since the database of the public archive was chosen to be Wikipedia, then the same data structure should be used to simplify the work. Hence, the data structure for the public archive should probably be hierarchical model.

4.3 Data Contents

From a General View

From a general view, we would probably think that a public archive should consist the humanity studies-based information such as history, civilization, culture and the scientific-based information like the naming of species, bacteria and so on, which depicts the environment of natural life. To pick which to be included in the public archive, big data processes would be a good tool to use, therefore, let us look at the statistics provided by Wikipedia (English).

Wikipedia's Statistics

Wikipedia ranks articles according to their quality and importance, as shown below:

All rated articles by quality and importance						
	Importance					
Quality	Top	High	Mid	Low	???	Total
FA	1,159	1,775	1,675	1,013	187	**5,809**
FL	141	560	640	596	116	**2,053**
A	211	410	569	362	72	**1,624**
GA	2,031	4,623	9,032	9,659	1,708	**27,053**
B	11,792	22,384	34,234	26,818	13,463	**108,691**
C	9,925	28,541	63,316	85,682	41,907	**229,371**
Start	16,888	74,168	298,739	749,889	284,444	**1,424,128**
Stub	4,295	30,327	221,651	1,796,410	834,941	**2,887,624**
List	2,925	10,882	33,055	87,671	60,255	**194,788**
Assessed	49,367	173,670	662,911	2,758,100	1,237,093	**4,881,141**
Unassessed	141	426	1,878	16,002	488,327	**506,774**
Total	**49,508**	**174,096**	**664,789**	**2,774,102**	**1,725,420**	**5,387,915**

In the table above, the column from top to bottom indicates the ranking by quality from high to low: the articles of best quality are marked as FA, Featured Articles. The row from right to left indicates the importance from top to low, which are generated according to the click rate.
The quality of articles is assessed by Wikipedia's editors. For example, featured articles are considered to be the best articles used by editors as examples for writing other articles. Before being marked as FA, articles are reviewed for accuracy, neutrality, completeness, and style according to Wikipedia's article criteria. From the table, we could tell that there are 5,809 featured articles out of 5,387,915 articles on the English Wikipedia (0.1% are featured). Need to mention, Wikipedia has internal processes about the assessing articles which no longer meet the criteria can be proposed for improvement or removal at the featured article review.

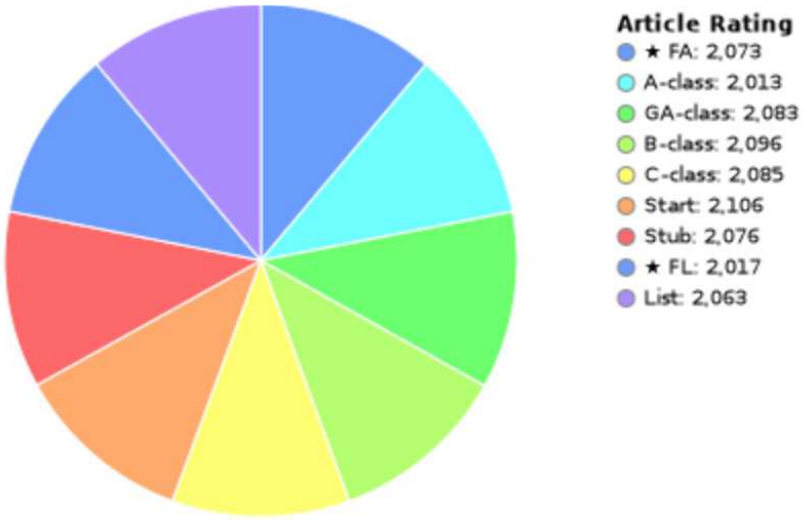


Figure 4.1: Article Rating Piechart

Hence, according to the big data provided, the articles be chosen should be the ones of best quality and top importance.

General Data Content

From the Wikipedia's statistics, I chose to use the articles of best quality and top importance, which are 1,159 out of 5,387,915. However, it does not mean only the articles with top importance (i.e. the top click rate) reflect the key information of human development. However, there are too many articles in the category of FA (5,809); this leads me to consider about how to effectively select articles, which is a key problem. Therefore, I decided to use categories–the key of the hierarchical structure–even more to sort out suitable articles.
In Wikipedia, **The Category:Wikipedia Did you know articles that are featured articles** were used on the main page of English Wikipedia and are a hidden category.
Wikipedia:Did you know (DYK) is the project page for the ' Did You Know ' section on the Main Page. The DYK section showcases new or expanded articles that are selected through an informal review process, where the choice of articles is subject to a set of criteria. Under this category, there are 1,384 articles in total, including inclusive portals that provide important information of human, such as arts, biography, geography, history, mathematics, science, society, technology and so on. By using articles from

these categories, a better and more comprehensive image of human beings would be reflected in the public archive for Lunar Mission One.

From Education's Point of View

In order to fulfill the education purpose of the public archive, something could be done with the data content.
Wikipedia for Schools is a website that includes a selection of articles from Wikipedia that matches the UK National Curriculum and can be used by school children around the world.
It is of a much smaller capacity than Wikipedia that is easier for downloading and using to educate. The picture below demonstrates the topics included in the 6,000 articles available in Wikipedia for Schools.

Figure 4.2: Topics in Wikipedia for Schools

We could therefore engage children in the school to learn this information as well as encouraging them to create their own ' history ' as a human on the big blue dot[2]. They could write up their own information under the categories of art, business studies, citizenship, countries, design and technology, everyday life, geography, history, IT, language and literature, mathematics, music, people, religion and science. These categories are selected based on the subjects they study at schools as well as the categories used in the database of ' Wikipedia for Schools ' .

[2]This phrase is a phrase from the Voyager space program. The blue marble phrase is from the Moon program.

Chapter 5

Summary of the Data Structure

Language: English;
Structure: multi-layered hierarchical model;
Content:
1: Humanities-based and science-based information: 1,384 articles from 'The Category:Wikipedia Did you know articles that are featured articles'.
2: Education-based information: children around the world create their own text and articles under the categories of art, business studies, citizenship, countries, design and technology, everyday life, geography, history, IT, language and literature, mathematics, music, people, religion and science.
Hence, the data structure for the public archive is like:

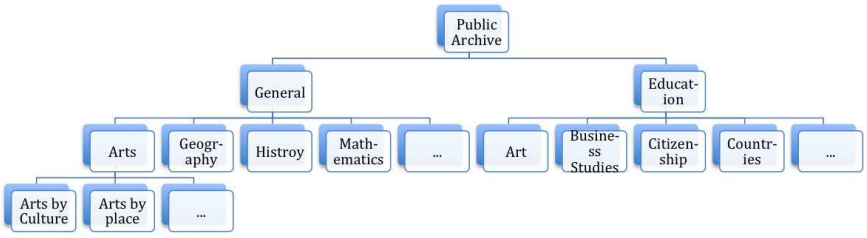
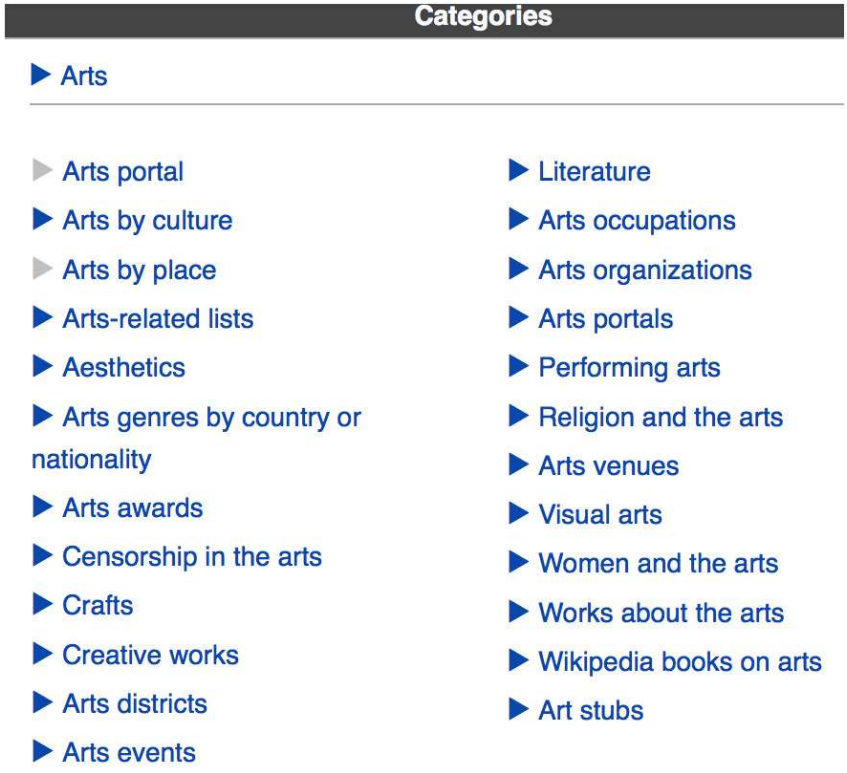


Figure 5.1: Data structure for the public archive

Take the Category:Arts as an example, the sub-categories are like:



And sub-sub-categories could be followed the same way until the article is found.

Impact:
The data structure above allows searching for information under different categories, which relates to relevant information closely.

Chapter 6

Applications and Limitations

6.1 Applications

This data structure is designed for a public archive of Lunar Mission One, which is dedicated to both record human information and engage the public in space exploration.
The usage of hierarchical model manages to create a multi-layered approach to organize the data efficiently, while the usage of Wikipedia as the database helps the data to be organized and inclusive–from highly qualified and officially approved material, to children's contributions with only teacher approval.

6.2 Limitations and Improvements

1. Language: the language used for the output of this data structure is English, but since the public archive might be read by a far future intelligence who does not understand English, the language might be transferred into other symbols recognized by intelligence species.
Improvement: It might be necessary to research into designing of symbols that are commonly recognized by intelligence beings.
2. Capacity: the content of the general information is 1,384 articles, which consumes lots of capacity of the public archive. Besides, the content written by children may also consume lots of space, so the actual size of the public archive needs to be taken into account.
Improvement: The size of the public archive need to be considered, if the capacity is not big enough, the more articles need to be further selected.

Chapter 7

Reflections

Doing this project has been a long but exciting process for me. I was fascinated by mathematics and hence decided to design a data structure for the Lunar Mission One using my mathematical knowledge. I did background researching in data modeling as well as learning Wikipedia with the help of my mentor Gerald Shields, who has a deep understanding about Wikipedia. Those researching process helped me come up with an idea of the data structure of the public archive.

I have self-studied the concepts and applications of data models and data structure, reading books and searching online; as well as studying the way data has been organized in Wikipedia. It was at that moment I knew that there were so many resources available in Wikipedia for us to use as the database, and the problem laid on how to pick up useful articles from millions of articles online. I learned the ' category ' function in Wikipedia, by playing around it and searching for effective grouped information to be used. Besides selecting data of published material in terms of humanities and science, I also learned about Wikipedia for Schools, which helped me generate ideas for the data structure of an unpublished record written by children. This project definitely helped me learn not only about designing data structure but also about devoting to public interests, using STEM knowledge to help the public know more and engage more in science and human history.

To improve my project, I would have started with investigating the way Wikipedia organized data so that it would save me a lot of time for designing the data structure, since I first looked into different types of data structure but finally realized that using the same data structure as the database (Wikipedia) would be the easiest way to do.

There are limitations as I stated above, further improvements could be made. In the future, I would like to investigate more about the capacity of the public archive in order to select the articles and information more effectively. Lunar Mission One seemed like an enormous project for me and I was lucky enough to be involved and designed something myself. I hope one day we could send something useful about human to the Moon, leaving something in the space to prove our existence and persistence.

Bibliography

[1] Matthew West and Julian Fowler (1999). Developing High Quality Data Models. The European Process Industries STEP Technical Liaison Executive (EPISTLE).

[2] Simison, Graeme. C. and Witt, Graham. C. (2005).Data Modeling Essentials.3rd Edition. Morgan Kauffman Publishers. ISBN 0-12-644551-6