Data Structuring and Organisation in the Public Archive

Kajetan Champlewski

Introduction

Lunar Mission one intends to place an archive of data on the Moon to act as a time capsule, containing data regarding humanity and our civilization, intended to last for up to one billion years. The estimated size of this archive is 200TB, and it is to be composed of two parts – the public archive and the private archive. This project focuses specifically on the public archive, as the contents of the private archive are curated by the owners of the data in the archive, whereas the larger public archive is curated and organised by Lunar Mission One and accepts submissions from the general public, and so must be appropriately structured and organised by Lunar Mission One.

In order to ensure the success of such an archive, it is imperative that it be readable by its discoverers. This has been relatively simple to achieve for similar projects in the past, such as the Voyager space probes, which relied on engraved images to explain basic information as well as a process for reading the analogue data attached to the probes. Lunar Mission One faces a more complex challenge, as the data is digital and therefore reading it will require more complex technology. Furthermore, the entire archive is physically located within the lunar rock, in a shaft only a few centimetres in diameter, meaning its contents must be carefully organised and structured.

Another issue this archive faces is limited storage. The 200TB of storage would be sufficient for over 4500 copies of the English Wikipedia [1] (including only the text). This storage would therefore be sufficient for a curated archive, composed primarily of text. However, Wikipedia is not an adequate measure of size as it has strict submission guidelines and is primarily composed of text, whereas the approach of Lunar Mission One consists of using social media and collecting data from the public to be submitted to the archive. An example of a site where the public uploads data freely is YouTube, where over 400 minutes of video are uploaded every minute [2]. At this rate, an uncompressed archive of 200TB would be full in just under two hours [3]. It is therefore clear that compression and specific curation of the data will be required in order to ensure that the archive is not filled up immediately.

The existence of compression in the archive poses another problem: the process of decompressing the data also has to be explained to the discoverers of the archive. This means that the compression methods used have to be limited to ones which can easily be explained to the archive's discoverers, while also maintaining a high degree of compression to ensure that as much data as possible can be stored in the archive. This compromise between compression and readability is the primary challenge of how the data in the archive should be structured and organised within the archive.

Abstract

The report is composed of four main sections, each analysing a particular type of data which we can expect to be present in large quantities in the public archive – text, images, audio, and video. There is also a fifth section which covers recommendations for how the archive should be curated by Lunar Mission One to produce the maximum possible public engagement while conforming to the file standards suggested. The following are the abstracts of each of these subsections:

Text

Text is the most plentiful content submitted to the internet and social networks. As such, we may expect a large quantity of text to be submitted. However, it takes very little data to store a very large amount of text, so the conclusion reached is that text content does not require compression. The characters which are represented by the digital data in a text document also need to be explained to the discoverers, and two methods for this are suggested: the association of the digital data with physical images of the characters, and the association of the digital data with digital images of the characters. A combination of both methods is recommended for maximum readability but the latter method is more practical.

Images

Images are not submitted as plentifully as text data, but take up much more data and as such are the main use of data on social networks. Therefore, images are our priority for compression, as the compression used here will make the biggest overall impact on the data

capacity of the archive. A compression method consisting of a simplified version of the JPEG standard is recommended as it combines a high compression ratio with a relatively simple, easily explained algorithm. The method of explaining how image storage and compression works to the discoverers of the archive is to use physical descriptions of light at various wavelengths (red, green and blue) and the associated digital data in order to demonstrate how raw image data is stored, and following this with a digital step-by-step example of the compression process from a raw image to a compressed image, to ensure the discoverers can reverse the process to read the image data in all compressed images.

Sound

A surprisingly low amount of sound content is submitted to social networks, and it is likely that fairly little sound content will be considered relevant for submission to the public archive. Uncompressed sound at low bitrates also uses relatively little data, so compression would do more harm by reducing readability than good by decreasing data usage. The method of explaining sound data is similar that of image data – a visual description of sound waves and the associated digital data.

Video

Video storage is the second greatest use of data on social media, and so would likely be a large proportion of the archive. However, video compression is extremely complex and cannot be explained as simply and reliably as image compression, and uncompressed video takes up vast amounts of storage space. Therefore, the recommendation for video is to allow only very small quantities of video with only simple image compression (which will still take up large quantities of data) and to disallow highly compressed video altogether as it is practically impossible to ensure it will be readable.

Structuring

It is unfeasible for Lunar Mission One to directly curate all submissions to the public archive. Therefore, it would be reasonable for Lunar Mission One to assign quantities of data for each of its chapters to curate, who would then sub-assign the data to educational organisations and other reputable sources of submissions. In this way, the data can still be centrally curated by Lunar Mission One and can be standardised to conform to the compression algorithms suggested, without requiring Lunar Mission One to directly curate every piece of data.

Methodology

The majority of the literature research was performed online. Encyclopaedias were used to provide an idea for where to find sources and journal articles on the subject and for a general understanding of the topics involved. For locating journal articles, the main resource used was Google Scholar, and also occasionally Web of Knowledge.

The recommendations described in this report were created by selecting the best balance between readability and compression of data where such a balance was immediately apparent, for instance in the case of text, and by reviewing the relevant literature and building upon the recommendations and methods described therein in order to determine the balance in cases where it was not immediately apparent.

No specific experiments were performed to determine the readability of the formats suggested, as these are designed to be readable to any discoverer of the archive which may conceivably be any form of life. It is therefore practically impossible to determine an objective test for this form of readability. Therefore, while the recommendations are designed to ensure the highest possible readability, and certain forms of data are objectively more readable than others (uncompressed data is always more readable than compressed data), no objective standardized measure of readability is available for any of the formats suggested.

Text

Compression

Uncompressed text takes up a very small quantity of information. The complete works of William Shakespeare can be contained in a text file 5.3MB in size [4], which can fit into the 200TB archive over 37 million times. We can therefore assume that text, even if submitted in vast quantities, will not require compression, especially as this compression would drastically reduce readability without increasing the archive's effective data capacity by even 1%.

Readability

Two main standards of digital text encoding are currently in use – ASCII and the Unicode standard. ASCII only covers the English alphabet, numbers, and a handful of special characters. Unicode in its current version covers most characters in most languages that are currently in use, and totals 128,172 characters [5]. The digital sequence of bits representing each character must be related to the character it represents in some way. Two main ways of doing this exist:

Firstly, a physical representation of the characters along with their binary counterparts could be placed alongside the archive, most likely in the form engraved characters and binary data physically located above the archive. This method would be effective for ASCII as it contains only 95 character codes (that have visual representations as individual characters). However, it would be highly impractical for Unicode due to the much larger number of characters involved.

Secondly, the binary code representing each character could be stored digitally alongside the image representing the character. This relies on the discoverers of the archive being capable of reading the stored image data in order to ensure the text is readable, but it is a method which is practical for Unicode as well as ASCII, as 128,172 small images would not take an overly excessive amount of memory.

The most practical method for text would likely be a combination of physical and digital explanations of text, with the physical component limited to the character set of ASCII. This also provides some reference images in both physical and digital form, which may assist in the understanding of the image data.

Images

Compression

Uncompressed images take up a large quantity of memory. Therefore, as images are expected to constitute the majority of the data stored in the archive, image compression is a requirement. Currently two main forms of compressed image formats are in use – JPEG and PNG. Of these, JPEG uses lossy compression whereas PNG is lossless. As a result, JPEG has a significantly higher compression factor and is therefore the compression method that will be used as a base for the file format of images in the archive. Due to the lossy compression, JPEG causes artefacts when compressing text and other high frequency image data (see Figure 1), but because text is being stored separately and most stored images will be photographs, JPEG is preferable.

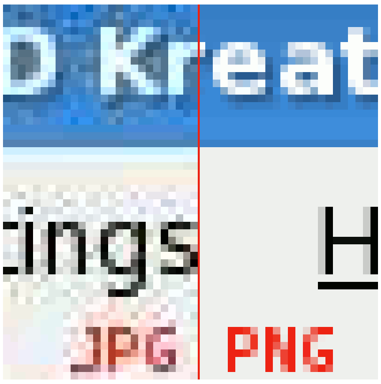

Figure 1: comparison of text compressed with JPG and PNG.

Attribution: By en:User:Toniht, cropped by en:User:Plugwash (English Wikipedia) [GPLv2 (https://www.gnu.org/licenses/old-licenses/gpl-2.0.html)], via Wikimedia Commons

The specific format recommended here is a more restrictive version of the JPEG compression standard and the JFIF file format, created with the specific intention of maintaining as much of the compression and file size reduction as possible through JPEG, while limiting the variety of options available in the compression and thus limiting how much needs to be explained to the discoverers of the archive. Such a standard may be developed by determining which combination of currently available options in the JPEG standard provide the most effective compression on average, and restricting the standard to these options.

Readability

All image files submitted to the public archive must be converted to the simplified JPEG format or to a raw image format before being included in the archive, as any other format would not be explained to the discoverers of the archive and therefore be effectively unreadable.

The explanation of how image data is stored will consist of two parts, the physical explanation of how raw images represent light, and the digital explanation of how compressed image data represents raw images.

Physical Explanation

The physical explanation requires a representation of the red, green and blue colours and the associated binary data that encodes each of these colours. For example, the piece of physical data representing green would have below it the binary inscription 000000001111111100000000, representing the binary data that encodes a green coloured pixel –one with 0/255 red and blue components and 255/255 green components. More than just the three base colours may be used, with accompanying binary data, in order to provide more examples for the archive's discoverers.

The physical representation of each colour must also be considered – it would appear reasonable to simply place a material of the appropriate colour next to the inscription. However, colours can fade, particularly so over one billion years, and we cannot assume that the discoverers of the archive will see colour using red green and blue receptors as we do – the species on Earth generally all have vision sensitive to wavelengths that can pass easily through water (something we cannot guarantee if the discoverers originate from a species which has not evolved in water), but even these species vary greatly in their range of vision. The mantis shrimp for instance has 12 or more different colour receptors [6], as opposed to the 3 present in humans, and birds are capable of seeing in ultra-violet [7]. Therefore, a representation of colour that is understandable to any vision-capable species. As such, the wavelength and/or frequency of the electromagnetic radiation should be communicated, as this is something any species sufficiently advanced to reach the moon will be able to comprehend. This will require a standardised unit of length or time that is recognisable to any potential discoverer of the archive. The unit that appears to suit these requirements is the frequency and wavelength of the light produced by the hyperfine transition of atomic hydrogen, as used by Carl Sagan et al. [8] (see Figure 2). This provides a unit of distance of approximately 21cm, and a unit of time of 1420 Mhz^{-1}. These base units may then be used to represent any wavelength of light, with its accompanying digital representation (see Figure 3). However, more colours can be represented via RGB than via wavelengths of light, as certain colours that can be created through RGB need combinations of multiple wavelengths. Therefore, several wavelength inscriptions may be required with a single RGB inscription to describe some colours.

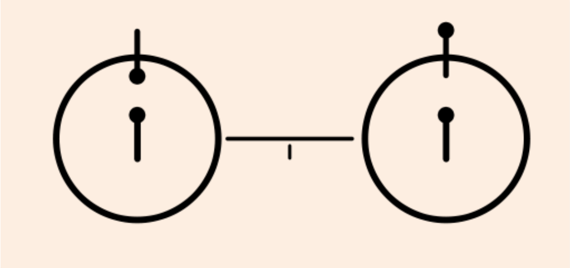

Figure 2: The description of the hyperfine transition of atomic hydrogen as used in the plaques on the Pioneer space probes. A binary one, represented as a vertical line, is used as the base unit. Binary zeroes are represented in the rest of the plaque using horizontal lines.

Attribution: NASA

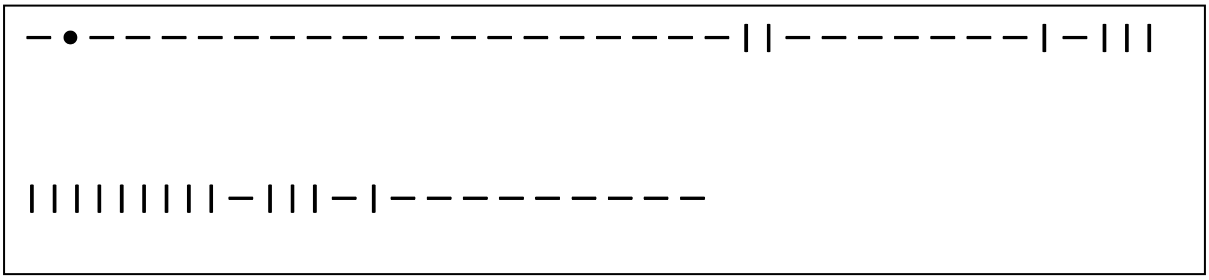

Figure 3: An example description of orange light of a wavelength of 605nm. The data in the upper part of the inscription is the binary data $1.1000000010111_{(2)} \times 2^{-19}$, which is the value that the time period or wavelength of the base wave must be multiplied by in order to obtain a 605nm wave. The lower part of the inscription is the binary data for the colour represented as $\text{FFBA00}_{(16)}$. As such, an association may be formed between the 605nm wavelength and the binary colour code our image storage system uses to represent it.

The positioning of the data is also important in communicating it, as it must be made clear which section of the inscription refers to the wavelength of the light and which section refers to our binary representation of it (though this should be apparent upon analysis of the digital archive). This should be ensured by locating the inscription representing the wavelength of the light in terms of the base units in the direction of the description of the base unit seen in Figure 2. The inscription also uses a binary point – something which may not be understood as easily as standard binary. In order to explain the binary point, it would be best to 'calibrate' the discoverers' understanding of the base units by providing them with data they will already possess, such as an inscription depicting the size of the archive, with binary data (including a binary point) that represents the size of the archive in base units. This will ensure the discoverers understand that binary information to the right of the binary point is in negative powers of two, and binary data to the left is in positive powers of two.

Digital Explanation

While the physical explanation will allow the discovers to understand individual pixels, certain pieces of data, such as the dimensions of an image, may still not be immediately apparent. In order to ensure these can also be read and understood, a simple raw image containing only pixel and dimension data and any other data deemed absolutely necessary for the image should be included at the start of the image storage section. This should be an image of the physical explanation of how to read images, as it can be assumed that the discoverers of the archive have this in their possession and are capable of comparing it to the digital image data.

The majority of images in the archive will be compressed, and it must also be ensured that the discoverers are capable of understanding this compression. The digital explanation of this compression should consist of the base information needed for compression along with a raw image, and should be followed by copies of the image and associated data, one at each stage of compression. This should be repeated with several different images, including at least a few images of the archive or probe itself, and should contain all of the various options available in the simplified JPEG compression format. This will ensure that all images compressed in the archive can be decompressed by reference to these examples.

This digital explanation should be included at the start of the image archive and separated from the remainder of the image archive by a significant quantity of empty bits, so the explanation can be clearly presented as being separate from the remainder of the content, and can be understood more easily. The number of images used as examples may be chosen depending on the space available in the archive, and the options available in the compression algorithm used, the recommended one being simplified JPEG.

Sound

Compression

Uncompressed audio data with a sample rate of 44100 Hz, a bit depth of 16 bits, and without stereo will take up 310.1MB of data per hour of audio [9]. At this sample rate and bit depth, the sound is of a very high quality, and it is reasonable that lower sample rates and/or bit depths could be used should a large quantity of sound data be submitted to the public archive.

As a low quantity of sound data is expected to be submitted, and the quality of this data can be decreased if additional storage space is needed, compression is not recommended for sound data, in particular because it would increase the amount of explanation of the data needed and would thereby decrease the likelihood that the data will be readable to the discoverers of the archive.

Readability

Sound is composed of longitudinal waves of various frequencies. Similar to how colours are physically explained, the sound data would require a physical explanation. In this case, the data would take the form of a set of waves, described using the base units as in the case of colour data, with the binary data representation of these waves inscribed below the description of the waves themselves. The amplitude of each wave would also be provided alongside it, as different waves can have different amplitudes in a single sound sample. The sampling rate will also need to be specified, once again in base units, so that it is clear what time period is represented by each set of waves. As the sampling rate and bit depth of the data must be explained physically, these values will be fixed for all the sound data throughout the archive, and as such must be chosen to be sufficiently high to ensure the sound is of a high enough quality to convey information, while also being sufficiently low to ensure the sound data does not take up an excessive portion of the archive.

Sound data will contain little digital explanation – no new lines or dimensions must be explained digitally, as the relevant information of how combinations of sound waves are combined into binary data is already explained physically. Also, unlike with images, there is no reference material we can assume the discoverers will have in their possession. It would therefore be reasonable for the sound archive to contain no digital explanation of image data and to simply contain the sound files that were submitted, in their uncompressed format at the sampling rate and bit depth described in the physical explanation.

Video

Compression

Raw video data is composed of a series of images, each displayed for a particular time period, for instance 1/25 of a second, in order to produce the illusion of a moving image. A raw 25 frame per second video would therefore take up as much data for each second as 25 images of the same size. Compressed video takes up significantly less space, so compression must be used if large quantities of data are to be uploaded to the archive. However, video compression is extremely complex, relying on encoding key frames followed by changes to the image over a period of time before presenting another key frame. This cannot be explained in a simple way physically or digitally. As such, compressed data will have to be disallowed in the public archive as it will likely be unreadable to its discoverers and therefore useless to include in the archive.

These limitations mean that only uncompressed video, or video composed of compressed images (without any compression that relates the images to each other) can be included with a high likelihood of being readable to the archive's discoverers. This means that the submission of video data will need to be severely limited in order to ensure the archive does not rapidly become filled up with video.

Readability

The readability of raw video is dependent on the readability of images, and requires no complex physical explanations beyond those already created for images. A digital explanation of video may simply consist of a set of compressed images, separated by the time period of a single frame represented in base units, in binary. Following this explanation, the video section of the archive may follow, containing video files which consist of series of compressed images.

Structuring

Separate Archives

As each individual data type will require its own physical explanation, it would be most useful to combine this explanation with the digital data it refers to. Therefore, the archive should be physically structured as separate archives for each of the data types, with the relevant physical and digital data clearly combined. While this would most likely decrease the amount of storage available in the archive somewhat, it would vastly increase the readability of the archive as it would be clear which part of the archive the physical explanations refer to.

Hierarchical Centralised Data Sourcing

A large quantity of data must be sourced and processed for the public archive. As strictly defined file formats must be used for the archive, the data must be centrally collated using a system capable of processing these file formats. It would therefore be best to use a centralised platform where users may submit text, images sound and video, which will allow for the data to immediately be processed into the formats which will be used in the archive.

As data is to be sourced globally, it would be most efficient to distribute the data quota of the archive between the chapters of Lunar Mission One located in various countries, and allow them to curate individual submissions and to sub-distribute data quotas to educational institutions and other sources of submission. These sources could then use the centralised platform to fill their data quotas with their submissions, ensuring that the total data quota is not exceeded, while also ensuring as many various sources as possible have an opportunity to submit their own information.

Conclusion

The content of the archive should be separated into text, images, sound and video. Each of these should use their own file format and come with their own explanations:

- Text: uncompressed, characters physically and digitally explained.
- Images: simplified JPEG compression, physically and digitally explained.

- Sound: uncompressed (though potentially decreased quality), physically explained.
- Video: uncompressed, digitally explained.

Of these, it is expected that images will be the primary use of data in the archive, followed by video, sound and text.

The physical explanations will take the form of engraved data on metal, each located with the relevant part of the archive, with details described using binary, in units based on the hyperfine transition of the hydrogen atom, as used on the Pioneer plaques developed by Carl Sagan et. al. [8]. The digital explanations will be located at the beginning of the relevant digital part of the archive.

Overall, this structure and format of data will ensure that all data remains as readable as possible while also allowing for adequate compression so that a large number of submissions may be accepted.

As file formats and compression algorithms are constantly changing, it is not possible to state conclusively the best file format or compression method possible for any given application. However, for the purposes of communicating data to the discoverers of the lunar archive, the simplest compression method with a reasonable compression ratio is likely preferable, and this is what the recommendations of this report are based around. The explanation of how data is stored follows a similar rule of prioritising simplicity, but the specifics of how this is done are far less likely to change than the specifics of the file formats, as they are based on establishing fundamental units and descriptions based on the laws of physics, which tend to remain consistent over time.

Overall, the recommendations of this report should be taken as suggestions for how the public archive should be designed, but as data storage technology changes their specific details may need to be changed to remain relevant.

Evaluation

While creating messages which can be understood by any potential receiver, such as the Pioneer plaques or the digital messages sent out by SETI, has been a regularly researched topic, the challenge presented by the public archive is unique in that it requires compression of the data to be combined with ensuring its readability, and increasing one of these quantities generally decreases the other. Furthermore, modern file standards are often extremely complex and convoluted, as they must include all the changes the standard underwent throughout its existence. This further increases the challenge of ensuring the readability of such standards to an audience (in this case the potential discoverers of the public archive) who do not know any of the details of the standard.

Evaluation of Compression and File Standard Recommendations

In this project, I endeavoured to describe the file standards used for each of the various data types in the project in as great detail as possible. In particular, I devised guidelines for a simplified format of JPEG which can be used to maintain the readability of images while ensuring their compression. However, the details of the JPEG standard are not freely available and the standard itself is extremely long and contains a wide variety of different options [10]. Therefore, while I believe the guidelines recommended here (limiting the options as much as possible to simplify the standard) are valid, a lot of work remains to be done in reviewing the JPEG standard, determining which elements of it produce the most simply and effectively compressed image, and formally creating a subsection of the standard which unambiguously describes the format which should be used.

For the other data types, no additional compression is in use. so no standards need to be revised – instead, further work needs to be done in order to formally develop simple, unambiguous standards for storing this data which contain as little metadata as possible in order to maximise readability. While the concept of these standards is simple – a lack of complexity – their formal implementation will require a large amount of further worker and this is an area which needs to be pursued further.

Evaluation of Readability and Data Explanations

The second main aim of this project, beyond developing guidelines for easily readable file standards, was to ensure that the data was readable. The problem of communicating data to an unknown receiver who may be of a different civilisation of species has been considered in the past, so literature was available in this area. Utilising the work of Carl Sagan et. al. [8] in developing a standardised base unit that can be understood by any civilisation with sufficient technology to reach the lunar archive enabled me to conceive of a relatively simple method of physically explaining image data as well as sound data. However, further work in this area is also needed, as the specific physical explanations must be developed and collated in a way that ensures readability (a challenge given that the physical descriptions must be placed in a shaft a few centimetres wide along with the archive) and that conforms to the file standards that will be developed.

Ultimately there is no one objective way to judge readability of content intended for unknown recipients of an unknown civilisation or culture, so further work in this area is particularly useful in order to allow a consensus on the best combination of compression and readability to develop, as this is not something that can be objectively determined and therefore a selection of subjective opinions on the topic would be required to develop a sound judgement.

Improvements

Some modifications to the way this project was carried could have led to its improvement. I feel that the scope of this project was too broad – had I instead focused on specifically the format and description of image data, for example, I may have been able to develop a more detailed set of guidelines for the simplified JPEG file format, and may have been able to produce more detailed information and methods of explaining this file format to the archive's discoverers.

Another improvement that could have been made, specifically in the area of images, is to base the file format on Google's WebP standard instead of the JPEG standard, as WebP compression results in files 25-34% smaller than JPEG [11], although its compression is more complex than that of JPEG and therefore may impede readability.

References

[1] "Size of Wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia. [Accessed 23 August 2016].

[2] “300+ HOURS OF VIDEO UPLOADED TO YOUTUBE EVERY MINUTE,” Tubular Insights, [Online]. Available: http://tubularinsights.com/youtube-300-hours/. [Accessed 10 September 2016].

[3] “Video filesize calculator,” [Online]. Available: http://toolstud.io/video/filesize.php?imagewidth=1280&imageheight=720&framerate=25&timeduration=400&timeunit=minutes. [Accessed 10 September 2016].

[4] “The Complete Works of William Shakespeare,” 1 September 2011. [Online]. Available: http://www.gutenberg.org/cache/epub/100/pg100.txt. [Accessed 10 September 2016].

[5] “Unicode® 9.0.0,” Unicode, 21 June 2016. [Online]. Available: http://www.unicode.org/versions/Unicode9.0.0/. [Accessed 10 September 2016].

[6] H. H. e. a. Thoen, “A different form of color vision in mantis shrimp,” *Science,* vol. 343, no. 6169, pp. 411-413, 2014.

[7] I. C. e. a. Cuthill, “Ultraviolet vision in birds,” *Advances in the Study of Behavior,* vol. 29, pp. 159-214, 2000.

[8] C. L. S. S. a. F. D. Sagan, “A message from Earth.,” *Science,* vol. 175, no. 4024, pp. 881-884, 1972.

[9] “Audio Bit Rate and File Size Calculators,” The Audio Archive, [Online]. Available: http://www.theaudioarchive.com/TAA_Resources_File_Size.htm. [Accessed 10 September 2016].

[10] “ISO/IEC 10918-1:1994,” 17 February 1994. [Online]. Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=18902. [Accessed 10 September 2016].

[11] “WebP Compression Study,” Google, 26 February 2016. [Online]. Available: https://developers.google.com/speed/webp/docs/webp_study. [Accessed 11 September 2016].

Acknowledgements

I would like to thank the Nuffield Foundation for funding this project and providing me with the opportunity to carry out this research. I would also like to thank Lunar Mission One, and Paul E. Bennett in particular for providing a research topic and detailed assistance and information regarding all aspects of LM1 whenever it was needed. I would like to express my gratitude also to Andrew Ross, who assisted greatly in the process of producing my presentation and report, and in locating information for further research. Thanks are also due to Sue Dimond and Gillian O'Carroll, who ensured that the resources necessary for this project were provided and who helped ensure things ran smoothly and were an enjoyable experience.