THE UNIVERSITY of York

Future-proofing of archives

By Yu Lin

Supervisor: Alistair Edwards

15 September 2010

This dissertation is submitted for the degree of Master of Science in Computing in the Department of Computer Science at the University of York

Number of words = 14939, as counted by the MS word count command. This includes all the body of the report and Appendix.

Signed Declaration

All sentences, phrases and illustration quoted to other people's work have been referenced and acknowledged in this project. All regarding people have been involved and acknowledged in the project as well. I can guarantee that there is no plagiarism and making up results in this project.

Name: Yu Lin

Signature: Yu Lin

Date: 13 September 2010

Acknowledgements

Firstly, I would like to express my sincere thanks to my supervisor Dr. Alistair Edwards for his continuous help and support in my project. He was always there to help me solve technique questions and give me some advice.

Secondly, many thanks to Sarah Walpole, who is a staff of the Royal Anthropological Institute and offered me some photographs for my experiments in this project.

Finally, I also would like to thank the lab administrator of the department of Computer Science for installing corresponding applications for me.

Abstract

The emergence of archives is to preserve documents with lasting social, scientific or artistic value, so that they can be accessed by generations in the future like today. However, countless valuable archives are disappearing due to the vulnerability of storage medium and the elimination of various storage formats. Clearly, it is extremely necessary to find out a suitable approach to preserve our precious documents.

The aim of the project is to find out a robust file format for long-term preservation of digital archives from a set of potential candidates, such as TIFF, PDF, JPEG, etc. The worked out format will be the option for long-term preservation of digital documents for the Royal Anthropological Institute (RAI). Through deep study and research, the SVG image format has been chosen for long-term preservation. As most of the documents are digitized in the TIFF format in RAI, the TIFF files are converted to SVG files by two possible methods. In addition, a separate report for non-professional users should be created as guidelines for the RAI to help them establish a complete preservation system of digital documents.

Keywords: archives, long-term preservation, SVG, TIFF

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

This chapter introduces the outline of the project. The background of long-term preservation of digital archives will be introduced in section 1.1. This is followed by the objectives of the project in section 1.2. Finally, the structure of the dissertation will be described in section 1.3.

1.1 Background

The emergence of archives is to preserve documents with lasting social, scientific or artistic value, so that they can be accessed by generations in the future like today. Innumerable archives have been created since the cultures of human beings came. However, most documents were produced by using rag paper or industrial paper, whose duration is limit to centuries as a medium. It is clear that the lifetime of these archives is too short to be suitable for long-term preservation.

Thanks to the rapid growth in digital imaging technologies, a vast number of archives are digitalized to various computer file formats aiming for long-term preservation without loss of information, such as TIFF, PDF, and JPEG, by scanner and other translation tools. After digitalization, these digital archives are stored on all kinds of digital media, including CD, DVD and hard disk and so on.

For the moment, there a larger number of organizations are doing such things. The Royal Anthropological Institute (RAI) is one of them. The RAI is a small and longest-established institution specialized in furtherance of anthropology (the study of humankind), which is dedicated to the process of digitizing a number of classic documents at the present time. The objective of digitization is to allow visitors access to high-fidelity digital copies of artifacts without damaging the originals. The RAI digitized artifacts into TIFF and JPEG format by scanner. The TIFF files are produced in a higher resolution and created for long-term preservation of archives, but they are very large. And the JPEG files are lower resolution and more convenient for day-to-day use [1]. However, there is no guarantee of long-term availability of current TIFF and JPEG format.

Although archives are able to be digitized and stored in digital media for accessing in the future, there are two facing problem that the storage medium is very vulnerable and the data format is easy to become obsolete. So, will this digital format or media still be accessible and readable in 20 years? There have been a set of tragedies happened. Take the Blue Peter as an example, Doomesday Books were invented by using a laser disc in 2000, and the preservation format rapidly became obsolete and soon no players are capable of performing it [2].

Based on above example, we found that it is extremely important to select a suitable digital file format for long-term preservation of archives. Actually, there a lot of time and effort has been invested on choosing appropriate format for preservation over the last ten years. Nevertheless, no suitable solution has been sorted out for that so far. Different organizations are still using various digital file formats to preserve archives.

But within the digital preservation community, a model called OAIS Reference Model has been well-established for managing the life-cycle of digital resources during long-term preservation of digital archives. The OAIS, stands for Open Archival Information System, is denoted on the long-term preservation of digital documents as an ISO standard reference mode. It aims both to clarify the very fundamental terms of long-term preservation and to identify the processes required in the lifetime of implementing preserving the documents. The OAIS reference model will be described in chapter 3 in detail.

1.2 Objectives

The objective of this project will be to find a file format which is as robust as possible for long-term preservation of digital archives. After an exact format has been worked out, it will need to find out methods for converting TIFF files to the chosen format. In addition, a report for non-professional users should be created as guidelines for the RAI to help them establish a complete preservation system of digital documents.

1.3 Structure of the dissertation

This project addresses the problem of long-term preservation of archives. The dissertation is divided in 8 chapters.

Chapter 1- this chapter is an introduction to our project. It starts with illustrating the background about long-term preservation of digital archives. This is followed by the objectives of the project. Finally, the structure of dissertation will also be described.

Chapter 2- in this chapter, some useful and relative issues on long-term preservation for digital archives will be reviewed, which includes the OAIS reference model, as a standard model of long-term preservation for digital archives, the XML technology and raster and vector images. An inspirational case study on the NEDLIB project for long-term document deposit will be illustrated in the end.

Chapter 3- both functional and non-functional requirements of the potential digital format for long-term preservation will be listed in this chapter, including the specific oriental format of digital archives, the management and backup requirement, etc.

Chapter 4- in this chapter, in order to find out the most suitable format for long-term preservation, comprehensive comparisons are conducted between a set of potential images and ubiquitous file formats, which are SVG, TIFF, JPEG, PNG and PDF. And the result turns out that SVG is the best format to be applied for the permanent storage.

Chapter 5- the SVG is the format chosen for the long-term preservation, but a majority of digital archives are preserved by using TIFF format in the RAI institution at the present time. This chapter focuses on two main ways of converting archives from TIFF format to SVG format.

Chapter 6 - in this chapter, three experiments are conducted in order to evaluate the outcomes of various applications, such as Vector Magic, Adobe Illustrator, CorelDraw and Inkscape.

Chapter 7- in the beginning of this chapter, some technical approaches of preserving the SVG archives will be introduced and then the most suitable approach, Migration, will be recommended to be applied to the project.

Chapter 8- this chapter concludes the achievement of the research as far, followed by some limitations of the project, and summarized recommendations are illustrated in the end of this chapter.

Chapter 2 – Literature Review

2.1 Introduction

In this chapter, the OAIS (Open Archival Information System) Reference Model will be introduced in section 2.2, which provides a framework for long-term preservation of digital documents. And the XML file format will be outlined as well in section 2.3. In section 2.4, two kinds of computer graphics, the familiar raster image and the less well-known vector image, will be described in detail. Finally, a relevant research on long-term preservation of digital archives will be illustrated.

2.2 The OAIS Reference Model

The OAIS, which stands for Open Archival Information System, is denoted on the long-term preservation of digital documents as an ISO standard reference mode. It aims both to clarify the very fundamental terms of long-term preservation and to identify the processes required in the lifetime of implementing preserving the documents.

Tracing back to the first version of OAIS Reference Model, which is in May 1999, it was firstly developed as a beginning step towards creating formal standards for long-term archiving for the Space Science data by the Consultative Committee for Space Data Systems (CCSDS). In the following years, the standard of OAIS was revised several times until ISO standard (14721:2003) was finally accepted in February 2003.

Although the OAIS Reference Model is initially generated form the space research, it is also being applied in many other fields nowadays. Actually, the model gives a very generally theoretical model description of the organization of an archive. It is compatible enough to be used in both conventional and digital documents. There are two models to be applied to develop long-term preservation, which are Information Model and Process Model [3].

2.2.1 Information Model

In the information model, all objects are gathered in a package, called information package. All the information objects are encapsulated and only can be moved in the range of the package. The whole package is wrapped in package information, which includes both the content information and the preservation description information. Getting into the inner side, the content information is composed with both representation information and the data objects. In addition, preservation description information (PDI), which covers all the information needed for preservation, including provenance information, context information, reference information and fixity information. In the information model, required information is packaged, and the OAIS Reference Model should hold descriptive information about the package to facilitate search and retrieval. For a graphical representation of this, see Figure 2-1.

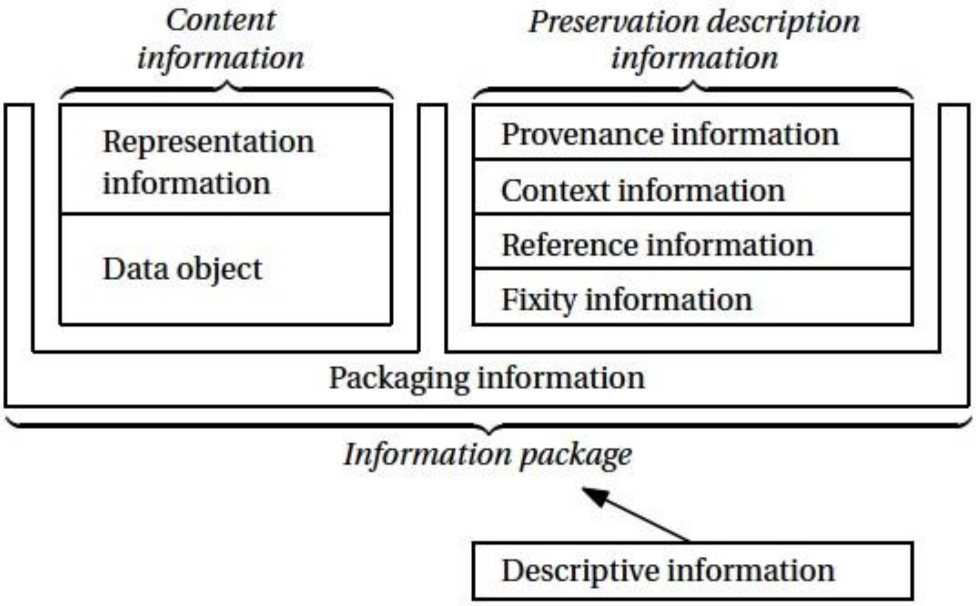

Figure 2-1: The Information Package Model [4]

There are three kinds of information packages that the OAIS Information Model recognizes. They are:

- Submission information packages (SIP). It is submitted to the archive by produces.
- Archival information package (AIP). It is the variant that the OAIS actually preserves.

- Dissemination information packages (DIP). It is a tailored version of the API based on consumer requirements.

These variants of information packages may include other information packages. For example, AIP may be packed into a single DIP for dissemination to a consumer and SIP can also contain information that planned to be spilt into several AIPs.

2.2.2 Process Model

As illustrated in Figure 2-2, the process model is modeled and grouped into the following six processes: Administration, Preservation Planning, Ingest, Data Management, Access and Archival Storage.

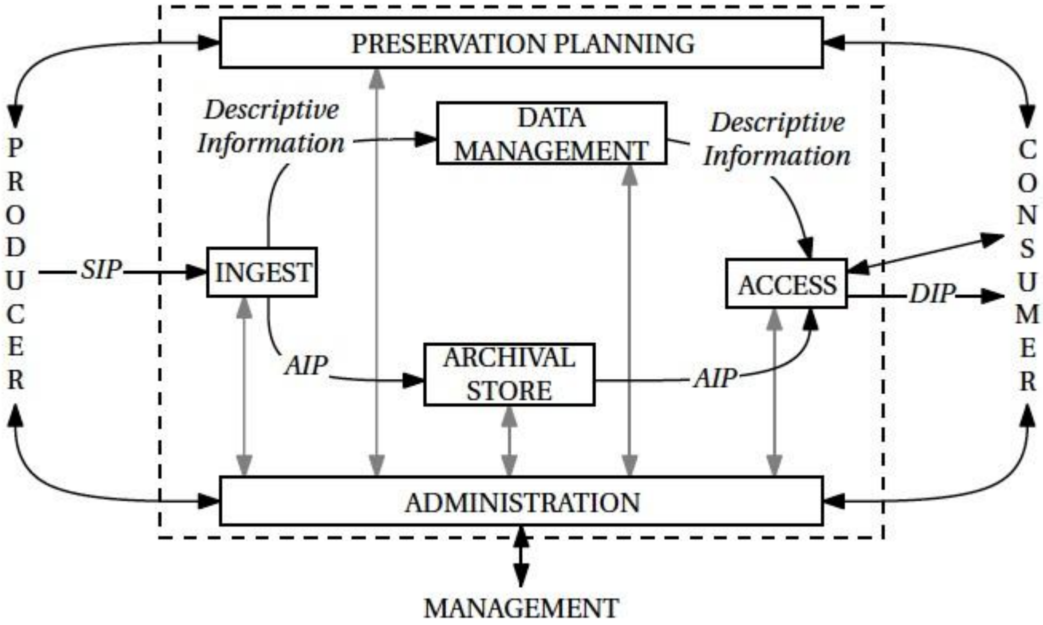

Figure 2-2: The OAIS Process Model [4]

Ingest - This process gains the submitted information from the producer in a form of SIP package. The process aims to store the information from SIP package and prepares the information package for administration; produce AIPs for archival store and generate descriptive information for data management process;

Data Management - The process covers the data required to run the system and the data is always stored in a form of database. The process aims to: store and update the data; save the data; and create descriptive information to the Access process.

Access - This process performs the searching and accessing of archived information. The functions of the process are: enable the customers to search the archive's contents; allow retrieval of the archival information; receiving APIs and Descriptive Information from Archival Store and Data Management process; producing DIPs and delivering to users.

Archival Storage - This process obtains AIP packages from the Ingest process and it focuses on the storage of the AIPs. The functions of the process are: to take charge of the storage hierarchy; error detecting; disaster recovery; to replace media as necessary; and it also producing copies of AIPs to the Access process on request.

Preservation Planning - This process protects the procedures and policies at the OAIS covering the technological changes. The functions of the process is managing technology, platforms and standards; monitoring requirements changes; interacting with producers and customers; designing preservation standards and strategies; and packaging planning and designing.

Administration - This process denotes all the activities required to make the OAIS running smoothly. This process aims to: cover the configuration of the whole system; manage the relationships between all the processes; update archival information; activating requests and customer service and performing access control in physical aspect.

The outlined activity flow in the OAIS Reference Model is well illustrated through the six processes, showing both the flows from archive to user and form producer to archive. The OAIS Reference Model is being considered as an effective standard and guidelines for establishing a complete preservation system of digital archives at the present time. Additionally, it has been applied in many projects all over the world, such as the Planets project [5], The NEDLIB project [6], the InterPARES Project etc. [7].

This project will need to base on the OAIS reference model as well, especially in the phase of preservation of electronic archives.

2.3 Extensible Markup Language (XML)

Extensible Markup Language, abbreviated XML, is produced by W3C organization. It is restricted form the Standard Generalized Markup Language (SGML), and it is an application profile of SGML. XML gives descriptions of both the partial behaviour of the processing computer programs and a set of data objects, which are called XML documents.

XML is that it is a *Meta* language, a language for describing language. You can therefore effectively describe anything in XML; you just need the right DTD. For instance, you might want to have a typeset document (with different fonts and styles) and that could be described in XML – even though the XML file itself would contain only plain (Unicode) text.

The appearance of XML provides generality, usability and simplicity over the Internet. Among the world languages, XML is strongly supported by Unicode as a textual data format. XML is mostly used due to its flexible representation of data, although its design concentrated on documents. For example, it is widely applied to web services.

Over the last ten years, XML has been applied in many aspects, even in the aspect of images. For example, the SVG image format, which is vector image, is created based on XML language. More details about the SVG format will be introduced in chapter 4.

2.3.1 What's so great about XML?

There are many reasons for XML to be such popular, such as ease of exchange and data handling. In the following, it will give a brief description of some strengths of XML [8].

Easy Data Exchange

The markup and data stored in XML can be configured. Because XML is a plain text, developers can have a direct access, and they can easily examine and modify the data. As it is just text, all the data does not need to be encoded in some patented or copyrighted way, and this will make the XML more accessible. Moreover, comparing with other formats, it is very efficient to store most data.

Customizing Markup Languages

Developers can customize markup languages in XML, and it is very easy to create. For different makeup languages, people can create their customized applications and browsers.

Self- Describing Data

The entities called data units in XML documents are self-describing. See the following simple document of self-describing data:

```
<? xml version="1.0" encoding="ISO-8859-1"?>
<FILE>
   <GREETING>
            Hello XML
   </GREETING>
</FILE>
```

In the above example, a root element named FILE was created. There is a customized markup element called GREETING inside the root FILE. It is easy that we can understand meaning of document by customizing our tag.

Structured and Integrated Data

Another powerful feature of XML is that it is not only a representation of data, but also the structure of the data and how these elements can be well integrated into others. Structure and integration are very important when dealing with complex data.

2.3.2 XML in the future

XML has a rapid development in recent years; it has been applied to many aspects. As the outcome of standard for XML, more and more software vendors have adopted it. XML has now been frequently used in Web and will continue to be loved in the future. XML is a plain text format, and it is believe that more research and experiment will be carried out to find out more practical usage.

XML is everywhere. Obviously, between all sorts of applications, it is the most common tool used for data transmissions. XML will become more and more popular in the aspect of describing and storing information.

In fact, a number of XML-based languages have been generated since 2009, including Atom [9], SOAP [10], and RSS [11]. Moreover, most office-productivity tools use XML-based formats as the default format, including Apple's iWork [12], OpenOffice.org (Open Document) [13], and Microsoft Office (Office Open XML) [14]. According to previous investigation, some digital file format based on XML seems a likely candidate for long-term preservation of archives due to its openness and expansibility [15].

2.4 Raster and vector images

There are two kinds of computer graphics, the familiar raster images (composed of pixels) and the less well-known vector images (composed of paths). The former comprises scanning with file formats such as JPEG, PNG, TIFF and the products of digital photography. The latter can be created by some translation applications and professional drawing applications such as Adobe Illustrator, CorelDraw and Inkscape, but it is not that widely used so far.

2.4.1 Raster images

Raster images, namely bitmap images, are composed of horizontal and vertical rows of pixels, which are grids of small rectangular cells. The quality of the image will be better when these cells are together closer. The pixels are composed of numbers representing magnitudes of brightness and color, and they may carry on the different arrangement and the dyeing constitutes the pattern. Raster images are resolution dependent, and one of its features is that the size of uncompressed file is independent of image complexity. As the raster image illustrated in Figure 2-3, the entire innumerable single blocks can be seen, when raster images are enlarged. Because the enlarging images result in expanding these single pixels, the enlarged images appears irregular and the quality will be reduced.

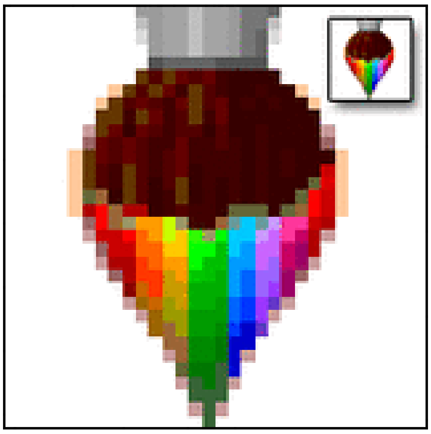

Figure 2-3: A Bitmap Image-shown zoomed in with the original size in the upper right corner [16]

There are some common raster images formats, including JPEG, TIFF, GIF and PNG etc. Some of them will be introduced at length in chapter 4. Nowadays, raster images are being used widely on the Internet and other fields. For example, all images from digital cameras are raster image and scanned images are bitmaps as well.

The advantage of raster images is that it is simple and thereby the images bit plane representation is fast on the raster imaging viewers or devices. As mentioned above, the disadvantage of raster images is that they cannot be intentionally scaled due to the loss of pixel rows by decreasing the size of the images. However, an effective shrinking algorithm for raster document compression is very important. Although the size of raster documents is quite large, raster type of image is very practical for pictorial graphics archiving. Document storage and archiving require the use of raster technology for scanning representation of documents.

2.4.2 Vector images

Vector images are draws up according to the geometry characteristic. Points and paths are connected by using mathematical relationships in order to describe an image, which can only depend on the software production. The images take a relatively small intrinsic space, because this type's image document contains the independent separation image that can be freely and unlimitedly combined again. Thus the quality of the images will be retained, as vectors images are resolution independent.

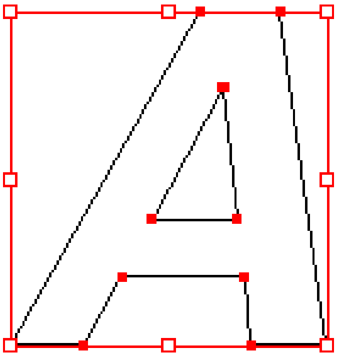

Figure 2-4: An example of vector image [17]

Vector images are composed of separate lines, namely, straight or curved. Figure 2-4 shows an example of vector image. These graphic elements are *spot, line, rectangle, polygon, circle* and *arc* and so on, all of which are obtained through the mathematical formula computation. The mathematical formulas determine the positions of the dots in order to get the best results when illustrating an image and it enables an image to be scalable to any size and detail with the unchanged volume. Because the vector graph may obtain through the formula computation, therefore the vector graphic file volume is generally small.

Some common vector images formats are SVG [18], AI and CDR. SVG is a graphics description language based on XML, which will be presented in detail in chapter 4. At the present time, vector images are used mainly in professional cartography.

The biggest merit of vector graphs is that they will not be distorted whenever they are enlarged or shrunk; because vector images are resolution independent. It can be zoomed in and out to any degree without losing quality of original image. The biggest shortcoming is that it is difficult to display rich color images with lifelike effect due to vector images are usually composed of solid areas of color or gradients. Even so, vector images are becoming more and more advanced, and it has been applied to drawings a lot now than we could a decade ago.

2.4.3 Comparison between Raster and Vector images

Raster and vector images have different features, file formats and rules that they obeyed. The following Table 1 shows the difference between Raster images and Vector images.

	Raster images	**Vector images**
Features	● Composed of pixels ● Resolution dependent	● Composed of objects, each of which has its own color, size, shape, etc. ● Resolution independent
File format	*.bmp、*.pcx、*.gif、*.jpg、*.tif, etc. The size of a same image may vary from each other due to different compression algorithms.	*.AI、*.EPS, *.dwg, *.cdr, etc. According to the mathematical formula of each element, software programs calculate the graphic and display it.
Rules	● Resolution losses when scaling the documents. ● The bigger the graphic area is, the more bytes the document requires. ● The richer the color is, the bigger the size of the document is. ● The process of converting raster images to vector images is complicated, which requires tremendous data processing.	● Resolution remains when scaling the documents. ● The size of the document relates with the number of the elements and the complexity of these elements. ● The size of the document has no relationship with the graphic area or the color. ● The process of converting vector images to raster images is easy by using converting applications.

Table 1: Comparison between Raster and Vector images

2.5 Relevant research and development

The NEDLIB project [6], one of the famous projects about long-term preservation of digital documents, is chose as a basic tutorial for further study before this project starts.

The NEDLIB project, which stands for Networked European Deposit Library, was established on January 1, 1998. There are eight European national libraries participated in this project, one national archive, three publishers and two ICT organizations.

There are three main objectives of the NEDLIB project.

First, it aims to design fundamental tools and establish a general architectural framework for building deposit systems for electronic publications (DSEP). The project dedicated to the technical part of extending the deposit into digital works. In order to come out with a highly designed DSEP for local implementation by the eight deposit libraries, the project consortium agreed to take advantage of a genius Reference Model, the Open Archival Information System (OAIS) model. This model has already been widely used by other projects, such as PANDORA in Australia and CEDARS in the UK.

The project also focuses on addressing the issue of long-term digital preservation, which includes the preservation requirements of various categories of digital deposit. Conventions and standards for digital deposit will be provided in order to guide these deposit libraries when implementation a DSEP.

The last objective is to construct a demonstrator system covering the whole DSEP with the existing tools and software developed by NEDLIB. The external existing library systems, such as the cataloguing and library acquisition systems, will interface to the demonstrator by interacting with a DSEP [6].

A modeling of a DSEP on the basis of OAIS was created in 1999. The model was called "process model". In fact, the process of the model was mapped to the OAIS set of functional entities. See Figure 2-5 as follows:

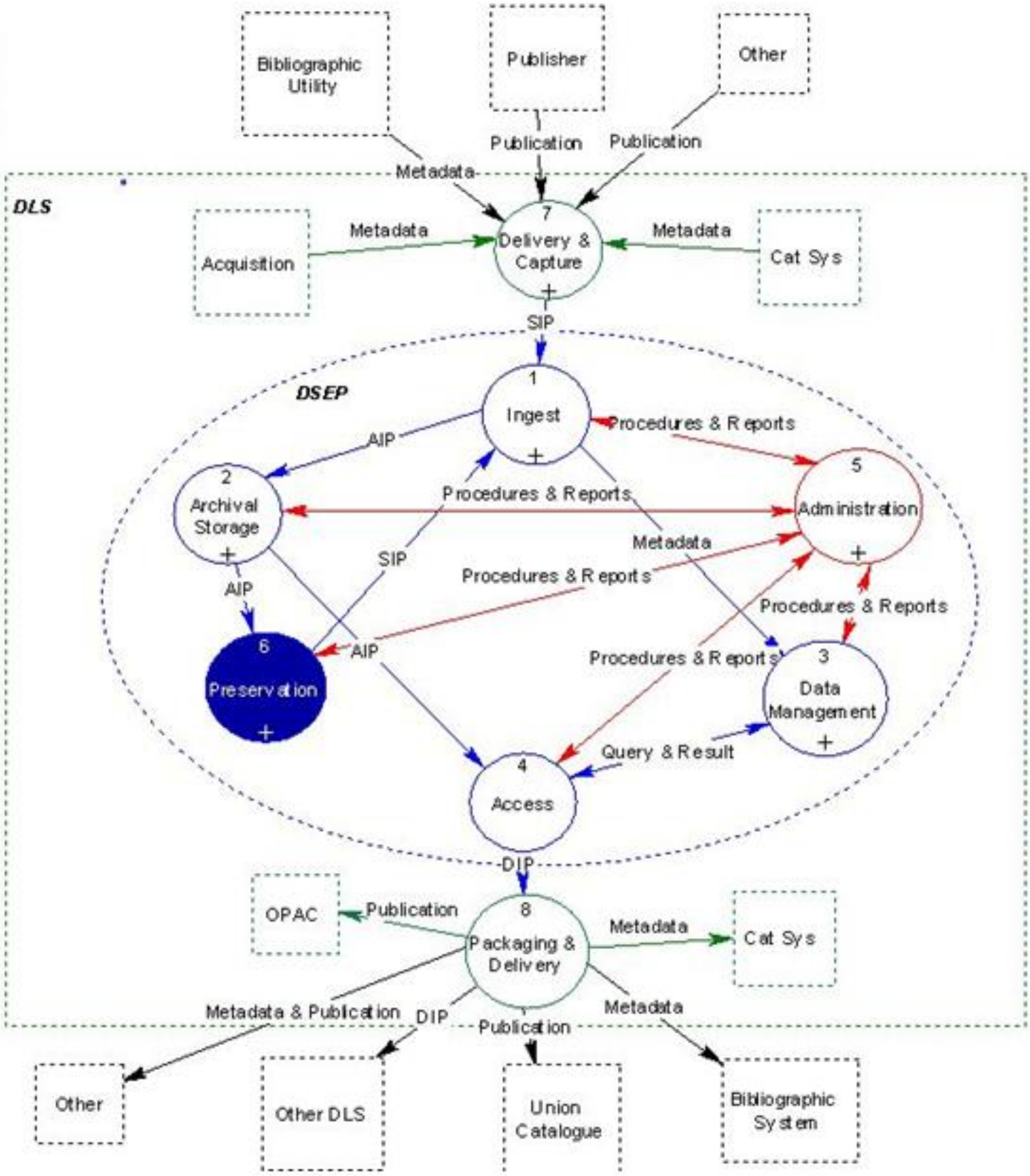

Figure 2-5: DSEP process model for handling electronic publications [19]

The above figure shows the workflow of the process model for the DSEP. It detailed how this model works for depositing electronic documents, from selection stage to end-user access.

2.6 Conclusion

In this chapter, the OAIS reference model for long-term preservation of digital documents is briefly reviewed, with illustration of its information and process models. XML technology is also described due to its wide application in various images format, such as SVG. Additionally, the advantage and disadvantages of raster and vector images are comprehensively compared in order to get a clear picture between these two formats. Finally, a case study on the NEDLIB project is illustrated and the DSEP process model gives some inspiration on the long-term preservation.

Chapter 3 – Requirements Specification and Analysis

3.1 Introduction

This chapter will elicit and analyze both functional and non-functional requirements of the project, which is to look for a future digital format for long-term preservation of digital documents. And some potential problems that may happen in the system will be discussed based on the elicited requirements as well. Besides that, proper recommendations should be given on management of document preservation.

3.2 Requirements Overview

Information technology is updating quickly over time. There are a number of digital document formats that have been emerged in the past, such as TIFF, JPG for image and TXT, RTF for plain text. They can still be accessible by current software, but what about 10 years later?

An appropriate format should be chosen or designed in advance. And then, a format translation system will be designed to convert the TIFF format to the found format. In addition, this worked out format should be able to display its content in specified software or in some mainstream browsers, for example, Opera, Firefox, IE, Chrome and Safari.

The following Figure 3-1 shows the overview of system requirements. Files in TIFF format will be converted into the chosen or created format by the designed converter. The produced files should be stored based on the OAIS reference model for long-term preservation for further access and be able to be displayed on the mainstream browsers as listed above.

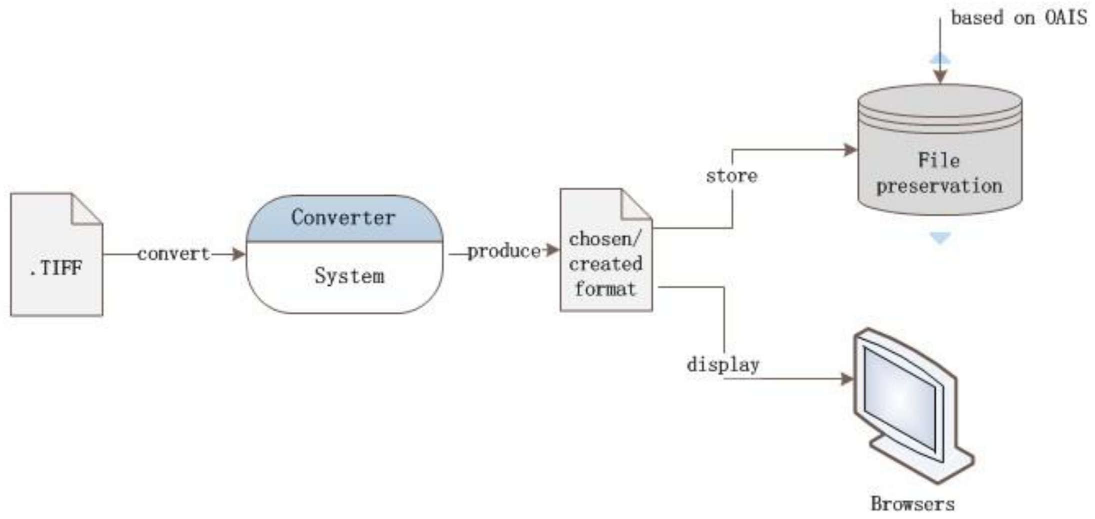

Figure 3-1: Requirements Overview

On the whole, the project for future-proofing of digital archives must be able to support the following high-level goals.

- Choosing or designing a robust and accessible digital format for Long-term preservation.
- Building system that can convert the TIFF format to new appropriate format.
- Reading the worked out format in main browsers or specified software.
- Preserving digital archives for long-term based on the OAIS reference model.
- Generating a separate report for non-professional users from the RAI institution to establish a complete preservation system for digital documents.

3.3 Functional Requirements

Functional requirements consider features required from a system, all the functionality of the system should be well performed without consideration of physical constraints. The functional requirements of this project will be listed below.

3.3.1 Long-term format

To find a robust future format to replace the TIFF format that is vulnerable to obsolescence for long-term preservation. However, if there is no appropriate existing format, a new format has to be devised and implemented for long-term preservation of digital document.

Browsers are playing an important role when users are using computers and they are not possible be out of date in the future. Due to this characteristic, the found format must be supported by the mainstream browsers, such as Opera, Firefox, IE, Chrome and Safari.

3.3.2 Converting system

An appropriate converting application tool should be chose or created for converting the TIFF image format to the new future format. The system must concentrate on following aspects:

- The speed of file format translating
- The quality of produced files
- The size of created files. It is also a key criterion of performance. Clearly, smaller size is more suitable for preservation.

This system should be able to convert TIFF image format to the selected format at least. The TIFF, a common image type, is used widely by RAI and numerous organizations all over the world, especially in terms of preservation of digital documents. They digitized valuable artifacts into TIFF for long-term preservation of archives. In a word, the system should be able to convert TIFF format in priority.

3.3.3 File preservation based on OAIS

The produced files for long-term preservation should be preserved based on the OAIS reference model. In fact, techniques for achieving file preservation are widely used and contain data backup, replication and media refreshment. In the OAIS model, all above activities are fallen into the following six processes: Ingest, Administration, Data Management, Preservation Planning, Access and Archival Storage.

3.4 Non-functional Requirements

Non-functional requirements can be considered as constrains in functional requirements. It is the specified criteria used for judging the performance of the system, rather than the required behavior that the system must implement. The Non-functional requirements of this project will be listed below.

3.4.1 Backup and recovery

Long-term preservation of digital archives will require regular reviews. In addition, these archives may suffer from wrack and ruin due to some uncertain natural calamities. Consequently, a complete backup and recovery planning must be stated for data safety.

3.4.2 Permissions and Copyright

Permissions and copyright of digital archives have caused a lot of issues in the information age. Additionally, these issues are quite difficult to be resolved based on existed law. Hence, relevant advice should be given to deal with the permissions and copyright

3.4.3 A separate report for the RAI

Besides the project report, a separate report should be created as a user manual to be presented to the RAI, guiding the non-professional users how to reformat these documents for long-term preservation. Moreover, it should also provide advice on how to implement data migration regularly [1].

3.5 Conclusion

This chapter illustrates the functional and non-functional requirements elicited from the project. It is required that a robust future format for long-term preservation of digital archives needs to be worked out; a converting system from the TIFF to the found format has to be established; the new formatted archives should be able to display in various browsers; etc. Further research will be carried out based on all the requirements in order to discover the strategy for the project.

Chapter 4 –Research of Digital Formats

4.1 Introduction

In this chapter, we will consider some well-known image formats and ubiquitous file format like PDF (Portable Document Format) to find a suitable format for long-term preservation of digital documents. The image formats include both raster images and vector images, which have been introduced in detail in literature review. Raster images have some alternatives such as JPG, TIFF and PNG. Vector images, like SVG, may be less known so far but it is becoming more and more popular at the present time.

An appropriate format will be chosen for long-term preservation by analyzing and assessing characteristics of these main digital formats.

4.2 The Principal Digital Formats

Nowadays, thousands of digital file formats have been created according to various requirements and needs. However, due to the limited time, only a small number of principal digital formats will be focused on in order to find out a suitable one.

Generally speaking, file formats can be classified into three types, standard, open and proprietary. Standard formats are the format that develops to become an international standard, which are still stable and public till next release of the standard. If the software producers publish their format specifications to all the people, the formats are called an open file format. Proprietary file formats are those unpublicized formats developed by software producers.

As a successful file format for long-term preservation should be based on international openness, high stability and good compatibility. Some possible alternatives will be listed as following based on investigation. We will analyze and assess these digital formats to find out their characteristics and potential shortcomings for long-term preservation both the near and distant future.

4.2.1 Scalable Vector Graphics – SVG [18]

SVG, short for Scalable Vector Graphics, is an XML–based vector image file format that is to describe two-dimensional images, both in static and animation.

The SVG specification is an open standard, developed by the W3C (World Wide Web Consortium) from 1999. The latest version is 1.2, but it is still a working draft. Therefore, the current recommendation version is 1.1 that has been developed since 2001[18].

SVG format is defined based on XML text files. This means that content inside SVG image can be searched, edited and scripted, even compressed. Based on these characteristics, SVG image can be created and edited by text editor. Moreover, some professional drawing software such as Adobe Illustrator, CorelDraw and Inkscape also can handle it natively.

The following code segment is a simple example of SVG file with .svg suffix. See Table 2:

```
<? xml version="1.0" standalone="no"?>

<!DOCTYPE svg>

<svg version="1.1" id="Olympic_Rings" viewBox="-34 -12 68 33" width="1020"
height="495" xmlns="http://www.w3.org/2000/svg">

<g fill="none" stroke-width="2" stroke="white">
   <circle cx="-11" cy="9" r="9" stroke-width="3"/>
   <circle cx="11" cy="9" r="9" stroke-width="3"/>
   <circle cx="-11" cy="9" r="9" stroke="#f4c300"/>
   <circle cx="11" cy="9" r="9" stroke="#009f3d"/>
   <circle cx="-22" cy="0" r="9" stroke-width="3"/>
   <circle cx="0" cy="0" r="9" stroke-width="3"/>
   <circle cx="22" cy="0" r="9" stroke-width="3"/>
   <circle cx="-22" cy="0" r="9" stroke="#0085c7"/>
   <circle cx="0" cy="0" r="9" stroke="black"/>
   <circle cx="22" cy="0" r="9" stroke="#df0024"/>


<g transform="translate(-11,9)">
    <circle r="9" stroke-width="3" clip-path="url(#clip_white)" cx="0" cy="0"/>
    <circle r="9" stroke="#f4c300" clip-path="url(#clip_color)" cx="0" cy="0"/>
</g>

<g transform="translate(11,9)">
    <circle r="9" stroke-width="3" clip-path="url(#clip_white)" cx="0" cy="0"/>
    <circle r="9" stroke="#009f3d" clip-path="url(#clip_color)" cx="0" cy="0"/>
</g>
</g>
</svg>
```

Table 2: The code of an SVG image [20]

SVG file is composed of a root element <svg> and some defined elements, like <circle>, <path>, <text> and so on. All opening tags must have closing tags!

The SVG format has been developed with the cooperation of a number of important organizations, including Adobe, Sun, HP, Apple, IBM, and Kodak so far. It is also being supported and rendered directly by all mainstream web browsers, such as Opera, Safari, Google Chrome, Firefox, etc. However, Microsoft Internet Explorer (IE) cannot support SVG format natively so far. IE can only view SVG images after installation of a plugin called SVG viewer, which is developed by Adobe Company. Fortunately, Internet Explorer 9, the next version of IE, will support SVG directly in 2011.

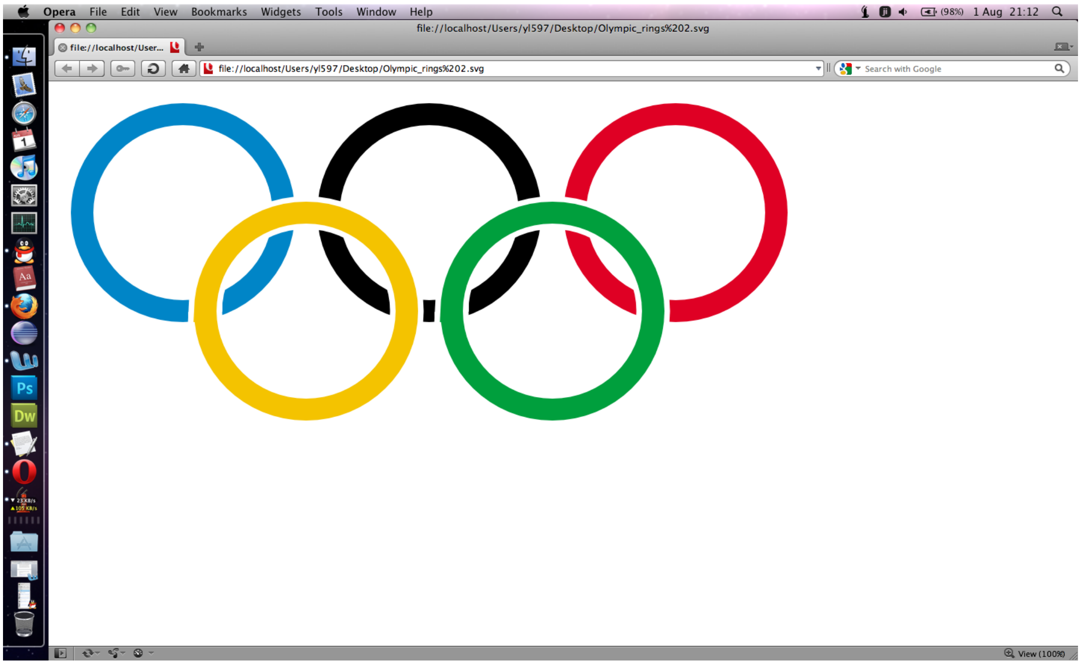
Opera File Edit View Bookmarks Widgets Tools Window Help
(98%) 1 Aug 21:12
file://localhost/Users/yl597/Desktop/Olympic_rings%202.svg
file://localhost/User...
file://localhost/Users/yl597/Desktop/Olympic_rings%202.svg
Search with Google
View (100%)

Figure 4-1: The visual image based on code of Table 2

Figure 4-1 is the outcome of the SVG code displayed in Table 2, which can be well supported by the Opera browser.

In contrast to a raster image (like TIFF and JPG), which is composed of pixels, SVG image has a number of important strengths follows:

- SVG is an open standard
- SVG files can be created and modified by many tools, even text editor.
- Text in SVG file can be searchable and indexed.
- SVG images can be scalable unrestrainedly with lossless of quality.
- SVG supports complex animation.
- SVG allows linking and embedding a graphic object, including raster image.
- The size of SVG files is smaller than common bitmap images.

Scalability is one of the most excellent advantages of the SVG format. Taking advantage of this feature, we can easily produce high-resolution image by scaling up a SVG image. So far, SVG files have been applied in plenty of aspects, especially in printing and mapping.

Here is an example about scalability of SVG file. Figure 4-2 is a small unscaled SVG file of the Olympic flag. The Figure 4-3 is the SVG file after scaling up by several times. Comparing above two images, we will find out that the latter image is zoomed with lossless of quality. That is a major attraction of SVG.

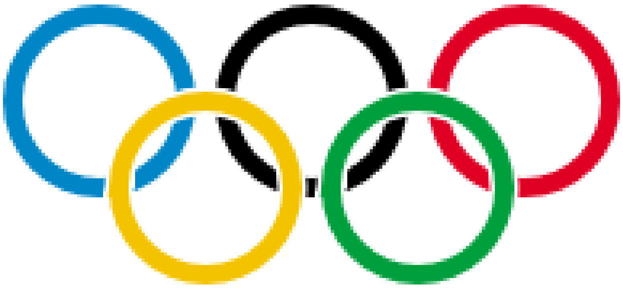

Figure 4-2: The initial Olympic flag [20]

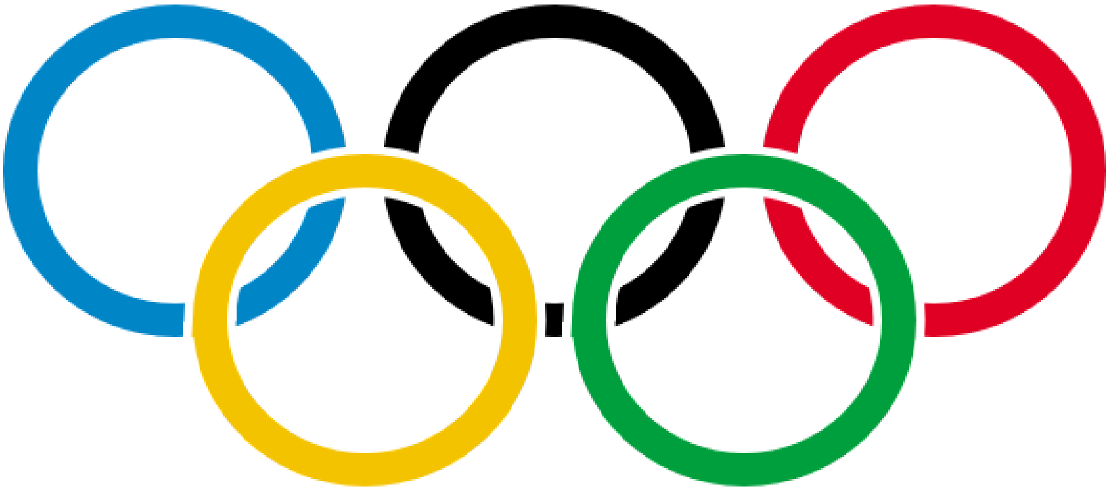

Figure 4-3: The changed Olympic flag

The SVG is an open standard format and is based on XML technology. Obviously, it is a potential alternative for long-term preservation of digital documents.

4.2.2 Tagged Image File Format – TIFF [21]

TIFF, stands for Tagged Image File Format, is one of the most common and widely used bitmap image formats for preservation today. It is also a standard in the widespread industry such as publishing and printing.

TIFF is a proprietary image format. The format was invented by Aldus Company in 1986 but was later acquired by Adobe Company, which holds officially the copyright to its specification now. It is one of earliest image file format, even though it is still very popular at this time. The TIFF format is comprehensively supported across all platforms such as Windows, Mac, and Linux. What is more, it is also widely sustained by numerous image applications, by printing and publishing applications, by word processing, optical character recognition and scanning.

There are two file extensions for TIFF, which are .tif and .tiff. The former is used a lot today. The latter is applied occasionally as well.

There is a major difference between the TIFF and other image formats that TIFF supports most of the color spaces and a wide range of various compression algorithms, which is defined in dedicated tags. The lossless compression schemas supported by TIFF are listed as following:

- LZW (Lempel-Ziv-Welch) [22]
- CCITT Fax group 3 & 4, Used a lot in fax and multi-page line art document

In fact, the TIFF format is very flexible that can be modified and enhanced to allow more additional functionality and extension, according to software developers' needs. However, such extensibility comes forth the problem of format mismatch. Images produced by one software developer may not be used by other software developers. A typical example is that the popular image editor Photoshop cannot read the TIFF image made by SCITEX products, which is one of most famous scanner companies in the world [22].

The TIFF is used as the popular preservation format at this time and it is expected to be the right format for digital documents preservation in the foreseeable future. Numerous companies and organizations are using TIFF format as long-term preservation format of digital documents at the moment.

Although TIFF is a very popular preservation format due to its high quality so far, it is not the most suitable format for long-term preservation according to investigation. There are some reasons for that. First, TIFF files take up a more space in a memory disk and more time when software read them, by contrast to other image formats such as JPEG, PNG. Second, the TIFF is not an open format, but a propriety image format that belongs to Adobe Company. Moreover, some cutting edge digital cameras cannot support TIFF format, which means that it is being abandoned inch by inch. All these points make it unsuitable to be a candidate for long-term preservation for digital archives.

4.2.3 Joint Photographic Experts Group – JPEG [23]

Joint Photographic Experts Group (JPEG) is a raster image format that utilizes lossy compression method. The format supports a selectable tradeoff between file size and image quality. By using the efficient lossy compression, JPEG can reduce file size to 5% of its original size with almost perceptible loss in image quality.

The most common file extensions for JPEG are .jpeg and .jpg. However, .jif, .jpe and .jfif are used occasionally as well. JPEG can support a wide range of bitmaps, such as gray-scale JPEGs, full-color images and real-world scenes, even at 24 bits per pixel. Moreover, it can also compress these bitmaps to reduce image size for storage and transmission.

The aim of JPEG is to be used in highly photographic or detailed graphics, and typically, is applied to digitized and rendered images. But, JPEG is not suitable for screen captures, line drawings and other image types, like regions that used sharp lines and colors. For high image quality, these graphics should be saved in some lossless image format such as TIFF, PNG and GIF.

Today, due to small amount of JPEG image data, the format has been the most popular image format that used for web graphics and storing pictures, used by most digital cameras and image capture device. But, these JPEG variations are not different actually. They all simply called JPEG.

Strictly speaking, JPEG is just a family of compression algorithms, not a “true” image format as an official standard that defined by the international standard organizations. Out of question, JPEG is very popular on the web, but there are still some arguments about it so far. For example, format compatible. We can exchange images with anyone else without a common agreement. This means that JPEG program maybe cannot read the JPEG file that created by other programmer.

There is no doubt that JPEG will be still accessible in the near future. However, it is not suitable for long-term preservation due to its application of lossy compression technique. This means that some of image data will be lost in the compression process. The characteristic makes JPEG not suitable for archives preservation.

4.2.4 Portable Network Graphics – PNG [24]

PNG, short for Portable Network Graphics, is a bitmap image format similar to GIF. It is a fully open source image format, is supported well by W3C.

PNG format was developed to replace older GIF and complex TIFF format in 1995. On one hand, PNG is a license-free and patent-free image format. This means that developers do not need to pay any fees when they are using PNG lossless compression. However, GIF uses the LZW compression method, which is a patented compression algorithm owned by Unisys. On the other hand, unlike TIFF, the PNG is applying lossless compression, which allows user to own these two key characteristics, high quality and small size.

There are three major image types that can be supported by PNG such as Palette-based (8-bit), Grayscale and Truecolor. GIF only supports the first one. JPEG only sustain the last two. Comparing to GIF, PNG can even use a bit depth of up to 48-bit true color, which allows PNG to be created for high quality images.

Nowadays, PNG has been used primarily in two fields, the web and image editing. It has been approved as a standard on the web today. For the web, the PNG has some apparent advantages over GIF format as follows:

- Control image brightness in cross-platform
- Variable transparency
- Progressive display method

Besides above advantages, PNG can also compress better than GIF with lossless compression algorithm. However, PNG cannot support multiple-image and animations, which can be implemented in GIF format.

For image editing, PNG has been widely used as intermediate format for image editing and image transformation because of its special characteristics, lossless compression and a large range of bit depth. Based on that, we can resave PNG file to other image formats without degrading its quality.

The PNG compression is one of the best compression methods that can be had with lossless compression and without paying patent fees. In addition, PNG is an open image format, supported by W3C, which make PNG format more suitable for long-term preservation. Today, the most major web browsers and image editing software can also support it very well.

The PNG format is acceptable for long-term preservation, and it could be an option for a long time in the future.

4.2.5 Portable Document Format – PDF [25]

PDF is short for Portable Document Format, is created by Adobe Company for exchanging documents over the Internet in 1993. The advantage of PDF file is that it allows us to convert documents and completely keep the formatting of original documents in print and on the screen, including images and fonts.

PDF is designed for representing documents regardless of operating system, software and hardware. It performs very well in multiplatform such as Windows, Mac OS and Linux. Today, Mac OS can support the PDF format natively without installing extra applications. However, PDF reader software needs to be installed in Windows system to read the format.

Before January 2007, a number of organizations resist to use PDF, as it was a patent file format, owned by Adobe Company. But, after Adobe released the PDF 1.7 specification for ISO certification in 2007, PDF had been an open standard file format under supported by International Organization for Standardization, known as ISO 32000.

Today, a lot of organizations select PDF as a format for long-term preservation, similar to TIFF format. They choose a special PDF version, namely PDF/A, which has been optimized for long-term preservation of digital archives.

It seems that the PDF format is a good choice for long-term preservation, but compared to the SVG, PDF is absolutely not better than SVG. There are some reasons as follows:

- Difficulty in edition

  PDF files are very difficult to be edited after created. This will depress the efficiency of preservation work.

- Software required

  In order to read PDF files, corresponding software is required to be installed in advance. Such as Adobe Reader that belongs to Adobe Company. Although the Adobe Reader is free of charge at the moment, there is no guarantee that it could be cost free forever.

4.3 Comparison of alternatives

Table 3 summaries these five alternative formats for long-term preservation of digital archives, which are SVG, TIFF, JPEG, PNG and PDF. Further comparisons between these formats have been sorted out.

	Description	Compression method	Browser Support	Format License	Recommendation
Scalable Vector Graphics (SVG)	SVG is XML based W3C open standard, is a vector image that has a number of advantages such as scalability. And most web browsers have supported SVG so far. It is suitable for long-term preservation of digital documents.	Already Compact	Most	Open	★★★★★
Tagged Image File Format (TIFF)	TIFF is not an open standard format, but a propriety image format that belongs to Adobe Company. TIFF file takes up a more space in a memory disk and more time when software read it, it is not suitable for long-term preservation	Lossy or Lossless	Limited	Proprietary	★★
Joint Photographic Experts Group (JPEG)	JPEG applies lossy compression technique, which results in losing some image data during the compression process. That makes JPEG not applicable for long-term archives preservation.	Lossy	100%	Standard	★
Portable Network Graphics (PNG)	PNG compression is one of the best that can be compression lossless and without paying patent fees. The main problem is lack of major supported companies. It is not suitable for preservation at this time, but it has bright future.	Lossless	Most	Open	★★★
Portable Document Format (PDF)	PDF has good strength, but also has obvious drawbacks. By comparing to SVG, it is not good enough. However, it may be a good alternative in the foreseeable future.	Lossy or Lossless	No	Standard	★★★

Table 3: Comparison of alternative image formats

Clearly, the proprietary format is not applicable for this project. Take jpeg for example, it is covered by a patent, and in 2002 the patent owner threatened to enforce the patent, so that anyone with a jpeg image would have had to pay a royalty [26]. This example illustrates that why not to use proprietary formats.

TIFF and PDF are quite popular to be the preservation formats at this time, but both of them have the same vital disadvantage that is format license. JPEG causes information loss in the images due to the compression algorithms, thus it is not applicable. PNG may be an acceptable option as well in the future, but the main problem is lack of major supported companies at present.

SVG is the best choice for long-term preservation of digital documents, by comparing five alternative image formats in terms of characteristics, format license, and web browser support and so on. It is an open standard, editable, text searchable and indexed, scalable without quality loss, etc. Apparently, SVG will have a promising future and bright perspective.

4.4 Conclusion

Much work has been done in research on long-term preservation of digital archives in this chapter. A comprehensive introduction of five potential future-proofing formats is conducted. Thought the comparison, SVG is chosen as the best choice for long-term preservation of digital archives.

Chapter 5 - TIFF to SVG format

5.1 Introduction

The SVG format has been proved to be the best format for permanent storage so far based on the previous investigation. As we said, a number of organizations and companies are using TIFF format for long-term preservation of digital archives. Thereby, it is necessary to seek for a way to convert the TIFF format to SVG format.

In this chapter, two possible ways of conversion from TIFF to SVG will be introduced for recommendation.

5.2 Ways of conversion

There are two types of graphics images, raster images (also known as bitmap images) and vector images, which has been reviewed in section 2.4. Images in the SVG format belong to vector images, while images in the TIFF formats fall into raster images.

Thus, methods capable of converting bitmap images to vector images are needed to be worked out. So far, there are two major appropriate conversion methods, namely, the tracing bitmap and the embedding bitmap into SVG file. Applying both of them, images in the TIFF format will be able to be converted into the SVG format. These two methods will be described in detail in the following part.

5.2.1 Tracing bitmap using "Vector Magic"

Vector Magic [27] is an application that allows you to convert bitmap images to clean vector images directly. It is a patent application and users can purchase and use it by two ways, online web and desktop application. Actually, other applications also have functionality of tracing bitmap to vector images, like Adobe Illustrator, CorelDraw and Inkscape, which will be introduced at length in next section. However, these applications will not be considered in this part, because the quality of generated result is a bit poor by the method of tracing bitmap.

Vector Magic is extremely easy to use, because the process of converting is automatic. It traces pixels of original bitmap first, and then draws it as a vector image by lines, polygons and so on. The produced image inherits all the features of vector images, such as scalability with lossless of quality, editability, with a smaller size than the raster image.

Here are two examples of comparisons between original bitmap images and the vectorized results produced by Vector Magic application.

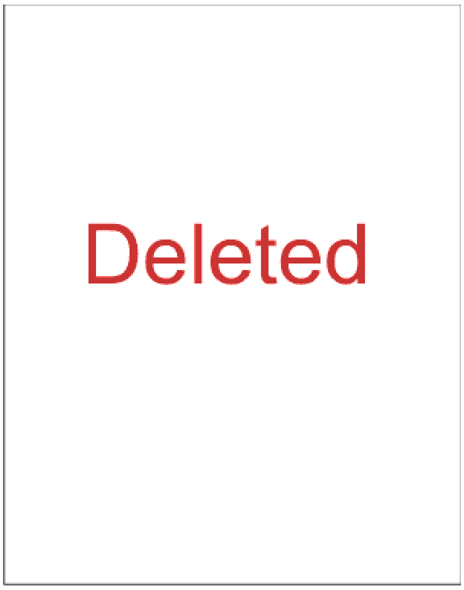

Figure 5-1: The original simple bitmap image [28]

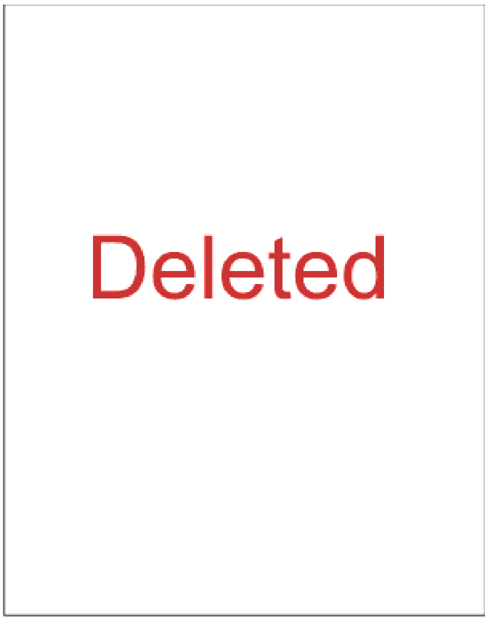

Figure 5-2: The vectorized image [28]

As shown in Figure 5-1, it is a simple bitmap image, which is in the TIFF format. The vectorized image in the SVG format is shown in Figure 5-2.

Although the vectorized image looks extremely similar as the original image, a bit difference would be found when zoomed in. As highlighted in each of the images, the produced vector image still has a bit blemish in terms of presenting the details.

As shown in Figure 5-3, the second example uses a more complicated bitmap image. As this image contains more presentation in the details of the image, the vectorized result has more obvious difference of the details.

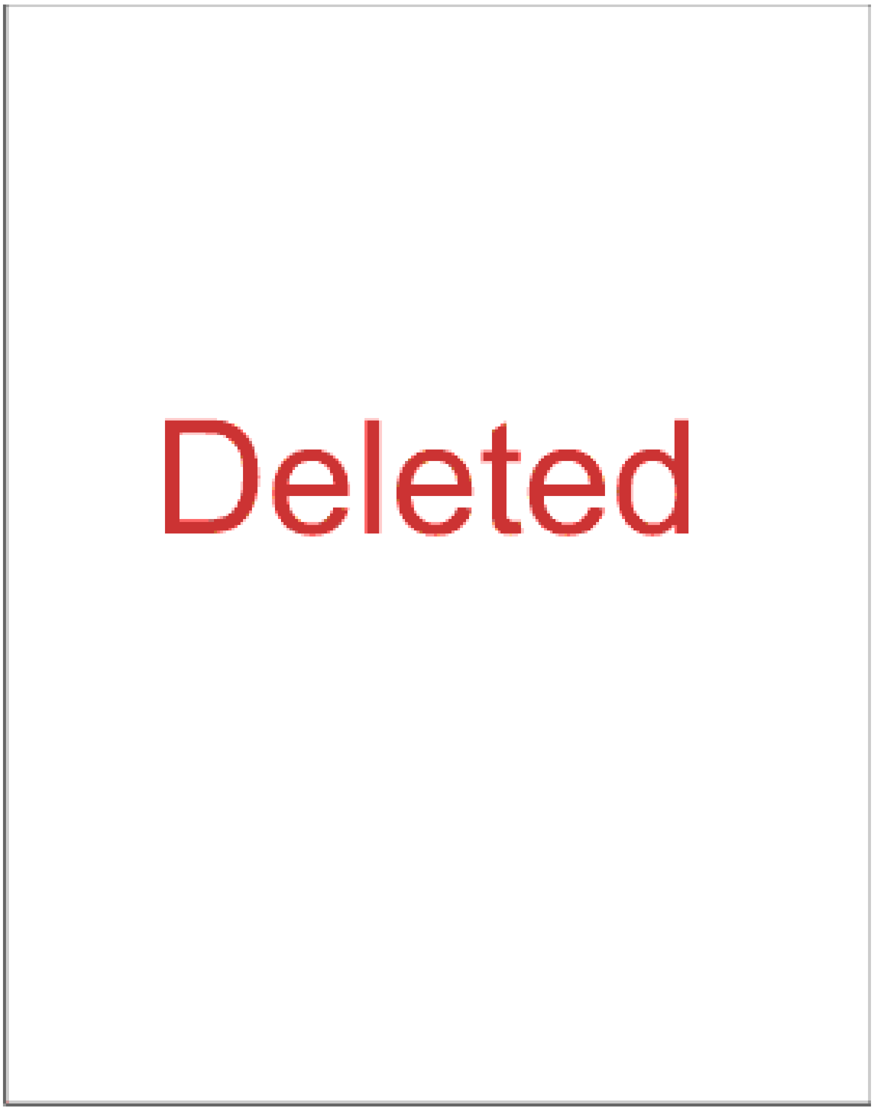

Figure 5-3: Left one is original bitmap image, right one is vectorized image [28]

To summarize, the Vector Magic application has the following advantages:

- It does a great job in converting bitmap images to clean vector images when tackling images without complicated details.
- It supports batch processing.
- The converting process is automatically performed.

The Vector Magic also has some drawbacks as shown below:

- Vector Magic is not a free application.
- Time consuming. It needs to take about 8 minutes to convert a 40 MB TIFF image to vector image.

In a word, Vector Magic performs very well in converting the TIFF to SVG so far, especially when tackling simple images. Furthermore, the technology of tracing bitmap is developing rapidly. It is a good option for converting bitmap images to vector images for the moment.

5.2.2 Embedding bitmap into SVG

Besides the method of tracing bitmap, there exists another feasible way to convert the TIFF to the SVG format, namely embedding bitmap into SVG files.

Images converted by embedding bitmap method have the <svg> tag, which is similar like SVG format files. Nevertheless, the content of the <svg> tag is not composed with tags like <path>, <circle> and <line>, but the Base64 code, which is encoded from original bitmap image. Actually, the created SVG file is a fake SVG file, because it losses some characteristics of vector images. For example, it cannot be scalable with lossless quality.

However, it still has some other features of vector image, such as smaller size and can be modified by text editor.

There are a number of applications can convert bitmap images to vector images so far. But among of them, just minority of applications can get acceptable results. Here are three outstanding applications, including Adobe Illustrator, CorelDraw and Inkscape.

Adobe Illustrator

Adobe Illustrator is a high-price graphics editor applicable for vector images, which is developed by Adobe Company, and it is a proprietary application with high price.

By contrast to other applications, Illustrator has some advantages:

- The size of created SVG file is so small that is good for long-term preservation.
- The generated SVG file possesses high quality.
- The created SVG file is supported by most of the mainstream browsers.

Illustrator supports converting the TIFF format to the SVG format by embedding bitmap into SVG, but the only applicable embedding bitmap is JPEG. As mentioned in chapter 4 that JPEG may not be rendered by browsers in the future. Hence, embedding JPEG into SVG is not suitable for long-term preservation. Additionally, Illustrator cannot support converting of large file (more than 100 MB) very well.

CorelDraw

Likewise, CorelDraw is also a priced vector graphics editor that is designed and marketed by Corel Corporation. CorelDraw is a commercial application, which was designed for professional and aspiring designers alike [29].

Similarly, CorelDraw can also support converting the TIFF format to the SVG format by embedding bitmap into SVG. However, CorelDraw is better than Illustrator in terms of various bitmaps embedding. It can be embedded three types of bitmap, JPEG, GIF and PNG. As mentioned in chapter 4, PNG is an open and stable image format, and it can be the alternative format for long-term preservation. Therefore, embedding PNG into SVG file is a good option for preservation of electronic archives at the present time.

Inkscape

Inkscape is an open source and vector graphics editor that possesses the similar capabilities as Illustrator and CorelDraw. However, similar with Illustrator, it is only applicable for embedding JPEG into SVG file. Moreover, it cannot support large file as well.

Inkscape is an application with free of charge, but it is not a good choice for converting the TIFF format to the SVG format using the method of embedding bitmap.

Although the embedding bitmap is also a possible approach to convert the TIFF to the SVG, a key point is that to use embedding bitmap loses the advantage of using SVG. In other works, if you have an embedded Jpeg, then you need a jpeg interpreter to render it.

5.3 Conclusion

There are two major appropriate conversion methods available now, tracing bitmap and embedding bitmap into SVG file. In this chapter, tracing bitmap technology introduces the Vector Magic application for converting the TIFF format to the SVG format. And embedding bitmap technology takes advantage of three existed converting software, Adobe Illustrator, CorelDraw and Inkscape.

Comparing to embedding bitmap, tracing bitmap is more suitable for preservation of electronic documents at the moment as embedding bitmap loses the advantage of using SVG.

Chapter 6 – Experiments, Results & Discussions

6.1 Introduction

In this chapter, three experiments will be conducted focusing on conversion from TIFF images to SVG images by using either the tracing bitmap or embedding bitmap. The experiments are conducted by Vector Magic, Adobe Illustrator, CorelDraw and Inkscape. The performance and quality of the output SVG files will be evaluated.

6.2 Hardware and software environment

Before conducting these experiences, the required applications must be installed. Vector Magic v1.14 and CorelDraw X4 are purchased by the department of Computer Science and have been installed in the departmental labs.

Tested hardware configuration as follows:

CPU: Intel P4, 3.4 GHz
RAM: 1-GB Memory
Hard disk: 320-GB hard drive

Tested software configuration as follows:

Windows XP (32bit)
Vector Magic v1.14 (desktop application)
Adobe Illustrator CS 5(trial version)
CorelDraw X4
Inscape v0.48 (open source)

6.3 Experiments

Experiment #1: Convert a TIFF file to a SVG file by using tracing software

In this experiment, a TIFF (raster image) image file is converted to real SVG (vector image) image files by using tracing technique of various applications separately, including Vector Magic, Illustrator CS5, CorelDraw and Inkscape. The "real" means that the created SVG files inherit all the attributes from the SVG file such as scalability, editability and so on. Figure 6-1shows the results of this experiment.

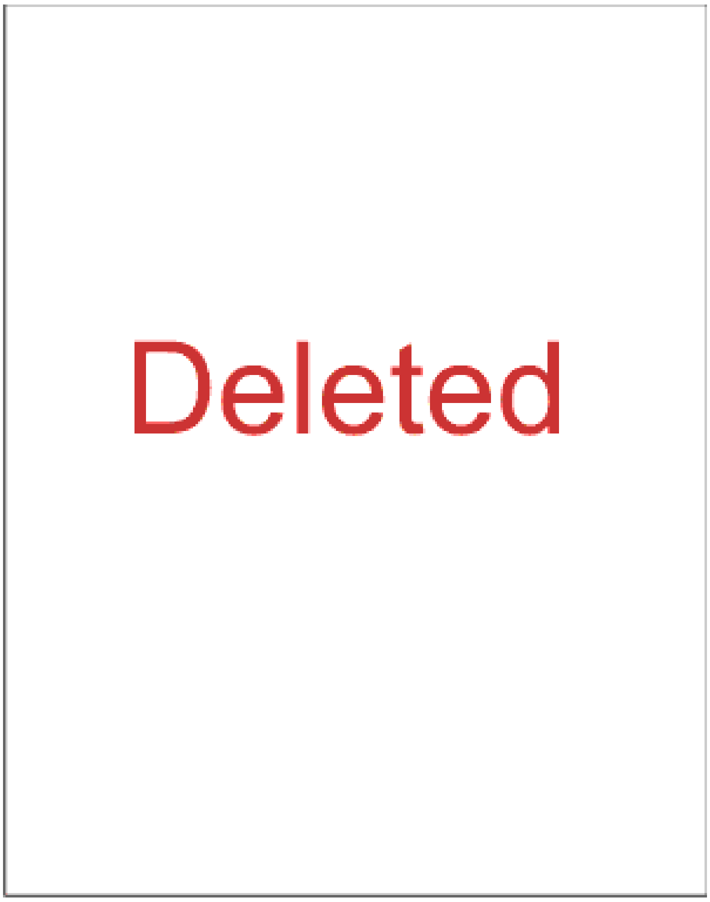

Figure 6-1: the results of experiment 1 by using tracing software

From the outcomes of these different tracing programs, Vector Magic, Adobe Illustrator CS5, CorelDraw X4 and Inkscape, it is clear that the SVG file, which produced by Vector Magic, has the best shape and detail of the bitmap original. In order to give an intuitive view of these results, Table 4 summaries the important factors of each tracing software, which includes size of the converted images, length of the time during the processing, accuracy level of the converted images and also the level of ease of use. The ease of use is based on how many steps to convert the TIFF to SVG. There are three levels, easy, medium and difficult. The criterion of each level is shown as follows:

- Easy: Open TIFF files-> converting-> Save SGV files
- Medium: Open TIFF files->manual settings->converting-> Save SGV files
- Difficult: pre-setting (e.g. environment configuration, required plug-in and tools installing)->Open TIFF files->manual settings->converting-> Save SGV files

	Size (KB)	Time (S)	Accuracy	Ease of Use
Vector Magic	13.6	1.0	Very high	Easy
Adobe Illustrator CS5	8.2	0.3	High	Medium
CorelDraw X4	10.6	0.5	Low	Medium
Inkscape	15.4	0.5	Low	Medium

Table 4: Detailed data of the produced files (size of original TIFF image is 78 KB)

Compare to the other tracing tools, Vector Magic may take a little longer to process any individual image. However, it is more probably to produce high accuracy of results than others. In addition, Vector magic is extremely easy to use by several clicks to create passable result.

Experiment #2: Convert a set of TIFF files to SVG files by using Vector Magic

In this experiment, some TIFF files will be converted to SVG files by Vector Magic with medium quality. The experiment is used to evaluate performance and stability of Vector Magic by processing a set of files with a wide range of file sizes. The Table 5 shows a part of results of experiment 2. Some samples from the RAI are selected to make comparisons.

	Original size (MB)	New size (MB)	Time (Min)
Sample 1	29.6	2.68	1.9
Sample 2	34.5	4.56	5.5
Sample 3	36.2	3.44	4.8
Sample 4	36.3	5.61	6.0
Sample 5	36.5	0.81	3.5
Sample 6	38.5	0.61	3.1
Sample 7	38.6	0.64	3.5
Sample 8	38.8	8.18	13.1
Sample 9	40.3	0.76	4.0
Sample 10	40.3	0.64	4.4
Sample 11	42.8	0.5	4.6
Sample 12	50.9	3.37	7.2
Sample 13	169	2.88	8.3

Table 5: The results of experiment 2

According to the experience, it proved that Vector Magic can support a wide range of file sizes, even huge TIFF files (169 MB). Generally, the bigger the size of the file is, the longer the converting takes. However, the converting time depends on two factors, the size of the image and the detailed colors level of image.

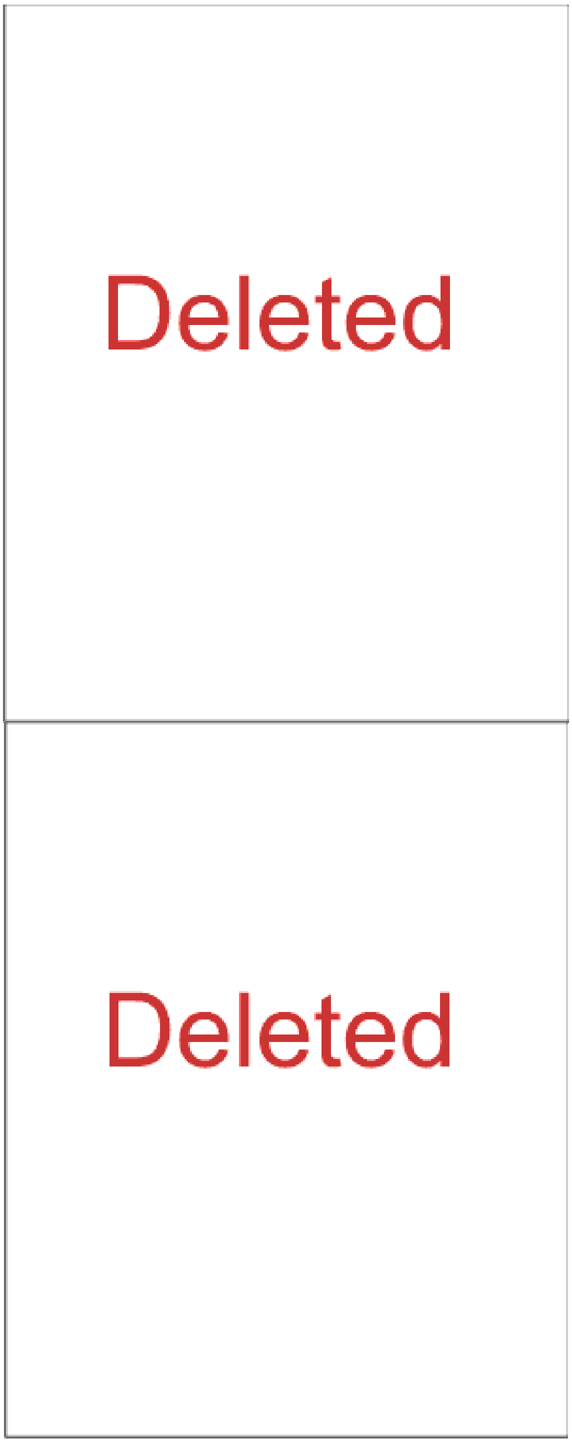

Figure 6-2: Comparison of original raster image and produced vector image [28]

Figure 6-2 (one of examples) shows the difference between the original and the created images. The top one is an original TIFF image. The bottom one is the created SVG file. The above figure illustrates how well Vector Magic traces the bitmap original. In fact, it is very difficult to make a distinction between them if they are not zoomed in a lot.

There is another example to present the superiority of vector images. As shown in Figure 6-3, the original TIFF file becomes blurry and causes the loss of quality after zoomed in. However, the generated SVG file, as illustrated in Figure 6-4, is still in a high quality after scaled up. This also reveals that the created SVG file is a clean SVG file.

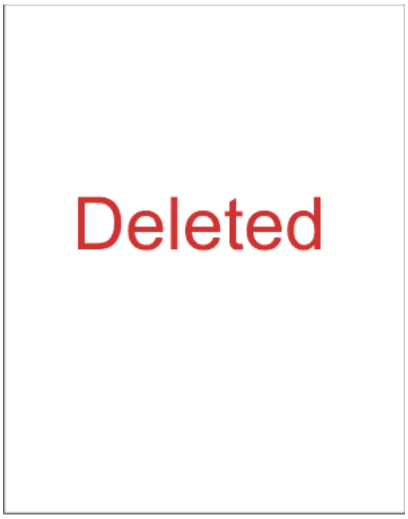

Figure 6-3: The original TIFF file is zoomed in times

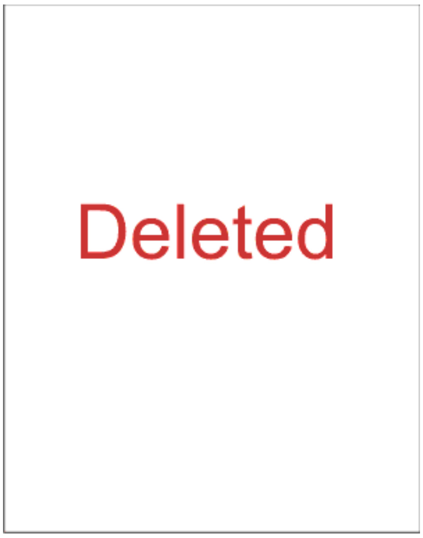

Figure 6-4: The created SVG file is zoomed in times

Experiment #3: Convert TIFF files to SVG files by embedding bitmap into SVG

Adobe Illustrator, CorelDraw and Inkscape will be used to convert TIFF format to SVG format by embedding bitmap into SVG file in this experiment. The aim of this experiment is to evaluate and compare performance of above three applications on processing a TIFF file (original size is 29.6 MB). Table 6 shows the results of comparison.

	Size (MB)	Time (S)	Support big file	Ease of Use
Adobe Illustrator CS5	0.46	0.8	No	Easy
CorelDraw X4	1.88	2.0	Yes	Easy
Inkscape	0.77	1.1	No	Medium

Table 6: The results of experiment 3

Compare to other two applications, CorelDraw X4 has some advantages. For example, X4 can support big image file (more than 100MB). In addition, it also can embed three types of bitmap into SVG file such as PNG, JPG and GIF. As mentioned in chapter 4, PNG image format has been widely used on the web so far, and it will be continually supported by W3C in the future. This implies that CorelDraw X4 is better than Illustrator CS5 and Inkscape for converting TIFF file to SVG file by using embedding bitmap.

However, there are also some disadvantages of CorelDraw X4. For example, it may take longer to produce SVG file by contrast to others. Moreover, the size of created SVG file is bigger than Illustrator CS5 and Inkscape.

6.4 Conclusion

Through the three experiments, it is clear to understand both advantages and disadvantages of each application by comparison of output data when applications are being using to convert TIFF files to SVG files.

Clearly, the best option for the former is using Vector Magic to trace bitmap into vector image at this time. For the latter, CorelDraw is a good choice for converting TIFF files to SVG files by the method of embedding bitmap.

Comparing these two possible conversion methods, the tracing bitmap is better to be used to convert the TIFF to SVG than the embedding bitmap based on above experiments. There are two major reasons for it. Firstly, the files created by tracing bitmap are clean SVG files, but embedding bitmap is not. Secondly, the size of generated files by tracing bitmap is smaller than embedding bitmap.

Chapter 7 – Preservation Approaches

7.1 Introduction

This chapter describes some common preservation approaches and one of them will be chosen for data migration in the cooperative project with the RAI. A separate report will be created as guidelines for the RAI institution to help them establish a complete preservation system of digital documents.

7.2 Technical Preservation Approaches

There are a number of digital preservation approaches for long-term purpose that has been proposed by individuals and organizations so far (see Figure 7-1). However, no definitive strategy has been adopted to solve the problem of preservation of digital documents. Moreover, information on likely potential problems, cost and accessibility in the future of various approaches is still extremely limited at the present time.

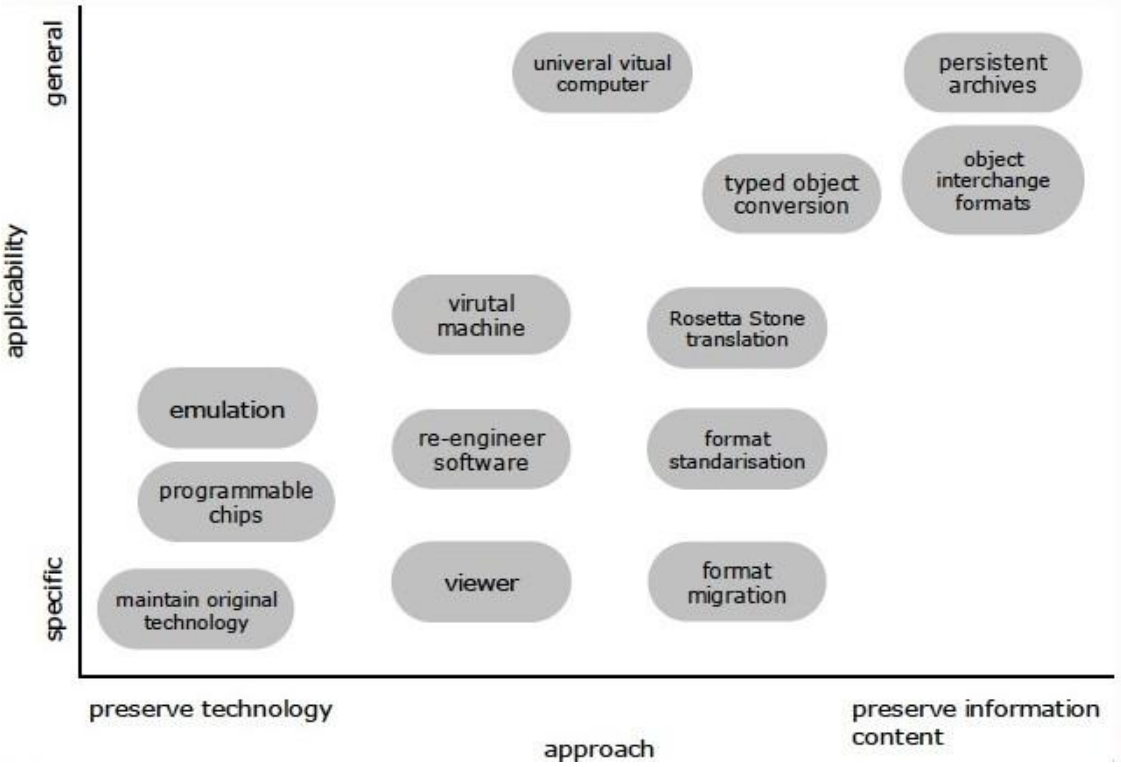

Figure 7-1: Digital preservation approaches [30]

In this chapter, some possible alternatives will be selected to make a comparison, including hardware museums, migration, universal virtual computer and emulation. Then, one of these preservation approaches will be chosen to apply to this project.

7.2.1 Hardware museums

The approach of collecting and maintaining all relevant original hardware and software that used to access old digital archives is called hardware museums [31]. By hardware museums, future generations can view previous documents in the original computer environments. See Figure 7-2, it is a picture of a history computer museum in Mountain View, California.

Figure 7-2: A computer history museum [32]

Unfortunately, this preservation approach has been proven that it is not applicable for long-term preservation of digital documents in practice.

There are some reasons for it.

- Huge cost - Require a large number of older machines and corresponding software. This means that a great deal of funds needs to be invested on hardware and software.

- Knowledge need - Future generations, who want to use these machines and software, need a lot of understanding of the preserved systems.
- Hard to maintain – Numerous parts of old hardware will not be possible to be replaced in the generations.
- Too many items – It is impractical to collect all kinds of computers all over the world. This implies that it can only read certain digital documents.

In fact, experts from the digital preservation community also agree that hardware museum is not suitable for long-term preservation because of huge cost and hard maintainability.

7.2.2 Migration

Generally speaking, migration can be divided into two sub-migration, format migration and media migration.

The process of converting a digital data from an original file format into a new format that remains fully accessible and functional is called format migration [33]. For example, convert Microsoft word to Adobe's Portable Document Format (PDF) By migration technique, old digital data can still be accessed completely in current system environments without maintain contemporary hardware and software. Figure 7-3 illustrates a simple process of file format migration.

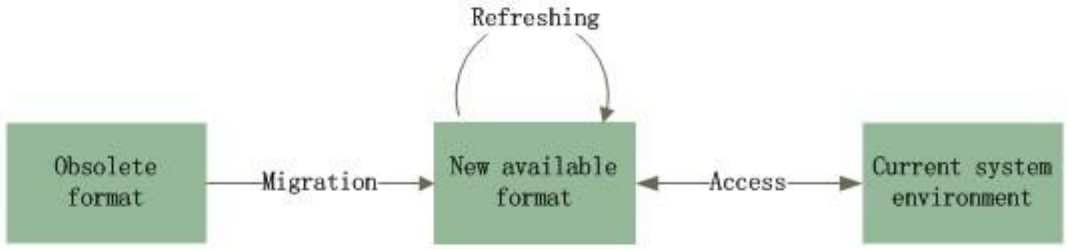

Figure 7-3: the simple process of format migration

Media migration, often is called data refreshment, is used to ensure that data remains forever by media degradation over time. For media migration, the exact lifetime of media should be estimated after a media migration. Then, another migration will be performed before the threshold is reached.

Actually, migration is a well-known preservation strategy for long-term preservation of digital archives and is used almost everywhere in IT industry at the present time. It has been proven that migration is a saved and efficient preservation method, by contrast to others. In addition, there is a great deal of knowledge about the migration technique; a set of tools can be used to do migration.

Using migration for long-term preservation has some advantages as follows.

- Migration always happens on a basic premise, which is that an obsolete file format needs to be converted into a new available format in order to being accessed in the future by generations in the simplest possible way. Hence, this implies that migration can make data be accessed over immutability.
- Using migration for preservation can reduce cost of preservation without maintaining the system environment all the time for original files.
- The quality of documents potentially may be improved by migration. Because the new format need to satisfy current quality standards after migration. There is no doubt that the quality of future format will be improved step by step.

There are also some disadvantages of migration as follows.

- Only simple migration can be automated completely. Therefore, it will cost a great deal of time to migrate a large number of files without automation.
- Some properties of the original files may be lost without sufficient understanding of original and target file format when migration is being processing.
- Authenticity of file data may be reduced after too many times migration.

Numerous experts agree that migration preservation is an obvious and suitable candidate for long-term preservation of digital documents according to its characters at the present time.

7.2.3 Emulation

Emulation is the process of simulating an older system environment on current system, including emulating applications, hardware platforms and operating systems. For example, Writing a windows application to run on a Mac OS by emulation. The emulation approach was proposed to use for long-term preservation of digital archives by Jeff Rothenberg (1999), who provides a diagram (see Figure7-4) that illustrates the encapsulation in an emulating system [34].

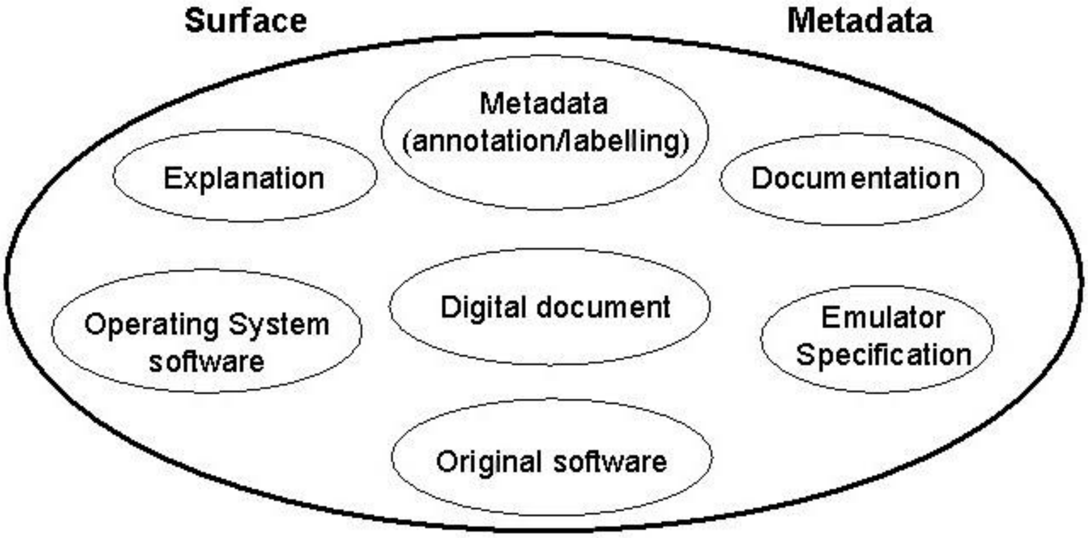

Figure 7-4: The encapsulation in an emulator [35]

Emulation is a strategy that keeps digital data in its original state but create a simulating system environment to read the original data on current computers. This means that using emulation will maintain the exact “look and feel” of the data and not lose authenticity of the data.

The following steps are showing how emulation works.

1. An emulated system environment encapsulated a contemporary access environment for the original data, including applications, operating systems and hardware configuration.

2. The emulated system environment should be accessed based on current software and hardware platforms.

3. The emulated system environment can access the target data in the complete software of the rendition system.

Comparing to migration approach, emulation is more complex and technical. Nevertheless, a number of people advocate using emulation for long-term preservation of digital documents. Here are some reasons for it.

- Using emulation will not lose authenticity and "look and feel" of the original data because the file format does not need to be changed.
- Only one emulator needs be created if there is a large number of same format documents, which can reduce time and cost.

There are also some shortcomings of emulation as follows.

- As a preservation method, the process of emulation is so complex that needs more technical support.
- It is extremely difficult to get a precise old hardware specification (that may contain screen size, CPU clock speed and resolution, etc.) for creating an emulating system environment.

Actually, whether emulation approach is suitable for long-term preservation of digital documents is still an open question now. There are still some pending questions need to be solved before an agreement sorted out. Therefore, emulation just can be a potential preservation strategy for long-term preservation of digital archives.

7.2.4 Universal Virtual Computer

UVC, short for Universal Virtual Computer, is a new and viable preservation approach based on combination of migration and emulation, which was proposed by Lorie of IBM (2000). UVC is a preservation method of platform-independent. This means that it can be applied to run any software in the foreseeable future, even in a yet unknown platform [36].

Lorie described the UVC as "Universal" because of its basic definition. He said that the approach would last forever without any disadvantages. "Virtual" because it is a virtual machine. In addition, it will not be physically built and it is a "computer" in its functionality.

Actually, the UVC was designed as a substitute of migration and emulation approaches. On one hand, UVC inherits all the advantages of them. For example, it allows the file data keeps in its original state without any change. On another hand, it also takes out disadvantages of migration and emulation, such as loss of data authenticity.

The UVC system (see Figure 7-5) is composed of four different components, including UVC program, UVC interface, UVC and logical data schema (LDS) with information type description. In the UVC system, UVC program also called format decoder. UVC interface, namely logical data viewer, contains two items such as UVC interpreter and restoration program [37].

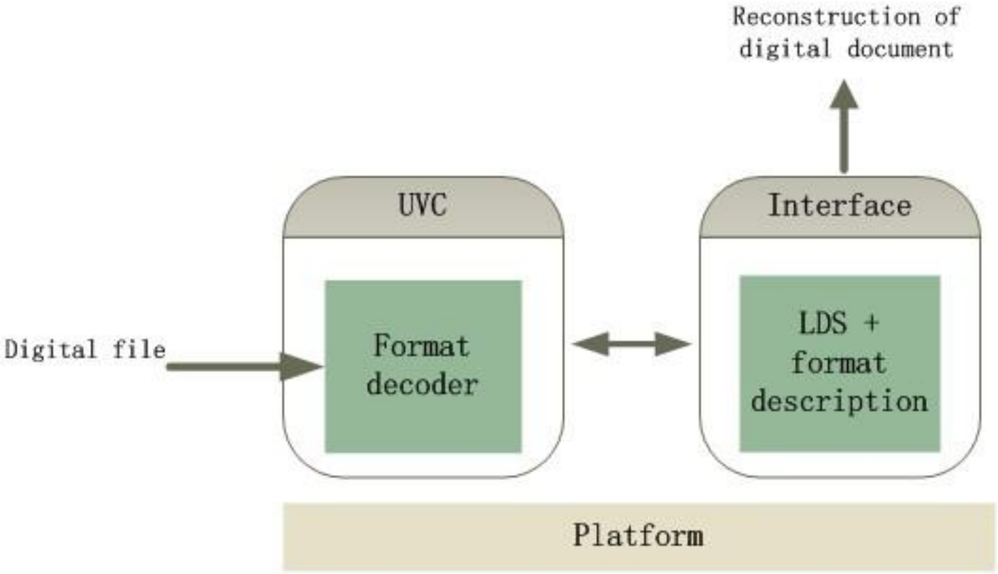

Figure 7-5: The UVC system and components

The process of UVC system is that format decoder, which runs on UVC, decodes the input digital file into some element tags. These elements build new data, which is very similar to XML file. Finally, a viewer inside UVC system can read the created data.

The UVC approach is attractive to developers for the reasons as follows.

- The data for preservation will keep its original state with unchanged.
- There is no need to emulate real hardware environments.
- UVC is platform-independent and can be run on any platform in the future.

It seems that the UVC strategy is a perfect preservation approach for long-term preservation of digital archives. However, it has not been used widely so far, because the UVC method is still in an experimental phase. Hence, no statements about required time and cost have been made.

7.3 Approach applied in this project

One objective of this project is to create a separate report which will be guidelines and a user manual for helping the RAI establish a complete preservation system of digital archives for long-term.

Based on the situation of the RAI and the comparisons of various preservation approaches, migration method is more suitable for long-term preservation than other preservation approaches in this project. After all, comparing to emulation, migration does not need too much professional knowledge but some converting tools for file format migration and simple operations.

It has been mentioned that migration approach can be divided into two sub-migration, format migration and media migration. Issues about format migration have been discussed in chapter 5. The TIFF, which may become an obsolete format in the future, has been converted to a vector image format, SVG. The SVG format is an open format and is becoming more and more popular nowadays. And the SVG has been proven that it is suitable for long-term preservation of digital documents.

However, the detailed approaches of media migration and production of advice on data management will be described in the report for the RAI (See the report in appendix). This report will also illustrate the rationale behind the claim for future-proofing. In addition, advice on other aspects of preservation will be given finally, such as database system for managing the produced SVG files.

7.4 Conclusion

By comparing various preservation approaches, we find out that both migration and emulation are more suitable for long-term preservation of digital documents at the present time.

Hardware museum has been proven that it is not impossible to be an option for long-term preservation. And the Universal Virtual Computer is still in an experimental phase. It is still not mature enough for digital preservation.

The migration and emulation approaches have some limitations such as “look and feel” and authenticity of original data, but these two approaches have been applied to digital preservation by the digital preservation community and have been proven to work so far.

Chapter 8 – Conclusion

8.1 Introduction

This chapter summaries the format research for long-term preservation of digital archives as well as the recommended strategy in file storage for the ARI. Firstly, the achievement of the research will be described, followed by the discussion of limitations of the project. Finally, recommendation based on current research will be illustrated.

8.2 Achievement

Searching for the format for future-proofing archives is a long-term study. Due to limited time, the achievement of the project so far has been listed as below. There are three aspects of this project that has been fulfilled completely:

- The SVG is recommended as the most likely and robust digital format for long-term preservation of digital archives, which is a type of the vector image formats.

- Two feasible approaches of converting the TIFF to the SVG have been worked out, the tracing bitmap and the embedding bitmap. The former takes advantage of the application of Vector Magic to trace TIFF to clean SVG. The latter converts TIFF to SVG by using Adobe Illustrator, CorelDraw and Inkscape. Tracing bitmap by Vector Magic is the recommended approach derived from the research.

- A separate report for the RAI. The report has been created as a user manual and guidelines to be presented to the RAI institution to help them establish a complete preservation system of digital archives.

8.3 Limitation

Although all objectives of this project have been implemented so far, there are still some limitations for long-term preservation of digital archives as follows:

- Although Vector Magic performs very well in converting the TIFF to SVG so far, it still needs to be improved to make the results more perfect in the future.

- Current storage media is still too vulnerable and unstable to be used for long-term preservation. Moreover, the lifetime of these media is so short that we need to refresh them regularly to keep data safe.

8.4 Recommendation

A complete preservation system of electronic documents may appear a simple task, but it does comprise a lot of follow-on work such as future access and readability.

Based on our experience of this project, the following recommendations are given:

- The tracing bitmap method is better to be used for converting TIFF to SVG, because embedding bitmap loses the advantage of using SVG. Furthermore, if you have an embedded Jpeg, then you need a jpeg interpreter to render it.

- Browsers can be used for accessing and reading the generated SVG files now and future. Opera is the best one nowadays in terms of supporting aspect.

- The OAIS reference model is designed to clarify the very fundamental terms of long-term preservation and identify the processes required in the lifetime of implementing preserving digital documents. It has become an international standard for preservation of digital archives. Thereby, the whole preservation system should base on it. (See the report in appendix)

Reference

[1] Alistair Edwards. (2010) "Project with the Royal Anthropological Institute notes on the meeting on 24 June", University of York, June 2010.

[2] Alistair Edwards. (2010) "Project Proposals 2010-11" Available: http://www-users.cs.york.ac.uk/~alistair/projects2010.html. Last accessed July 2010.

[3] Michael Day. (2006) "The OAIS Reference Model" Reference Models meeting, One Great George Street, London, 25 January 2006

[4] Alex Ball. (2006) "Briefing Paper: the OAIS Reference Model" UKOLN, University of Bath, Page 18.

[5] Farquhar, A., Hockx-Yu, H (2007) "Planets: Integrated Services for Digital Preservation", International Journal of Digital Curation, Vol. 2, No. 2 (2007).

[6] NEDLIB web-site (2010) "The NEDLIB project" Available: http://www.konbib.nl/nedlib/. Last accessed August 2010.

[7]Authenticity Task Force (2000) "InterPARES Project: Research Methodology Statement" International Reasearch on Permanent Authentic Records in Electronic Systems, DRAFT, November 7, 2000.

[8] Holzner, S. (2001) "Inside XML Indianapolis", Ind. New Riders, Page 9-11.

[9] Ben, Hammersley (2005). "What Is Atom"

http://www.xml.com/pub/a/2005/10/26/what-is-atom.html. Last accessed September, 2010.

[10] W3C Recommendation (2007). "SOAP Version 1.2 specification"

http://www.w3.org/TR/soap12. Second Edition, 27 April 2007.

[11] "All about RSS" http://www.faganfinder.com/search/rss.php. February 19, 2004.

[12] Apple (2009). http://www.apple.com/asia/iwork/. Last accessed September, 2010.

[13] "Language localization status". OpenOffice Language Localization Project, http://wiki.services.openoffice.org/wiki/Languages. Last accessed September, 2010.

[14] Microsoft (2005) "Q&A: Microsoft Co-Sponsors Submission of Office Open XML Document Formats to Ecma International for Standardization". https://www.microsoft.com/presspass/features/2005/nov05/11-21Ecma.mspx. Last accessed September, 2010.

[15]Harold, E. R. & Means, W. S (2002) "XML in a nutshell: a desktop quick reference" Cambridge, O'Reilly.

[16] Sue Chastain. (2010) "Vector and Bitmap Images" About.com Guide, http://graphicssoft.about.com/od/aboutgraphics/a/bitmapvector.htm. Last accessed August 2010.

[17]Sketchpad.net. (2009) "Two Kinds of Computer Graphics" Available: http://www.sketchpad.net/basics1.htm. Last accessed August 2010.

[18] Wikipedia, "Scalable Vector Graphics" Available: http://en.wikipedia.org/wiki/Scalable_Vector_Graphics. Last accessed July 2010.

[19] Titia van der Werf-Davelaar (1999) "Long-term Preservation of Electronic Publications" D-Lib Magazine, ISSN 1082-9873, Volume 5 Number 9, September 1999.

[20] Wikipedia, "Olympic_flag" Available: http://en.wikipedia.org/wiki/File:Olympic_flag.svg Last accessed July 2010.

[21] Adobe Developers Association (1992). "TIFF Revision 6.0 specification"

http://www.adobe.com/Support/TechNotes.html. June 3, 1992

[22] S. Anderson, M. Pringle, M.Eadie, T.Austin and A. Wilson, "Digital Images Archiving Study", the Computer Journal, Final version, PP 56-60 March 2006.

[23] Chris, Lilley (1996). http://www.w3.org/Graphics/JPEG/.

[24] W3C Recommendation. “Portable Network Graphics (PNG) Specification” Second Edition, ISO/IEC 15948:2003 (E), 10 November 2003.

[25] Adobe Developer Connection (2009) “PDF Reference and Adobe Extensions to the PDF Specification”http://www.adobe.com/devnet/pdf/pdf_reference.html. Last accessed September, 2010.

[26] Wikipedia. http://en.wikipedia.org/wiki/Jpeg#Patent_issues. Last accessed September, 2010.

[27] James, Diebel. Jacob, Norda (2007). “What Is Vector Magic?” http://vectormagic.com/home/about. Last accessed September, 2010.

[28] The RAI. “The RAI collection of photographs”, June 20, 2006.

[29] Troidl, David. "SVG – From CorelDraw to Your Browser". Graphics Unleashed. http://www.unleash.com/davidt/svg/index.asp. Last accessed August 2010.

[30] Thibodeau K. (2002). “Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years” in proceedings of The State of Digital Preservation: An International Perspective，Conference Proceedings. Washington. 2002.

[31]Borghoff, U. M., P. Rödig, et al. (2006). Long-Term Preservation of Digital documents : Principles and Practices. Berlin, Heidelberg Springer-Verlag

[32] http://prisonphotography.wordpress.com/2009/05/12/museum-self-promotion-moma-vs-sfmoma/Images Computer History Museum by David Glover. Last accessed August 2010.

[33] Garrett, J (1996). "Preserving digital information: Report of the task force on archiving of digital information" http://www.rlg.org/legacy/ftpd/pub/archtf/final-report.pdf Last accessed August 2010.

[34] Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation, Jeff Rothenberg, 1999. Available at: http://www.clir.org/pubs/reports/rothenberg/contents.html

[35] Stewart Granger (2000). ”Emulation as a Digital Preservation Strategy” University of Leeds, D-Lib Magazine, ISSN 1082-9873, Volume 6 Number 10 Available at: http://www.dlib.org/dlib/october00/granger/10granger.html#fig2

[36] Lorie, Raymond A. (2001). "Long Term Preservation of Digital Information", Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'01) Roanoke, Virginia, USA. pp. 346–352.

[37] Wikipedia, “UVC-based_preservation” Available: http://en.wikipedia.org/wiki/UVC-based_preservation. Last accessed August 2010.

Appendix

Abbreviation

TIFF	Tagged Image File Format
PDF	Portable Document Format
JPEG	Joint Photographic Experts Group
SVG	Scalable Vector Graphics
BMP	Bit Map Picture
GIF	Graphic Interchange Format
PNG	Portable Network Graphics
RAI	Royal Anthropological Institute
OAIS	Open Archival Information System
ISO	International Organization for Standardization
CCSDS	Consultative Committee for Space Data Systems
PDI	Preservation Description Information
SIP	Submission Information Package
AIP	Archival Information Package
DIP	Dissemination Information Package
NEDLIB	Networked European Deposit Library
XML	Extensible Markup Language
SGML	Standard Generalized Markup Language
XHTML	Extensible Hypertext Markup Language
SOAP	Simple Object Access Protocol
RSS	Really Simple Syndication
DSEP	Deposit Systems for Electronic Publications
IE	Internet Explorer
W3C	World Wide Web Consortium
LZW	Lempel-Ziv-Welch
UVC	Universal Virtual Computer

Table 7: Abbreviations for terminologies

A draft report for the RAI

Introduction

This report aims to offer advice about establishing a complete preservation system of digital archives for your institution, including the recommended format and software, data (digital archives) migration, files accessing and data management.

Background Study of the RAI

As I know, your institution is a small and longest-established institution specialized in furtherance of anthropology (the study of humankind), which is a non-profit registered charity and is entirely independent, accountable to the Council straight.

There are a larger number of archives and manuscripts collections in your institution so far. These artifacts are digitized to either the TIFF or JPEG format in order to prevent them from natural calamities and obsolescence, and the generated files are preserved and maintained by proprietary database software, called FileMaker.

Data Migration

The process of data migration can be divided into two sub-migrations, format migration and media migration.

For archives preservation, format migration refers to the process of converting the digital archives from one format to the other format. The objective is to convert these electronic documents into a better format that will be more robust for long-term preservation. Most documents are digitized to the TIFF format in your institution; In fact, the TIFF format is near to obsolescent. However, the SVG format is recommended as the most likely file format for long-term preservation. There are two feasible conversion approaches, namely, the tracing bitmap and the embedding bitmap into SVG file. Comparing to embedding bitmap, tracing bitmap is more suitable for preservation of electronic documents at the moment as embedding bitmap loses the advantage of using SVG. The

Vector Magic software is recommended for converting the TIFF to the SVG.

Media migration is also called data refreshment, which is used to protect data from media degradation. At the present time, there has some common stable media with a very limited life span for data preservation such as, tapes, optical discs, hard disk etc. At the present time, hard disk is the best option for preserving digital data by comparing to optical discs, because a corresponding driver needed to access data if data is stored in optical discs. Relevant advice on media migration as follows:

- Review digital data in hard disk every year.
- Refresh hard disk every three years.
- Make two more backups in different places.

Data Management

Actually, the current database index to digital files is not robust enough by applying the proprietary application FileMaker for long-term. The recommended solution is applying a web-based XML database system, which is more efficient and easier to search the specified data. The database system should be built based on XML and JavaScript technology. By the database, users can use keyword to search the content of generated SVG files and download the required files through the Internet all over the world.

File Accessing

The digital documents are recommended to be preserved in the web-based XML database system in the SVG format for long-term storage. At the present time, the SVG format has been well supported and accessible directly by most mainstream web browsers, such as Opera, Safari, Google Chrome, Firefox, etc. Clearly, browser can be used for accessing the files in the future, because it is not possible to be out of date.

Conclusion

To summary up, all recommendations for your institution will be clearly illustrated as follows:

- **Data Migration**
  - Format migration: converting TIFF documents to SVG documents by using CorelDraw software.
  - Media migration: Hard Disk. Relevant advice as follows:
    - Review digital data in hard disk every year
    - Refresh hard disk every three years
    - Make two more backups in different places
- **Data Management:** conducting web-based XML database system
- **File Accessing:** using mainstream web browsers, such as Opera, Safari, Google Chrome, Firefox and so on