Chapter 10
Digital File Formats for Digital Preservation

Mayank Yuvaraj
Central University of South Bihar, India

ABSTRACT

During the course of planning an institutional repository, digital library collections or digital preservation service it is inevitable to draft file format policies in order to ensure long term digital preservation, its accessibility and compatibility. Sincere efforts have been made to encourage the adoption of standard formats yet the digital preservation policies vary from library to library. The present paper is based against this background to present the digital preservation community with a common understanding of the common file formats used in the digital libraries or institutional repositories. The paper discusses both open and proprietary file formats for several media.

INTRODUCTION

In the process of digital preservation, file formats play very important role. These are the standards which incorporate the quality issues relating to information storage, file size, type of data compression and also information retrieval. Since we are moving towards 'Paperless society', the functions and strategies for achieving the basic objectives have been changed in modern Library and Information Centre. The same principal applies to the managers of archive drawings and photographic records, who have the facilities for the storage of their archives in digital format. The conversion of the original hard copy records into digital records is known as digitization. In the process of digital preservation, it is primarily essential to examine basic issues to achieve the required end product. For line drawings the question needs to be asked as to whether the drawing is likely to be used and modified or is retained as an archive. If it belongs to an archival collection, the drawing is scanned and stored as a digital file on CD-ROM or DVD. If the drawing is to be used as a working drawing on a Computer Aided Design (CAD) system, then the scanned data needs to be converted into vector data in an intelligent form i.e. in a way that replicates the same drawing if it had been produced originally on CAD.

DOI: 10.4018/978-1-5225-6921-3.ch010

Photographs are also scanned using the original negatives or transparencies if possible, if not then photographic prints. Old and damaged photographs can be cleaned and digitally repaired.

Three major issues related to storage of Digital information:

- The format of the digital data and any compression has a major effecting file size and a minor effect on quality.
- The resolution of the scanning is a balance between quality and file size.
- The storage medium can have an effect on retrieval times.

As computer storage is becoming less expensive, it supports to move towards higher standards, i.e. higher resolution. But it should be remembered that a relatively small increase in the resolution results in a proportionately high increase in file size.

Several raster-based file formats are available to store digitized information. Several new image file formats are also emerging. Many are limited to particular applications, such as digital capture devices or image manipulation programs. Others are destined for wider use, with their developers intending them to become official or de facto standards. It is a long and difficult process to create such standard formats and longer still for them to become widely used and supported (Ingels, 1976).

Presently we have mainly six formats, which may be divided into two categories:

- **Open (Non-Proprietary) Formats:** TIFF, PNG, GIF and JPEG etc.
- **Close (Proprietary) Formats:** MrSID, DjVu, Genuine Fractals and Pixel Live/ VFZoom etc.

Detailed description of some of the prominent file formats is as follows:

OPEN (NON-PROPRIETARY) FORMATS

Any digitization project will need to consider the long-term usefulness and accessibility of the images and this means choosing a file that is both an established industry ‘standard’ as well as a non-proprietary format. Some of these formats are discussed below:

Tagged Image File Format (TIFF)

The first version of the TIFF specification was published by Aldus Corporation in the fall of 1986. TIFF describes image data that typically comes from scanners, frame grabbers, and paint- and photo-retouching programs. TIFF is not a printer language or page description language. The purpose of TIFF is to describe and store raster image data. A primary goal of TIFF is to provide a rich environment within which applications can exchange image data. This richness is required to take advantage of the varying capabilities of scanners and other imaging devices. Though TIFF is a rich format, it can easily be used for simple scanners and applications as well because the number of required fields is small.

By the mid-nineties there was some discussion about replacing CompuServe’s GIF format, but the immediate prompt for PNG’s development was a patent dispute. Unisys asserted their rights to the LZW compression that lay at the heart of the GIF format, forcing those developing software to pay royalties whenever they made use of GIF/LZW. TIFF supports LZW compression, which can reduce the file size