Compression and digital preservation: do they go together?

Judith Rog; National Library of the Netherlands; The Hague; the Netherlands

Abstract

This paper discusses whether compression can be used on objects that need to be preserved for the future. In the field of digital preservation compression has always been considered as one of the worst things to do. This article discusses the arguments against compression from a digital preservation point of view and tries to distinguish myths from facts.

The e-Depot

The Koninklijke Bibliotheek, National Library of the Netherlands (KB) has the responsibility to collect and preserve all Dutch publications. Traditionally the preservation of publications concentrated on the preservation of paper. Nowadays, publications are increasingly published in digital form rather than on paper. For this reason the KB needed a digital archiving system. In 2003 the e-Depot was taken into production. The e-Depot is a digital archiving system that was originally dedicated to the digital preservation of academic journals. The system was developed as a joint effort of the KB and IBM and resulted in the IBM DIAS (Digital Information and Archiving System) [1] core, complemented with some custom-made software for the KB. The system design is based on the OAIS Reference Model [2], an ISO certified standard for digital archiving that has been widely accepted in the digital preservation community. The model offers a framework and a common language to discuss digital preservation and archiving systems dedicated to long-term preservation. It does not, however, offer ready-to-use solutions on how to put such a system into practice. And although in the last few years many archiving systems have been developed that are based on the OAIS model, most of them provide only the safe storage of objects and their metadata and do not provide solutions on how to keep the stored objects accessible for the future. Digital preservation is not only about the storage of bits, but also about keeping the objects accessible. In the same year the e-Depot became operational, the KB set up the Digital Preservation Department. This department is responsible for the long-term access to the objects in the e-Depot.

Currently, the e-Depot contains mainly academic journals by national and international publishers. Today (February 2007), approximately 8 million articles in PDF take up 8 TB of storage space. In the near future the KB will also start with the storage of digital master images in the e-Depot: the output of mass-digitisation projects currently in progress. Preliminary calculations predict that if all images from the mass-digitisation projects are stored in uncompressed form, another 33 million files consisting of more than 600 TB of data will be added to the e-Depot within the next 5 years. On top of that, large quantities of data from a web archiving project [3] that has started in 2006, will be added to the e-Depot as well. With these huge amounts of data that we will have to store, storage has become a real issue for the KB. Given that so many other cultural institutions have started mass-digitisation projects as well or will start them in the near future, we expect that storage is not only an issue for the KB, but also for many others.

Although popular belief suggests that costs for storage per storage unit (GB/TB) will continue to decrease, a more likely scenario is that only the costs for the media on which the data is stored will decrease, but the costs for maintenance, administration and preservation of the stored objects will actually increase, especially in the long-term. According to the National Archives of Sweden [4] the costs for storage media amount to only 5-10% of the total costs. Storing an increasing amount of data for the long-term will therefore lead to considerable higher costs.

In view of the enormous growth in storage volume the e-Depot will face during the next five years, some form of compression would be very attractive; it would lead to a drastic reduction in costs. When considering compression to reduce the costs of data storage, we must however keep in mind that the e-Depot was not simply designed for storage, but more specifically for long-term preservation. This implies that the KB not only focuses on bit preservation, but also on strategies to keep the stored documents accessible in the future.

Compression

Compression is reducing the number of bits used to store data. This can either be done with or without loss of information, using lossy or lossless compression algorithms. Lossless compression algorithms reduce redundant information in the original data without the loss of information. A file that has been compressed using lossless compression can always be restored to its original. Using lossy compression means there will always be some loss of information, although this loss may not necessarily be perceptible by the user.

This paper will discuss whether compression has any consequences for the long-term sustainability of the digital objects. More specifically we will concentrate on the compression of images as the output of the mass-digitisation projects that have been started or will be started in the KB within the next five years. The choice for either lossless or lossy compression and possibly the degree of lossy compression, is a choice that has impact on the visible quality of the image. It can be compared to the choice for a particular resolution or tonal capture. These choices are not relevant from a digital preservation point of view and should be made by the quality managers of the digitisation project.

When discussing compression within this article, we will concentrate on compression that is used within an image file format (e.g. using JPEG compression within a TIFF file or using compression within a JPEG 2000 file). For digitisation projects this is the first kind of compression to consider. We will not discuss the so-called archive formats that can be used as a wrapper on top of the original format (e.g. ZIP or RAR). They may be considered at a later stage.

Arguments against compression

In order to preserve digital objects for future use we have to do more than bit preservation. Because of the rapid changes in technology, the software and hardware that were used at time of creation may have become obsolete. Digital preservation strategies and tools are needed to keep ahead of the changes in technology and keep the objects accessible. Several strategies exist to do so. The digital objects can be adapted to the new environment (migration). The old environment can be simulated on the new environment (emulation). Or the objects can be stored in a way that is as much as possible independent of a particular environment. All strategies have their pros and cons. The choice for a strategy depends on many factors, but whatever strategy is chosen, it is essential that we keep so-called preservation metadata on the objects. These metadata hold information essential for the future interpretation of the objects. Summarized; digital preservation consists of the preservation of bits, the strategies and tools to keep the objects accessible and the creation and storage of metadata describing the objects.

Now, why is compression so much frowned upon in the digital preservation field? What are the risks for the long-term sustainability of the objects? The main disadvantages of compression that have been mentioned in the digital preservation community can be summed up as follows:

1. Compression adds an extra layer of complexity to the object, while the interpretation of the bit stream should be as simple as possible.
2. To render the object in the future another source of information (the (de)compression algorithm) is needed, adding another dependency while dependencies should be avoided whenever possible.
3. Performing the compression is an additional operation on the digital object and every additional operation can potentially cause errors.
4. Compressed objects are more affected by bit errors leading to an irretrievable loss of data.
5. Compression impedes preservation actions.

Considering these arguments, it seems that the need for mass-digitisation projects to reduce their costs for storage conflicts with KB's long-term preservation policy. On the one hand we want to reduce costs, but on the other hand if we do so, by compressing the images, we go against the commonly accepted rules of digital preservation. But are these arguments against compression for long-term preservation really irrefutable? A report [5] from the Gemeentearchief Amsterdam (Municipal Archives Amsterdam) shows that JPEG 2000 compressed objects are in fact more robust. The report also argues that when the compression algorithm is carefully stored, the extra layer of complexity of the compressed objects does not cause additional difficulties. As long as the compression algorithm is available, the documents can always be decompressed again. The report claims that storing a compression algorithm is straightforward and similar to storing the file format specifications. This report was another trigger for the KB to reconsider the proposition that compression should not be used on objects that need to stay accessible for future use.

In this paper we will discuss the five arguments against compression and propose to test their validity.

1 An extra layer of complexity

A digital object is manifested as a file. The bit stream of that file has to conform to the specifications laid down in the file format specifications. To render the digital object a computer is required that has to interpret the stored bit stream and translate it to a representation that can be interpreted by the human user. The file specifications are therefore a crucial link to keep the stored files accessible. The interpretation of the bit stream is a complex task. From a preservation point-of-view, it is argued that the interpretation of a bit stream should therefore be kept as simple as possible. The simpler the specifications that are used to store the file, the simpler the task of interpreting the bit stream in the future and therefore the least chance of problems during interpretation. When we follow this line of reasoning, an uncompressed TIFF file is more suitable for long-term preservation than a JPEG compressed TIFF file. To render the latter, both the TIFF specifications as the JPEG algorithm must be used to interpret the file.

This extra layer of complexity will only be a vulnerability if in the future, we have to build a reader from scratch. This presumes that one day we will discover certain image objects in a digital archive for which there is no reader application available anymore. But is that a plausible scenario? When considering the suitability of a file format for long-term preservation not only the complexity of the format is a criterion to consider, but also the ubiquity of the format and the availability of reader applications. A JFIF/JPEG file will score much higher on these last two criteria than an uncompressed TIFF file.

2 Another source of information to store

As described, the file specifications are the key for interpreting the file and rendering it in a human 'readable' form. Hence it is crucial that the specification will be stored and remain accessible for a future programmer to use. Therefore adding compression to a file not only means that interpretation of the file can be more complex. It also implies that both the file format specifications and the decompression algorithm should remain accessible for future programmers to use.

The TIFF 6.0 specifications do not include the specifications of the algorithms that can be used for compression within the TIFF file. For a TIFF file that uses JPEG compression, not only the TIFF specifications will have to be preserved, but also the JPEG algorithm and possibly all other specifications that are referenced in either one. Both references have to be carefully analysed to see if they are self-containing or also refer to external references that are essential to interpret the bit stream and render the digital object.

File format specifications and compression algorithms should not have to be stored by each institution that has an archiving system for long-term storage. One or preferably several repositories should keep these specifications and all other external references needed to interpret a file safely for the long-term. A logical place to do this could be the file format registries that are emerging like PRONOM [6] and GDFR [7]. Currently however, this is not within the scope of these registries. PRONOM refers to an external location where the file specifications can be found, like for example to the Adobe website for the TIFF 6.0 specifications. But what if Adobe for some reason stops publishing the specifications at this location? PRONOM also mentions the

compression schemes that can be used within TIFF 6.0, but it does not point out where the specifications of these algorithms can be found and whether they are fully self-containing or refer to other external references as well. We plead for a repository that would actually include all relevant specifications of these formats and all other external dependencies within the repositories themselves. On top of that such a repository could include all relevant software to render, migrate, emulate etc. the objects as well. Whether this should be a registry like the file format registries as mentioned above, or another repository or archive is not important, as long as all specifications, algorithms and software are safely stored and remain accessible for future use. Even if shared repositories are not available yet, the institution could store the information in a local repository as long as it carefully investigates what information has to be preserved and preserves that information for the future. If specifications would be stored and maintained like described above, the risk of using compression is mitigated.

3 Error introduction

Not only are compressed files more complex to interpret, they are also more complex to create. When performing compression on an image, errors could be introduced and as a consequence the file may not fully adhere to its specifications. This may have consequences for current or future use of the object. Applications may not be able to render the object if it does not adhere to its full specifications. The more complex the procedure to create the file, the bigger the chance of making mistakes. However, this theoretical risk has never been tested in reality.

4 More affected by bit errors

Another often heard argument against compression is that compressed files are more affected by bit errors than uncompressed files. This means that the loss of bits has a much higher impact on compressed images and as a consequence would more often lead to problems in the representation of the image than it would for uncompressed images. However, a report [5] from the Gemeentearchief Amsterdam (Municipal Archives Amsterdam) claims that JPEG 2000 compressed objects are, in fact, more robust than uncompressed images. Random bits in compressed and uncompressed images were corrupted and as a result the JPEG 2000 images were less affected than the uncompressed TIFF images.

Theoretically, one could also argue that the fewer bits to store in a bit stream, the less chance to loose a bit. Compressed images are less likely to be affected by bit changes as a result of errors occurring on the storage. Suppose we have a carrier that can contain 1,000,000 bits and we have an uncompressed image that uses 500,000 bits and a compressed one that uses only 5,000 bits. Now assume that a hundred, randomly distributed, bits would be corrupted for some reason, chances are far higher for the uncompressed image to be affected than for the compressed one.

5 Impedes preservation actions

Allegedly, the compression of images might impede preservation actions that need to be performed on an object. A preservation action is any action that has to be performed to keep the objects safely stored and accessible for future use. Examples are: migration of digital objects to a new storage medium when the older medium is reaching its end-of-life, the migration of digital objects to a more open or newer format when the old format does not have open specifications or when current software no longer supports this format and it becomes obsolete [8] or emulation [9] for more complex digital objects that cannot easily be migrated. Such actions could supposedly be hindered by compression.

When we keep in mind that we are focussing on compression within a file format, this last argument actually comes down to the same point as was made before when describing the difficulties that may arise when trying to read the compressed objects in the future. It may be more complex for migration tools to migrate a compressed object because of the complex compression algorithm that was used. But in essence this is the same problem that rendering applications have to face, because they both need to interpret the file before either rendering or migrating it. The same pros and cons can therefore be repeated here. Furthermore, when considering for example uncompressed TIFF versus JFIF/JPEG the chances that preservation action is needed at all may be higher for TIFF than for JPEG because the latter is less likely to become obsolete.

Other forms of preservation actions needed, like media migration, may actually benefit from compression because there are less bits to copy from one medium to the other. Especially when dealing with large archives, the enormous volume of data storage could lead to problems when the need for migration to new storage media arises. Compression could drastically reduce time and costs for such a migration. Also, because there are fewer bits to copy, there is less chance for creating errors during this kind of migration.

Next steps

As argued above, it is not so straightforward that compression is a real danger for objects that need to be preserved and remain accessible for the long-term. There are many theoretical pros and cons that can be brought up for discussion. Now it is time to put these theoretical ideas to the test. As far as we know this has not been done before for the purpose of digital preservation. Below we describe how we propose to do this.

To test whether it is a problem that for compressed files not only the file format specifications, but also the compression algorithms need to be safely stored, the KB wants to research exactly which references need to be preserved to keep uncompressed or compressed image formats accessible for future use and whether there is a difference in storing file format specifications as opposed to compression algorithms.

Uncompressed and compressed images will be compared to investigate whether more deviations from the specifications occur in compressed than in uncompressed images. In this way we will test whether error introduction is more likely to happen in compressed than in uncompressed images. Possibly we could use a tool such as JHOVE [10] to do so.

In the coming months the KB will extend the experiments that have been performed by the Gemeentearchief Amsterdam (Municipal Archives Amsterdam) to reconsider the proposition that tells us that compressed file are more affected by bit errors. Bit corruption will be simulated on compressed and uncompressed images and the effect on either type will be investigated. We will extend the experiments by not only changing a number of random bits but also simulating the loss of a number of bits. Furthermore the KB will research whether the change or loss of bits is actually

representative of the errors that could occur in real-world situations.

To perform these tests we will need a large set of both compressed and uncompressed image files. As our digitisation projects have not started yet, we do not have such a large collection readily available. The KB is however one of the participating partners in the European project Planets (Preservation and Long Term Access through Networked Services) [11]. Planets aims to develop tools and services to put the preservation planning module of the OAIS model into practice. Among other things, Planets will have a testbed available in the coming months. This testbed will provide tools, benchmark data and a controlled environment that safeguards the reproducibility of the tests. The testbed could be used for our testing.

Conclusion

With the vast amounts of data that the KB will be storing in the next few years, some form of (lossless) compression would be very attractive because it leads to a drastic reduction in costs without the loss of information. However, the KB e-Depot was not simply designed for storage, but more specifically for long-term storage. This implies that the KB not only focuses on bit preservation but also on strategies to keep the stored documents accessible in the future. In the coming months the KB will start some experiments and desktop research to see whether the often heard arguments against compression for the digital preservation point-of-view still hold water. We suspect that some form of compression can be performed on objects intended for long-term preservation as long as you keep to some common sense rules like:

- Use open, well-documented, preferably standardized compression schemes;
- Choose algorithms that are not patented;
- Choose compression algorithms that are widely used;
- Choose algorithms that have forms of error detection (or even error correction) and perform consistency checks in the archive;
- Use metadata to document all choices that were made or settings that were done.
- Safely store all specifications and other external dependencies needed to interpret the compressed files in the future. Preferably in a shared repository for everyone to use.

These criteria are actually very much similar to the criteria that are used to assess the sustainability of file formats. However, as mentioned, further research in the form of structured tests is needed to confirm this opinion.

This paper is intended to start discussion so any ideas, comments or help will be very much appreciated.

References

[1] IBM DIAS, Digital Information and Archiving System. Information available at: http://www.kb.nl/dnp/e-depot/dm/dias-en.html

[2] "Reference Model for an Open Archival Information System (OAIS)" Blue Book, Issue 1 (January 2002). Also available at: http://public.ccsds.org/publications/archive/650x0b1.pdf

[3] Web archiving, Digital Preservation Department National Library of the Netherlands. Information available at: http://www.kb.nl/hrd/dd/dd_projecten/projecten_webarchivering-en.html

[4] J. Palm, The Digital Black Hole. Available at: http://www.tape-online.net/docs/Palm_Black_Hole.pdf

[5] R. Bourgonjen, M. Holtman, E. Fleurbaay, Digitalisering ontrafeld. Technische aspecten van digitale reproductie van archiefstukken (April 2006). Available in Dutch only at: http://gemeentearchief.amsterdam.nl/algemeen/organisatie/projecten/digitalisering_ontrafeld_web.pdf

[6] PRONOM, UK National Archives. Information available at: http://www.nationalarchives.gov.uk/pronom/

[7] GDFR, Global Digital Format Registry, Harvard University Library. Information available at: http://hul.harvard.edu/gdfr/

[8] Migration research, Digital Preservation Department National Library of the Netherlands. Information available at: http://www.kb.nl/hrd/dd/dd_projecten/projecten_migratie-en.html

[9] Hoeven, J.R. van der en Wijngaarden, H.N. van, "Modular emulation as a long-term preservation strategy for digital objects", International Web Archiving Workshop 2005 (IWAW'05), Vienna, Austria, 2005. Available at: http://www.iwaw.net/05/papers/iwaw05-hoeven.pdf

[10] JHOVE, JSTOR/Harvard Object Validation Environment. Information available at: http://hul.harvard.edu/jhove/index.html

[11] Planets, Preservation and Long Term Access through Networked Services. Information available at: http://www.planets-project.eu/

Author Biography

Judith Rog (1976) completed her MA in Phonetics/Speech Technology in 1999. After working on language technology at a Dutch Dictionary Publisher she was employed at the National Library of the Netherlands/Koninklijke Bibliotheek (KB) in 2001. She first worked in the IT department of the KB for four years before joining the Digital Preservation Department in 2005. Within the Digital Preservation Department she participates in several projects in which her main focus is on file format research.