ISO/IEC JTC 1/SC 29/WG 1
(ITU-T SG16)

Coding of Still Pictures

JBIG
Joint Bi-level Image Experts Group

JPEG
Joint Photographic Experts Group

TITLE:	**Use Cases and Requirements for DNA-based Media Storage version 0.1**
SOURCE:	Marc Antonini (editor) and Touradj Ebrahimi (editor)
PROJECT:	JPEG DNA Exploration
STATUS:	**Approved**
REQUESTED ACTION:	For information and feedback
DISTRIBUTION:	Public

Contact:
ISO/IEC JTC 1/SC 29/WG 1 Convener – Prof. Touradj Ebrahimi
EPFL/STI/IEL/GR-EB, Station 11, CH-1015 Lausanne, Switzerland
Tel: +41 21 693 2606, Fax: +41 21 693 7600, E-mail: Touradj.Ebrahimi@epfl.ch

Use Cases and Requirements for DNA-based Media Storage

Version 0.1

1. Introduction

JPEG standards have been used in the storage and archival of digital pictures as well as moving images. The most popular format for storage and archival of digital pictures is the popular legacy JPEG format as described in ISO/IEC 10918 and, in particular, in parts 1, 3, and 5 of the latter standards.

While the legacy JPEG format is widely used for photo storage in SD cards, as well as archival of pictures by consumers, JPEG 2000 as described in ISO/IEC 15444 is used in many archival applications, notably for the preservation of cultural heritage in the form of visual data as pictures and video in digital format. Examples include the Library of Congress, Library and Archives Canada, Chronicling America website, and the Google Library Project. Because of its use in digital cinema, JPEG 2000 is also used for archival of movies in digital form.

In terms of technology, both legacy JPEG and JPEG 2000 formats are based on a transform-quantization-entropy coding pipeline with JPEG using the Discrete Cosine Transform (DCT) and JPEG 2000 using the Discrete Wavelet Transform (DWT), followed by quantization, coefficient reordering, and entropy coding. The legacy JPEG format has been extended to define JPEG XT, as described in ISO/IEC 18477, to include features attractive for archival applications such as lossless coding, while being backward compatible with the popular legacy JPEG format.

The latest JPEG image coding format called JPEG XL, as described in ISO/IEC 18181, also offers a number of attractive features important to archival applications, such as lossless compression and lossless transcoding from legacy JPEG to JPEG XL, resulting in smaller file sizes without numerical loss in the pixel values.

The ongoing effort in JPEG AI to produce a learning-based image coding standard is yet another potential tool that can be used in archival where the content is stored in its original form, but any post-processing such as denoising, super-resolution and enhancements are carried out without impacting the recorded content.

2. Scope

The scope of JPEG DNA is the creation of a standard for efficient coding of images that considers biochemical constraints and offers robustness to noise introduced by the different stages of the storage process that is based on DNA synthetic polymers.

3. DNA-based Media Storage: Use Cases

DNA-based representations of media data might provide efficient means for storage of huge data. Synthetic DNA provides a very high storage density compared with the traditional electronic and magnetic based methods. Furthermore, it also provides long-term support for data, which is not comparable with the traditional storage devices. According to [1], DNA has the theoretical ability to store more than 450 Exabytes in 1 gram, which is well beyond the current HDD technology requiring 600 grams for a 10TB of storage.

Moreover, DNA can last for centuries, which is not comparable with the typical duration of the current storage devices. Finally, it is becoming fast, easy and cheaper to perform in-vitro replications of DNA. In fact, DNA-based storage is considered as one of the solutions to the growth of digital data that some believe could reach over 170 zettabytes in 2025 [2]. Most of this data is related to the proliferation of media information over social networks. However, most of this media information is almost never accessed (the so-called cold data) and its storage does not require rapid access. Currently, DNA still faces the lack of random access which limits efficient access times.

While storage is the key denominator, there are different relevant use cases depending on specific requirements in terms of storage longevity, target quality, etc.

3.1. Long Term Media Archives and Cultural Heritage Preservation

Considering the complexity of the storage/synthesis and reading/sequencing processes, DNA-based storage seems well suited for large scale, long-term preservation archives with DNA-based storage confined to one or a few central storage units where information is intended to be accessed only infrequently [3]. In this case, longevity is a key requirement and no quality degradation would be acceptable. Lossless coding may also be a relevant requirement. National archives of audiovisual media and cultural heritage artefacts clearly fit in this use case. Cultural heritage artefacts archival covers a wide field of types of digital items from scanned ancient books, maps and photos to three-dimensional models of small objects like statues, flat objects like textile samples and entire ancient carpets, zoology and geologic specimens and ancient scientific devices. According to archive curators besides high-fidelity coding, scalability, progressiveness and random access should be supported for efficient browsing and detailed inspection. Storage of metadata jointly with the artefacts records is also seen as very important, as it is common practice to have artefacts expert analysis report and commentary stored as metadata.

3.2. Social Networks Cold Media Storage

With the explosion of social networks, huge amounts of personal media assets are created, which should be stored for long periods, e.g., the lifetime of users. However, most such data, getting old with time, are infrequently accessed, thus justifying the so-called cold storage. In this case, some quality degradation may be tolerated over time. Companies like Facebook, Twitter, etc., may fit in this use case.

3.3. Preservation of Medical Images

Medical images are typically represented by huge amounts of data. Several medical diagnosis systems generate images that require very high resolution and high dynamic ranges, while multiple systems represent volumes through several longitudinal slices. In multiple medical image applications, lossy storage is not acceptable. One of the reasons is related to legal issues that can arise if a wrong diagnosis is issued, that could be considered as caused by the loss due to compression technology. Furthermore, radiologists and physicians of some specific specialities are trained to use almost imperceptible textures for diagnosis, that should not be affected by any lossy coding mechanism.

Nowadays, because of the huge storage requirements, hospitals usually just delete the medical records of patients after some time, typically after the patient's recovery. This behaviour limits the possibility of this information being used in follow up studies or as case studies.

DNA-based storage would be a natural option for medical images that could solve this challenge.

Furthermore, hospitals typically have special conditions for DNA preservation, which makes the DNA storage a perfect solution for medical imaging long term storage. Due to the type of data in question, lossless coding and high error resilience are thus critical.

3.4. Preservation of Large-scale Repositories of Biomedical Data: Beyond Local Data Storage

The availability of a standardized effective compression and coding strategy would be essential for setting the basis for a unified framework for the collection, sharing and processing of biomedical big data. Medical data are mostly acquired and represented in higher dimensional spaces: from volumetric data, such as structural MRI and CT, to 3D+time, such as functional MRI, where the acquisition consists of a temporal series of 3D volumes, to 4D, as it is the case for diffusion weighted MRI (dMRI) where many volumes are acquired in a single scan, ranging from 32 for basic clinical acquisitions to 256 or more for advanced acquisition schemes. The exploitation of the native dimensionality of the data would lead to remarkable improvements in compression performance also in a lossless mode. This would follow the spirit of JP3D, the extension in Part 10 of JPEG 2000 targeting volumetric data, expanding to higher number of dimensions and accounting for inter-volume-redundancy and potentially leading to a tremendous gain in compression. In addition, ensembles of multimodal data are increasingly available for the same subject. The exploitation of intra-subject correlations across modalities could in principle be also exploited.

Metadata allowing to effectively retrieve the entire set of information concerning a specific query, for instance, a given subject or a given feature (type of pathology, acquisition site, etc.) would be of great help to the research community. Indeed, such big data repositories are multimodal (including imaging data from many different acquisition modalities), multiscale (including genetic, imaging, behavioural and lifestyle data), and multidimensional, ranging from 1D (neurophysiological signals, genetic data), to 2D (images), 3D (structural MRI, CT etc), and 4D (volume sequences such as diffusion MRI, functional MRI, ASL), and an effective indexing policy allowing random access and fast retrieval is still missing. Needless to say, data from gene sequencing and expression is growing exponentially, and the DNA coding strategy could be a very elegant solution to the long-term storing of such data as well.

The increasing need of large-scale studies have boosted a number of initiatives targeting multicentric data collection and sharing calls for new and efficient means of long-term data storage, even before the recent pandemics. Standardization comes naturally into play in order to facilitate and enhance data sharing and algorithm benchmarking.

Remarkable examples are the UK BioBank (https://www.ukbiobank.ac.uk/), collecting a large-scale biomedical database and research resource, containing in-depth genetic and health information from half a million UK participants, the Alzheimer's Disease Neuroimaging Initiative (http://adni.loni.usc.edu/), including MRI and PET images, genetics, cognitive tests, CSF and blood biomarkers as predictors of the disease of both healthy and diseased subjects, the Human Connectome Project (http://www.humanconnectomeproject.org/), that represents the first large-scale attempt to collect and share data of a scope and detail sufficient to begin the process of addressing deeply fundamental questions about human connectional anatomy and variation of healthy individuals. In addition, standardized effective coding would support initiatives such as ENIGMA (Enhancing Neuro Imaging Genetics through Data Analysis), a consortium bringing together researchers in imaging genomics to understand brain structure, function, and disease, based on brain imaging and genetic data.

3.5. DNA Coding for Traceability

One important use case where DNA data storage could be used is for traceability purposes. In other words, the ability to identify and trace elements of a product as it moves along the production chain from raw material to finished products. Traceability is highly important as it can allow easy identification of produced goods revealing information which is fundamental for protecting consumers and increasing their trust on different brands. Traceability is the process of marking products with some special barcode containing all information regarding the product's authenticity by listing all the ingredients as well as the ingredients' origin and characteristics. This barcode should therefore follow the product throughout its lifespan.

When traceability is applied on materials such as textiles, gold, diamonds, or construction materials such as concrete, it is essential that the barcode can remain intact for decades or even hundreds of years in some cases. This barcode longevity can be achieved using DNA data storage. More precisely, the information regarding the product can be encoded into DNA and stored into the concerned material to be retrieved in the long-term without loss of information. Regarding the storage of the DNA into the raw materials there are multiple solutions that can be used according to the different use cases of this application. The DNA containing the barcode needs to be protected from contacts with water and oxygen to ensure reliability of the molecule. To this end, it can be either inserted into sealed capsules [4] which can ensure reliability for thousands of years in room temperature or other options such as encapsulation of DNA into silica glass beads [5] that can protect DNA for decades in room temperature and can go up to 2000 years if kept in low temperature. The encapsulated DNA is then integrated into the different materials to signify the authenticity and the list of ingredients of the referring material.

4. DNA-based Media Storage: Requirements

The following presents a potential list of requirements for the identified use cases:

1. **Compression efficiency -** The standard **shall** offer significantly increased compression efficiency when compared to solutions in the literature, e.g. based on binary transcoding. This includes various media modalities including those where redundancy has to be exploited across multiple components such as volumetric and multi-spectral images.

2. **Lossless coding -** The standard **shall** offer lossless coding at a state-of-the art compression performance.

3. **Composite media assets -** The standard **shall** offer the capability to represent composite sets of multiple, related elementary media assets, notably images (and associated metadata), e.g. by means of appropriate auxiliary data.

4. **Metadata -** The standard **shall** offer the capability to efficiently code relevant metadata, associated to the relevant media assets, available or not in a binary representation.

5. **Privacy and Security –** The standard **shall** offer the capability to efficiently code appropriate privacy and security related data such as integrity protection, associated to the relevant media assets.

6. **Random access -** The standard **shall** allow the access to specific parts of the information without having to decode the full coded information.

7. **Biological constraints -** The standard **shall** consider the relevant biological constraints on the coding

process to avoid affecting the stability of the sequence and synthesising and sequencing errors, e.g. avoiding long homopolymers (repeats of the same nucleotides > 3) and extreme G-C content.

8. **Error resilience -** The standard **shall** offer some degree of error resiliency regarding reading/sequencing errors, including unequal protection against errors and those taking into account the special nature of errors in DNA-based storage

9. **Scalability -** The standard **shall** allow scalable/progressive representations of the information where reading only part of the full information would offer a lower quality or lower resolution of the complete information.

10. **Ambiguity -** The standard **shall** allow decoding without any ambiguity, i.e. a decoded bit may not be both '0' and'1'.

11. **Artificial recognition -** The standard **shall** allow the encoding output to be unambiguously recognized as artificial DNA; this is relevant as the artificial DNA stream should not be confused with natural DNA streams.

12. **Biosafety -** The standard **shall** prevent encoding outputs which constitute any danger in terms of biosafety. The standard **shall** define mechanisms and conditions to ensure biosafety.

References

1. *M. Dimopoulou, M. Antonini, P. Barbry and R. Appuswamy, "A biologically constrained encoding solution for long-term storage of images onto synthetic DNA," European Signal Processing Conference (EUSIPCO), Sept. 2019.*

2. *M. Campbell, "DNA data storage: automated DNA synthesis and sequencing are key to unlocking virtually unlimited data storage," Computer, vol. 53, no. 04, pp. 63-67, 2020.doi: 10.1109/MC.2020.2967908.*

3. *"Making DNA data storage a reality", https://www.the-scientist.com/cover-story/making-dna-data-storage-a-reality-30218.*

4. *Kevin Washetine, Simon Heeke et al. "DNAshell Protects DNA Stored at Room Temperature for Downstream Next-Generation Sequencing Studies" Biopreservation and Biobanking 2019 doi.org/10.1089/bio.2018.0129*

5. *Paunescu, D., Puddu, M., Soellner, J. et al. Reversible DNA encapsulation in silica to produce ROS-resistant and heat-resistant synthetic DNA 'fossils'. Nat Protoc 8, 2440–2448 (2013). https://doi.org/10.1038/nprot.2013.154*