Original post

As Archivematica users know, a standard installation of Archivematica includes a Format Policy Registry (FPR) that contains rules, commands and tools for a wide variety of preservation actions that are performed automatically during ingest. One type of rule is normalization: there are hundreds of rules for normalizing (converting file formats to a select set of preservation formats) during ingest. If the user chooses to normalize during ingest, these rules are invoked automatically on any ingested file for which there is a normalization rule.

There are valid reasons to normalize extensively upon ingest. First, it means narrowing your holdings down to a smaller number of formats for long-term preservation, formats that are today considered to be sustainable and "preservation-friendly". This means keeping an eye on, say, a dozen formats rather than several dozen or even hundreds of formats, depending on the diversity of your content producers. Second, it allows you to spot and address issues with formats during ingest, rather than discovering them years down the road when they may be harder to address. For example, that image file may not normalize properly because it has a colourspace issue; better to fix that issue now, with current tools and knowledge, than discover and attempt to fix it sometime in the future. Third, it means a certain amount of work up front, permitting a higher level of confidence that a lot of the heavy lifting on digital preservation has been done by the time the content is placed into long-term storage - that AIP is DONE and it won't have to be touched for a long time.

The downside of extensive use of normalization is the size of your AIPs, particularly when it comes to video files. Nearly all ingested born-digital video files are compressed, and when Archivematica runs the default normalization rule - convert to ffv1/lpcm in an mkv wrapper - a small video file can produce a very large master derivative. If you're interested to find out more about why this happens, see Ashley Blewer's blog post at https://bits.ashleyblewer.com/blog/2019/09/19/ffv1-bigger-than-before/. The same can be true for raster images - a JPEG file can be highly compressed, and an uncompressed TIFF preservation copy can be much larger than the JPEG file. On a small scale this might not make much of a difference, but JPEGs are ubiquitous, and a few thousand JPEGS across a few SIPs can have a noticeable impact on processing time and storage.

Ubiquity is the key here, and this brings us to the main point of this post. Should we change the default settings in Archivematica to skip normalization for highly ubiquitous files like JPEGs and h264-encoded mp4 files? Keep in mind that the settings could always be changed: the normalization rules would still be there but they would just be disabled for certain formats. However, we are aware that not all users edit FPR rules, and that the defaults Archivematica ships with are often considered de facto recommendations by Artefactual Systems.

We would love to hear from digital curators and preservationists out there. What is your opinion on normalizing everything that can be normalized? Do you edit the default FPR rules, and if so, why? Would such a change in Archivematica's default rules have a negative impact on you, or, in your opinion, on the wider community of users? Do you have opinions about specific formats? An open discussion on this discussion list would be great, but if you're feeling shy, please email me at evelyn[at]artefactual[dot]com.

Regards,
Evelyn McLellan
Systems Archivist & Metadata Specialist
Artefactual Systems

Excerpts from responses

In my opinion, normalization for digital video should not be the default, especially for ubiquitous file formats.

I think that ffv1/pcm/mkv is the best format for long term preservation of analog video migrated to a digital format. But, the format shouldn't be employed as a one size fits all option. As you and Ashley point out, transcoding an h.264 encoded video file to ffv1/pcm/mkv is going to occupy more storage space, and given the ubiquity of the codec, likely doesn't reduce the risk of obsolescence. In fact, were you to heavily weight broad adoption of a format when assessing potential obsolescence, h.264 could be viewed as *less* likely to face obsolescence than ffv1, given that the user community for h.264 is broader than that of ffv1.

You could argue that JPEG is better than TIFF for similar reasons (only if JPEG is what you have to begin with!).

The U.S. National Archives (NARA) has released its digital Preservation Framework which also covers video file formats:
https://github.com/usnationalarchives/digital-preservation/blob/master/Moving%20Image%20Formats/NARA_PreservationActionPlan_DigitalVideo_20190801.pdf
There are a few video file formats at moderate risk with the recommendation to transform them. Interestingly enough, no mention of ffv1 at all.

Very very few institutions consider the information that is actually being lost from the original when normalizing.

I think normalization poses a risk in form of a false warm-&-fuzzy-feeling of being “safe” when it comes to files formats. I know that normalized versions are typically kept alongside the original within an AIP – but, really, if an archive then bases all technology / community watch and all profiling of their archive on the normalized file formats, what good does it do?

I wrote this back in 2007:

https://blog.dshr.org/2007/05/format-obsolescence-prostate-cancer-of.html

I still stand by it. Projections of the relative doomed-ness of formats are unreliable. The real question is whether you are confident enough in your projection and the quality of the migration to throw away the original. Unless you are you aren't migrating, you are creating an access surrogate.

There are of course a number of moving parts. One thing that should be acknowledged is that migration is almost always imperfect, so keeping the chain of formats from the original to the current is critical.

Another is that different migrations are for different purposes and with no naming convention there is some confusion about normalisation vs migration vs transformation which may or may not be the same thing. The NARA Guidelines for example talk about producing an ODT of each document for preservation and PDF for access which seems sensible to me. So when the policy for Word changes to ODT2 do you automatically re-migrate from the original and replace the ODT which you can do because you still have the Word. Or keep them all?

Migration validation is another area that is under-discussed. How do we prove at scale that a migration is successful. Which properties do we compare and how much wiggle room do we allow?

Lastly, render tools are under-discussed. If there is a well-supported render tool does that eliminate the need for migration? Or do you do both just in case. Is emulation a special type of render tool?

The question for the community is how do we share best practice for migration for all of the different types of asset across products so we can all compare, learn and implement. NARA have done a great job articulating their internal rules as have others but this are big documents to read and are machine readable.

So that's why Preservica, Artefactual and Arkivum are backing PAR to ensure that we can all read each other's rules and quickly learn from each other. More here http://parcore.org/ .

This is an interesting topic and one that the UK Archivematica group discussed a few years back. There is a blog post about the discussions here (http://digital-archiving.blogspot.com/2017/12/how-would-you-change-archivematica_3.html) that may be of interest.

[From the blog post]:

We had a discussion about the benefits (or not) of normalising a compressed file (such as a JPEG) to an uncompressed format (such as TIFF). I had already mentioned in my presentation earlier that this default migration rule was turning 5GB of JPEG images into 80GB of TIFFs - and this is without improving the quality or the amount of information contained within the image. The same situation would apply to compressed audio and video which would increase even more in size when converted to an uncompressed format.

If storage space is at a premium (or if you are running this as a service and charging for storage space used) this could be seen as a big problem. We discussed the reasons for and against leaving this rule in

the FPR. It is true that we may have more confidence in the longevity of TIFFs and see them as more robust in the face of corruption, but if we are doing digital preservation properly (checking checksums, keeping multiple copies etc) shouldn't corruption be easily spotted and fixed?

Another reason we may migrate or normalise files is to restrict the file formats we are preserving to a limited set of known formats in the hope that this will lead to less headaches in the future. This would be a reason to keep on converting all those JPEGs to TIFFs.

The FPR is there to be changed and being that not all organisations have exactly the same requirements it is not surprising that we are starting to tweak it here and there – if we don't understand it, don't look at it and don't consider changing it perhaps we aren't really doing our jobs properly.

However there was also a strong feeling in the room that we shouldn't all be re-inventing the wheel. It is incredibly useful to hear what others have done with the FPR and the rationale behind their decisions.

In tandem with the reasons outlined, I'd also run shy of default normalisation yet I see where it might be desirable to build it into a specific preservation workflow e.g. the conversion of pdf files to pdf/a. I think Trevor Owens is spot on in framing the normalisation decisions of ubiquitous formats such as JPEG in terms of risk management and threat mitigation: '...just because a format might theoretically be more ideal in terms of sustainability, all that is relevant is the likelihood that the risks will be a problem worth responding to.' [https://osf.io/preprints/lissa/5cpjt]

We are making some changes to the default FPR to make them conform with our local file format policies. There are slight differences between the format policies we are implementing for Special Collections and for our research repository, but the common changes we are making include disabling normalization to ffv1 in mkv for H264-encoded mp4s and disabling normalization to tiff for png images. We have discussed potentially also disabling normalization to tiff for jpegs in the future, a move that I am personally inclined toward but we have not yet committed to. We've found the Bentley Historical Library's look at their own FPR really helpful in thinking through our own situation[1].

It's a bit difficult to discuss format normalization in a vacuum, as the risks are in some cases related to other digital preservation practices such as infrastructure for bit preservation. As one example, one of the reasons uncompressed and losslessly compressed formats are typically preferred to lossy formats is because the impact of a bit flip is much more pronounced in the latter. Yet in many institutions, a considerable amount of resources are spent ensuring that such bit flips are (1) unlikely to happen, and (2) detectable and correctable when they do. We store multiple copies of content, often on different media and services to increase their resiliency, use filesystems and hardware that integrate native checksumming and error correction, and/or calculate and periodically verify file checksums for fixity. I would guess that many Archivematica users fall into this camp...In cases where an institution has invested so heavily in bit preservation, I don't personally find arguments around the potential damage of bit flips in lossy compressed, ubiquitous formats like jpeg and png very compelling.

As my co-authors and I argue in a recent paper on environmental sustainability and digital preservation[2], it's also worth considering whether format migrations need to happen at the moment of

ingest or whether migration at the time of access/use might suffice for some use cases. Widely used FOSS migration tools used by Archivematica such as ImageMagick and ffmpeg are not likely to become unavailable any time soon, meaning that we may be able to responsibly defer some migration in the interests of not increasing the environmental footprint of our activities (keeping in mind that all of those bit preservation activities mentioned above tend to act as multipliers on our storage and energy footprints). The particular value and local context of collections may also mean that these trade-offs are worth it for some collections or content and not others, particularly when curators and preservationists employ tiered digital preservation models (as many of us do in practice, whether formally or just to maximize the impact of our resources).

I don't know that there are easy answers here for the default rules. I wonder if it might be useful to have a few easily selectable FPR "profiles" available to users. I can imagine a future where users are asked on setup to select how conservative they want to be with formats generally (say "Normalize everything possible" vs "Do not normalize ubiquitous formats"), and based on that selection the individual FPR rules for common formats like jpeg, png, H264 mp4 are enabled/disabled accordingly in one action.

[1] Bentley Historical Library. "Customizing Archivematica's Format Migration Strategies with the Format Policy Registry (FPR)" (October 2016).
http://archival-integration.blogspot.com/2016/10/customizing-archivematicas-format.html.
[2] Pendergrass, K., Sampson, W, Walsh, T., and Alagna, L "Toward Environmentally Sustainable Digital Preservation." The American Archivist, Vol 82, Issue 1 (Spring/Summer 2019).
https://doi.org/10.17723/0360-9081-82.1.165. Open access copy avaialable via Harvard DASH:
https://dash.harvard.edu/handle/1/40741399.

On the Institutional side of things we have been working with the default preservation normalization rules on a case-by-case basis. In a recent example, we received an accession of convocation videos in .mov and .mp4 video containers. Needless to say these videos are very large in size. We have made the decision not to do any normalization up front and only create access copies in an “on demand” scenario. As a secondary example, we also recently received a transfer of approximately 1,000 files of administrative records from a faculty at the school. There was very little media included in this accession so in this case we did attempt preservation normalization on this accession.

I think (and I hope) as we receive more and more content clearer patterns will emerge and we will be able to react more consistently. We have also been including a document in the submission documentation sub-directory of the AIP which we hope rationalizes our decision making a little.

The other opportunity that presents itself on the institutional records side is our ability to work with and have some influence on records creators within units to create and send us files in a more standardized way that would fall in line with digital preservation best practice. With private donors this would likely never be an option.

As a service provider, we have provided both options for government organisations and institutions for born digital collections: either normalising to a standard format, or organising original digital video as preservation master files.

If there is a consensus to change default settings of video to skip normalisation, it should be codec and format dependant. As an example, you probably don't want a highly proprietary format to be the preservation master.

I would also recommend that the video files have gone through some type of format validation and quality control process before being presented to Archivematica to ensure the integrity of the file, technical parameters meet the particular specification, and video and audio characteristics are valid and agree with file technical metadata (eg frame rate, bit depth, interlaced or progressive etc)