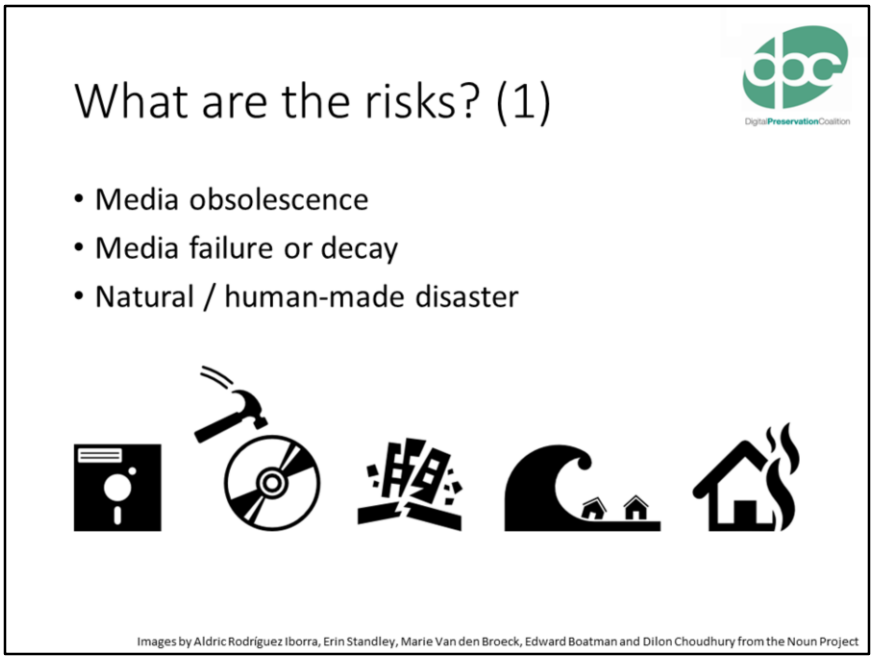
Just keep the bits: an introduction to bit level preservation

Just keep the bits…

- At a fundamental level, all digital data is stored as a series of 0s and 1s. These are “binary digits” or “bits”.
- There are many challenges that digital preservation seeks to solve, not least of which is preserving our ability to decode digital data and gain access to the useful information encoded within it
- But at its most basic and perhaps critical level, we need to make sure that that this encoded data, our files or streams of bits, are not lost or damaged
- This session will describe some approaches to make sure that we can simply *keep the bits*



- Our data faces a variety of risks. If left alone, it's not likely to survive intact on into the future.
- Worse still, any loss may not be entirely clear to the casual observer. In fact it may require massive manual effort just to work out if any of your data has become damaged. Unless some care is taken to manage and preserve the data properly
- What are the risks?
    - Media obsolescence – when storage media, such as tape, floppy disks or CDs become obsolete. You no longer have the hardware needed to read them. Example – my laptop doesn't even have a disc drive.
    - Media failure – storage media is commodity product and tends to have a reasonably short lifespan. Most hard disks tend to have a reliable lifetime of around 5 years. A commonly cited example of media failure is 'bit rot' – though all forms of storage media are subject to different forms of decay, bit rot refers to the loss of data due to the small electronic charge of a bit, or alternatively, by cosmic rays or other high energy particles. This can be avoided by using different forms of storage media and refreshing data onto new storage media over time.
    - Disaster – Fire, flood, etc.

What is the result?

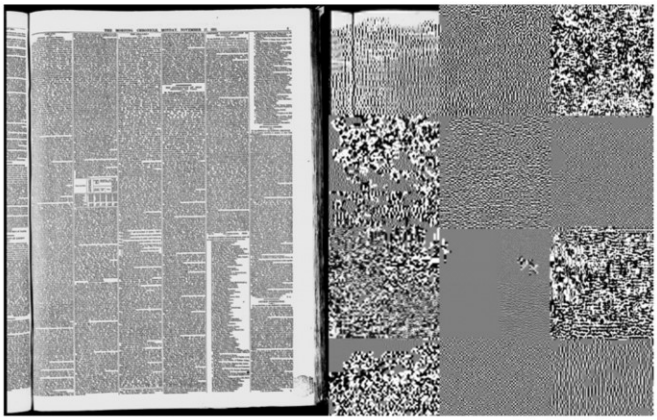
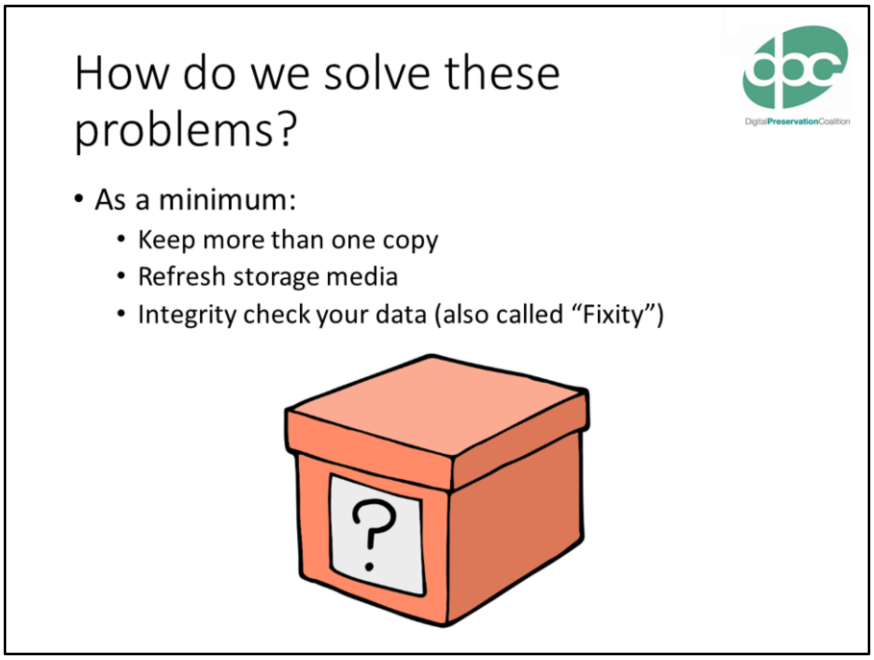


Image courtesy of the British Library

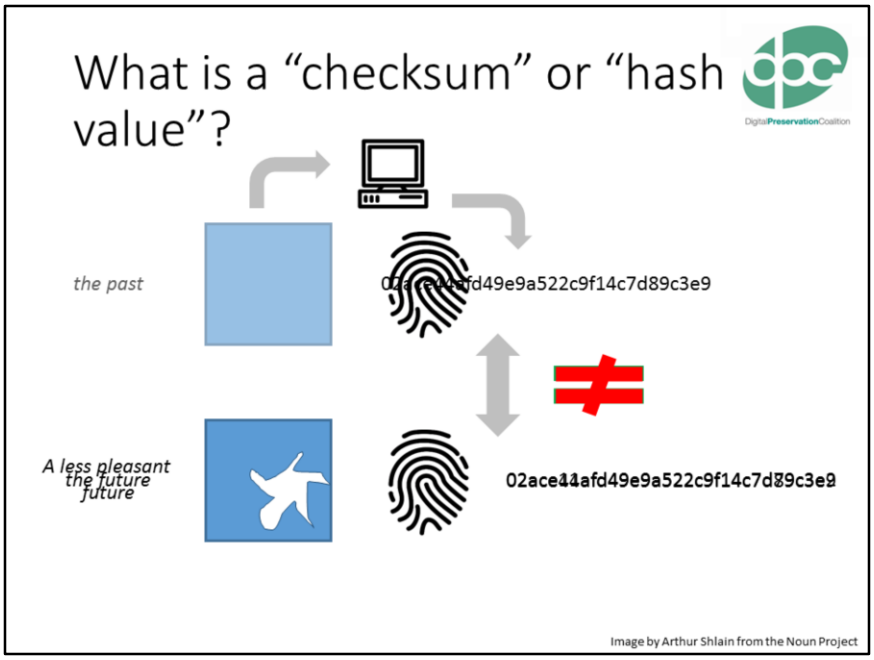
So what happens when these risks bite? The outcome is often unpredictable:

- Media degradation will often lead to a complete failure of the storage device. In other words, you can't read back any of the data stored on it.
- In some cases, damage might be more subtle. Some of the bits in a bitstream might become lost or damaged. This might lead to an obvious result as in the case of this before (left) and after (right) screenshot of a digitised newspaper page.
- Alternatively, damage might be less difficult to recognise visually. Some of the newspaper pages that were also damaged looked fine until you zoomed in, and they became fuzzy. Although the bitstream was damaged, the viewer software did it's best to render the image without informing the user. Things are not always as they seem!



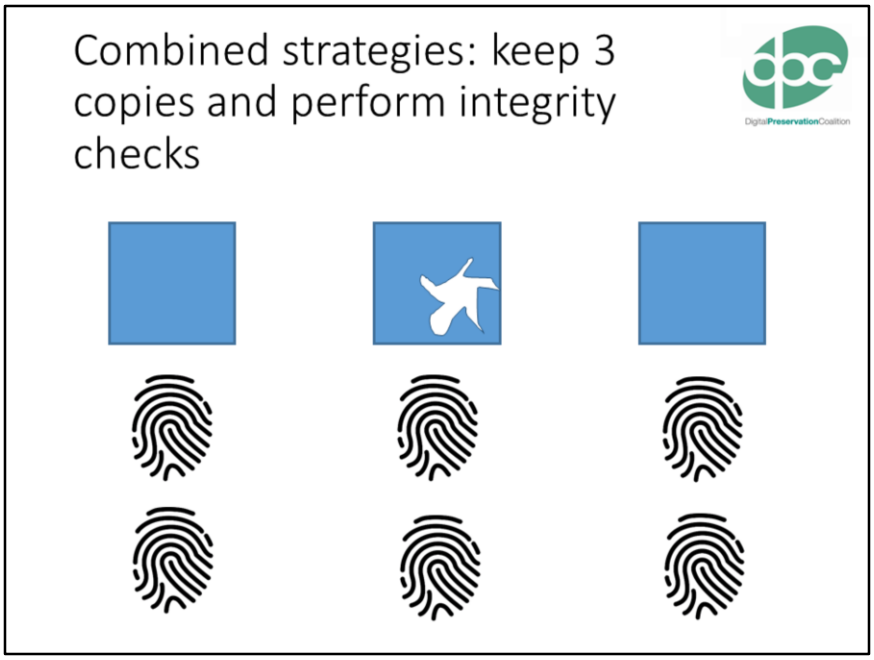
So how do we solve these problems?

- Keep more than one copy of the data. 2 is ok, 3 is great. Some organisations store 4 copies.
- Note that digital preservation is a trade-off between risk and cost. The more copies the better, but more are costly. There is no precise answer for the perfect number of copies as the sweet spot is likely to depend on your own circumstances.
- Keep one copy in a different geographical location – provides some insurance against natural or unnatural disaster
- Storage media will degrade over time, so be prepared to periodically migrate data to new storage media
- Things will still go wrong, so implement a process of integrity (or fixity) checking so you can automatically tell if your bitstreams are still intact.



So to perform integrity checking we need to use a simple technology called checksums or hashes. This animation shows how they work.

1. Let’s say we have a bitstream, or digital file, that we want to preserve.
2. We begin by creating a checksum. This is simply a fairly unique short number derived from the file using a software tool.
3. Think of a checksum as finger print.
4. At some point in the future, we want to verify that our file remains exactly as it was, back when we first created the checksum at the top of the screen.
5. We generate a new checksum from the file.
6. We then compare the new checksum with the old one.
7. In this case, the checksums are identical, so we know the file is undamaged and exactly as it was.
8. However, if the file had become damaged, the checksum we would generate from it would be different.
9. On comparing the two checksums, we can see that they are different from each other. This confirms that our file is no longer identical to how it was. Perhaps it has been damaged by media failure or “bit rot”.



So if we combine our two strategies of keeping multiple copies of each file and applying integrity checks:

1. We make 3 copies of the file we want to preserve, ideally placing one file offsite to protect against natural disasters
2. We generate checksums from each file, and we can see that they are all the same, and each file is good.
3. Over time we can then recalculate our checksums, and see that the three copies of the file are still exactly as they were
4. Until at some point in the future we recalculate our checksums and discover that one of them is different!
5. Straight away we know that the middle copy has become damaged
6. So we then discard the damaged file
7. And replace it with a copy of one of the others

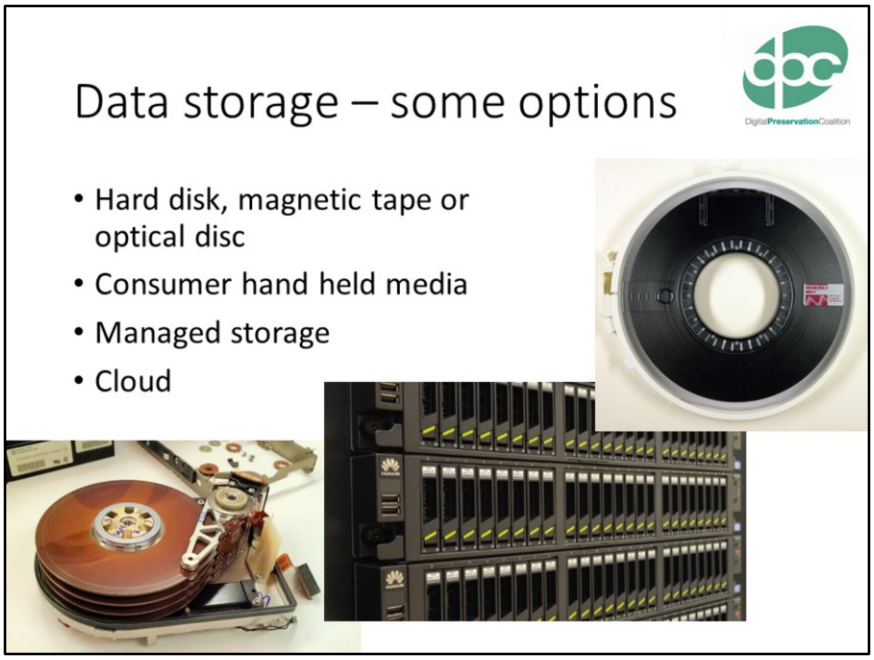
Using these techniques we can dramatically reduce the chance of losing any of our data.

Integrity checking - tools

- Fixity
  - https://www.avpreserve.com/tools/fixity/
- Auditing Control Environment (ACE)
  - https://wiki.umiacs.umd.edu/adapt/index.php/Ace
- For alternatives – see COPTR
  - http://coptr.digipres.org/Category:Fixity

Here are some software tools you can use to perform integrity checking:

- The tool “Fixity” is a good place to start for generating checksums.
- ACE is a more advanced tool, ideal for performing scheduled integrity checks.
- More options can be found on COPTR.



Here are some options for technologies and approaches to storing your data:

- Hard disk (sometimes known as “spinning disk”), magnetic tape or optical disk technologies all have their advantages and disadvantages. For example, magnetic tape is cheaper than hard disk, but will not provide as fast access to the data
- Except at very small volumes of data storage, consumer hand held media, such as external hard drives, DVD discs or Blu Ray discs should be avoided.
- A better option is proper managed storage. As is shown in the middle photo, a server mounted rack of hard disks that is located in a climate controlled server room and managed by an expert will be much more reliable, if a little more costly. This is typically what might be provided by the IT department of a larger organisation.
- Alternatively managed storage can be outsourced to a third party. This is typically called “cloud storage”.

Watch out for the red herring that is so called long lived media!

- Frequent announcements will be seen in the news for new types of storage media that are guaranteed to last 100 (or more years).
- None of these technologies has ever yet taken off, as they don't solve digital preservation for us.
- There are a number of reasons behind this, including their expense and lack of any real guarantee about reliability.

What are the risks? (2)

- Common mode failure (technology)
- Human error and malicious damage

There are some further risks to be aware of:

- A single storage technology (either the storage media itself, or the software and hardware that manages and provides access to it) could have some kind of flaw. Perhaps a software bug or manufacturing fault. Keeping copies of your data on at least two different kinds of storage technologies should mitigate this risk considerably.
- Bit rot:
- Human error or perhaps even malicious damage is a potentially very high risk, if low in probability. Ensuring that no single member of staff has write access to every copy of your data could mitigate this risk, but is often a difficult challenge in organisational terms.
- Ultimately many of our techniques for tackling bit preservation are about avoiding keeping all your digital eggs in one basket.

Recap

- Keep several copies of all your files
- Keep one copy in a different geographical location
- Don't use the same storage technologies for each copy and don't rely on a single vendor to store all your data
- Create checksums
- Perform periodic integrity checks
- Avoid having any one person with write access to all copies of data

- Ultimately – you need a digital repository that manages all this for you (and more!)

Further information

- Digital Preservation Handbook
  - http://www.dpconline.org/advice/preservationhandbook
- David Rosenthal's blog
  - http://blog.dshr.org
- AVPreserve cloud storage profiles
  - https://www.avpreserve.com/papers-and-presentations/cloud-storage-vendor-profiles