When I was very kindly asked to do this talk, Pip Laurenson said “just talk about the need for change that you’ve seen during your time working on digital preservation and archiving” - “you’ve worked on European projects, you’ve been involved in technology transfer, you’ve spun out a business from a University, and you’ve established Arkivum- there’s lots to say”. I thought no problem, I can talk about all that.

Then I started thinking a bit harder about what I’d say. And I kept coming back to the same theme. “Digital preservation is bloody hard work”.

I don’t mean doing EU projects like PERICLES - that’s the fun bit. What I mean is that it’s bloody hard making stuff live on after funding runs out in EU projects. For example, it’s bloody hard trying to make a business case for preservation and archiving that stacks up to get commercial investment in a new business. And, in general, it’s bloody hard trying to sell digital preservation solutions outside of the traditional hunting grounds of museums, libraries, archives and special collections.

But, I can’t really stand here and do a talk where I moan a lot about things being bloody hard – especially after all the inspirational things we heard last night at the Digital Preservation Coalition awards!

So I thought about why it seems so hard. I thought about whether there’s an alternative way of doing things. So I thought I’d talk about where I think there are some great opportunites for digital preservation to be applied in new ways and to new problems.

In particular, new opportunities for digital preservation to be adopted and used by a far bigger community than it is today. And new opportunities for us all to benefit from these new uses and these new communities.

I’m calling this ‘preservation through the power of many’ – which is at least a better title than “digital preservation: it’s bloody hard work!”.
Let me start by taking about a different world - the NHS.

The National Health Service in the UK is the world’s fifth largest employer. The NHS is topped only by US department of defense, the Chinese Military, Walmart and, somewhat ironically, MacDonalds, which all by itself has probably helped boost the NHS rankings!

The NHS employs 1.7 million people and has a budget of nearly £140 billion pounds.

The NHS is simply huge.

The NHS provides free publicly funded healthcare to the UK’s 65 million residents.

The NHS deals with a staggering 1 million people every 36 hours.

The NHS does a fantastic job.
But the NHS is in crisis.

The NHS cannot cope with current workloads.

Staff are up in arms about pay and conditions.

Rarely does a day go by without some form of news story about waiting lists, staff shortages, strikes by junior doctors, lack of beds, and a multitude of other problems.

Despite this, the NHS remains one of the world’s best healthcare systems. Probably everyone in this room has been thankful to the NHS at one time or other.

The problem the NHS faces is a population that is both growing and living longer.

We now survive the diseases that used to be killers only a few decades ago. Instead we live on to face heart disease, cancer, strokes, dementia, alzheimers and other horrors of old age. These diseases can be complex, hard to treat, drawn out and of course extremely expensive.

It’s no wonder the NHS is struggling.
Is the solution to increase staffing?

Is the solution to keep increasing budgets?

Is the solution to increase efficiencies?

Is the solution to new technologies?

The answer is both yes and no.

These things will all help, but they are not the whole solution. They are like treating the symptoms and not the cause.
Perhaps the most important part of the solution is for fewer people to get ill in the first place.

Healthier lifestyles mean fewer ailments, fewer trips to the doctors or to hospital, and a huge reduction on the workload of the NHS.

Just to put this into context very briefly, here’s some examples.

According to Cancer Research UK, an unhealthy lifestyle is the root cause of about a third of all cancers.

There are over 350,000 new cases of cancer each year.

Experts say most cases of premature death from heart disease are completely preventable.

160,000 people die from heart disease each year.

http://www.cancerresearchuk.org/health-professional/cancer-statistics
It’s clear that if only people took better care of their health then the NHS wouldn’t be in crisis. And we’d all feel a lot better for it. This is what I mean by the ‘power of many’. 65 million people doing just a little bit more to look after themselves would have a transformative effect on the 1.6m people in the NHS that look after them.

My argument is that the same is true for digital preservation.
If only people creating digital content took better care of their data.

The power of many applies just as much to the long-term health of digital content and our collective digital legacy as it does to the long-term health of our nation and the healthcare system that supports it.
Let me put this into some stark numbers.

There are 65 million people in the UK with 680,000 qualified healthcare professionals in the NHS to look after them.

The same 65 million people all generate digital data in one way or another, either privately as individuals, through their employment and work, or by being part of records or content created by others.

But my estimate is that there’s probably only 10,000 professionals in the UK who have any formal training in digital preservation, i.e. people who know how to look after the digital legacy that we are all creating. For every one person who is trained in how to keep digital content alive, there are 6,500 people generating that digital content day-in and day-out.

That’s a staggering gap. That’s also a huge opportunity. The ‘power of many’ could have an enormous effect if we could get that 65 million to do just a little bit more digital preservation themselves.

http://www.nhsconfed.org/resources/key-statistics-on-the-nhs

“In 2015, across Hospital and Community Healthcare Services (HCHS) and GP practices, the NHS employed 149,808 doctors, 314,966 qualified nursing staff and health visitors (HCHS), 25,418 midwives, 23,066 GP practice nurses, 146,792 qualified scientific, therapeutic and technical staff, 18,862 qualified ambulance staff and 30,952 managers.”

678,912 qualified staff (not including managers).
This is the crux of what I’m proposing.

We need to apply digital preservation ‘upstream’ into the world of those creating content.

And if we don’t do this and take advantage of the ‘power of many’, then we’ll face all sorts of problems ‘downstream’ in our archives and memory institutions.

I’ll spare you the stats and leave the ‘zettabytes digital universe’ to IDC, but if you think about social media, the internet of things, cloud computing and a host of other transformative technologies then three things are clear:

- We are generating ever larger volumes of digital content
- We are generating ever more complex digital content
- Our history is becoming ever more digital and ever more fragile

http://www.idc.com/downloads/where_is_storage_infographic_243338.pdf
If it were easy to move digital preservation upstream then we’d have done it already. Realizing the ‘power of many’ requires three things.

1. Digital preservation tools or techniques need to provide immediate benefit in the world of content creators. The traditional arguments used in memory institutions about safeguarding a digital legacy so it’s accessible in the future won’t work. The argument needs to be that there is benefit right now by solving a business problem right now and not about some benefit that might accrue at some point in the future.

2. Digital preservation tools or techniques need to reduce costs. The traditional argument of investing in digital preservation by spending a bit of extra money now in order to reduce costs of access or loss in the future won’t work. The argument needs to be that digital preservation can not only solve a business problem right now, but it can also save money - at the same time.

3. Digital preservation tools and techniques need to be integrated and automated into the environment where content is created. This is actually just another aspect of lowering the costs, but in this case the ‘people costs’. There needs to be no extra barriers in place, no hurdles to jump, no extra processes and procedures to follow. People need digital preservation to work for them automatically in the background – almost as if it wasn’t there at all.

My argument is that if we can do those three things then it is entirely possible to apply digital preservation tools and techniques to solve immediate business problems and as a result achieve at least some of the ‘power of many’. And most importantly, if we do this then we as a digital preservation community will all benefit as a result. More people will be making use of digital preservation tools and techniques, including industries where there is a lot more money available, and not only will we be on the receiving end of digital content that is a lot more ‘healthy’ but we’ll have a lot more people with a vested interest in sustaining and advancing digital preservation as a discipline.
Pipedream or possibility?

I want to spend the rest of this talk looking at some examples that I see in my day-to-day work at Arkivum of how we can achieve these things.

I’m deliberately going to pick examples outside the traditional areas of heritage and collecting institutions.

Instead I’m going to pick examples in finance, healthcare, pharmaceuticals and data protection.

I’m going to pick examples where I think there really is potential and not just a figment of my imagination!

I’m going to pick examples where the digital preservation technologies and techniques we need already exist – it’s a case of how to apply them to new problems and in new ways.
First up is GDPR, which is the forthcoming EU General Data Protection Regulation which applies in May 2018.

Think of this as expanding the requirements of the Data Protection Act in the UK.

And no, Brexit doesn’t mean it won’t apply – if we want to continue trading with Europe then we’ll need to support a compatible system of Data Protection, which means GDPR.

GDPR applies to Personal Data with a broad definition and has requirements on those that control and process that personal data. These requirements include the need to properly store and secure personal data to prevent unauthorised disclosure and use.

And if you fail to meet these requirements then the penalties can be huge. This includes up to €20M or 4% of global turnover in fines. Yes, 4% of global turnover – which could be massive – think about multi-national businesses such as large banks! This is why industry will care about GDPR because (a) it’s becoming the law and (b) not following the rules could be very costly.


For this reason, estimates are that there will need to be a lot of Data Protection Officers – DPOs – to support the correct implementation and governance of GDPR by an organisation. That’s 75,000 officers – that doesn’t include the people who’ll be executing GDPR on a day to day basis let alone all the people creating personal data in the first place. A lot of people in a lot of industries with a lot of money. Surely an opportunity for the ‘power of many’.

And that’s indeed the case. Whilst initial attention will likely be mostly on data confidentiality and preventing data breaches, which after all is what will trigger the big fines, the GDPR also affords a set of rights to individuals. The right to access your data and the right to have your data are just two of these.

These rights will be getting attention too from organisations, their DPOs, and their suppliers.

I want to highlight one right in particular. The right to data portability.

This says that as an individual I have the right to access my personal data as held by an organisation and I should expect it to be given to me in a form that is easy not just for me to use and understand but is also machine readable and can be transferred to other organisations who need to process it.

As shown here on an extract from The Information Commissioner’s Office website, this requires the use of open, commonly used and machine readable formats.

We know how to do that!

- File format identification
- File format normalisation
- File format validation
- Metadata standards

*Imagine a world where all data mentioning people is in open and preservation ready formats!*

Guess what? We know how to do that in the digital preservation community. We have tools for file format identification, conversion and validation. We have metadata standards for interoperability and exchange. We have just the sort of knowledge and experience that people need to support this part of GDPR.

If we worked with those people implementing GDPR including the vendors of the software and systems that they’ll use, then we could see automated and embedded digital preservation tools and techniques being used to solve the immediate business problem of being ready for GDPR.

We could help save big businesses a lot of money and we could help them be better prepared and we could help this happen sooner.

The benefits then flow downstream to those responsible for the long-term retention of that data. If you indulge yourselves a bit, then you can imagine a world where you might be an archivist or records manager in a large bank or other big organisation and the stuff coming at you for long-term retention is already in a preservation ready format. No need for large-scale re-guard actions. No struggling lots of unknown format types. Surely that would be a good place to be?

http://openpreservation.org/technology/products/fido/
https://www.archivematica.org/en/
http://verapdf.org/
The next example I wanted to use is Digital Pathology and Genetic Sequencing.

The use of digital technologies is transforming clinical diagnosis in healthcare. Benefits include faster turn-round times from pathology labs, the ability for genomics to identify cancers or rare diseases with greater precision and hence drive much more targeted treatments, and the ability to share pathology data between organisations so that a wider set of specialists can be brought to bear on a given patient and their condition.
The problem is that these new disruptive technologies are generating an unprecedented amount of data.

If you look at the fall in cost for sequencing a whole human genome then it’s dropped by a factor of more than 10,000 over the last 10 years.

This means whole genome sequencing is now affordable on a population scale and is on the verge of being part of routine healthcare. You can now sequence a full human genome for approximately $1000.

We’re now getting very close to the point where whole genome sequencing is adopted in clinical practice and routinely applied to those visiting hospital. We’re talking only a few years away, not decades.

https://wwwgenome.gov/sequencingcostsdata/
You can see the potential when you look at the UK’s 100,000 genome project. The whole genome of 70,000 patients and their family members is being sequenced for a range of cancers and rare diseases. This is creating an enormous dataset that scientists can use to better understand the cause of these diseases and how best to treat people at an individual level.

The dataset will top 20PB of data and that’s just for 100,000 genomes. Imaging scaling this up to the whole population.

There are already some pretty basic challenges to address in storing this data and then making it accessible for use, let alone what’s coming down the tracks.
For the 100,000 genome project it’s not so much of a problem – they have a dedicated team and infrastructure built from the ground up to address data at this scale.

What’s much tougher is coping with this data volume in the NHS. The data volumes already being generated by genomics and pathology data is already outstripping the capacity of NHS hospital infrastructures. The NHS already spends $3Bn a year on IT.

Forget about the fact that the data has value to the patient and in some cases their children too, which means keeping it for decades, the problem is here and now.

First, how to store this stuff at scale and at a low enough cost – whilst maintaining a good level of data safety.

Second, how to provide access across healthcare organisations in a secure and managed way.

Lots of data, big files, content that doesn’t change, large scale low cost storage, providing access to a designated community in a controlled way. Sound familiar?
Guess what? We know how to do that in the digital preservation community.

We know how to build large scale data storage systems that keep data safe for the long-term.
We know how to analyse costs and risks.
We how to balance safety, security and ease of access in storage solutions.
We know how to implement storage that makes multiple copies of the data, stores them in different locations, migrates media and checks for corruption.
We know how to provide access to data to a designated set of users – it’s called a trusted digital repository.

There’s a lot that the preservation community has to offer.

But the important point is that what we have is something that can be applied to the immediate and pressing problem of data storage and how to keep the costs down. It’s not about trying to convince the NHS of the long-term benefits of digital preservation – it’s actually about solving the short term pain of storage overload and un-manageable budgets. And if we help the NHS crack that using digital preservation techniques and technologies then the good news is that we’ll help to solve the long-term problem as part of the process.

This is what I mean by using digital preservation to tackle immediate business problems and do so in a way that saves money. The ‘power of many’ in this case is getting one of the largest organisations in the world to make use of and benefit from digital preservation. This ‘power of many’ is then further amplified by the whole population having a deep and vested interest in their data being held safely and securely.

The benefits then flow downstream to those with long-term responsibility for that data. Imagine you’re in a NHS hospital responsible for records management and long-term retention of patient data and that all the genomics data generated or stored by your hospital was held on preservation-grade storage from day one. No more trying to wrangle data from multiple servers and USB drives. No more worrying about data loss. Everything stored with checksums with multiple copies – that would be nice!

http://dpn.org/
https://www.lockss.org/
http://www.arkivum.com/
http://4cproject.eu/
Continuing with the medical theme, I want to look next at drug trials by Pharmaceutical companies.

It takes on average $2.5Bn to bring a new prescription drug to market as part of a process that takes more than a decade.

Part of that process includes clinical trials at various stages to test their efficacy, i.e. whether they work or not, and their safety, e.g. whether they have any nasty side effects.

https://www.sciencemag.org/content/early/2020/03/05/science.337.6171.1225.t3d5v
As you can imagine, drug trials are a highly regulated area and submissions have to be made to regulatory bodies such as the FDA to provide detailed records of a trial and its results.

There is currently a transformation happening which is seeing old paper-based drug trial submissions being replaced with new electronic versions.

A drug trial is documented in what is called a Trial Master File. This is basically a very large collection of PDF documents that is assembled to document the trial and all the tests done on the participants in the trial.

https://tmfrefmodel.com/
As you can imagine, the regulators impose stringent requirements on how drug trials are performed and the handling of the records of the trial.

These requirements include all the usual aspects of records management plus some digital preservation related requirements, for example the ongoing readability and useability of drug trial documentation. These requirements last for the lifetime of the drug on market and then some, i.e. multiple decades.

Not sticking to these standards can be very costly – a failure to produce the documentary evidence could result in a drug having to be taken off market and the loss of revenue from that could be billions. And that’s not including the more important concerns of protecting the health and wellbeing of the public from drugs that haven’t been tested properly or are subsequently found to have problems.

Addressing these challenges is a natural place to target digital preservation.

But that’s not actually the scenario that I want to talk about.

http://www.fda.gov/
http://www.ich.org/home.html
Where I think there is an opportunity to address a more immediate problem is in the way that drug trial information is collected and transferred between organisations.

What happens in a drug trial is that a Pharma company will typically recruit one or more Contract Research Organisations to conduct the trial. These CROs then work with hospitals and doctors to recruit participants to take part in the trials. The documentation is assembled and fed back up the chain ultimately back to the pharma company who has sponsored the work.

Whilst there are formats and systems for creating and storing Trial Master Files within the CRO or pharma organisation, there isn’t a commonly agreed way to transfer these between organisations. Given that a drug trial can last years and contain 100,000 documents, the robust exchange of TMF content between organisations becomes important.

Currently, work is underway to look at standardising the exchange of TMF content.

We know how to do that!

- Metadata standards
- Event models
- Containers
- Chain of custody

*Imagine a world where all drug trial data has checksums and is described using PREMIS & METS!*

Guess what? We can do that.

We have standards for metadata and how to exchange it in a robust way between organisations. It’s called METS.

We have ways of bagging up and transferring documents and metadata files with the ability to check that nothing has been corrupted or lost in the process. It’s called BagIt.

We have ways of describing events for the creation and transfer of content. It’s called Premis.

We have all the components needed by the Pharma industry to underpin robust and standardised transfer of clinical trial documentation between organisations.

This know-how and technology could solve an immediate problem in the pharma industry. There are substantial cost savings from both not re-inventing the wheel and starting out with a solid solution from the outset.

The ‘power of many’ effect comes from the large number of people and big organisations involved in drug trials – plus the very large budgets at stake. Small changes that improve and speed-up the process can have a huge effect – getting a drug to market earlier is worth something like £1M per day to a pharma company.

The benefits then flow downstream to those with long-term responsibility for that data. Imagine you are an archive manager at a pharma company and all the drug trial records you were responsible for keeping came ready packed in PREMIS and METS and had a full set of checksums!

http://www.loc.gov/standards/premis/
http://www.loc.gov/standards/mets/
One final example. This time financial markets.

When you think of financial markets you often think of furious activity on a trading floor full of people in blazers shouting at each other. Or people behind huge banks of screens with telephones jammed against each year.

But in reality there are a lot people in a lot of banks and other financial organisations who are sat behind desks and make trades day-in day-out when managing funds and providing investment services to their clients.

For example, on the London Stock Exchange, there are 1 million trades per day creating a daily turnover of over $5Bn.

As you would expect financial markets and trading are heavily regulated, for example driven by the events that lead to the market crash in 2007.

The EU has created something called MiFID II, which is the second version of the Markets in Financial Instruments Directive.

This puts a load of regulations in place including that any form of communication that could lead to a trade should be recorded and retained for 5-7 years. It’s not just recording the final decision to make a trade but includes everything in the run-up that could have lead to a trade, so for practical purposes this means recording pretty much everything.

https://www.fca.org.uk/markets/mifid-ii
Recording means telephone calls, emails, instant messaging over channels such as Bloomberg, and more recently things like Skype for Business. There’s a lot of channels that need to be captured.
The problem is that much of this information is in silos, e.g. proprietary voice call recording software and formats.

Much of the information is audio. A large financial institution can generate 100,000 call recordings each day.

The problem is that financial organisations now need to extract all these audio files and other types of media from their various systems and store them in a such a way that they have confidence that nothing has gone missing and they can access and use the files whenever they need to. There’s a lot of systems to capture content from and a lot of file formats to deal with.

Basically financial institutions need to create and manage their own AV archives. But in their case it’s so they can satisfy immediate compliance requirements – there’s a very short window of time for making sure call recordings are safely stored – you can consider it as archiving at the point at which data is created. Any delays results in increased risk of non-compliance and a backlog of content to handle that can build up very quickly.
Guess what? In the digital preservation community we know how to build big AV archives and preserve audio visual formats.

We know about codecs and wrappers and how to normalise and convert files.

We know the benefits of using a small handful of ‘pivot’ formats that use open specifications and how this saves the pain of having to manage a proliferation of proprietary formats that come from legacy systems.

We know the cost savings of taking an automated factory approach to large AV collections – it’s 50% cheaper.

All this can be applied to the huge volumes of stuff that banks and other financial institutions are now going to have to collect and manage. It can be applied in a way that helps with the immediate problem of an organisation knowing up-front that it’s compliant and doing so in a way that is faster and cheaper than current approaches.

The ‘power of many’ comes into play because there are a lot of people who care about compliance, information governance and litigation readiness in financial institutions – it’s big money and involves whole teams of people.

But as with the other cases I’ve shown, the benefits also flow downstream to those responsible for the long-term retention of that data. Imagine you’re an archivist or records manager in a financial institution and have responsibility for long-term retention and access to recordings and other evidence under MiFID. Life would be a lot easier if those recordings were already a minimal set of open formats.

http://dx.doi.org/10.7207/twr12-01
http://www.iasa-web.org/tc04/audio-preservation
https://www.ffmpeg.org/
http://www.videolan.org/vlc/index.html
https://mediaarea.net/MediaConch/
https://www.prestocentre.org/
Conclusions

Preservation *can* be moved upstream into the world of content creators.

1. Provide immediate benefit
2. Reduce costs
3. Integrated and automated

Everyone benefits from the ‘power of many’

In this talk, I’ve argued that digital preservation tools and techniques can be applied ‘upstream’ in the world where content is first created and has the potential to be used to solve a wide range of current and pressing problems.

I’ve given just a few examples that I’ve come across and there are for sure going to be many more.

But in order to be successful, I’ve also tried to show that three things are needed.

1. Preservation techniques need to be applied to a problem that needs an immediate solution and provides an immediate benefit. This isn’t to say that digital preservation isn’t also needed in the conventional sense of helping to ensure digital content is accessible and usable in the future, rather that there are some problems in the ‘here and now’ that are often a lot more pressing and mean there’s far more likely to be interest from an organisation or their suppliers.

2. Preservation techniques need to help reduce costs – which includes saving time as well as money. It’s not enough to show that there’s a benefit if it can’t be achieved at a lower cost. This is simply about removing trade-offs and making the decision ‘easy’ – if the application of digital preservation results in an immediate problem being solved and at a lower cost then it’s kind of a no-brainer.

3. Preservation techniques need to be integrated and automated - this is actually just another aspect of lowering the costs. There needs to be no extra barriers in place, no hurdles to jump, no extra processes and procedures to follow. People need digital preservation to work for them automatically in the background – almost as if it wasn’t there at all.

And if these things can be achieved then there’s a good chance of digital preservation getting its foot in a lot more doors and most importantly in industries where there are a lot of people who would be in effect using digital preservation on a daily basis – even if they didn’t know it by that name.

The result should also benefit all of us. There would be more investment into and sustainability of some of the tools and techniques that we all rely upon. Digital content would also be in a lot better shape when it does make it to archivists, records managers and others responsible for its long term preservation.
And that brings me to the end of the talk. I’ll finish with one of my favourite quotes, well actually an African proverb “if you want to go fast then go alone, but if you want to go far then go together”. This is where I think we are in the digital preservation community – we have the opportunity to get a lot more people using digital preservation techniques and tools and through the ‘power of many’ we have the opportunity to go a lot further as a result.