Introduction to PrePARe project 'lightning' modules

The PrePARe project lightning modules are designed to provide a basic introduction to fundamental issues in Digital Preservation. They are aimed primarily at post-graduate students and early-career researchers, but the approach is also suitable for more experienced researchers who want to start learning more about digital preservation. As many digital preservation issues are a good fit to other areas frequently covered by training courses, such as Information Management, Reference Management or Project Management, rather than provide a standalone training course, we would develop a series of short training modules – one on each area – which could be slotted into related courses already being run by university training providers. These could be librarians, research skills trainers or lecturers.

The digital preservation issues we selected were:

- Storage ('Store It Safely'). This covers some issues around backing up data, but the focus is on making sure that files can be accessed in the future through being stored on appropriate hardware and the file formats have been carefully chosen.
- Documentation and metadata ('Explain It'). This describes the importance of providing suitable documentation and metadata to digital files, and the types of information that may need to be provided in this metadata.
- Data sharing and re-use ('Share It'). This looks at why researchers may want to consider making their data more widely available to other researchers or the wider public, and how they might go about doing this.
- Planning ('Start Early'). This examines the importance of starting to think about digital preservation issues at the start of the project rather than at the end, and highlights areas which need to be considered.

Slides and explanatory notes have been produced to form a module on each topic. The notes are included in the PowerPoint presentation (using Office 2010) and are also appended to this document. Each module is expected to last around 10 minutes. The modules are independent so they do not all have to be used in the same course or in any particular order; presenters can pick those with the most relevance or interest to their intended audience.

These modules complement a leaflet 'Sending your Research into the Future' produced in collaboration with the LSE and University of London.

A further set of slides provides a brief introduction to data to give some additional context if required (the 'What is 'data'?' module).

The layout of the slides is deliberately non-branded; the presenter can therefore chose to put them into an existing template and integrate them into an existing presentation in a more seamless way, or leave them as distinctive independent modules, depending on preference.

Differences in presentation style are also recognised and presentations have the following options:

- Mainly pictures. For most slides, all explanatory information is to be provided by the presenter.
- Pictures and bullet points. This is a more guided style which provides prompts for the presenter.
- Text only. The most basic style allows for easiest adaptation to other PowerPoint templates and styles.

Each style is available as a separate set of slides.

Note that there may be differences in display between different versions of PowerPoint, or if an alternative presentation package such as Keynote, is used. Note that some slides have a small amount of animation.

With some small adaptations, the five presentations could be combined to provide one digital preservation module. As 'digital preservation' as a phrase is not likely to generate much interest, a possible title would be 'Sending your research into the future'. A suggested order is:

1. What is 'data'?
2. Explain It
3. Store It Safely
4. Share It
5. Start Early

The presenter should choose the combination and order that they feel best suits their target audience.

Discussion questions

For each topic, we suggest some questions that will allow the presenter to go into more depth or to initiate interactive group work if they wish. These may be particularly helpful if the modules are combined to a single course.

What is 'data'?

Participants are encouraged to think about what they mean by data, to explore what their own data are, and to discuss what limitations and/or problems are posed by their definitions. Questions for this area could include:

- Does it matter how 'data' is defined?
- Are some types of data (words, numbers, pictures, audio, video…) more 'worthy' than others?
- What is your data?
- Does data have to be digital?
- What are the differences between looking after digital and physical/analogue data?

Explain It

Good issues to explore in this area are the pitfalls of not understanding data at a later date, how to decide what to include and how this might affect future use, and producing documentation in a time-efficient way. Suitable questions for discussion may be:

- Do you document your data? If so, how? If not, why not?
- How important are the original reasons for data collection for future research?

- Does whether you document your data, or the way in which you do it, affect:
    - how you go about your work (e.g. analysis)?
    - the way that you communicate your research?
- What are potential advantages to producing documentation as you go? What are possible disadvantages?

Store It Safely

This area is particularly good for participants to share tips, tools and experiences. While back-up is only one aspect of this module, it is a good area to explore; however discussions should – ideally – look beyond back-up to more long-term storage issues. Another good discussion point here is selection and appraisal. Possible questions are:

- Should you store everything? Why? Why not?
- How do you go about identifying what to discard?
- Is it enough simply to store a file securely?
- Digital vs physical: do they need to be treated differently?
- Does it matter if things get lost? How would data loss affect you now? How could this affect research in the future?

Share It

Data sharing and open access issues can provoke strong opinions, and experiences are often heavily determined by the subject of the research and the conventions in that field. These questions may be controversial:

- Should data from publicly-funded research be publicly available?
- Do you want to make your data available? Why? Why not?
- Does the timeframe for making data available matter?
- Is it enough to make publications available? Are the data needed too?

Start Early

This subject is a good opportunity to get participants to think about the whole of their research, particularly if it is the last module to be used. Questions may therefore focus on individual issues and concerns rather than the philosophical questions around data management. This would depend on the audience and the amount of time allotted for additional discussion. Possible areas for questions and discussion are:

- Why is this the last module?
- What is the end point of your research?
- What is the start point?
- Where on the timeline of your project are you?
- Does it matter if you don't follow a plan?
- What concerns do you have about data management planning? What else do you need to know?

Module scripts

What is 'data'? script

Slide 1: What is 'data'? (And why you should care) [Title slide]

Slide 2: What is 'data'? [1]
[A series of questions for group discussion]

Slide 3: What is 'data'? [2]
One of the first things to consider is what is data?
Typically, definitions vary, and may depend on what is being studied. But when considering data management it is good to take a broad definition.
Most people think of 'things involving numbers' as data (such as spreadsheets, numerical experimental outputs), but data is far from being limited to just numbers. Examples can include:

- Emails
- Videos
- Audio files (e.g. oral history, interviews, focus groups)
- Websites (including all sorts of sites such as social media sites, as well as established academic sources)
- Computer source code
- Books
- Papers
- Works of art
- Catalogues, concordances and indexes

It doesn't matter what your area of research is, you will be dealing with data in one form or another, and that data might not be digital.

Slide 4: What is 'data' [3]
That's why the 'any information you use in your research' definition works well for thinking about data management – it makes sure you don't miss out something important!

Slide 5: Why you should care [1]

Slide 6: Storage media are fragile
Digital data are fragile – and can be much more so than physical data such as paper.

Slide 7: Don't let this be you [lost data poster]
Losing crucial research material is the stuff of nightmares… but nightmares come true sometimes.

This is a genuine poster from a pub in Cambridge [I have only altered the picture to straighten it, change the contrast to make it easier to read, and remove some of the details, e.g. the address of the pub and the person's contact information]

You might think 'Ah, but I would take more care of my laptop/external hard-drive/back up disks' …

Slide 8: What would happen to your data if there was a fire in your office, department or home?
…but sometimes things are out of your control. Fires in university departments and buildings seem to occur every couple of years (there have been at least two so far this year). If you do most of your work at home, or in a library, then fire, flood or other catastrophe are still risks, and can take a long

time to sort out. While you're busy salvaging what you can and cleaning up, you don't want to have to worry about whether you've lost significant amounts of work and the underlying data for your research. There are also all too common instances of loss, theft and computer viruses, even if you are careful; it's good to be prepared. Thinking about good information management will help you to deal with these events, should they happen to you.

Slide 9: Can you find things you know you have when you need them?
Data can also be hard to find again when you need them, just because you're not sure where they are. Having a system in place to look after your data is another way of looking after them.

Slide 10: Why you should care [2]
When you finish a project, others may want to build on the work that you've done. If you're doing a PhD, maybe a project student or another PhD student will want to reproduce some of your work to learn techniques and then be able to develop them further. Or you might carry on the work as a post-doc. Researchers from different groups, different institutions, or different disciplines, may also be interested in making use of your data. This is something that can enhance your reputation – and it's much easier to make a good impression when you information is well managed.

Slide 11: Send your research into the future
Looking after your data gives them a life when you've finished using them for your own research. This might be something that you want to do, or something that you need to do to fit with requirements posed by your department, university or funding body.
Another way to think of it is that the future doesn't have to be very far away – it could be you looking at earlier research a few months or years after you created it. So you can get tangible personal benefits (if you're not in an altruistic mood!)

Slide 12: Credit and license information

Store It Safely

Slide 1: Store it safely [title slide]
You'll be aware of the importance of backing up the files on your computer. But are you aware of some of the key things you need to consider when you want to store your digital material safely so that you can still access it 6months, 1 year, 3 years, 10 years from now?

These slides highlight some of the important considerations, and why they're important to you – no matter what your research is on. And they apply to your personal digital records (like music, documents and photos) too.

Slide 2: Make multiple copies… and keep them in different places
Back up is probably one data management thing that most people are aware that they should be doing, or doing better. It's actually a good idea to have more than one back up copy, particularly of important and/or irreplaceable material; this is part of the LOCKSS principle (Lots Of Copies Keeps Stuff Safe).

It's also a good idea to keep these copies in different places, for example you might keep a copy of some material in a cloud-based service (WARNING: if your research deals with sensitive data you may not be able to do this), on an external hard-drive or on DVDs/CDs. Consider asking a friend/colleague or family member to look after one copy, or keep one copy at home and one in your office, so your material is physically in separate places. This minimises the risk of data loss in the case of flood, fire or theft.

But remember that back-up isn't the same as preservation – it's just one aspect of it! If you have made a back-up copy of your data, that means you now have two copies in total to look after. But the good news is that this greatly reduces the risks to your data, and goes a long way to helping it stay safe over time.

Slide 3: Use open/standard file formats where possible…
You might not have given too much thought to the format you're saving it in; perhaps you mainly use the standard packages on your computer or the default outputs from equipment (such as a digital camera).

You might use the formats that other people in your research group/at your work use as this makes it easier to share information.

Slide 4: …they're more likely to be readable in the future
Just be aware that over the long term, proprietary formats can become obsolete and you might find yourself going back to an old file and not being able to open it. Some formats are better because they are open, in widespread use, or conform to standards, and it's best to use these for the long-term preservation of a file once you're no longer working on it (even if you use a different format while you're actually working on the data).

And remember that storage media don't have infinite lives either. As well as becoming obsolete – when was the last time you saw a computer with a floppy disc drive? - hard drives die, CDs and DVDs fail, memory pens get lost down the back of the sofa or end up in the wash… Be aware of the expected lifespans of various types of backup media and use this to help you choose the best option for you. It's a good idea to contact your IT support person about what secure, backed-up storage is available; remember that this can vary between departments. You may be able to arrange secure

access both on and off campus, and this is much more secure than relying on portable storage options such as data sticks.

Slide 5: Organise your files [animated slide]
The books in this photo are actually in vague order around their subject, but finding and accessing any particular book is likely to be difficult. It's also very hard to tell quickly if the book you want is even in one of those piles.

Having a system, particularly one that you've explained and documented, means that it's easier to find things again in the long term.

Key things to consider here are:

- the names you give files and folders
- use of tags, or information in file properties
- whether you need to conform to any conventions in your field

Slide 6: Don't keep everything. Be selective about what you keep and what you throw away
It's very easy to fall into the trap of keeping everything because digital file storage is now so cheap to buy.
This is OK (though probably not very efficient) - as long as you can find things again later AND BE SURE THAT YOU'VE FOUND THE RIGHT VERSION.

It can be a lot easier to find files you need if you've made sensible decisions about selection as you go. Selection is important, but how do you go about doing it?
Ultimately much of the choice will be personal, but here are some things to consider. Remember that these are questions that you need to ask yourself about your material – there might not be a right or wrong answer, and if you're not sure you should get a second opinion:

- Law/policy/mandate – Do you actually have a choice over whether to keep or throw away? Are you obliged to keep the data? Are you obliged to destroy it?
- Emotional attachment – a surprisingly powerful force – don't underestimate it; just be aware of it as a factor
- Cost – would it cost more to keep than to get again? Did you have to pay for it in the first place (or would you have to pay for it now?)
- Do you need to keep both raw and processed data? Do you need to keep the means to process raw data?
- Should you discard anything before the end of a project?
- Should you keep drafts of documents? You might want evidence of the process, or it might be the final copy that counts. You might need to put a pre-print of a journal article in an open access repository rather than the final published version.

Remember that for everything you keep, you should be able to find and access it again in the future.

Slide 7: Store it safely [summary slide]

Slide 8: Credit and license information

Explain It script

Slide 1: Explain It: Why your research deserves good documentation and metadata (title slide)

Slide 2: Why create documentation?
Good documentation does take time to produce – and so it's very tempting not to bother.

While it might seem like a waste of time, there are several reasons why it's important to do, and these will probably save you time in the future!

Slide 3: Make material understandable
First of all, because documentation should be thorough it will contain a lot of information that might seem obvious. But will that same information still be obvious in a few months, years, decades, centuries… time?

It's very easy to assume that you will remember it, but it's quite easy to forget crucial information. It also means that other people can understand what you've done and why. It's important to include context (why you did your research, how it fits into other contemporary research, or follows on from previous work), as well as explaining your methods and analytical techniques. This is related to the next point…

Slide 4: Make material reproducible
By providing documentation, you can provide the methodology of how you generated, collected or produced your data (for example information about collection strategies, interview methods, survey techniques, algorithms, database searches), and how you reached your conclusions from your data (for example any statistical methods you used). This is useful for you if you need to replicate or adapt or re-purpose an aspect of your research method later on.

This is important as it means that people can reproduce your research, either to verify your conclusions or as a starting point to develop your work further. In many research groups, this could be a student or post-doc who continues work started by a previous group member. Replicating methodology can also be a useful training tool.

Key points:

- Detailing your methods helps people understand what you did (and why)
- Explaining your algorithms, search methods etc makes your work reproducible
- Conclusions can be verified

Slide 5: Make material reusable
One of the main advantages of creating documentation is that it makes data re-usable. This doesn't have to be altruism – it can be by you at a later date. Besides, making your data available has benefits for your reputation, so documentation doesn't have to be altruistic even if you don't intend to re-use the data yourself.

Slide 6: Documentation and metadata
Documentation is information about the data that is human readable. Metadata is the same thing, but machine readable. This has important implications for searching for data. The structured machine-readable form of metadata means that it can make things easier to find. Think of it like tagging a photo in facebook or on flickr. The more comprehensive it is the easier it is to find things, and you can never be quite sure what other people will be looking for. But providing better metadata increases the chances of finding relevant information.

Slide 7: Make material findable
Producing good metadata means that it's easier to find your data, as it highlights the important aspects in a machine-readable way. This makes computer-based searches, whether on your searching your own hard drive or looking for something on a database online, work better for you – they're more likely to find relevant files and information more quickly.

If you're working on a large project you might be interested in crowd-sourcing metadata production. This works well with niche communities who are active online (such as transport, or local history).

It's easier to produce good metadata when files have also been documented!

Slide 8: What to include (I)
These slides outline the standard information that are generally included in documentation and metadata; you should think of these as minimum requirements, and add anything else that you think is relevant to understanding your data.

A good way to decide what to include is to think of the information that you wish you'd known before you started collecting your data, or the sort of information you would want to know about a digital data set (whether it's generated by a machine, a hand-crafted database, a series of images or audio files, or any other type of data) without having to look at the data set itself. You should also think about how you would go about searching for relevant material for your research and the types of keywords that you'd use. This will help you assign metadata to your own material.

Some of these areas are to provide context around the data being collected, which might include information about funding bodies and grant numbers, prevailing social, political or scientific developments, customs or practices. It might be relevant to include further details about the project not just the data set.

The description of the item can cover both a description of the digital item itself and a description of any physical object that it represents; you could think of this as telling someone what the item is and what it looks like without them having to look at the digital files themselves.

The methodology doesn't just apply to science and social science data sets, as it will include issues like your theoretical approach, or critical framework, as well as experimental details. You might want to include information about why you used the method that you did, and why you chose it over other possible methods or approaches.

If you take any measurements, it's important to include units; it may not be obvious to other users of your data what you used, even in context, particularly in the future. There are cultural norms in measurements, and so it's important not to make assumptions about what would be used (think of inches vs cm).

It's also a good idea to include any references to any related datasets (collected by you or other people).

Slide 9: What to include (II)
The way that jargon and acronyms are used can change over time, and sometimes very rapidly. It's important to be explicit about what an acronym stands for. It's particularly important to explain terms that have different meanings in other disciplines or if there is any ambiguity of meaning within your discipline.

Much jargon is discipline-specific but has been in use for a long time; it is more important to define newly-coined terms.

If you have used any codes (maybe to help with statistical analysis, e.g. 1=male 0=female) it's important that these codes are made clear.

Technical information about a file can help it's long term preservation; some of this may be generated automatically by your digital equipment, e.g. camera details can be included in the metadata for a digital photograph.

Slide 10: Explain it [summary]

Slide 11: Credit and license information

Share It script

Slide 1: Share it: Making your research available for re-use [title slide]

Slide 2: Why? [1]
Reputation
Sharing data can build your reputation in number of ways. Laying your work open to scrutiny means that you will get credit for high quality research, increased understanding of your methods and allowing your work to be verified by others. Sharing allows you to make a greater contribution to your community. It can also help extend your reputation beyond that community.

Key points:

- Get credit for high quality research
- Increased understanding of your methods
- Allows work to be verified by others
- Recognition for contribution to research community
- Extend research beyond your discipline

Slide 3: Why? [2]
Funding
Many funders require publications and supporting data to be made open access, or be put in a suitable data centre or repository. So you might have to do it to meet your funder's requirements; doing it may make your funding proposal more attractive in cases where it is not essential.

Key points:

- Making data and/or publications available may be a requirement of your funding body
- It may make your funding proposal more attractive when sharing data is not essential

Slide 4: Why? [3]
Impact
Sharing your research makes it more visible: easier to find and easier to access. In fact, there is evidence that making your publications open access leads to increased citations

Key points:

- Sharing makes your data:
    - Easier to find
    - Easier to access
- Open data/publications lead to increased citations

Slide 5: Why? [4]
Reuse
Sharing your research allows it to be re-used; this might be within your field, for example using the data as a starting point for a complementary study, or as test data for new software and algorithms. It might be useful for teaching purposes. It might be re-used in contexts not currently envisaged – for example in new developments several years down the line, or in completely different fields. And you will get credit as your work will be cited each time.

Key points:

- Starting point for a complementary study
- Test data for new software and algorithms
- Teaching purposes

- Contexts not currently envisaged
- Completely different fields

Slide 6: How? [1]
Remember that sharing your data doesn't have to happen while you're still working on it (though it can if you want); it can be made available to be adapted, re-used and built upon when you've moved on to something else (though be aware that funders can specify when data has to be made available).

So how can you make it available? There have been some people who have done all their research 'in the open' using blogs for dissemination; this is a bit unusual though, and there are plenty of other things that you can do.

It is more usual to deposit data in a repository or data centre at the end of a project. You can choose open or controlled access, depending on what makes the most sense for your data. For example, DSpace@Cambridge (the institutional repository at the University of Cambridge) is open access, but the UK Data Archive, which specialises in social science data, is controlled access and users need to go through an authorisation process. In both cases, there are end-user agreements in place, and you will not lose copyright on your material by depositing it. What you will get is skilled data curation.

Slide 7: How? [2]
You can also redact material, for example 3rd party copyrighted material in a PhD thesis, or place embargoes so that it cannot be accessed for a certain period, for example because of publisher requirements or applying for a patent. Such measures may also be necessary with some confidential information.

Slide 8: Share it [summary slide]

Slide 9: Credit and license information

Start Early script

Slide 1: Start Early: Save time [title slide]
Digital preservation might seem irrelevant, or not something that you have to think about until you get to the end of a project.

BUT it's A LOT easier to plan for how you are going to look after your digital material when you create it (or even before) than at the end. There are a number of reasons for this: you might forget why you made certain decisions about some of the material, there might even be things that you now can't understand. Leaving it till the end might mean that it becomes someone else's responsibility, but they will not understand your data as well as you do. You might have made decisions about how to store your material that make it hard to preserve. Going back through lots of data and sorting it out will take time (possibly quite a lot of time) and you will probably be keen to move on to the next project.

So planning to preserve right from the start will save you time in the long run.

Slide 2: Make a plan for what you are going to do with your material (digital and analogue)… both during the project and once it's finished
Planning will help your research in several ways:

- It can act as a road map for your research
- It can save you time (for example by minimising the need to re-collect data)
- It can save you money (for example by helping you decide on the most cost-effective safe way of storing data)
- It can help streamline collaborative work (as you will all have a common set of guidelines)
- It can help you focus on the key questions of your research

Slide 3: Make plans to: Store it Safely, Share it, Explain it
These three areas will help with the long-term preservation of your research. Ask for more information!
[Separate sets of slides are available on each of these areas.]

Slide 4: Get into good habits early in your research… it will save you time in the end

Good habits to cultivate:

- Organise your data from the start. Exactly how you do this will depend on the norms in your discipline or research group, and personal preference (e.g. the Lego could have been organised by colour, but has been organised by type). The system that you use can affect how easy it is to find things. If your system mirrors your thought processes, this will probably help find information again later.
- Give your files sensible names; doing this when you save a file the first time takes hardly any time at all. Going back and renaming files takes a lot longer (and that's assuming you don't need to open them to find out what they are). Be particularly careful with files that are given names automatically, e.g. from digital cameras or experimental outputs as these tend not to be meaningful.
- Back up regularly. Get into the habit early on so it is second nature. This also means that your back up will be comprehensive and there is less chance of missing an important file.
- Document, label and tag as you go. If you take measurements, what units were they in? What equipment did you use? What search parameters did you use on a database search? Who/what is in the photo, and when and why did you take it? Make sure spreadsheets have row and column labels. These sorts of things can seem obvious but can be surprisingly easy

to forget, particularly after a few (3) years. And of course you should also document the less obvious things too! If in doubt on what to include, err on the side of too much rather than too little information.

- Be careful about using copyrighted material; get copyright clearance on third party copyrighted material as you go along (copyright clearance is necessary for any published work, including e-theses) and make sure the permissions for use are clearly associated with the relevant material. Also look out for alternative licenses (such as Creative Commons) and make sure you retain the necessary information to give appropriate attribution.

Remember that these tips will make it easier and quicker for you to preserve your digital material AND make things easier for you as you work on your data! It's a win-win situation!

Slide 5: "The perfect is the enemy of the good" - Voltaire
It is better to start off doing *something* than not to start at all. Don't think that your plans have to be perfect - they almost certainly won't be. And the very nature of research means that the unexpected can, and probably will, happen; and then you will probably need to change your plan in response.

"A good plan implemented today is better than a perfect plan implemented tomorrow" – George Patton.

Slide 6: Start Early [summary]

Slide 7: Credit and license information