How Long is Long-Term Data Archive?

R.D. McDowall, McDowall Consulting, Bromley, Kent, UK.

The Electronic Records and Electronic Signatures Final Rule (21 CFR Part 11) requires data to be held electronically for the record's retention period. How long is this and what are the implications for you?

In the last "Questions of Quality" column we covered back-up and recovery (1) but in this article I'd like to look at the closely related archive and restore. Let's get the definitions sorted out.

- Back-up and recovery is used for short-to-medium-term data storage for immediate recovery of data (from files to disks) in the event of file loss, corruption or system failure. This is a prerequisite for disaster recovery or business continuity planning.
- Archive and restore is concerned with the long-term storage of data and electronic records.

Back-up is a regular process usually occurring daily and performed by the computing or IT department and is concerned with the whole of the system or the data disk. In contrast, archive and restore occurs infrequently and is driven by the users (the archiving process may be performed by the IT department but it is user specified). However the data are usually more carefully selected and will be by work package rather than the whole or part of a disk.

Some of the issues we will look at in this column are

- which electronic records to archive
- how long to archive
- will the application be around at the end of the retention period?
- file formats
- technology changes: will you pick a long-term winner?

What Should I Archive?

As mentioned earlier, archive is different to back-up and you will need to organize the archiving of data around specific work packages that will depend on the type of laboratory. This can vary but some typical examples can be the analytical results and electronic records from the following types of work package:

- a specific analytical method
- a specific batch(es) of the same material
- a specific stability study
- a pharmacokinetic study or protocol
- analytical work for a specific method across a defined date.

These are merely some suggestions because most laboratories tend to work slightly differently. I'm sure you'll be able to find approaches that will suit your needs better other than those listed here.

Note that the suggestions are around specific rather than general analytical work packages. This is important as once you have archived the data you may have to retrieve this at some time. Spend time in designing a simple way of defining your requirements and get it right first time; alternatively hope you have retired/the company has merged or it is someone else's problem if you don't. The issue is that you could spend a long time before it emerges that you cannot easily restore data and electronic records as they are stored in several places or, worse still, the system performing the archive does not restore correctly.

Why Bother?

There is a group of people who think that we should not archive data off-line but archive it on-line instead. As hardware is more resilient and fault tolerant, servers can be purchased relatively cheaply and then we can just buy new disks. As the electronic records that we generate grow in number and total volume, we just keep pace with extra disk storage. As we approach the disk capacity, the simple answer is to increase either the size or number of disks, or both. This is one way of looking at archiving; however, there can be problems with this approach. Usually this approach is taken when there is uncertainty with defining how and what to archive.

System performance may be an issue as the number and volume of electronic records and data files increase:

- Can the file management or database system cope with the size?
- Was this tested and supported by the vendor?
- Is the security and integrity of the records maintained? For instance, is it easy to change records that are approved or final?
- What happens if there is a disaster? Can you recover all of the records? How do you know and have you tested this?

My view is that you need to keep records on-line for a certain amount of time and then archive them off-line. The reason for this is more to do with protection of the records but also overall system performance.

How Long to Retain the Record?

The length of time that a record should be stored is defined in the predicate rule under which 21 CFR Part 11 operates (2).

- For manufacturing laboratories working under good manufacturing practice (GMP), the retention period is the expiry date of the batch of material plus one year; therefore, if you make a product with an expiry of five years you'll need

"Changing file format can have a major impact on the ability of a program to replay data."

to store the records for six years before you can legally dispose of them.
- On the research and development front, the situation is that you will have patent issues to consider as well as the time for product development. Usually you will be keeping the electronic records for 15 years after the last launch in the last country, plus the development time that can take seven or more years from the date of compound synthesis.

However, that's just the regulations to look at. What about reality? Many organizations will retain paper records for much longer than that, mainly for defence against litigation.

What about access to the records? It's likely that most requirements for access to any electronic record will be made early in the life of a record rather than much later. However there are still the regulatory requirements to consider especially in the research and development area of the pharmaceutical industry where development and retention times are the longest.

Can I Replay the Data?

The regulatory requirement is for ready replay of data. Note that this is not instant replay because with an effective archive process, the archived data will be stored in one or more remote locations and will have to be restored back to the appropriate computerized system.

The key issue here is to ensure that whatever archive process you have, it not only works from the computer system but you can restore the electronic records whenever you wish. There are a number of questions concerning changes to your data system that will have an impact on the ability to reprocess chromatographic data. Look at any differences between when you originally acquired the electronic records and when you want to reprocess:

- Are any file formats the same?
- If your electronic records are stored in a database, is the structure the same?
- Have the integration algorithms changed?
- Has the operating system changed?
- Has the application software changed?
- Has the hardware platform changed?
- Has the archive medium changed?
- Has the archive software changed?

As you can see there are a number of changes that can impact the ability to develop an effective archive. Changes in any one of the above can limit or destroy the ability to restore data from an off-line archive. The people who vote for an on-line archive will still have the same problems with many of these questions, except that they are immediate (on-line) rather than delayed (off-line).

This area is one of the most technically challenging problems that we are presented with to comply with 21 CFR Part 11. We'll look at two in the next two sections; those of file format and archive media followed by the issue of data migration.

Changed File Format?

What impact does a file format have on archive and restore? There are a number of issues to consider; however, to give you a better understanding of the problem think back to when you were using a word processor. If you have used Microsoft® Word for a long time you'll remember the problems with migrating from Word 95 to Word 97 that illustrate what we'll face with chromatographic data: the file formats were not the same and the migration route originally used conversion to rich text file format. Even then the problems did not end there, as the conversion was not 100%. A number of features did not work, such as pagination; and the table of contents transferred adequately but all the page numbers were migrated as 1. This was manageable because all you had to do was delete the table of contents and reinsert it, but imagine the audit trail entries for regulated data. Transpose this to your chromatographic records; how would you feel if all your peak areas came out as 1 after a migration from one version to another?

Changing file format can have a major impact on the ability of a program to replay data. Therefore, one of the issues from the 21 CFR Part 11 perspective will be to tie you to a specific vendor unless there are universal standards across all data systems. Although we have ANDI (now ASTM) data file standards for chromatography data systems (CDS), there are still problems with this as ready replay from an archive of ANDI files does not appear to be acceptable to the Food and Drug Administration (FDA). The following is a question and answer session from a conference in Berlin, Germany, in September 1999 and includes the answer from Paul Motise of the FDA. The full conference comments are available from the Labcompliance website at www.labcompliance.com.

Why is chromatographic ANDI data not an alternative for long-term archiving and ready retrieval?

Question from Dr Kiechle, Novartis: My question is regarding analytical instruments and chromatographic equipment, such as liquid or gas chromatographs. In my department, chromatographic data are electronically stored and archived in original format and are available for reprocessing any time, such as when inspectors are visiting.

However, after 10, 15 years or so the vendor company of this equipment may have disappeared and the software or hardware is no longer available. In these situations we file all this original information as ANDI converted files because we think that ANDI conversion has the broadest future and in 15 or 20 years from now we can get the data back into ANDI format but maybe not into the original raw data format if the company no longer exists.

Unfortunately these ANDI data can no longer be reprocessed in the same way as the original ones can. We will have these data on the screen, together with the sequence of injection, with the integration methods, with the operator's name and the audit trail. So we have everything in the ANDI format, but just for viewing. We cannot reprocess the data any more. I would like to get your interpretation as to whether these data will be accepted as original data in that sense of Part 11 in 12 or 15 years time?

Answer from Paul Motise, US FDA: This is something that we are going to have to develop agency guidance to explain our expectations in more depth. In the preamble to Part 11 we explained that the agency did not expect companies to save computer hardware and software for the sole purpose of recreating events. We anticipated that it would be possible to make an accurate and complete copy of those electronic records. Now there are a couple of things involved here, first of all, the length of time: in a good manufacturing arena you would be required to keep data for one year after the expiry date, a typical expiry date might be 3 to 5 years. So a projection of 10–15 years is probably more than what is required by the FDA. Consider the nature of the record and the corresponding predicate regulation will tell you how long

you have to keep that record. Part 11 says if you keep it in electronic form you must preserve it in electronic form.

Now what does preserve mean? This is something that we will address in further guidance. My own perspective goes something like this: when you have an electronic file, you have data, straight numbers, you also have meta-data, you have something that turns that information, that data, into knowledge, something that you can use. You have the hardware and software and operating systems. Add all of those pieces up and you have the bottom line: knowledge. This turns bits, ones and zeros into something that makes sense.

When you convert from one system into another as part of your archiving because a vendor may be going out of business it's important to be able to preserve that knowledge. You want to take a look at the method of conversion to make sure that you are not making things look better or worse than they really are, to ensure that that you can still make sense of that information and use it the way you need to use it. In the GMP context, GMPs requires you to keep all laboratory data for as long as the batch record must be kept and that includes the chromatographic raw data itself. That's nothing new even aside from Part 11. Firms find it to their advantage to be able to keep that data in such a way that they could run the sample again and perhaps they find impurities in the future that they did not find before. There is real value in doing that. So I would tell companies to make every effort to preserve that ability for as long as they possibly can. Again we will address this further in agency guidance. That's my impression from now, and I hope that makes sense.

The problem is that we have no universal data format for all electronic records created by a CDS. Until we do, the impact is to restrict users to a single vendor who has the responsibility of either:

- Maintaining stable file formats or
- Providing fully working and documented conversion routines to enable data migration between one version of a file and another.

In either situation, my view is that this will restrict you to a specific CDS vendor until users press vendors for interoperability between different vendors' data systems. Most CDS users can't see past the taps on their laboratory bench so asking them to put pressure on vendors is unlikely to happen and vendors won't develop these systems. The pain of moving between vendors has to be less than the pain of staying with your current one or you won't move. Select the right system and the right vendor or you will be paying a high price in the future.

Did You Buy a VHS or Betamax Video?

Technology is advancing at an amazing rate; however, what is the impact of selecting the wrong hardware? Remember the video wars in the 1980s: the choice was either VHS or Betamax. Technically better, the Betamax format lost out to VHS. Fine if you just recorded from the TV but if you had bought films you had nowhere to go as there were no easy data conversion options to VHS.

The same problem is occurring now with archive media: what options would you choose? Some options to ponder are

- floppy disk
- zip disk
- magneto-optical drives
- CD-ROM disks
- DVD disks.

Let's reject the floppy disk as not an effective option (many have tried though!) and look at the other options.

- Zip disk is proprietary and only has a capacity of either 100 or 250 MB. It uses the same technology as the floppy disk and can be erased by a magnetic field.
- Magneto-optical drives have larger capacity but don't have a universal standard and are stable against magnetic fields.
- CD-ROM has a de facto standard based on the Sony-Phillips co-development of the technology with a reasonable capacity of 650 MB.
- DVD has larger capacity but there are many standards and capacities and it is not known which one will succeed.

Not very encouraging, so the safest approach is to use CD-ROM as this is the best standard. However, it is a relatively mature technology with relatively limited capacity. Unless you are creating large numbers of diode array detector files then a CD-ROM disk capacity is usually adequate for most CDS work packages.

Data Migration Issues

Before we conclude this column, let's consider data migration, as this may be a key factor in the ability of a laboratory to replay data and get the same result as originally obtained.

Usually data migration is considered when you move from one system to another (3); however, now you'll have to consider the impact of data migration within the same vendor's system if a change is made that impacts the ability to acquire and calculate results.

The most difficult will be migrating data from one vendor's system to another and is discussed in reference 3. However, don't be complacent; even if you are installing the next release of your current data system, ensure that you read the release notes to see what has changed in the new system; if there is something that impacts the ability of the system to repeat the calculations you have performed earlier, then you need to assess the situation.

Equally vendors must make efforts to reduce the impact of such changes and provide either full information on the impact of the change and, if there is a major impact, provide working and fully documented solutions.

Summary

We've looked briefly at archive and restore and the ability to comply with the requirements of 21 CFR Part 11. This is a problem area and will continue to be for a few years until encompassing solutions are provided. To minimize the impact ensure that you are aware of the changes that may impact your data system and test them fully before implementing them.

References

(1) R.D. McDowall, *LC•GC Eur.*, **14**(6), 282–287 (2001).
(2) 21 CFR Part 11, Electronic Records, Electronic Signatures Final Rule, Federal Register 62 13430–13466 (1997).
(3) R.D. McDowall, *LC•GC Eur.*, **13**(1), 30–35 (2000).

Bob McDowall is Visiting Research Fellow in the department of Chemistry at the University of Surrey, and Principal of McDowall Consulting, Bromley, Kent, UK. He is also a member of the Editorial Advisory Board of LC•GC Europe.