Challenges of Long-Term Archiving in the Pharmaceutical Industry

Anita Paul (Roche, Basel, Switzerland); Juerg Hagmann (Novartis, Basel, Switzerland)

Abstract

What is unique in terms of "trusted digital preservation" in the pharmaceutical industry? What are the specific legal and regulatory requirements and what are the typical types of data concerned in the core business processes of research, development and manufacturing? Life science is a "high-risk industry". While products are targeted to treat illness and prolong lives they can also be life-threatening if used wrongly. In order to gain approval for marketing authorizations, terabytes of data and documentation are being generated to prove efficacy and safety of a compound. This data will be subject to regulatory review before an approval is granted. Global Good Practice and resulting national regulations require the retention of vital data up to several generations of software and hardware. This paper discusses some of these problems, drawing on the experience of building practical solutions in the context of an example in the area of registration documentation.

Data in Pharmaceutical Industry

The life cycle of a typical pharmaceutical product spans over several decades from first discovery through development to marketing authorization and distribution until the product is withdrawn from the market. Even after decades old compounds may be rediscovered for use in new indications. In addition, during the life time of a compound, terabytes of data and resulting documentation (records) are being generated to support the marketing authorization and further extensions to this authorization. Multiple partners are normally included in the development of a drug including third parties. An actual discovery may have happened in a small biotech company or university off-spring, development activities are then spread over several Contract Research Organizations (CROs), while the marketing may be shared by two Pharma companies and production could be done as toll manufacturing. This shows how broad the variety of sources and originators of data production may be, indicating potential issues of custody and management of data ownership through time. Not less is the variety of types of data which are generated by very specific (and often proprietary) systems and high tech equipment. Examples include: Compound screening systems (incl. spectral data), Laboratory information management systems (for analytical raw data e.g. chromatograms), ECG devices, mathematical modeling and statistical data systems (SAS), genome data systems (producing high volumes), manufacturing batch control systems, clinical trials databases all including highly specialized hardware, software and output generation devices.

Considering these aspects two major problems become evident: how to manage the variety of data and how to identify and control the large amount of vital data requiring long-term retention?

Today a typical submission to a Health Authority still consists of a large amount of paper which was scanned to PDF format, generated from and/or summarizing some of the source data described above. However paper submissions will in near future no longer be accepted by major Health Authorities [1] and regulatory requirements are often more advanced as the business environment in the real world. The use case of "Registration Documentation" below will show the difficult balance between business and regulatory requirements. Already the common rule to keep registration related documentation as long as the product is on the market plus 10-15 years is ambitious (possibly indefinite for some products) but there are other challenges interfering like the management of digitally signed documents from third party providers or the new requirement of many countries requesting now a "multiple generation" approach, namely to monitor the safety of biotech products on the genetic pool.

The following section describes why the industry has to adhere to strict legal and regulatory requirements related to data protection and retention.

Legal & Regulatory Requirements (Compliance)

The pharmaceutical industry is highly regulated by international so called Good Practice (GxP) guidelines, also called "predicate rules", that mandate what records must be maintained, the contents of those records, whether signatures are required and how long the records must be retained. GxP guidelines are primarily Good Laboratory Practice (GLP), Good Clinical Practice (GCP) and Good Manufacturing Practice (GMP)[2], ensuring the quality of the processes leading to the final product. In addition to these GxP requirements which all contain provisions about set time periods for records retention, the industry has to comply with regulations from national law and health authorities, namely the U.S. Food and Drug Administration (FDA) which has become a quasi formal global standard for the requirements of Computerized System Validation (CSV) [3]. What is particularly challenging in the focus of our long-term archiving context are the requirements outlined in the original provisions of CFR 21 part 11 for computerized systems regarding electronic record and signature handling [4].

The industry is facing three major implementation challenges for this rule: (a. Records generated in electronic form must be stored and retained in electronic form. Printed material is not a substitute for an e-record. (b. Records must be

stored as complete and accurate copies. A complete copy includes metadata, such as processing and integration parameters and audit trail logs. (c. Records must be readily available throughout the entire retention period. Inspectors want to be able to replay data using the same process as when the data were initially generated [5]. The FDA is also enforcing predicate rules related to record keeping requirements [6].

Despite of the fact that the FDA has re-examined some of their original requirements in 2003 [7] which allow more discretion in terms of hybrid situations (paper and e-records), generic file formats and even allow to archive required e-records "in nonelectronic media such as microfilm, microfiche, and paper, or to a standard electronic file format", the mentioned implementation challenges remain, as predicate rules and risk considerations of pharmaceutical companies prevail due to the reasons described above. The FDA has also increased the controls related to part 11 from 2007 onwards [8].

Based on such requirements it becomes obvious that digital preservation does not just mean to be able to read specific data in a long-term (rendition) format after several years but to be able to "ready retrieve" and rework e.g. analytical raw data for a reanalysis of old information using newer techniques or visualization techniques ("signal detection").

The FDA and its representatives have defended their position several times. "FDA wants to use the same tools to evaluate the data that the operator uses to create the data. (…) FDA wants to take advantage of modern electronic search tools, which are expected to make inspection work more efficient. Without such tools, inspections would take longer to complete, resulting in delays in the approval of new medical products."[9]

In addition, the FDA has clearly stated that the part 11 regulations extend beyond the retirement of a computerized system [10]. This requirement has lead to the fact that many in the industry are retaining old computer hardware instead of a structured approach to system migration.\

Standard Approaches

In terms of IT solutions for long-term archiving the industry faces the same problems as all Health Authorities and other institutions that are responsible for the custody of large amounts of data and records. The dream of a robust and infrastructure independent platform has not become much closer as we also see how NARA is struggling with its ERA project [11].

One approach which is still being further developed is a long-term archiving platform and service based on the OAIS model and VERS (self-describing objects) [12]. Such a solution (provided it is validated) fulfils a lot of current functional requirements of long-term archiving and is capable to cover the variety problem of data as all kinds of business applications incl. ERP systems can feed the archive. It's the core purpose of such an archive platform to live much longer than all the applications which feed it.

- Formats: compressed PDF/A, TIFF, ANDI (Analytical Data Interchange Protocol [13]) and others; PDF/A compliant PDF 1.4 is for example the required submission format by US FDA, mainly to support long-term retention of the FDA's archival copy . [14]
- Storage media: optical disk (juke boxes) and write once hard disk with time stamps to ensure integrity.
- Redundancy/integrity: Checksums are calculated to ensure the integrity of the data. The checksum values are stored together with the archive packages.
- Automated replication of archive packages to remote location.
- High-availability of data (online). Ability to rework a package. Automated package submission for selected content (interface).
- Browser front-end for submission and download of archive packages.
- Identity management with an authorization module; allows over all search for legal discovery users.

Due to shortcomings in terms of package size the following approach allows the solution of the volume problem to a certain extent.

With **File System Archiving** (FSA) you can migrate content out of file systems and replace with a shortcut to the archived content. When users click the shortcut, the archived documentary records are retrieved, creating a seamless experience for end users. FSA is a method to conserve space on a file server and store content indefinitely and securely, applying strengths of archiving capabilities. The solution also enables to capture files or directories with multiple files from an ECM or E-Mail system. Advantages are: Well and reliable performing transfer (submission or retrieval) of high-volume packages (e.g. for studies) up to 300 GB or more, with large files (max. 4 GB currently) and an almost unlimited number of files within a package (folder). Other important advantages are that the IT departments can offer robust and quick solutions to knowledge workers, such as: 1. Enough, affordable space for these documents and 2. To relieve the knowledge workers from all unnecessary, administrative tasks concerning their documents.

FSA enables you to automate the process of storing content safely in multiple physical locations or on hot stand-by devices. In addition, you can automatically render content into standardized formats such as PDF or PDF/A and TIFF to ensure future readability by customizing the pipeline. FSA solutions also include versioning, controls for single instance and full-text search across archived file systems into a single result set.

Concerning a structured approach to system migration for laboratory data in order to meet the requirements of reprocessing old data for inspections, we refer to the article of McDowall [15].

Beside of the approaches described above there are widespread storage solutions for "ready retrieval" which do hardly fulfill best-practice requirements of archiving. They just offer some pragmatic protection of the integrity of the data, secure access and minimal sets of metadata.

Outsourcing of critical data to third party providers (digital vault) is neither a common approach mainly for confidentiality and availability reasons.

Use Case "Registration Documentation"

While other Health Authorities have not been that explicit in establishing their computer system and e-records

expectations in the past as the FDA has been, the move to accept only electronic formats for new drug approval submissions has now raised the need for those authorities to specify how they can accept data to facilitate a consistent review and long term retention according to national archiving requirements.

The International Conference of Harmonization (ICH), bringing together representatives from Industry and Health Agencies of major regions, has set up specifications on how submission data will need to be presented [16]. This specification is based on the content structure previously defined for paper submissions ("Common Technical Document").

The eCTD is defined as an interface for industry to agency transfer of regulatory information while at the same time taking into consideration the facilitation of the creation, review, lifecycle management and archival of the electronic submission.

The eCTD defines an XML DTD backbone with attached leafs of files. The purpose of the XML backbone is firstly to manage meta-data for the entire submission and each document within the submission. It also constitutes a comprehensive table of contents and provides navigation aids.

Allowed formats for content files are PDF for narrative data (documents), XML for structured data. For graphics it is also recommended to use PDF whenever possible. As appropriate or when PDF is not possible, JPEG, PNG, SVG, and GIF are accepted.

Details for the file format production (i.e. for graphic formats) are also provided as part of the specification. The integrity of the overall submission package is guaranteed via a defined checksum value. A stylesheet should be included into the submission to facilitate the viewing.

The specification is not only targeted at defining one time submission requirements but intends to support the full life-cycle management of a pharmaceutical product via a specific command set specifying allowed actions on the nodes of the backbone (append, modify, delete). This allows that subsequent information submissions loaded into the agency's repository can be linked to the initial submission. Information items can be amended and replaced through the commands included in the meta information without the need to resubmit the entire documentation/data.

Global Pharma companies are now faced with the business need to know the proper life-cycle status for the different agencies (current view "valid data for product x and country y"?). In addition they need to maintain an archives copy for each record submitted (historic view "which data segment for product x in country y at what time point"?).

Conclusions

Regulators and best-practice frameworks require a lot of high quality standards of records, for which the real world business in the pharmaceutical industry has some difficulties when it comes to long-term preservation. In general a lot of pragmatic solutions prevail hardly meeting expert requirements of archiving, depending on the size and resources of an organization. Research and literature in this field is not yet advanced as one could think maybe also due to the lack of international communication across the industry. The authors propose a more active promotion/coordination of industry specific long-term archiving standards through associations such as the ICH or via the professional industry association "Drug Information Association" (DIA) which runs a special interest group dedicated to electronic document and records management. However there are some reasonable approaches as shown above. They all help to make progress in achieving the following benefits:

"- Increasing the quality of decisions by making up-to-date information based on historical data available to all who need it
- Allowing information reuse regardless of geography
- Reanalysis of old information using newer techniques or visualization techniques
- Extraction of new value from old records via data mining
- Less time wasted wondering about data sources: one search will provide it all" [17]

These are issues and challenges around which all records managers and archivists are struggling in the pharmaceutical industry and which have not changed since the last Pharma specific archive conference in France in 2003, organized by the French Association of Archivists [18].

References

[1] The European Medicine Agency has recently released the new requirement , that from July 2008 **only** electronic submissions will be accepted for new drug applications. See: http://www.emea.europa.eu/pdfs/human/regaffair/56336607en.pdf

[2] http://en.wikipedia.,org/wiki/Good_Manufacturing_Practice, http://en.wikipedia.org/wiki/Good_clinical_practice, http://en.wikipedia.org/wiki/Good_Laboratory_Practice

[3] http://en.wikipedia.org/wiki/Computerized_system_validation

[4] Code of Federal Regulations, Food and Drugs, Title 21, part 11, sections 11.10(a) and 11.10(b), "Electronic Records; Electronic Signatures; Controls for closed systems, Washington 1999

[5] L. Huber, W. Winter: Implementing 21 CFR Part 11 in Analytical Laboratories, part 4, Data Migration and Long-Term Archiving for Ready Retrieval, in: BioPharm, 13(6), June 2000, p.58

[6] "We will enforce all predicate rule requirements, including predicate rule record and recordkeeping requirements". (see Ref. [7] below)

[7] FDA's final "Guidance for Industry: Part 11, Electronic Records; Electronic Signatures - Scope and Application", August 2003 http://www.fda.gov/cder/guidance/5667fnl.htm, (last visited 28.3.2008); see also: W. Winter, L. Huber: Part 11 Is Not Going Away. The new electronic records draft guidance, in: BioPharm International, May 2003, pp.28

[8] http://www.labcompliance.com/newsletter/2008/03_mar-newsletter.htm (last visited 28.3.2008)

[9] see Ref. [5], p.58/59

[10] see Ref. [5], p.59

[11] http://www.archives.gov/era/index.html (visited 30.3.2008)

[12] Victorian Electronic Records Strategy (Australia): http://www.prov.vic.gov.au/vers/vers/default.htm

[13] see: E1947-98 Standard Specification for Analytical Data Interchange Protocol for Chromatographic Data (www.astm.org)

[14] FDA CDER Portable Document Format specification. Version 1.0 - 2005 http://www.fda.gov/CDER/regulatory/ersr/PDF_specification_v11.pdf

[15] R.D. McDowall, Chromatography Data System V: Data Migration and System Retirement, LCGC Europe 13(1), 2000, p.30-35; a summary is found in Huber L. (see ref. no. [5], p.63)
[16] Electronic Common Technical Document Specification V3.2 http://estri.ich.org/eCTD/
[17] Robert Sharpe, Digital archiving in the pharmaceutical industry, (Tessella), Oct 2006
[18] see: La gestion des archives dans l'industrie pharmaceutique: spécificités et similitudes. Journée d'études de la section archives d'entreprises, Lettre des Archivistes AAF, No.70, nov. 2003-janvier 2004

Authors' Biographies

Anita Paul received her BA in library science from the University of Frankfurt (1984) and her MA in information science from the University of Konstanz, Germany (1990). She has worked in several assignments in the library and information science field, including the University of Kassel and the German National Library, working on open standards for library interconnectivity. Since working with Roche Pharmaceuticals Division she has focused on the challenges of Records and Information Management in regulated industries, charing and participating in various internal and industry-wide groups dedicated to records management.

Juerg Hagmann has received his MA in Economic and Social History and Constitutional Law from the university of Berne (Switzerland) (1988). Since then he has worked as an archivist and librarian in different companies in the private sector. Since 2003 he works for Novartis in Basel; first as a global archiving program manager and then as a global education and training manager for records management in the team of Group Records Management. He acts as a chairman for the eArchive committee of the Swiss Association of Archivists.