Preserving Digital Records in Industry

DPC Forum with Industry

5th June 2002, London

Philip Lord

Digital Archiving Consultancy

philip@philiplord.com

Programme

- A review of problems with digital archiving in industry
- Based on experience establishing a data archive in a commercial, science- based enterprise
- Discussion of:
  - Business Drivers
  - Data issues
  - Management issues
  - Current status
- Examples from pharmaceuticals

First -

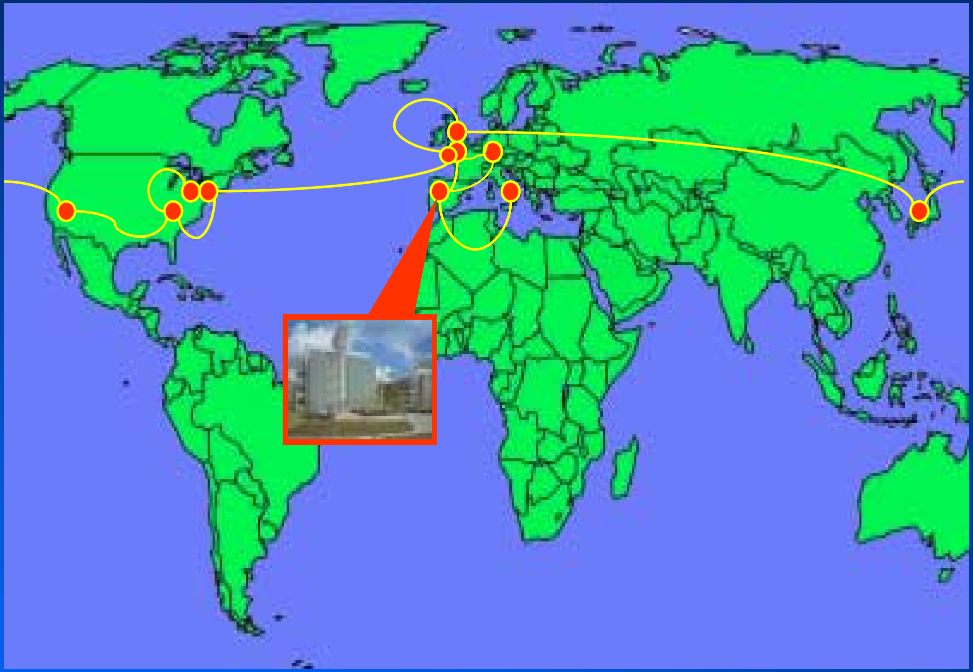
– quick tour of the scientific and technical data generation environment in larger organisations.

A tour of the local environment

A typical, large modern commercial R&D facility

- Many laboratories with different instrumentation and systems
- But part of a larger R&D organisation

To the wider organisation



These facilities will be distributed internationally (perhaps in clusters) and connected by networks.

Into the laboratories

An imaging application – fairly simple file structures; large quantities of data.

A robotic analytical application – likely to produce huge quantities of files of varying type in complex relationships to each other

and elsewhere

Of course people – scientists and technicians – will be creating the rather more familiar office documents.

A manufacturing plant that will produce very many standardised records.

Drivers for industry

Enforced

- Regulation
- Legal obligation

Voluntary

- Contractual
- IP protection
- Legal protection
- Operational efficiency
- Preservation for future re-use

Summary of drivers

- Mainly risk mitigation
  - Regulatory most important
- Cost minimisation
- Rarely: sentiment or the historical record

An example of a regulatory driver

USA FDA’s 21CFR Part 11

- Electronic Signature; Electronic Records
  - Good electronic records management (in FDA’s view)
  - Electronic signatures accepted as equivalent to ink on paper
- Protecting American consumer from fraud and sub-standard drug products
- Introduced August 1997
  - No retrenchment by FDA

21CFR Part 11: Electronic Records

Definition:

> “Any combination of text, graphics, data, audio, pictorial information or any other information representation in digital form, that is created, modified, maintained, archived, retrieved or distributed by a computer system.”

21CFR Part 11: Archiving Requirements

- Requires electronic records to be MAINTAINED ELECTRONICALLY throughout their whole lifecycle, including long-term retention
- No option to print copies and retain these
  - Even if possible
  - But rule does expect human readable versions

Reinforcement from other rules

- World-wide Good Laboratory Practice (GLP) regulations require data to be archived
  - Require responsibility of an identified individual
  - Only authorised personnel can enter the archive
  - Materials must be indexed to expedite retrieval
  - Logging of materials removed and returned
  - Provide appropriate storage conditions to minimise deterioration
- Good Manufacturing and Good Clinical Practice too require data to be retained and readily retrievable

21CFR Part 11: Impact

For industry:

- Significant business sector mandated to create e-archives
- Expensive & changes processes
- Progress slow, industry push-back

Regulatory environment:

- May be used as a model by other US government agencies
- May influence regulations outside USA

21CFR Part 11: Implications

- The industries covered by this rule must establish electronic archives for all relevant data created electronically
  - – FDA is not forthcoming on how to achieve this – industry must solve it
- Wide coverage: includes medical device manufacturers and cosmetics companies as well as pharmaceuticals.

Part 2: Scientific Data Archiving

The drivers exist – but there are formidable problems, such as:

- Heterogeneity of data types
- Complex data structures
- High data volumes
- Systems and instrument configuration dependencies
- Geographical dispersion

Heterogeneity of data sources

Situation:

- Thousands of different systems
  - Operating under a variety of operating systems
  - And in multiple versions
- Applications often very specialised
  - Proprietary data formats
  - Companies small and often short-lived

Response:

- Assumptions nothing about data sources
- Collect as much “systems metadata” as possible

Complex data structures

Situation:

- Heterogeneity rules here too
  - Rather like multi-media documents
- Complex, proprietary file formats
- Records may comprise just one or many thousands of files

Response:

- Where possible use neutral standards
  - Capture renditions in alternative formats
- Collect as much “applications metadata” as possible

High data volumes

Situation:

- Heterogeneity of data volumes per record, from a few kilobytes to many gigabytes per record
- High aggregated volumes – terabytes per annum in a large organisation

Response:

- Design for scalability
- Rule of thumb- data input doubles every year

Systems and instrument configuration dependencies

Situation:

- Content interpretation dependent on system or device settings
  - Calibrations
  - Dip switches
  - Environmental settings

Response:

- Demanding requirement: no satisfactory solutions yet
- Collect as much “context metadata” as possible

Geographical dispersion

Situation:

- Departments geographically spread
  - Span time-zones (which one gives the right time?)
  - Span legal and regulatory jurisdictions (where is responsible authority?)
- Data dispersion too – records consist of dispersed parts

Response:

- Be aware of this when designing or specifying systems
- Consult compliance and legal officers

Part 3: Science-based industry issues

There are also organisational and structural problems:

- Lack of suitable systems and services
- Management issues
- Structural issues
- Costs

Lack of suitable systems and services

- No commercially available systems available yet for this environment?
- There are software solutions which claim archiving, but:
  - Only address a small subset of the environment
  - Solve only part of the problem – not long-term preservation
- They may be good “archiving middleware” in the domains they address

Management issues

- No clear role with responsibility to act
- Few professional records managers
  - Generally have little influence at senior levels
- Lack of expertise and understanding
- Not a central management concern, even if problem clear to the wider organisation
- The geographical problem again

Structural issues

- Where is long-term responsibility to be bequeathed?
  - Often not clear
  - Who owns the data?
  - Who understands the data?
- Long-term infrastructure required, but:
  - Mergers and acquisitions
  - Company failures

Costs

- Seen as a cost with no return
- Costs can be substantial
  - Acquisition of a system and infrastructure
  - Resources – human and capital – to operate it
  - And to preserve and keep accessible data for use or inspection

Part 4: Status report

- Overall little progress so far
- A very few major companies in the lead (e.g. GlaxoSmithKline)
- Still significant issues to be solved:
  - Preservation strategies to address the data fragility/obsolescence problem
  - Long-term cost containment
- No commercially available products
- Lack of awareness, skills and experience improving but still to be overcome

Questions?

Philip Lord, MSc MCLIP MIMA
Digital Archiving Consultancy
2 Wayside Court
TWICKENHAM
TW1 2BQ
UK

☎ +44 (0)20 8607 9102
📠 +44 (0)70 5067 5010
philip@philiplord.com
www.philiplord.com