



GOOD PRACTICE GUIDE:

Data Integrity by Design

GOOD PRACTICE GUIDE:

Data Integrity by Design

Disclaimer:

The *ISPE GAMP® RDI Good Practice Guide: Data Integrity by Design* provides practical guidance to support pharmaceutical organizations achieve good data governance together with the efficient and effective implementation and operation of compliant GxP computerized systems. This Guide is solely created and owned by ISPE. It is not a regulation, standard or regulatory guideline document. ISPE cannot ensure and does not warrant that a system managed in accordance with this Guide will be acceptable to regulatory authorities. Further, this Guide does not replace the need for hiring professional engineers or technicians.

Limitation of Liability

In no event shall ISPE or any of its affiliates, or the officers, directors, employees, members, or agents of each of them, or the authors, be liable for any damages of any kind, including without limitation any special, incidental, indirect, or consequential damages, whether or not advised of the possibility of such damages, and on any theory of liability whatsoever, arising out of or in connection with the use of this information.

© Copyright ISPE 2020. All rights reserved.

All rights reserved. No part of this document may be reproduced or copied in any form or by any means – graphic, electronic, or mechanical, including photocopying, taping, or information storage and retrieval systems – without written permission of ISPE.

All trademarks used are acknowledged.

ISBN 978-1-946964-34-2

Preface

Data Integrity by Design is the concept that data integrity must be incorporated from the initial planning of a business process through to the implementation, operation, and retirement of computerized systems supporting that business process. It promotes the application of critical thinking to identify how data flows through the business process, and to proactively assess and mitigate risks across both the system and data lifecycles. It emphasizes data integrity as foundational to protecting patient safety and product quality.

This *ISPE GAMP® RDI Good Practice Guide: Data Integrity by Design* supports organizations as they embrace and implement a holistic approach by leveraging data governance and knowledge management activities to drive continual improvement in data integrity. The Guide fosters a patient-centric mindset, focusing resources and management attention on quality best practices that inherently facilitate meeting regulatory compliance requirements.

This Guide provides a bridge between the system lifecycle approach defined in *ISPE GAMP® 5: A Risk-Based Approach to Compliant GxP Computerized Systems*, and the data lifecycle approach in the *ISPE GAMP® Guide: Records and Data Integrity*. Data integrity can only be achieved when both lifecycle approaches are adopted, understood, and actively managed.

Part of the *ISPE GAMP® Guide: Records and Data Integrity* series, this Guide replaces and significantly expands upon the *ISPE GAMP® Good Practice Guide: Electronic Data Archiving* from 2007.

Acknowledgements

The Guide was produced by a Task Team led by James Henderson (Eli Lilly and Company, USA), Lorrie Vuolo-Schuessler (Syneos Health, USA) and Charlie Wakeham (Waters Corporation, Australia). The work was supported by the ISPE GAMP Community of Practice (CoP) and sponsored by Michael Rutherford (Syneos Health, USA).

Core Team

The following individuals took lead roles in the preparation of this Guide:

Petch Ashida-Druar	GlaxoSmithKline	USA
Matt Brawner	Data Integrity and Compliance Consultant	USA
Roger Buchanan	Buchanan Consultancy Ltd.	United Kingdom
Rob Dillman	Eli Lilly & Company	USA
Stuart Jones	Merck & Co. Inc.	Ireland
Paige Kane	Merck & Co. Inc.	USA
Heather Longden	Waters Corporation	USA
Charles H. Maher III	CAI	USA
Bob McDowall	R D McDowall Ltd.	United Kingdom
Matt McMenamin	GlaxoSmithKline	USA
Anubha Mukherjee	CSV and IT Compliance Consultant	India
Mark Newton	Heartland QA	USA
Levi Schenk	CSL Behring	USA
Rob Stephenson	Rob Stephenson Consultancy	United Kingdom

Contributors and Team Reviewers

The Leads wish to thank the following individuals for their valuable contribution during the preparation of this Guide.

Elissa Hilton	Eli Lilly & Co.	USA
Brandi Stockton		USA

CSA Industry Pilot Team

The Team Leads wish to thank the following members of the CSA Industry Pilot Team for their exceptional efforts in producing the CSA Appendix.

Senthil Gurumoothi	Gilead Sciences, Inc.	USA
Shana Kinney	REGENXBIO	USA
Kevin Martin	Azzur Group LLC	USA
Khaled Moussally	Compliance Group Inc	USA
Raechelle Raimondo	Abbvie (formerly Allergan)	USA
Ken Shitamoto	Gilead Sciences, Inc.	USA
Francisco Vincenty	U.S. Food and Drug Administration (FDA)	USA

Special Interest Appendix on AI

The Team Leads wish to thank the following members for their hard work on the Special Interest Appendix on Artificial Intelligence: Machine Learning.

Ravi Krishna	Assembly Consulting	USA
Urmi Parkar	PRA Health Sciences	USA
Levi Schenk	CSL Behring	USA
Eric Staib (Lead)	Genpact	USA
Brandi Stockton		USA



Subject Matter Expert Input and Review

The Team Leads wish to thank the following individuals for their expert contribution to the document.

Peter Baker	Green Mountain Quality Assurance	USA
Monica Cahilly	Green Mountain Quality Assurance	USA
Judy Samardelis	Thermo Fisher Scientific	USA

Regulatory Input and Review

Particular thanks go to the following for their review and comments on this Guide:

Kevin Bailey	MHRA	United Kingdom
Gaye Camm	Therapeutic Goods Administration (TGA)	Australia
Robert Tollefsen	U.S. Food & Drug Administration (FDA)	USA
Francisco Vincenty	U.S. Food and Drug Administration (FDA)	USA

Special Thanks

The Leads would like to give particular thanks to Chris Clark (TenTenTen Consulting Limited, United Kingdom), and ISPE Technical Advisor, Sion Wyn (Conformity Ltd., United Kingdom) for their efforts during the creation process of this Guide. The Team would also like to thank ISPE for technical writing and editing support by Jeanne Perez (ISPE Guidance Documents Technical Writer/Editor) and production support by Lynda Goldbach (ISPE Guidance Documents Manager).

The Team Leads would like to express their grateful thanks to the many individuals and companies from around the world who reviewed and provided comments during the preparation of this Guide; although they are too numerous to list here, their input is greatly appreciated.

Company affiliations are as of the final draft of the Guide.



600 N. Westshore Blvd., Suite 900, Tampa, Florida 33609 USA
Tel: +1-813-960-2105, Fax: +1-813-264-2816

www.ISPE.org

AVA Documents
Document Sharing Hub

Table of Contents

1	Introduction	9
1.1	Background	9
1.2	Case for Quality Program	9
1.3	Purpose.....	10
1.4	Scope.....	10
1.5	Structure of the Guide.....	11
1.6	Key Terms	11
1.7	Key Roles and Responsibilities.....	14
2	Foundations to Data Integrity by Design	15
2.1	Governance to Achieve Data Integrity by Design	15
2.2	Data Ownership	19
3	Retention Strategy	23
3.1	Retention Periods	24
3.2	Readability	25
3.3	Availability	28
3.4	Access	30
3.5	Protecting Records and Data.....	31
3.6	Managing System Retirement.....	36
3.7	Records Management and Retention through Mergers, Acquisitions, and Divestments	41
4	Implementing Data Integrity by Design	43
4.1	A Process to Achieve Data Integrity by Design.....	43
4.2	Business Process	47
4.3	Data Flow Diagrams	48
4.4	Data Classification and the Intended Use of the Data	49
4.5	Business Process Risk Assessment	50
4.6	Data Lifecycle	51
4.7	Data Nomenclature	55
5	System Planning	59
5.1	Planning Computerized Systems to Efficiently Support the Optimized Business Process	59
5.2	Addressing Individual Systems	61
5.3	System Risk Assessment.....	63
6	Active Records.....	67
6.1	Creation	67
6.2	Processing	69
6.3	Review, Reporting, and Use	75
7	Semi-active and Inactive Records	81
7.1	Semi-active Records.....	81
7.2	Retention of Inactive Records.....	81
7.3	Return to Active State (Retrieval).....	86
7.4	Destruction.....	87

Management Appendices

8	Appendix M1 – Knowledge Management.....	89
8.1	Introduction	89
8.2	Key Concepts.....	89
8.3	Managing Knowledge	92
8.4	Mindsets and Behaviors.....	94
8.5	Conclusion	95
9	Appendix M2 – Understanding Data Integrity Compared to Data Quality	97
10	Appendix M3 – Third-Party Data	99
10.1	Introduction	99
10.2	Assessments and Responsibilities.....	99
10.3	Data Governance for CxO	99
10.4	Data Storage Off-Premise.....	100
10.5	Conclusion	101

Development Appendices

11	Appendix D1 – Example Business Process Mapping: Laboratory System.....	103
12	Appendix D2 – Instrument Devices with Electronic Record Storage.....	105
12.1	Introduction	105
12.2	Background.....	105
12.3	Instrument Use.....	106
12.4	Accounting for Original Data	106
12.5	Challenges	107
12.6	Remediation Strategies.....	107
12.7	Risk Register.....	109
12.8	Example Data Integrity Risks, Interim Controls, and Actions to Consider	109
12.9	Conclusion	110

Operation Appendices

13	Appendix O1 – Data Analytics and Technical Solutions Supporting Data Integrity.....	111
13.1	Introduction	111
13.2	Detecting Data Integrity Issues with Business Rules.....	112
13.3	Technical Solution Using Computer Lockdown via Software Shells	116
14	Appendix O2 – Good Practices for General Archiving	119
14.1	What is Archiving?	119
14.2	What to Archive.....	119
14.3	How to Archive	119
14.4	Managing an Archive	120

15 Appendix O3 – GLP Archiving Considerations	123
16 Appendix O4 – Example Retention Periods and Requirements	125
17 Appendix O5 – Maintaining Legacy Software	139
17.1 Introduction	139
17.2 Non-Disposal of Retired Systems	139
17.3 Compatibility to Modern Operating Systems	139
17.4 Virtual Machine Solution	139
17.5 The Hardware Museum	140
17.6 Conclusion	140

Special Interest Topics Appendices

18 Appendix S1 – Artificial Intelligence: Machine Learning	141
18.1 Introduction	141
18.2 Background	141
18.3 Scope	141
18.4 Data Lifecycle (Iterative, Autonomous, and Adaptive)	142
18.5 Concept Phase (Understanding the Business Case)	142
18.6 Project Phase (Data Modeling and Evaluation)	144
18.7 Operation Phase (Deployment and Monitoring)	145
18.8 Further Reading	146
19 Appendix S2 – Computer Software Assurance	147
19.1 Introduction	147
19.2 Establishing a Lifecycle-based Approach and Basic Assurance	152
19.3 Risk-based Assurance	153
19.4 Example: Applying Risk-based Approach from <i>ISPE GAMP® 5</i> and CSA	157
19.5 Example: Applying <i>ISPE GAMP® 5</i> and CSA Using Direct Leveraging of Testing Throughout the System Lifecycle	160
19.6 Conclusion	162

General Appendices

20 Appendix G1 – References	163
21 Appendix G2 – Glossary	169
21.1 Acronyms and Abbreviations	169
21.2 Definitions	171

1 Introduction

1.1 Background

To date, many regulated life science organizations have been addressing data integrity on a system by system basis. The initial focus has often been on achieving good data integrity outcomes by the remediation of technical gaps in existing systems, and by developing good practice cultures and behaviors to ensure that compliance is being achieved “in real time.” This focus may have resulted in missed opportunities to work with suppliers on the development of enhanced functionalities supporting data integrity, for both existing and new systems.

The requirement to achieve data integrity is not new; data integrity expectations and holistic approaches have been implied within regulations globally for some time, for example:

EU GMP 1.4 [1]:

“(i) Product realisation is achieved by designing, planning, implementing, maintaining and continuously improving a system that allows the consistent delivery of products with appropriate quality attributes”

“(xi) Continual improvement is facilitated through the implementation of quality improvements appropriate to the current level of process and product knowledge.”

US GMP 21 CFR 211.160 (b) [2]:

“Laboratory controls shall include the establishment of scientifically sound and appropriate specifications, standards, sampling plans, and test procedures designed to assure that components, drug product containers, closures, in-process materials, labeling, and drug products conform to appropriate standards of identity, strength, quality, and purity.”

US GMP 21 CFR 211.194 (a) [3]:

“Laboratory records shall include complete data derived from all tests necessary to assure compliance with established specifications and standards, including examinations and assays”

Achieving data integrity requires a holistic approach across the entire organization, leveraging critical thinking, and implementing data governance across all regulated business processes.

1.2 Case for Quality Program

The US FDA Center for Devices and Radiological Health (CDRH) Case for Quality program [4] promotes a risk-based, product quality-focused, and patient-centric approach. This initiative has been endorsed by the ISPE GAMP® CoP (Community of Practice) Leadership [5] who believe that:

“such an approach is appropriate throughout the regulated life science industries, including pharmaceutical, biological, and medical devices, and throughout the complete product life cycle, regardless of the specific applicable predicate regulation. GAMP® strongly supports the adoption of new and innovative computerized technologies and approaches throughout the product life cycle to support product quality, patient safety, and public health.”

Computer Software Assurance was born from CDRH's Case for Quality [4] which enhances and incentivizes the adoption of practices and behaviors to improve medical safety, responsiveness, and how patients experience products. As part of that effort, Computerized System Validation (CSV) was identified as a barrier to new technologies due to the lack of clarity on risk-based effort, compliance-focused approaches, and perceived regulatory burden. In actuality, the FDA supports and encourages the use of automation, information technology, and data solutions throughout the product lifecycle in the design, manufacturing, service, and support of life sciences [6]. An industry team was formed and work begun on the development of an FDA draft guidance on Computer Software Assurance (CSA).

Appendix S2 contains a detailed discussion of CSA, and a case study based around a data integrity technical control demonstrating the application of CSA in a real-life situation.

1.3 Purpose

This *ISPE GAMP® RDI Good Practice Guide: Data Integrity by Design* builds on the guidance contained in the various *ISPE GAMP® Good Practice Guide Series* [7], including the *ISPE GAMP® Guide: Records and Data Integrity* [8], and *ISPE GAMP® 5: A Risk-Based Approach to Compliant GxP Computerized Systems* [9].

The focus of this new Good Practice Guide is on unifying system and data lifecycles, from data creation to destruction. The purpose of this new Guide is to provide a practical “bridge” between achieving good data governance for regulated business processes and the efficient and effective implementation and operation of compliant GxP computerized systems.

This Guide emphasizes the interrelationships between these operational imperatives and presents a common approach, providing a harmonized framework that will help organizations achieve/enhance product quality and patient safety.

1.4 Scope

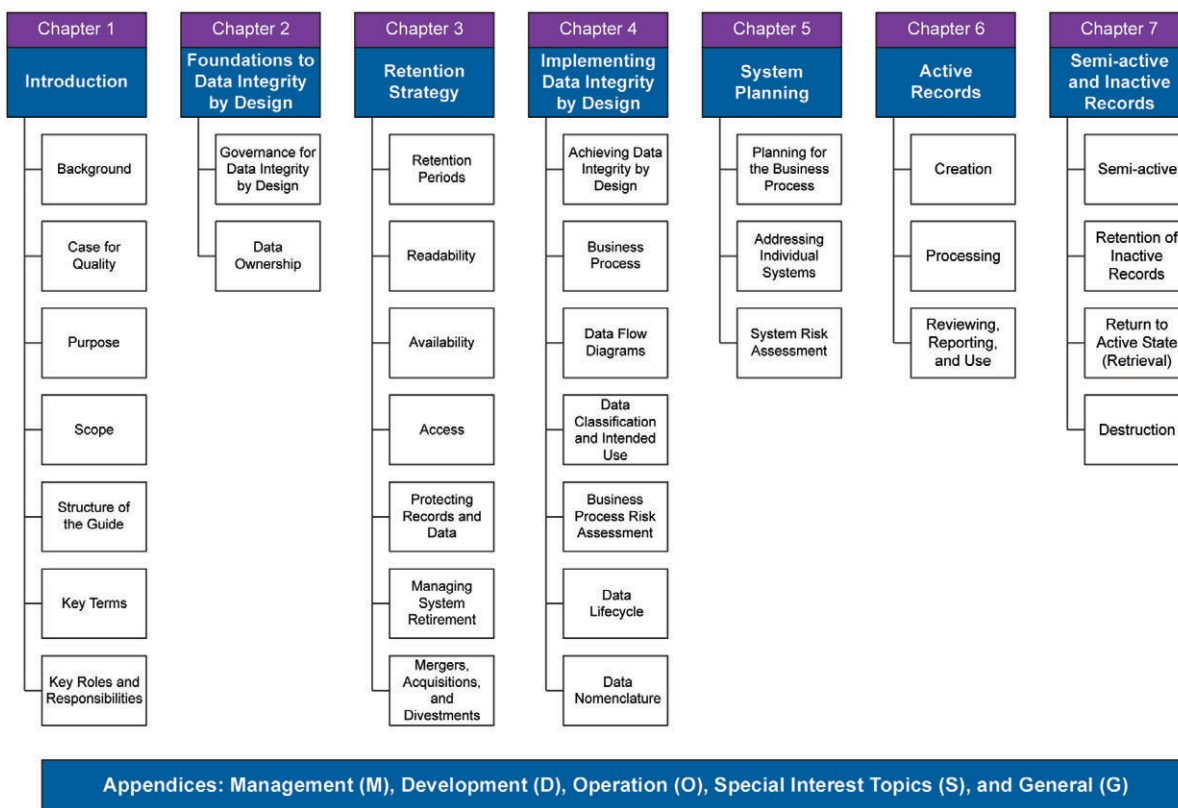
This Guide is intended to encourage organizations to adopt a risk-based approach to ensuring data integrity in support of patient safety and product quality.

The scope of this Guide is as broad as the scope of *ISPE GAMP® 5* [9] and includes all computerized systems used in support of GxP regulated business processes within the life science industries. While computerized systems feature strongly within this Guide, as these systems provide technical controls for data integrity, data governance also requires procedural and behavioral controls supported by a quality culture that leverages critical thinking and a continual improvement mindset.

1.5 Structure of the Guide

This Guide is divided into seven chapters and related sub-sections, plus a set of appendices, as shown in Figure 1.1.

Figure 1.1: Guide Structure



1.6 Key Terms

1.6.1 ALCOA+

ALCOA+ is the acronym for the key concepts that can help to support record and data integrity [8]. See Table 1.1.

Table 1.1: ALCOA+ Principles [8]

Principle	Data Expectation
Attributable	- Attributable to the person or system generating the data - Identify the person or system performing an activity that creates or modifies data - Linked to the source of the data
Legible	- Readable and permanent - Accessible throughout the data life cycle - Original data and any subsequent modifications are not obscured
Contemporaneous	Recorded or observed at the time the activity is performed
Original	Original data is the first recording of data, or a “true copy” which preserves content or meaning

Table 1.1: ALCOA+ Principles [8] (continued)

Principle	Data Expectation
Accurate	- Free from error - No editing performed without documented amendments - Conforming to truth or standard
Complete	All data, and relevant metadata, including any repeat or re-analysis performed
Consistent	- Application of good documentation practices throughout any process - The application of date and time stamps in the expected sequence
Enduring	Recorded in a permanent, maintainable form for the retention period
Available	Available and accessible for review, audit, or inspection throughout the retention period

1.6.2 Audit Trail

In this document the term audit trail refers to a data audit trail of operator entries and actions that create, modify, or delete regulated records, as required by 21 CFR Part 11 [10] and EU Annex 11 [11], as distinguished from other system and technical logs.

1.6.3 GxP

The term “GxP” is used within this Guide to represent the encompassing regulations (Good Practices) to which different aspects of regulated companies must adhere. It is not intended to imply that all regulatory requirements are the same across Good Manufacturing Practice (GMP), Good Clinical Practice (GCP), Good Laboratory Practice (GLP), Good Distribution Practice (GDP), and Good Pharmacovigilance Practice (GVP, also known as GPvP), etc.

1.6.4 Data Integrity and Quality

Data integrity and data quality are sometimes used interchangeably, but there is actually a difference between these two terms.

It is the awareness of the similarities and differentiation between data integrity and data quality that is less prevalent and therefore a number of examples are provided in Appendix M2 Table 9.1 to aid in this clarification.

1.6.4.1 Understanding Data Integrity

The need for data integrity is well established through predicate rules, through data integrity guidance, and through industry acceptance of data integrity as an essential component to protecting patient safety and ensuring product quality.

Data integrity is the assurance that the data is original and trustworthy, and that this assurance has been maintained throughout the data lifecycle. The requirements for data integrity are discussed comprehensively in the *ISPE GAMP® Guide: Records and Data Integrity* [8].

1.6.4.2 Understanding Data Quality

The *ISPE GAMP® Guide: Records and Data Integrity* [8] states that:

“Data quality relates to the data’s fitness to serve its intended purpose in a given context within a specified business or regulatory process. Data quality management activities address aspects including accuracy, completeness, relevance, consistency, reliability, and accessibility.”

MHRA [12] defines data quality as:

“The assurance that data produced is exactly what was intended to be produced and fit for its intended purpose. This incorporates ALCOA.”

The Organisation for Economic Co-operation and Development (OECD) [13] defines data quality as:

“Data quality is the assurance that the data produced are generated according to applicable standards and fit for intended purpose in regard to the meaning of the data and the context that supports it. Data quality affects the value and overall acceptability of the data in regard to decision-making or onward use.”

Data quality requires that the data is organized and able to be accessed, sorted, and searched to enable the business to effectively use the data, and is reflected in the list below:

- **Data Accuracy:** The extent to which the data is free of identifiable errors
- **Data Accessibility:** The level of ease and efficiency at which data is legally obtainable, within a well-protected and controlled environment
- Data Comprehensiveness: The extent to which all required data within the entire scope are collected, documenting intended exclusions
- Data Consistency: The extent to which the data is reliable, identical, and reproducible by different users across applications
- **Data Currency:** The extent to which data is up-to-date; a datum value is up-to-date if it is current for a specific point in time, and it is outdated if it was current at a preceding time but incorrect at a later time
- Data Nomenclature: A consistent approach to metadata entry that facilitates identification of data relating to the same product or process for use in trending and/or data analytics, for example, a data lake. Discussed more fully in Section 4.7
- Data Granularity: The level of detail at which the attributes and characteristics of data quality are defined
- **Data Precision:** The degree to which measures support their purpose, and/or the closeness of two or more measures to each other
- Data Relevancy: The extent to which data is useful for the purposes for which it was collected
- **Data Timeliness:** The availability of up-to-date data within the useful, operative, or indicated time

Note: Items in **bold underline** are reasonably addressed in ALCOA+ as part of data integrity attributes.

Data quality is defined by the Data Management Body of Knowledge (DMBOK) [14] as:

“...the planning, implementation, and control of activities that apply quality management techniques to data, in order to assure it is fit for consumption and meet the needs of data consumers.”

See Figure 8.2 in Appendix M1 for the concept of data producers and consumers within knowledge and data management.

1.6.5 Types of Data

In this Guide, data is classified as regulated, operational, or unnecessary, as originally defined in *ISPE GAMP® RDI Good Practice Guide: Data Integrity – Manufacturing Records* [15]:

- **“Regulated:** *used for a regulated decision or to support a regulated process, that is, data as required by, or in support of, the predicate rules – what we have to keep*
- **Operational:** *non-regulated data used for business process decisions such as performance analysis and management of maintenance schedules – what we want to keep*
- **Unnecessary:** *data not needed due to either the circumstances of its creation (e.g., during a non-regulated activity or process) or because that data does not provide additional context, metadata, or meaning for the activity or process – what we do not need”*

Note that some data may need to be retained for financial, health and safety, or other non-life science regulations.

1.7 Key Roles and Responsibilities

The key roles recommended for data integrity by design are those identified in the following ISPE GAMP® publications:

- *ISPE GAMP® 5: A Risk-Based Approach to Compliant GxP Computerized Systems (GAMP 5) [9]*
- *ISPE GAMP® Guide: Records and Data Integrity (RDI) [8]*
- *ISPE GAMP® RDI Good Practice Guide: Data Integrity – Key Concepts (DI-KC) [16]*

The definitions of these roles and their assigned responsibilities are not repeated here, however, the roles and where they are described are listed below:

- Senior Management (DI-KC)
- Chief Data Integrity/Governance Officer (DI-KC)
- Process Owner (GAMP 5, DI-KC)
- System Owner (GAMP 5, DI-KC)
- System Support and Administration Personnel (DI-KC)
- Database Administration (DI-KC)
- Data Steward (RDI, DI-KC)
- Data Owner (RDI, DI-KC)
- Quality Unit (GAMP 5)/Quality Function (DI-KC)
- Subject Matter Expert (GAMP 5, DI-KC)
- End User (GAMP 5, DI-KC)
- System Developer (DI-KC)
- Supplier (GAMP 5)

See Figure 2.3 for a schematic representation of the relationship between various owner roles, data lifecycle, and record phase.

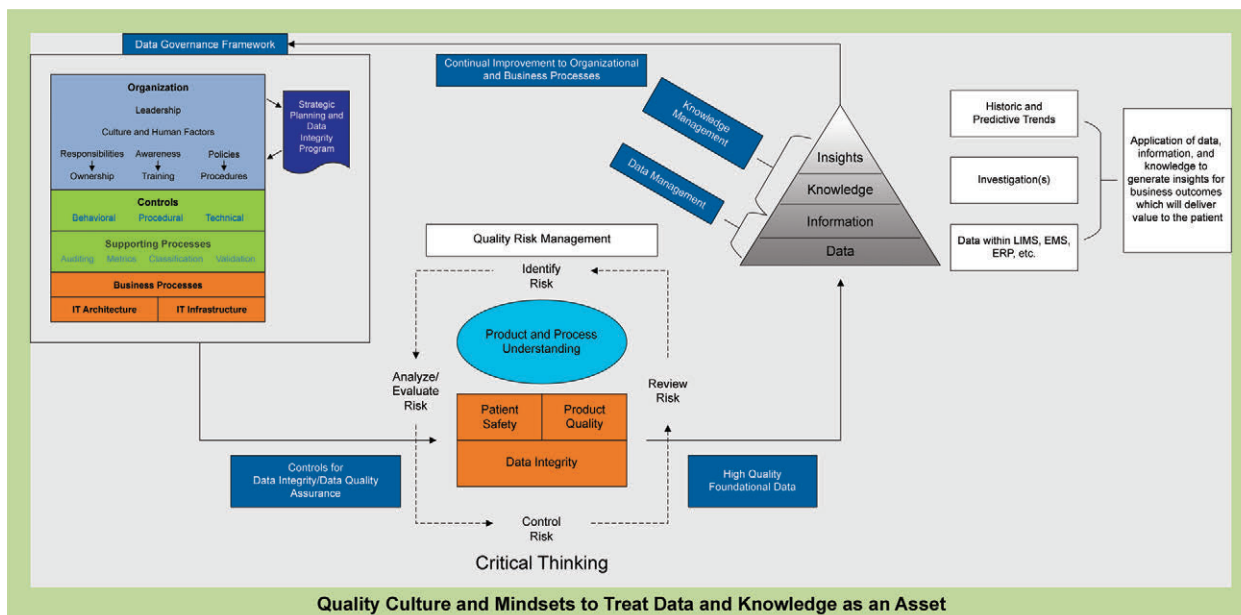
2 Foundations to Data Integrity by Design

Data integrity by design is a critical-to-quality initiative for GxP regulated organizations that ensures public health, patient safety, and product quality. This is supported by the FDA Case for Quality initiative [4] as introduced in Section 1.2, and by current (and emerging) regulations and guidelines for compliant records and data. Collectively these, along with warning letters and other interventions, are acting as drivers to achieve the adoption of fully electronic recordkeeping (even for manual tasks). There are also business efficiency gains that can be achieved with moving away from paper records.

Data integrity by design can be achieved through a governance framework combined with Quality Risk Management (QRM) and Knowledge Management (KM).

The data governance framework provides the controls for data integrity and quality assurance. QRM is collectively applied to product and process understanding to achieve patient safety, product quality, and data integrity resulting in high quality foundational data. This is an input to KM to drive continual improvement to organizational and business processes. Throughout all activities, critical thinking must be leveraged within a quality culture to strive for operational excellence. This all-encompassing approach is shown in Figure 2.1. KM is explained in detail in Appendix M1.

Figure 2.1: Foundations to Data Integrity by Design



2.1 Governance to Achieve Data Integrity by Design

Governance to achieve data integrity by design requires the following elements to be established:

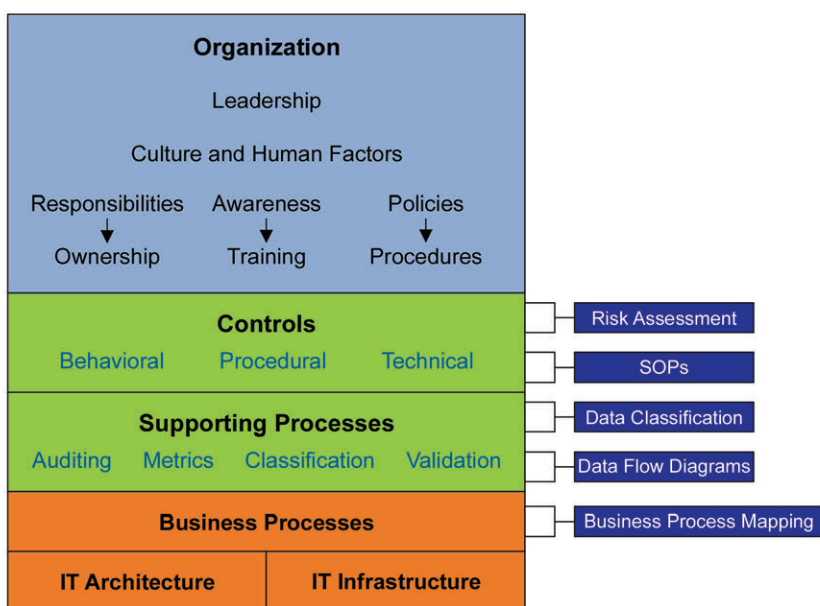
- A long-term vision
- A unifying strategy for data and system governance
- A roadmap to deliver the strategy for a specific process or project
- Provision of resources (budgetary and personnel)

- Supporting policies and procedures

Figure 2.2 expands on the data governance framework with tools to aid in the understanding of the process and to optimize system planning to achieve data integrity by design. These tools are discussed in detail in Chapter 4.

Figure 2.2: Leveraging Governance Elements to Ensure Data Integrity by Design

Adapted from ISPE GAMP® Guide: Records and Data Integrity [8]



Some components of this governance approach may already be in place within the regulated organization; however, these should be reviewed and expanded as necessary to achieve a fully comprehensive governance framework. Effective data governance is the ultimate aim of any data integrity assurance program.

2.1.1 Establishing a Long-term Vision

In order to achieve an effective data integrity by design strategy, a high-level vision for record and data management, “what this might look like,” needs to be established at an organization level.

The high-level vision needs to leverage critical thinking across all data governance activities, including developing the data flow through the business process and identifying critical-to-quality data and its owner(s). It also needs to apply a risk-based approach to identifying and mitigating the risks and vulnerabilities of manual intervention, which otherwise could result in inadvertent or unauthorized data manipulation.

Without an organization-wide vision for data integrity by design, there is the risk that important decisions relating to the introduction or modification of processes and systems may be taken without proper consideration of the data integrity implications. This may result in the organization adopting suboptimal solutions that prevent or inhibit the ability to achieve and demonstrate good data integrity practices.

For Example, a strong company vision with respect to data integrity could be that:

As part of our commitment to support public health, patient safety, and product quality, our five-year vision is that all regulated records will be created and maintained in electronic format according to good practice ALCOA+ principles throughout their lifecycle (creation to destruction).

This vision should form a strategic approach to defining an appropriate data architecture and corresponding data lifecycle that are considered when acquiring new systems and remediating existing record and data lifecycle processes. This strategic approach sets the expectation that eventually eliminates all paper and hybrid records within the organization that are difficult to maintain in a compliant state throughout their retention period.

The vision can be translated into annual goals and continual improvement objectives that funnel down to individual contributors (who are the data generators) to provide assurance that integrity and quality are built into the process.

2.1.2 A Unifying Strategy for Data and System Governance

Governance can take place at many levels within an organization, for example, global, regional, group, national, site wide, departmental, process, project.

From a governance point of view, “data integrity by design” can be restated as “data integrity by intent.” There needs to be a clear mandate at every level of governance that achieving data integrity through intentional design is a business-critical requirement, one that will be supported, endorsed, and resourced. This should be a commitment not just for the quality unit, but for all functions involved in the design, implementation, validation, and operation of the business process.

Governance to achieve data integrity and governance for the validation of computerized systems within a regulated organization have been discussed in the *ISPE GAMP® Guide: Records and Data Integrity* [8] and *ISPE GAMP® 5* [9] respectively.

Validation of computerized systems for intended use is necessary to achieve data integrity, while data integrity requirements must be considered as part of each validation effort in order to ensure that ALCOA+ imperatives are delivered.

When determining and establishing governance approaches to achieve data integrity by design, there is an integral relationship between data integrity and computerized system validation that must be recognized in the corporate Quality Management System (QMS). The data needs will in turn create the risk scenarios that ultimately drive the configuration and validation activities for the computerized system. Data, and the regulated use of the data, are the drivers for many project activities within the system lifecycle. The system lifecycle ensures that ongoing controls are established and verified to make certain the data is appropriately and effectively managed throughout the data lifecycle.

Table 2.1 shows a list of practices that are shared by good governance frameworks for both data integrity and computerized system validation.

Table 2.1: Common Governance Elements

Elements Common to Data Integrity and Computerized System Validation
• Senior management sponsorship and engagement • Defined roles and responsibilities • Provision of adequate resources • Training support • Focus on patient safety, product quality, and data integrity • Risk-based approach • Leveraging supplier involvement • Part of the QMS • Continual improvement • Scalability • Change control and configuration management • Access control

Table 2.1: Common Governance Elements (continued)

Elements Common to Data Integrity and Computerized System Validation
• Information security and cybersecurity • Periodic review and evaluation • Disaster recovery/business continuity plans • Backup and restore • Archive and retrieve • Active management of data retention and destruction • IT support

A company-wide data integrity by design strategy should be established as early as possible in the development of any significant data integrity initiative. It may also be appropriate to determine a data integrity by design strategy at a lower level, such as for a specific business process or project.

Effective data integrity by design should follow a risk-based approach that encompasses the full business process lifecycle. This lifecycle addresses data (creation, processing, reporting and use, retention and destruction) and, by implication, the lifecycle phases of all associated computerized systems (concept, project, operation, retirement).

For a regulated organization, the scope of governance activities to achieve data integrity by design should cover GxP processes, data lifecycles, and computerized systems. The same governance elements are required whether the component(s) of the computerized system are on-premise or leverage a service-based architecture (e.g., Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS)).

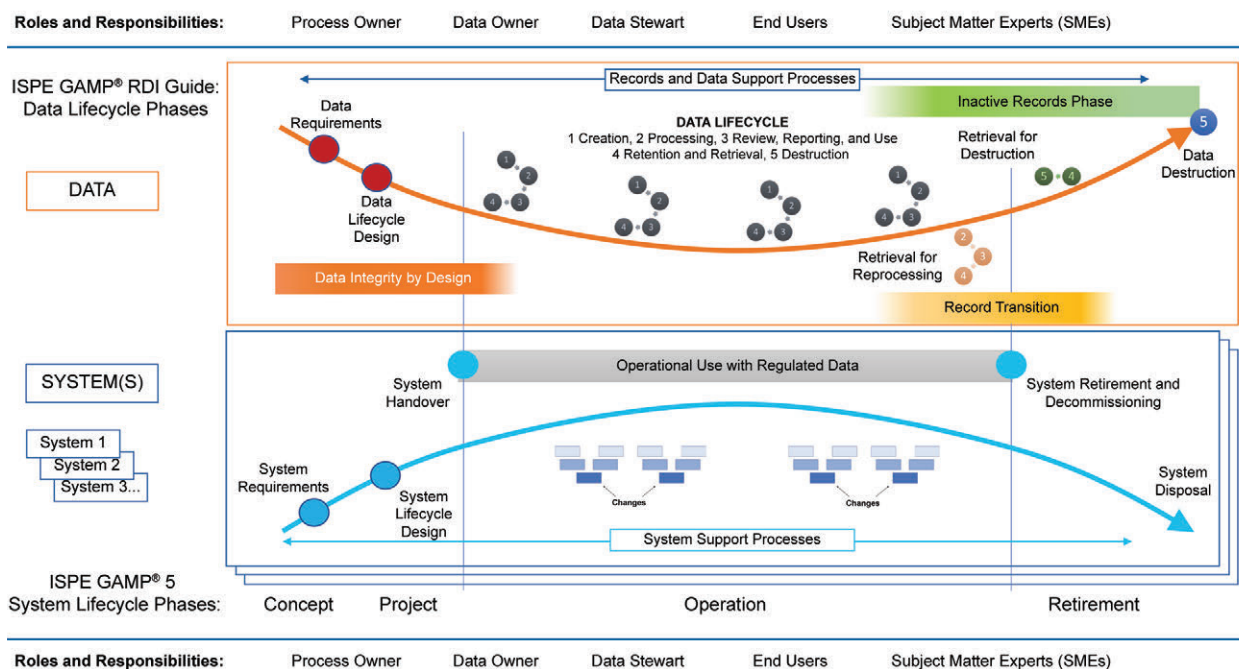
This governance scope is shown schematically in Figure 2.3. This diagram is an abstract representation of the need for, and relationship between, the data lifecycle and system lifecycle for the computerized systems involved in the creation, processing, review, reporting and use, retention and destruction of the data within a business process. The data lifecycle should be considered both within the context of the system lifecycle and separate to it. The lifecycles run on independent timelines in that the data may outlive the system and need alternative arrangements for retention after system retirement, or conversely, the system may continue in operation long after the retention period has expired for a specific data set.

Within the operational phase of the system lifecycle, change control and configuration management are essential to maintain the validated state of the system.

The roles and responsibilities depicted at the top and bottom of the diagram in Figure 2.3 need to be defined and assigned to the appropriate individuals, and may only be meaningful through specific phases of the two lifecycles. Of these roles, data ownership is a critical concept to ensure data integrity throughout the data lifecycle. Regardless of where the data resides (on- or off-premise) or where it is generated (Contract Manufacturing Organization (CMO), Contract Research Organization (CRO), etc.), data ownership remains with the regulated company.

Figure 2.3 shows that processes to manage records and data throughout the data lifecycle must be considered in the initial data integrity by design phase. Data lifecycles can transcend multiple systems, with data flow between the systems. Overall, a data flow diagram of the complete business process is an essential requirement for data integrity by design.

Figure 2.3: Data and System(s) Lifecycle



To consistently achieve an effective design that delivers compliance throughout the data lifecycle, the retention requirements must be understood when defining the business process specification.

In many circumstances it may be ineffective to consider only the lifecycle of the data within a single system because data may be transferred across multiple systems (computerized or paper-based) as part of the data flows required to construct the regulated record. Where copies of the data exist in more than one system (e.g., source and target systems), it is important to understand:

- The scope of the data use
- Which record will be used to make any decision under the predicate rules
- How it will be assured that the data is current and complete

It is also possible that a single system supports multiple business processes, for example, a document management system. Consequently, the retention and other data integrity aspects from several processes may need to be considered in the system design and implementation. A holistic approach to data integrity by design is therefore required over the full data lifecycle(s) for all impacted regulated data.

2.2 Data Ownership

2.2.1 Considerations on Ownership

Consider a business process that is not computerized. The (business) process owner is accountable for all aspects of the process.

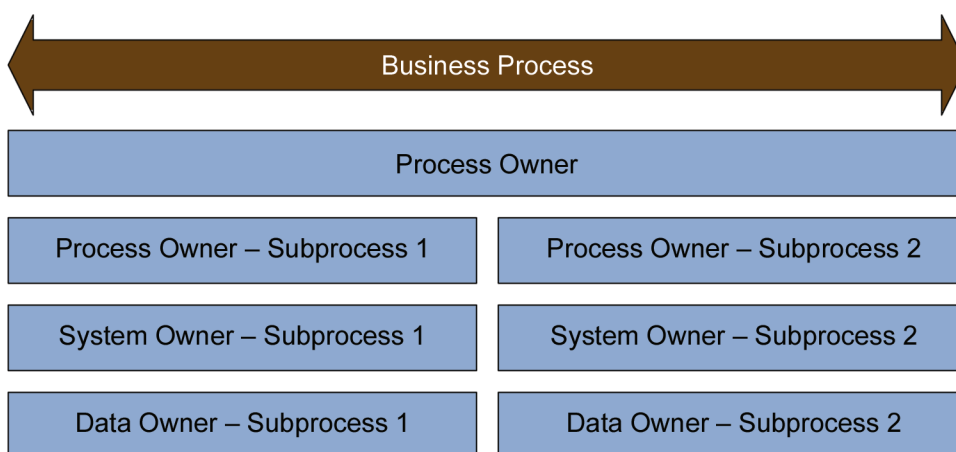
Once the business process is supported by a computerized system, accountabilities are divided into two separate aspects: process ownership and system ownership (responsible for the computerized system, as discussed below).

Those activities associated with the performance of the business process remain with the process owner; the separate role of system owner has accountability for those activities related to the technical support and maintenance of the computerized system.

The process owner has accountability for the quality and integrity of the entire business process including the process data throughout its data lifecycle, but may delegate data responsibility to data owner(s) at specified phases of the data lifecycle.

As a further level of complexity, the business process itself may be divided into subprocesses. The process owner may delegate some or all their responsibilities to sub-process owners, which may also include delegation of system and data ownership. A hierarchy of process, system, and data owners is thus established, as depicted in Figure 2.4. The subprocesses can each contain one or more computerized systems.

Figure 2.4: Simple Example Showing Business Process Divided into Two Subprocesses



Data ownership and system ownership creates a matrix of responsibilities as delegated data ownership may be aligned with, for example, specific clinical trials, sites, or products, while the data may be processed in multiple systems.

Data should never be without a designated owner at any point in the data lifecycle; effective handover of data ownership responsibilities between individuals transitioning to/from an organization role (with designated data ownership responsibility) is critical to successful maintenance of data integrity and data quality.

When considering the computerization of a complex process, the overall process owner should understand the intended scope of the computerization and ensure the involvement of all subsidiary process, system, and data owners.

2.2.2 Principles of Data Ownership

It is important for an organization to establish and maintain a culture that supports data integrity and to have clear policies for data governance. Data governance should address data ownership throughout the data lifecycle to make certain data integrity risks are addressed throughout the data lifecycle. Electronic systems have become integral to most business processes within the life science industry, and with this evolution comes the need to understand what data is captured and who owns it. The concept of data ownership is not new and applies equally to paper data and electronic data. The ease of movement of electronic data makes data ownership more complex.

Data ownership is often defined at a company level, but is then refined to the system level (documented in the roles and responsibilities section of the system Standard Operating Procedures (SOPs)) or even to a specific project (documented in the quality and project plan) or clinical trial (documented in the trial master file or other trial configuration documentation). The data owner must understand the business value of the data and often has a position of responsibility within the business unit generating the data. Individual organizations should allocate roles and responsibilities based on organizational structure and the specific system involved. The organization needs to establish policies and procedures to support the role of data owner and ensure it is appropriately resourced for each business process or system.

Data ownership can be a subset responsibility of the process owner (see *ISPE GAMP® RDI Good Practice Guide: Data Integrity – Key Concepts* [16]). As defined within *ISPE GAMP® 5* [9], the process owner is:

“the person ultimately responsible for the business process or processes being managed. This person is usually the head of the functional unit or department using that system, although the role should be based on specific knowledge of the process rather than position in the organization. The process owner is responsible for ensuring that the computerized system and its operation is in compliance and fit for intended use in accordance with applicable Standard Operating Procedures (SOPs) throughout its useful life....Ownership of the data held on a system should be defined and typically belongs to the process owner.”

2.2.3 Responsibility of the Data Owner

The data owner is responsible for the integrity of the data, and for defining the quality requirements for the data and how the data is to be used. Because electronic data is easily shared, there is a need to make sure that the data owner maintains awareness of the use of the data. In the case of clinical trial data and medical records, the data owner must protect the interests of the patient and their confidential data. Non-GxP regulations such as GDPR [17] and HIPAA [18] may also impact patient data.

The data owner must be knowledgeable about how the data is used throughout the data lifecycle (data that may span the organization), and how and where it is retained in a secure archive for the required period. Data ownership is typically associated with a particular job title or function in the organization, and each new person to that job inherits the data ownership. Where an internal reorganization impacts job roles and responsibilities, special attention should be paid to ensure that data ownership is not lost or overlooked.

Where data crosses organizational boundaries within the process, data ownership may be transferred to a new data owner, although the original data owner may always remain accountable for the original data generated within their scope. For example, analytical data owned by the Quality Control (QC) laboratory may be incorporated into a manufacturing batch record, which has its own data owner. In the event of an investigation or audit, any queries about the analytical data within the batch record requires assistance from the QC laboratory data owner.

Uncertainty in data ownership can be introduced during mergers and acquisitions and should be resolved as part of the integration process, see Section 3.7.

It is important to separate the data owner from the system owner, who is responsible for system maintenance roles (for example, IT groups responsible for the operating system, application, and database), to prevent a conflict of interest. With current working practices, interest in the data often spans different departments (e.g., IT security is concerned about safeguarding the data, and the legal department may be concerned with data privacy.) Data owners have a direct interest in the data and therefore should not have system administration and/or privileged accounts.

Many organizations create a data steward role for the technical people with day-to-day responsibility for the data. Data stewardship activities may be embedded in the responsibilities of other roles, rather than being a new and specific individual role. The Data Steward role is defined in the *ISPE GAMP® Guide: Records and Data Integrity* [8] as:

“A person with specific tactical coordination and implementation responsibilities for data integrity, responsible for carrying out data usage, management, and security policies as determined by wider data governance initiatives, such as acting as a liaison between the IT department and the business. They are typically members of the operational unit or department creating, maintaining, or using the data, for example personnel on the shop floor or in the laboratories who actually generate, manage, and handle the data.”

2.2.4 Data Owner Involvement in Infrastructure Changes

In IT Service Management (ITSM), a platform change (e.g., database migration or cloud hosting provider migration) may often be considered an infrastructure change, which involves the system owner but not necessarily the business process owner. At a minimum, the data owner (as delegated by the business process owner) should be consulted before any change occurs and be kept informed through its implementation.

At all times, the data owner should know where the data resides, and the process followed to verify any data migrations, considering the mechanism used, reliance on network bandwidth during migration, error reporting tools used, any data transformation etc., such as non-Unicode™ to Unicode™ conversion, or database changes from SQL to Oracle®. The level of verification should be commensurate with the risks of the infrastructure change and any migration involved.

2.2.5 Data Ownership Across Multiple Organizations

Data ownership may be simple when a single organization owns and uses the data throughout the data lifecycle, but when the creation, processing, and/or reporting of data is contracted to a second organization, the concept of data ownership becomes more complex. The contract giver or sponsor company legally owns the data (even if they only use data in summary form), while the contract acceptor may be the entity responsible for the creation, processing, and reporting of the data.

The contract acceptor should have a data steward to address the day-to-day data responsibilities within the contract organization, and report all concerns and risks to the data owner within the contract giver company who is ultimately responsible for data ownership. The contract acceptor may be responsible for the data throughout the data lifecycle. All of these responsibilities and expectations must be clearly defined within the quality contract and/or technical agreement and verified through the vendor qualification process. In an environment involving multiple organizations, a robust communication plan is necessary to ensure potential data integrity issues are reported to the “legal” data owner. In this context, potential data integrity issues are not limited to instances of suspected data manipulation but also include cyberattacks or security breaches, and the report should address the impact of these events.

3 Retention Strategy

A strategy should be developed to ensure the long-term retention of regulated or operational data output by a business process. The retention strategy should focus on the management of the data throughout its lifecycle, not just on the initial creation of data within the process.

Use of electronic records as the original/official records supporting business processes continues to increase. Therefore, it is imperative to apply critical thinking to develop the retention strategy in the planning phase for a new process or system. This will help ensure data is maintained in its dynamic format throughout the data lifecycle as long as reasonably possible.

The risk is that as technology advances, old file types become obsolete, driving companies to:

- Maintain obsolete technology to continue to read and access the files, which can increase data integrity risks
- Migrate the data so that new technology can continue to read the record

There are risks associated with any strategy, and assessing those risks and planning a broad methodology for record and data retention at the beginning of a system's lifecycle allows the plan to be adjusted as technology changes or as systems go through their lifecycle. As such, it is imperative for a company to have a strategy in place that will lead their employees through the process; without this, the company increases its risk as the system moves toward its end of life. The longer a system goes without a retention strategy, the greater the likelihood the company will lose the ability to migrate or maintain records to the end of their retention period. Typically, as data ages, the need for it to be processible decreases until at some point, a risk-based decision to change to a static format may be appropriate. This is discussed in Section 7.2.

In order for data integrity to be ensured within the system and data lifecycles, a retention strategy should be crafted during the design phase and ensure that data remains secure and protected from deliberate or accidental changes, manipulations, or deletions throughout the retention period. The strategy for archiving GxP data also needs to be planned, whether it is archived electronically or using a paper-based system.

An electronic archiving system, both hardware and software, must be designated as a computerized system and validated to ensure the integrity of the data, and make certain that it is accessible and readable. See Appendices O2 and O3 for more details on archiving requirements. The data governance for the records in any new computerized system and the archive solution should be aligned early. Where problems with long-term access to data are envisaged, or when computerized systems must be retired, mitigation strategies for ensuring continued readability of the data should be established. These may include virtualization of the original computerized systems, which is discussed in Appendix O5.

As regulated companies increasingly utilize cloud-based services, a SaaS provider may need to offer appropriate strategies to ensure adequate controls are in place for data retention. See *ISPE GAMP® Good Practice Guide: IT Infrastructure Control and Compliance (Second Edition)* [19] for additional considerations around cloud-based services and data retention.

The high-level elements of the strategy should be defined in a corporate data governance document as discussed in *ISPE GAMP® Guide: Records and Data Integrity* Section 3.3.6 Policies and Standards [8], listing the necessary elements, plus ones to consider, for the record retention of a system or project. Refer to Appendix O1 – Retention, Archiving, and Migration [8] of the same Guide for information on how to document the retention strategy for a particular system.

The key points to a retention strategy are knowledge and consideration of data classification and intended use, readability, availability, access, and data ownership. A good place to start when developing a strategy is to document this information in the form of a risk assessment with the goal of documenting the business needs, associated risks, and current mitigation plans. As new information becomes available over the system lifecycle, the assessment can be updated easily to reflect such changes. See Figure 4.1 Data Integrity by Design Process Flow Diagram for an overview of the assessment during the system's lifecycle. This approach keeps the retention strategy up-to-date and representative of the current needs for the system against the ability and feasibility in technology. The following sections provide a more thorough breakdown of the details to consider.

3.1 Retention Periods

A regulated company generates many records with varied retention requirements during the course of its business processes, for example:

- Regulated records subject to a mandated retention period
- Regulated records retained beyond the minimum for business reasons (see Section 7.4.1)
- Records of scientific innovation or invention (including drug development data) that may be needed in perpetuity for legal reasons
- Financial records falling under the Sarbanes-Oxley Act 2002 (US) [20] (or similar legislation in other countries)
- Records containing personally identifiable information (for example, human resources records and patient enrollment records for clinical trials)
- Operational records needed solely for process improvements and business decisions (see Section 1.6.5).

Each of these records are required to be available for a different duration depending on what, if any, legislation applies to that data classification (see Sections 1.6.5 and 4.4 for simple classification, and the more detailed example in *ISPE GAMP® RDI Good Practice Guide: Data Integrity – Key Concepts*, Appendix 8 [16]). Even within the regulated records category, different retention periods are specified. Additionally, there may be legal and business reasons for extending retention periods, including possible future use of the record.

It is therefore important to identify what records fall under what legislation, and the applicable retention periods. Where different records have drastically different retention periods (they can vary from 6 months to 30 years), it is important to segregate their storage. For example, a database containing a mixture of data (GLP, GCP, GMP, GDP, and business records) needs to be kept for the duration of the longest retention period of any of the records, giving rise to long-term storage of an excessively large amount of data, much of which is no longer required.

Where an organization is creating data that is relevant to a number of other entities (e.g., training records within a Contract "x" Organization (CxO)), it is important that the organization has a clear approach for identifying and managing the data – both for operational use and when archiving/destroying data – for the individual entities. Implications concerning third-party data are discussed in Appendix M3.

The retention period, if not specified, should be sufficient to support any challenges to data integrity. In the absence of a required retention period, the final disposition should be documented.

Records should be protected to ensure they are accurate and readily retrievable throughout the retention period. Readability is implied in retention period and discussed in Section 3.2. Readability of audit trails and other metadata is required under predicate rules to preserve the GxP content and meaning of the record, and should be retained and available for review and copying for at least as long as the original record. During system planning and selection, it is essential to include a user requirement for the ability to archive the complete data including audit trails and metadata.

Failure to meet this requirement could result in either an inability to include the audit trails and metadata in the archive, or a need to preserve the entire original database to allow recreation of the audit trails.

There may be business needs that require data to be retained beyond the GxP-mandated retention period; these business needs are discussed in Section 7.4.1.

Appendix O4 Tables 16.1 and 16.2 contain examples of retention requirements and periods. Appendix M3 discusses the important considerations around managing data generated by third parties, for example, CROs.

3.2 Readability

3.2.1 Importance of Readability

Data should remain readable during the complete data lifecycle, that is:

- Able to be displayed in a human-readable format
- For dynamic data, retained in such a way that it can be further interacted with if required by the intended use and/or specific predicate rule requirements (see Section 4.4 on intended use)

With paper records, no special tools are required to read the data. If the paper record has endured (survived), it can be read as a stand-alone complete record. It can be converted to a static electronic image, that is, a PDF or JPG file by a scanning process, and easily and reliably read on any system having PDF or image viewing software. This should be a verified process to ensure the electronic copy is a true copy, and is discussed in detail in Section 7.2.1.

In a computerized system, original records generated electronically often require additional resources such as databases to supply the record content, along with associated metadata to provide the context and meaning of the record. Typically, data created in an electronic system is readable by that same system, and often readable only by that system, due vendor-proprietary formats and database structure.

If the record is static in nature (i.e., does not require user intervention or processing to be meaningful), then it can be converted to a PDF or similar format by a verified process to provide ease of readability. Section 7.2.1 discusses the differences between static and dynamic data and implication thereof.

3.2.2 Impact of Upgrades or Retirement on Readability

Changes to the software application such as a software upgrade or database structure modifications can result in a loss of readability for data created in earlier software versions. For readability considerations around inactive data stored in a dedicated archiving system, see Section 7.2.2.

It is recommended to ensure software is backward compatible prior to the upgrade. If the software does not provide backward compatibility, then confirm that the data can be migrated as and when required. Controls around data migration are discussed in Section 3.5.3.

Changes to a GxP computerized system should be implemented under formal change control and configuration management, which is detailed fully in *ISPE GAMP® Good Practice Guide: A Risk-Based Approach to Operation of GxP Computerized Systems* Chapter 10 [21]. When reviewing a change proposal for a computerized system, it is important to review vendor release notes for:

- Changes to the data structure or file format in the new version, and vendor claims around record compatibility between old and new versions

- Changes to the storage approach, e.g., change in the database version or even to the database type (Oracle® versus MySQL™, etc.)

Where there will be an impact on the readability of the existing data in applying the software upgrade, remediation activities are needed, such as:

- Using vendor tools (if provided) to migrate the existing data into a newer format readable by the upgraded software, and verifying the migrated data remains complete (including metadata), accurate, and consistent compared to the original records. Data migration testing, including advice on statistical sampling, is covered in detail in the *ISPE GAMP® Good Practice Guide: A Risk-Based Approach to Testing of GxP Systems (Second Edition)* Appendix T9 [22].
- If a vendor tool is not available, determine if there is a commercially available tool that can provide the needed data migration.

Data migration is discussed in detail in Section 3.5.3.

Even where there is no explicit vendor statement on the impact to the data readability, it is still recommended to verify that a sample of existing data is readable (and, for dynamic data, can be interacted with) after the upgrade.

Data sampling should be risk-based and leverage statistical approaches to ensure sufficient data is sampled to identify (and correct) any missing or corrupt data. This is in addition to, and separate from, validating the change to the computerized system.

Where there is an upgrade to a computerized system containing master data (e.g., manufacturing recipes, customer contact details), formal verification may be required to confirm that the data is unaffected and remains aligned across integrated systems.

In the event of a database change, an extensive software upgrade with no migration tool or backward compatibility, or at system retirement, an evaluation should be made of the feasibility of maintaining a functional copy of the original computerized system (either as a physical or virtual machine) that will be used to read and interact with existing data throughout the remaining retention period for that data.

Considerations for maintaining functional copies of the software include:

- Newer Operating Systems (OS) have better backward compatibility features that may be able to run legacy software without much trouble if it was produced in the last several years and if the data storage remains in the same structure (e.g., flat file throughout or the same relational database).
- For laboratory and process control software, it is important to determine if the software application can be run without connection to instrumentation and process equipment.
- The practical implications and limitations of maintaining a functional copy of an obsolete system using virtual machine technology, as covered in detail in Appendix O5.

This should be combined with an assessment of the risk of not keeping the records in a dynamic format (see Section 7.2.1). At some point, as discussed in Section 7.2, maintaining readability of the old dynamic records may become unfeasible and conversion to static format is inevitable.

One risk associated with long-term readability of records is personnel turnover. The personnel that generated and interacted with the data may no longer be with the company and operation of the legacy system may be challenging with inexperienced operators. It is advisable to archive training materials with the software so there is at least some information available when accessing the old data by inexperienced personnel. Any information that can be provided to a new user about the data also helps the data owner answer questions during a regulatory review.

3.2.3 Recommendations for New Systems

When selecting a new system for use with GxP data, incorporate the retention and archiving requirements into the user requirements.

When considering new systems, it is important to evaluate:

- The available functions to export and report data out of the system for archiving purposes
- The system's ability to mark data as "Archived" and how archived data will be removed from the originating system
- How existing active data is impacted during a system upgrade:
 - Is the data automatically upgraded as needed to suit the new software version as part of the system upgrade?
 - Does the data have to be revised or modernized by a separate step to keep it readable?
 - Does the data need to be imported into the upgraded system?
- The history of earlier versions of the system to understand the depth and frequency of changes to data structure, file formats and databases, and also to assess the level of backward compatibility routinely maintained in the system.

These could be addressed in a vendor evaluation; minimally, there should be a discussion with the vendor to understand their vision and plans for the product going forward.

When a system is upgraded (upgrade of the application or operating system, or implementation of a later-model system), the existing data may be automatically revised to be readable within the system, either during the upgrade or when transferred into the upgraded system. This relies on the capability of the new system to read data created in the older version, possibly by converting it into the newer format. This type of transfer is the simplest form of migration but still requires a risk-based level of verification to confirm the integrity of the data. Special attention needs to be paid to any changes in the processing algorithms that could produce a different result.

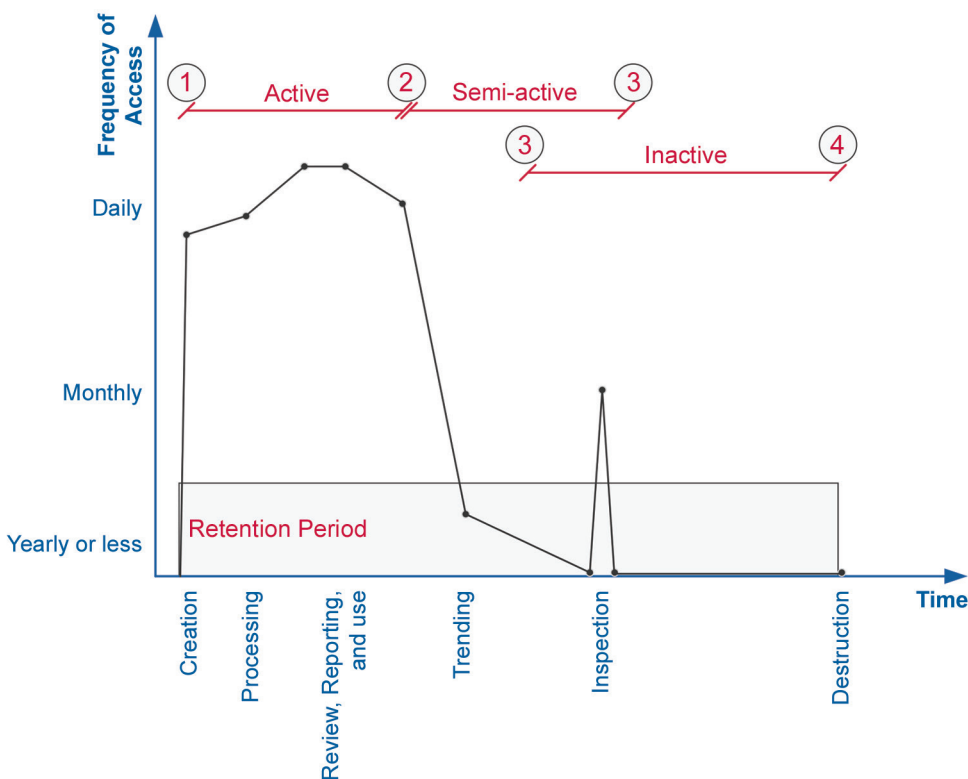
A system that internally manages its data revision during an upgrade reduces the effort required to keep data readable in the system through future system upgrades during its operational life. A system with the inherent ability to read data created in all earlier versions of the system, including data restored from the archives, greatly simplifies the ongoing data retention burden.

The capability to export data for archiving and retention is critical when using SaaS solutions hosted off-premise. There needs to be a mechanism to extract the data if and when the regulated company moves away from that solution; this requirement should be defined in the Service Level Agreement with the SaaS provider. Many of the considerations discussed in Appendix M3 – Third-Party Data may also be relevant for SaaS solutions.

3.3 Availability

The need for the availability of records is determined by the current state of the records: active, semi-active, or inactive.

Figure 3.1: Example Frequency of Records Access over Time



In the Figure 3.1 diagram:

- ① Active records start from the first creation of data. Generally the record remains active through the processing and review, reporting and use phases of the data lifecycle, where access is needed frequently to progress the data to complete its business purpose, that is, its use. Active records are introduced in Section 3.3.1 and discussed in detail in Chapter 6.
- ② Semi-active describes the situation where the record is no longer routinely accessed, but may be needed periodically such as for use in trending and control charts, and during annual product review or complaints investigations. Semi-active records are introduced in Section 3.3.2 and discussed in detail in Section 7.1.
- ③ Inactive records are records no longer expected to be accessed other than for exceptional circumstances, for example during an inspection or investigation. At this point in the record lifecycle, the records will be infrequently or minimally accessed for the remainder of their retention period. Inactive records are introduced in Section 3.3.3 and discussed in detail in Section 7.2.
- ④ At the end of the retention period, the inactive record can be destroyed as long as there is no legal hold or business reason to keep the data. Destruction is covered in Section 7.4.

There are no hard-and-fast rules as to when a record should transition from active to semi-active, or from semi-active to inactive; it is a spectrum of decreasing need to frequently access the data and a desire to maintain performance in the live system(s). Points ② and ③ can move within the graph based on the regulated company's choice.

It should be noted that a record entering the semi-active or inactive state should not impact its static or dynamic state – the regulatory expectations to maintain dynamic data in dynamic format remain.

3.3.1 Active Records

Data stored in the computerized system that originally generated the data (the “live system”) is immediately available to authorized users of that computerized system (although there may be restrictions on which users can access what data – see Section 3.4). This immediate availability is efficient where data is active (being generated, processed, reported, reviewed, and/or used) and being used by a range of interested parties (production operators generating data, production supervisors collating and reporting data, quality unit reviewing data, authorized person releasing the batch based on the data). Active records are discussed in detail in Chapter 6.

3.3.2 Semi-active Records

As it moves through the lifecycle, data may be needed less frequently; for example, batch release data from the previous month may not be needed until the annual product review when it will be trended with other release data. In this semi-active phase, where the data is needed infrequently, the options available to an organization are to:

- Leave the data in the live system
- Leave the data in the live system but prevent further changes by segregating the semi-active (or inactive) data from the active records, that is, the data is restricted as read-only or is rendered inaccessible to the majority of users
- Look at off-line storage solutions, for example, electronic archives as discussed below and accept the consequently slower access to the data inherent in moving it out of the live system. This should be balanced against the need for data to be rapidly available in the case of a recall scenario.

Based on the business process and regulatory requirements, a read-only or reduced access approach may be used to limit the potential for data changes, in which case ideally the system should offer incremental levels of control that can be implemented over time, such as:

1. No new data can be written into the data folder but allowing users to work with existing data
2. Progressing later to preventing further processing or editing of the data, but the data is still available for viewing and approval
3. Finally, after the business process cycle time is complete, moving to fully read-only

There may even be a “no access” level of control whereby the data can no longer be viewed, pending its transfer to the archiving system.

It is important to limit which users can initiate the restrictions, with potentially a smaller subset of users able to reverse the restrictions in case the data is needed for an investigation or audit. This is particularly important for chromatography data where an inspector can request the data to be reprocessed to assess the impact of alternate processing parameters.

Semi-active records are discussed in more detail in Section 7.1.

3.3.3 Inactive Records

When an organization determines that the data is no longer needed to be readily available for instant access, or the amount of data in the live system is degrading system performance, moving the data from the live system to an electronic archive is recommended. This brings multiple advantages:

- Reduced workload for managing active records
- The amount of data in the live system does not become sufficiently large to impact system performance (e.g., long search times, delays in reading data from a database)
- The time required to back up the live system remains manageable (e.g., less than 24 hours, thus allowing daily backups)
- The storage space required for backups of the live system is minimized
- The archived data is only accessible to a limited number of independent personnel and no longer available to the originating users, thus reducing the risk of alteration during the retention period [23]; this is a specific regulatory requirement for some areas. (See Appendix O4 for examples of requirements for record retention.)

There are more general regulatory requirements (listed in Appendices O2 and O3) for the use of archives in support of long-term data retention. In its simplest form, archiving electronic data requires the creation of a complete and accurate copy of the data (including metadata to preserve GxP content and meaning) in an off-line storage location. Once the off-line copy has been verified as complete and capable of being restored into the live system when needed, the original record can be deleted from the live system. As with any other regulated data, provision should be made for backing up the archived data to guard against data loss.

Where validated automated archiving processes cannot be provided, manual archiving may be required. Manual archiving can only be controlled by procedure/protocol and requires verification that all the required records are archived. Manually administering the electronic archiving process, especially when dealing with flat files, carries an inherent risk that more data may be deleted than was copied to the archive location, for example by choosing to copy only the passing results for long-term retention. For this reason, validated automated electronic archiving systems are preferred.

See Section 7.2.2 for a discussion of the important features for an electronic archiving system, including advanced indexing capabilities to facilitate record searching. Archive requirements are included in Appendices O2, O3, and O4.

If a manual archive process is all that is available for a system, an archive protocol must be written to ensure all data is appropriately archived.

Some companies choose to maintain the data in a secured area of the live system in a “no access” state for the full retention period. If that approach is used, the data may be secure, but the advantages listed above will not be achieved.

Inactive records are discussed in more detail in Section 7.2.

3.4 Access

Inherent in considerations regarding the availability of the data is the consideration of who should be able to access the data during its lifecycle. Access controls should ensure that only authorized personnel can access data. This may be controlled by business role and organization.

A detailed discussion of security management for computerized systems is contained in *ISPE GAMP® Good Practice Guide: A Risk-Based Approach to Operation of GxP Computerized Systems* Chapter 15 [21] and in *ISPE GAMP® Good Practice Guide: IT Infrastructure Control and Compliance (Second Edition)* Appendix 5 [19] for infrastructure security.

3.4.1 Access to Live Systems

When active records are stored in the live system, access controls in that system are based on the principles of:

- **Segregation of Duties:** This is the general concept of having more than one person required to complete a process or workflow as an internal control intended to prevent fraud and error. For example, segregation of duties requires that authorization to generate data should be separate to authorization to verify data (e.g., a user can review their peers' data but not their own), and that users conducting normal work tasks in the system should have access rights to do so while users with elevated access rights (e.g., administrator privileges, engineer roles) should not conduct normal work tasks on the system [24].
- **Least Privileges:** A further aspect of assigning access to functionality is the principle of least privileges, whereby each user is given sufficient access for their routine tasks and no more. Senior users may be given access to higher risk functionality (e.g., creating new data storage folders or adding new users), with the highest risk functionality (e.g., system administration, database administration or other enhanced access) reserved for users outside of the department's reporting structure (for example, members of the corporate IT department). Careful consideration should be given to how and where passwords are managed, including resetting forgotten passwords; there is potential for fraud if someone in the business process can reset a password and then conceivably execute tasks under someone else's account.

Assigning access controls, including database administrator access, is discussed in detail in *ISPE GAMP® RDI Good Practice Guide: Data Integrity – Key Concepts* Section 4.5 [16].

In addition to controlling which functionality in the system is accessible to a particular user, there may also be a need to control what data can be accessed by that user. This is especially relevant where a computerized system is shared across multiple business processes and/or organizations, for example sharing a chromatography data system across both the research and development and QC laboratories, or where a system contains confidential patient information.

3.4.2 Access to Archive Systems

In an electronic archiving system, access to the archived data can be restricted to a limited number of data management personnel and no longer available to the originating users thus reducing the risk of data alteration during the retention period [23].

If there is a need to restore the archived data to the original system or a functional copy thereof, special care must be taken to manage any access to or editing of the restored data. This is discussed in Section 7.3.

System lifecycle considerations for archive systems are covered in Section 7.2.2; archiving requirements are discussed in more detail in Appendices O2 and O3.

3.4.3 Physical Security

Logical security (system access controls) alone are not sufficient [11, 25]. Physical access controls such as perimeter security, building and facility access control should be in place in addition to logical security. The storage location must be physically secured against unauthorized access, for example, by instigating key card door controls on the IT server room such that only authorized IT personnel can gain access to the system server or archive storage media.

3.5 Protecting Records and Data

It is important to protect records supporting GxP decisions to ensure that the information is secure from unauthorized modification or deletion. The procedures and processes implemented to address risks to active records should be commensurate with the importance of the data, the process, and complexity of the system. All data must be in a format that permits secure storage and available for review and reporting.

As discussed in the *ISPE GAMP® Good Practice Guide: IT Infrastructure Control and Compliance (Second Edition)* [19]:

“Lack of security may compromise availability of applications and services, record integrity and confidentiality, reputation with stakeholders, and may lead to unauthorized use of systems that would ultimately impact product quality.

Information security may be characterized as the preservation of:

- *Availability: ensuring that authorized users have access to information and associated assets when required*
- *Integrity: safeguarding the accuracy and completeness of information and processing methods*
- *Confidentiality: ensuring that information is accessible only to those persons authorized to have access.”*

3.5.1 Site Storage and Controls

There are processes, controls, and procedures necessary to protect the safety, confidentiality, integrity, and availability of the data in the operational environment. Such safeguards include (but are not limited to):

- Backup and restore procedures
- Business continuity planning
- Disaster recovery planning

It is important to develop the backup approach and disaster recovery strategy based upon the risk of the system and the criticality of the process and data. Backup and restore, and disaster recovery as part of business continuity management are discussed in detail in Chapters 13 and 14 of the *ISPE GAMP® Good Practice Guide: A Risk-Based Approach to Operation of GxP Computerized Systems* [21].

Backup and restore procedures ensure that the accurate and reproducible copying of digital assets, including the data and software, are protected against loss of the original data so that the data can be restored, if required due to a disaster.

Conventional backup processes typically involve creating a duplicate copy of system data on a fixed-time basis, such as a nightly scheduled backup. Other controls that can be implemented utilize sophisticated network architectures including database mirroring and redundancy across diverse and geographically dispersed datacenters, which can result in close to 100% uptime. Synchronous replication – where data is written to the primary storage and a remote replica simultaneously ensuring that both versions are always identical and where one replica is always available – is another approach to lessen the risk of loss of data. Data transfer speed should ensure that replication is performed at a necessary rate to prevent degradation or loss.

The Business Continuity Plan and the Disaster Recovery Plan collectively address how to continue or resume operation after a disaster. A disaster is defined within the *ISPE GAMP® Good Practice Guide: IT Infrastructure Control and Compliance (Second Edition)* [19], as:

“Any event (i.e., fire, earthquake, power failure, etc.) which could have a detrimental effect upon an automated system or its associated information.”

The Business Continuity Plan (BCP) describes the steps the business must follow to restore the critical business process following a disruption and addresses what the business must do to continue without the computerized system(s).

A Disaster Recovery Plan (DRP) is defined by the *ISPE GAMP® Good Practice Guide: A Risk-Based Approach to Operation of GxP Computerized Systems* [21] as:

“A sub-set of Business Continuity Management that focuses on regaining access to an IT system, including software, hardware, and data following a disaster.”

Disaster recovery must address the provision for replacing or restoring the computerized system. The DRP must provide details of how to restore the complete computerized system including the infrastructure, hardware, application, and database. The recovery site is usually chosen a geographical distance from the main site in case the disruption is due to a natural occurrence such as an earthquake or flooding. If there is a need to restore specific instruments, as in the case of laboratory systems, the DRP should contain procedures for obtaining replacement instruments.

It is important to document how much downtime is acceptable for a specific computerized system/business process. The process owner must define the allowed maximum system down time expressed as the Recovery Time Objective (RTO) and the maximum loss of data that can be tolerated by the business process defined as the Recovery Point Objective (RPO). This information is used by the IT specialist to design the necessary backup process and architecture to support these requirements. Companies must ensure response times are described within their internal or external Service Level Agreements (SLAs).

All of these processes, controls, and procedures should be periodically verified and may be assessed during periodic review to provide assurance that they still meet organizational and regulatory expectations, including doing trial restores of the backup data. Consideration should be given to perform walk-throughs and mock-executions of the DRP and BCP.

3.5.2 Off-Premise Storage and Controls

The same data integrity requirements (including confidentiality and security) apply to systems managed internally by the regulated company or externally. The expectations from the regulators are the same whether organizations use physical infrastructure on-premise, virtualized servers on-premise, hosting at datacenters, or leveraging off-premise cloud-based services. The regulated company has the ultimate responsibility for patient safety, product quality, and data integrity, whatever and however the data is stored.

In cases where the system is supplied using cloud-based infrastructure (SaaS, PaaS, or IaaS), the responsibility for the procedures related to disaster recovery are shifted in whole or in part to the cloud provider. One of the benefits of a cloud-based infrastructure is its resiliency, that is, the ability of the service to respond to issues, although the positive achieved in resilience is often challenged by a decrease in security controls. One of the advantages of using a cloud-based infrastructure is system uptime, which providers often report as 99% to 100% [26].

The *ISPE GAMP® Good Practice Guide: IT Infrastructure Control and Compliance (Second Edition)* [19] provides some insight into the advantages and risks of using various infrastructure services.

“Advantages for IaaS, SaaS, and PaaS solutions include:

- *Large flexibility in capacity and services*
- *Easy to expend or reduce*
- *Reduced internal hardware foot print*
- *Reduced capital costs*
- *Outsourced control and maintenance*

- *High availability*
- *Less need for IT knowledge and IT investment”*

“Cloud computing introduces a flexibility in resource capacity, but also introduces new risks to regulated companies. These risks include:

- *Less or no control over the datacenter*
- *Several suppliers working together to provide the infrastructure*
- *Less control over the infrastructure*
- *Less control over the data*
- *Less control over applied services*
- *Data and systems are outside the companies’ network*
- *Different qualification approach required.”*

The use of infrastructure outsourcing has also presented compliance challenges including security, availability, integrity, confidentiality, and should the contract be terminated, data deletion or an inability to access the data.

The two critical activities/considerations for the users of cloud-based infrastructure services are the supplier assessment and the contract or Service Level Agreement (SLA).

As with any supplier assessment, it is critical to assess the services delivered by the cloud provider and the controls and procedures established to ensure the operations. It is important to define the expectations of the regulated company and the corresponding responsibilities of the provider. One of the considerations is the notification of changes to the hosted service, including upgrades to infrastructure, platform, or software application. At a minimum, the notification should include details of the change to allow the regulated company to evaluate any potential impact to their systems and data.

As part of this assessment, it is important to evaluate the disaster recovery process at the supplier including review of the DRP and the frequency and conditions of the disaster recovery testing. The DRP should address the communication strategy during a disaster. It is important to document how a disaster at the cloud provider will be communicated to the system users.

It is also important to determine if the cloud-service provider subcontracts any part of their service. If this is the case, such as a SaaS provider using a third-party PaaS or IaaS provider, it is important to understand how the cloud-service provider evaluated their subcontractors and how problems with service at the subcontractors will be prioritized and communicated.

The *ISPE GAMP® RDI Good Practice Guide: Data Integrity – Key Concepts* [16] states:

“Even if the third party is not subject to healthcare regulations when they enter into a Quality Contract/Technical Agreement with a regulated company, they must be aware of the requirements of the environment in which they are working.”

Ongoing performance monitoring of service delivery from the provider is important.

Appendix M3 contains additional guidance on managing off-premise data.

3.5.3 Migration

The MHRA 'GXP' Data Integrity Guidance and Definitions [12] provides significant guidance on controls required for data transfer and data migration. Fundamentally, any transfer or migration process must be validated and must preserve the GxP content and meaning of the data.

The *ISPE GAMP® Guide: Records and Data Integrity* Appendix O1 [8] discusses the option of storing electronic records in formats other than the original record format. Appropriate points when considering record transfer or migration include:

- *"Creation of the record"*
- *The point at which a record is to be archived*
- *At system upgrade, especially if electronic record conversion is needed*
- *At system replacement when contemplating data migration to the new system*
- *At system retirement, especially if electronic record conversion or development of rendering software is needed*
- *When a media refresh is needed"*

The decision to convert electronic records into a different format (an alternative file type, or dynamic to static) should be based upon a documented risk assessment considering the requirements for record retention, access, and use. *ISPE GAMP® Guide: Records and Data Integrity* Appendix O1 [8] includes a comprehensive table, Table 18.1, that describes multiple risk factors for the conversion of electronic records to an alternative format.

Transferring data between different applications will also likely require a conversion of format. When this is needed for active data, discussions with the vendor of the target system will determine if the entire active data set can be converted directly:

- Via direct conversion and transfer
- Via an export from the current system and import into the new application

One challenge with retention of records is that the records may be in a format proprietary to an individual vendor and therefore the data may only be read in the originating application.

The conversion of inactive records to a vendor-neutral format was introduced in the *ISPE GAMP® RDI Good Practice Guide: Data Integrity – Key Concepts* Section 4.3.6.3 [16]. The vision of vendor-neutral formats has been a desire for industry for many years, especially in the realm of analytical laboratory data. Leveraging a vendor-neutral format could help to break the reliance between the data lifecycle and the system lifecycle. Data would no longer be exported from one system and imported into another, but exist in file structures that live in a common data lake with the ability of various applications to open, view, interrogate, and mine the data for new scientific insights, as well as provide the ability to easily share data between collaborating companies. This is an ongoing industry initiative and not yet fully implemented or available.

Any "temporary" storage locations of data need to be tightly secured and controlled. There should be a record of transferring data between storage types, formats, or computerized systems. A Data Migration Plan should define the scope of the migration, the personnel and data involved, and the strategy to be applied. The *ISPE GAMP® Good Practice Guide: A Risk-Based Approach to Operation of GxP Computerized Systems* Chapter 17 [21] examines migration in detail and offers guidance on creating a Data Migration Plan. Migrating data of questionable integrity does not improve the integrity of the data but may protect the data from further risks going forward.

As with all migration/conversion activities, it is essential to consider if all data, metadata (including audit trails and review and approval signatures), and relational traceability between records is preserved after migration to the new application. Care must be taken when migrating IT systems with databases to ensure that the associations are not disrupted and that the data viewing capabilities remain. Risk assessment and management for any changes in data or metadata should be evaluated and documented. It is important that migration or data transfer processes are designed and validated to confirm that data integrity is preserved (the accuracy, completeness, content, and meaning of the data) following a risk-based approach.

When the migration/transfer involves cloud-based solutions, there are additional considerations required as described in the *ISPE GAMP® Guide: Records and Data Integrity* [8]:

“As part of an outsourcing process regulated companies need to:

- *Understand and accept which aspects of control are being delegated to a provider*
- *Assess and accept the controls implemented by the provider*
- *Contractually define the level and frequency of reporting*
- *Agree on the need for supplier support during regulatory inspections, depending on the architecture and services provided”*

Often using PaaS or SaaS solutions, system upgrades may be driven by the service provider and not by the regulated company. If the upgrade forces the need for data migration, a risk-based approach should be undertaken to managing the migration. Appendix O5 offers additional considerations around off-premise solutions.

3.6 Managing System Retirement

As stated in the *ISPE GAMP® RDI Good Practice Guide: Data Integrity – Key Concepts* [16], system retirement:

“involves decisions about data retention, migration, or destruction, and the management of these processes. Retiring a computerized system has major implications about how the data created in the system during its operational life remains available, enduring, and readable in the remaining phases of the data lifecycle.”

When retiring a system, it is important to determine:

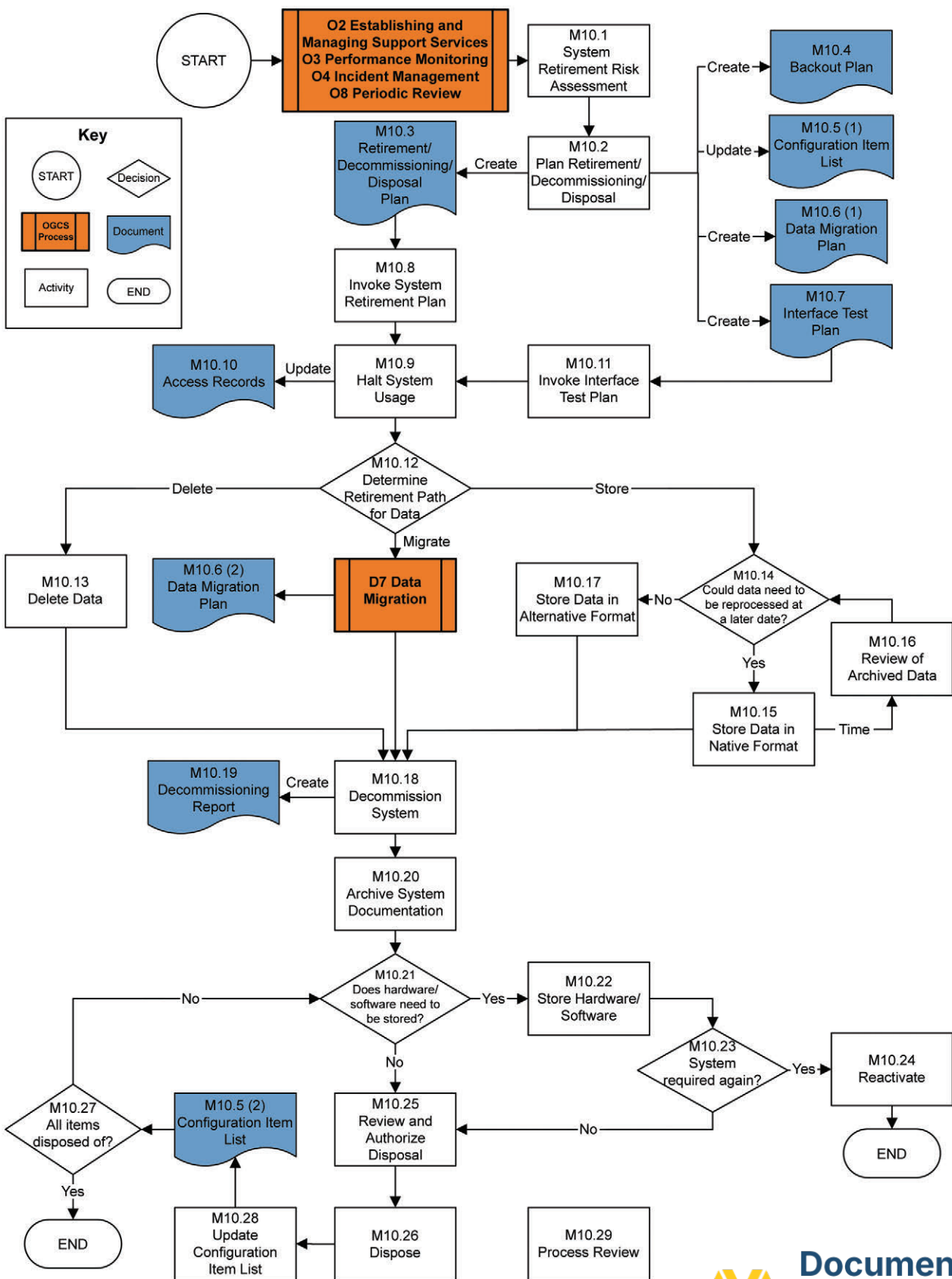
- What data in the system needs to be retained after the system retirement (it may need administrator access to view all data within the system)?
- What data is within its retention period that needs the system to read and display the data?

For dynamic data, a functional copy of the legacy software may be needed to read or interact with the data if the data has not been migrated to a format readable in a current system. This is discussed more in Section 3.2.2, with details around maintaining legacy software covered in Appendix O5.

Appendix M10 of *ISPE GAMP® 5* on System Retirement [9] presents an overview of the system retirement process including data integrity considerations.

Managing system retirement is comprehensively discussed in the *ISPE GAMP® Good Practice Guide: A Risk-Based Approach to Operation of GxP Computerized Systems*, Chapter 18: System Retirement, Decommissioning, and Disposal [21], which contains a detailed process flow diagram showing the critical activities and records to be produced (see Figure 3.2), plus a RACI matrix (Responsible/Accountable/Consulted/Informed) of the roles and responsibilities involved.

Figure 3.2: Managing System Retirement Workflow [21]



The primary consideration during system retirement is to ensure that data and metadata generated and stored in a system are retained and readily available and readable as defined in the system retirement plan. The system retirement process should have been considered during the creation of the organization's retention strategy. The strategy should be based on risk, taking into account regulatory expectations and business requirements, data criticality, and the complexity of the computerized system. The strategy should require the identification of record type (GMP, GLP, GCP, etc.), record retention period, and archival and retrieval processes for each system.

The 2010 *ISPE GAMP® Good Practice Guide: A Risk-Based Approach to Operation of GxP Computerized Systems* [21] remains valid in every respect, but consideration of subsequent regulations and guidance on data integrity, specifically expanding roles and responsibilities to include data owner and data steward, generates a slightly modified RACI matrix as presented in Table 3.1.

Table 3.1: Managing System Retirement RACI Roles and Responsibilities

Adapted from ISPE GAMP® Good Practice Guide: A Risk-Based Approach to Operation of GxP Computerized Systems [21]

Role	RACI Code	Responsibilities
Process Owner (can also be the Data Owner)	A	- Accountable for the Retirement Plan - Accountable for the execution and verification of the Retirement Plan - Accountable for the data and documentation related to the system being retired and ensuring compliance with relevant retention policies - Responsible for approval of the Retirement Plan and Report
Data Owner	R	- Responsible for preserving the quality and integrity of data and documentation related to the system being retired and ensuring compliance with relevant retention policies - Consulted on content of the Retirement Plan
System Owner	R	- Responsible for preparing aspects of the Retirement Plan - Responsible for approval of the Retirement Plan and Report
Quality Unit	C	- Consulted on content of Retirement Plan for regulatory and compliance aspects - Responsible for approval the Retirement Plan and Report
End User	I	- Informed of the plan to retire the system - Responsible for verifying aspects of the Retirement Plan, e.g., ongoing data accessibility, impact on interfaced systems
Platform Support (SME)	R	- Responsible for preparing aspects of the Retirement Plan - Responsible for executing platform dependent aspects of the Retirement Plan
Application Support (SME)	R	- Responsible for preparing aspects of the Retirement Plan - Responsible for executing the Retirement Plan
System Administrator (SME)	R	Responsible for preparing aspects of the Retirement Plan
Archivist (SME)	R	- Consulted on content of Retirement Plan for organization records retention policies - Responsible for executing the archive aspects of the Retirement Plan - Acts as Data Steward for archived data - Responsible for preserving the quality and integrity of data and documentation related to the retired system and ensuring compliance with relevant retention policies
Project Manager	R	Responsible for managing retirement, decommissioning, and disposal
Data Steward	R	Responsible for ensuring that good data integrity practices are followed during the planning and execution of the retirement, decommissioning and disposal phases
Where: R = Responsible, A = Accountable, C = Consult, I = Inform		

3.6.1 Data Considerations

Prior to beginning the retirement process, there needs to be a documented risk assessment to understand the concerns and difficulties to be addressed. The retirement process should be defined, and any transfer or migration process to be used validated, to ensure that no data or metadata is lost or compromised during the process. The records should be copied and verified as complete and accurate before the destruction of the original equipment. It is essential to ensure that the metadata, such as the audit trail or other information has transferred completely, as loss of such information permanently compromises the integrity of the copy of the record.

Transferring inactive records that have already been archived from the system into a secure off-line archive location may require additional steps, such as restoration back to the original application prior to conversion and transfer. This is common when the inactive records have been archived in the original existing format, which is likely to be vendor dependent. Therefore, it is critical to plan for this migration before decommissioning the original computerized system application.

If some of the data has reached the end of its retention period and can be discarded there should be a clearly defined process requiring data owner, process owner, and quality (and maybe legal) approvals prior to disposal. This requires the ability to segregate such data from the rest of the data without creating issues and that the date of data creation is stored with the record. Naming conventions that include the date of creation may be beneficial because when data is migrated, the file creation dates may be reset to the date of migration. Data destruction is discussed further in Section 7.4.

When a system reaches the end of life and is to be decommissioned, a retention strategy is needed for data that has not reached the end of its retention period. As discussed in Section 3.2.2, ideally, dynamic data should be maintained in a dynamic format, but sometimes this is not possible. If data is required to be retained for decades (e.g., traceability records for tissue donors, as contained in Appendix O4), there may come a point where a risk-based decision to save it in a static format may be justifiable (i.e., the need to sort or reprocess it is very unlikely).

For example, a risk assessment may find that there is little or no foreseeable need for reprocessing after 10 years, while retaining a computer (or even a virtual environment) with an unsupported operating system and application gets increasingly problematic. It is important to consider and document the risk of creating static records from the dynamic records [12] and to ensure that the records are still able to support the reconstruction of the activities performed and decisions made. This approach should not be considered for active records. This is further discussed in Section 7.2.

Information stored in PDF files may be substituted and transformed when transferred to a new server if the fonts were not embedded when the PDF was generated. Such substitution or transformation can alter the record permanently and may alter the meaning, such as when symbols are transformed or changed to standard fonts, rendering the information useless if it becomes unreadable. It is important to consider this during the initial transfer plan.

Another significant problem with PDF files is that they are limited to a static representation of the data and may not contain the complete and accurate record of what was originally recorded by an electronic system; this is extensively discussed in Section 7.2.1. Finally, PDF files also have very limited searching and indexing capabilities, and indexing within a document management system is needed to ensure the records can be found when needed.

Readability of the retained data depends on the format of the data. If the data has remained in a vendor-specific format, a functional copy of the legacy software may need to be retained to allow for data readability (as discussed in Section 3.2.2 and Appendix O5). Training and operating materials should be archived along with the legacy system to assist access by inexperienced personnel.

3.6.2 System Considerations

The scope of the system retirement strategy should include everything from simple devices and non-networked systems through to and including enterprise systems. Different scenarios relating to system retirement are described below. Note that for Step 1 these are broken down by system types, with common activities listed first, and system-specific activities listed under system type.

Step 1 – System Retirement

The system is removed from active operations, normal end user access is withdrawn, and interfaces with other systems are deactivated. SMEs involved in decommissioning activities are allowed access and permissions sufficient to perform their allotted tasks. During the development of the retirement plan, it is essential to consider all associated records, qualification data, calibration records, etc., as well as the regulated data and metadata required to be retained.

System Types

System not connected to a computer: This system generates electronic data but does not permanently store the data (titrator, osmometer, etc.). The retirement plan should document the archival and retrieval of the temporary data stored on the system. When retiring this system, no additional data will be generated from the system. Data typically generated from these systems is static in nature and therefore may be retained in printed or electronic format. These files are often saved in a read-only format (e.g., PDF) for ease of archival. Appendix D2 discusses some of the implications of managing the data during the use of these systems.

System connected to a non-networked computer: Systems are connected to the computer for system control and data acquisition. Data from these systems use temporary media to transfer data to a more permanent secure location. When retiring such a computerized system, no additional data will be generated from the system. Data typically generated from these systems is dynamic in nature. The retirement plan should document the archival and retrieval of the data stored on the system and permanent storage location. This plan should allow for data to be maintained in its original dynamic format. In most situations, it is necessary to retain a copy of the application software to retrieve data in its dynamic format.

System connected to a networked computer: These are systems connected to a computer for system control and data acquisition, where the computer is connected to a network allowing direct data transfer to a secure location. When retiring such a system, no additional data will be generated from the system. Data typically generated from these systems is dynamic in nature. The retirement plan should document the archival and retrieval of the data stored on the system. This plan should allow for data to be maintained in its original dynamic format. In most situations, it is necessary to retain a copy of the application software to retrieve data in its dynamic format should this be required. A virtualized copy of the system can be retained for this purpose until the end of the retention period is reached – see Section 3.2.2 and Appendix O5.

Enterprise Systems: Enterprise systems are typically used to meet the needs of multiple sites. When retiring an enterprise system, no additional data is added to the system from the specified retirement date. An enterprise system may be retired at a single site or across all sites:

- If only a specific site is retiring the system and it will remain in operational use at other sites, then data availability is simplified since the system is still active.
- It is more complicated if the whole enterprise system is to be retired from all sites as alternative arrangements will be needed to maintain data availability and readability.
- When one site pilots a new system, then there may be an interim negative impact on data quality and nomenclature across the organization, as the new system may store data differently and with an enhanced nomenclature compared to the existing system. This incompatibility continues until all sites have migrated to the new system. See Sections 1.6.4.2, 4.7, and Appendix M2 for further explanation of data quality and nomenclature.

The retirement plan should allow site data to be retrievable in its dynamic format with the permission of the designated data owner. An enterprise system typically remains in use until data is migrated to a new enterprise system or to a long-term archive. When retiring a complete enterprise system to move to a new one, no additional data is added to the system from this point forward. The retired system typically remains in use until the data stored is either migrated to the new system or has reached its record retention requirement.

Retirement of the Archive System: When an archive system is due for replacement (reaches the end of its lifecycle or an improved archiving solution is found), special care must be taken to preserve the integrity of the data throughout the process. If the archive system is connected to other systems, replacement interfaces need to be created to link to the new archive and validated to ensure they are transferring complete data. Retirement of an archive system requires all of the same considerations as retirement of any other type of system but is complicated by the diversity of the records and the possibility that the new archive system does not support all of the file formats. This may be the time to convert dynamic data to static data based on unsupported file formats. Refer to Section 7.2.1 for details about this conversion.

Step 2 – System Decommissioning

Decommissioning is the controlled shutdown of a retired system. A system may be stored if required to be reactivated at a later date, such as to retrieve regulatory data or results.

Step 3 – System Disposal

Data, documentation, software, or hardware can be permanently destroyed. Each may reach this stage at a different time. Data and documentation should not be disposed of until they have reached the end of the record retention period as specified in the record retention policy.

3.7 Records Management and Retention through Mergers, Acquisitions, and Divestments

During mergers, acquisitions, and divestments, it is essential to plan and define the approach to managing the impacted systems and data. A fixed timeframe may be imposed by legal contracts to review associated and impacted data as part of the transition period.

Some factors to consider for records management and retention include:

- What data is needed and impacted?
- What systems are involved, which will be kept and which will be retired?
- Is a new system purchase required as part of consolidation or divestment? If so, this will need validation for intended use before data can be imported.
- Can data generated by the same business process across the two companies be transferred into a single system, e.g., all batch record data can be merged within the chosen Manufacturing Execution System (MES)?
- How will both the active data from the live system and the inactive data from an archive system be managed?
- What legacy systems are needed to read the archived data and how will they be managed?
- Who will be the data owners at the end of the transition phase?
- There may be changes in the underlying business processes during the transition period; how will these impact the data flow ongoing?

For divestments, the transition period should be used to identify how to segregate data between the companies. An independent party may provide the data segregation process and provide a dataset to the recipient. In that situation, it is vital to identify the data requirements early, ensure that data segregation is performed at least twice, so there is an opportunity to rehearse data segregation and verify the resulting data set, and address any issues before the final cut of data is taken.

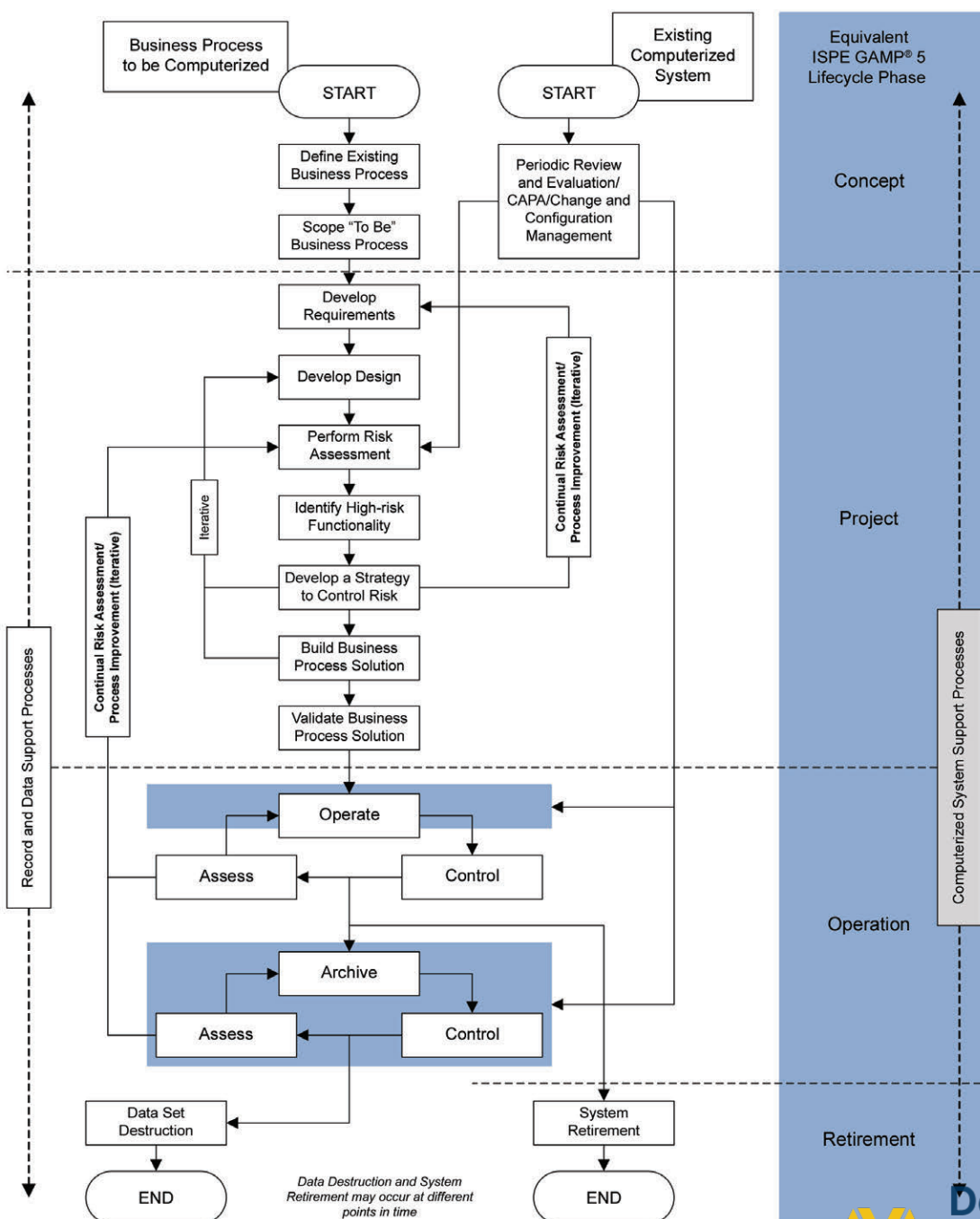
4 Implementing Data Integrity by Design

4.1 A Process to Achieve Data Integrity by Design

This section presents an approach to gaining a more systematic and enhanced understanding of the data set/data system under development with respect to effectively managing the data lifecycle for regulated records.

A process to achieve data integrity by design throughout the data lifecycle is presented schematically in Figure 4.1.

Figure 4.1: Data Integrity by Design Process Flow Diagram



This process supports the development, maintenance, and retirement of compliant data sets and data systems.

The data integrity by design process facilitates delivery of a design that ensures the computerized system(s) is/are fit for intended use with respect to the data lifecycle and the business process.

The key development related activities for a process to be computerized are described or referenced in more detail below.

- **Define Existing Business Process:** In addition to the process workflow it is critical to capture the business process and end to end data flows (incorporating data transfer across interfaces) in order to provide assurance that the full data lifecycle is understood, including high-risk areas and compliance gaps.
- **Scope “To Be” Business Process:** Use critical thinking to identify opportunities for end to end business process improvement, with a focus on prevention and detection opportunities to minimize existing risks to data integrity and data quality for regulated and other business-essential records.

Business process mapping is explained in detail in Section 4.2, while Section 4.3 introduces data flow diagrams. Once the business process map and data flow diagrams have been generated, a business process risk assessment (see Section 4.5) can be used to identify potential data integrity risks inherent in the process that must be addressed as far as possible by the computerized systems selected.

- **Develop Requirements:** Critical thinking should be applied to analyze the regulatory requirements for data integrity, and determine the necessary and most effective controls to meet the intended use. The intended use inherently impacts the controls needed; for example, a system intended for use in early stage drug discovery may have a somewhat different set of data integrity requirements than one intended for use in product release testing with potential direct impact on patient safety.

It is also important to realize that the computerized system is comprised of people, process, and technology, and therefore the requirements should inherently address not just the system technical controls but also the applicable data governance aspects (e.g., user access controls and segregation of duties) and the ability of the system to be configured to meet the intended use.

Sections 4.4, 4.6, and 4.7 provide guidance around defining the intended use of the data, the data lifecycle, and the importance of creating data nomenclature to be used throughout the process or organization.

Establish high-level functionality and data requirements for the “to be” business process and data lifecycle (see the *ISPE GAMP® RDI Good Practice Guide: Data Integrity – Key Concepts* Appendix 6 [16]). In purchased systems, the user requirements may be a combination of the business process requirements and the vendor specification.

The historical focus of a typical User Requirements Specification (URS) has been on the required functionality of the system to support the business process, including the interfaces to other systems and the ability to collate data from multiple systems into a cohesive data set. Data and data lifecycle considerations should be carefully considered and specified with equal rigor in order to ensure that good practice expectations for data quality and data integrity can be incorporated into the design for a new system at the earliest opportunity.

To properly develop the design, a more detailed process map and data flow diagrams may need to be developed at this stage in order to facilitate the identification of data integrity requirements. As part of data integrity by design, how the system is designed to be available to a user in their environment should be considered. For example, identifying the physical location of terminals and determining if they are sufficiently robust (e.g., to cleaning processes in the operating environment) and maintained so that the terminal is available to an operator for recording entries contemporaneously.

The preparation of the URS for a computerized system should indicate the priority of each requirement, for example, “must have,” “should have,” and “could have.” As part of this process, data integrity-specific requirements should be prioritized as a part of an iterative process of QRM (according to ICH Q9 (pharmaceuticals) [27], ISO 14971 (medical devices) [28], and *ISPE GAMP® 5* [9]) in order to assess each requirement’s potential to impact the quality and integrity of data within the system. Known data integrity trouble spots based on industry and company experience (e.g., interfaces) should be explicitly examined for potential risks.

Data-related requirements should be clearly cross-referenced to process step activities and data flow considerations, and to risks to public health, patient safety, and product quality. Requirements must be clear, correct, and unambiguous. Cross-references to specific regulatory requirements (such as the predicate rules of US GMP [29] as well as US FDA Guidance for Industry: Part 11 [30] and EU EudraLex Chapter 4 [31] and Annex 11 [11]) may also be helpful. An example of regulated record retention periods is contained in Appendix O4.

For computerized systems, data integrity requirements address availability, performance, resilience, security, deployment model/architecture, and stability. These requirements are technical or procedural and include specific controls such as second person verification, secure time stamp, compliant electronic signature, electronic audit trail, backup and restore, etc. There may be other data integrity related considerations for a particular system, for example, that static¹ records are searchable or that dynamic records are able to be reprocessed by duly authorized users.

Minimally, systems that generate static data in a flat-file format may need to be networked with centralized data storage to protect the data. An automated and validated mechanism to capture the flat files into a database system for increased security and traceability is desirable.

- **Develop Design:** Identify and evaluate any commercially available systems with the potential to meet the requirements. Where there is no commercially available solution, establish a design project to develop a system. The initial design of a purchased, commercially available system may be configuring a single system with a workstation or may include combining multiple systems with a workstation to automate a process. The automated system could be from a different vendor but configured to work with the same software, for example, connecting a dissolution bath, fraction collector, and a UV/Vis spectrophotometer. (Chapter 5 offers guidance on planning computerized systems.)

This design activity may be performed with help from the supplier or using internal support but should be presented to system owner’s Subject Matter Experts (SMEs) and quality representatives for review.

- **Perform Risk Assessment:** Data integrity focused risk assessment should be performed early in the project phase and repeated as more information becomes available and greater knowledge is obtained (see *ISPE GAMP® RDI Good Practice Guide: Data Integrity – Key Concepts*, Chapter 4 – Risk Management Approaches [16] for more information). System level risk assessments are covered in Section 5.3.
- **Identify High-Risk Functionality:** (see Section 5.3) – A “must have” requirement for data integrity is typically a property or characteristic of the developed system that assures data set/data system quality and data integrity within the business process. “Must have” requirements for data integrity are not automatically high risk. While “must have” status may infer a high severity of harm if the requirement is not met, the additional factors of likelihood of occurrence and probability of detection in the *ISPE GAMP® 5* [9] risk assessment methodology may result in an overall medium or low risk priority for the requirement.

Having identified the high-risk functionality, QRM can be used to prioritize subsequent activities.

- **Develop Strategy to Control Risk:** A strategy to control end to end process and data flow ensures that a data set of the required quality will be produced consistently. (This is discussed throughout the system lifecycle considerations in Chapters 6 and 7.)

¹ Static and dynamic records are defined and discussed in detail in Section 7.2.1.

The elements of the control strategy should describe and justify how the chosen configuration and technical and procedural controls contribute to the final quality and integrity of the data output. These controls should be based on product and process understanding as defined in the key concepts of *ISPE GAMP® 5* [9]. A detailed understanding of the system should be supported by the quality system, such as incident management processes. The quality system requires adequate tools to support critical thinking to assess the impact of these types of events (and evolves as more knowledge of the system is gained).

Sources of process variability, such as manual intervention, that can impact patient safety, product quality, and data integrity should be identified as risks, appropriately understood, and subsequently controlled. Understanding sources of variability and their impact on the data within the system can provide an opportunity to implement additional or alternative controls that reduce the risks. Science-based product and process understanding, in combination with QRM, supports the control of data sets and data systems such that any variability can be compensated for in an adaptable manner to deliver consistent data quality.

The adoption and implementation of innovative technologies and control paradigms, for example, process analytical technology, may allow the design of an adaptive process step (a step that is responsive to input) with appropriate process controls to ensure consistent product quality and data integrity.

Enhanced understanding of the data system and data set performance can justify the use of alternative approaches to determine that data is fit for purpose.

Continuing to work through Figure 4.1:

- **Build Business Process Solution:** It is uncommon now for systems to be developed as custom solutions. Build process solution more typically represents the application of a selected system configuration to meet intended use and the provision of supporting procedures for that use. In the case of some SaaS solutions, configuration may not be available, in which case build solution is limited to the implementation of use procedures.

- **Validate Business Process Solution**

Supporting processes include:

- **Knowledge Management** (see Appendix M1)
- **Supplier and System Selection** (see *ISPE GAMP® 5*, Appendix M2 – Supplier Assessment [9])

With the move to more purchased and built solutions, including SaaS, benefits can be gained by collaboratively working with the supplier to improve the technical controls built into the system to prevent and detect data integrity issues.

- **Managing Continual Risk Assessment and Process Improvement throughout the Data Lifecycle:** (see Figure 4.1) Throughout the data lifecycle, companies have opportunities to evaluate innovative approaches to improve data quality and integrity (see ICH Q10 [32]). This includes re-evaluating controls to determine if additional controls or adjustments to those controls are needed. Even after risk mitigation there may be a level of residual risk that should be periodically reviewed and reassessed.

In the operational phase of the system, process performance can be monitored to ensure that it is working as anticipated to maintain data integrity as expected. This monitoring includes:

- System periodic review
- Procedures and instructions
- Roles and responsibilities (RACI)

- Record media, e.g., electronic or paper
- Critical aspects of the process impacting patient safety, product quality, and data integrity
- Process step parameters, including changes and the cause of those changes

Upon gaining additional process knowledge and performance monitoring data, the data integrity-specific configuration of the computerized system and/or control elements can be revised accordingly.

It should be noted that each of the data lifecycle processes may require a different focus of design such that appropriate controls are in place. These controls may vary for different records based on data classification and intended use. For example, in-process checks as part of a batch record need to be retained until the batch expiration date plus 1 year, compared with biological tissue traceability records [33, 34], which need to be retained for at least 30 years. The design of the retention solution for these two examples is likely to be very different. A more detailed listing of retention periods is contained in Appendix O4.

4.2 Business Process

The ability to map the business process is important to enable understanding of the business activities, and decision points in a business process. The business process mapping or modeling may be depicted as flowcharts or tables, or in combination. If a table is used to describe the business process, the location (where), responsible person/role (who), the proper time to perform the action (when), and the output(s) of the action (what) should be included, depending upon the needs [8]. Business process mapping illustrates the steps performed to fulfill the business purpose, and should include both manual and computerized system aspects of the process.

As shown in the *ISPE GAMP® RDI Good Practice Guide: Data Integrity – Key Concepts*, Section 3.2 [16], a business process can be described in very simple terms using a block diagram. A single block can involve one or many manual operation(s) or computerized system(s). The level of granularity of the process map is dependent upon the level required for the business to identify the associated risks to data integrity, product quality, and patient safety.

Business process maps are useful to help the business identify the risks associated with the use of the system including data integrity risks. Depending upon the needs, the process map may include details such as:

- Sequence of process steps
- Location of the activity
- Time of the activity
- Critical decision points, including actions (e.g., review/approval/disposition) mandated under the predicate rules
- System functions
- Responsible person/roles
- System interfaces
- Infrastructure including servers for production, application, database
- Inputs of the activity
- Outputs of the activity

- Risks associated with the process
- Documentation involved in the process (used and generated)

If any parts of the process are outsourced, it is important to include this information in the process map. Data created by an outsourced facility remains the responsibility of the marketing authorization company, and, as described in the *ISPE GAMP® RDI Good Practice Guide: Data Integrity – Key Concepts* [16], needs to be available for review, such as during a clinical trial. A process map used to help IT support the system should contain information about the hardware and servers supporting the system.

The process map can help with the creation of SOPs, work instructions, and training materials before the operational phase. It is also an essential basis for the end to end testing of the business process before the system is released for use.

4.3 Data Flow Diagrams

It is essential for a data flow diagram to capture the level of detail that will truly identify all data activities and the subsequent identification of their potential risks. For example, the use of a data flow diagram incorporating ancillary systems could identify the use of unvalidated and uncontrolled spreadsheet to evaluate specification limits. However, a separate data flow diagram may not be needed for a very simple business process.

Depending on the complexity of the computerized system involved and/or the business model (e.g., outsourcing services or systems), the data flow could be simple or very complex. It is crucial not to overcomplicate the data flow as this can inhibit the clarity and understanding of the diagram. The more controls and interfaces in place, the more complex the data flow diagram becomes.

A data flow diagram can be created by identifying for each step in the business process map:

- From where will information come into the process step (e.g., sample or batch ID, limits and specifications) and how (e.g., electronic interface/manual transcription)
- What data will be generated during that step of the process and how is it captured? What calculations or processing must be performed on the data and how and where will those be done? What metadata is needed to provide the GxP content and meaning to the data?
- For manufacturing processes, what are the Critical Process Parameters (CPPs) and Critical Quality Attributes (CQAs), and how are they managed and fed into the process control?
- Where does the data need to go next (e.g., to the next system used in the process, or to a laboratory information system or Enterprise Resource Planning (ERP) system, or both)? How will it be transferred? How is the transfer verified? What data cannot be transferred? How and where is it stored?

Figure 4.2 shows a simplified outline from which to build a data flow diagram. Split out all the data sources and downstream systems/processes, and add detail on the data being transferred.